REFORMA DE LA EDUCACIÓN MATEMÁTICA EN COSTA RICA

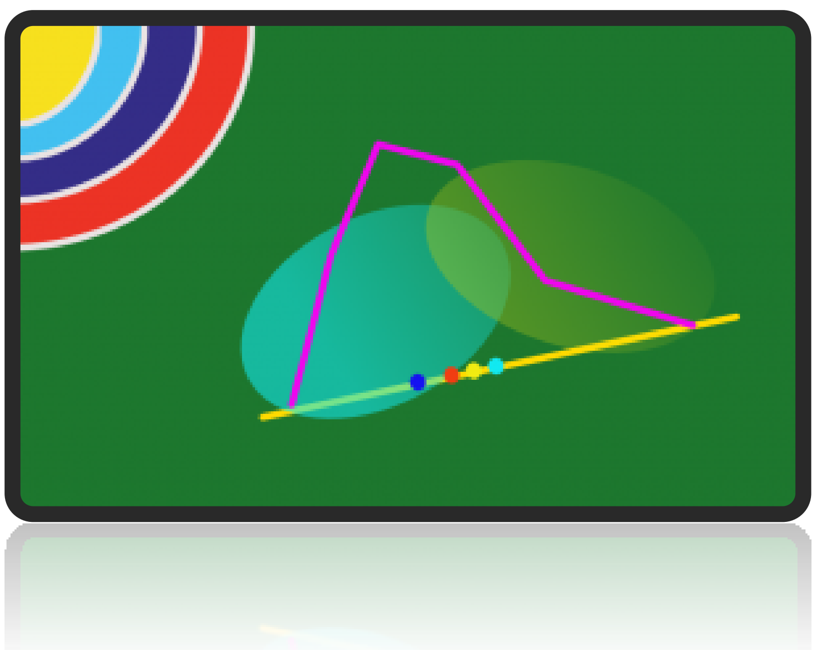
mep
Ministerio de Educación Pública

www.reformamatematica.net

Curso MATΣPJA en línea

ESTADÍSTICA Y PROBABILIDAD

Material de consulta



Costa Rica
2020

Contenido

Índice de conceptos

I. Introducción

El Este documento fue elaborado por el Proyecto Reforma de la Educación Matemática en Costa Rica (www.reformamatematica.net).

La primera versión de este documento fue ofrecida en el 2016 como apoyo a un curso virtual con la modalidad MOOC denominado Preparación Matemáticas Bachillerato. Varias de sus partes también fueron usadas en 2017 y 2018 para cursos Mini-MOOC ofrecidos por el Proyecto. La presente versión está dirigida especialmente a apoyar el curso Curso MAT∑PJA en línea ofrecido por el Ministerio de Educación Pública para estudiantes que desean obtener su Bachillerato por Madurez.

Se describen diferentes conocimientos vinculados al uso de las medidas estadísticas, de posición y variabilidad, y a conceptos básicos de probabilidad para la resolución de problemas; de acuerdo con las temáticas incluidadas en los Programas de Estudios de Matemáticas para la Educación Diversificada.

Al inicio del documento se le proporciona un índice alfabético en el que se da un listado, en orden alfabético, de los temas o contenidos con el número de página donde aparecen. Si usted hace clic sobre dicho número, será remitido a la página donde se proporciona el concepto, tema o contenido correspondiente. Se puede regresar al índice alfabético desde cualquier página haciendo clic sobre la palabra Índice que aparece en el encabezado de todas ellas.

Es importante aclarar que el presente documento no es un libro de texto y tampoco es exhaustivo. Procuramos que sea autosuficiente pero no está pensado para ser utilizado como un medio para organizar la acción de aula.

Más materiales educativos se pueden acceder en el sitio web Recursos Libres de Matemáticas (https://recursoslibres.reformamatematica.net).

II. Estadística

1. Conceptos previos:

Considere el siguiente ejemplo:

El Instituto Nacional de Estadísticas y Censos (INEC) es la institución oficial que tiene a cargo la elaboración de censos y encuestas nacionales, entre otras funciones que realiza. Periódicamente efectúa la denominada “Encuesta de ingresos y gastos”. A continuación se presentan las preguntas incluidas en uno de los módulos o secciones de esta encuesta que se denomina “Características de la vivienda”:

A. Características de la vivienda

Reg. 2

1. Tipo de vivienda

- Casa independiente .... ○1
- En fila o contigua .... ○2
- En edificio o apartamento .... ○3
- Tugurio .... ○4
- Otro ________ especifique ○5

2. ¿Esta vivienda es...

- ...propia, totalmente pagada? .... ○1
- ...propia, pero la están pagando? .... ○2
- ...propia, por regalo o donación? .... ○3
- ...alquilada? .... ○4
- ...cedida o prestada por trabajo? .... ○5
- ...cedida o prestada por familiar, amigos u otra persona? .... ○6
- ...precario? .... ○7
- Otro tipo ________ especifique ○8

pase a 5

3. ¿Cuántos metros cuadrados de construcción tiene esta vivienda?

- Menos de 30 m^2 .... ○1
- De 30 a 40 m^2 .... ○2
- De 41 a 60 m^2 .... ○3
- De 61 a 100 m^2 .... ○4
- De 101 a 150 m^2 .... ○5
- De 151 a 200 m^2 .... ○6
- Más de 200 m^2 .... ○7

4. ¿Hace cuántos años fue construida esta vivienda?

- Menos de un año .... ○1
- De uno a menos de 5 años .... ○2
- De 5 a menos de 10 años .... ○3
- De 10 a menos de 20 años .... ○4
- De 20 a menos de 30 años .... ○5
- Hace 30 años o más .... ○6

5. ¿En esta vivienda se desarrolla alguna actividad que proporcione ingresos por...

- ...comercio? .... ○1
- ...artesanía? .... ○2
- ...industria? .... ○3
- ...servicios? .... ○4
- Ninguna .... ○5

6. ¿Cuál es el material predominante en las paredes exteriores?

- Block o ladrillo .... ○1
- Zócalo .... ○2
- Madera .... ○3
- Prefabricado .... ○4
- Otro ________ especifique ○5
- Material de desecho .... ○6

7. ¿Cuál es el material predominante en el techo?

- Lámina de metal o zinc .... ○1
- Fibrocemento .... ○2
- Entrepiso .... ○3
- Otro ________ especifique ○4
- Material de desecho .... ○5

8. ¿La vivienda tiene cielo raso?

Sí ...... ○1 No ...... ○2

9. ¿Cuál es el material predominante en el piso?

- Mosaico, cerámica, terrazo .... ○1
- Cemento (lujado o no) .... ○2
- Madera .... ○3
- Otro ________ especifique ○4
- No tiene (piso de tierra) .... ○5

10. ¿Cómo es el estado de...

	Malo	Regular	Bueno
...las paredes exteriores?	○1	○2	○3
...el techo?	○1	○2	○3
...el piso?	○1	○2	○3

11. ¿Esta vivienda se abastece de agua por...

- ...tubería dentro de la vivienda? .... ○1
- ...tubería fuera de la vivienda? .... ○2
- No tiene agua por tubería .... ○3

12. ¿El agua que consume proviene de...

- ...un acueducto del A y A? .... ○1
- ...un acueducto municipal? .... ○2
- ...un acueducto rural? .... ○3
- ...una empresa o cooperativa? .... ○4
- ...un pozo? .... ○5
- ...un río, quebrada o naciente? .... ○6
- ...otro ________ especifique ○7

13. ¿Esta vivienda tiene servicio sanitario...

- ...conectado a alcantarilla o cloaca? .... ○1
- ...conectado a tanque séptico? .... ○2
- ...de pozo negro o letrina? .... ○3
- Otro sistema? ________ especifique ○4
- No tiene .... ○5

14. ¿Esta vivienda tiene baño?

Sí ..... ○1 ¿Cuántos? ____ No ..... ○2

http://www.inec.go.cr/

En la encuesta anterior se mencionan diferentes conceptos estadísticos que van a ser citados en el presente documento.

Unidad elemental o unidad estadística

En todo estudio estadístico se requiere tener definida la unidad básica que va a proporcionar la información necesaria para resolver el problema. Tal como se evidencia en el ejemplo anterior, la información recolectada proviene de los hogares, entonces son ellos las unidades estadísticas de interés.

La unidad elemental o unidad estadística puede corresponder a personas, animales, objetos, hogares, entre muchas otras posibilidades y está en función del estudio o problema que se desea realizar.

Características o variables

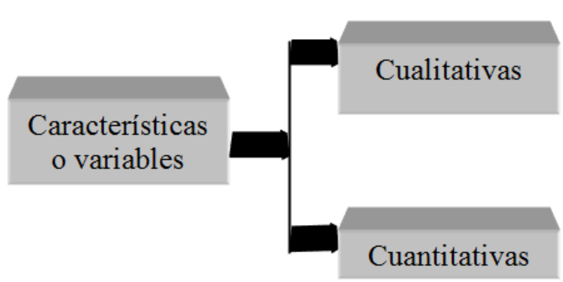
En una situación o estudio estadístico en particular, una vez que se ha definido la unidad estadística, se deben establecer las particularidades que son de interés para el estudio, es decir, todos aquellos aspectos que caracterizan a la unidad elemental y que se relacionen con el propósito del estudio.

En el ejemplo anterior, dentro de la encuesta se incluyen varias características de la vivienda que son objeto de estudio: tipo de vivienda, condición de la vivienda, metros cuadrados de construcción, entre otros.

Volver al índice de conceptos

Observación o dato

En el ejemplo previo, al momento de aplicar el cuestionario, cada uno de los hogares de la muestra seleccionada, proporciona información para las variables consideradas en el estudio, estos valores se denominan observaciones o datos. Es decir un dato representa el valor numérico, la cualidad o categoría que se obtiene de una unidad estadística para una variable en particular. Por lo anterior, para cada característica o variable de interés en un estudio, los datos estadísticos son conjuntos de números o de categorías correspondientes a las observaciones o respuestas obtenidas en el análisis de las diferentes unidades estadísticas incluidas en el estudio.



Variables cuantitativas

Corresponden a aquellas características que generan datos u observaciones que involucran números. A manera de ejemplo, algunas variables cuantitativas pueden ser: número de personas por hogar, tamaño del lote donde se encuentra la vivienda, entre otras.

Variables cualitativas o categóricas

Corresponden a aquellas características que generan datos u observaciones cualitativas. Por ejemplo, material predominante en el techo de la vivienda (mosaico, cerámica, terrazo, cemento, madera, otro); estado de las paredes (malo, regular bueno).

A las variables cualitativas se les acostumbra llamar *variables categóricas*, pues los datos que generan pertenecen a diferentes categorías. Por ejemplo, cada uno de los posibles materiales predominantes en el techo se les llama categorías.

En algunos casos las variables cuantitativas se agrupan en clases y pueden ser analizadas en forma categórica.

Población

Por lo que se ha venido señalando, un estudio de naturaleza estadística involucra una cantidad grande de unidades estadísticas; todas ellas son objeto de estudio. En este sentido, la totalidad de unidades estadísticas recibe el nombre de población en estudio. En el ejemplo de las características de la vivienda, la población está constituida por todos los hogares del país.

En general, una población puede ser finita, infinita o indeterminada.

- Un ejemplo de un caso de población finita corresponde a la población de viviendas de un cantón.

- Un ejemplo de población indeterminada se presenta en la siguiente situación: una empresa que fabrica enlatados de palmito desea analizar si la cantidad de producto por enlatado se ajusta a lo que se incluye en la etiqueta, que son 500 ml. El estudio es urgente pues se ha presentado una demanda, donde se afirma que están vendiendo menos producto del que se supone. Debido a que la empresa no puede parar el proceso de producción para hacer el estudio, las unidades estadísticas son los enlatados producidos y la variable de interés es la cantidad de palmito por enlatado, resulta imposible determinar la cantidad total de enlatados, entonces se dice que la población es indeterminada.

- Un ejemplo de una población infinita se puede observar en la siguiente situación hipotética: con la intención de determinar el uso que se le da al suelo en el país, se ha decidido seleccionar aleatoriamente diferentes puntos de coordenadas (x,y) (x: latitud, y: longitud) sobre el territorio nacional e identificar el uso que se le da a la tierra en ese punto particular.

Aunque el problema es hipotético, puede notarse que la unidad estadística es un punto de coordenadas (x,y) sobre el territorio nacional. Debido a que este punto se elige dentro del continuo de puntos territoriales de Costa Rica, la población de interés es entonces infinita.

Muestra

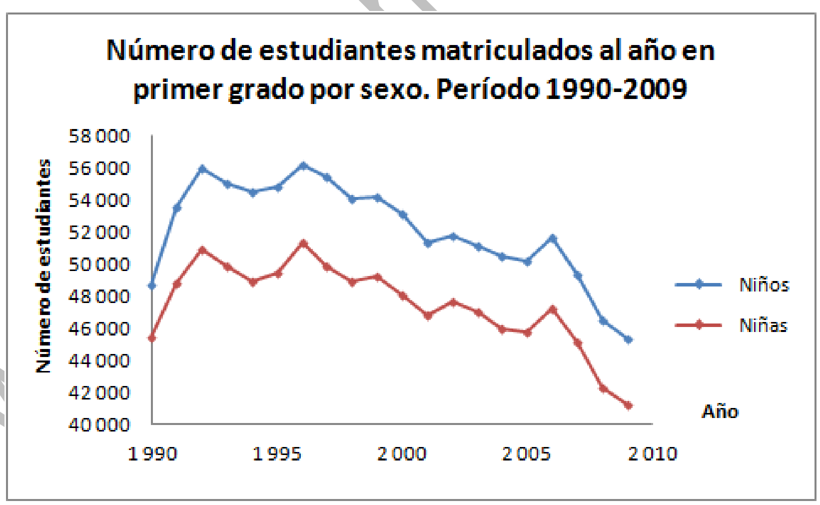
Una muestra corresponde un subconjunto de la población que se selecciona para realizar el estudio estadístico de interés. Normalmente resulta imposible analizar toda la población, debido a que es muy grande, infinita o indeterminada. Por ejemplo, si se desea realizar un estudio de las viviendas de Costa Rica, ante la dificultad de visitar todas las viviendas del país, se puede seleccionar una muestra de ellas. Se dice que una muestra es aleatoria, si para seleccionar los elementos se utiliza algún procedimiento aleatorio que posibilita que todos los elementos de la población tienen posibilidad de ser seleccionados.

Volver al índice de conceptos

2. Análisis de información resumida en representaciones estadísticas

Problema 1. Disminución de matrícula escolar

La matrícula anual de estudiantes en primer grado ha disminuido en los últimos años, la siguiente gráfica refleja ese comportamiento.



Fuente: http://www.estadonacion.or.cr/index.php/estadisticas/costa-rica/compendio-estadistico/estadisticas-sociales

Además, se observa una importante diferencia entre la matrícula de hombres y de mujeres, la cual se mantiene durante todo el período. Sin embargo, el director de una pequeña escuela rural indica que la matrícula en primer grado ha venido aumentando en los últimos años, y normalmente se matriculan más niñas que niños. Considere las siguientes proposiciones que se elaboraron de acuerdo con la información que suministra del gráfico. Identifique aquellas que son verdaderas:

I. La información del director es falsa, pues es de esperar que los resultados generales de la gráfica deben reflejarse en cada una de las escuelas.

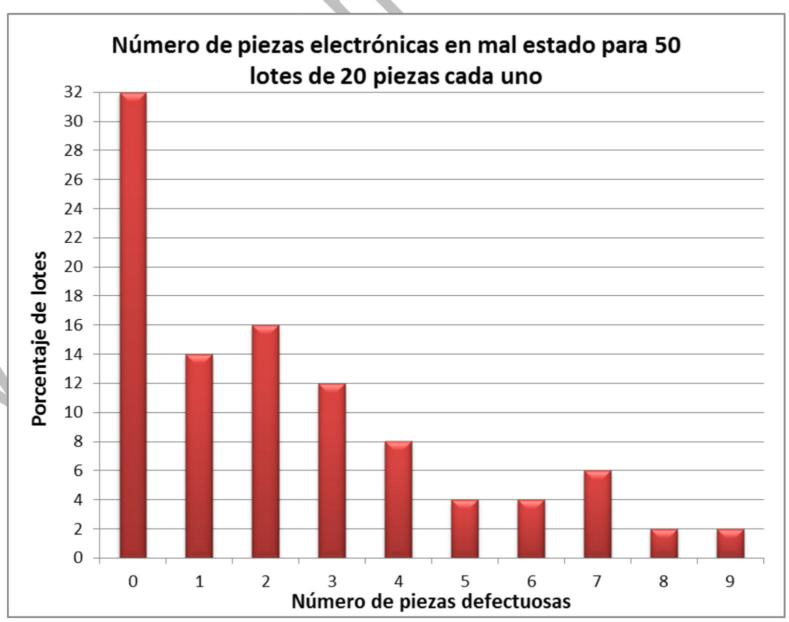
II. Los resultados proporcionados por el director reflejan que la información de la gráfica debe estar incorrecta.

III. Los resultados proporcionados por el director de la escuela no contradicen la información de la gráfica.

Solución
La representación gráfica resume el comportamiento general del número de estudiantes matriculados en primer grado de primaria en el país según sexo, lo cual corresponde al comportamiento de la población total del país. Una pequeña escuela rural representa un caso específico o una muestra particular, por lo que puede no cumplir con el patrón representado en la población. De acuerdo con esto, los resultados proporcionados por el director de la escuela no contradicen la información de la gráfica. Por ello, la proposición III es la única correcta.

Problema 2. Producción de piezas electrónicas

Una fábrica empaca en lotes de 200 unidades las piezas electrónicas que produce. Se establece un plan de inspección por muestreo que consiste en examinar 20 piezas de cada lote elegidas al azar. Se rechaza el lote completo si de las 20 piezas aparecen tres o más defectuosos. Se revisan 50 lotes y la distribución del número de piezas defectuosas por lote se muestra en la siguiente gráfica:



Determine el número de lotes que deberían ser rechazados.

Solución:

En la representación gráfica se muestra el comportamiento de 50 lotes de piezas electrónicas. En cada uno de ellos se han seleccionado 20 piezas aleatoriamente y se determina si cada una es defectuosa o no lo es. Debido a que se rechazan aquellos lotes para los cuales hay tres o más piezas defectuosas, del gráfico se extrae la siguiente información:

Número de piezas defectuosas	Porcentaje de lotes
0	32
1	14
2	16
3	12
4	8
5	4
6	4
7	6
8	2
9	2
Total	**100%**

Los valores en azul representan los porcentajes de lotes que deberían ser rechazados, esto quiere decir que el 38% de los lotes tienen tres o más piezas defectuosas. Entonces de los 50 lotes, el 38% debería ser rechazado, esto equivale a 19 lotes en total.

Modelos simétricos y asimétricos

En el análisis de los problemas anteriores, interpretar la forma o patrón de variación de los datos fue clave para dar respuesta a los cuestionamientos planteados. La distribución de los datos puede tomar diversas formas, pero en términos generales existen tres patrones básicos (aunque pueden existir otros):

Modelo simétrico   Modelo de asimetría positiva   Modelo de asimetría negativa

Cuando la forma de la distribución de los datos es simétrica o aproximadamente simétrica, significa que la mayor cantidad de datos se ubica en el centro, con un comportamiento similar a ambos lados. Mientras que en las distribuciones asimétricas las mayores frecuencias de datos tienden a concentrarse hacia uno de los extremos. En ellas se observa la presencia de pocos valores en la dirección contraria, pero que tienen un importante efecto en el mensaje de los datos, tal como veremos más adelante.

Volver al índice de conceptos

Problema 3. Vida útil de los bombillos

Suponga que le envían a comprar unos bombillos a una ferretería, la persona que le atiende indica que de acuerdo con las características del bombillo que usted está solicitando, tiene dos marcas con el mismo precio: A y B. Le pregunta que cuál de ellas desea. Usted recuerda que en un artículo de una revista que leyó recientemente salió publicado la siguiente representación gráfica sobre la vida útil de estos bombillos:

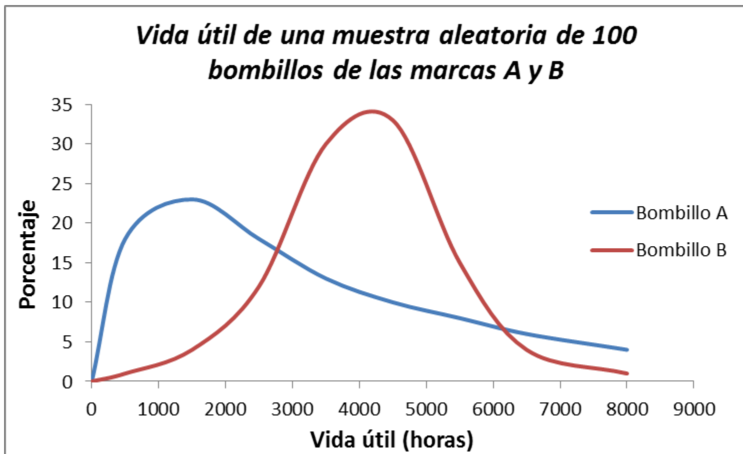


Imagen propia elaborada con fines didácticos

Utilice esta información para decidir qué marca de bombillos compraría y argumente esta decisión.

Solución:

Cuando se compra un bombillo, como con cualquier otro producto electrónico no tenemos certeza de cuánto tiempo va a funcionar, esta duración o vida útil varía de un bombillo a otro. Para resumir la variabilidad generada entre los bombillos, se pueden utilizar diferentes herramientas estadísticas. La representación gráfica es una de las herramientas que permite modelar la variabilidad en la vida útil de los bombillos y que es muy utilizada en diferentes estudios.

La representación gráfica anterior corresponde a una aproximación de este modelo y se basa en muestras aleatorias de 100 bombillos de los tipos A y B. En el eje "*X*" se ha representado la vida útil (duración) y en el eje "*Y*" el porcentaje de bombillos. El área que está encerrada bajo las curvas representa el porcentaje de bombillos que tienen la vida útil correspondiente, entonces en los intervalos donde hay más área significa que hay un mayor porcentaje de bombillos que tiene una vida útil en ese rango.

Por ejemplo, se presentó un mayor porcentaje de bombillos de la marca A que tienen una vida útil menor de 3000 horas, con respecto a los bombillos de la marca B. Asimismo un mayor porcentaje de bombillos de la marca B tuvo una vida útil entre 3000 y 6000 horas respecto a la marca A. También se observa que hay mayor porcentaje de bombillos de la marca A que duran más de 6000 horas.

De acuerdo con lo anterior, debido a que no podemos estar seguros de cuánto nos va a durar de un bombillo que vamos a comprar, entonces deberíamos adquirir bombillos de la marca B, debido a que según la representación gráfica anterior hay más posibilidad de que estos bombillos nos duren más.

Hay que tener claro que el análisis anterior se fundamenta en el hecho de que la información de estas muestras debe ser una buena representación de lo que ocurre en la producción total de bombillos.

Problema 4. Selección de adolescentes

Suponga que un asistente médico requiere hacer un estudio con adolescentes (jóvenes entre 12 y 17 años) en una clínica de salud. Debe escoger entre dos clínicas A y B, pero no sabe en cuál de ellas tiene más opciones de seleccionar adolescentes. Solicita colaboración a la administración de las clínicas para que le indiquen la edad (en años cumplidos) de los menores que asistieron a la clínica en la última semana. Debido a que las personas encargadas de proporcionar la información les gustan las representaciones estadísticas, en vez de darle los datos en forma simple, se pusieron de acuerdo para proporcionar la siguiente información al asistente médico.

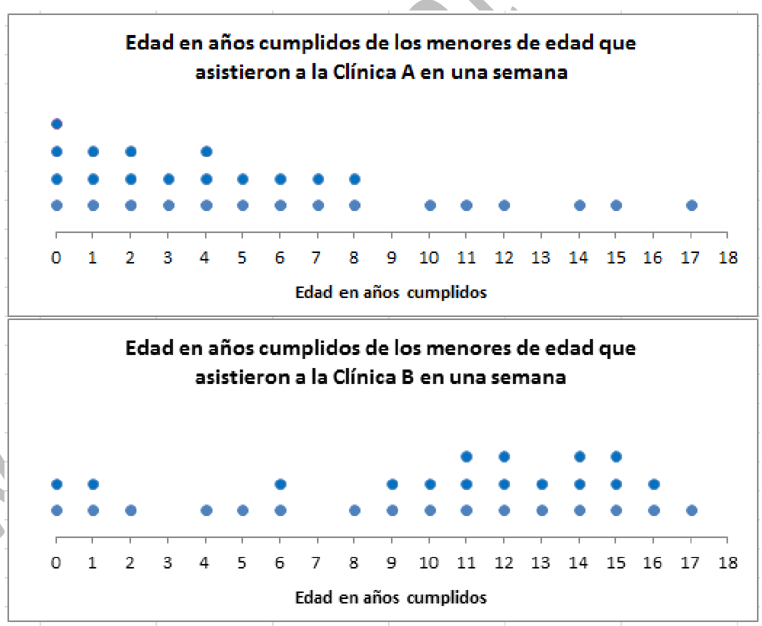


Imagen propia elaborada con fines didácticos

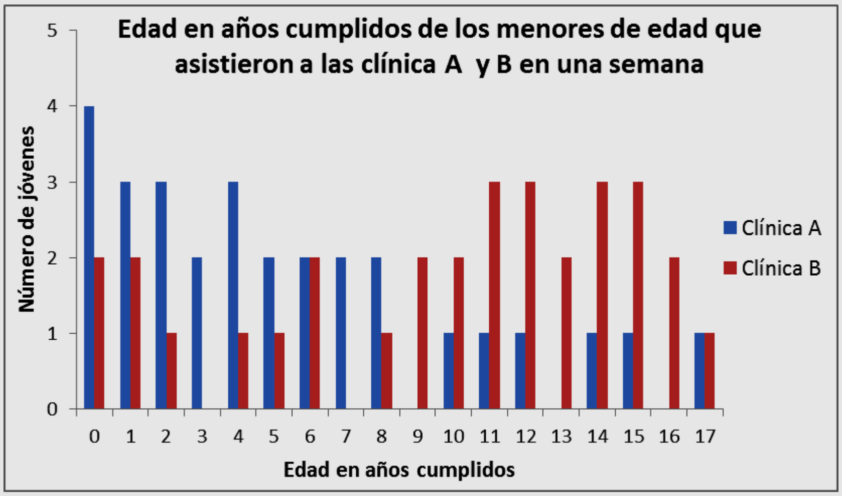
Ayude al asistente médico a tomar una decisión respecto a la clínica en la que debería realizar el estudio.

