Fundación POLAR

Estadística

Distribuciones bidimensionales

En Birongo, estado Miranda, un grupo de artesanos capacitados y dotados técnicamente por Fundación Polar, institución que los acompaña en la gestión de sus procesos desde el año 2000, se dedica a la fabricación de barras y bombones de chocolate, elaborados en diferentes presentaciones y a partir de la producción de cacao que tiene lugar en más de 300 plantaciones que se ubican en los alrededores de esta población. A mediano plazo, la consolidación de los procesos productivos, control estadístico y administrativo de cada uno de ellos y la introducción de mejoras tecnológicas, permitirán aumentar la producción diaria de chocolate a más de cuatro veces la actual (70 kg/día).

La asociación funciona en un centro de producción –diseñado y construido por Fundación Polar– adaptado a las exigencias propias de la elaboración de chocolate. Las potencialidades de atracción turística de esta planta han servido de base para la formulación de un plan de desarrollo turístico para Birongo y la región, gestionado por la comunidad organizada.

www.fpolar.org.ve

Fotografía: Alejandro Reyes.

Últimas **Noticias**

EMPRESAS POLAR

Estos fascículos están disponibles en línea, visitando la página web: **http://www.fpolar.org.ve/matematica2**

Distribuciones bidimensionales

Si se mide una variable X en un objeto en dos momentos distintos (antes y después de algún tratamiento a dicho objeto), o si se mide X en dos objetos distintos (padre e hijo), se obtienen pares de valores con los que se pueden formar distribuciones de **frecuencia bidimensionales** de dicha variable. También se pueden formar **distribuciones bidimensionales** cuando se mide X en un primer objeto e Y en un segundo objeto que se supone relacionado de alguna manera con el primero.

Por ejemplo, en 10 ciudades del mundo que presentan problemas de transporte urbano, se recogieron datos acerca de la modalidad de desplazamiento de los trabajadores entre su hogar y el trabajo.

Sea X la modalidad de transporte utilizado e Y el número de lugares de trabajo de las ciudades en estudio tenemos el siguiente cuadro:

Nº de lugares de trabajo/ciudad	Modalidad de desplazamiento Auto propio	 Autobús público	 Bicicleta	Total
16-65	77	18	5	100
65-120	42	37	21	100
120-175	16	59	25	100
Total	135	114	51	300

Al observar el cuadro podemos extraer información acerca de las modalidades de transporte que se utilizan en estas ciudades, por ejemplo en las ciudades que tienen entre 120 a 175 lugares de trabajo la modalidad de desplazamiento utilizada es el autobús público.

En el cuadro siguiente se muestra otro ejemplo de distribución bidimensional.

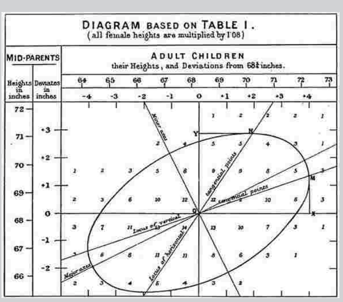
Distribución de 1 000 adultos mayores de 20 años según el nivel de educación y nivel de ingreso

Ingreso Y (Bs)	X nivel de educacion Primaria	 Secundaria	 Post secundaria	Total por filas
$Y < 3 \cdot 10^5$	$n_{11} = 420$	$n_{12} = 210$	$n_{13} = 70$	$n_{1\bullet} = 700$
$3 \cdot 10^5 \leq Y < 10^6$	$n_{21} = 120$	$n_{22} = 60$	$n_{23} = 20$	$n_{2\bullet} = 200$
$Y \geq 10^6$	$n_{31} = 60$	$n_{32} = 30$	$n_{33} = 10$	$n_{3\bullet} = 100$
Total por columnas	$n_{\bullet 1} = 600$	$n_{\bullet 2} = 300$	$n_{\bullet 3} = 100$	$n = 1\,000$

En este cuadro, la **fila** y la **columna de totales** corresponden a las distribuciones de frecuencias unidimensionales de las variables X (Nivel de educación) e Y (Nivel de ingreso). Estos totales, por estar en los márgenes del cuadro, se denominan **distribuciones marginales**. Las otras filas y columnas son también distribuciones unidimensionales de X e Y, pero calificadas de **condicionales.**

