UNIVERSIDAD PEDAGÓGICA NACIONAL

UNIDAD AJUSCO

AREA ACADÉMICA 3

APRENDIZAJE Y ENSEÑANZA EN CIENCIAS HUMANIDADES Y ARTES

CUERPO ACADÉMICO EDUCACIÓN MATEMÁTICA

GUIA DE ESTADÍSTICA DESCRIPTIVA

MTRA. MA. DEL ROSARIO JIMÉNEZ HERNÁNDEZ

MTRO. JUAN DE DIOS HERNÁNDEZ GARZA

2017

PRESENTACIÓN

Esta guía está elaborada con el propósito de que los alumnos que adeudan Estadística Descriptiva de la Licenciatura en Pedagogía cuenten con un material de consulta que apoye el desarrollo de los contenidos temáticos, para que avancen de forma independiente en el aprendizaje de esta asignatura y en su preparación para el examen extraordinario.

CONTENIDOS TEMATICOS

I. IDEAS BASICAS

Tema1. Conceptos generales: población, muestra, variables.
Tema 2. Representatividad de las muestras.
Tema 3. Tipos de variables.
Tema 4. Escalas de medición.

II. DESCRIPCIÖN DE VARIABLES

Tema 1. Descripción gráfica de los datos.
Tema 2. Descripción numérica de los datos.

III. DISTRIBUCIONES DE PROBABILIDAD DE UNA VARIABLE

Tema 1. Distribución de probabilidad para una variable discreta.
Tema 2. Distribución de probabilidad para una variable continua.

IV. DESCRIPCIÖN CONJUNTA DE DOS VARIABLES

Tema 1. Variables categóricas.
Tema 2. Variables numéricas.

I. IDEAS BÁSICAS

Tema 1. Conceptos generales: población, muestra, variables.

Población

De manera general, una población (N) se puede definir como “el mayor conjunto de unidades en el cual se tiene un cierto interés”.

En términos estadísticos, una población se define considerando varios elementos como: las unidades bajo estudio (U_i), los factores comunes que comparten estas unidades (A, B, C,...), la característica (variable: X, Y, ..) que se pretende medir en las unidades de estudio, la forma de medición de las variables (Escala) y los valores observados (O_i) en las unidades de estudio, en una muestra aleatoria y representativa.

La muestra y su selección

Cuando no se tiene información tan extensa como en un censo (estudio de toda la población), resulta más eficiente y práctico analizar una porción de la población, es decir, una muestra (n), la aplicación de la Estadística conlleva ventajas operativas, económicas y en términos de tiempo.

Selección de muestras

El objetivo final de un estudio estadístico es hacer inferencias (extrapolaciones) sobre la población de interés y un paso previo es tener una muestra que sea representativa. Una manera práctica de lograr la representatividad de las muestras es mediante el proceso de aleatorización.

Una muestra aleatoria es aquella en la que todos los individuos de la población tienen la misma probabilidad ($\frac{1}{N}$) de ser seleccionados (muestra autoponderada).

Extrapolación muestra a población

Una solución para el problema de inferir de muestra a población, la extrapolación, en presencia de aleatoriedad es tomar una muestra grande con igual probabilidad de que cada elemento de la población esté en la muestra (Diseño Autoponderado). En este caso opera la teoría de probabilidad y tendremos la consistencia y normalidad de estimadores de promedios y proporciones. Esto aunque la población sea finita y se tome sin reemplazo la muestra, si $n < N$.

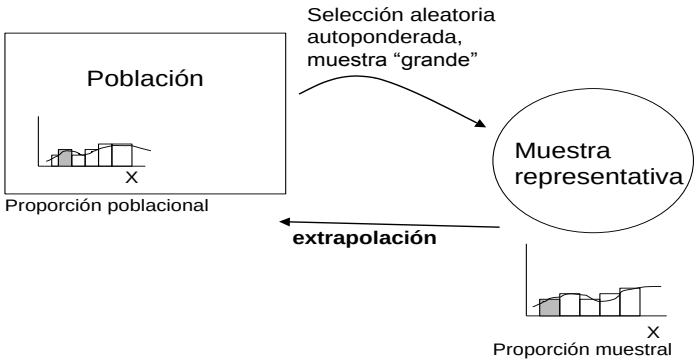
Tema 2. Representatividad de las muestras.

En el caso de muestras autoponderadas (*n* “grande”) los promedios muestrales se parecen mucho a los poblacionales. Se puede afirmar que esto ocurre porque en la muestra, la distribución de los valores de la(s) variable(s) de estudio también se parece a la de la población. Se dice entonces que la muestra es representativa de la población.

La representatividad es para la(s) variable(s) de interés en el estudio, aunque no se tenga para otras variables. Ejemplo, un grupo de 45 alumnos de la Especialidad de Estadística Aplicada del IIMAS-UNAM, es posible que pueda ser considerado como una muestra de sujetos entre 23 a 45 años, de clase media, en México y clínicamente sanos para el estudio del contenido de hemoglobina en sangre. Sin embargo, si el estudio pretende evaluar los conocimientos de Estadística de la población de la UNAM, ciertamente el grupo no es una muestra representativa, porque fueron seleccionados por su interés y conocimientos de estadística, cosa que no ocurre en otros programas educativos de la Institución.

Otra situación común en que la muestra no es representativa de la población de interés es el caso de las encuestas por teléfono. Si se hace una encuesta, seleccionando a las personas a entrevistar a partir del directorio telefónico, se está haciendo a un lado a una porción importante de habitantes (aquellos que no tienen teléfono). En este caso en particular, no todas las personas tienen la misma probabilidad de ser seleccionadas.

Lo que importa es que las variables de interés en la muestra presenten una distribución de frecuencias semejante a las de la población. Si la muestra es grande y autoponderada se tendrán elevadas probabilidades de que esto suceda. En este caso la extrapolación (inferencia) tendrá errores pequeños.



Una muestra aleatoria permite estimar características de la población. Cuando nos referimos a las características de una población, estamos pensando en los valores numéricos que la caracterizan. A esto valores se les llama *parámetros*. Cuando hacemos un cálculo basándonos en los valores de la muestra, tenemos un valor llamado *estimador o estadístico muestral o simplemente*

estadístico. Los estimadores sirven, a nivel descriptivo, para resumir información de un conjunto de datos y a nivel inferencial para estimar los parámetros o valores poblacionales.

Para diferenciar ambos conceptos-*estimador y parámetro* - podemos pensar en que nuestra población de interés está definida como los alumnos (unidades de estudio) de la UPN (Factor común A), del turno vespertino (Factor común B) que cursan Estadística (Factor común C), y que nos interesa conocer el tiempo (X) promedio que tardan en llegar de su casa la universidad. Existen valores reales de esa variable, aunque no los conozcamos o aunque sea difícil medirlas. Estos valores reales en la población de alumnos son los *parámetros o valores poblacionales como la media de la población (μ)* o *la varianza de la población (σ^2).*

Por otra parte, si consideramos que los alumnos de un grupo son una muestra (n) representativa de la población (N), entonces podemos calcular el tiempo promedio ($\bar{x}$) en llegar de su casa a la universidad y este valor promedio sería el *estimador* de la media poblacional. Al estimador de la varianza poblacional (σ^2) se le llama varianza muestral (s^2).

.
Una idea básica e importante es el hecho de que **el tamaño de la muestra no depende del tamaño de la población, sino de la variabilidad presente en la población.** Por ejemplo, si deseamos estimar la edad de alumnos de bachillerato y si además se define la población como alumnos de quinto semestre, basta una muestra pequeña (menor a 30) porque la edad en este nivel es bastante homogénea. En contraparte si deseamos estimar la proporción de alumnos zurdos, la muestra debe ser bastante grande (mayor a 30), porque la característica ser zurdo es bastante rara.

Tema 3. Tipos de variables

Las variables son la herramienta fundamental de la Estadística porque dependiendo del tipo de variable es el análisis que se realiza con cada una de ellas. Los datos de una variable numérica se pueden analizar calculando promedios como las medidas de tendencia central (media aritmética, mediana y moda) y las medidas de dispersión (varianza y desviación estándar). Si la variable es categórica ordinal solo se le puede analizar calculando la mediana y la moda (como medidas de centralización); pero si la variable es nominal el único valor que se le puede calcular es la moda.

Por ejemplo, una maestra, lleva a cabo un estudio socioeconómico de sus alumnos. Cada familiar del alumno entrevistado reporta entre otras el nivel socioeconómico que puede ser alto (A), medio alto (MA), medio (M), medio bajo (MB), o bajo (B). Investiga el número de hijos por familia, que puede ser desde cero hasta cualquier número entero positivo que corresponda a la magnitud observada. El nivel académico de los integrantes de la familia. El tipo de vivienda donde se pregunta si es propia o paga renta, el tipo de piso si es de

tierra de cemento u otro; el número de cuartos con que cuenta y cuantos se utilizan para dormir, el gasto en transporte, etc. Todas estas características son variables.

Estas características de interés no presentan un solo valor determinado y predecible con exactitud en cada medición observada.

Se concluye que una característica de interés que tienen en común todos los elementos de un conjunto de individuos de tal manera que al medirla se obtienen valores diferentes e impredecibles se le llama ***variable*****.**

A continuación se presenta una tabla con las escalas de medición y sus características para las diferentes variables:

Escala de medición	**Operaciones básicas**	**Cambios permitidos**	**Ejemplos de variables**	**Algunos valores**
Nominal	Determinación de igualdad o pertenencia a una categoría	cambios en los nombres de las categorías	Género Religión	M, F C, P, A
Ordinal	Determinación del grado de intensidad	Cambios que mantengan las relaciones de orden	Calificación	NA, S, B, MB
Intervalo	Determinación de igualdad de intervalos o diferencias	Se puede cambiar la unidad de medida y el origen	Tiempos de traslado	Números enteros y fraccionarios
Razón	Determinación de igualdad de razones o proporciones	Se puede cambiar la unidad de medida pero no el origen	Porcentajes	Números enteros y fraccionarios

Las variables numéricas continuas surgen al hacer mediciones de una cierta magnitud, como medir el tiempo, el peso o la estatura.

Una variable continua puede tomar cualquiera de los infinitos valores de un intervalo (valores tan próximos como se quiera).

Las variables numéricas discretas surgen al hacer conteos o enumeraciones. Por ejemplo número de hermanos que tienen los alumnos de un grupo o el número de veces que asisten semanalmente a la biblioteca.

Las variables discretas pueden tomar un número finito o infinito numerable de valores.

Actividad

Clasifica cada una de las siguientes variables y determina sus posibles valores o algunos de ellos.