Solución:

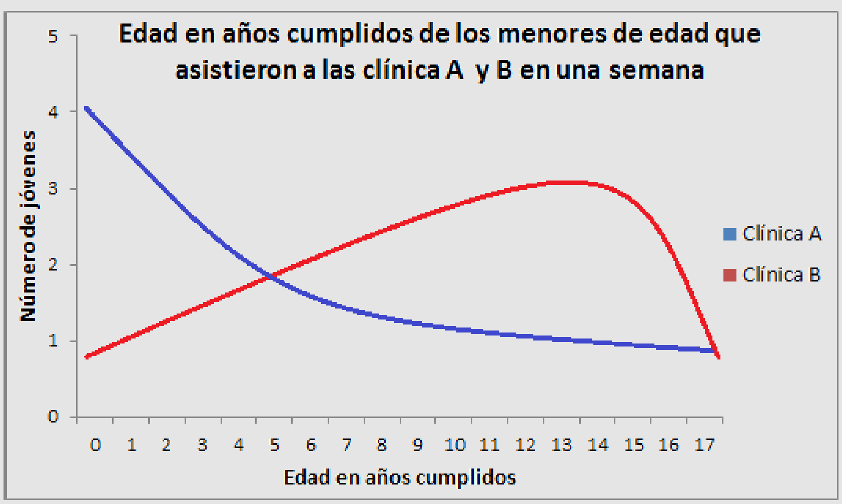
Esta representación se conoce como *diagrama de puntos*: los puntos representan la edad de los jóvenes y el número de puntos en cada edad (frecuencia) representa el número de jóvenes que tienen dicha edad.

De acuerdo con la información recabada en esa semana, el número de adolescentes fue mucho mayor en la clínica B; por su parte, en la clínica A se tuvo una mayor presencia de niños menores de ocho años. Por lo tanto, se debe indicar al asistente médico que, si lo acontecido en esa semana es un reflejo de lo que ocurre regularmente en las clínicas, el estudio debería realizarlo en la clínica B.

Nota: En vez de un diagrama de puntos, la información se pudo haber resumido mediante gráficas de barras, tal como se indica a continuación.



Debido a que el número de jóvenes menores de 18 analizados en las dos clínicas es diferente (29 en la clínica A y 31 en la B), entonces no se deben comparar las barras individuales, para responder el problema debe observarse el patrón de la variabilidad en los datos.



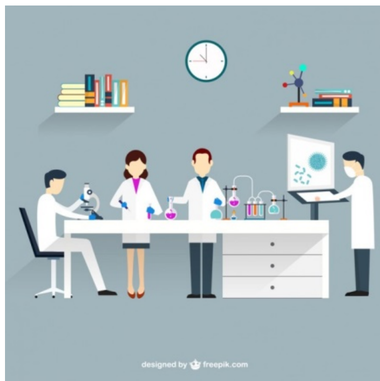
La visualización de estos patrones permite identificar que en esa semana se presentaron más adolescentes en la clínica B que en la clínica A.

3. Análisis de datos cualitativos: valores absolutos y porcentuales

Problema 5. Deserción y selección de área científica

Suponga que usted estudia en el Liceo El Progreso, que es una institución académica en la cual en el año 2015 se matricularon 192 estudiantes en undécimo año, pero 12 de ellos se retiraron antes de concluir el año académico, los demás aprobaron el curso lectivo.

De ellos, 18 se inscribieron para presentar Física como área científica para las pruebas de bachillerato, mientras 46 optaron por realizar la prueba de Química y el resto se inscribieron para presentar Biología. El director del colegio desea resumir esta información de manera que sea comprensible para entregar un informe a los padres de familia.



http://www.freepik.es/

Ayude al señor Director a resumir esta información para que la incluya en dicho informe y resulte comprensible para los padres de familia.

Solución:

En este caso la unidad estadística es el estudiante de noveno año del Liceo el Progreso, las variables de interés del estudio son:

- *Condición de desertor*, cuyas respuestas pueden ser: (1) desertor o (2) no desertor
- *Especialidad científica* que va a presentar en bachillerato: (1) Física, (2) Química y (3) Biología.

Los datos que generan estas variables no son numéricos, debido a que para cada una se anota lo siguiente. Observe que se pueden utilizar números para representar las categorías o respuestas vinculadas con cada estudiante, pero es meramente simbólico, cualquier número que se puede emplear se utiliza para simplificar la escritura.

En primer lugar, cuando se analiza la condición de desertor solamente caben dos posibilidades para cada uno de los estudiantes matriculados (desertor o no desertor), entonces la descripción del comportamiento es sencilla, basta con determinar el número de estudiantes que abandonaron la institución, en este caso fueron 12, por lo que 180 no se retiraron y aprobaron el curso. Sin embargo, estos datos por sí solos no son suficientemente ilustrativos desde el punto de vista estadístico. Se requiere determinar un valor relativo o porcentual que permita hacer comparaciones con otras instituciones o con otros años. Entonces el Director podría utilizar porcentajes en el informe:

- *En el año 2015, en el Liceo El Progreso desertaron 12 estudiantes de un total de 192 que matricularon undécimo año, por ello el porcentaje de deserción fue* $\frac{12}{192} \cdot 100\% = 6{,}25\%$. *Esto es equivalente a decir que de los 192 estudiantes que se matricularon en undécimo año en esta institución, el 93,75% logró aprobar el año.*

En segundo lugar, en cuanto al área de ciencias que escogieron los estudiantes, de los 180 jóvenes con derecho a realizar las pruebas de bachillerato, 18 seleccionaron Física, 46 Química y por lo tanto $180 - 18 - 46 = 116$ escogieron Biología. Para dar una mejor interpretación a estos números se pueden determinar los porcentajes correspondientes. Se recomienda al Director indicar en el informe lo siguiente:

- *En el Liceo El Progreso para el año 2015, de los 180 estudiantes que debían realizar las pruebas de bachillerato, 18 estudiantes seleccionaron el área de Física es decir el* $\frac{18}{180} \cdot 100\% = 10{,}0\%$, *por su parte 46 estudiantes seleccionaron el área de Química que corresponde al* $\frac{46}{180} \cdot 100\% = 25{,}6\%$, *(aproximadamente), y finalmente, 116 estudiantes seleccionaron el área de Biología, que representa el* $\frac{116}{180} \cdot 100\% = 64{,}4\%$ *(aproximadamente)*

Otra forma de presentar esta información corresponde a un pequeño cuadro:

Liceo El Progreso: área científica seleccionada por los 180 estudiantes que realizaron la prueba de Bachillerato, año 2015

Área científica seleccionada	Número de estudiantes	Porcentaje de estudiantes
Física	18	10,0
Química	46	25,6
Biología	116	64,4
Total	**180**	**100,0**

Información ficticia utilizada con fines didácticos

En el problema anterior, aunque los datos originales eran cualitativos (no numéricos), se debió realizar un conteo de cada categoría lo que a la postre generó la suma total en cada caso y los porcentajes correspondientes, de modo que la variabilidad de los datos fuera comprensible para un lector. En estos casos los porcentajes constituyen una medida estadística adecuada para resumir la información. Esta alternativa se vuelve fundamental cuando se requiere realizar comparaciones entre grupos con diferentes variables.

Problema 6. Selección de área científica según sexo del estudiante

Con base en el problema anterior, suponga ahora que el Director del Liceo El Progreso desea presentar la información sobre el área científica seleccionada por los estudiantes; pero desea realizar una comparación por sexo (hombres y mujeres)

www.freepik.es

Decide presentar en el informe el siguiente cuadro:

Liceo El Progreso: área científica seleccionada por los 180 estudiantes que realizaron la prueba de Bachillerato según el sexo, año 2015

Área científica seleccionada	Número de estudiantes	
	Hombres	Mujeres
Física	9	9
Química	16	30
Biología	37	79
Total	**62**	**118**

Información ficticia utilizada con fines didácticos

¿Qué críticas haría usted al señor Director en relación con el cuadro anterior, si lo que se pretende es realizar una comparación por sexo en la escogencia del área científica, de modo que sea de fácil comprensión para los padres de familia?

Solución:

Cuando se desean realizar comparaciones entre grupos, no es conveniente utilizar los números absolutos, debido a que normalmente ocurre que los números totales entre los grupos son diferentes (en el ejemplo anterior hay 62 varones contra 118 mujeres), entonces los valores parciales de cada categoría no son comparables. Por ejemplo, sería un error indicar que la preferencia por el área de Física fue igual entre hombres y mujeres basados únicamente en el hecho que nueve hombres y nueve mujeres seleccionaron esta área. Pero también es error señalar que las mujeres presentaron más preferencia al área de Química que los hombres.

Para realizar una lectura adecuada de los datos del cuadro se requiere hacer una comparación porcentual, tal como se indica en el cuadro siguiente:

Liceo El Progreso: área científica seleccionada por los estudiantes que realizaron

la prueba de Bachillerato según el sexo, año 2015

Área científica seleccionada	Hombres Total	Hombres Porcentaje	Mujeres Total	Mujeres Porcentaje
Física	9	14,5	9	7,6
Química	16	25,8	29	24,6
Biología	37	59,7	80	67,8
Total	**62**	**100,0**	**118**	**100,0**

Información ficticia utilizada con fines didácticos

Al comparar los porcentajes se puede notar que hubo mayor preferencia de los hombres por el área de Física y de las mujeres por el área de Biología; pero en el área de Química fue similar la preferencia. Entonces se debe recomendar al Director que sustituya el cuadro que quería incluir por este otro que resulta más ilustrativo.

Los problemas anteriores realizan un análisis de variables cualitativas (no numéricas) por medio de porcentajes. Sin embargo, si las variables son cuantitativas (numéricas) entonces se requiere recurrir a otras técnicas porque los porcentajes no aplican para ellas.

4. Uso de medidas estadísticas de posición: medidas de tendencia central

Volver al índice de conceptos

La media aritmética o promedio

Corresponde al valor numérico que se obtiene de sumar todos los datos y dividirlos por el total de datos. La fórmula es:

$$media\ aritmética\text{:}\ \frac{Suma\ de\ datos}{Total\ de\ datos}$$

La mediana

Divide al grupo de datos en subconjuntos y se ubica en la mitad, cumple la siguiente propiedad: el 50% de los datos toma un valor numérico menor o igual que la mediana y otro 50% tiene un valor numérico mayor o igual que la mediana.

Por ejemplo, si se tienen n datos ordenados de menor a mayor $X_1 \leq X_2 \leq \dots \leq X_n$, (el subíndice representa la posición de cada dato, de menor a mayor, X_1es el menor dato, le sigue X_2, así sucesivamente) entonces la mediana se determina por:

$$\text{Mediana:}\begin{cases} X_{\left(\frac{n+1}{2}\right)} & \text{si el n es inpar} \\ \dfrac{X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n}{2}+1\right)}}{2} & \text{si el n es par} \end{cases}$$

Nota: Si n es impar, entonces $n+1$ es un número par y entonces $\frac{n+1}{2}$ es un número entero, por esta razón el dato $X_{\left(\frac{n+1}{2}\right)}$ corresponde al dato que se encuentra en la posición $\frac{n+1}{2}$.

Por otro lado, si n es par, entonces $\frac{n}{2}$ es un número entero y también su $\frac{n}{2}+1$ que sería el número entero consecutivo, entonces $X_{\left(\frac{n}{2}\right)} + X_{\left(\frac{n}{2}+1\right)}$ representa la suma de los datos que están en estas posiciones

Volver al índice de conceptos

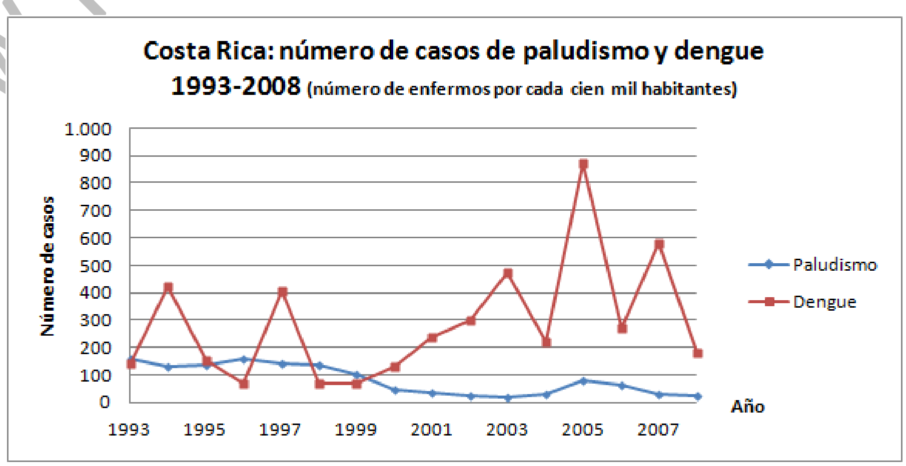
La moda

Corresponde al valor numérico que más se repite dentro de un grupo de datos. En un grupo de datos puede presentarse una única moda, más de una moda, o podría no existir este valor.

Nota: Las medidas estadísticas: media aritmética, mediana y moda se les llama medidas de tendencia central, porque se ubican hacia el centro de la distribución de los datos.

Problema 7. Enfermos por picadura de mosquito

Existen diversas enfermedades que se adquieren por la picadura de mosquitos, dos de las más conocidas son el paludismo y el dengue. La siguiente gráfica muestra el número de casos de estas enfermedades que se presentó en el país entre 1993 y 2008.



Fuente: http://www.estadonacion.or.cr/index.php/estadisticas/costa-rica/compendio-estadistico/estadisticas-sociales

Considere las siguientes proposiciones que se relacionan con la lectura realizada al gráfico:

I. Para el período 1993-2008, el número promedio anual de enfermos por dengue es superior al número promedio anual de enfermos por paludismo.

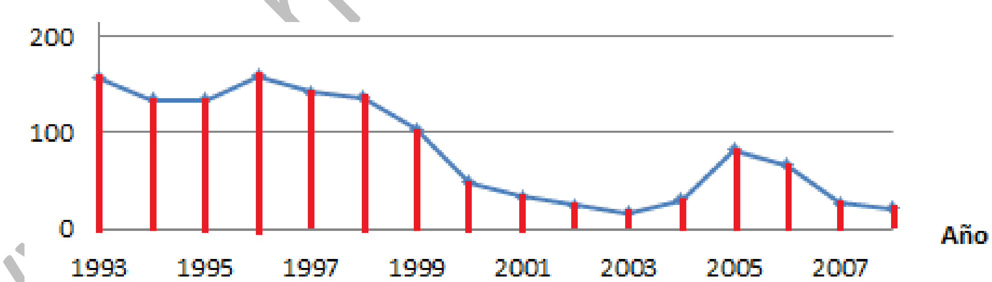
II. Para el período 1993-2008, el número promedio anual de enfermos por paludismo fue menor de 100 casos por cada cien mil habitantes.

Determine si las proposiciones anteriores son verdaderas o falsas.

Solución:

Hay que destacar que la representación gráfica anterior muestra un índice de salud que corresponde al número de enfermos por cada cien mil habitantes para un período de 18 años (de 1993 hasta el 2008). Según esta representación en la mayoría de los años analizados el número de enfermos por dengue superó al de enfermos por paludismo (esto último se puede observar ya que hay una mayor cantidad de años en el período en donde la gráfica del dengue está por encimo que la gráfica del paludismo), estas diferencias fueron más grandes después de 1999. Entonces para este período del estudio, al sumar el número de enfermos por dengue se obtiene un valor mayor que la suma del número de enfermos por paludismo. Por ello es claro que el número promedio anual de enfermos por dengue superó al número promedio anual de enfermos por paludismo.

Por otro lado, si observamos solamente el comportamiento de la curva correspondiente a los casos de paludismo, vemos que entre 1993 y 1999 se presentaron más de cien casos por cada cien mil habitantes, pero después de ese año el número de casos fue menor a cien como se muestra en la siguiente imagen.

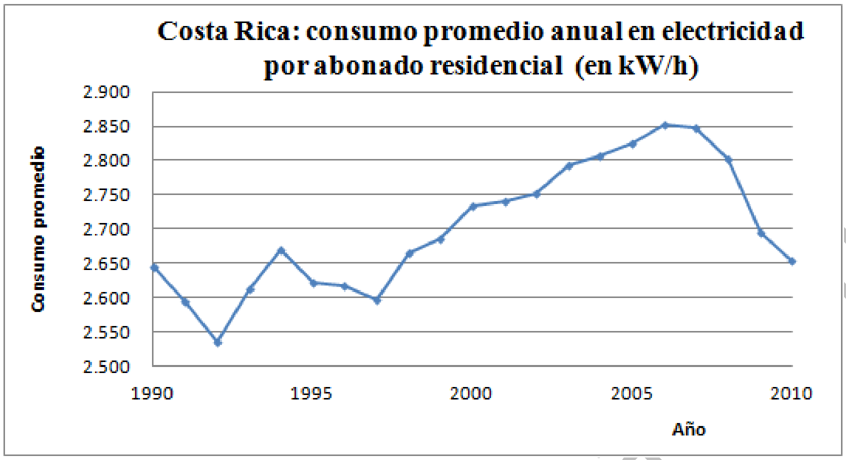


Bajo el supuesto que el gráfico está construido con una escala adecuada, al extraer el promedio de las magnitudes de los segmentos que se representan con rojo que corresponden al número anual de casos de paludismo por cada cien mil habitantes, se puede observar visualmente que este promedio debería ser menor de 100.

Por lo anterior se puede decir que las dos proposiciones son verdaderas.

Problema 8. Consumo promedio anual de electricidad por abonado residencial

Observe la siguiente representación gráfica.



Fuente: http://www.estadonacion.or.cr/index.php/estadisticas/costa-rica/compendio-estadistico/estad-ambientales

Analice las siguientes afirmaciones vinculadas con la información que presenta el gráfico anterior:

I. El consumo total de electricidad entre los abonados residenciales fue aproximadamente igual en 1990 y en el 2010.

II. El máximo consumo de electricidad presentado en una vivienda fue de 2850 kW/h.

Determine cuál o cuáles de estas afirmaciones se puede asegurar con certeza que sean verdaderas:

Solución:

Es evidente que en 1990 había mucho menos abonados residenciales que en el 2010. Por esta razón, no es suficiente saber que el consumo promedio anual de electricidad (en kW/h) fue el mismo para 1990 y 2010 (como lo presenta la gráfica) para concluir que el consumo total también haya sido aproximadamente igual en esos años, puesto que ello depende del número total de abonados residenciales en cada uno de dichos años. Hay que recordar que el promedio de cada año se determina por la fórmula:

$$Promedio = \frac{\text{Consumo total en electricidad de los abonados residenciales}}{\text{Total abonados residenciales}}$$

Entonces si los consumos promedio por abonado residencial de 1990 y 2010 son iguales y el número total de abonados residenciales son diferentes, también el consumo total de electricidad de los abonados residenciales en ambos años son diferentes.

Por otro lado, la gráfica ilustra el consumo promedio anual de electricidad y ello no permite asegurar el valor del máximo consumo de electricidad en una vivienda particular, basta con analizar la fórmula de cálculo del promedio, para identificar que con la información del gráfico no es posible determinar el consumo total de electricidad por vivienda ni tampoco el acumulado para todas las viviendas.

Por lo anterior, ambas proposiciones son falsas.

Problema 9. Rendimiento en combustible de los vehículos

Pilar y Beatriz discuten acerca del rendimiento de sus automóviles con respecto al gasto de combustible. Cada una de ellas indica que su vehículo es más económico.

http://www.diariomotor.com/

Deciden medir el rendimiento de sus vehículos en los días laborales (5 días por semana). Al iniciar cada día llenan completamente el tanque de combustible y miden el kilometraje recorrido. Al día siguiente vuelven a realizar el proceso para determinar los litros de combustible consumidos. Después de dos semanas, tienen información correspondiente a 10 días. Al dividir los kilómetros recorridos por los litros consumidos obtuvieron el rendimiento en kilómetros por litro de combustible.

Rendimiento en kilómetros por litro de los vehículos de Pilar y Beatriz en 10 días laborales.

Vehículo de Pilar			Vehículo de Beatriz		
Km recorridos	**Litros consumidos**	**Rendimiento (Km por litro)**	**Km recorridos**	**Litros consumidos**	**Rendimiento (km por litro)**
28,62	2,37	12,1	43,15	3,96	10,9
28,68	2,17	13,2	34,84	2,54	13,7
30,05	2,40	12,5	35,99	3,76	9,6
30,50	2,62	11,6	37,77	3,12	12,1
30,73	2,05	15,0	38,85	2,96	13,1
28,42	2,14	13,3	40,92	2,88	14,2
27,85	2,13	13,1	38,77	4,07	9,5
27,26	2,33	11,7	38,01	2,71	14,0
32,96	2,44	13,5	39,36	3,04	12,9
29,10	2,02	14,4	40,84	3,36	12,2

Información ficticia utilizada con fines didácticos

Sin embargo, no saben cómo analizar esta información para realizar una comparación que permita identificar cuál de los dos autos tuvo un mejor rendimiento en las dos semanas observadas. Utilice las medidas de tendencia central para ayudar a Pilar y Beatriz a resolver el problema que enfrentan.

Solución:

En este problema vemos el efecto de la variabilidad en las situaciones de la cotidianidad, pues aunque tenemos los mismos autos y se someten a recorridos similares, el rendimiento varía de un día a otro, lo cual puede deberse a múltiples causas además del funcionamiento del vehículo: embotellamientos vehiculares, estilos de conducción, estado de las carreteras, entre otras. El reto para resolver el problema consiste en resumir la información por medio de alguna estrategia que permita comparar el rendimiento entre los autos.

En estos casos las medidas de posición constituyen una buena alternativa. Regularmente, las personas recurren al promedio o media aritmética de datos, la cual se define de la siguiente manera:

En el problema que estamos analizando, para el auto de Pilar la suma de los rendimientos es $130{,}4\ \mathrm{km/L}$. Entonces la media aritmética del rendimiento sería: $\frac{130{,}4}{10}\ \mathrm{km/L} = 13{,}04\ \ \mathrm{km/L}$.

Para el auto de Beatriz la suma de los rendimientos es $122{,}2\ \mathrm{km/L}$. Entonces la media aritmética del rendimiento sería: $\frac{121{,}2}{10}\ \mathrm{km/L} = 12{,}22\ \ \mathrm{km/L}$.

De acuerdo con lo anterior, para los diez días observados, el rendimiento promedio del auto de Beatriz superó el rendimiento promedio del auto de Pilar.

Sin embargo, desde el punto de vista estadístico no es adecuado recurrir a una sola medida de posición, sino que conviene analizar otras medidas que nos ayuden a complementar el mensaje. En este sentido mediana permite complementar el mensaje que comunica la media aritmética. Seguidamente se presenta la definición de la mediana:

Para nuestro problema, si ordenamos los rendimientos de menor a mayor en cada caso se tendría lo siguiente:

Rendimiento en kilómetros por litro

Auto de Pilar		Auto de Beatriz	
1	11,6	1	9,5
2	11,7	2	9,6
3	12,1	3	10,9
4	12,5	4	12,1
5	13,1	5	12,2
6	13,2	6	12,9
7	13,3	7	13,1
8	13,5	8	13,7
9	14,4	9	14,0
10	15,0	10	14,2

Como puede notarse en cada caso $n = 10$ datos, que corresponde a un valor par, entonces la mediana se calcula mediante la fórmula:

$$\frac{X_{\left(\frac{10}{2}\right)} + X_{\left(\frac{10}{2}+1\right)}}{2} = \frac{X_5 + X_6}{2}$$

Nota: Como *n* es 10, entonces la mediana se encuentra entre los datos que están en la posición $\frac{10}{2} = 5$ y en la posición $\frac{10}{2} + 1 = 5 + 1 = 6$. En el caso del auto de Pilar, en la posición cinco está el número 13,1 y en la posición seis está el número 13,2.