Interesante:

Si tomas cualquier valor del segundo cuadro de distribución bidimensional, observarás que al multiplicar el total de la fila a la que pertenece dicho valor con el total de la columna donde está ubicado y lo divides entre el numero total de datos representados en el cuadro, el resultado coincide con el elemento seleccionado. Por ejemplo tomemos el elemento n_{32} =30, el total de la fila tres $n_{3\bullet} = 100$; el total de la columna dos $n_{\bullet 2} = 300$; el producto de estas dos cantidades dividido entre el número total de elementos n=1000 es igual a 30 que fue el valor seleccionado. Esto queda resumido en la siguiente expresión $n_{ij} = \frac{n_{i\bullet} \cdot n_{\bullet j}}{n}$ **.Esto no es usual que ocurra, cuando ocurre se dice que las variables X e Y son independientes.**



¿SABÍAS QUE...?

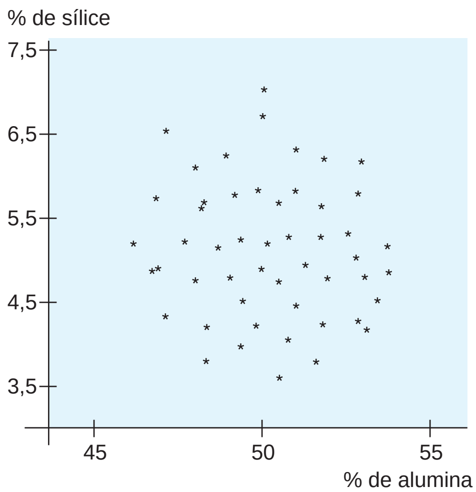
Karl Pearson (inglés 1857-1936), cuando trabajó asociado a Francis Galton (inglés 1822-1911), desarrolló el denominado coeficiente de correlación lineal. Pearson hizo otros aportes a la estadística que al igual que el coeficiente mencionado son de interés permanente.

Correlación

Palacio de las Academias anteriormente Universidad Central de Venezuela (1920). Caracas, Venezuela.

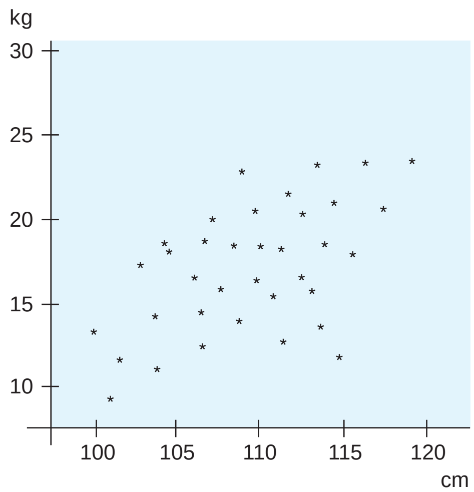
Cuando se mide una variable Y o dos variables X e Y y se obtienen -como se explicó al tratar acerca de las distribuciones bidimensionales- pares de valores de X o de X e Y; éstos pueden ser representados como puntos en gráficos cartesianos, llamados **gráficos de dispersión** o "**nubes de puntos**". La forma en que aparecen estos puntos sugieren la existencia de relación entre las variables lo que se denomina **correlación**. Los gráficos siguientes muestran situaciones interesantes que pueden presentarse.

Gráfico 1. % de alúmina y % de sílice en la composición de la bauxita



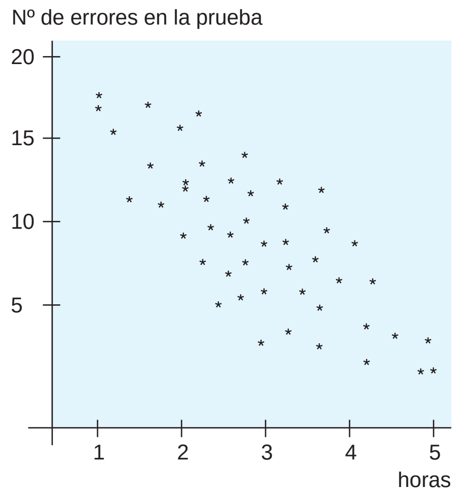
El gráfico 1 muestra evidencia de independencia entre el % de alúmina y de sílice que contiene la bauxita ya que al tomar cualquiera valor de uno de sus ejes tiene varios valores en el otro.