1. Grado que cursan los alumnos en una escuela primaria.

2. Número de hijos que tendrá un matrimonio.

3. Número de puntos de la cara superior al lanzar un dado legal una vez.

4. Peso atómico de los elementos químicos.

5. Calificación obtenida por un estudiante al final del curso de Estadística.

6. Género de los alumnos que cursarán el sexto semestre este ciclo escolar.

7. Número de teléfono celular de los alumnos de la Licenciatura en Psicología Educativa de la UPN.

8. Fecha de los próximos eclipses solares visibles en México.

9. Número de alumnos a admitir en el bachillerato de la UNAM para el próximo año lectivo.

10. Edad de los alumnos de secundaria del municipio de Naucalpan

Tema 4. Escalas de medición

Cuando las variables son numéricas, se utilizan, en su medición, las escalas de intervalo y de razón. En la escala de intervalo se puede cambiar el origen y la unidad de medida, por ejemplo en el tiempo (en minutos) que hacen los alumnos de su casa para llegar a la universidad, el origen puede ser de 15 minutos y la unidad de medida puede cambiar a "unidades de 10 minutos". Si el objetivo es conocer el número de hermanos, se usa la escala de razón (no se puede cambiar la unidad de medida ni el origen).

Los valores de una variable continua se suelen agrupar en intervalos llamados *intervalos de clase.* El punto medio entre los extremos de cada intervalo se llama *marca de clase, punto medio de clase o punto medio del intervalo.* Siempre que se agrupe una variable por intervalos se produce una pérdida de la información, pues lo que se tiene en cuenta es la pertenencia o no de cada dato al intervalo y no su valor exacto.

La escala ordinal se usa en situaciones donde los valores de la variable, comúnmente categórica (ordinal), se pueden jerarquizar u ordenar, asignando valores como por ejemplo Excelente, Bueno, Regular o Malo, pero no se pueden realizar operaciones aritméticas entre estos valores.

La escala nominal se usa cuando se tienen variables categóricas (nominales) como por ejemplo el tipo de música preferido o preferencia por algún refresco.

A continuación se presentan tres variables medidas en la **escala de intervalo** (Edad), **de razón** (No. de hijos) y **Nominal** (Edo. Civil).

Los datos siguientes proporcionan información del Perfil Socio-económico de 305 estudiantes de la NILITS (Nivelación en la Licenciatura de Trabajo Social). Universidad de Guadalajara (2009).

Variable Edad. Variación: de 26 a 66 años.

Edad (años)	No. de estudiantes (Frecuencia)
26-30	33
31-35	47
36-40	72
41-45	48
46-50	57
51-55	26
56-60	18
61-66	4

Variable No. de hijos. Variación: de 0 a 6 hijos

Número de hijos	No. de estudiantes (Frecuencia)
0	56
1	53
2	83
3	62
4	30
5	15
6	6

Variable estado Civil

Estado Civil	No. de estudiantes (Frecuencia)
Soltero	81
Casado	164
Divorciado	50
Viudo	10

II. DESCRIPCIÓN DE VARIABLES

Tema 1. Descripción gráfica de los datos

Cuando se está tratando con una gran cantidad de datos es conveniente agruparlos en intervalos, para lo cual es necesario considerarlos ordenados dentro de ese intervalo de acuerdo a su frecuencia que corresponde al número de veces que los datos considerados se repiten.

- Los intervalos o clases deben ser del mismo tamaño o amplitud.

- Los intervalos deben construirse de manera que no haya datos que pertenezcan a dos intervalos diferentes, es decir, los intervalos deben ser ajenos y no traslaparse.

- Los límites de clase que corresponden, el inferior al menor valor de la variable en cada intervalo y el superior al mayor valor de la variable en el intervalo.

- Límites reales de clase que se localizan en medio del límite superior de un intervalo y del límite inferior del siguiente.

Además es necesario determinar algunos valores que servirán para analizar y representar al conjunto de datos agrupados en intervalos, tales como:

- Marca de clase o punto medio del intervalo.

Es el valor representativo de cada intervalo y corresponde al valor de la variable situado exactamente en el centro de cada uno de ellos.

- Tamaño o amplitud del intervalo.

Es el tamaño que corresponde a cada intervalo y que se obtiene como la diferencia del límite real superior menos el límite real inferior de cada intervalo

- Frecuencia absoluta

En una variable estadística *la frecuencia absoluta (o simplemente frecuencia)* de un determinado valor, es el *número de veces que la variable toma dicho valor*.

- Frecuencia acumulada

La *frecuencia acumulada* de un valor *x* de la variable, *es la suma de las frecuencias absolutas de los valores de la variable menores o iguales a x*.

- Frecuencia relativa

La *frecuencia relativa* de un determinado valor de la variable *es el cociente entre la frecuencia absoluta de dicho valor y el número total de valores*

observados. A la frecuencia relativa también se le llama proporción, probabilidad estimada o porcentaje.

- Frecuencia relativa acumulada.

La *frecuencia relativa acumulada* de un determinado valor de la variable, *resulta de sumar a su frecuencia relativa las frecuencias relativas de sus valores anteriores*.

Las cuatro frecuencias anteriores, se suelen agrupar en cuadros llamados *tabla de distribución de frecuencias*.

Tabla de Distribución de frecuencias

Una distribución es una descripción de cómo están repartidos los elementos de una muestra (o de una población). La Estadística se interesa por las distribuciones de los posibles valores de una variable, relacionando cada valor con la frecuencia a la que sucede. La frecuencia puede observarse experimentalmente o calcularse a partir del concepto de frecuencia relativa.

Cuando la información estadística se coloca en una tabla de distribución de frecuencias, debe de reunir las siguientes características:

a) Debe ser una matriz con tantas entradas como sean necesarias.
b) En cada línea o columna se debe de especificar la variable que se está midiendo.
c) Los datos tienen que estar descritos con claridad y precisión.
d) En su caso, deben indicarse las unidades: miles, cm, kg, etc.
e) Indicar la fuente de los datos.
f) Y, en fin, *todas las características que contribuyan a que las tablas resulten claras y no se presten a confusión.*

Ejemplo de una Tabla de Distribución de Frecuencias

Los datos agrupados presentados en la tabla siguiente corresponden al gasto en pasajes por día (en $), de una muestra de 65 alumnos de una escuela secundaria.

La variación (variabilidad o dispersión) de esta variable va desde $9 a $43.

Gasto ($)	Límites reales del intervalo	Punto medio del intervalo (X_j)	Número de alumnos (Frecuencia)	Frecuencia acumulada	Frecuencia relativa
9 - 15	8.50 - 15.50	12.00	7	7	0.1077
16 - 22	15.50 - 22.50	19.00	10	17	0.1538
23 - 29	22.50 - 29.50	26.00	25	42	0.3846
30 - 36	29.50 - 36.50	33.00	13	55	0.2000
37 - 43	36.50 – 43.50	40.00	10	65	0.1538

Gráficas

Las gráficas (o gráficos) son muy utilizados en la prensa, en la televisión y en los libros para presentar los datos de una forma más vistosa. Además, también se consigue que, de un solo vistazo, podamos darnos cuenta de los detalles fundamentales.

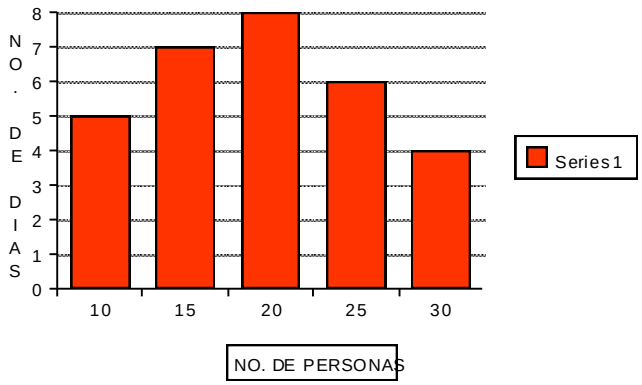
En ocasiones, cuando se nos habla de una persona o lugar, que no conocemos, preferimos que nos muestren una fotografía además de las características que nos puedan platicar. Así pues, resulta conveniente, además de tabular un conjunto de datos, proveer una imagen gráfica que sea explicativa por si sola. Cuando los datos son cualitativos resultan adecuadas las gráficas de barras o circulares. Si los datos son cuantitativos, pueden ser adecuadas el polígono de frecuencias o los histogramas de frecuencias.

Gráfica de una variable numérica discreta

Si los datos corresponden a una variable numérica discreta, puede ser adecuada una gráfica de barras que proporciona información acerca de la forma de la distribución de los datos (simétrica o asimétrica) y la variabilidad.

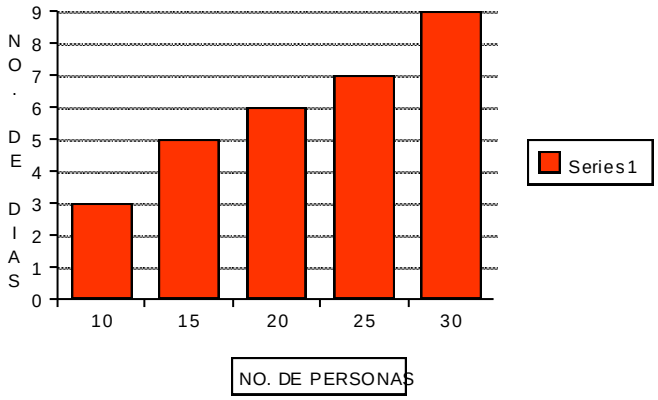
Por ejemplo se define la variable X: número de personas atendidas diariamente durante 30 días en una oficina:

No. de personas	10	15	20	25	30
No. de días	5	7	8	6	4



La distribución de frecuencias es aproximadamente simétrica, por lo que los valores se agrupan alrededor de 20 personas atendidas por día, y se puede decir que, en promedio (media aritmética, mediana o moda), se atienden aproximadamente 20 personas por día.

Si en el ejemplo anterior, la atención de las personas por día fuera 3, 5, 6, 7, 9, la frecuencia más alta (9 días) correspondiente al valor 30 personas. Modifica el promedio anterior.



Cuando la distribución de frecuencias es asimétrica, se debe reflexionar acerca del valor que se utilizará como promedio.

Gráfica de una variable numérica continua

Un histograma de frecuencias es la gráfica más común para representar datos cuantitativos. Esta gráfica nuestra como es la distribución en cuanto a la forma de los datos (simétrica, asimétrica, bimodal, concentraciones o huecos en los datos, etc.). Cuando el histograma se basa en datos provenientes de una muestra, la gráfica solamente describe el comportamiento de los datos en la muestra, pero podría sugerirnos que la población tiene una forma similar, sin embargo no se puede afirmar que la población tenga la misma forma (no se pueden hacer inferencias). Por lo tanto, el histograma es una técnica solamente descriptiva.

Procedimiento para la construcción del histograma

a) Ordene los datos de manera creciente, de menor a mayor.
b) Obtenga el rango de sus mediciones: diferencia entre el dato mayor y el menor.
c) Decida cuántas clases o intervalos tendrá el histograma (entre 5 y 10).
d) Divida el rango entre el número de clases (o número de intervalos), para obtener la amplitud del intervalo.
e) Forme los intervalos comenzando con el menor valor observado y sumando a este valor la amplitud del intervalo (ningún dato debe estar en la frontera o límite de dos clases).
f) Cuente cuántos datos caen en cada clase, (para conocer la frecuencia absoluta). Puede también calcular la frecuencia relativa, es decir, que proporción o porcentaje de casos cae en cada intervalo.
g) Construya la gráfica ubicando en el eje horizontal los intervalos y en eje vertical las frecuencias absolutas o las frecuencias relativas.