La mediana en el rendimiento del auto de Pilar es $\frac{13{,}1 + 13{,}2}{2}$ $\text{km/L} = 13{,}15\ \text{km/L}$. Esto quiere decir que en la mitad de los días observados el rendimiento en el auto de Pilar fue de $13{,}15\ \text{km/L}$ o menos, y en la otra mitad de días el rendimiento fue mayor o igual que ese valor.

La mediana en el rendimiento del auto de Beatriz es $\frac{12{,}2 + 12{,}9}{2}$ $\text{km/L} = 12{,}55\ \text{km/L}$. Entonces en la mitad de los días observados el rendimiento en el auto de Beatriz fue de 12,55 km/L o menos, y en la otra mitad de días el rendimiento fue mayor o igual que ese valor.

También la mediana en el rendimiento del auto de Pilar toma un valor más alto que la correspondiente al auto de Beatriz, para los diez días observados.

Finalmente se tiene que para los datos del problema no hay modas.

En resumen, se tiene lo siguiente:

Rendimiento en kilómetros por litro de los autos de Pilar y Beatriz, información recabada por día, en 10 días consecutivos

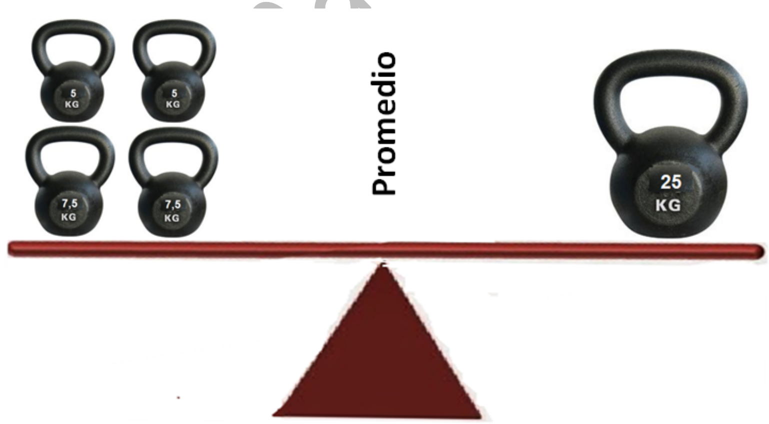
Medida estadística de posición	Rendimiento de auto de Pilar Km/L	Rendimiento de auto de Beatriz Km/L
Media aritmética	13,04	12,22
Mediana	13,15	12,55
Moda	No tiene	No tiene

Tanto la media aritmética como la mediana de los rendimientos fueron superiores para el auto de Pilar, lo que da la impresión que este auto tiene un mayor rendimiento, no puede asegurse que sea así en definitiva, porque se basa en resultados de unos pocos días, podría ser que si se observan más días la situación cambie.

5. Relación gráfica entre las medidas de tendencia central

La media aritmética, la mediana y la moda tienen una posición estratégica en el eje x cuando se visualizan gráficamente. Cuando se tiene un número grande de datos y se analiza su patrón de variabilidad entre las medidas de tendencia central (media aritmética, mediana o moda), existe una relación que se vincula directamente con la asimetría de la distribución de los datos.

Si la distribución de los datos es asimétrica (a la derecha o a la izquierda), los valores extremos provocan que la media aritmética se aparte un poco del lugar donde está la mayor concentración de datos



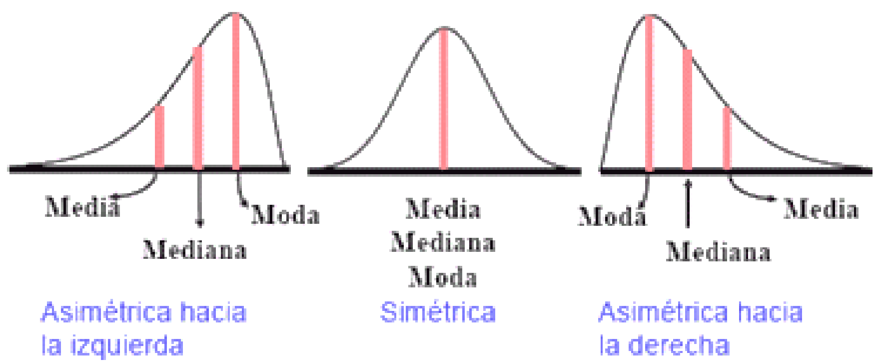
Observe que el peso promedio de estas cinco pesas es $\frac{5{,}0+5{,}0+7{,}5+7{,}5+25{,}0}{5}\text{kg} = 10\text{ kg}$. Sin embargo, este valor de 10 kg no es representativo de la tendencia central en los pesos de las cinco pesas. En este caso el peso promedio se ve influenciado por la pesa de 25 kg, la cual provoca que el promedio sea más grande que el común de las pesas. Por esta razón, se dice que el promedio o media aritmética no es una medida adecuada para representar la tendencia central de los datos cuando existe fuerte asimetría positiva (a la derecha como en el caso de las pesas) o negativa (a la izquierda).

En cambio, independientemente del tipo de asimetría que presente la distribución de los datos, la mediana, siempre se ubica en el centro, de modo que no más del 50% de los datos son menores y no más del 50% son mayores.

Observe que para el caso de las pesas, al existir cinco datos, la mediana en peso estaría dada por el valor: $Mediana = X_{\left(\frac{5+1}{2}\right)} = X_3 = 7{,}5$ kg.

Por su parte la moda, se ubica donde hay más concentración de datos; sin embargo, es común que la moda tienda a variar mucho entre muestras diferentes. En nuestro ejemplo hay dos modas que son 5 kg y 7,5 kg.

En términos generales, para muestras suficientemente grandes en donde se presenta una única moda, la relación entre estas tres medidas de tendencia central se visualiza en el siguiente esquema:



Nota: La relación anterior es útil cuando los datos han sido resumidos previamente en una distribución de frecuencias y se tiene un patrón general de la forma de la distribución, pero cuando se tienen pocos datos la moda podría tener un comportamiento diferente del que se muestra en el esquema, por esta razón es conveniente concentrarse solamente en la comparación entre la mediana con la media aritmética.

Analice el siguiente problema que se vincula con los conceptos anteriores:

Problema 10. Comparación del rendimiento entre grupos

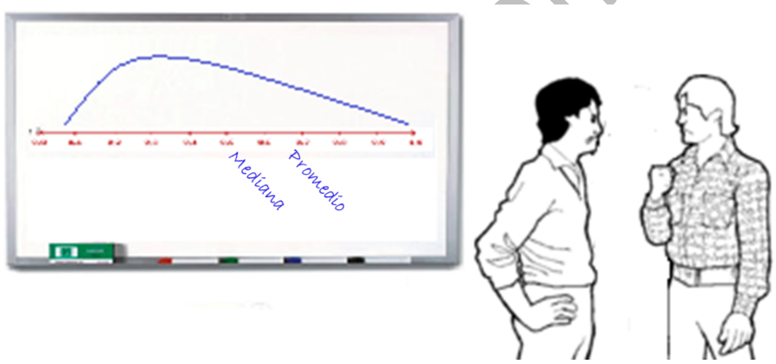
Dos profesores analizan los resultados de un examen en un curso universitario, para los cuales se presentan a continuación dos medidas de tendencia central.

Calificaciones (en escala de 0 a 100) de un examen parcial de los grupos A y B

Medida estadística	Grupo A	Grupo B
Media aritmética	85	77
Mediana	76	81

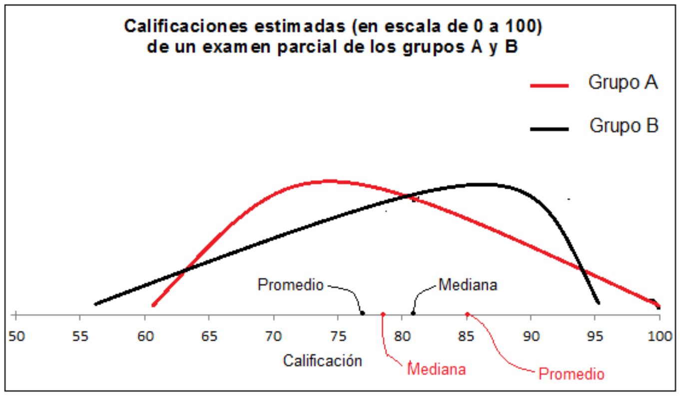
El profesor del grupo A se muestra muy orgulloso porque la calificación promedio de su grupo fue de 85, mientras que en el grupo B que corresponde al profesor fue apenas de 77.

Sin embargo, el profesor del grupo B le indica que, en términos generales, su grupo tuvo un mejor rendimiento debido a que la mayoría de sus estudiantes tuvieron una calificación por encima de 80, lo cual no ocurrió en el grupo A. El docente del grupo A le responde que eso no puede ser posible y que está manipulando la información. Usted debe decidir si el profesor del grupo B tiene o no tiene la razón.



Solución:

Con tan poca información resulta muy difícil determinar cuál de los grupos tuvo un mejor rendimiento. Además, el concepto de mejor rendimiento podría ser ambiguo, debido a que depende de qué quiere decir *un mejor rendimiento*. Sin embargo, las medidas estadísticas son claras en cuanto a su interpretación. Por ejemplo, en términos relativos, la media aritmética establece que en el grupo A la suma de las notas es mayor que en el grupo B. Por otro lado, la mediana señala que el 50% de los estudiantes del grupo B lograron calificaciones mayores o iguales a 81, lo que es un buen indicador para señalar que la mayoría de los estudiantes (más de la mitad) tuvo notas mayores o iguales a 81. Mientras que la mediana del grupo A señala que la mitad de los estudiantes tuvo calificaciones mayores o iguales a 76. Con esta información el docente del grupo B tiene razón en su afirmación. Un esquema muy básico que podría aproximar la distribución de calificaciones sería similar al siguiente:



En este esquema se observa que en el grupo B hay más estudiantes que tuvieron notas mayores que 80. Pero se debe tener claro que esto es simplemente una aproximación de la distribución de las calificaciones. Las distribuciones reales pueden ser diferentes, pero la afirmación del profesor B es correcta.
En este problema, las diferencias en el promedio pueden obedecer a la presencia de valores extremos hacia uno de los lados. Posiblemente, unas pocas calificaciones muy altas provocan que la calificación promedio del grupo A se incremente, mientras que unas pocas calificaciones muy bajas en el grupo B podrían haber provocado que la calificación promedio del grupo B disminuyera.

Problema 11. Años de experiencia en labores docentes

Considere la información del cuadro y las proposiciones dadas.

En el cuadro siguiente se muestran los años de experiencia que tienen 196 docentes de educación primaria que fueron encuestados en el año 2009.

Distribución de los docentes encuestados según número de años de experiencia en labores de docencia. Año 2009

Años de experiencia	Cantidad de docentes	Porcentaje de docentes
De 1 a menos de 5	28	14,3
De 5 a menos de 10	31	15,8
De 10 a menos de 15	37	18,9
De 15 a menos de 20	40	20,4
De 20 a menos de 25	48	24,5
De 25 a menos de 30	12	6,1
Total	**196**	**100**

Con base en los datos del cuadro anterior analice las siguientes proposiciones:

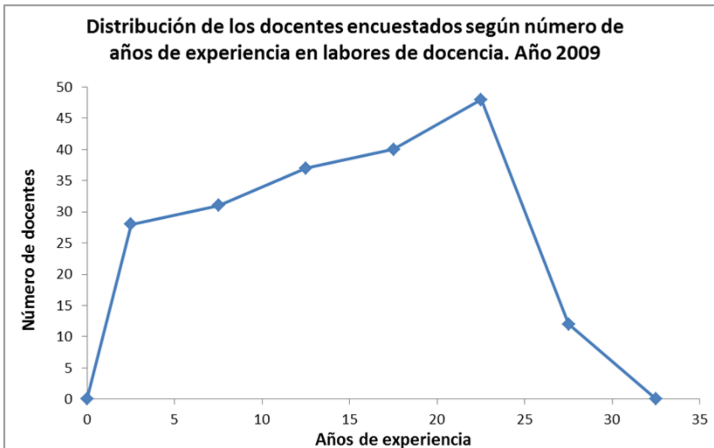
I. El número promedio de años de experiencia de los docentes encuestados es un valor entre 20 y 25 años.

II. El valor de la mediana de la experiencia de los docentes encuestados se encuentra entre 15 y 20 años.

Determine cuál o cuáles de las proposiciones anteriores son verdaderas.

Solución:

Los datos sobre los años de experiencia de los docentes fueron resumidos en el cuadro que se presenta, por lo que no se sabe cuáles fueron los datos individuales. Sin embargo, al observar el ordenamiento de los datos y las frecuencias correspondiente, el cuadro permite identificar una asimetría negativa.



En este sentido, de acuerdo con la relación que debe existir entre el promedio y la mediana, se supone que el número promedio de años de experiencia de los profesores debe ser menor que la mediana de los años de experiencia. Por otro lado, de acuerdo con la definición de la mediana, esta medida acumula el 50% de los datos, si analizamos los porcentajes acumulados, puede notarse que hasta la clase de 15 a 20 años de experiencia, se ha acumulado este 50%, lo cual significa que la mediana en los años de experiencia está en esta clase.

Años de experiencia	Cantidad de docentes	Porcentaje de docentes	Porcentaje acumulado
De 1 a menos de 5	28	14,3	14,3
De 5 a menos de 10	31	15,8	30,1
De 10 a menos de 15	37	18,9	49,0
De 15 a menos de 20	40	20,4	69,4
De 20 a menos de 25	48	24,5	93,9
De 25 a menos de 30	12	6,1	100,0
Total	**196**	**100**	

Por lo tanto la primera proposición sería falsa y la segunda verdadera.

Nota: Los análisis anteriores evidencian que en distribuciones asimétricas, positivas o negativas el promedio o media aritmética se ve afectado por la presencia de los valores extremos, entonces no es un buen indicador del comportamiento central de los datos. La mediana puede ser un mejor indicador de esta tendencia central. Pero si se tienen las tres medidas, lo ideal es realizar un análisis integral.

A pesar de lo anterior, desde un punto de vista estadístico la media aritmética sigue teniendo una gran importancia en los análisis más elaborados.

6. Otras medidas estadísticas de posición: medidas de orden

En algunos análisis estadísticos se requiere destacar una medida o valor hasta el cual se acumula un porcentaje o un cierto porcentaje de datos. En este sentido existen diversas medidas estadísticas destinadas a determinar estos valores. Unos de los más conocidos son los cuartiles:

Volver al índice de conceptos

Los cuartiles

Son tres valores que dividen al grupo de datos en cuatro partes. Tienen una función similar a la de la mediana. El primer cuartil o C_1 cumple que el 25% de los datos son menores o iguales a C_1 y el 75% son mayores o iguales a C_1. El segundo cuartil o C_2 corresponde a la mediana. El tercer cuartil o C_3 cumple que el 75% de los datos son menores o iguales a C_3 y el 25% son mayores o iguales a C_3.

Existen diferentes fórmulas de cálculo para los cuartiles, normalmente se presentan pequeñas diferencias en el resultado de aplicar estas fórmulas, pero esas diferencias no son importantes para los problemas que se aplican en secundaria, siempre que se utilice una misma fórmula al hacer comparaciones.

Al igual que con la mediana, lo primero que se debe considerar para calcular un cuartil es que los datos estén ordenados de menor a mayor. Entonces si se tienen *n* datos y los representamos con $X_1, X_2, X_3, \cdots, X_n$ se cumple que:

$$X_1 \leq X_2 \leq X_3 \leq \cdots \leq X_n$$

Una de las fórmulas más utilizadas para determinar los cuartiles es la siguiente:

Si *k* toma los valores 1, 2 o 3, entonces se calcula el valor $\frac{k \cdot (n+1)}{4}$:

a) Si $\frac{k \cdot (n+1)}{4}$ en un entero, el cuartil número *k* sería el dato ordenado que se encuentre en esa posición, o sea:

$$Cuartil\ k{:}C_k = X_{\frac{k \cdot (n+1)}{4}}$$

b) Si $\frac{k \cdot (n+1)}{4}$ no es entero, se determina la parte entera de este número, digamos M (corresponde al mayor número entero inferior a $\frac{k \cdot (n+1)}{4}$). Se dice que el cuartil k es un promedio entre los datos ordenados que están en la posición M y $M + 1$, o sea.

$$\text{Cuartil k: } C_k = \frac{X_M + X_{M+1}}{2}$$

Nota: De acuerdo con lo anterior el primer cuartil acumula el 25% de datos menores, el segundo cuartil acumula el 50% de los datos menores y el tercer cuartil acumula el 75% de los datos menores.

Nota: Observe que el segundo cuartil coincide con la definición de la mediana.

Nota: además de los cuartiles existen otras medidas similares tales como: los quintiles los cuales dividen el conjunto de datos en cinco grupos, los deciles que dividen el conjunto de datos en diez grupos, los percentiles que dividen el conjunto de datos en cien partes. Todos estas medidas cuartiles, quintiles, deciles y percentiles se les llama cuantiles. Para efectos de la educación secundaria solamente se analizan los cuartiles.

En muchos casos la identificación del valor más bajo o del más alto es de interés práctico, este hecho provoca que estos valores sean identificados también como medidas estadísticas.

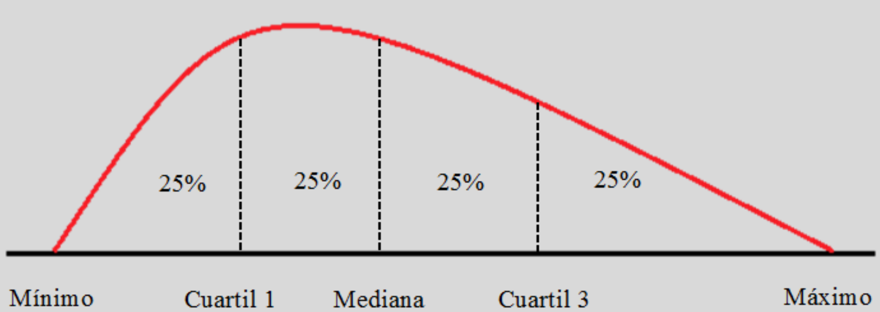
Volver al índice de conceptos

El máximo y el mínimo:

Corresponde al mayor valor y al menor valor de grupo de datos respectivamente.

Nota: A las medidas estadísticas: mediana, cuartiles (en general a todos los cuantiles) y valores máximo y mínimo se les llama medidas estadísticas de orden, porque ordenan el conjunto de datos.

Cuanto estos valores se analizan en una representación gráfica de una variable continua, se puede interpretar el porcentaje de datos encerrados entre las cuantiles como el área bajo la curva:



Problema 12. Entrenamiento para los 100 metros planos

Un atleta de alto rendimiento se entrena para participar internacionalmente en los 100 metros planos (competencia atlética en la cual los corredores deben correr 100 metros en una pista en el menor tiempo posible). En los últimos días realizó 50 competencias de entrenamiento que fueron cronometradas. Debido a la variabilidad de los resultados, se observan tiempos de recorrido muy buenos, pero otros no tanto. Junto con su entrenador coinciden en que posiblemente un 25% de los tiempos no responden a su verdadero potencial debido a circunstancias técnicas especiales: una mala salida (salió tarde), una caída o resbalón durante la competencia, un roce con un compañero, mal estado de la pista, entre otras. Por esta razón, deciden eliminar el 25% de los tiempos más altos alcanzados en estas competencias, y analizar el resto de los resultados.

http://caliescribe.com/deporte/mundial-atletismo-menores-2015

Los tiempos obtenidos por este atleta en segundos para las 50 competencias fueron los siguientes:

12,38	10,29	10,79	10,76	9,95
10,41	11,53	9,97	10,30	9,98
12,10	10,63	9,91	10,51	9,94
10,24	10,10	12,50	10,95	12,46
10,35	10,57	10,83	10,63	10,48
10,43	10,18	10,78	10,20	10,00
10,70	10,24	9,95	10,09	12,18
9,86	11,94	10,85	10,84	10,25
10,03	10,30	10,43	10,29	10,13
9,92	13,54	12,51	10,30	10,31

Ayude al atleta y a su entrenador a determinar el valor hasta el cual se acumula el 75% de los mejores tiempos logrados por este (el 25% de los tiempos son de mayor magnitud) y al mismo tiempo determine el mejor tiempo de competencia logrado.

Solución:

Como puede notarse las fórmulas de los cuartiles pueden ayudarnos a resolver este problema. Específicamente, se requiere determinar el tercer cuartil relacionado con los datos de estos tiempos. Pero hay que tomar en cuenta que para determinar cualquiera de estos cuartiles, se requiere ordenar los datos de menor a mayor.

Posición	Tiempo	Posición	Tiempo	Posición	Tiempo	Posición	Tiempo	Posición	Tiempo
1	9,86	11	10,09	21	10,30	31	10,57	41	10,95
2	9,91	12	10,10	22	10,30	32	10,63	42	11,53
3	9,92	13	10,13	23	10,30	33	10,63	43	11,94
4	9,94	14	10,18	24	10,31	34	10,70	44	12,10
5	9,95	15	10,20	25	10,35	35	10,76	45	12,18
6	9,95	16	10,24	26	10,41	36	10,78	46	12,38
7	9,97	17	10,24	27	10,43	37	10,79	47	12,46
8	9,98	18	10,25	28	10,43	38	10,83	48	12,50
9	10,00	19	10,29	29	10,48	39	10,84	49	12,51
10	10,03	20	10,29	30	10,51	40	10,85	50	13,54

Se tiene n = 50 datos. Este número es par, además, como buscamos calcular el tercer cuartil, se tiene que k = 3,

Al aplicar la fórmula $\frac{\mathrm{k}\cdot(n+1)}{4}$ se tiene que $\frac{3\cdot 51}{4}=38{,}25$, entonces el mayor número entero inferior a 38,25 es 38, esto quiere decir que M = 38, con lo cual el tercer cuartil está entre el dato que está en la posición 38 ($\mathrm{X}_{38}=10{,}83$) y el dato de la posición 39 ($\mathrm{X}_{39}=10{,}84$):

$$\text{Cuartil 3: } \mathrm{C}_3=\frac{\mathrm{X}_{38}+\mathrm{X}_{38+1}}{2}=\frac{10{,}83+10{,}84}{2}=10{,}835$$

Esto quiere decir que en el 75% de las competencias, los tiempos fueron menores o iguales que 10,835 segundos; mientras que en el otro 25% de las competencias, los tiempos fueron mayores o iguales a 10,835 segundos.

Para responder el problema les podemos decir al atleta y a su entrenador que valoren las competencias en las cuales el tiempo obtenido fue menor o igual a 10,835 segundos.

Por otro lado, para determinar el mejor tiempo de competencia basta con identificar el valor más bajo o tiempo mínimo obtenido en las 50 competencias, este valor fue de 9,86 segundos.

Nota: Anteriormente se han analizado diferentes medidas estadísticas de posición (algunas de tendencia central y otras de orden), cada una de ellas tiene una interpretación diferente, la comparación entre las diferentes medidas permite generar un análisis más integral del mensaje que comunican los datos para la resolución de un problema.

Problema 13. Rendimiento en combustible de los vehículos

Para ejemplificar las medidas de orden, volvamos a analizar el problema 9, en donde Pilar y Beatriz discutían sobre el rendimiento de combustible de sus vehículos (en kilómetros recorridos por litro). Para los 10 días observados, los datos de los rendimientos de estos vehículos ordenados de menor a mayor son:

Rendimiento en kilómetros por litro

Auto de Pilar		**Auto de Beatriz**	
1	11,6	1	9,5
2	11,7	2	9,6
3	12,1	3	10,9
4	12,5	4	12,1
5	13,1	5	12,2
6	13,2	6	12,9
7	13,3	7	13,1
8	13,5	8	13,7
9	14,4	9	14,0
10	15,0	10	14,2

Analice el problema de acuerdo con la información que proporcionan las medidas de orden para indicar cuál de los vehículos tiene un mejor rendimiento.

Solución:

Para responder el problema se debe determinar las siguientes medidas de orden: mínimo, primer cuartil, mediana, tercer cuartil y máximo (hay que recordar que el segundo cuartil es la mediana)

Según la fórmula de los cuartiles, el primer cuartil: $\mathrm{k} = 1$, como $\mathrm{n} = 10$ entonces debemos calcular la expresión $\frac{\mathrm{k} \cdot (n+1)}{4} = \frac{1 \cdot 11}{4} = 2.75$, debido a que no es entero entonces la parte entera 2,75 es $\mathrm{M} = 2$, quiere decir que el primer cuartil es el promedio de los datos dos y tres:

Para el auto de Pilar: $\text{Cuartil 1: } \mathrm{C}_1 = \frac{\mathrm{X}_2 + \mathrm{X}_3}{2} = \frac{11{,}7 + 12{,}1}{2}\ \mathrm{km/L} = 11{,}9\ \mathrm{km/L}$, lo que significa que en el 25% de los días el rendimiento fue menor o igual a 11,9 km/L, y en el 75% de los días el rendimiento del vehículo fue mayor o igual a este valor.

Para el rendimiento del auto de Beatriz: $\text{Cuartil 1: } \mathrm{C}_1 = \frac{\mathrm{X}_2 + \mathrm{X}_3}{2} = \frac{9{,}6 + 10{,}9}{2}\ \mathrm{km/L} = 10{,}25\ \mathrm{km/L}$, la interpretación es similar.