Gráfico 2. Talla y peso de niños de 6 años de edad



El gráfico 2 sugiere que existe una relación de dependencia entre la talla y el peso. A medida que aumenta la talla aumenta el peso.

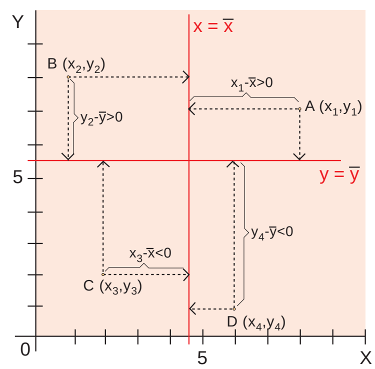
Gráfico 3. Tiempo dedicado a prepararse para una prueba y número de errores cometidos al realizarla



El gráfico 3 sugiere que el número de errores disminuye con el tiempo dedicado al estudio por lo que existe una relación de dependencia.

¿Cómo analizar una nube de puntos?

En el gráfico de dispersión se trazan las rectas $x = \bar{x}$ e $y = \bar{y}$, siendo $\bar{x}$ e $\bar{y}$ las medias de los valores de las variables X e Y. Se han seleccionado cuatro puntos de la nube (A, B, C y D) indicando las diferencias $(x_i-\bar{x})$ e $(y_i-\bar{y})$.



Tomando punto por punto tenemos:

$(x_1-\bar{x})\ (y_1-\bar{y}) > 0 \qquad (x_2-\bar{x})\ (y_2-\bar{y}) < 0$

$(x_3-\bar{x})\ (y_3-\bar{y}) > 0 \qquad (x_4-\bar{x})\ (y_4-\bar{y}) < 0$

La $\sum_{i=1}^{n} (x_i-\bar{x})\ (y_i-\bar{y})$ indica como es la correlación entre las variables X e Y.

La correlación entre las variables de los gráficos anteriores se expresa en:

	Gráfico 1	Gráfico 2	Gráfico 3
$\sum_{i=1}^{n} (x_i-\bar{x})\ (y_i-\bar{y})$	Igual o próxima a 0	Notablemente mayor que 0	Notablemente menor que 0
Indicio de correlación	Inexistente	Positiva	Negativa

Regresión

Bob Abreu y Omar Daal.

Se ha visto antes que en las nubes de puntos se pueden notar indicios de la existencia de relación entre dos variables X e Y. Podemos avanzar más en el conocimiento de la relación entre dos variables guiándonos por deducciones derivadas de algunas observaciones en otras experiencias, e intentar hacer predicciones de los valores de una de ellas en función de los valores de la otra.

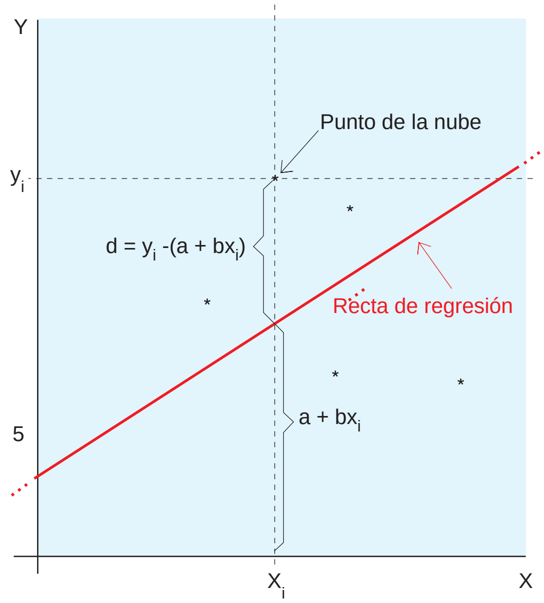
En ocasiones se puede suponer que es posible predecir los valores de Y basándonos en el conocimiento de valores de X, mediante el uso de una función y = a + bx, como la representada en el gráfico.