Características de los gráficos

Las principales características que debe reunir un gráfico son:

a) Debe ajustarse a la realidad de los datos que representa.

b) Ha de ser claro: fácil de leer y entender.

c) Debe de llevar el título y todas las indicaciones necesarias para una correcta interpretación.

Los gráficos pueden ser *simples*, si representan directamente las frecuencias absolutas o las frecuencias relativas.

Los gráficos son *acumulativos* si representan los valores de las frecuencias acumuladas.

Diagramas de tallo y hojas.

El diagrama de tallo y hojas puede considerarse como un '' hibrido'' de histograma y tabla de frecuencias, que permite organizar gráficamente los valores, destacando características tales como simetría, dispersión, casos extraordinarios, concentraciones y huecos. Los diagramas de tallo y hojas constituyen una alternativa innovadora que facilita el análisis de datos, sin necesidad de realizar pesados cálculos, aprovechando el hecho de que los datos se encuentran ordenados.

Para ejemplificar esta situación, analicemos los datos siguientes referidos a la edad de 55 personas

27	23	22	38	43	54	35	26
25	23	22	52	31	30	41	45
29	28	27	25	29	28	24	37
26	33	25	27	25	34	32	36
21	23	24	18	48	23	16	38
28	18	20	29	27	43	28	29
18	22	32	33	26	31	23	

a) Se forma el tallo con la cifra de las decenas (las cifras de las decenas varían de 1 a 5).

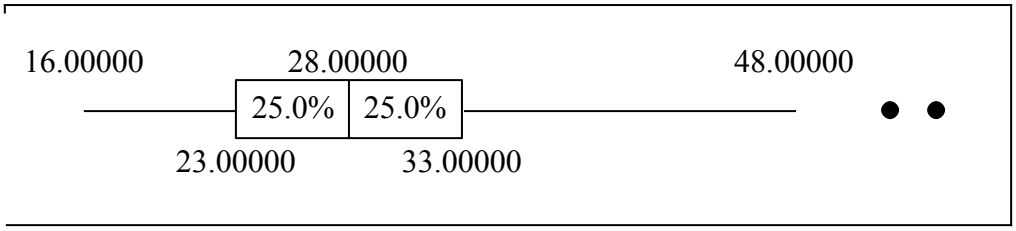
TALLO

1
2
3
4
5

b) Con las cifras de las unidades de los datos se forman las hojas, ordenadas de menor a mayor:

1 6888
2 01222333334445555666777788889999
3 012233456788
4 13358
5 24

A partir del diagrama de tallo y hojas, se puede construir otra representación gráfica llamada diagrama de caja. Este diagrama se construye utilizando tres valores descriptivos (cuartiles) que dividen al conjunto de datos ordenados en cuatro partes que contienen el 25% de los datos cada una. El diagrama muestra la siguiente información: valor mínimo (16 años), primer cuartil (Q_1=23 años), segundo cuartil o mediana (Q_2=28 años), tercer cuartil (Q_3=33 años) y valor máximo (58 años). También se observan dos puntos, que corresponden a dos datos (edades) atípicos (52 y 54 años).



Actividad

Representar gráficamente la siguiente información correspondiente a alumnos universitarios:

Número de Hermanos	Nivel Socio-económico	Estatura (cm)
7	Alto (A)	180
2	Bajo (B)	175
2	B	178
7	B	183
3	Medio Alto (MA)	180

1	A	180
1	Medio (M)	180
1	B	182
1	M	177
0	M	184
3	MA	172
4	A	173
0	A	162
4	M	194
1	M	174
1	MA	165
1	B	167
2	B	155
2	B	174
2	M	160
1	M	155
1	A	174
1	MA	162
3	M	180
1	M	160
2	A	170
1	A	155
1	M	150
1	B	166
3	MA	170
1	A	160

1	M	183
2	M	163
3	B	162
2	B	154
5	A	155

Tema 2. Descripción numérica de los datos.

Las medidas numéricas descriptivas, resumen la información de un conjunto de datos.
En una población, los parámetros más importantes son los *que ubican el centro de la distribución* y ***los que describen la dispersión o variación de los datos****.* A estos se les llama respectivamente, Medidas de Tendencia Central y **Medidas de Dispersión, Variación o Variabilidad**, por tal motivo resulta necesario, en primera instancia, calcular estos tipos de medidas a los datos de la muestra y, en segundo lugar, cuando se pretende hacer inferencias sobre los parámetros de la población, estas medidas muestrales serán los estimadores para tal efecto.

Medidas de Tendencia Central

Las medidas de tendencia central más comunes son: el promedio aritmético o media aritmética (o simplemente media), la mediana y la moda. Estas medidas sirven para localizar el centro de una distribución de datos, es decir, ubican el valor alrededor del cual se encuentra un conjunto de datos. Aunque tienen un mismo propósito, estas medidas, de manera general, tendrán un valor diferente (sólo en algunos casos muy particulares, se da que la media, la mediana, y la moda sean iguales, o que dos de ellas coincidan).

Si los datos que se tienen fueran de una población, las medidas de centralización se calculan de la misma manera que en la muestra, solamente es necesario tener presente si se habla de un parámetro o de un estimador, según sea el caso.

La media aritmética o promedio aritmético.

La suma de n datos x_1, x_2,,,,,x_n, se simboliza: $\sum_{i=1}^{n} Xi$. Si los datos se repiten (están asociados a frecuencias), la suma se simplifica si se multiplica cada uno de los datos por su frecuencia: $(x_1)(f_1)+(x_2)(f_2)+\ldots..+(x_k)(f_k)$; esto se simboliza como $\sum_{i=1}^{k}(X_i)(F_i)$. La media se define como el resultado que se obtiene al dividir la suma de todos los valores de la variable entre el total de valores (n).

$$\bar{x} = \frac{\sum_{i=1}^{k}(X_i)(F_i)}{n}$$

Cuando los datos se agrupan en intervalos, para calcular la media, los valores de la variable se sustituyen por los puntos medios de cada intervalo, se multiplican por las frecuencias respectivas, se suman estos productos y se divide la suma entre el total de valores (el valor de la media, obtenido de esta manera, es aproximado).

Ejemplo

Los datos agrupados presentados en la tabla siguiente corresponden al gasto en pasajes por día (en $), de una muestra de 65 alumnos de una escuela secundaria. Calcular el promedio del gasto semanal de los 65 alumnos.

Gasto ($)	Límites reales del intervalo	Punto medio del intervalo (X_j)	Número de alumnos (Frecuencia)	$(X_i)(F_i)$
9 - 15	8.50 - 15.50	12.00	7	(12)(7)=84
16 - 22	15.50 - 22.50	19.00	10	(19)(10)=190
23 - 29	22.50 - 29.50	26.00	25	(26)(25)=650
30 - 36	29.50 - 36.50	33.00	13	(33)(13)=429
37 - 43	36.50 – 43.50	40.00	10	(40)(10)=400

$\bar{x} = \frac{84+190+650+429+400}{65}$=26.97. Este promedio se interpreta como: los 65 alumnos, en promedio, gastan $26.97 o los 65 alumnos gastan cada uno $26.97, de manera que (65)(26.97)=1753 pesos de gasto total.

Si los datos no se repiten el cálculo de la media se simplifica a la expresión $\bar{x} = \frac{\sum_{i=1}^{k}(X_i)}{n}$, que también se puede expresar como $\sum_{i=1}^{n} X_i = n\bar{x}$

Por ejemplo, calcular la media para la edad y la estatura de 9 alumnos. Los datos son

Alumno	Edad (años) (X)	Estatura promedio (m) (Y)
1	9	1.37
2	10	1.42
3	11	1.50
4	12	1.58
5	13	1.60
6	14	1.69
7	15	1.71
8	16	1.73
9	17	1.74
Suma	117	14.34

$\bar{x} = \frac{9+10+11+\cdots 16+17}{9} = \frac{117}{9} = 13\ años$, este resultado significa que, teóricamente, los 9 alumnos tienen 13 años cada uno, por lo que (9)(13)=117 años en total.

$\bar{y} = \frac{1.37+1.42+\cdots 1.73+1.74}{9} = \frac{14.34}{9}$ = 1.593 m, con este promedio podemos decir que, la estatura de los 9 alumnos es de 1.593 m, la estatura total es (9)(1.593)=14.337 m

Propiedades de la Media Aritmética.

A continuación se proporcionan las propiedades más importantes de la media aritmética.

Propiedades numéricas.

1. *La media de un conjunto de datos es siempre un valor perteneciente al rango de la variable. En cualquier caso (por rara que sea la distribución de los datos, simétrica o asimétrica, por ejemplo), tanto la media como la mediana y la moda, se encuentran entre los valores máximo y mínimo de los valores observados.*
2. *La media puede no coincidir con ninguno de los valores de los datos. Es decir, puede ser un número que no tenga sentido en el contexto propuesto, por ejemplo, si el número de hermanos para 5 personas es 1, 4, 3, 0 y 5, el promedio es 2.6 hermanos.*
3. *En el cálculo de la media intervienen todos los valores de los datos.*
4. *La media se ve afectada por cualquier cambio extremo en los valores de los datos. Si en el ejemplo anterior existiera una persona con 13 hermanos (en vez de 5), este valor extremo modifica la media de 2.6 a 4.2 hermanos*

Propiedades algebraicas.

1. *La media conserva el cambio de origen y escala: si el promedio de calificación de un alumno es por ejemplo, 8.75 en la escala de 1 a 10, en la escala de 1 a 100, el promedio es 87.5*
2. *La media de la suma de dos o más variables es la suma de las medias (en el caso de la mediana y la moda, esta propiedad no se cumple).*
3. *La media no está definida para datos ordinales o nominales (la media no tiene sentido si la variable es categórica o cualitativa).*
4. *La media, la mediana y la moda, consideradas como operación, no tienen ningún elemento neutro, ni la propiedad asociativa.*

Propiedades estadísticas.

1. *La media es un valor representativo de un conjunto de datos. La media es menos resistente (se ve afectada por cualquier cambio en los datos) que la mediana y la moda.*
2. *La media coincide con el centro de gravedad del conjunto de datos.*
3. *La suma de las desviaciones de un conjunto de datos con respecto a la media es cero.*
4. *En distribuciones simétricas, la media, la mediana y la moda coinciden.*
5. *Es respecto a la media cuando la suma de los cuadrados de las desviaciones es mínima.*

Actividades

1. Hay 10 personas en un ascensor, 4 mujeres, 4 hombres y 2 niños. El peso medio de las mujeres es de 60 kg, el peso medio de los hombres es de 80 kgs. y el peso medio de los niños es de 35 kg, ¿cuál es el peso medio de las 10 personas en el ascensor?
2. Cada estudiante de un grupo de 20 estudiantes pesa 86 kg. en promedio. Se sabe que 9 personas del grupo pesan en promedio 75 kg. cada una. Del grupo de los 11 estudiantes restantes, ¿cuánto pesa en promedio cada uno?
3. De los 200 alumnos que presentaron un examen de 12 reactivos, el 10% responde correctamente a 3 reactivos, el 50% a 7 reactivos, el 30% responde correctamente a 10 reactivos y el resto al total de reactivos del examen. Organice los datos en una tabla de distribución de frecuencias y calcule el número promedio de reactivos resueltos correctamente.
4. En un equipo de futbol, el promedio de altura de los jugadores es de 165 cm. En el último torneo se incorporaron tres jugadores suplentes que miden 170, 175 y 180 cm ¿Cuál es ahora el promedio de altura de los 11 jugadores?
5. El promedio de 9 alumnos es 7 y el promedio de otros 7 alumnos es 9. Encontrar el promedio de los 16 alumnos.