Tercer cuartil: $\mathrm{k} = 3$, como $\mathrm{n} = 10$ entonces debemos calcular la expresión $\frac{\mathrm{k} \cdot (n+1)}{4} = \frac{3 \cdot 11}{4} = 8{,}25$, debido a que no es entero entonces la parte entera de 8,25 es $\mathrm{M} = 8$. Quiere decir que el tercer cuartil es el promedio de los datos siete y ocho:

Para el rendimiento del auto de Pilar, $\text{Cuartil 3: } C_1 = \frac{X_8 + X_9}{2} = \frac{13{,}5 + 14{,}4}{2}$ $\text{km/L} = 13{,}95\ \text{km/L}$, lo que significa que en el 75% de los días el rendimiento fue menor o igual a 13,95 km/L y en el 25% de los días el rendimiento del vehículo fue mayor o igual a este valor.

Para el rendimiento del auto de Beatriz, $\text{Cuartil 3: } C_1 = \frac{X_8 + X_9}{2} = \frac{13{,}7 + 14{,}0}{2}$ $\text{km/L} = 13{,}85\ \text{km/L}$, la interpretación es similar.

Para determinar la mediana, como $n = 10$ y representa un número par, entonces la mediana se encuentra los datos $X_{\frac{10}{2}}$ y $X_{\frac{10}{2}+1}$ (revisar fórmula de la mediana) entonces:

$$Mediana = \frac{X_5 + X_6}{2}$$

La mediana en el rendimiento del auto de Pilar es $\frac{13{,}1 + 13{,}2}{2}\ \text{km/L} = 13{,}15\ \text{km/L}$. Esto quiere decir que en la mitad de los días observados el rendimiento del auto de Pilar fue de $13{,}15\ \text{km/L}$ o menos y en la otra mitad de días el rendimiento fue mayor o igual que ese valor.

La mediana en el rendimiento del auto de Beatriz es $\frac{12{,}2 + 12{,}9}{2}\ \text{km/L} = 12{,}55\ \text{km/L}$. Entonces en la mitad de los días observados el rendimiento del auto de Beatriz fue de 12,55 km/L o menos y en la otra mitad de días el rendimiento fue mayor o igual que ese valor.

En cuanto a los valor máximo y valor mínimo se tiene que para el auto de Pilar el mayor rendimiento fue de $15{,}0\ \text{km/L}$ y el menor $11{,}6\ \text{km/L}$. Para el vehículo de Beatriz el mayor rendimiento fue de $14{,}2\ \text{km/L}$ y el menor $9{,}5\ \text{km/L}$.

Para una mejor comparación, las medidas estadísticas anteriores pueden ser resumidas en un cuadro (se incluyó también la mediana que fue calculada en el problema 9):

Rendimiento en kilómetros por litro de los autos de Pilar y Beatriz, información recabada por día, en 10 días consecutivos

Medida estadística de orden	Rendimiento de auto de Pilar Km/L	Rendimiento de auto de Beatriz Km/L
Primer cuartil	11,9	10,25
Mediana	13,15	12,55
Tercer cuartil	13,95	13,85
Máximo	15,0	14,2
Mínimo	11,6	9,5

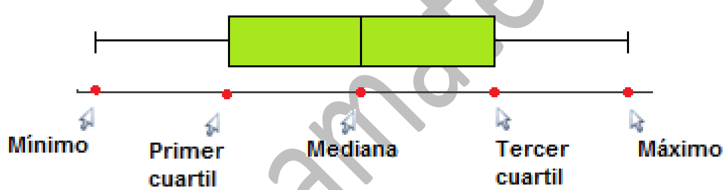
En conclusión, en las medidas estadísticas de orden analizadas, el auto de Pilar mostró un mayor rendimiento, por ello puede concluirse que el vehículo de Pilar tiene un mejor rendimiento en el consumo de combustible por kilómetro recorrido en los diez días en que se recolectaron los datos, según la información de los 10 días muestreados.

7. Representación de la distribución de datos mediante un diagrama de cajas

Volver al índice de conceptos

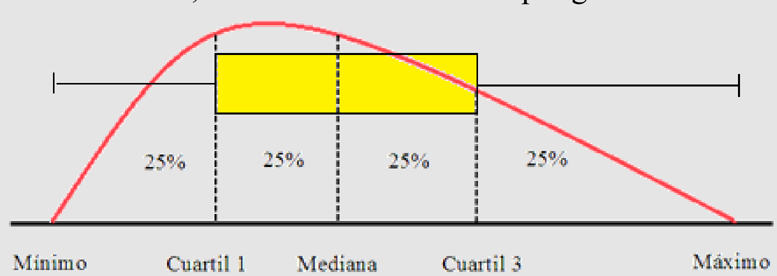
Diagrama de cajas

Como se mencionó anteriormente, las medidas estadísticas de orden tienen una interpretación gráfica que favorece una visualización práctica sobre el patrón de la distribución de los datos. En este sentido los diagramas de cajas se convierten en una importante herramienta estadística para llevar a cabo esta visualización. Un diagrama de cajas, también llamado gráfico de bigotes, consiste en una representación estadística en la cual se incluyen las principales medidas de orden que hemos analizado previamente. Cada caja debe estar en escala según el valor de la medida estadística en el eje x. Entre el primer cuartil y el tercer cuartil se dibuja una caja tal como se muestra. Esta caja se une con el mínimo y el máximo por medio de segmentos. Finalmente la mediana se representa con una línea que divide la caja.



De acuerdo con lo anterior, un diagrama de cajas proporciona información de estas cinco medidas de orden, pero además, son un indicador de la forma aproximada que tiene la distribución de datos. Sin embargo, no se puede perder de vista que estos diagramas son simplemente un resumen de los datos, por medio de cinco medidas. Es importance mencionar que hay que ser muy cuidadosos al interpretar la información que comunican.

Nota: Observe que por medio de diagrama de cajas se puede visualizar en forma aproximada la distribución general de los datos, tal como se hace con un polígono de frecuencias.



Observe que la distribución de los datos representada por la curva muestra una ligera asimetría positiva, esto se evidencia también en diagrama de cajas, debido a que la caja está sesgada hacia la izquierda.

Seguidamente el problema 3, será analizado de acuerdo con el comportamiento de un diagrama de cajas.

Problema 14. Vida útil de los bombillos

Suponga que le envían a comprar unos bombillos a una ferretería, la persona que le atiende indica que de acuerdo con las características del bombillo que usted está solicitando tiene dos marcas con el mismo precio: A y B. Le pregunta que cuál de ellas desea. Usted recuerda que en un artículo de una revista que leyó recientemente salió publicado la siguiente representación gráfica sobre la vida útil de estos bombillos:

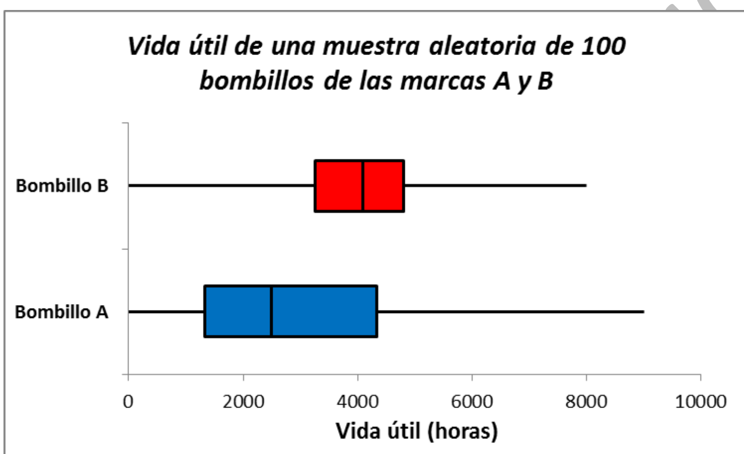
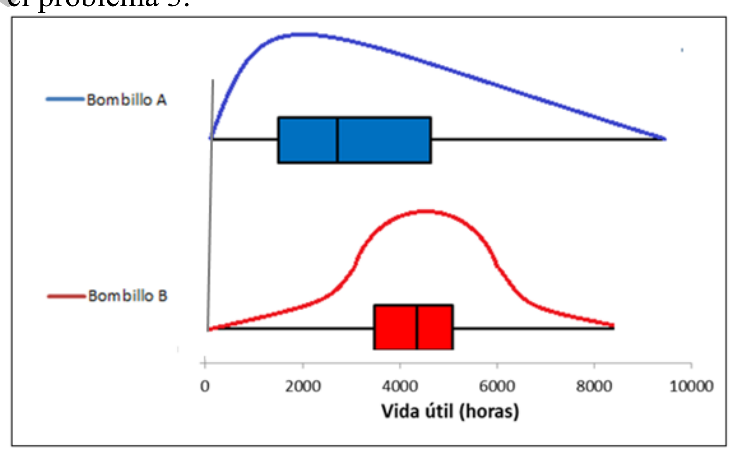


Imagen propia elaborada con fines didácticos

Utilice esta información para decidir qué marca de bombillos compraría y argumente esta decisión.

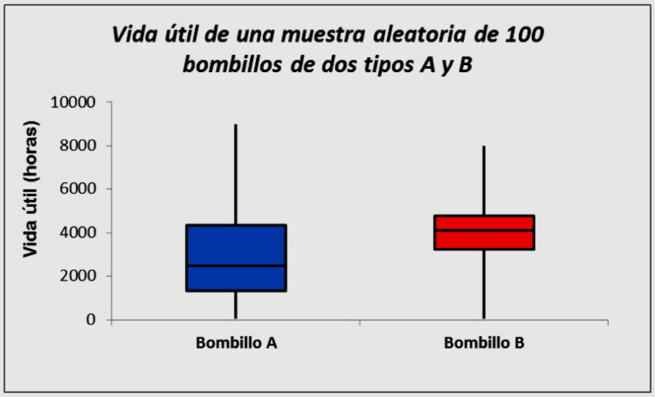
Solución

Note que en la caja correspondiente, a cada tipo de bombillo se puede visualizar el tipo de asimetría que presenta la distribución de los datos de acuerdo con la representación gráfica que apareció en el problema 3.



La información que proporciona el diagrama de cajas es evidencia suficiente para decidir adquirir los bombillos B debido a que existe mayor posibilidad de que tenga una más duración.

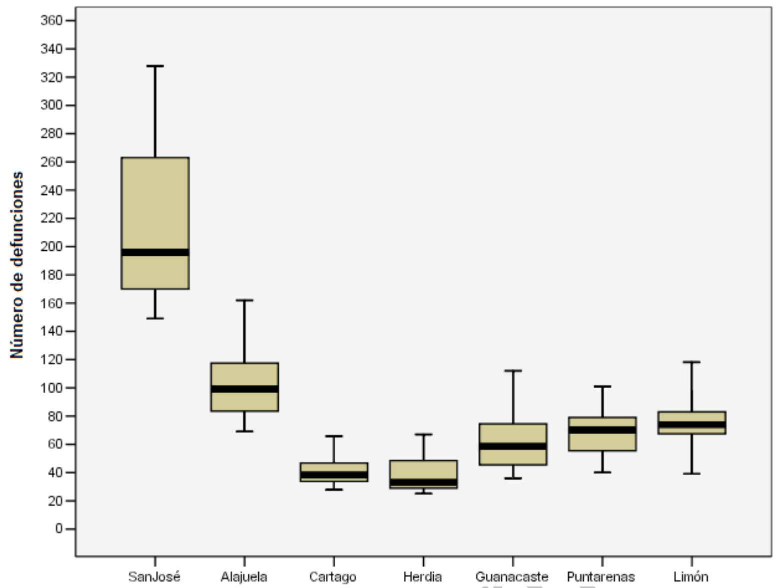
Nota: Los diagramas de cajas pueden expresarse en forma vertical tal como se muestra a continuación:



Problema 15. Defunciones por accidentes de tránsito según provincia

Considere el siguiente diagrama de cajas.

Costa Rica: Número anual de defunciones por accidentes de tránsito
Período 1991-2010



Fuente: http://www.estadonacion.or.cr/index.php/estadisticas/costa-rica/compendio-estadistico/estadisticas-sociales

De la información proporcionada por el diagrama, ¿cuál de las siguientes proposiciones son verdaderas:

I. Limón y Cartago son las provincias en las que se presentó la menor cantidad de muertes por años debidas a accidentes de tránsito en este período.

II. Para San José, en más de la mitad de los años del período 1991-2010 se presentó más de 220 defunciones anuales por accidentes de tránsito.

III. Heredia fue la provincia en que se presentó la menor cantidad de defunciones por accidentes de tránsito por año en este período.

Solución

Al observar las cajas que representan al número anual de defunciones por accidentes de tránsito en nuestro país en el período 1991-2010, se identifica que es en Heredia donde se presentó la menor cantidad de defunciones por accidentes de tránsito en dicho período: tiene menor mediana que todas las demás provincias y el tercer cuartil y valor máximo son similares a Cartago que sería la provincia que le sigue en este análisis de valores absolutos.

En este sentido se tiene que la proposición I sería falsa, debido a que las provincias con menor cantidad de defunciones son Heredia y Cartago. En San José la mediana es menor de 200 defunciones anuales, por ello no es cierto que "*en más de la mitad de los años del período 1991-2010, presentó más de 200 defunciones anuales por accidentes de tránsito*", debido a que el porcentaje de años en los que hubo más de 220 defunciones es mucho menor que la mitad.

Por lo tanto la proposición III es la única correcta.

8. Media aritmética ponderada o promedio ponderado

Hasta el momento los datos con que hemos trabajado no presentan diferencias en cuanto al valor o peso relativo de cada uno. En estos casos, se dice que los datos son simples porque todos tienen el mismo valor o peso dentro del análisis. No obstante, esto no siempre es así: podemos enfrentar situaciones donde los datos con que trabajamos tienen diferente valor o peso. Observe los siguientes problemas:

Promedio ponderado

Cuando se requiere determinar el promedio de datos que tienen pesos diferentes, es decir no todos ellos tienen el mismo valor para efectos del problema, entonces es necesario ponderar según el peso relativo (ponderación) de cada dato. En estos casos se aplica la fórmula:

$$Promedio\ ponderado\text{:}\frac{Suma\ de\ (datos \cdot ponderación)}{Suma\ de\ (ponderaciones)}$$

Volver al índice de conceptos

Problema 16. Ganancia por venta de zapatos

En una zapatería se promedian ventas diarias de 8 pares de zapatos de hombre y 12 de mujer. La ganancia media en los zapatos de hombre es de ₡2000, mientras que en los zapatos de mujer la ganancia media es de ₡2500. Con esta información ayude al gerente de la zapatería a determinar la ganancia promedio por par de zapatos.

Solución

De acuerdo con la fórmula del promedio ponderado, la ganancia promedio por par de zapatos debe calcularse de la siguiente manera

$$\frac{8 \cdot ₡2000 + 12 \cdot ₡2500}{20} = ₡2300.$$

Entonces este sería el monto solicitado.

Problema 17. Nota promedio de un curso

Suponga que usted y una amiga matriculan un curso que tiene el siguiente sistema de evaluación: tres exámenes cortos (o quices), dos exámenes parciales y un examen final. Los exámenes cortos tienen un valor de 5% cada uno, los exámenes parciales valen 20% cada uno y el examen final tiene un valor de 45%. Las calificaciones obtenidas por usted y su amiga son:

Evaluación	Calificación personal	Calificación de amiga
Primer examen corto	75	89
Segundo examen corto	84	93
Tercer examen corto	88	98
Primer examen parcial	71	87
Segundo examen parcial	73	87
Examen final	91	74
Suma total	**482**	**528**

Considerando estas calificaciones ¿cuál fue la nota promedio de cada uno?

Solución:

Si se consideran las calificaciones en forma simple, su amiga obtuvo notas más altas con excepción del último examen. Si calculamos el promedio simple de las calificaciones se obtiene: $\frac{482}{6} = 80{,}3$ para usted y la correspondiente a su amiga $\frac{528}{6} = 88{,}0$.

Sin embargo, determinar un promedio simple constituye un error para efectos de cálculo, porque el peso o valor relativo es diferente para los distintos tipos de evaluación. En estos casos cada dato debe ponderarse de acuerdo con el peso relativo (ponderación) de cada evaluación, de acuerdo con la fórmula:

$$Promedio\ ponderado: \frac{Suma\ de\ (datos \cdot ponderación)}{Suma\ de\ (ponderaciones)}$$

Debido a que las ponderaciones corresponden al valor porcentual de cada evaluación, entonces la calificación promedio puede determinarse tal como se muestra en el siguiente cuadro:

Evaluación	Calificación personal	Calificación de amiga	Ponderación	Nota· ponderación (personal)	Nota · ponderación (amiga)
Primer examen corto	75	89	5	$75 \cdot 5 = 375$	$89 \cdot 5 = 445$
Segundo examen corto	84	93	5	$84 \cdot 5 = 420$	$93 \cdot 5 = 465$
Tercer examen corto	88	98	5	$88 \cdot 5 = 440$	$98 \cdot 5 = 490$
Primer examen parcial	71	87	20	$71 \cdot 20 = 1420$	$87 \cdot 20 = 1740$
Segundo examen parcial	73	87	20	$73 \cdot 20 = 1460$	$87 \cdot 20 = 1740$
Examen final	91	74	45	$91 \cdot 45 = 4095$	$74 \cdot 45 = 3330$
Suma total			**100**	**8210**	**8210**

En las últimas dos columnas se incluyó el producto de cada calificación por la ponderación y al final de la columna aparece la suma, entonces su calificación promedio es $\frac{8210}{100} = 82{,}1$, que es igual a la de su amiga. Entonces se concluye que la nota final del curso es 82,1 para ambos.

Este promedio ponderado se utiliza también cuando los datos han sido agrupados en una distribución de frecuencias. Observe el siguiente problema:

Problema 18. Insecticida contra cucarachas

En un laboratorio químico se está realizando un experimento con un nuevo insecticida contra cucarachas.

www.freepik.es

Desean determinar el tiempo promedio de vida después de que el insecto ha estado en contacto con dicho insecticida. El experimento consiste en fumigar un sector del piso y exponer varias cucarachas que cruzan por el sector que fue rociado. Luego se mide el tiempo hasta que la cucaracha muere. El asistente de laboratorio anota el número de cucarachas que han muerto cada cinco minutos y forma la siguiente distribución:

Tiempo de sobrevivencia de una muestra de 110 cucarachas que fueron expuestas a un nuevo insecticida

Tiempo de sobrevivencia	Número de cucarachas
De 0 a menos de 5 minutos	7
De 5 a menos de 10 minutos	32
De 10 a menos de 15 minutos	46
De 15 a menos de 20 minutos	18
De 20 a menos de 25 minutos	5
De 25 a menos de 30 minutos	2
Total	110

Información ficticia utilizada con fines didácticos

El asistente de laboratorio debe entregar un informe en donde además del cuadro anterior debe incluir el tiempo promedio de sobrevivencia; sin embargo, no sabe cómo determinar este valor. Ayude al asistente de laboratorio a determinar este promedio de modo que pueda concluir el informe correctamente.

Solución:

El asistente del laboratorio enfrenta un problema típico de los análisis estadísticos. Aunque se expusieron 110 cucarachas al insecticida, no se tienen los 110 datos correspondientes al tiempo de sobrevivencia de cada una de ellas, entonces no es posible determinar el tiempo de sobrevivencia promedio utilizando la fórmula de cálculo simple. En estos casos únicamente se sabe que siete cucarachas murieron en menos de cinco minutos, pero no se sabe cuánto tiempo duró cada una, igualmente ocurre con las otras clases.

Para resolver el problema de la ausencia de los datos simples, se busca un valor que represente a cada una de las clases o agrupamientos, el punto medio de la clase (corresponde al promedio entre el límite inferior y el límite superior de esta) constituye el mejor representante. Esto quiere decir que el valor 2,5 segundos sería el representante de la primera clase, 7,5 segundos el representante de la segunda clase y así sucesivamente. Con estos valores entonces se puede determinar un promedio ponderado de la siguiente forma:

Tiempo de sobrevivencia	**Número de Cucarachas (f)**	**Punto medio (x)** (representante de clase)	$x \cdot f$
De 0 a menos de 5 minutos	7	2,5	17,5
De 5 a menos de 10 minutos	32	7,5	240
De 10 a menos de 15 minutos	46	12,5	575
De 15 a menos de 20 minutos	18	17,5	315
De 20 a menos de 25 minutos	5	22,5	112,5
De 25 a menos de 30 minutos	2	27,5	55
Total	**110**		**1315**

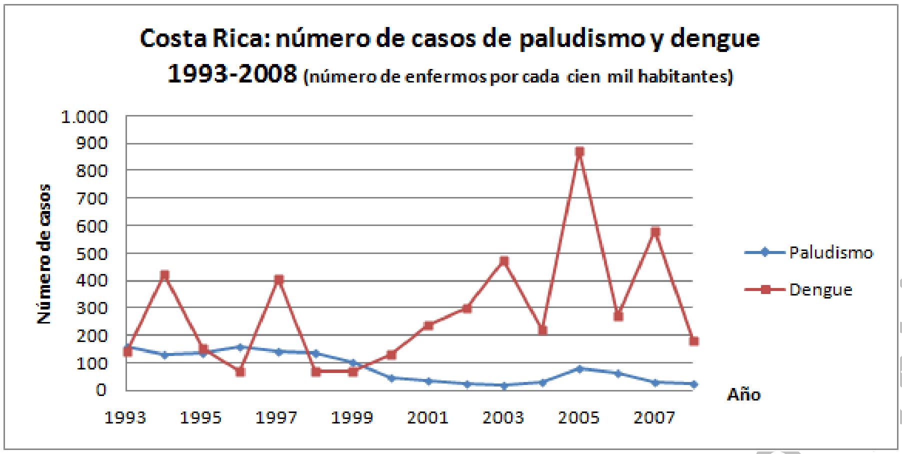
Si se aplica la fórmula del promedio ponderado se tiene que el tiempo promedio de sobrevivencia de las cucarachas una vez que fueron expuestas al insecticida fue de $\frac{1315\ segundos}{110}$=11,95 segundos aproximadamente.

9. Uso de medidas estadísticas de variabilidad

Los análisis anteriores se han concentrado en la posición de los datos, pero debido a que el propósito básico de la Estadística consiste en analizar su patrón de variabilidad, es común encontrar situaciones para los cuales se hace necesario medir esta variabilidad. Al respecto observe el siguiente problema.

Problema 19. Enfermos por picadura de mosquito

Volvamos al problema 7, en donde se analiza dos enfermedades que se vinculan con la picadura de mosquitos: el paludismo y el dengue. La siguiente gráfica muestra el número de casos de estas enfermedades en el país entre 1993 y 2008.



Fuente: http://www.estadonacion.or.cr/index.php/estadisticas/costa-rica/compendio-estadistico/estadisticas-sociales

De acuerdo con esta información, ¿en cuál de las variables se presenta mayor variabilidad: en el número anual de enfermos por paludismo, o en el número anual de enfermos por dengue?

Solución:

A simple vista puede observarse que el número de enfermos por paludismo fue menos variable en este período e incluso presentó una importante disminución. Por su parte, el número de enfermos por dengue fue muy inestable con grandes variaciones entre años y con una tendencia al aumento. En general se puede concluir que el número anual de enfermos por dengue fue mucho más variable durante el periodo 1993-2008.

Problema 20. Contrato de trabajo con empresa turística

Don Juan está actualmente sin empleo, pero acaba de recibir dos ofertas que le parecen atractivas y que generan un salario competitivo dentro de su especialidad laboral que es el turismo. Sin embargo, el salario varía en los meses dependiendo de que la temporada turística sea alta, media o baja. Por esta razón, debido a que don Juan tiene gastos mensuales fijos no le conviene tener muchas diferencias en su ingreso entre un mes y otro.

El siguiente cuadro resume los salarios mensuales (en colones) que le ofrecen las empresas:

Salario mensual en colones ofrecido a don Juan por las empresas A y B, según el mes del año

Mes	Temporada	Empresa A	Empresa B

Enero	Alta	1 300 000	1 050 000
Febrero	Alta	1 200 000	1 050 000
Marzo	Media	1 000 000	850 000
Abril	Media	800 000	850 000
Mayo	Baja	600 000	650 000
Junio	Baja	600 000	650 000
Julio	Alta	1 100 000	1 050 000
Agosto	Media	700 000	850 000
Setiembre	Baja	500 000	650 000
Octubre	Baja	500 000	650 000
Noviembre	Media	800 000	850 000
Diciembre	Alta	1 100 000	1 050 000

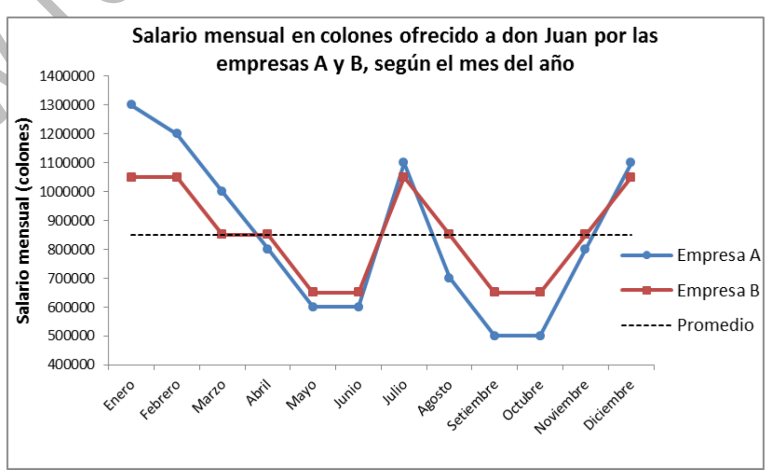
Información ficticia utilizada con fines didácticos

Ayude a don Juan a tomar una decisión sobre la empresa que debería seleccionar, según sus necesidades e intereses.