Los parámetros a y b se pueden determinar a partir de los datos disponibles, cumpliendo el requisito de que $\sum[y_i - (a+bx_i)]^2$ sea menor que la correspondiente a cualquier otra recta que se pudiera elegir con el mismo propósito.

La recta así determinada se dice que es de **regresión mínimo cuadrática**. Para obtener los parametros se aplican estas ecuaciones

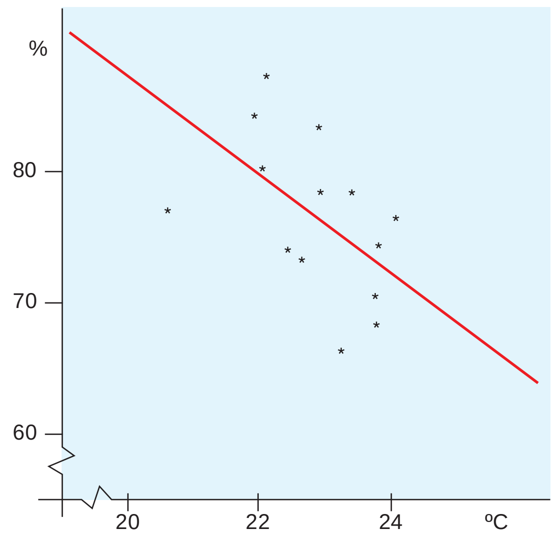
$$b = \frac{\sum_i x_i y_i - n\bar{x}\bar{y}}{\sum_i x_i^2 y_i - n\bar{x}^2} \qquad a=\bar{y}-b\bar{x}$$

Las diferencias y_i -(a + bx_i) son útiles en la búsqueda de argumentos acerca de la relación entre las variables X e Y. Por ejemplo algunas de las diferencias pueden indicar la presencia de puntos atípicos que influyen de forma extrema en la determinación de la ecuación.



Veamos el siguiente ejemplo: Si se planteara la suposición de que, conocida la temperatura atmosférica (X) se puede en función de ella determinar la humedad atmósferica relativa (Y), se obtuvo la línea de regresión mínimo cuadrática a partir de los datos que se expresan en la tabla siguiente:

Día	(X) Temperatura media (ºC)	(Y) Humedad relativa media (%)	(X^2)	(XY)
1	23,8	68	566,44	1 618,4
2	23,9	74	571,21	1 768,6
3	20,7	77	428,49	1 593,9
4	22,2	87	492,84	1 931,4
5	22,1	80	488,41	1 768,0
6	23,0	78	529,00	1 724,0
7	22,7	73	515,29	1 657,1
8	23,0	76	529,00	1 748,0
9	22,5	74	506,25	1 665,0
10	22,0	84	484,00	1 848,0
11	23,0	83	529,00	1 840,0
12	24,1	76	580,81	1 831,6
13	23,3	66	542,89	1 537,8
14	23,8	70	566,44	1 666,0
15	23,5	78	552,25	1 833,0
	$\sum x$=343,6	$\sum y$=1 144	7 882,32	26 100,9
	$\bar{x}$=22,9	$\bar{y}$=76,3		



$$b = \frac{\sum_i x_i y_i - n\bar{x}\bar{y}}{\sum_i x_i^2 y_i - n\bar{x}^2} = \frac{26\,100{,}8 - 26\,209{,}5}{7\,882{,}32 - 7\,866{,}15} = 0{,}67$$

$a=\bar{y}-b\bar{x}$ = 76,3 + 15,3 = 91,6

Recta de regresión y = 91,6 - 0,67x

Reto:

Un periodista deportivo afirma que el número de sencillos (S) es mejor predictor del promedio de bateo (AVG) que el número de extrabases (D + T + J). Utilizamos los datos de los cuadros correspondientes a jugadores venezolanos en la Liga Nacional de EE.UU. (mediados de mayo 2004):