La Mediana

Al valor central en una serie de n datos ordenados en forma creciente (o decreciente) se le llama la mediana

Cuando n es un número impar la mediana corresponde al valor central de los datos.

Ejemplo

Determinar la mediana para la edad y la estatura de los 9 alumnos en la tabla siguiente (observe que los datos ya están ordenado de manera creciente)

Alumno	Edad (años) (X)	Estatura promedio (m) (Y)
1	9	1.37
2	10	1.42
3	11	1.50
4	12	1.58
5	13	1.60
6	14	1.69
7	15	1.71
8	16	1.73
9	17	1.74
Suma	117	14.34

La mediana para la edad de los 9 alumnos corresponde al quinto valor: 13 años.

Este resultado se interpreta como: el 50% de los alumnos tiene 13 años o menos

La mediana para la estatura corresponde al valor central (1.60 m) La interpretación es: el 50% de los alumnos tiene estatura de 1.60 m o menos.

Cuando n es un número par, la mediana corresponde al promedio de los dos valores centrales.

Por ejemplo, los datos siguientes corresponden a los contenidos de azúcar y cafeína para 8 refrescos de cola (Revista del Consumidor, Profeco. 2003).

Marca	Azúcar (g/100 ml)	Cafeína (mg/100 ml)
Big Cola	10.9	12
CM	10.3	6
Coca Cola	10.6	15
Great Value	10.2	5
Hola Cola	10.4	13
Pepsi Cola	11.1	14
Pepsi Limón	11	16
Royal Cola	11	12

Ordenando los datos en forma creciente,10.2, 10.3,10.4, 10.6, 10.9, 11, 11, 11.1, la mediana para el contenido de azúcar es: $\frac{10.6+10.9}{2} = 10.75\ g/100\ ml$, lo que se interpreta como; el 50% de los refrescos de cola tienen contenido de azúcar de 10.75 g/100 ml o menos.

Ordenando los datos en forma creciente, 5, 6, 12, 12, 13, 14, 15, 16, la mediana para el contenido de cafeína es 12.5 mg/100ml. Este valor significa que el 50% de los refrescos de cola tienen contenido de cafeína de 12.5 mg/100 ml o menos.

Cuando la variable es categórica ordinal, también se puede determinar un valor para la mediana (los valores categóricos ordinales se pueden jerarquizar).

Ejemplo

Los datos siguientes corresponden a la opinión de n=11 personas acerca de la atención en una tienda de autoservicio: Excelente (E), Pésima (P). Buena (B). Regular (R), Mala (M), R, B, E, M, M, R

Ordenando los datos: P, M, M, M, R, R, R, B, B, E, E. En este caso la mediana es el valor R.

Para el caso de n par, por ejemplo, P, M, M, M, R, R, R, B, se tienen dos medianas: M y R.

Actividades

1. En una empresa solamente pueden ingresar aspirantes que obtengan calificaciones superiores a la mediana en el examen de conocimientos. Este año se presentaron 12 aspirantes que obtuvieron los siguientes puntajes: 7.5 9.5 7.5 9.7 7.8 9.2 8 9.2 8.1 9 8.2 8.8, ¿cuáles son los puntajes aceptados?
2. Las calificaciones obtenidas por un estudiante en 7 asignaturas son: S S MB B S B MB.

   a) ¿Cuál es el valor de la moda?
   b) ¿Cuál es el valor de la mediana?

Propiedades de la mediana

a) Al calcular la mediana no usamos todos los valores observados de la variable, lo que la limita como medida de tendencia central.

Ejemplo

Supongamos que medimos la estatura de tres personas, de las cuales la primera mide 160 cm y la segunda 165 cm. Si la mediana es 165 cm, ¿cuánto mide la tercera persona? Nadie podría dar un valor exacto como respuesta. Sin embargo, si la media aritmética es 165 cm, podemos afirmar que, puesto que la media es un valor representativo, la suma de la estatura de las tres personas es (3)(165)=495 cm, por lo tanto la tercera persona mide 170 cm.

b) No puede ser aplicada a distribuciones de variables cualitativas nominales.

c) Como medida de tendencia central, presenta ciertas ventajas frente a la media en algunas distribuciones ya que no se ve afectada por valores extremos de las observaciones. *La mediana es invariante si se disminuye una observación inferior a ella o si se aumenta una superior,* puesto que sólo se tienen en cuenta los valores centrales de la variable. Por ello es adecuada para distribuciones asimétricas o cuando existen valores atípicos.

d) *Conserva los cambios de origen y de escala.* Si sumamos, restamos, multiplicamos o dividimos cada elemento del conjunto de datos por un mismo número esta operación se traslada a la mediana. Ello hace que ésta se exprese en la misma unidad de medida que los datos.

e) La mediana es *un estadístico resistente:* con pequeñas fluctuaciones de la muestra no cambia su valor. Se pueden cambiar uno o varios datos sin que por ello cambie el valor de la mediana, basta con no modificar las dos partes del mismo tamaño en que ésta divide a la distribución.

f) *Si los datos son ordinales la mediana existe,* mientras que la media aritmética no tiene sentido, puesto que su cálculo se basa en los valores (numéricos, necesariamente) de los datos.

g) *Para datos agrupados en intervalos con alguno de ellos abierto también es preferible la mediana a la media.* En estos casos, o bien se prescinde del intervalo abierto, o no es posible calcular la media ya que faltaría una de las marcas de clase, la correspondiente a este intervalo.

Comparación de los tres promedios

Actividades

Un sindicato y una empresa sostienen un debate respecto a los salarios de los trabajadores. El sindicato reporta que los obreros reciben en promedio $ 4000 por mes. El gerente dice que el pago promedio es de $ 8364 mensuales. Un inspector de impuestos afirma que es de $ 7000 por mes. ¿Quién tiene la razón?

Salario mensual	Número de empleados
$ 3000 a $ 5000	5
$ 6000 a $ 8000	1
$ 9000 a $ 11000	0
$ 12000 a $ 14000	5

En esta situación es necesario observar que existen 5 datos atípicos (de $12000 a $14000) que se alejan demasiado de los otros 6, lo que implica que la media aritmética se altera (lo que no ocurre con la mediana).

a) Calcule el sueldo medio ($\bar{x}$).
b) Calcule el sueldo mediano (Me).
c) Calcule el sueldo modal (Mo).

Las medidas anteriores se pueden interpretar de la siguiente manera:

- La media ($\bar{x}$) indica que si el dinero fuera repartido equitativamente, cada trabajador recibiría la misma cantidad.
- La mediana (Md) hace notar que la mitad de los empleados recibe menos que el valor de ella, y que la otra mitad tiene salarios más altos.
- La moda (Mo) solamente señala el sueldo más común.

Es importante reflexionar que ante un valor promedio (Media, Mediana o Moda), resulta necesario preguntarse:

1. ¿Qué tipo de promedio se utilizó?
2. ¿Se incluyen todos los datos?
3. ¿Qué se pretende con este promedio?
4. ¿Quién lo dice?
5. ¿La forma de la distribución de los datos es más o menos simétrica?

La forma de la distribución de los datos es una característica importante para elegir una medida de tendencia central adecuada.

Si la distribución de los valores de la variable es aproximadamente simétrica: la media y la mediana son casi iguales.

Medidas de centralización para datos agrupados

Se dice que los datos agrupados son aquellos que se agrupan en intervalos de clase y que se analizan considerando a la marca de clase (Punto medio del Intervalo) como el valor que corresponde a todos los datos del intervalo, es decir, el análisis ya no se realiza con los datos brutos.

Para que un mejor entendimiento del cálculo de las medidas de centralización para datos agrupados, utilizaremos el siguiente:

Ejemplo.

Se tiene que el siguiente conjunto de datos corresponde a la edad (en años) de los habitantes de una colonia de la delegación de Tlalpan, que asisten a la escuela (a partir de la primaria).

Edad (años)	No. de habitantes
7 a 10	5
11 a 14	11
15 a 18	23
19 a 22	31
23 a 26	16
27 a 30	10
31 a 34	4

Inicialmente, se obtienen las columnas correspondientes a los puntos medios de los intervalos (marcas de clase) (x_i), límites reales de clase ($LRI \ y \ LRS$), frecuencias acumuladas (fa) y las necesarias ($f_i x_i$) para el cálculo de la media aritmética.

x_i	LRI	LRS	fa	$f_i x_i$
8.5	6.5	10.5	5	42.5
12.5	10.5	14.5	16	137.5
16.5	14.5	18.5	39	379.5
20.5	18.5	22.5	70	635.5
24.5	22.5	26.5	86	392.0
28.5	26.5	30.5	96	285.0
32.5	30.5	34.5	100	130.0

$$\sum f_i x_i = 2002$$

La media aritmética se obtiene con $\bar{x} = \dfrac{\sum_{i=1}^{n} f_i x_1}{n}$

Donde: f_i es la frecuencia i-ésima.

x_i es la marca (o intervalo) de clase i-ésima.

n es el número total de datos.

Se tiene que: $\bar{x} = \frac{2002}{100} = 20.02\,años$

La mediana se obtiene con $Me = L_I + \frac{\frac{n}{2} - fa}{f_i} \times c$

Donde:

L_i es el límite real inferior de la clase o intervalo mediano.

n es el número total de datos (tamaño de la muestra).

f_a es la frecuencia acumulada anterior a la del intervalo mediano.

f_i es la frecuencia absoluta del intervalo mediano.

C es el tamaño o amplitud del intervalo mediano

La clase mediana es el intervalo de clase donde se encuentra el $\left(\frac{n}{2}\right)^{avo}$ dato, siendo en este caso el intervalo de clase donde está el $\left(\frac{100}{2}\right)^{avo} = 50^{avo}$ dato, es decir, el cuarto intervalo, donde $c = 22.5 - 18.5 = 4\,años$, así la mediana es:

$$Me = 18.5 + \frac{\frac{100}{2} - 39}{31} \times 4 = 18.5 + 1.4193 = 19.9193\,años$$

La moda se obtiene con $Mo = L_I + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times c$

Donde:

L_i es el límite real inferior de la clase o intervalo modal.

Δ_1 es la diferencia de frecuencias entre la clase o intervalo modal y el anterior.

Δ_2 es la diferencia de frecuencias entre la clase o intervalo modal y la siguiente.

C es el tamaño o amplitud del intervalo modal.

La clase modal es el (los) intervalo(s) de clase de mayor frecuencia, siendo en este caso el cuarto intervalo, de donde se obtiene que $\Delta_1 = 31 - 23 = 8$ y $\Delta_2 = 31 - 16 = 15$ y como $c = 22.5 - 18.5 = 4\ años$, así la moda es:

$$Mo = 18.5 + \frac{8}{8+15} \times 4 = 18.5 + 1.3913 = 19.8913\ años$$

La Dispersión, Variabilidad o Variación

Con respecto a la variabilidad, Behar (2009) menciona que la variación es una realidad observable. Está en todas partes y en todas las cosas. La variabilidad afecta todos los aspectos de la vida.