Solución:

Lo primero que debería realizar don Juan consiste en determinar el ingreso promedio mensual que se le está ofreciendo en cada empresa. Para ello debe sumar los salarios de los 12 meses y dividirlo por 12. En ambos casos el total es de 10 200 000 colones, y entonces su salario promedio mensual será de $\frac{10\ 200\ 000}{12}$ colones = 850 000 colones.

Una vez determinado el salario promedio, se debe analizar la variabilidad, ya que don Juan requiere que no haya muchas variaciones en su salario mensual a lo largo del año, debido a los gastos fijos que debe cubrir mensualmente. En este sentido, basta con observar los datos para darse cuenta que en la empresa B los salarios son más estables a lo largo del año, presentan menos diferencias entre las temporadas y entre los meses, por ello la recomendación sería que acepte el empleo de la empresa B. Para facilitar el análisis la información del cuadro puede ser representada mediante un gráfico de línea como el siguiente:



Puede notarse que la línea correspondiente a los salarios ofrecidos por la empresa A presenta mayor variabilidad en relación con el salario promedio, debido a que en términos generales se aleja más del promedio que la línea correspondiente a los salarios ofrecidos por la empresa B. Por esta razón es que le conviene seleccionar la empresa B.

En estos dos problemas resultó relativamente fácil determinar cuál de los grupos de datos tenía menos variabilidad; sin embargo, esto no siempre ocurre debido a que muchas veces las diferencias de variabilidad no se logran visualizar a partir de los datos y se debe recurrir a análisis más elaborados para resolver estos problemas. Para simplificar este trabajo se han diseñado indicadores para medir la variabilidad de los datos. Las medidas estadísticas de variabilidad más utilizadas son:

- Recorrido
- Recorrido intercuartílico
- Variancia y desviación estándar

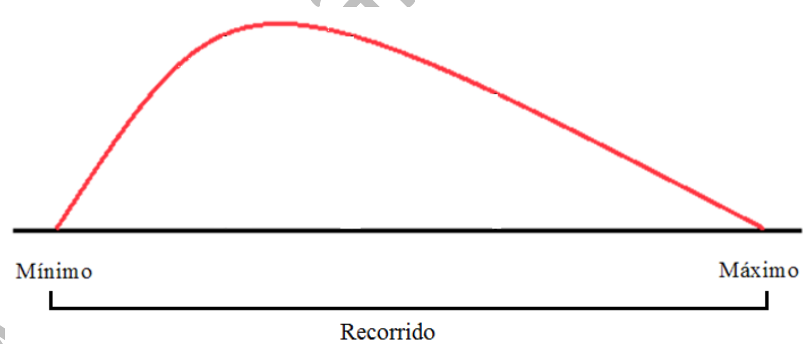
Al igual que con las medidas de posición, las medidas de variabilidad tienen una connotación muy particular y conceptualmente miden diferentes relaciones entre los datos.

El recorrido

Se define por:

$$Recorrido = máximo - mínimo$$

Mide la mayor diferencia numérica que se presenta entre los datos.



Volver al índice de conceptos

Problema 21. Constancia en el lanzamiento del martillo

María y Raquel son dos lanzadoras de martillo que se preparan para participar en diferentes eventos internacionales. En las competencias sobre lanzamiento de martillo, el atleta se ubica en un círculo de 2,15 m de diámetro, el cual está instalado dentro de una jaula de seguridad, protegida con redes y tiene solamente espacio sobre el cual debe salir el martillo. En el proceso de clasificación se realizan solamente tres lanzamientos y se toma como válido el mayor. La técnica debe practicarse mucho porque un alto porcentaje de los lanzamientos son nulos debido a que el atleta se sale del círculo al hacer el lanzamiento o no logra enviar el martillo a la zona de competencia.

https://commons.wikimedia.org/wiki/File:EVD-martillo-011.jpg

Producto de lo anterior, además de que es muy importante que los atletas lancen a gran distancia el martillo, también deben ser muy constantes (haya poca variabilidad en los lanzamientos) debido a que, muchas veces solamente logran obtener un lanzamiento válido (de los tres posibles).

De 10 lanzamientos de entrenamiento se generaron los siguientes resultados (en metros):

María	69,38	66,72	70,46	65,52	66,18	69,39	71,75	76,04	70,12	52,32
Raquel	76,21	70,76	64,61	60,88	74,63	72,45	65,09	58,24	56,34	72,62

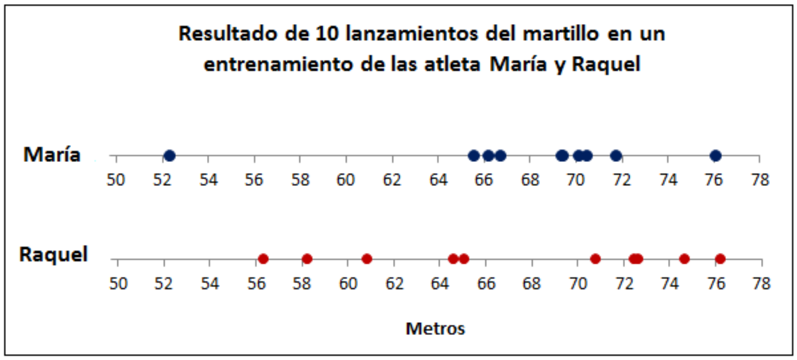
De acuerdo con esta información, analice si el recorrido será una buena medida para determinar cuál de las atletas tuvo lanzamientos menos variables.

Solución:

Para determinar cuál de las atletas es más consistente en los lanzamientos de entrenamiento (son menos variables), al utilizar el recorrido se puede caer en un grave error. Observe que para este caso se tienen los siguientes resultados:

Atleta	**Recorrido en los lanzamientos**
María	$76,04 - 52,32 = 23,72$
Raquel	$76,21 - 56,34 = 19,87$

El valor del recorrido pareciera indicar que los lanzamientos de María son más variables, pero observando los datos se puede notar que, con excepción del lanzamiento de 52,32 metros, el resto de lanzamientos de María son menos variables entre sí que los lanzamientos de Raquel. Esto también se puede observar mediante una representación gráfica aproximada de los datos.



Hay que recordar que el recorrido si interpreta como la mayor diferencia entre lanzamientos: para María fue de 23,72 metros y para Raquel fue de 19,87 metros.

Nota: El ejemplo anterior muestra que se debe tener cuidado al comparar la variabilidad entre grupos de datos por medio del recorrido, debido a que no necesariamente cuando un grupo de datos tiene un mayor recorrido significa que sea el más variable. El recorrido se ve muy afectada por la presencia de uno o más valores que sea extremadamente grande o extremadamente pequeño (o ambos).

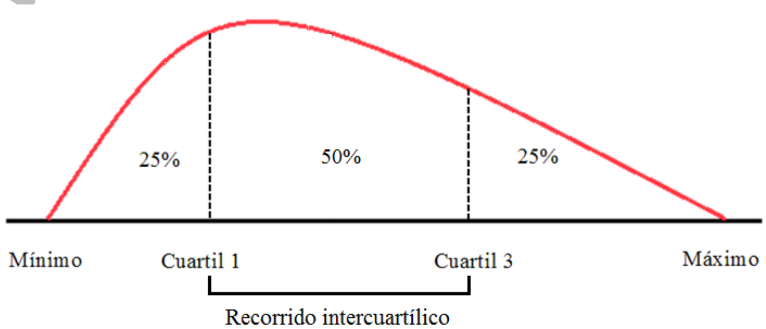
El error que se podría cometer consiste en creer que el grupo de datos con mayor recorrido es el más variable, esto no siempre ocurre, lo que el recorrido nos indica es la mayor diferencia entre dos datos del conjunto.

El recorrido intercuartílico

Permite corregir el problema de los valores extremos que presenta el cálculo del recorrido, porque para su cálculo se considera únicamente el 50% de los datos centrales o intermedios. Por esta razón, su fórmula de cálculo es:

$$Recorrido\ intercuartílico = Tercer\ cuartil - Primer\ cuartil$$

Observe que con esta medida se indica la longitud del ámbito o rango en el cual varía el 50% de los valores centrales de la distribución de datos.



Una forma visual muy práctica para analizar la variabilidad por medio de la articulación del recorrido y el recorrido intercuartílico son los diagramas de caja. Aunque un diagrama de caja incluye medidas estadísticas de orden (mínimo, cuartiles y máximo), permite valorar visualmente la variabilidad de los datos por medio de estas figuras.

Volver al índice de conceptos

Problema 22. Rendimiento en combustible de los vehículos

Para ejemplificar la interpretación del recorrido intercuartílico, volvamos a analizar el problema 9, en donde Pilar y Beatriz discutían sobre el rendimiento de combustible de sus vehículos (en kilómetros recorridos por litro). Para los 10 días observados, los datos de los rendimientos de estos vehículos ordenados de menor a mayor son:

Rendimiento en kilómetros por litro

Auto de Pilar		Auto de Beatriz	
1	11,6	1	9,5
2	11,7	2	9,6
3	12,1	3	10,9
4	12,5	4	12,1
5	13,1	5	12,2
6	13,2	6	12,9
7	13,3	7	13,1
8	13,5	8	13,7
9	14,4	9	14,0
10	15,0	10	14,2

Analice la variabilidad en el rendimiento de estos vehículos desde según la información que proporciona el recorrido intercuartílico.

Solución:

Para determinar el recorrido intercuartílico de ambos rendimientos, se puede recurrir a las medidas de orden que fueron calculadas en el problema 12, los valores se resumen en el siguiente cuadro:

Rendimiento en kilómetros por litro de los autos de Pilar y Beatriz, información recabada por día, en 10 días consecutivos

Medida estadística de orden	Rendimiento de auto de Pilar Km/L	Rendimiento de auto de Beatriz Km/L
Primer cuartil	11,9	10,25
Mediana	13,15	12,55
Tercer cuartil	13,95	13,85
Máximo	15,0	14,2
Mínimo	11,6	9,5

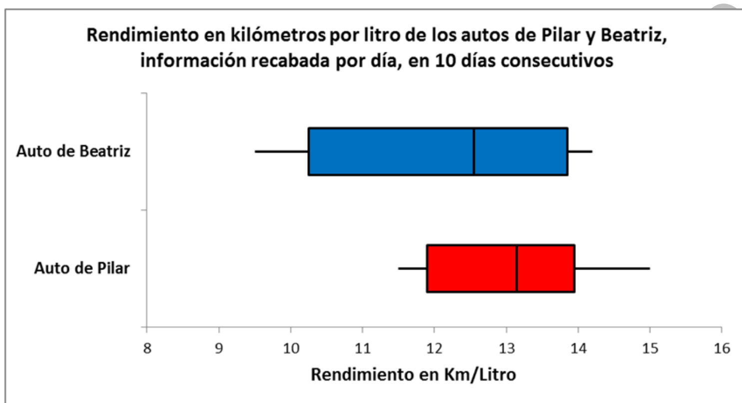
De acuerdo con esta información el recorrido y el recorrido intercuartílico vienen dados por:

Medida	Rendimiento de	Rendimiento de

	auto de Pilar Km/L	auto de Beatriz Km/L
Recorrido	3,4	4,7
Recorrido intercuartílico	2,05	3,6

Para interpretar estas medidas podemos decir que la mayor diferencia en el rendimiento del auto en dos días muestreados fue de 3,4 Km/L para el auto de Pilar y de 4,7 Km/L para el auto de Beatriz. Por otro lado, el 50% de los valores intermedios del rendimiento variaron en un rango de 2,05 Km/L para el auto de Pilar y de 3,6 para el auto de Beatriz. Ambos valores son mayores en el auto de Beatriz.

Estos valores pueden ser analizados en forma visual por medio en un diagrama de cajas:



En general puede notarse que en el período observado el auto de Beatriz refleja una mayor variabilidad en el rendimiento, solamente en el último 25% de los datos el rango fue mayor en las mediciones del auto de Pilar; sin embargo, hay que tener claro que este análisis de variabilidad es parcial porque se basa solamente en las medidas de orden.

Aunque el recorrido y el recorrido intercuartílico se complementan para ilustrar el patrón de variabilidad de un conjunto de datos, sus fórmulas de cálculo no incluyen todos los datos, por lo que podrían no ser suficientemente ilustrativas para todos los problemas. Por esta razón se han definido medidas de variabilidad más completas.

Variancia poblacional:

En términos sencillos, cuando se trabaja con todos los datos de una población (n datos), para el cálculo de la variancia se determina el cuadrado de la diferencia de cada dato con respecto al promedio $(dato - promedio)^2$ y se aplica la siguiente fórmula:

$$Variancia = \frac{Suma\ (dato - promedio)^2}{n}$$

Variancia muestral:

Cuando se trabaja con los elementos de una muestra (n datos) en vez de toda la población, la fórmula de la variancia viene dada por:

$$Variancia = \frac{Suma\ (dato - promedio)^2}{n-1}$$

Volver al índice de conceptos

Desviación estándar:

La variancia se calcula con la intención de generar un indicador general de la variabilidad de todos los datos (se elimina el efecto de valores negativos); sin embargo, las unidades de medida que corresponden al contexto del problema quedan al cuadrado. Por esta razón si se extrae la raíz cuadrada de la variancia se genera otra medida de variabilidad que se llama desviación estándar, la cual tiene las mismas unidades de medida que los datos originales.

$$Desviación\ estándar\ poblacional = \sqrt{\frac{Suma\ (dato - promedio)^2}{n}}$$

$$Desviación\ estándar\ muestral = \sqrt{\frac{Suma\ (dato - promedio)^2}{n-1}}$$

Nota: El cálculo de la variancia tiene una particularidad que no la tiene ninguna de las otras medidas utilizadas previamente, la cual consiste en que su fórmula de cálculo varía si los datos corresponden a toda la población o a una muestra. La diferencia entre la variancia de una muestra o de una población consiste en que para la población la suma de las desviaciones al cuadrado se divide entre el número de datos y en una muestra se divide entre el número de datos menos uno. Las razones por las que se hace esta distinción obedecen a principios de teoría estadística que no conviene discutir en el presente documento (dado que está vinculado con propiedades de los estimadores muestrales para estimar los valores poblacionales desconocidos).

Para hacer patente esta diferencia los libros de texto utilizan diferentes símbolos que incluye también la media aritmética o promedio (aunque se utiliza la misma fórmula). Si se tiene un grupo de datos $X_1, X_2, \ldots, X_n$ que corresponden a variable cuantitativa, entonces se utiliza los símbolos:

- μ representa la media aritmética de una población
- $\overline{X}$ representa la media aritmética de una muestra
- σ^2 representa la variancia de una población
- S^2 representa la variancia de una muestra

Para la variancia se cumple que:

$$\sigma^2 = \frac{Suma\ (dato - promedio)^2}{n} = \frac{Suma\ (X_j - \mu)^2}{n}$$

$$S^2 = \frac{Suma\ (dato - promedio)^2}{n-1} = \frac{Suma\ (X_j - \overline{X})^2}{n-1}$$

Incluso se tienden a utilizar otras formas de simbolizar lo anterior; pero no es necesario que el estudiante profundice en estos temas, lo que tiene que se tiene que tener presente corresponde a diferenciar si los datos pertenecen a toda la población o una muestra, lo cual debe extraerse del contexto donde se plantea el problema, es decir, la redacción del problema planteado debe dar las pautas para que el estudiante pueda establecer si los datos pertenecen a una muestra o a una población.

Problema 23. Enfermos por picadura de mosquito

Volvamos al problema 7, en donde se analizó el número de enfermos por paludismo y dengue que se presentaron entre 1993 y 2008.

Año	**1993**	**1994**	**1995**	**1996**	**1997**	**1998**	**1999**	**2000**
Paludismo	157	134	134	159	141	137	104	49
Dengue	143	422	153	67	407	70	68	129

Año	**2001**	**2002**	**2003**	**2004**	**2005**	**2006**	**2007**	**2008**
Paludismo	34	25	17	30	82	66	27	22
Dengue	235	299	473	221	874	274	583	184

Fuente: http://www.estadonacion.or.cr/index.php/estadisticas/costa-rica/compendio-estadistico/estadisticas-sociales

Por medio de la desviación estándar en el número de defunciones, determine ¿en cuál de las variables se presenta mayor variabilidad: en el número anual de enfermos por paludismo o en el número anual de enfermos por dengue?

Solución:

Para determinar la desviación estándar de los datos en cada caso, primeramente hay determinar el número promedio de enfermos por año. Según el cuadro que se muestra abajo, en este período de 16 años se presentaron 1318 enfermos por paludismo y 4602 enfermos por dengue, entonces el promedio de enfermos para el período es:

Promedio anual de enfermos por paludismo

$$\frac{1318}{16} = 82{,}375$$

Promedio anual de enfermos por dengue

$$\frac{4602}{16} = 287{,}625$$

De acuerdo con la fórmula de la variancia, para cada año debe calcularse la diferencia del número de defunciones con respecto al promedio y luego elevar al cuadrado. Por ejemplo:

Para paludismo:

Para 1993 se tiene $157 - 82{,}375 = 74{,}625$, entonces $(74{,}625)^2 = 5568{,}89063$

Para 1994 se tiene $134 - 82{,}375 = 51{,}625$, entonces $(51{,}625)^2 = 2665{,}14063$

Para 1995 se tiene $134 - 82{,}375 = 51{,}625$, entonces $(51{,}625)^2 = 2665{,}14063$

Del mismo modo, se calculan los valores de los otros años:

Para dengue:

Para 1993 se tiene $143 - 287{,}625 = -144{,}625$, entonces $(-144{,}625)^2 = 20916{,}3906$

Para 1994 se tiene $422 - 287{,}625 = 134{,}375$, entonces $(134{,}375)^2 = 18056{,}6406$

Para 1995 se tiene $153 - 287{,}625 = -134{,}625$, entonces $(-134{,}625)^2 = 18123{,}8906$

En forma idéntica se calculan los otros valores. En el cuadro de abajo aparecen todos los cálculos. Para determinar la variancia la suma de todos los cuadrados debe dividirse por el número de años menos uno, bajo el supuesto de que la información corresponde a un período de 16 años, la cual representa una muestra de la situación de estas enfermedades.

Para el caso del paludismo, como la suma de las diferencias al cuadrado es de 44401,75, entonces la variancia viene dada por:

$$Variancia\ para\ paludismo = \frac{44401{,}75}{15} = 2960{,}11667$$

Para el caso del dengue, la suma de las diferencias al cuadrado es de 724487,75, entonces la variancia viene dada por:

$$Variancia\ para\ dengue = \frac{724487{,}75}{15} = 48299{,}1833$$

Hay que recordar que la desviación estándar es la raíz cuadrara de la variancia, entonces se tendría que la desviación estándar en el número anual de enfermos por paludismo para el período 1993-2008 viene dada por $\sqrt{2960{,}11667}$ que equivale aproximadamente a 54,4 enfermos. Mientras que para este mismo período la desviación estándar en el número anual de enfermos por dengue fue de $\sqrt{48299{,}1833}$ equivale aproximadamente a 219,8 enfermos. Es claro entonces que para el período 1993-2008, el número anual de enfermos por paludismo fue menos variable que el número anual de enfermos por dengue. El siguiente cuadro muestra todos los cálculos:

	Número de enfermos		**Diferencia respecto al promedio**		**Diferencias al cuadrado**	
Año	**Paludismo**	**Dengue**	**Paludismo**	**Dengue**	**Paludismo**	**Dengue**
1993	157	143	74,625	-144,625	5568,89063	20916,3906
1994	134	422	51,625	134,375	2665,14063	18056,6406

1995	134	153	51,625	-134,625	2665,14063	18123,8906
1996	159	67	76,625	-220,625	5871,39063	48675,3906
1997	141	407	58,625	119,375	3436,89063	14250,3906
1998	137	70	54,625	-217,625	2983,89063	47360,6406
1999	104	68	21,625	-219,625	467,640625	48235,1406
2000	49	129	-33,375	-158,625	1113,89063	25161,8906
2001	34	235	-48,375	-52,625	2340,14063	2769,39063
2002	25	299	-57,375	11,375	3291,89063	129,390625
2003	17	473	-65,375	185,375	4273,89063	34363,8906
2004	30	221	-52,375	-66,625	2743,14063	4438,89063
2005	82	874	-0,375	586,375	0,140625	343835,641
2006	66	274	-16,375	-13,625	268,140625	185,640625
2007	27	583	-55,375	295,375	3066,39063	87246,3906
2008	22	184	-60,375	-103,625	3645,14063	10738,1406
Total	1318	4602			44401,75	724487,75
Promedio	82,375	287,625				
Variancia					2960,11667	48299,1833
Desviación estándar					54,4069542	219,770752

Nota: Como puede notarse, el cálculo de la variancia es tedioso, sobre todo en aquellos casos en que hay muchos datos. Realizar todos estos cálculos es de poco interés práctico desde un punto de vista estadístico. Lo adecuado es utilizar una calculadora científica con funciones específicas que generan el valor de la deviación estándar, ya sea correspondiente a una muestra o a una población. En este sentido, el estudiante debe limitarse a incluir los datos a la calculadora y solicitar la medida requerida.

Se debe revisar el manual de usuario de la calculadora para determinar la forma en que se incluyen los datos y en que la máquina devuelve los resultados. En el caso de la desviación estándar, las calculadoras emplean diferentes símbolos para el caso de poblaciones o muestras, por ello debe identificar cada símbolo y su significado. Sobre este tema se profundiza en uno de los videos del curso.

Problema 24. Comparación de notas en Español

Las notas en Español durante el primer trimestre de los dos grupos de décimo año en cierto colegio, vienen dadas por:

Grupo 1	40	67	68	69	70	70	70	72	74	74	75	75	75	76	81	81	82	84	87	98
Grupo 2	48	51	53	68	70	70	71	72	75	75	75	78	79	79	82	89	91	92	93	

El profesor responsable de atender estos grupos desea conocer en cuál de ellos se generó el mejor rendimiento y en cuál las notas fueron más variables. Colabore con el profesor de Español para resolver el problema. Para ello analice y compare los valores del recorrido, recorrido intercuartílico y la desviación estándar.

Solución:

Debemos observar que las notas ya fueron ordenadas de menor a mayor, por lo que es posible determinar las medidas estadísticas de orden con más facilidad.

Posición	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**Total**
Grupo 1	40	67	68	69	70	70	70	72	74	74	75	75	75	76	81	81	82	84	87	98	**1488**
Grupo 2	48	51	53	68	70	70	71	72	75	75	75	78	79	79	82	89	91	92	93		**1411**

La media aritmética de las calificaciones o calificación promedio es

Grupo 1	$\frac{1488}{20} = 74{,}40$
Grupo 2	$\frac{1411}{19} = 74{,}26$

Como en el grupo 1 hay 20 estudiantes (número par) y en el grupo 2 hay 19 estudiantes (número impar), las fórmulas de cálculo de la mediana son diferentes:

- Para el grupo 1 la mediana se calcula por: $\frac{X_{\frac{20}{2}} + X_{\frac{20}{2}+1}}{2} = \frac{X_{10} + X_{11}}{2}$ (el promedio de los datos 10 y 11).
- Para el grupo 2 sería el dato $X_{\frac{19+1}{2}} = X_{10}$ (el dato 10).

Grupo 1	$\frac{74 + 75}{2} = 74{,}5$
Grupo 2	75

Esto quiere decir que también la mediana de las calificaciones es similar en ambos grupos. Para el grupo 1 la mediana significa que el 50% de los estudiantes tuvo una nota de 74,5 o menos y el otro 50% tuvo una nota de 74,5 o más. Del mismo modo se interpreta la mediana del grupo 2. Además, el hecho de que la mediana y el promedio tengan pocas diferencias, es un indicador de que la distribución de las notas es bastante simétrica en ambos casos.

En cuanto a la moda, el grupo 1 tiene dos modas 70 y 75 (con frecuencia de tres estudiantes cada una), mientras que en el grupo 2 es 75 (con frecuencia de tres estudiantes).

> **Nota**: en los casos en que hay pocos datos, como ocurre en este problema, la moda no es una buena medida de tendencia central, porque tiene un comportamiento irregular tal como se indicó antes.

De acuerdo con la fórmula de cálculo de los cuartiles, para determinar el valor del primer cuartil, con $n = 20$, su posición es $\frac{k \cdot (n+1)}{4} = \frac{1 \cdot 21}{4} = 5{,}25$, esto quiere decir que en el grupo 1, primer cuartil se encuentra entre los datos cinco y seis y se cumple que:

$$C_1 = \frac{X_5 + X_6}{2} = \frac{70 + 70}{2} = 70.$$

Esto significa que el 25% de los estudiantes del grupo tuvo una nota de 70 o menos, y el 75% tuvo una nota de 70 o más.

Para el grupo 2, $n = 19$, su posición es $\frac{k \cdot (n+1)}{4} = \frac{1 \cdot 20}{4} = 5$, entonces el primer cuartil corresponde al dato cinco:

$$C_1 = X_5 = 70.$$

Se interpreta de la misma forma que en el grupo 1.