S	D+T+J	AVG
47	8	0,290
46	20	0,284
45	22	0,302
43	18	0,276
37	8	0,264
30	6	0,256
29	13	0,204
28	9	0,243
26	7	0,236
26	5	0,276

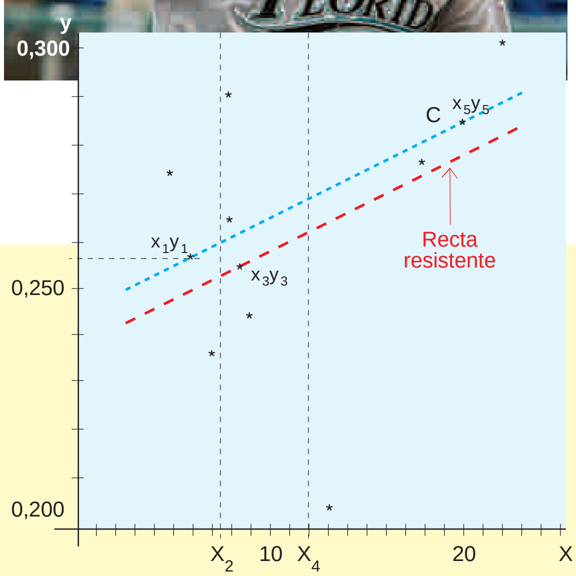

César Izturis (AVG = 0,290)


Miguel Cabrera (AVG = 0,284)

- Obtenga los parámetros a y b de las rectas de regresión mínimo cuadrático del AVG respecto de los sencillos (S) y también respecto de los extrabases (D+T+J).
- Represente gráficamente esas rectas y calcule $\sum_{i=1}^{n} [y_i - (a + bx_i)]^2$ para cada una de ellas usando una calculadora.
- ¿Le parece acertado lo que afirma el periodista? ¿Por qué?

Interesante:

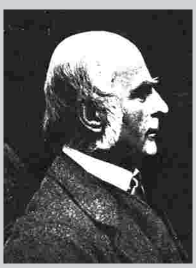
Para considerar la posible relación en dos variables X e Y expresable por una función y= a + bx, puede resultar conveniente valerse de una línea recta, denominada **recta resistente**, cuya determinación se basa en el cálculo de las medianas, tal es el caso de la recta que se muestra en el gráfico siguiente, en el cual X es el número de extrabases e Y es el promedio de bateo de los 10 jugadores de la Liga Nacional que aparecen en el cuadro de arriba. El calificativo “resistente” alude a la poca variación de los parámetros a y b que pueden causar los puntos muy alejados.



¿SABÍAS QUE...?

La idea de la **recta resistente** es junto con la de los “tallos y hojas”, la de los “gráficos de cajas” y otras, una de las integrantes de la proposición del análisis de datos exploratorio que, en los años sesenta del siglo XX, formularon Tukey y Mosteller. Tal proposición está siendo acogida, dándole flexibilidad e intensificando y extendiendo el uso del análisis estadístico de datos.


Frederick Mosteller
Estadístico norteamericano
(1916-)



Sir Francis Galton, estimulado por su primo Charles Darwin, fundó la Eugenesia, que consiste en la reproducción selectiva de los seres humanos para mejorar la especie. Ideó algunos métodos importantes usados por la psicología de la adaptación. Fue el primer autor en aplicar, de forma sistemática, la estadística a la resolución de problemas estudiados en psicología. Estudió hermanos gemelos para separar la contribución de la naturaleza y la crianza del carácter, la inteligencia y la conducta humana. Inventó la técnica de la asociación libre para indagar en la memoria.

Orientaciones metodológicas
Sugerencias para los docentes

El docente debe presentar la estadística como un medio para lograr conocimiento y no sólo como un medio para resumir gráfica o numéricamente conjuntos de datos.

En la búsqueda de conocimiento en áreas sustantivas, la estadística se asocia con el método científico y debe, por tanto, participar en cada una de las siguientes fases:

1. desarrollo de la pregunta de investigación
2. diseño de la investigación
3. recolección de datos a través de observación y medición
4. disposición de datos en registros adecuados para su procesamiento
5. análisis estadístico de datos y
6. discusión de resultados.