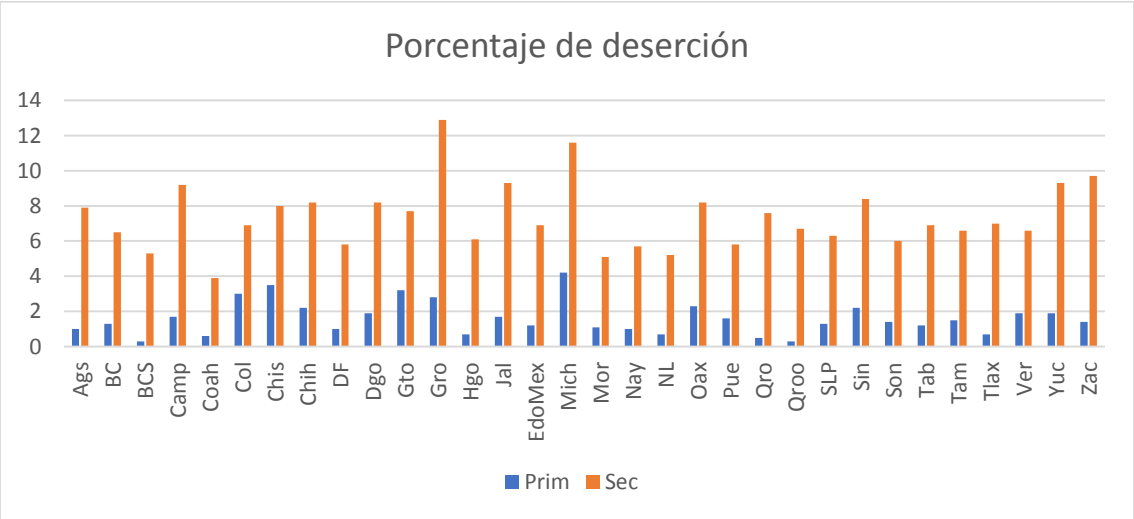
La variabilidad es una característica inherente a los datos y la Estadística proporciona los métodos para cuantificarla. La reflexión sobre la omnipresencia de la variabilidad permite comprender que aun cuando las muestras seleccionadas sean de la misma población y del mismo tamaño, las medias o las proporciones obtenidas en cada muestra son diferentes.

Por ejemplo, en la siguiente información referida a porcentajes de deserción en los estados del país, lo relevante es observar la variación y preguntarse, digamos, ¿en cuál nivel educativo existe mayor variación?, ¿qué relación hay entre la deserción de los niveles educativos?

Entidad	Primaria	Secundaria
Ags	1	7.9
BC	1.3	6.5
BCS	0.3	5.3
Camp	1.7	9.2
Coah	0.6	3.9
Col	3	6.9
Chis	3.5	8
Chih	2.2	8.2
DF	1	5.8
Dgo	1.9	8.2
Gto	3.2	7.7
Gro	2.8	12.9
Hgo	0.7	6.1
Jal	1.7	9.3
EdoMex	1.2	6.9
Mich	4.2	11.6
Mor	1.1	5.1
Nay	1	5.7
NL	0.7	5.2
Oax	2.3	8.2
Pue	1.6	5.8
Qro	0.5	7.6

QRoo	0.3	6.7
SLP	1.3	6.3
Sin	2.2	8.4
Son	1.4	6
Tab	1.2	6.9
Tam	1.5	6.6
Tlax	0.7	7
Ver	1.9	6.6
Yuc	1.9	9.3
Zac	1.4	9.7

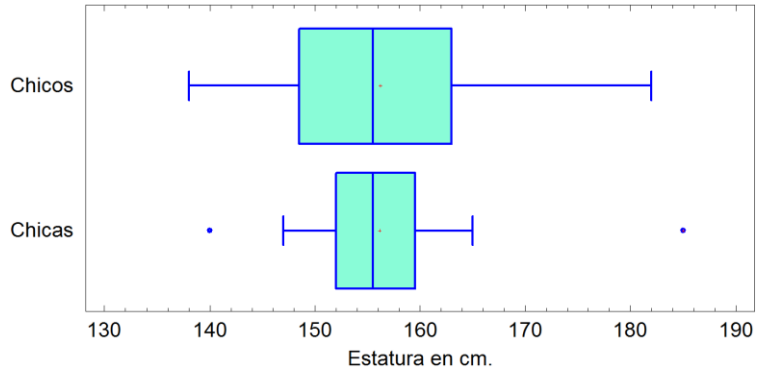
Fuente: Estadísticas Básicas del Sistema Educativo Nacional.



En la gráfica se observa que para el nivel primaria el promedio es aproximadamente de 1.5 %, y que hay poca variación con respecto a la deserción. Para el nivel secundaria el promedio es aproximado a 7%, pero la variación es mayor en este nivel, lo que indica que hay estados con poca deserción con aproximadamente 4% (Coahuila) y estados con mucha deserción como Guerrero con aproximadamente 13%.

El análisis descriptivo de los datos no puede restringirse exclusivamente al cálculo de las medidas de tendencia central porque, puede ocurrir que, dos distribuciones de frecuencias con igual media o con igual mediana pueden tener diferentes gráficas, es decir, si solamente se consideran las medidas de tendencia central, se pueden obtener conclusiones erróneas al no tomar en cuenta la dispersión (variación o variabilidad) de los datos.

En el siguiente diagrama de caja, por ejemplo, se observa que la estatura media de chicos y chicas es igual a 155 cm, sin embargo las estaturas de las chicas se alejan menos de la media, en contraparte, las estaturas de los chicos se alejan más de la media. De acuerdo con esto hay mayor variabilidad en la estatura de los chicos (Batanero, C. et al, 2015)



Otro caso es: Roberto y María forman una pareja con una estatura media de 1.70 m y Ana y Luis también son pareja con una estatura promedio de 1.70 m. Si solamente conocemos esta medida de centralización, nos inclinaríamos a pensar que los 4 tienen una estatura muy parecida. Sin embargo si aparte del promedio nos dicen que la desviación media de Roberto y María es de 0.01 m y que la desviación media de Ana y Luis es de 0.25 m, entonces llegaríamos a la conclusión de que Ana y Luis forman una pareja muy “dispareja”.

Las medidas de dispersión (variación o variabilidad) indican, en promedio, cuánto se alejan los datos de la media aritmética. Si los datos se alejan poco de la media entonces su dispersión o variación es menor que si alejan mucho de la media.

Según esta idea, si se aplica un examen de conocimientos a un grupo de alumnos, podemos pensar en tres situaciones:

Primera: si el examen es fácil, entonces se esperaría una distribución de frecuencias (número de alumnos) asimétrica y sesgada a hacia las calificaciones altas y mayor variación.

Segunda: si el examen es difícil, la distribución esperada sería asimétrica y sesgada hacia las calificaciones bajas y más variación.

Tercera: si el examen es rutinario, entonces se esperaría una distribución de frecuencias aproximadamente simétrica, con calificaciones cercanas a la media y poca variación.

Medidas de variación, dispersión o variabilidad

Las medidas de variación más comúnmente utilizadas son el rango, la varianza o la desviación estándar y el coeficiente de variación que mide la dispersión relativa.

La varianza muestral (s^2) se define como la suma de los cuadrados de las diferencias de los datos con respecto a la media, dividida entre el total de datos

menos uno. Esta medida tiene el inconveniente de que transforma las unidades de medición en cuadrados, por lo que no se puede comparar con la media aritmética. Por esta razón se define la desviación estándar (s) como la raíz cuadrada de la varianza.

El coeficiente de variación se utiliza cuando se desea comparar dos distribuciones de frecuencia que tienen diferente unidad de medida, se calcula dividiendo la desviación estándar entre la media.

El Rango (R)

El rango es la más simple de las medidas de dispersión y se define como la diferencia entre la medida mayor y la menor, pero no informa cuántos valores abarcan los datos.

El rango es muy utilizado en procesos industriales. En mucho, su utilidad en este campo se debe a lo sencillo y rápido que es calcularlo. El rango provee información útil cuando la muestra es pequeña, cuando la muestra es grande, no resulta una medida adecuada.

La Varianza (S^2) y la Desviación Estándar (S)

La desviación estándar es la medida de dispersión más utilizada para medir la variabilidad en una muestra (o si fuera el caso en una población). Para calcularla, primero se obtiene la varianza y después se extrae su raíz cuadrada.

Por ejemplo, los datos siguientes representan el contenido de azúcar (en g/100 ml) y el contenido de cafeína (mg/100 ml) de 8 refrescos de cola. En la cuarta y quinta columna de la tabla se ilustra el procedimiento para calcular la suma de los cuadrados de las diferencias utilizados para el cálculo de la varianza del contenido de azúcar (Revista del Consumidor, Profeco. 2003).

Marca	Azúcar (g/100 ml)	Cafeína (mg/100 ml)	Xi-Media	$(Xi\text{-}Media)^2$
Big Cola	10.9	12	10.9-10.6875	0.04515625
CM	10.3	6	10.3-10.6875	0.15015625
Coca Cola	10.6	15	10.6-10.6875	0.00765625
Great Value	10.2	5	10.2-10.6875	0.23765625
Hola Cola	10.4	13	10.4-10.6875	0.08265625
Pepsi Cola	11.1	14	11.1-10. 6875	0.17015625
Pepsi Limón	11	16	11-10.6875	0.09765625
Royal Cola	11	12	11-10.6875	0.09765625
				0.88875

La varianza es $s^2 = \frac{0.88875}{7}$ =0.12696, para conocer la desviación estándar se extrae la raíz cuadrada de la varianza: s = $\sqrt{0.12696}$=0.3563

La varianza muestral se calcula con la siguiente fórmula: $S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{x})^2}{n-1}$

Si se calcula la varianza de la población (σ^2), el denominador de la fórmula anterior se cambia por N (tamaño de la población).

Si la muestra es de datos agrupados, la varianza se calcula con:

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

Donde: f_i es la frecuencia i-ésima y x_i es la marca de clase i-ésima.

También se puede calcular la varianza para datos agrupados de una muestra y consecuentemente la desviación típica o estándar, mediante:

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

Otro ejemplo de la importancia de la cuantificar la variación es proporcionado por Flores y Díaz (2013), que presentan los resultados de PISA (2012) mencionando que es importante considerar dos aspectos. El primero, que el promedio OCDE es la media de sus 34 países o economías, los cuales se ponderaron por igual, a fin de evitar que dicho valor esté inclinado hacia los países con mayor población escolar de 15 años. En el ciclo 2012 el promedio de la OCDE en Matemáticas fue de 494 puntos, con una desviación estándar de 92.

El segundo aspecto es que para el cálculo del promedio de AL se adoptó la misma metodología que para el promedio OCDE, otorgando igual peso a los ocho países latinoamericanos (Argentina, Brasil, Chile, Colombia, Costa Rica, México, Perú y Uruguay) que participaron en PISA 2012. La media de desempeño en Matemáticas para los países latinoamericanos fue de 397 puntos y la desviación estándar de 80.