En cuanto al tercer cuartil, para el grupo 1 la posición es $\frac{k \cdot (n+1)}{4} = \frac{3 \cdot 21}{4} = 15{,}75$, entonces el tercer cuartil se encuentra entre los datos 15 y 16. Entonces:

$$C_3 = \frac{X_{15} + X_{16}}{2} = \frac{81 + 81}{2} = 81.$$

Entonces, significa que el 75% de los estudiantes tuvo una nota de 81 o menos y el 25% tuvo una nota de 81 o más.

Para el grupo 2, la posición es $\frac{k \cdot (n+1)}{4} = \frac{3 \cdot 20}{4} = 15$, entonces el tercer cuartil corresponde al dato 15, o sea:

$$C_3 = X_{15} = 82.$$

Se interpreta en forma similar al correspondiente del grupo 1.

Por último la nota más baja del grupo 1 fue de un 40 y en el grupo 2 fue un 48. Por su parte, la nota más alta del grupo 1 fue un 98 y del grupo 2 fue un 93.

El siguiente cuadro presenta un resumen de las principales medidas de posición:

Calificaciones del primer trimestre en la materia de Español de los grupos 1 y 2

Medida	**Grupo 1**	**Grupo 2**
Promedio	74,40	74,26
Mediana	74,50	75,00
Moda	70,00 y 75,00	75,00
Primer cuartil	70,00	70,00
Tercer cuartil	81,00	82,00
Mínimo	40,00	48,00

Máximo	98,00	93,00

En resumen, las medidas estadísticas de posición muestran que el rendimiento es similar en ambos grupos en cuanto a las calificaciones del primer trimestre.

Para analizar la variabilidad se tiene que el recorrido en cada grupo fue:

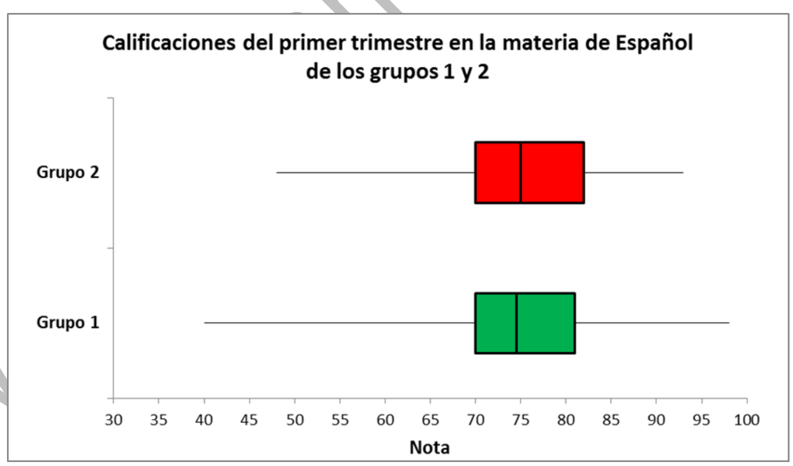
Grupo 1	$98 - 40 = 58$
Grupo 2	$93 - 48 = 45$

Quiere decir que en el grupo 1 la mayor diferencia en las calificaciones fue de 58 puntos, mientras que en el grupo 2 fue de 45 puntos. Por su parte, el recorrido intercuartílico viene dado por:

Grupo 1	$81 - 70 = 11$
Grupo 2	$82 - 70 = 12$

El 50% de las calificaciones intermedias varían en un rango entre 11 puntos para el grupo 1 y 12 puntos para el grupo 2.

Con estos valores y el de la mediana se puede construir un diagrama de cajas.



Nota: De acuerdo con lo anterior, se podría creer que la variabilidad en las calificaciones es similar en ambos grupos. Sin embargo, hay que tener presente que el recorrido intercuartílico es un indicador de variabilidad que nos muestra el rango de variación del 50% de los valores centrales. Entonces este es el mensaje que debe prevalecer al interpretar el recorrido intercuartílico.

Seguidamente se presenta el cálculo de la suma de desviaciones al cuadrado para el grupo 1

$Dato$	$Dato - promedio$	$(Dato - promedio)^2$

40	$40-74{,}4=$ -34,4	$(-34{,}4)^2=$ 1183,36
67	$67-74{,}4=$ -7,4	$(-7{,}4)^2=$ 54,76
68	-6,4	40,96
69	-5,4	29,16
70	-4,4	19,36
70	-4,4	19,36
70	-4,4	19,36
72	-2,4	5,76
74	-0,4	0,16
74	-0,4	0,16
75	0,6	0,36
75	0,6	0,36
75	0,6	0,36
76	1,6	2,56
81	6,6	43,56
81	6,6	43,56
82	7,6	57,76
84	9,6	92,16
87	12,6	158,76
98	23,6	556,96
Total	**0**	**2328,8**

Entonces la varianza del grupo 1 sería: Variancia $=\frac{2328{,}8}{20}=116{,}44$

Para el grupo 1 la desviación estándar de las calificaciones viene dada por $\sqrt{116{,}44}\approx 10{,}79$. Entonces la desviación estándar de las calificaciones del grupo 1 es 10,79 puntos. Se puede repetir el procedimiento para el grupo 2:

$\boldsymbol{Dato}$	$\boldsymbol{Dato-promedio}$	$\boldsymbol{(Dato-promedio)^2}$
48	$48-74{,}26=$ -26,26	$(-26{,}26)^2=$ 689,59
51	$51-74{,}26=$ -23,26	$(-23{,}26)^2=$ 541,03
53	-21,26	451,99
68	-6,26	39,19
70	-4,26	18,15
70	-4,26	18,15
71	-3,26	10,63
72	-2,26	5,11
75	0,74	0,55
75	0,74	0,55
75	0,74	0,55
78	3,74	13,99
79	4,74	22,47
79	4,74	22,47
82	7,74	59,91
89	14,74	217,27
91	16,74	280,23
92	17,74	314,71
93	18,74	351,19
Total	**0**	**3057,73**

Entonces la variancia del grupo 2 sería:

$$\text{Varianca} = \frac{3057,73}{19} = 160,93$$

Y la desviación estándar es $\sqrt{160,93} \approx 12,69$ puntos.

Puede notarse que al utilizar una medida de variabilidad mucho más precisa, las calificaciones del grupo 2 se muestran más variables que las del grupo 1, la desviación estándar es prácticamente dos unidades mayor.

Nota: si utilizamos la calculadora, no sería necesario realizar los cálculos repetitivos y tediosos al determinar la variancia.

Problema 25. Contrato de trabajo con empresa turística

Retomemos nuevamente el problema 20 donde don Juan debe escoger entre dos empresas de turismo que le ofrecen salario que varía de un mes a otro. Recordemos que, en igualdad de condiciones, debe escoger la empresa que le ofrezca los salarios menos variables.

Mes	Temporada	Empresa A	Empresa B
Enero	Alta	1 300 000	1 050 000
Febrero	Alta	1 200 000	1 050 000
Marzo	Media	1 000 000	850 000
Abril	Media	800 000	850 000
Mayo	Baja	600 000	650 000
Junio	Baja	600 000	650 000
Julio	Alta	1 100 000	1 050 000
Agosto	Media	700 000	850 000
Setiembre	Baja	500 000	650 000
Octubre	Baja	500 000	650 000
Noviembre	Media	800 000	850 000
Diciembre	Alta	1 100 000	1 050 000

Dado que el salario promedio que ofrecen ambas empresas es de 850 000 colones, entonces determine la empresa que le ofrece los salarios más variables con base el cálculo de la desviación estándar y ayude a don Juan a tomar la decisión.

Solución:

Si no se utiliza la calculadora para determinar directamente la desviación estándar, entonces se deben se deben realizar los siguientes cálculos:

Mes	Empresa A	Empresa B	Diferencias A	Diferencias B	Diferencias al cuadrado A	Diferencias al cuadrado B
Enero	1300000	1050000	450000	200000	202500000000	40000000000
Febrero	1200000	1050000	350000	200000	122500000000	40000000000
Marzo	1000000	850000	150000	0	22500000000	0
Abril	800000	850000	-50000	0	2500000000	0
Mayo	600000	650000	-250000	-200000	62500000000	40000000000
Junio	600000	650000	-250000	-200000	62500000000	40000000000
Julio	1100000	1050000	250000	200000	62500000000	40000000000
Agosto	700000	850000	-150000	0	22500000000	0
Setiembre	500000	650000	-350000	-200000	122500000000	40000000000
Octubre	500000	650000	-350000	-200000	122500000000	40000000000
Noviembre	800000	850000	-50000	0	2500000000	0
Diciembre	1100000	1050000	250000	200000	62500000000	40000000000
Total	**10200000**	**10200000**	**0**	**0**	**870000000000**	**320000000000**

Debido a que los datos corresponden a todos los salarios del año, entonces la fórmula de cálculo empleada corresponde a la variancia de la población. Entonces la variancia y la desviación estándar para los salarios ofrecidos por cada empresa serían:

Empresa	Variancia	Desviación estándar
A	$\frac{870\ 000\ 000\ 000}{12} = 72\ 500\ 000\ 000$	269 258,24
B	$\frac{320\ 000\ 000\ 000}{12} = 26\ 666\ 666\ 666{,}67$	163 299,32

El valor de la desviación estándar deja en evidencia lo que se había observado visualmente en el problema 20: los salarios de la empresa B son menos variables y debería seleccionar dicha empresa.

10. Uso de medidas relativas

La mayoría de los problemas analizados previamente se vincularon con análisis estadísticos que involucraron medidas absolutas de posición o variabilidad. Sin embargo, tal como se estudió en los problemas 5 y 6, se requirió de porcentajes para realizar comparaciones relativas, existen fenómenos que requieren de un análisis relativo mediante el uso de las medidas de posición o de variabilidad.

Estandarización

Consiste en una estrategia que se utiliza para realizar comparaciones con datos que pertenecen a diferentes contextos o diferentes magnitudes. Para que los datos sean comparables se les debe estandarizar: para cada uno se calcula su diferencia respecto al promedio y el resultado se divide por la desviación estándar:

$$dato\ estandarizado = \frac{dato - promedio}{desviación\ estándar}$$

Con esta medida se elimina el efecto de la unidad de medida de los datos debido a que se modifica su posición, llevándolos a un estándar comparativo. Un dato estandarizado puede ser positivo, negativo o incluso nulo.

Volver al índice de conceptos

Problema 26. Comparación entre notas de examen de admisión

Pedro es un estudiante que está por ingresar a la universidad y realiza el examen de admisión en dos universidades. En el examen correspondiente a la universidad A obtuvo un 683 en una escala de 0 a 800; mientras que en el examen correspondiente a la universidad B obtuvo un 413 en una escala de 0 a 500. La calificación promedio y la desviación estándar de las calificaciones de todos los estudiantes que realizaron la prueba fueron:

Examen	Nota promedio	Desviación estándar
Universidad A	578	110
Universidad B	394	78

Pedro decide realizar trámites de admisión solamente en aquella universidad en la que obtuvo un mejor resultado en relación con todos los estudiantes que presentaron las pruebas, pero no sabe cómo puede realizar la comparación y efectuar la escogencia apropiada. Ayude a Pedro a resolver este problema.

Solución:

El problema que enfrenta Pedro es que la escala de calificación es diferente en las dos pruebas, por ello las notas que obtuvo no son comparables entre sí, a no ser que se realice una valoración relativa. Hay que tener presente que Pedro desea comparar sus notas en relación con todos los estudiantes que realizaron la prueba. Una forma de efectuar la comparación consiste en *estandarizar* o *tipificar* sus calificaciones tomando como referencia la calificación promedio y la desviación estándar de todas las notas.

Según la fórmula para estandarizar se tiene:

$$Nota\ estandarizada = \frac{nota\ obtenida - promedio}{desviación\ estándar}$$

Aplicando esta fórmula para las notas que obtuvo Pedro se tiene:

Examen	Nota promedio	Desviación estándar	Nota de Pedro	Nota estandarizada
Universidad A	578	110	683	$\frac{683-578}{110}=0{,}95$
Universidad B	394	78	413	$\frac{413-394}{78}=0{,}24$

De acuerdo con estos resultados, Pedro debería realizar el proceso de admisión en la Universidad A, debido a que tuvo un mayor rendimiento relativo tomando como referencia a los estudiantes que realizaron los exámenes.

Nota: Se debe tener presente que si se desea comparar la posición de datos que pertenecen a contextos distintos y que no son comparables en términos absolutos (tal como se presentó en este problema), entonces se debe realizar un análisis relativo, en donde la estandarización es una posibilidad para efectuar dicha comparación.

Otro tipo de problemas en los que se requiere realizar una comparación relativa, son los de análisis de la variabilidad en grupos de datos que tienen magnitudes o escalas diferentes. Observe el problema siguiente:

Problema 27. Salarios relativos de técnicos en refrigeración

Manuel es técnico en refrigeración y trabaja en la empresa Manolitos y Asociados, tiene un salario mensual de ₡585 000, y un amigo suyo llamado Felipe, que tiene la misma profesión, trabaja para la empresa Libertad R.L. y tiene un salario mensual de ₡650 000. Felipe le recrimina a Manuel indicándole que la empresa Manolitos y Asociados, no valora su trabajo pues le mantiene un salario muy bajo.

Por su parte Manuel se defiende al indicar que lo que ocurre es que la empresa Manolitos y Asociados tiene salarios más bajos que Libertad R. L.; pero le dice a Felipe que en términos relativos al comparar el salario con el resto de empleados de la empresa él tiene un mejor salario. La siguiente información representa las principales medidas estadísticas para todos los empleados de las dos empresas.

Medidas estadísticas relacionadas con los salarios de los trabajadores de las empresas Manolitos y Asociados y Libertad R.L (cantidades en colones)

Empresa	Promedio	Mediana	Desviación estándar
Manolitos y Asociados	510 250	515 000	35 450
Libertad R. L.	593 035	589 500	41 202

Utilice esta información para determinar si Manuel tiene razón de que, en términos relativos, al comparar el salario con el resto de empleados de la empresa, él ostenta un mejor salario que Felipe:

Solución

Al comparar el salario de Manuel y Felipe, es claro que Felipe tiene un mejor salario. No obstante, el problema solicita que se realice una comparación relativa de los salarios respecto al resto de empleados de cada empresa, por lo que se deben estandarizar dichos valores.

Para Manuel: $\frac{585\,000 - 510\,250}{35\,450} = 2{,}11$

Para Felipe: $\frac{650\,000 - 593\,035}{41\,202} = 1{,}38$

En términos relativos, al comparar los salarios con respecto al salario promedio de los trabajadores en cada empresa se concluye que Manuel tiene razón, en el sentido de que, en relación con el resto de empleados de la empresa su salario es mayor que el de Felipe.

Nota: Observe que la comparación relativa por medio de la estandarización permite realizar un análisis de la posición relativa de cada dato, pero elimina las unidades de medición, por lo que muchas veces puede ofrecer un resultado de interés práctico. Por ejemplo, de qué le sirve a Manuel tener un mejor salario relativo que Felipe, si en términos absolutos su salario es mucho menor.

No obstante, hay problemas donde este análisis relativo tiene importancia práctica, tal como se evidenció en el problema 26.

Problema 28. Rendimiento en Matemáticas en dos trimestres

Juan es un estudiante de noveno año y obtuvo un 88,0 en el primer examen de Matemáticas del segundo trimestre, su madre le llama la atención pues obtuvo una menor calificación que la obtenida en el primer examen del primer trimestre, que fue un 90,9. Juan responde que este examen estuvo más difícil y que su rendimiento, en términos relativos más bien mejoró. Para justificarse le muestra el siguiente cuadro, donde se incluye la calificación promedio y la desviación estándar de las calificaciones de todos los novenos de su colegio, en cada trimestre.

Medidas estadísticas absolutas	**I Examen**	
	Primer trimestre	*Segundo trimestre*
Promedio	86,5	83,3
Desviación estándar	9,18	8,51

Analice la información que suministró Juan para determinar si tiene la razón o simplemente intenta justificarse con su madre.

Solución

Se requiere hacer una comparación relativa de la nota obtenida en cada trimestre respecto a la calificación promedio de todos los estudiantes de noveno año. Para ello se recurre a la fórmula:

$$Nota\ estandarizada = \frac{nota\ obtenida - promedio}{desviación\ estándar}$$

En el siguiente cuadro se resumen los resultados de los cálculos:

Medidas estadísticas	I Examen	
	Primer trimestre	Segundo trimestre
Calificación de Juan	90,9	88,0
Promedio	86,5	83,3
Desviación estándar	9,18	8,51
Nota estandarizada	**0,48**	**0,55**

Las notas estandarizadas demuestran que efectivamente Juan tiene razón, debido a que al comparar sus calificaciones con la calificación promedio del todos los estudiantes de noveno año de la institución, su rendimiento su fue relativamente mejor en el examen del segundo trimestre que en el del primer trimestre.

En los problemas anteriores se analizó la posición relativa, seguidamente se analiza la variabilidad relativa.

Volver al índice de conceptos

Coeficiente de variación

Consiste en una relación estadística que permite comparar la variabilidad de diferentes grupos de datos que provenientes de contextos diferentes o que poseen diferentes magnitudes. Se calcula mediante la fórmula:

$$Coeficiente\ de\ variación: \frac{desviación\ estándar}{promedio} \cdot 100$$

Al igual que en los análisis de posición relativa, el coeficiente de variación elimina la unidad de medida para poder hacer una comparación de la variabilidad en forma equitativa.

Problema 29. Variación en la producción de plantas industriales

En tres plantas industriales que pertenecen a una misma compañía, se producen ciertos componentes electrónicos. La producción mensual de cada una de ellas se resume en el siguiente cuadro:

	Planta A	Planta B	Planta C
Promedio mensual	140 500	83 200	254 300
Desviación estándar	45 325	33 456	65 350

El gerente de la compañía está realizando un análisis de la producción de las plantas, y desea establecer una estrategia que optimice la producción y que reduzca la variabilidad. Para iniciar el trabajo desea comparar la variabilidad en la producción de las plantas, para determinar en cuál de ellas se genera la mayor dispersión, pero enfrenta el problema de que las desviaciones estándar no son comparables debido a que las producciones promedio son muy diferentes entre las plantas. Utilice sus conocimientos estadísticos para apoyar el gerente a resolver el problema.

Solución:

Al igual que en el análisis anterior enfrentamos un problema para el cuál las medidas estadísticas absolutas no son comparables. Entonces se requiere encontrar una relación estadística que permita comparar en forma relativa la variabilidad de los datos.

Si se utiliza la fórmula del coeficiente de variación para las tres plantas se tiene que:

	Planta A	Planta B	Planta C
Promedio mensual	140 500	83 200	254 300
Desviación estándar	45 325	33 456	65 350
Coeficiente de variación	$\frac{45\ 325}{140\ 500} \cdot 100 = 32{,}3$	$\frac{33\ 456}{83\ 200} \cdot 100 = 40{,}2$	$\frac{65\ 350}{254\ 300} \cdot 100 = 25{,}7$

De los cálculos anteriores puede notarse que en la Planta B se presenta la mayor variabilidad relativa y en la planta C la menor.

Problema 30. Defunciones según causa de muerte

La siguiente información corresponde algunas medidas estadísticas vinculadas con el número anual de defunciones por cada 10 000 habitantes, que se presentaron en Costa Rica en el período 1990-2010.

Costa Rica: Medidas estadísticas del número de defunciones anuales según la causa de muerte para tres grupos de enfermedades. Período 1990-2010 (defunciones por 10 000 habitantes)

	Causa de muerte		
Medidas estadísticas	**Aparato respiratorio**	**Aparato digestivo**	**Tumores (cáncer)**
Promedio	3,8	2,6	8,3
Mediana	3,8	2,6	8,2
Desviación Estándar	0,46	0,29	0,46

Fuente: http://www.estadonacion.or.cr/

Algunos expertos han manifestado que la variabilidad relativa en el número de defunciones anuales es menor en aquellas causas que obedecen a problemas con el aparato digestivo. Determine si esta afirmación es verdadera.

Solución

Se requiere comparar la variabilidad en términos relativos. Entonces se utiliza el coeficiente de variación. La fórmula de cálculo es:

$$Coeficiente\ de\ variación\colon \frac{desviación\ estándar\ anual\ de\ defunciones}{promedio\ anual\ de\ defunciones} \cdot 100$$

El cuadro siguiente incluye, para cada una de las tres enfermedades, el valor que tomó dicho coeficiente.

Medidas estadísticas	**Causa de muerte**		
	Aparato respiratorio	**Aparato digestivo**	**Tumores (cáncer)**
Promedio	3,8	2,6	8,3
Desviación Estándar	0,46	0,29	0,46
Coeficiente de variación	12,1	11,2	5,5

Puede notarse que la causa de muerte con menor variabilidad relativa son los tumores, por lo que la afirmación que hacen algunos expertos es falsa.

Nota: En resumen, en cuanto a los análisis de variabilidad el estudiante debe observar dos posibles análisis: el de variabilidad absoluta y el de variabilidad relativa.

1) **Variabilidad absoluta**: al medir la variabilidad absoluta se utilizan indicadores de la variabilidad general de los datos. Las medidas más utilizadas son el recorrido, el recorrido intercuartílico y la desviación estándar (raíz cuadrada de la variancia). Cada una de estas medidas mide diferentes propiedades de la variabilidad presente en los datos

2) **Variabilidad relativa:** la variabilidad relativa normalmente se mide por medio del coeficiente de variación, aunque es posible definir otras medidas. Se utiliza para comparar la variabilidad entre grupos que incluyen datos de diferente naturaleza o diferente magnitud, normalmente se aplica cuando los promedios de los grupos de datos son muy diferentes.

III. Probabilidades

1. Conceptos previos

Situación aleatoria

Una situación o experimento se dice que es aleatorio si los resultados que puede generar no pueden ser predichos sino que dependen del azar. Por ejemplo:

a) El resultado de un sorteo de la Lotería Nacional.

b) El resultado obtenido al lanzar al aire una moneda o un dado.

c) Predecir si va a llover o no en un día en particular.

d) Adquirir una enfermedad si se expone al contagio de algún virus.

Situación determinista

Una situación o experimento se dice que es determinista si los resultados que se puede generar pueden ser predichos sin necesidad de realizar la experiencia. Por ejemplo:

a) Precio a pagar al comprar cinco litros de leche si el precio por litro de es 750 colones.

b) Determinación de día que le sigue al domingo.

c) Determinar el número de meses del año que tiene exactamente 30 días.

Espacio muestral

Es el conjunto de los posibles resultados simples de un experimento aleatorio.

a) Al lanzar al aire un dado numerado de uno a seis, el espacio muestral está dado por el conjunto: $S = \{1, 2, 3, 4, 5, 6\}$.

b) Si se lanza una moneda nacional y un dado numerado de uno a seis, el espacio muestral correspondiente es el conjunto: $S = \{E1, E2, E3, E4, E5, E6, C1, C2, C3, C4, C5, C6\}$. La letra representa el resultado de la moneda y el número el resultado del dado.

Volver al índice de conceptos

c) De un grupo de décimo año de cierto colegio se selecciona aleatoriamente un estudiante para que represente al grupo en una actividad general de la institución. El espacio muestral está constituido por todos los estudiantes de ese grupo.

Punto muestral

Los puntos muestrales son los resultados simples de un experimento. En términos más simples, los puntos muestrales son los eventos de un espacio muestral. Por ejemplo, al lanzar un dado numerado de uno a seis, cada uno de los posibles resultados se considera un punto muestral de este experimento.

Eventos aleatorios

Los eventos aleatorios se consideran subconjuntos de un espacio muestral, un evento se considera un resultado posible de un experimento.

a) Al lanzar al aire un dado numerado de uno a seis, los siguientes son eventos aleatorios
A: obtener un número par, se tiene que $A = \{2,4,6\}$.
B: obtener un número primo, entonces $B = \{2,3,5\}$.

b) De un grupo de décimo año de cierto colegio se selecciona aleatoriamente un estudiante para que represente al grupo en una actividad general de la institución. Considere los eventos:
A: seleccionar una mujer. Eeste evento estaría constituido por todas las mujeres del grupo
B: seleccionar un estudiante que haya nacido en el mes de febrero. Este evento está constituido por todos los estudiantes que cumplen años en febrero.

Evento imposible

Representa al evento que no tienen puntos muestrales, es decir dicho evento no puede ocurrir. Normalmente se representa con ϕ. Por ejemplo:

a) En el experimento en que se debe seleccionar un estudiante de décimo año de cierto colegio, se considera el evento de que el estudiante haya nacido en el mes febrero. Sin embargo, si no hay estudiantes en esta condición, entones el evento se dice que es imposible debido a que no tiene puntos muestrales.

b) Si se lanzan dos dados numerados de uno a seis cada uno de ellos, y se suman los puntos obtenidos, el evento de obtener el número uno es imposible.

Evento seguro

Representa el evento que se tiene la seguridad absoluta de que va a ocurrir.

a) Se lanzan dos monedas nacionales y se considera el evento de obtener menos de tres escudos. Este evento es seguro o cierto debido a que al lanzar dos monedas nacionales se pueden obtener dos escudos, un escudo o ningún escudo, por ello se sabe con certeza que el número de escudos es menor que tres.

b) Se lanzan dos dados numerados de uno a seis y se suman los puntos obtenidos. El evento de obtener un número no mayor de 12 es una evento seguro, debido a que el número máximo que se puede obtener en este experimento es un 12 que se obtiene si los dos dados dieron por resultado un seis.