Las experiencias didácticas basadas en esta propuesta favorecen el logro de una visión más completa de la estadística que la que se puede lograr con el estudio detallado de sus partes constituyentes.

El aula de clases ofrece muchas altemativas para realizar una enseñanza viva de la estadística. A continuación presentamos una de estas altemativas, en la que los estudiantes pueden participar en el diseño estadístico de la investigación, el resumen de los datos y en la interpretación de los resultados obtenidos.

Por ejemplo, en un aula de clase se desea conocer **si existe una relación lineal entre la altura de las personas y su peso**.

1. Desarrollo de la pregunta de investigación

Para ello se puede proponer como universo de estudio a todos los estudiantes de la clase. Una vez identificado el universo, habrá que precisar la población estadística sobre la que responderemos la pregunta. A tal efecto, el docente iniciará una discusión sobre si basta con tomar en cuenta únicamente la altura y el peso de los alunmos o ¿habrá que tomar en cuenta el sexo? o si ¿jugará algún papel en esta relación la altura y el peso de los progenitores? Una vez ratificada la pregunta o modificada para ampliar el objetivo de la investigación, se seleccionarán las variables cuyo registro permitirá responderla.

2. Diseño de la investigación

El docente aclarará que lograr acceso a toda la población es poco común y, en general, será necesario recurrir a la seleccion de una muestra para la investigación del problema de interés. En este momento el aula de clase se puede convertir en un laboratorio de estadística, aprovechando la ventajosa situación de conocer la población para investigar sobre los procedimientos de muestreo y, sobre todo, para apreciar el desempeño de las muestras aleatorias en la estimación de características poblacionales.

Para ello se pueden formar varios equipos de muestreo, de dos o tres alumnos, que seleccionen muestras aleatorias simples de diversos tamaños (por ejemplo muestras de tamaño 5 y muestras de tamaño 10). Con este fin será conveniente que se prepare colectivamente el mecanismo de selección aleatoria, por ejemplo, una caja de cartón que contenga una papeleta con el número de lista de cada estudiante. El muestreo se podrá realizar sin reemplazo o con reemplazo. En este último caso la papeleta elegida se devuelve a la caja, pero no así en el primer caso. Las cinco o diez papeletas seleccionadas de la caja indican los alumnos seleccionados en la muestra y, por tanto, aquellos para los que se registrarán el peso, la altura y el sexo.

3. Recolección de datos a través de la observación y medición

La primera parte del ejercicio consistirá en medir la altura, el peso y registrar el sexo de cada estudiante. Para ello se procurará un metro y una balanza y se entrenará a los estudiantes en su uso. Esto constituye la población de interés de la investigación.

4. Disposición de datos en registros adecuados para su procesamiento

Para ilustrar con cierto detalle algunos de los procedimientos involucrados en una experiencia como la que proponemos consideraremos una clase hipotética conformada por cuatro alumnos. El resultado de este primer ejercicio se presentará en un cuadro como el de la siguiente página.

Cuadro 1.
Población en estudio

Sujeto	Peso (kg)	Altura (cm)	Sexo
1	46	154	Masculino
2	34	150	Femenino
3	42	155	Femenino
4	50	160	Masculino

5. Análisis estadístico de datos

Después de llevar a cabo el proceso de medición y registro correspondiente de los compañeros de clase incluídos en las muestras seleccionadas, los alumnos deberán calcular medias, desviaciones estándar y coeficiente de regresión, para completar un cuadro como el que se describe a continuación.

Cuadro 2.
Medias, desvíaciones estándar y coeficiente de regresión para muestra de tamaño n=2

Muestra	Peso promedio (kg)	DE Peso	Altura promedio (cm)	DE Altura	Coeficiente
(1,2)	40,00	8,49	152,00	2,83	0,33
(1,3)	44,00	2,83	154,50	0,71	-0,25
(1,4)	48,00	2,83	157,00	4,24	1,50
(2,3)	38,00	5,66	152,50	3,54	0,63
(2,4)	42,00	11,31	155,00	7,07	0,63
(3,4)	46,00	5,66	157,50	3,54	0,63
Población	43,00	6,83	154,75	4,11	0,55

DE= Desviación estándar

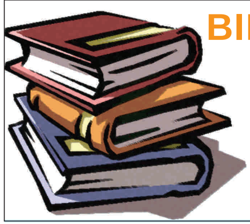
El efecto del tamaño de la muestra sobre la precisión de los estimadores, se podrá apreciar, también, comparando los resultados de las tablas para n=5 y para n=10.