Si solamente se proporcionara la media se puede concluir que la diferencia entre las medias de los países que integran la OCDE y AL es bastante grande: 494-397=97 puntos, pero el hecho de conocer la desviación estándar (DE) nos indica que en los países de la OCDE hay mayor variación en la calidad educativa de sus escuelas, ya que la DE=92 puntos. En los países de AL esta información nos indica que existe menor variación en la calidad de la educación.

Actividades

a) Si se tienen dos muestras de estudiantes con pesos promedio de 68 kg y de 85 kg respectivamente y con la misma desviación estándar, ¿qué se puede decir con respecto a la variabilidad de cada muestra?

b) En un estudio se encontró que el gasto promedio anual para atención médica de dos muestras de familias de clase media fue el mismo con una desviación típica de $700.00 para la primera muestra y de $450.00 para la segunda muestra, entonces, ¿qué se puede decir con respecto a la variabilidad de cada muestr

c) La siguiente información corresponde a datos sobre carreras profesionales por área. Encuesta Nacional de Ocupación y Empleo. Segundo Trimestre de 2017. STPS-INEGI

Educación

carrera	Profesionistas ocupados	Hombres (%)	Mujeres (%)	Ingreso mensual promedio ($)
Ciencias de la educación, programas multidisciplinarios o generales	212,919	34.8		$ 12,696
Didáctica, pedagogía y currículo	156,559	25.4		$ 11,974
Formación docente para educación básica, nivel preescolar	165,811	2.9		$ 9,336
Formación docente para educación básica, nivel primaria	348,433	34.6		$ 11,974
Formación docente para educación básica, nivel secundaria	79,915	43.9		$ 13,679
Formación docente para educación de nivel medio superior	6,613	44.6		$ 8,369
Formación docente para educación física, artística o tecnológica	88,607	74.9		$ 13,184

Formación docente para la enseñanza de asignaturas específicas	31,928	50.0		$ 18,887
Formación docente para otros servicios educativos	42,671	11.9		$ 11,261
Formación docente, programas multidisciplinarios o generales	29,892	27.1		9,719
Orientación y asesoría educativa	22,805	24.5		$ 10,887

Ciencias Biológicas

carrera	Profesionistas ocupados	Hombres (%)	Mujeres (%)	Ingreso mensual promedio ($)
Biología y bioquímica	111,585		52.5	$ 16,713
Ciencias ambientales	11,162		40.2	$ 12,309
Diagnóstico médico y tecnología del tratamiento	12,648		37.5	$ 16,096
Enfermería y cuidados	238,929		84.8	$ 12,633
Estomatología y odontología	123,511		56.5	$ 16,597
Farmacia	15,339		55.9	$ 5,928
Medicina	245,952		43.4	$ 19,987
Psicología	248,513		71.2	$ 18,934
Química	11,401		62.7	$ 24,596
Terapia y rehabilitación	64,961		69.9	$ 14,719
Veterinaria	67,186		18.9	$ 17,673

Comparar ambos grupos de carreras y concluir.

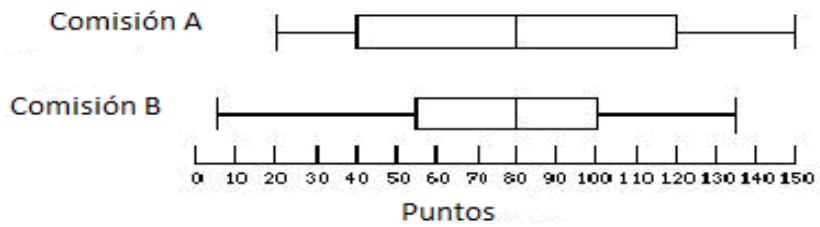
Medidas de posición

En ocasiones es necesario conocer proporciones de una población que cumple ciertos valores de la variable de interés lo cual no es posible describir fácilmente si solo se tienen las medidas de centralización y dispersión, por lo que es necesario determinar algunas medidas descriptivas mediante las cuales se pueda hacer esa descripción, estas medidas son llamadas de posición las cuales permiten determinar los valores de la variable que divide al conjunto de datos en partes iguales, tales medidas se llaman genéricamente cuantiles y de acuerdo al número de partes en que dividen a la población, así tenemos:

Mediana. Valor de la variable que divide en dos partes al conjunto de datos, que si bien es una medida de centralización también es de posición y su valor corresponde al central de una distribución de datos y describe que a su izquierda se encuentra el 50% de la distribución y a su derecha el otro 50%.

Cuartiles. Valores de la variable que dividen a la distribución en cuatro partes iguales y que describen que alrededor de cada uno de ellos se encuentra el 50% de la distribución, así el cuartil primero (Q_1) describe que a su izquierda se encuentra el 25% de la población con los valores más bajos de la variable y a su derecha otro 25% de los datos, el cuartil segundo (Q_2) que es equivalente a la mediana, describe que alrededor de él se encuentra el 50% de los datos con los valores centrales de la variable 25% a su izquierda y 25% a su derecha y el cuartil tercero (Q_3) describe que a su izquierda se encuentra el 25% de los valores centrales de la distribución y a su derecha el 25% de los datos con los valores mayores de la variable.

Para ejemplificar la interpretación de los cuartiles como medidas de posición analizaremos los siguientes diagramas de caja que muestran los puntajes obtenidos en un examen de un curso de Matemáticas por los alumnos de dos grupos (Rodríguez, 2012).



Con respecto al grupo revisado por la comisión A: el 25% de los alumnos obtuvo 40 puntos o menos, el 50% obtuvo 80 puntos o menos, el 75% de los alumnos obtuvo 120 puntos o menos, el 50% de los alumnos obtuvo entre 40 y 120 puntos. La variación entre los cuartiles 1 y 3 es: 120-40=80 puntos.

La interpretación del grupo revisado por la comisión B es semejante. Aquí es importante señalar que en este grupo existe menor variación con respecto a las calificaciones obtenidas por los examinados (100-55=45 puntos).

Medidas de posición para datos agrupados

Para efectuar el cálculo de las medidas de posición para datos agrupados se utiliza el mismo procedimiento de interpolación que para calcular la mediana (Q_2).

Ejemplo.

Los datos agrupados presentados en la tabla siguiente corresponden al gasto en pasajes por día (en $), de una muestra de 65 alumnos de una escuela secundaria.

Gasto ($)	Límites reales del intervalo	Punto medio del intervalo (X_j)	Número de alumnos (Frecuencia)	$(X_i)(F_i)$
9 - 15	8.50 - 15.50	12.00	7	(12)(7)=84
16 - 22	15.50 - 22.50	19.00	10	(19)(10)=190
23 - 29	22.50 - 29.50	26.00	25	(26)(25)=650
30 - 36	29.50 - 36.50	33.00	13	(33)(13)=429
37 - 43	36.50 – 43.50	40.00	10	(40)(10)=400

El cuartil k-ésimo se obtiene con: $Q_k = LI + (\frac{\frac{kn}{4} - f_a}{f_i})C$

Donde *LI* s el limite real inferior del intervalo donde está el k-ésimo cuartil (con k=1, 2 o 3).

Para el primer cuartil se tiene k=1, para saber en cual intervalo se encuentra, se sustituye k=1 y n=65: $\frac{(1)(65)}{4} = 16.25$ Este valor se ubica en el segundo intervalo.

$$Q_1 = 15.50 + \left(\frac{16.25 - 7}{10}\right)(7) = 15.50 + \left(\frac{9.25}{10}\right)(7) = 15.50 + 6.475 = 21.975$$

La interpretación es: el 25% de los alumnos gastan aproximadamente $22 o menos en transporte para llegar a la escuela.

El Coeficiente de Variación (CV).

El coeficiente de variación mide la dispersión relativa y permite comparar dos conjuntos de datos expresados en diferentes unidades de medida. Se calcula comparando (por cociente) la desviación estándar con la media. El coeficiente de variación no tiene unidades, lo que permite expresarlo en forma de porcentaje:

$$CV = \frac{s}{\bar{x}}(100)$$

Ejemplo 1

Si deseamos comparar el contenido de azúcar con la cafeína, en los refrescos de cola (Datos Profeco, 2003), necesitamos calcular el CV de cada variable.

El coeficiente de variación para el contenido de azúcar en los refrescos es

$CV = \frac{0.3563 g/100 ml}{10.6875 g/100 ml} = 0.0333$, o de manera equivalente CV = 3.33%

El coeficiente de variación para el contenido de cafeína en los refrescos es
$CV = \frac{4.0333 mg/100 ml}{11.625 mg/100 ml} = 0.3469$, o sea CV = 34.63%

A partir de la comparación de los valores anteriores, se concluye que existe menos dispersión o variabilidad en el contenido de azúcar de los refrescos (el contenido de cafeína es aproximadamente 10 veces mayor que el de azúcar).

Ejemplo 2

Después de registrar los datos correspondientes al peso y la estatura de 40 hombres, se colocaron en la siguiente tabla

	Media	Desviación estándar
Peso	172.55 libras	26.33 libras
Estatura	68.34 pulgadas	3.02 pulgadas

Calcular el coeficiente de variación de las estaturas, después el coeficiente de variación de los pesos; finalmente, comparar ambos resultados.

Solución.

Debido a que tenemos estadísticos muestrales, los dos coeficientes de variación se obtienen de la siguiente manera:

Estatura $CV = \frac{3.02 pul}{68.34 pul}(100\%) = 4.42\%$

Pesos $CV = \frac{26.33 libras}{172.55 libras}(100\%) = 15.26\%$

Se observa que las estaturas (con CV = 4.42%) tienen una variación considerablemente menor que los pesos con (CV = 15.26%). Lo anterior tiene sentido, ya que, por lo general, vemos que los pesos de los hombres varían mucho más que sus estaturas. Por ejemplo, es muy raro encontrar un adulto

que mida el doble que otro, pero es mucho más común ver a uno que pese el doble que otro.

Regla empírica: relación entre la media y la desviación estándar

En un gran número de estudios estadísticos, el uso conjunto de la media y la desviación estándar, permite conocer la distribución porcentual de una población.

Esta situación se verifica de manera general en distribuciones de datos con una sola moda y bastante simétricas (con forma de campana). El criterio es el siguiente:

- En el intervalo $(\bar{x}-s, \bar{x}+s)$ se encuentra aproximadamente el 68% de los datos.
- En el intervalo $(\bar{x}-2s, \bar{x}+2s)$ se encuentra aproximadamente el 95% de los datos.
- En el intervalo $(\bar{x}-3s, \bar{x}+3s)$ se encuentra aproximadamente el 99% de los datos.

Ejemplo 1

Los datos agrupados presentados en la tabla siguiente corresponden al gasto en pasajes por día (en $), de una muestra de 65 alumnos de una escuela secundaria.

Gasto ($)	Límites	Punto medio del intervalo (X_j)	Número de alumnos (f)
9 - 15	8.50 - 15.50	12.00	7
16 - 22	15.50 - 22.50	19.00	10
23 - 29	22.50 - 29.50	26.00	25
30 - 36	29.50 - 36.50	33.00	13
37 - 43	36.50 – 43.50	40.00	10

La distribución de probabilidad es aproximadamente normal con media de 26 97 pesos y desviación estándar de 8.21 pesos.