Volver al índice de conceptos

2. Operaciones con eventos

Cuando se vinculan dos o más eventos por medio de operaciones de conjuntos se generan nuevos eventos.

Unión de eventos

Si A y B son eventos de un espacio muestral S, la ocurrencia del evento A o del evento B (o de ambos), corresponde a los que se denomina *unión* de los eventos A y B, se denota con $A \cup B$, e incluye la reunión de los puntos muestrales de A y los de B.

a) Si se lanzan dos dados numerados de uno a seis cada uno y se suman los puntos. Se consideran los eventos:

   A: obtener un número par. Se tiene que $A = \{2,4,6\}$
   B: obtener un número primo. Entonces $B = \{2,3,5\}$

   Entonces la unión de los eventos A y B viene dada por $A \cup B = \{2,3,4,5,6\}$

b) De un grupo de décimo año de cierto colegio se selecciona aleatoriamente un estudiante para que represente al grupo en una actividad general de la institución. Considere los eventos:
   A: seleccionar una mujer.
   B: seleccionar un estudiante que haya nacido en el mes de febrero.

   La unión de estos eventos $A \cup B$ incluye a todas las estudiantes mujeres y también a los varones que cumplen años en febrero.

Intersección de eventos

Si A y B son eventos de un espacio muestral S, la ocurrencia de los eventos A y B al mismo tiempo se interpreta como la *intersección* de los eventos A y B, y se denotada con $A \cap B$. Esta intersección incluye los puntos muestrales que están en A y B a la vez.

a) Si se lanzan dos dados numerados de uno a seis cada uno y se suman los puntos. Se consideran los eventos:

   A: obtener un número par. Se tiene que $A = \{2,4,6\}$.
   B: obtener un número primo. Entonces $B = \{2,3,5\}$.

   Entonces la intersección de los eventos A y B viene dada por $A \cap B = \{2\}$.

b) De un grupo de décimo año de cierto colegio se selecciona aleatoriamente un estudiante para que represente al grupo en una actividad general de la institución. Considere los eventos:
   A: seleccionar una mujer.
   B: seleccionar un estudiante que haya nacido en el mes de febrero.

   La intersección de estos eventos $A \cap B$ incluye solamente las estudiantes mujeres que cumplen años en febrero.

Volver al índice de conceptos

Complemento de un evento

Si A es un evento de un espacio muestral S, la no ocurrencia del evento A se interpreta como la ocurrencia del *complemento* de A, y se representa con A^c. Este incluye los puntos muestrales que no están en A.

a) Se lanzan dos dados numerados de uno a seis cada uno y se suman los puntos. Se consideran los eventos:

   A: obtener un número par, se tiene que $A = \{2,4,6\}$.
   B: obtener un número primo, entonces $B = \{2,3,5\}$.

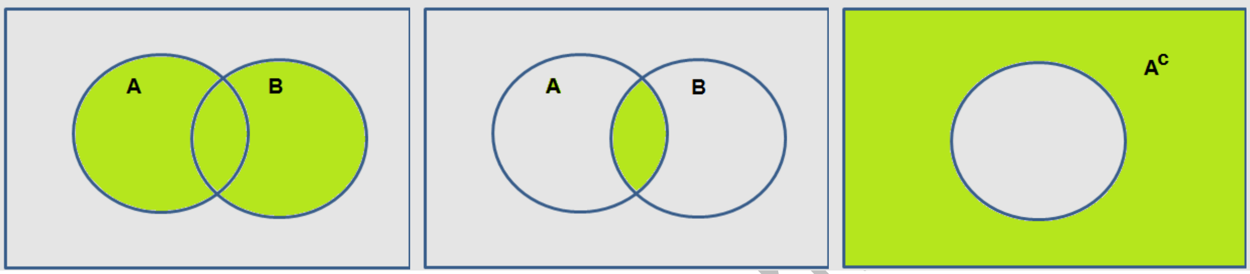
   Entonces el complemento del evento A viene dado por $A^c = \{1,3,5\}$, que significa obtener un número impar. El complemento del evento B viene dada por $B^C = \{1,4,6\}$, que significa no obtener un número primo.

b) De un grupo de décimo año de cierto colegio se selecciona aleatoriamente un estudiante para que represente al grupo en una actividad general de la institución. Considere los eventos:
   A: seleccionar una mujer.
   B: seleccionar un estudiante que haya nacido en el mes de febrero.

Entonces el complemento del evento A viene dado por A^c corresponde al evento de seleccionar un varón. El complemento del evento B viene dada por B^C corresponde al evento de seleccionar un estudiante que no cumple años en febrero.

Nota: Las operaciones anteriores también se pueden representar por los llamados *diagramas de Venn*, en los cuales el espacio muestral se representa con una figura cerrada (normalmente un polígono) y en su interior se incluyen los eventos por medio también como figuras cerradas (normalmente se utilizan círculos), las partes en verde representa el resultado de las operaciones:

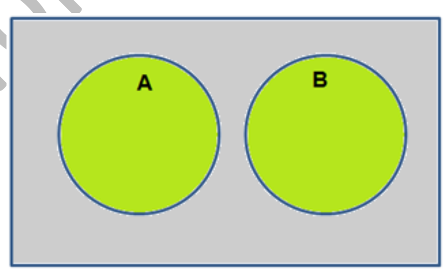
$A \cup B$ $\qquad$ $A \cap B$ $\qquad$ A^c



3. Eventos mutuamente excluyentes

Eventos mutuamente excluyentes

Si A y B son eventos de un espacio muestral S, se dice que los eventos A y B son mutuamente excluyentes si no tienen puntos muestrales en común, es decir $A \cap B = \phi$.



Se lanza un dado numerado de uno a seis y se consideran los eventos:

A: obtener un número par. $A = \{2,4,6\}$.
B: obtener un número impar. $B = \{1,3,5\}$.

Los eventos A y B son mutuamente excluyentes debido a que no tienen puntos muestrales en común.

Volver al índice de conceptos

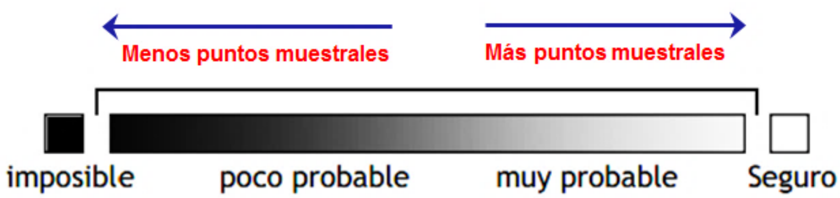
4. Eventos más probables, menos probables o igualmente probables

El término de probabilidad es utilizado cotidianamente, aunque no siempre con la precisión matemática que se requiere. En el lenguaje común se asocia el concepto de *probable* con el de *posible*, es decir se dice que un hecho es probable que ocurra si existe la posibilidad de que realmente ocurra. Sin embargo, esto puede resultar ambiguo y se requiere una connotación matemática más precisa.

Si se tiene un espacio muestral que tiene *n* puntos muestrales, para los cuales no hay preferencia de ocurrencia, se dice entonces que los puntos muestrales son igualmente posibles o igualmente probables (o equiprobables). Por ejemplo, cuando se lanza un dado numerado de uno a seis se supone que los seis posibles resultados son igualmente probables o equiprobables.

Eventos más y menos probables

Si los puntos muestrales de un espacio muestral *S* son equiprobables, además *A* y *B* son eventos de *S*, entonces se dice que *A* es más probable que *B* si posee más puntos muestrales que *B*. En caso contrario si *A* tiene menos puntos muestrales se dice que *A* es menos probable que *B*.

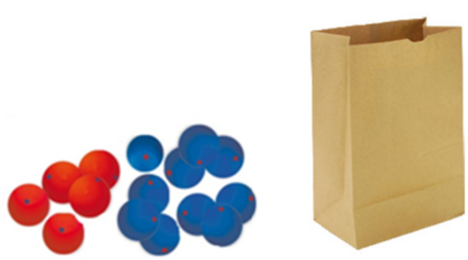


Eventos igualmente probables

Si los puntos muestrales de un espacio muestral *S* son equiprobables, además *A* y *B* son eventos de *S* entonces se dice que *A* es igualmente probable que *B* si posee la misma cantidad de puntos muestrales que *B*.

Problema 31. Selección de bolas rojas y azules

Suponga que en una bolsa de papel se incluyen cinco bolas rojas y diez bolas azules todas del mismo peso, textura y tamaño.



1. Si se extrae una bola en forma aleatoria (sin ver qué color se está escogiendo) ¿qué color es más probable que salga: azul o rojo?

2. ¿De qué manera se deberían variar las cantidades de bolas rojas y azules para que exista justicia o equidad en las posibilidades de selección?

Solución:

1. El espacio muestral correspondiente a este problema incluye las 15 bolas. Para responder la primer pregunta se debe suponer que todas las bolas tienen igual probabilidad de ser seleccionadas, entonces al existir más bolas azules en la bolsa y extraer una de ellas en forma aleatoria, sería de esperar que existan más probabilidades de que la bola seleccionada sea azul. Por ello se dice que la probabilidad de extraer una bola azul es mayor que la probabilidad de extraer una bola roja.

2. En la segunda pregunta, si se considera el mismo criterio o argumento empleado para responder la primera, entonces para que exista equidad se requiere que haya una misma cantidad de bolas azules y rojas en la bolsa, entonces también debería modificarse el total de bolas por un número par, donde la mitad sean rojas y la otra mitad azules. En este caso se diría que los eventos *extraer una bola azul* y *extraer una bola roja* son *igualmente probables*.

Nota: Para la solución del problema anterior, el requisito de que todas bolas incluidas en la bolsa tengan igual peso, textura y tamaño, resulta fundamental para justificar que todas las bolas tienen la misma probabilidad de ser escogidas, a este concepto se le llama equiprobabilidad o igual probabilidad.

Volver al índice de conceptos

Problema 32. Giro de una ruleta

Suponga que se hace girar la siguiente ruleta

1. ¿Qué evento tiene mayor probabilidad: obtener un resultado negro u obtener un número múltiplo de tres?

2. ¿Qué evento tiene mayor probabilidad: obtener un siete u obtener un cinco?

3. ¿Qué supuestos deben plantearse para que las respuestas anteriores sean correctas?

Respuestas:

1. En este caso se considera un espacio muestral que incluye ocho regiones igualmente probables. Entonces, al igual que en el problema anterior, para responder qué evento es más probable basta con determinar en cuál de los eventos existen más casos o resultados posibles a favor. Al revisar la ruleta se encuentran que existen cuatro regiones negras, mientras que los múltiplos de tres que aparecen son 3 y 6, por lo que solamente hay dos regiones a favor de este último evento. Por esta razón es más probable obtener un resultado negro.

2. Solamente existe una región que incluye al número siete y también solamente existe una región que incluye el número cinco, por ello se consideran los eventos obtener un siete y obtener un cinco como igualmente probables.

3. Tal como se indicó en el problema anterior, se requiere establecer condiciones al juego para que las respuestas dadas anteriormente sean válidas. Entre las condiciones básicas que deben estar presentes, la más importante es que las ocho regiones que incluye la ruleta deben ser igualmente probables. Obviamente esta condición requiere de otras tales como que la ruleta esté bien centrada para que gire uniformemente, la ruleta debe hacerse girar con suficiente fuerza para que el resultado no se pueda predecir, que la aguja se mantenga estática, entre otras condiciones.

5. Enfoque clásico de probabilidad

Volver al índice de conceptos

Concepto clásico de probabilidad

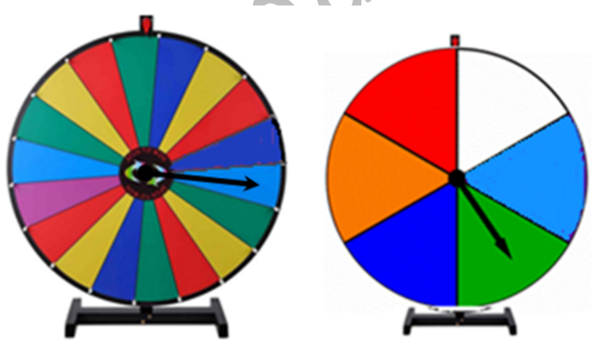
Si un experimento tiene n resultados igualmente probables (es decir el espacio muestral tiene n elementos) y un evento A cualquiera tiene a su favor k resultados ($k \leq n$) entonces se dice que la probabilidad de que el evento A ocurra (se representa con $P(A)$) viene dada por la razón:

$$P(A) = \frac{Número\ de\ resultados\ a\ favor\ de\ A}{Total\ de\ resultados\ del\ experimento} = \frac{k}{n}$$

Nota: La definición anterior se le llama *definición clásica de probabilidad*, por medio de ella se puede encontrar la probabilidad de cualquier evento siempre que se conozca el número total de resultados del experimento y el número de resultado a favor del evento.

Problema 33. Comparación entre ruletas

Suponga que se hacen girar las siguientes ruletas, las cuales están bien equilibradas y en cada una de ellas las regiones son equiprobables.



Ruleta 1 Ruleta 2

¿En cuál ruleta existe mayor probabilidad de que salga favorecido el color:

1. celeste?
2. rojo?
3. azul?

Solución:

Se debe tener cuidado para analizar este tipo de problemas, porque se incluyen dos ruletas con un número de regiones diferente, por lo que la cantidad absoluta de regiones a favor de cada color no es comparable entre las ruletas. Entonces se debe recurrir a comparaciones relativas por medio de la definición clásica de probabilidad.

Seguidamente se presenta el análisis para cada caso.

1. Para determinar en cuál de las ruletas el color celeste es más probable, se debe observar que en la primer ruleta hay dos regiones de color celeste de un total de 18 regiones que incluye la ruleta. Por ello la probabilidad del color celeste en esta ruleta es:

$$P(Celeste) = \frac{2}{18} = \frac{1}{9}$$

Mientras que en la segunda ruleta solamente una región es de color celeste de un total de 6 regiones. Entonces la probabilidad es:

$$P(Celeste) = \frac{1}{6}$$

Como se tiene que $\frac{1}{9} < \frac{1}{6}$, entonces es más probable obtener un color celeste en la ruleta 2.

2. En cuanto al color rojo, haciendo el mismo análisis de ítem anterior, se tiene que en la ruleta 1 hay cuatro regiones rojas de un total de 18. Por esta razón en esta ruleta la probabilidad es:

$$P(Rojo) = \frac{4}{18} = \frac{2}{9}$$

En la ruleta 2 hay una región roja de un total de seis, se tiene que:

$$P(Rojo) = \frac{1}{6}$$

Debido a que se cumple que $\frac{2}{9} > \frac{1}{6}$, entonces la probabilidad de seleccionar el color rojo es mayor en la ruleta 1.

3. Del mismo modo, para el color azul en la ruleta 1 se tienen tres regiones azules:

$$P(Azul) = \frac{3}{18} = \frac{1}{6}$$

La ruleta 2 incluye una región azul:

$$P(Azul) = \frac{1}{6}$$

Entonces la obtención de una región color azul es igualmente probable en ambas ruletas. Esto significa que los eventos son equiprobables.

Nota: En el problema anterior, ante la imposibilidad de realizar una comparación con los valores absolutos, se necesitó buscar una medida relativa que permitiera realizar la comparación. El problema de las comparaciones relativas ha sido objeto de mucha discusión en el presente documento dentro de los análisis estadísticos.

Este problema se pudo resolver también haciendo una comparación porcentual tal como se resume en el siguiente cuadro:

	Ruleta 1		**Ruleta 2**	
Color	**Número de regiones**	**Porcentaje**	**Número de regiones**	**Porcentaje**
Verde	4	22,2	1	16,7
Amarillo	4	22,2	0	0,0
Rojo	4	22,2	1	16,7
Azul	3	16,7	1	16,7
Celeste	2	11,1	1	16,7
Rosado	1	5,6	0	0,0
Naranja	0	0,0	1	16,7
Blanco	0	0,0	1	16,7
Total	**18**	**100**	**6**	**100***

* En la última columna la suma es mayor a 100 por criterios de redondeo

Observe que con este cuadro se pudo responder las preguntas planteadas en el problema. Este cuadro representa una forma de modelar el juego desde el punto de vista de las posibilidades que tiene cada color según la ruleta.

Problema 34. Lanzamiento de monedas y dado

Se lanzan dos monedas y un dado numerado de uno a seis, y se consideran los eventos

A: obtener solamente una corona y un número par.
B: obtener un múltiplo de tres.

Determine la probabilidad de que:

a) Ocurra el evento *A* o el evento *B*.
b) Ocurran los eventos *A* y *B* a la vez.
c) No ocurra el evento *A*.

Solución:

Si los resultados simples o puntos muestrales se representan como triadas o ternas de elementos tales como (E, C, 2), significa que la primera moneda cayó escudo, la segunda corona y el dado cayó en dos, tal como muestra la figura. En total el espacio muestral tiene 24 puntos muestrales que se resumen en el siguiente cuadro:

Resultado del dado	**Resultados de las monedas**			
	C,C	**C, E**	**E, C**	**E, E**
1	(C, C, 1)	(C, E, 1)	(E, C, 1)	(E, E, 1)
2	(C, C, 2)	(C, E, 2)	(E, C, 2)	(E, E, 2)
3	(C, C, 3)	(C, E, 3)	(E, C, 3)	(E, E, 3)
4	(C, C, 4)	(C, E, 4)	(E, C, 4)	(E, E, 4)
5	(C, C, 5)	(C, E, 5)	(E, C, 5)	(E, E, 5)
6	(C, C, 6)	(C, E, 6)	(E, C, 6)	(E, E, 6)

Para los eventos A y B,

A: obtener solamente una corona y un número par.
B: obtener un múltiplo de tres.

se tendría que:

$$A = \{(C, E, 2), (C, E, 4), (C, E, 6), (E, C, 2), (E, C, 4), (E, C, 6)\}$$

$$De\ este\ modo\ P(A) = \frac{6}{24} = \frac{1}{4}.$$

$$B = \{(\mathrm{C, C, 3}),(\mathrm{C, E, 3}),(\mathrm{E, C, 3}),(\mathrm{E, E, 3}),(\mathrm{C, C, 6}), (\mathrm{C, E, 6}),(\mathrm{E, C, 6}),(\mathrm{E, E, 6})\}$$

$$Con\ lo\ cual\ P(B) = \frac{8}{24} = \frac{1}{3}.$$

Las operaciones con eventos generan nuevos eventos.

a) La unión de los eventos A y B viene dado por:

$$A \cup B = \{(C, E, 2), (C, E, 4), (C, E, 6), (E, C, 2), (E, C, 4), (E, C, 6),(C, C, 3),(C, E, 3), (\mathrm{E, C, 3}),(\mathrm{E, E, 3}),(\mathrm{C, C, 6}),(\mathrm{E, E, 6})\}$$

Entonces la respuesta de a) es $P(A \cup B) = \frac{12}{24} = \frac{1}{2}$

b) La intersección de los eventos A y B es:

$$A \cap B = \{(E, C, 6),(C,E,6)\}$$

Con lo cual $\mathrm{P(A \cap B)} = \frac{2}{24} = \frac{1}{12}$

c) La no ocurrencia del evento A, o el complemento de A viene dado por:

A^c= {(C, C, 1), (C, E, 1), (E, C, 1), (E, E, 1), (C, C, 2), (E, E, 2), (C, C, 3). (C, E, 3), (E, C, 3), (E, E, 3), (C, C, 4), (E, E, 4), (C, C, 5), (C, E, 5), (E, C, 5), (E, E, 5), (C, C, 6), (E, E, 6)}

Por ello $\mathrm{P}(A^c) = \frac{18}{24} = \frac{3}{4}$

6. Propiedades básicas de las probabilidades

Según lo establecido en la definición clásica de probabilidad, se pueden deducir algunas propiedades básicas. Seguidamente se citan las más importantes:

Probabilidad del espacio muestral

Si *S* representa al espacio muestral de un experimento que tiene *n* puntos muestrales, se tiene que:

$$\mathrm{P(S)} = \frac{\text{Número total de elementos de S}}{\text{Número total de puntos muestrales}} = \frac{n}{n} = 1$$

Probabilidad del evento imposible

Normalmente se representa con ϕ al evento imposible, el cual no tiene puntos muestrales, entonces se tiene que:

$$\mathrm{P}(\phi) = \frac{0}{\text{Número total de puntos muestrales}} = 0$$

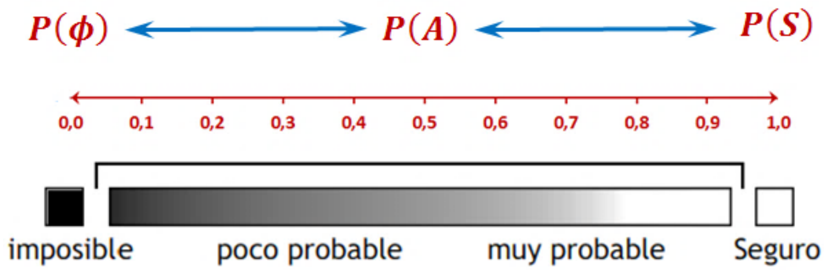
Probabilidad de un evento cualquiera

Se ha mencionado anteriormente que la probabilidad del evento imposible es cero y la probabilidad del evento seguro (que representa a todo el espacio muestral) es uno, entonces para cualquier otro evento *A* se cumple que:

$$0 \leq P(A) \leq 1$$

Lo anterior lo podemos observar en el siguiente esquema:

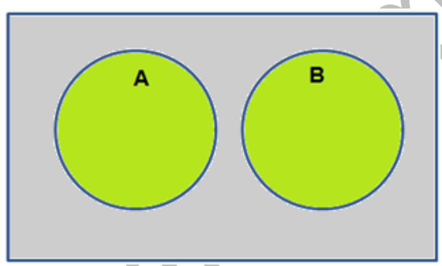
Volver al índice de conceptos



Probabilidad de la unión de eventos mutuamente excluyentes

Si tenemos dos eventos A y B en un espacio muestral S que son mutuamente excluyentes, es decir no hay puntos muestrales en común, se cumple que:

$$P(A \cup B) = P(A) + P(B)$$



Al no existir puntos muestrales en común entonces solamente se suman las probabilidades particulares.

Por ejemplo, si consideramos de nuevo del lanzamiento de un dado numerado de uno a seis y se consideran los eventos:

A: obtener un número par, se tiene que $\mathrm{A} = \{2,4,6\}$.
B: obtener un número impar, entonces $\mathrm{B} = \{1,3,5\}$.

$$\mathrm{P(A)} = \frac{3}{6} = \frac{1}{2} \text{ y } \mathrm{P(B)} = \frac{3}{6} = \frac{1}{2} \text{ entonces } \mathrm{P(A \cup B)} = \frac{1}{2} + \frac{1}{2} = 1 \; .$$

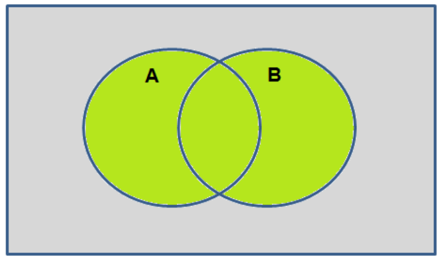
Observe que $\mathrm{A} \cup \mathrm{B}$ es el espacio muestral o sea el evento seguro.

Volver al índice de conceptos

Probabilidad de la unión de eventos cualesquiera

Si tenemos dos eventos A y B en un espacio muestral S, para los cuales existen puntos muestrales en común, entonces se cumple que:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$



Al existir puntos muestrales en común, cuando se suman las probabilidades de los eventos A y B, se suman dos veces los puntos muestrales en común, por esta razón a la suma de las probabilidades debe restarse la probabilidad de la intersección.

Consideremos nuevamente el lanzamiento de un dado numerado de uno a seis, y los eventos:

A: obtener un número par, se tiene que $\mathrm{A} = \{2,4,6\}$.
B: obtener un número primo, entonces $\mathrm{B} = \{2,3,5\}$.

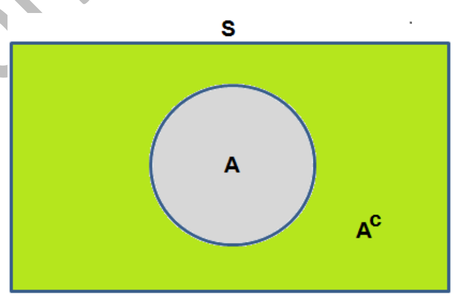
Debido a que $\mathrm{A} \cap \mathrm{B} = \{2\}$, además:

$$\mathrm{P(A)} = \frac{3}{6} = \frac{1}{2},\ \mathrm{P(B)} = \frac{3}{6} = \frac{1}{2} \text{ y } \mathrm{P(A \cap B)} = \frac{1}{6} \text{ entonces } \mathrm{P(A \cup B)} = \frac{1}{2} + \frac{1}{2} - \frac{1}{6} = \frac{5}{6}.$$

Probabilidad del complemento de un evento

Para un evento A cualquiera de un espacio muestral S, se cumple que:

$$P(A^c) = 1 - P(A)$$



El complemento de un evento A incluye los puntos muestrales de S que no están en A, por esta razón, como $\mathrm{P(S)} = 1$, entonces la probabilidad del complemento de A es $1 - \mathrm{P(A)}$.