Ahora la clase puede practicar con los métodos que permiten describir la población. Por ejemplo, el docente elaborará dispositivos de tallos y hojas para la altura y el peso, comparará la distribución de la altura y el peso para hembras y varones empleando gráficos de cajas, calculará resumenes de cinco números y la media y la desviación estándar para describir en forma resumida el centro y la variabilidad de los datos, y construirá un modelo de regresión lineal que relacione la altura con el peso empleando el método de mínimos cuadrados.

6. Discusión de resultados.

Los resultados de la descripción permitirán familiarizar al alumno con el comportamiento de cada variable, por separado, para toda la población . Así, por ejemplo, se contará con elementos para discutir sobre la forma de la distribución del peso empleando el dispositivo de tallos y hojas y el histograma. ¿Es asimétrica o simétrica? ¿Cómo resumen la media y la desviación estándar el centro y la variabilidad del lote de datos? ¿Cómo se compara este sumario con el que proporciona el resumen numérico de cinco números? ¿Cuál es el peso mínimo y el peso máximo? ¿Existen pesos atípicamente bajos o altos?

Todos los alumnos deberán elaborar el informe donde se recoge toda la experiencia realizada.

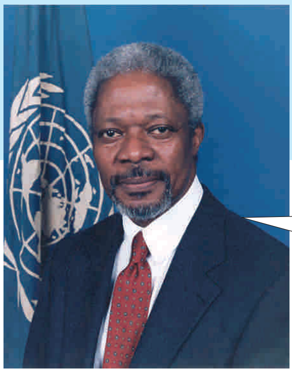


BIBLIOGRAFÍA

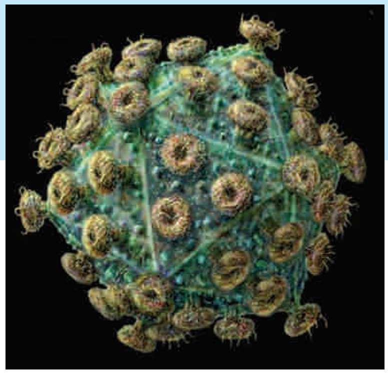
- Díaz-Batanero y Cañizares M. (1996) *Azar y probabilidad*. Editorial Síntesis. España.
- Garza T. (1996). *Probabilidad y estadística*. Editorial Iberoamericana. México.

Páginas web

- http://www.ine.gov.ve
- http://www.amstat.org/



Estadística y VIH/SIDA



"Una conspiración oficial de silencio sobre el SIDA en demasiados países, le ha negado a la gente la información que podría haberles salvado la vida"

Kofi Annan

Secretario General de la Organización de Naciones Unidas, ONU.

A pesar de todo lo que se ha aprendido en más de los 20 años transcurridos desde la descripción de los primeros casos del Síndrome de Inmunodeficiencia Adquirida (SIDA) y el desarrollo de esquemas terapéuticos efectivos, la epidemia causada por el Virus de Inmunodeficiencia Humana (VIH) se ha convertido en una pandemia la cual ha diezmado poblaciones enteras en todo el mundo. Según el Fondo de las Naciones Unidas para la Infancia (UNICEF), en todo el planeta Tierra hay 43 millones de portadores y han muerto 20 millones de personas hasta el presente.

De acuerdo al informe del Programa de Naciones Unidas contra el SIDA (ONUSIDA), una décima parte de las nuevas personas infectadas son menores de 15 años. Se cree que la mayoría de ellos (aproximadamente 90%) ha contraído la infección a través de sus madres por medio del embarazo, parto o lactancia.

Una de las características más resaltantes de la evolución de la pandemia en el mundo ha sido el incremento del número de mujeres afectadas. La razón hombre:mujer en el año 1990 fue de 10:1, mientras que en el año 1999 fue de 5:1.