La aplicación de la Regla Empírica porporciona los siguientes resultados

Intervalos de la Regla Empírica	Proporción o porcentaje aproximado de datos	Número aproximado de alumnos en la población
(26.97-8.21,26.97+8.21) (18.76,35.18)	0.68 o 68%	(0.68)(500)=340

(26.97-(2)(8.21),26.97+(2)(8.21)) (10.55,43.49)	0.95 o 95%	(0.95)(500)=570
(26.97-(3)(8.21),26.97+(3)(8.21)) (2.34,51.6)	1 o 100%	500

Si consideramos que la muestra de los 65 alumnos es representativa de la población de alumnos de la escuela (N=500), entonces, aproximadamente, 340 alumnos gastan entre $18.76 y $35.18, 570 alumno gastan entre $10.55 y $43.49.

Ejemplo 2

Flores y Díaz (2013) presentan los resultados de PISA (2012) señalando que en el cálculo del promedio de América Latina se adoptó la misma metodología que para el promedio OCDE, otorgando igual peso a los ocho países latinoamericanos (Argentina, Brasil, Chile, Colombia, Costa Rica, México, Perú y Uruguay) que participaron en PISA 2012. La media de desempeño en Matemáticas para los países latinoamericanos fue de 397 puntos y la desviación estándar (DE) de 80. Para los estudiantes mexicanos que participaron (33806), la media fue de 413 puntos.

La población de estudio es de 1472875 alumnos de 15 años, inscritos en séptimo grado (primero de secundaria) o superior. Los estudiantes elegibles nacieron entre el 1 de enero y el 31 de diciembre de 1996.

Si consideramos la media de los estudiantes mexicanos (413 puntos) y DE=80 puntos, entonces en el intervalo, (413-80,413+80)=(333,493) se puede estimar que aproximadamente (0.68)(1472875)=1001555 alumnos están comprendidos entre 333 y 493 puntos en la evaluación de PISA.

Actividad

En la tabla siguiente se muestra el total de pensionados y/o jubilados (IMSS-ISSSTE) de acuerdo con la Encuesta Nacional de Empleo (1998).

Xi: Edad (años)	$P(X_i)$
[30, 40)	0.016
[40,50)	0.056
[50,60)	0.183
[60,70)	0.382
[70,80)	0.263
[80,90]	0.1

El tamaño de la población es de 955000 personas

Una muestra aleatoria de 1251 pensionados y/o jubilados se observa en la siguiente tabla. Argumente si se verifica o no la Regla Empírica.

Xi: Edad (años)	F_i
[30, 40)	20
[40,50)	175
[50,60)	435
[60,70)	446
[70,80)	160
[80,90]	15

Tipificación o estandarización de una variable

La comparación entre puntuaciones directas puede llevar a conclusiones engañosas, es por eso que en la práctica se suele transformar las puntuaciones observadas en otras que, sin perder o distorsionar la información contenida en las puntuaciones originales, permitan una comparación directa de las mismas.

La tipificación o estandarización de un variable es un procedimiento que permite la comparación entre:

- Los valores de dos distribuciones distintas.
- Los valores de variables con unidades distintas.
- Reconocimiento de valores atípicos o dentro de ciertos valores de acuerdo a la regla empírica.

El procedimiento desarrollado para ello permite transformar un valor de la variable X en valores típicos o tipificados.

La puntuación típica de una observación indica el número de desviaciones típicas que esa observación se separa de la media del grupo de observaciones. Su media es 0 y su varianza y su desviación típica son iguales a 1. Su fórmula es:

$$z = \frac{X - \bar{x}}{s}$$

Que también se puede escribir de la siguiente manera; X=(z)(s)+$\bar{x}$

Ejemplo 1

Dos profesores de Matemáticas I presentan los siguientes valores descriptivos de sus grupos. Entre paréntesis se colocan las equivalencias de la media (0) y de la desviación estándar (1) en unidades tipificadas (estandarizadas)

	Media	Desviación estándar
Profesor 1	7 (0)	1.41 (1)
Profesor 2	7.9 (0)	1.12 (1)

Otras equivalencias son:

Para el Primer Profesor

Calificación (X)	Unidades estandarizadas (z)
2.77	-3
4.18	-2
5.59	-1
6.41	1
7.82	2
9.23	3

Para el Segundo Profesor

Calificación (X)	Unidades estandarizadas (z)
4.54	-3
5.66	-2
6.78	-1
9.02	1
10.14	2
11.26	3

Podemos también decir que una calificación de 8 con el Primer Profesor equivale a $z=\frac{8-7}{1.41}=0.709$

Una puntuación de z=0.709 equivale a: x=(0.709)(1.12)+7.9=8.69 con el Segundo Profesor.

Conclusión: asumiendo condiciones iguales en el proceso de enseñanza y aprendizaje en la asignatura de Matemáticas I, para aprobar con el Segundo Profesor es necesario ser más consistente en el estudio de esta asignatura.

Ejemplo 2

Se ha efectuado un examen de comportamiento agresivo a un grupo de adolescentes. El examen constaba de dos pruebas llamadas A y B.

	A	B
Media	15.5	75
Desviación estándar	2.5	30.6

Dos adolescentes denominados 1 y 2, han obtenido como resultado de cada prueba:

	A	B
Calificación de 1	16.7	14
Calificación de 2	77.5	82.4

¿Cuál tiene más agresividad?

Prueba A. Adolescente 1

$z = \frac{16.7-15.5}{2.5}$=0.48

Prueba B. Adolescente 1

$z = \frac{14 - 15.5}{2.5} = -0.6$

Prueba A. Adolescente 2

$z = \frac{77.5-77}{30.6}$=0.016

Prueba B. Adolescente 2

$z = \frac{77.5-82.4}{30.6}$=-0.16

Conclusión: en ambas pruebas el adolescente 1 es más agresivo.

El Teorema de Tchevichev

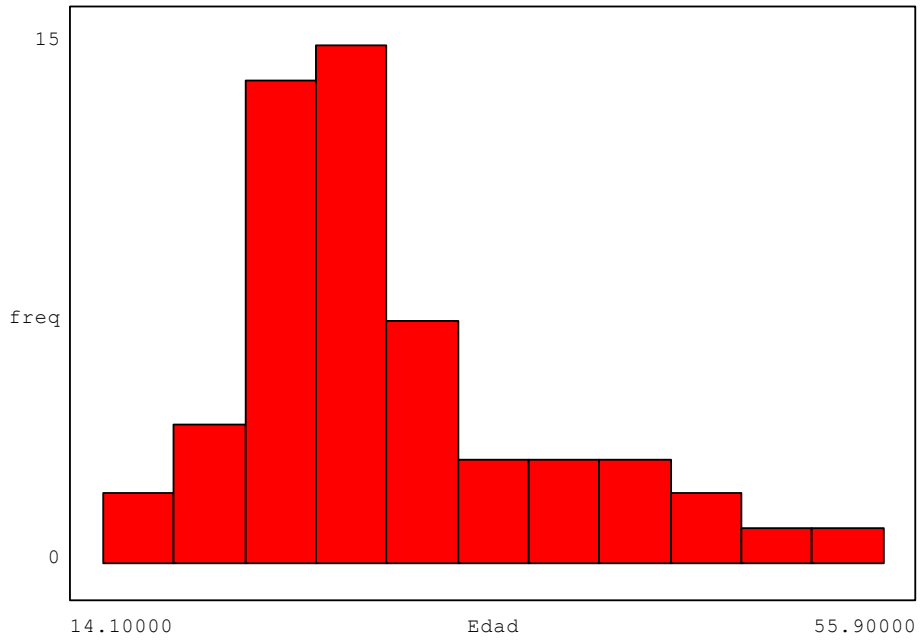
Si los datos de una muestra no tienen forma de campana no es posible aplicar la regla empírica. Para tener un idea de la distribución de los datos en la población se utiliza el Teorema de Tchevichev:

El intervalo $(\bar{x} - ks, \bar{x} + ks)$ contiene al menos el $(1 - \frac{1}{1^2})$(100%) de los datos de la población.

Específicamente:

- El intervalo $(\bar{x} - 2s, \bar{x} + 2s)$ contiene al menos el 75% de los datos de la población.
- El intervalo $(\bar{x} - 3s, \bar{x} + 3s)$ contiene al menos el 89% de los datos de la población.

Por ejemplo, el histograma siguiente muestra la edad (en años) de 55 personas. La media de la edad es 29.5 años y la desviación estándar es 9 años.



Puesto que la distribución de la edad es asimétrica, no es aplicable la regla empírica, lo adecuado es utilizar el teorema de Tchevichev:

El intervalo [29.5-(2)(9), 29.5+(2)(9)]=(11.5, 47.5). Si suponemos que el tamaño de la población es N=400 personas, entonces por este teorema, (0.75)(400)=300 personas tienen edades entre 11.5 años y 47.5 años.

El intervalo [29.5-(3)(9), 29.5+(3)(9)]=(2.5, 56.5), entonces podemos decir que (0.89)(400)=356 personas tienen al menos entre 2.5 años y 56.5 años.

Una reflexión sobre las medidas de tendencia central y las medidas de variabilidad.

Para finalizar el estudio de las medidas de tendencia central y de las medidas de dispersión, se presentan algunos comentarios y reflexiones acerca de estas medidas descriptivas.

Cabe señalar que lo más importante no es el obtenerlas, sino el interpretarlas. Por ejemplo, supongamos que un docente llega a una escuela y le piden que tienda dos de tres grupos vacantes, le dicen que el promedio de los tres grupos es prácticamente el mismo, digamos, 7.9, 8 y 8.2. Si no se le propociona otra información, el docente puede decir que le den dos cualesquiera de los grupos. Pero si se le proporcionan una medida de la variación, por ejemplo, la desviación estándar como 0.5, 1 y 2.5, entonces elegiría aquellos grupos que tienen menor variabilidad, ya que el grupo con mayor desviación estándar (2.5), presenta casos de alumnos con bajo rendimiento y alumnos con rendimiento alto.

Otra situación de interpretación de la media es: si el promedio del número de hijos en 10 familias es de 2.5 hijos, entonces se puede estimar que por cada 100 familias el número de hijos es de 250.

III. DISTRIBUCIÓN DE PROBABILIDAD DE UNA VARIABLE

Tema 1. Distribución de probabilidad para una variable discreta.

Comunmente la distribución de probabilidad para una variable discreta se representa por medio de una gráfica de barras.

Por ejemplo para la variable X: Número de hijos de 305 estudiantes, la probabilidad empírica se puede estimar con las frecuencias relativas,

X: Número de hijos	Número de estudiantes (Frecuencia)	Frecuencia Relativa (Probabilidad empírica)
0	56	$\frac{56}{305} = 0.1836$
1	53	0.1737
2	83	0.2721
3	62	0.2032
4	30	0.0983
5	15	0.0491
6	6	0.0196
Suma	305	0.9996

Utilizando las probabilidades empíricas (porque se calculan en una muestra), podemos contestar las siguientes preguntas (seleccionando un estudiante al azar).

a) ¿cuál es la probabilidad de que tenga 0 o 1 hijos?