Si volvemos al problema del lanzamiento de un dado numerado de uno a seis, considere el evento A que corresponde a obtener un múltiplo de tres, $\mathrm{A} = \{3, 6\}$, entonces como $P(A) = \frac{2}{6} = \frac{1}{3}$ entonces:

$$P(A^c) = 1 - \frac{1}{3} = \frac{2}{3}$$

Volver al índice de conceptos

Problema 35. Inventario de zapatos

En una zapatería se tiene un inventario de 1500 pares de zapatos, de los cuales 40% son negros y 30% son tenis (zapatilla deportiva), un 5% son tenis negras. Si se seleccionara aleatoriamente un par de zapatos, responda las siguientes preguntas:

1. ¿Cuál es la probabilidad de que sean negros o sean tenis?

2. ¿Cuál es la probabilidad de que no sean ni negros ni tenis?

Solución:

1. Para visualizar la información del problema y hacer más sencillo el análisis de las preguntas, los datos se pueden resumir en un cuadro o en un diagrama de Venn tal como se muestra. Los datos se pueden incluir en porcentajes o en probabilidades (en este caso los porcentajes se pueden convertir en probabilidades solamente dividiendo por 100). Lo importante es completar los valores faltantes por medio del complemento.

Tipo de Zapato	**Color del zapato**		**Total**
	Negro	Otro color	
Tenis	0,05	0,25	0,30
Otro tipo	0,35	0,35	0,70
Total	0,40	0,60	1,00

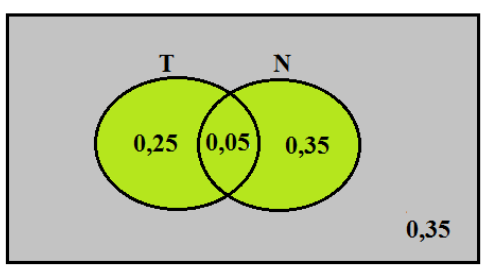
Observe que solamente los valores sombreados fueron dados en el problema, los restantes fueron obtenidos por diferencia. Si se representa con

T: al evento de que el par de zapatos seleccionado sean tenis, y
N: al evento de que el par de zapatos sean negros,

el evento de que los zapatos seleccionados sean tenis o sean negros viene dado por $T \cup N$, donde se sabe que $P(T) = 0{,}30$, $P(N) = 0{,}40$ y $P(T \cap N) = 0{,}05$. Por lo anterior, la probabilidad de $T \cup N$ se calcula por:

$$P(T \cup N) = P(T) + P(N) - P(T \cap N) = 0{,}30 + 0{,}40 - 0{,}05 = 0{,}65$$

El análisis anterior se puede representar mediante diagramas de venn



Observe que la probabilidad de $T \cup N$ corresponde a la suma de los valores del área sombreada con el color verde.

2. Si utiliza el análisis realizado en el ítem anterior, para determinar la probabilidad de que el par de zapatos seleccionado no sea negro ni tenis, se puede visualizar a partir del evento, que el par de zapatos sean tenis o sean de color negro. En el ítem anterior se demostró que:

$$P(T \cup N) = 0{,}65$$

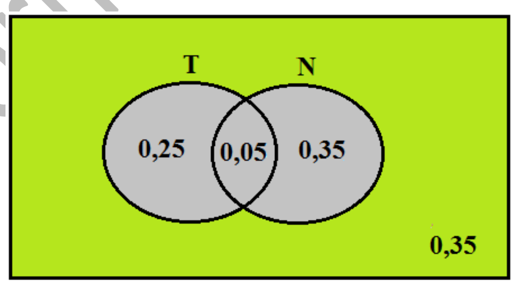
Entonces en la probabilidad de que el par de zapatos no sean ni tenis ni de color negro, se considera el complemento del evento $T \cup N$, con lo cual:

$$P(T \cup N)^C = 1 - P(T \cup N) = 1 - 0{,}65 = 0{,}35$$

Esto se visualiza fácilmente en el cuadro, la probabilidad corresponde al valor de la celda sombreada. Observe que dicha celda corresponde a otro tipo de zapato que no es tenis y otro color que no es negro.

Tipo de Zapato	**Color del zapato**		**Total**
	Negro	Otro color	
Tenis	0,05	0,25	0,30
Otro tipo	0,35	0,35	0,70
Total	0,40	0,60	1,00

Por medio de diagramas de Venn se puede visualizar también el análisis previo:



En este diagrama, el evento de que el par de zapatos seleccionado no sean tenis ni de color negro, se representa en la región verde.

Problema 36. Preferencia por carrera universitaria

A 260 estudiantes de undécimo año de cierto colegio se les consultó por su preferencia por las siguientes áreas para realizar estudios universitarios: Ingenierías, Ciencias Administrativas o Ciencias de la Salud; las respuestas se resumen en el siguiente cuadro:

Sexo	Ingenierías	Ciencias Administrativas	Ciencias de la Salud	Total
Masculino	30	35	35	100
Femenino	40	64	56	160
Total	**70**	**99**	**91**	**260**

Si se selecciona aleatoriamente a uno de estos estudiantes de undécimo año, responda las siguientes interrogantes:

1. ¿Cuál es la probabilidad de que el estudiante seleccionado sea del sexo masculino?

2. ¿Cuál es la probabilidad de que el estudiante seleccionado sea del sexo femenino y tenga preferencia por las Ingenierías?

3. ¿Cuál es la probabilidad de que el estudiante seleccionado sea del sexo masculino o tenga preferencias por las Ciencias de la Salud?

4. ¿Son igualmente probables los eventos tener preferencia por Ciencias Administrativas y por Ciencias de la Salud?.

5. Si el estudiante seleccionado es del sexo femenino ¿cuál es la probabilidad de que prefiera el área de Ingenierías?

6. ¿Es igualmente probable la preferencia por Ciencias de la Salud entre hombres y mujeres?

Solución:

1. Para responder esta pregunta se aplica la definición clásica de probabilidad. En total, de los 260 estudiantes de undécimo año, se tienen 160 estudiantes del sexo femenino, por lo que la probabilidad que el estudiante seleccionado sea de este sexo es:

$$\frac{160}{260} \approx 0{,}615$$

Nota: el símbolo $\approx$ significa aproximado, en el caso anterior el cociente $\frac{160}{260}$ tiene por resultado aproximado el 0,615, esto es diferente cuando se utiliza = porque en estos casos el valor es exacto.

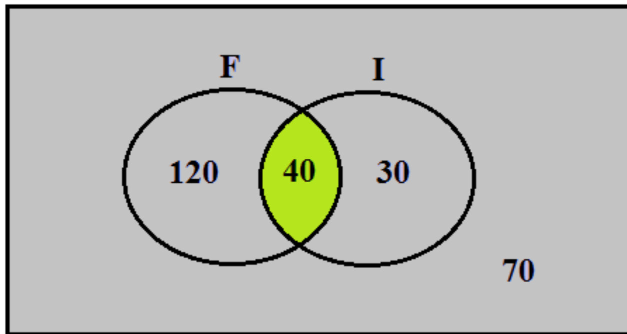
2. Si consideramos los eventos

F: el estudiante seleccionado es del sexo femenino,
I: el estudiante seleccionado tiene preferencia por Ingenierías,

entonces la probabilidad de que el estudiante seleccionado sea del sexo femenino y tenga preferencia por las Ingenierías, debe verse como $P(F \cap I)$. En total hay 40 estudiantes del sexo femenino que tienen preferencia por las ingenierías, entonces:

$$P(F \cap I) = \frac{40}{260} \approx 0{,}154$$

Este análisis se puede visualizar también mediante el uso de los diagramas de Venn. Puede notarse que en la intersección de F con I se encuentran 40 estudiantes de un total de 260 estudiantes de undécimo año.



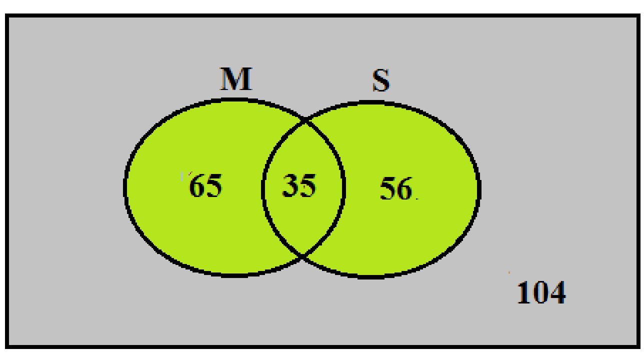
3. Para responder la tercera pregunta, consideremos los eventos

   M: el estudiante seleccionado es del sexo masculino
   S: el estudiante seleccionado tiene preferencia por Ciencias de la Salud

   Entonces la probabilidad que el estudiante seleccionado sea del sexo masculino o tenga preferencias por las Ciencias de la Salud sería $P(M \cup S)$ y viene dada por:

$$P(M \cup S) = P(M) + P(S) - P(M \cap S) = \frac{100}{260} + \frac{91}{260} - \frac{35}{260} = \frac{156}{260} = 0{,}6$$

   Mediante diagramas de Venn, puede notarse que en la unión de los conjuntos hay $55 + 35 + 56$ que suman 156 estudiantes, de un total de 260 estudiantes de undécimo año



4. Para determinar si ¿son igualmente probables los eventos tener preferencia por Ciencias Administrativas y por Ciencias de la Salud?, se requiere determinar las probabilidades individuales de cada evento y compararlas. Consideremos los eventos:

A: el estudiante seleccionado tiene preferencia por Ciencias Administrativas
S: el estudiante seleccionado tiene preferencia por Ciencias de la Salud

$$P(A) = \frac{99}{260} \approx 0{,}381$$

$$P(S) = \frac{91}{260} = 0{,}35$$

Entonces las probabilidades no son iguales: es más probable que un estudiante prefiera Ciencias Administrativas a que prefiera Ciencias de la Salud.

5. Si sabemos de antemano que el estudiante seleccionado es del sexo femenino, entonces el espacio muestral se debe cambiar para considerar únicamente las mujeres. Entre el total de mujeres se selecciona una de ellas y se requiere responder ¿cuál es la probabilidad de que ella prefiera el área de Ingenierías? Como en total hay 160 mujeres, de las cuales 40 tienen preferencia por ingenierías, la probabilidad solicitada es:

$$\frac{40}{160} = 0{,}25$$

6. Para determinar si ¿es igualmente probable la preferencia por Ciencias de la Salud entre hombres y mujeres?, al igual que en el ítem anterior, se debe realizar un análisis relativo tomando en cuenta solamente hombre o mujeres por separado. Observe la información:

Preferencia	**Hombres**	**Mujeres**
Ciencias de la Salud	35	56
Otra área	65	104
Total	100	160

Entre los hombres, la probabilidad de que un estudiante prefiera Ciencias de la Salud es:

$$\frac{35}{100} = 0{,}35$$

Entre las mujeres, la probabilidad que una estudiante prefiera Ciencias de la Salud es:

$$\frac{56}{160} = 0{,}35$$

Por lo tanto se tiene que efectivamente son igualmente probables los eventos.

Nota: Observe que las probabilidades anteriores no se calcularon sobre el total de 260 estudiantes, esto se pidió determinar la probabilidad de que el área preferida sean Ciencias de la Salud entre hombres y entre mujeres por separado, por ello se utilizó como denominador en el primer caso el número de hombres y en el segundo el número de mujeres.

7. Enfoque frecuencista o empírico de probabilidad

En los problema anteriores, para determinar la probabilidad de un evento hemos supuesto que conocemos el espacio muestral. Sin embargo, en la vida real muchas veces no es posible conocer este conjunto. En estos caso se puede obtener una aproximación de la probabilidad por medio de una muestra aleatoria de unidades estadísticas involucradas en el problema de interés. Hay que recalcar que si la probabilidad se calcula sobre una muestra y no sobre todo el espacio muestral, lo que se obtiene es una estimación o aproximación que puede variar si se utiliza otra muestra diferente. A este tipo de probabilidad se le llama probabilidad *empírica o frecuencista* y se define de la siguiente manera:

Volver al índice de conceptos

Concepto empírico o frecuencista de probabilidad

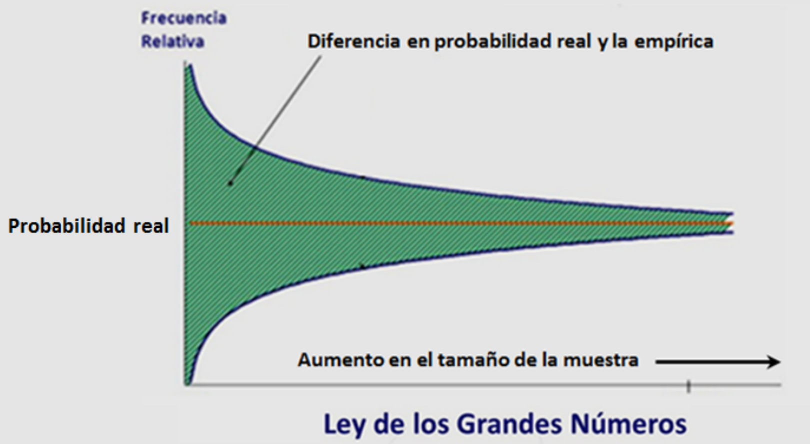
En una muestra aleatoria que incluye n elementos igualmente probables, de los cuales existe una frecuencia de k elementos a favor de evento A, se dice que la probabilidad de que el evento A ocurra (se representa con $P(A)$) y viene dada por la razón:

$$P(A) = \frac{frecuencia\ de\ resultados\ a\ favor\ de\ A}{tamaño\ de\ la\ muestra} = \frac{k}{n}$$

Nota: Observe que la definición anterior es similar a la definición clásica, con la salvedad anotada anteriormente de que el resultado empírico es una estimación de la probabilidad real y que puede variar de una muestra a otra.

A medida que el tamaño de la muestra se hace cada vez más grande la estimación de la probabilidad real se hace cada vez más precisa, esta propiedad se conoce como la ley de los grandes números.

Por medio de ella se puede encontrar la probabilidad de cualquier evento siempre que se conozca el número total de resultados del experimento y el número de resultado a favor del evento.



Puede notarse en el gráfico que conforme el tamaño de la muestra aumenta la probabilidad empírica se aproxima cada vez más a la probabilidad real del evento.

Problema 37. Condición de lateralidad según partes del cuerpo

En los últimos años en el fútbol nacional se ha observado una escasez de jugadores que pueda patear de igual manera con ambas piernas (*ambidiestros*). Algunas personas creen que esto ocurre porque la probabilidad de encontrar una persona ambidiestra ha venido disminuyendo. Según estudios publicados por la página web http://www.zurdos.cl/estadisticas.html para 1981 se tenía en el mundo que:

Tabla de probabilidades según la parte del cuerpo y su lateralidad

Parte del cuerpo	Muy Zurdo	Muy Diestro	Ambos por igual
Mano	0,05	0,73	0,22
Pie	0,04	0,46	0,50
Ojo	0,05	0,54	0,41
Oído	0,15	0,35	0,50

Porac C. & Coren S. Lateral preferences and human behavior. New York: Springer-Verlag, 1981
(Este cuadro fue modificado en cuanto a la notación con fines didácticos)

Para determinar si la creencia de que la probabilidad de encontrar personas ambidiestras ha disminuido respecto a los resultados que presentó esa investigación, se seleccionó una muestra aleatoria de 500 estudiantes de secundaria de Costa Rica y se realizó un estudio para determinar su condición de lateralidad según las partes del cuerpo:

Resultados de lateralidad de una muestra aleatoria de 500 estudiantes de secundaria en Costa Rica, 2016

Parte del cuerpo	Muy Zurdo	Muy Diestro	Ambos por Igual	Total
Mano	28	390	82	500

Pie	22	254	224	500
Ojo	22	280	198	500
Oído	71	170	259	500

Datos simulados con fines didácticos

Utilice esta información para responder las siguientes preguntas:

1. ¿Hay diferencias notorias en las probabilidades de que una persona sea ambidiestra en relación con las que se observaron en el estudio mundial de 1981?

2. ¿Qué se puede indicar respecto a la creencia de que ha disminuido la probabilidad de encontrar una persona ambidiestra de los pies?

3. ¿Qué supuestos o consideraciones deben tenerse para que las respuestas anteriores sean válidas?

Solución:

Utilizando los datos de la muestra se procede a responder las preguntas

1. En el siguiente cuadro se procede a determinar las probabilidades de que un estudiante de la muestra sea ambidiestro para las cuatro partes del cuerpo consideradas en el estudio:

Resultados de la probabilidad de que un estudiantes sea ambidiestro según la parte del cuerpo en una muestra aleatoria de 500 estudiantes de secundaria en Costa Rica, 2016

Parte del cuerpo	Número de estudiantes ambidiestros	Probabilidad de ambidiestro
Mano	82	$\frac{82}{500} = 0{,}164$
Pie	224	$\frac{224}{500} = 0{,}448$
Ojo	198	$\frac{198}{500} = 0{,}396$
Oído	259	$\frac{259}{500} = 0{,}518$

Al comparar estos resultados con los del estudio de 1981, se puede ver que:

Parte del cuerpo	Ambos por igual
Manos	0,22
Pie	0,50
Ojo	0,41
Oído	0,50

Se observan diferencias notorias en las probabilidades de ser ambidiestro en la mano y el pie, mientras que en el ojo o en el oído las diferencias no son muy grandes.

2. Tomando como base el resultado obtenido en el ítem 1, se tiene que para 1981 la probabilidad de que una persona fuera ambidiestra en los pies era del 0,50; mientras que la muestra aleatoria se obtuvo que la probabilidad que uno de los 500 estudiantes sea ambidiestro es 0,448; por lo que se podría considerar que ha existido un descenso.

3. El análisis realizado en este ejercicio parte de muchos supuestos, entre ellos:

   i. La información publicada en el estudio de 1981 era también válida para Costa Rica en ese mismo año.

   ii. Los resultados de la muestra aleatoria de estudiantes es representativa de lo que ocurre en la población joven de Costa Rica.

   Aunque se pueden señalar muchos otros supuestos, los anteriores son claves para fundamentar el análisis realizado.

Nota: Las propiedades básicas de probabilidad también son válidas en el caso de que estemos trabajando con probabilidades empíricas.

Problema 38. Tratamiento contra la influenza

Las personas enfermas de influenza constituyen la principal fuente de infección de la enfermedad, el virus responsable se trasmite principalmente por vía aérea. Los niños pequeños (entre seis meses y cinco años), adultos mayores (mayores de 65 años), enfermos crónicos, embarazadas e inmunosuprimidos pertenecen a los grupos de riesgo. Esto significa que al estar expuestos al virus, tienen mayor probabilidad de contagiarse de influenza que el resto de la población.

https://pixabay.com/es

http://www.apmadrid.es/

Suponga que una nueva cepa de virus afecta una pequeña comunidad (denominada comunidad A) que se encuentra aislada de otras poblaciones, este virus muestra un comportamiento diferente, por lo que se tomó una muestra aleatoria de 400 personas del grupo en riesgo que estuvieron expuestas al virus, de las cuales a 200 se les aplicó una vacuna y a los otros 200 se les aplicó un placebo (medicamento neutro sin acción farmacológica). Los resultados fueron los siguientes:

Condición de enfermedad para una muestra aleatoria de 400 personas en riesgo[1] de la comunidad A que fueron expuestas al virus de la influenza según tipo de tratamiento

Condición	Se aplicó vacuna	Se aplicó placebo	Total
Contrajeron influenza	112	171	283
No contrajeron influenza	88	29	117
Total	200	200	400

[1] Se considera en riesgo una persona que pertenece a los siguientes grupos: niños pequeños (entre seis meses y cinco años), adultos mayores (mayores de 65 años), enfermos crónicos, embarazadas e inmunosuprimidos.

Fuente: Esta información es ficticia, fue elaborada con fines didácticos.

Tomando como referencia la información anterior, se desea responder las siguientes interrogantes:

1. Si escoge una persona de esta muestra en forma aleatoria determine la probabilidad de que:

   a) no haya contraído la influenza.
   b) se le haya aplicado la vacuna.
   c) haya sido vacunada y contrajera la influenza.
   d) no contrajera la influenza o se la haya aplicado el placebo.

2. Para una persona de la comunidad A que pertenece al menos a uno de los grupos de riesgo:

   a) estime la probabilidad de que pueda enfermarse, sabiendo que se ha vacunado .
   b) estime la probabilidad de que no se haya vacunado, sabiendo que se ha enfermado

3. Según el estudio realizado, ¿hay evidencia de que la vacuna sea eficiente para disminuir la probabilidad de que una persona de la comunidad A se enferme de influenza?

Solución:

1. Para responder este punto se debe observar que las probabilidades solicitadas se calculan para una persona aleatoria que pertenece a la muestra. Por ello se aplica la definición clásica, no se realizan estimaciones sino que es un cálculo directo. El espacio muestral está constituido por las 400 personas. Sean:

   A: la persona fue vacunada B: la persona no es vacunada
   C: la persona contrajo influenza D: la persona no contrajo influenza

   a) En el inciso a) se pide la probabilidad el evento D, o sea $P(D)$. Debido a que de las 400 personas, 117 no contrajeron la enfermedad, $P(D) = \frac{117}{400} \approx 0{,}2925$. Entonces la probabilidad que la persona seleccionada no haya contraído influenza (redondeado a dos decimales) es 0,29.

   b) Acá se pide la probabilidad del evento A. Como la vacuna se aplicó a 200 personas, entones $P(A) = \frac{200}{400} \approx 0{,}5$. Entonces la probabilidad que la persona seleccionada no se le haya aplicado la vacuna es 0,5.

   c) En este inciso se solicita la probabilidad del evento $A \cap C$, en donde 112 personas fueron vacunadas y contrajeron la enfermedad. Entonces $P(A \cap C) = \frac{112}{400} = 0{,}28$. Con lo cual la probabilidad de que la persona seleccionada fuera vacunada y contrajera la enfermedad de 0,28.

   d) Por último, se solicita la probabilidad de $D \cup B$. Entonces, según las propiedades de probabilidades $P(D \cup B) = P(D) + P(B) - P(D \cap B) = \frac{117}{400} + \frac{200}{400} - \frac{29}{400} = \frac{288}{400} = 0{,}72$. Con lo cual, la probabilidad de que la persona seleccionada no contrajera la influenza o se la haya aplicado el placebo es 0,72.

2. A diferencia del punto 1., en el punto 2. se pide estimar probabilidades para una persona que pertenece a alguno de los grupos de riesgo en toda comunidad (ya no se refiere a la muestra). Por ello se aplica el enfoque empírico pues se utiliza la información de la muestra para estimar las probabilidades en toda la comunidad.

   a) Para responder el inciso a) hay que tener presente que la persona ya ha sido vacunada (son 200 en la muestra los que fueron vacunados) y se debe estimar la probabilidad de que la persona pueda enfermarse (de las 200 que fueron vacunadas se enfermaron 112), entonces la probabilidad estimada viene dada por $\frac{112}{200} = 0{,}56$.

   b) Aplicando el mismo principio del inciso a), si se supone que la persona se ha enfermado (283 personas de la muestra se enfermaron) y se pide estimar la probabilidad de que no se haya vacunado (de las 283 que se enfermaron 171 no recibieron vacuna), entonces la estimación de la probabilidad es $\frac{171}{283} = 0{,}60424\ldots$ La probabilidad estimada (redondeada a dos decimales) es 0,60.

3. Para saber si la vacuna disminuye la probabilidad de que una persona en riesgo se enferme, hay que estimar la probabilidad de que la persona se enferme entre los que se vacunaron y entre los que no se vacunaron.

   Entre las personas que se vacunaron la probabilidad de enfermarse fue de $\frac{112}{200} = 0{,}56$. Entre las personas que no se vacunaron la probabilidad de enfermarse fue de $\frac{171}{200} = 0{,}855$. Puede notarse que efectivamente la probabilidad de enfermarse es mucho menor entre las personas que se vacunaron que entre aquellos que no lo hicieron. Esto puede hacer suponer que la vacuna es efectiva.

IV. Bibliografía

Ministerio de Educación Pública (2015). Curso bimodal para el II Ciclo: Estadística mucho más que procedimientos y técnicas. Unidad didáctica Estadística. San José, Costa Rica: autor.

Ministerio de Educación Pública (2012). Programas de Matemáticas para la Educación General Básica y el Ciclo Diversificado. San José, Costa Rica: autor.

V. Créditos

Este documento ha sido elaborado por el *Proyecto Reforma de la Educación Matemática en Costa Rica,* del Ministerio de Educación Pública.

Autor
Edwin Chaves Esquivel

Revisores
Ángel Ruiz, Edison De Faria, Johanna Mena, Keibel Ramírez, Luis Hernández, Xinia Zúñiga.

Edición final de este documento
Edwin Chaves Esquivel, Luis Hernández Solís

Director del Proyecto Reforma de la Educación Matemática en Costa Rica
Ángel Ruiz

Para referenciar este documento

Ministerio de Educación Pública, Proyecto Reforma de la Educación Matemática en Costa Rica (2020). *Estadística y Probabilidad, Material de consulta*, San José, Costa Rica: autor.