La principal forma de transmisión del VIH/SIDA es la sexual. Un 90% de la adquisición de la enfermedad es por esta vía.

Gráfico 1. Distribución de edad según universidad

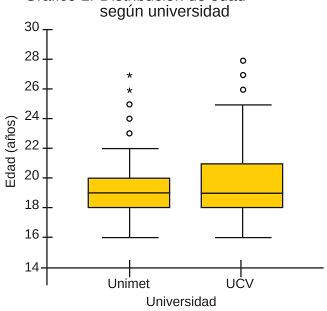


Gráfico 2. Distribución según conocimiento VIH/SIDA

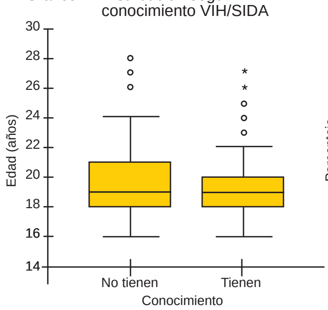


Gráfico 3. Conocimiento sobre VIH/Sida según instrucción jefe de hogar

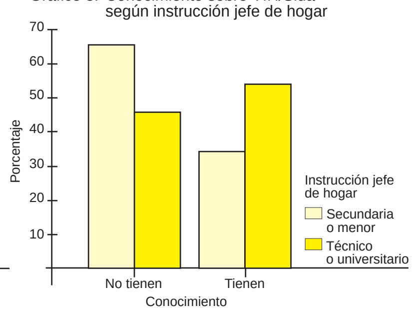


Gráfico 4. Edad: distribución por Universidad conocimiento VIH/SIDA

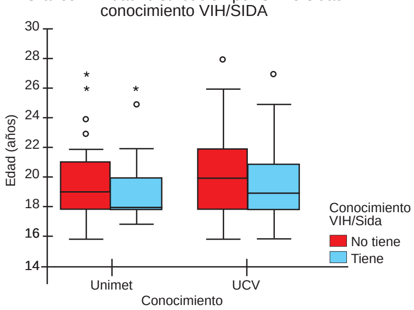
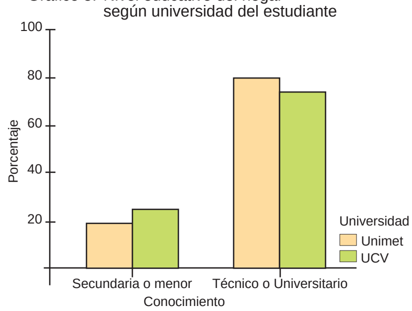


Gráfico 5. Nivel educativo del hogar según universidad del estudiante



Fuente: Dirección de Vigilancia Epidemiológica. MSDS. Venezuela 2001

Dada la trascendencia de este problema, investigadores de la Escuela de Medicina "Luis Razetti" y de la Facultad de Odontología de la Universidad Central de Venezuela (UCV), consideraron de interés estudiar la relación entre el conocimiento sobre VIH/SIDA de jóvenes y factores sociales. Para ello se escogieron aleatoriamente 268 estudiantes de la matrícula del ciclo básico de las Facultades de Ingeniería de la UCV y de la Universidad Metropolitana (UNIMET), en el período académico 2003-2004.

La obtención de información se realizó a través de una encuesta confidencial estructurada. Los datos se analizaron mediante dispositivos de cajas y barras.

Los resultados mostraron que aproximadamente la mitad de los estudiantes en la muestra, tanto de la UCV como de la UNIMET, no tenían informacion suficiente sobre VIH/SIDA. Adicionalmente, se establecieron diferencias entre el conocimiento que el estudiante posee sobre VIH/SIDA y el nivel educativo del jefe del hogar, en el sentido que a menor instrucción del jefe del hogar mayor es el desconocimiento del estudiante sobre VIH/SIDA.

Estos y otros resultados de diversas investigaciones estadísticas sobre VIH/SIDA, reafirman la importancia de llevar a cabo campañas publicitarias de información a través de medios de divulgación masiva que permitan atenuar los efectos de esta terrible pandemia moderna que a todos nos amenaza.