La probabilidad es 0.1737+ 0.2721=0.3573

b) La probabilida de que tenga 2 o 3 hijos es: 0.2721+ 0.2032= 0.4753

c) Es menos probable que tenga 4 o más hijos: 0.0983+0.0491+0.0196=0.1676

De manera general podemos decir que la distribución de probabilidad es asimétrica y que es más probable que un estudiante tenga 2 hijos y menos probable que tenga 6 hijos.

Ejercicio

Sea *X* una variable aleatoria que expresa el número de personas que habitan en una vivienda elegida al azar. La distribución de probabilidad de *X* es la siguiente:

X	1	2	3	4	5	6	7	8
P(X)	0.23	0.322	0.177	$P(x_4)$	0.067	0.024	0.015	0.010

a) Encuentre $P(x_4)$.

b) Representar gráficamente la distribución de probabilidad.

c) Hallar la probabilidad de que el número de personas que viven en un hogar sea menor o igual que cuatro.

d) Calcular la probabilidad de que al menos dos personas vivan en una vivienda.

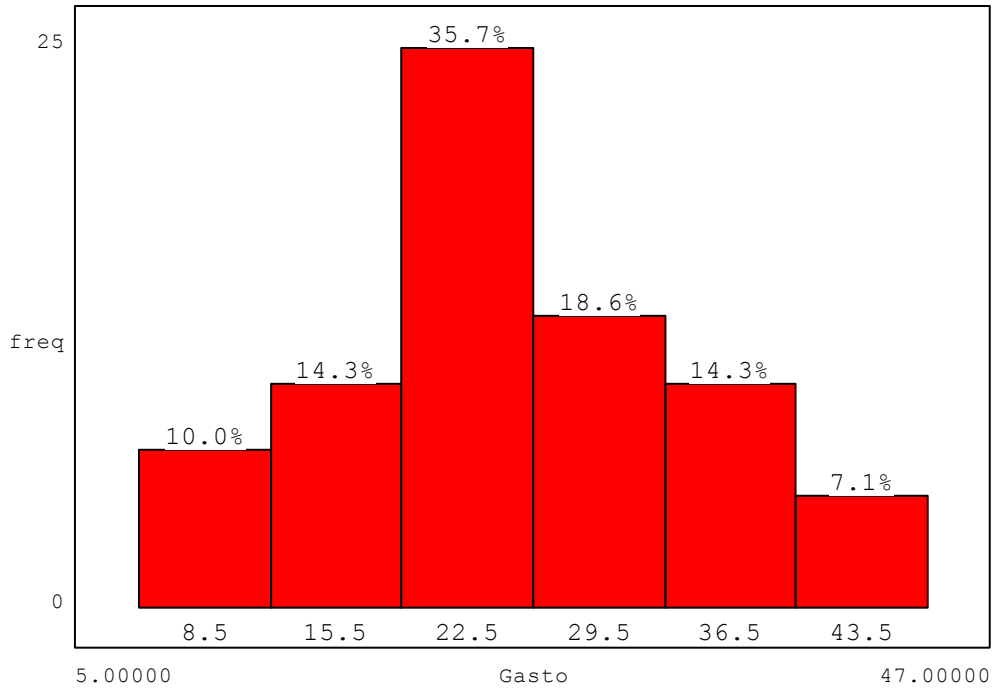
e) Obtener el número medio de personas que habitan en una vivienda.

Tema 2. Distribución de probabilidad para una variable continua.

La tabla siguiente muestra la distribución de frecuencias de 70 estudiantes de una escuela secundaria con respecto a la variable Gasto ($) en transporte para llegar a la escuela.

(X): Gasto ($)	No. de Estudiantes	Probabilidad
5 a12	7	0.1
12 a 19	10	0.14286
19 a 26	25	0.35714
26 a 33	13	0.18571
33 a 40	10	0.14286
40 a 47	5	0.07143

El histograma correspondiente se muestra a continuación



Si consideramos que las áreas de las barras son probabilidades empíricas (porque se calculan en una muestra) proporcionales (o iguales a las frecuencias relativas, entonces es posible determinar algunas probabilidades de que ocurran los valores de la variable en un cierto intervalo, algunos ejemplos son:

La probabilidad de que un estudiante gaste $19 o menos, es decir, P(x≤19)=0.1+0.143= 0.243, o de manera equivalente,10%+14.3%=24.3%

La probabilidad de que un estudiante gaste $29.5 o menos es, P(x≤29.5)=0.1+0.143+0.357+(29.526)($\frac{0.186}{7}$)=0.6+(3.5)(0.026571)=0.6+0.093= =0.693

La probabilidad de que un estudiante gaste $29.5 o más es (1-0.693)=0.307

La probabilidad de que un estudiante gaste $35 o más es: P(x≥35)= =(40-35)($\frac{0.143}{7}$)(0.071)=(5)(0.020429)+0.071=0.102145+0.071=0.173145

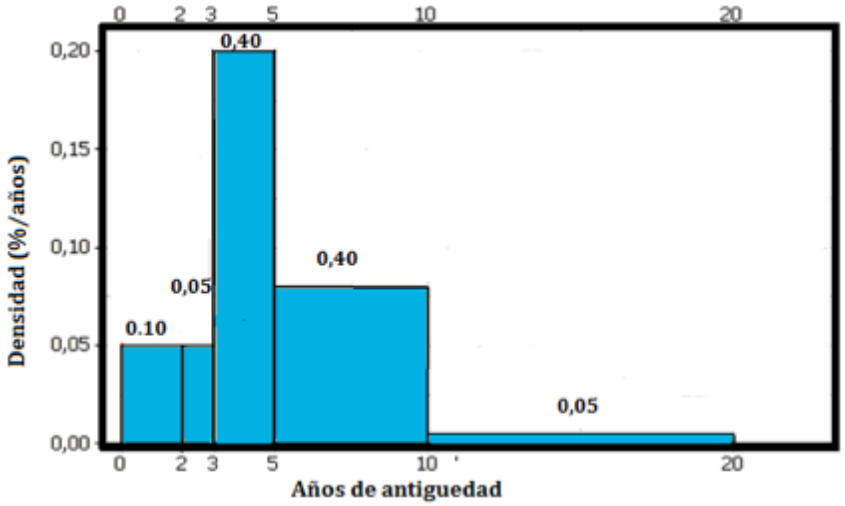
La probabilidad de que un estudiante gaste $25 o más: P(x≥25)=(26-25)($\frac{0.357}{7}$)+0.186+0.143+0.071=(1)(0.051)+0.4=0.451

La probabilidd de que un alumno gaste entre $15.45 y $34.35: P(15.45≤x≤34.35)=(19-15.45)($\frac{0.143}{7}$)+0.357+0.186+(34.35-33)($\frac{0.143}{7}$) = =(3.55)(0.020428)+0.543+(1.5)(0.020428)=0.0725+0.543+0.03064=0.646

Behar y Grima (2013) proponen el uso de intervalos de anchura desigual, considerando que en cada barra o rectángulo su área es proporcional (o igual) a la frecuencia relativa.

En un sector educativo se toma una muestra al azar de 500 docentes y se determina la antigüedad en este sector. Por razones de índole administrativo se representan los datos por medio de un histograma que considera los siguientes intervalos: 0.2 años, 2-3 años, 3-5 años, 5- 10 años y 10-20 años.
(Adaptado de Behar y Grima, 2013).

Años de Antigüedad	Frecuencia Relativa o Probabilidad Empírica
[0,2)	0.1
[2,3)	0.05
[3,5)	0.40
[5,10)	0.40
[10,20]	0.05



Ejemplos

Calcular la probabilidad de que al seleccionar un docente tenga entre 4 y 7.5 años en servicio

P(4≤x≤7.5)=(5-4)($\frac{0.4}{2}$)+(7.5-5)($\frac{0.4}{5}$)=0.2+0.2=0.4

Calcular la probabilidad de que al seleccionar un docente tenga entre 5 y 15 años

P(5≤x≤15)=0.4+(15-10)($\frac{0.05}{10}$) =0.4+(5)(0.005(=0.4+0.025=0.425

La Distribución de Probabilidad Normal

Existen algunas razones fundamentales por las cuales la distribución normal ocupa un lugar tan importante en la Estadística y la Probabilidad:

a) Tiene algunas propiedades que la hacen aplicable a muchas situaciones donde es preciso hacer inferencias al seleccionar muestras de una población.
b) Esta distribución llega a coincidir muy bien en las distribuciones observadas de frecuencias de muchos fenómenos, entre ellos las cracterísticas humanas (peso, estatura, coeficiente intelectual,etc.), en la producción de procesoso físicos (dimensiones, rendimientos), y en otras características de interés, tanto en el sector público como en el privado.

Parámetros de la distribución normal

Para definir una distribución normal de probabilidad, necesitamos conocer solamente dos parámetros: la media poblacional (μ) y la desviación estándar poblacional (σ).

A continuación se muestran tres poblaciones donde las variables consideradas tienen distinta distribución normal, descritas únicamente por la media y la desviación estándar, teniendo cada una de ellas una curva normal particular.

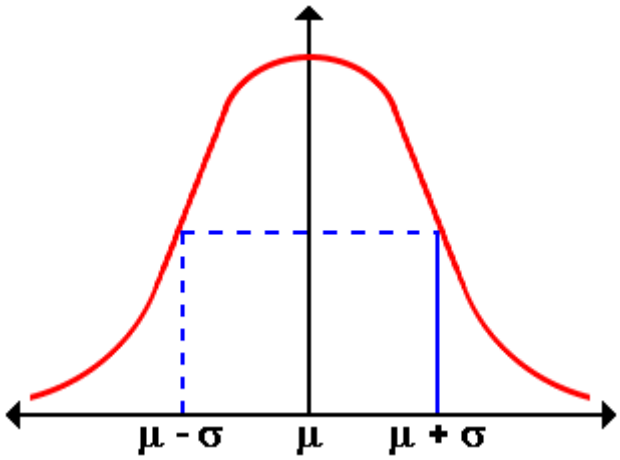
Variable: X	Media (μ)	Desv. Estándar (σ)
Calificación de una prueba estandarizada	500 puntos	100 puntos
Contaminación atmosférica en una ciudad.	2500 ppm	750 ppm
Peso de salmones pescados en un río.	15 Kg	3 Kg

Las distribuciones normales anteriores, muestran que la curva normal puede describir un amplio número de poblaciones, diferenciadas solamente por la media y la desviación estándar (o la varianza σ^2). Comunmente, esto se denota de la siguiente manera: N(μ,σ).

Áreas bajo la curva normal

No importa cuales sean los valores de μ y σ, para cualquier distribución normal de probabilidad, el área bajo la curva será de un total de una unidad cuadrada, por lo cual podemos considerar que las áreas bajo la curva normal, son probabilidades. En términos matemáticos, es verdad que cualquier distribución normal tiene las siguientes propiedades:

1. La curva tiene un valor máximo en la media (μ).
2. Es simétrica con respecto a la media.
3. El área (probabilidad) bajo la curva, correspondiente al intervalo $[\mu-\sigma,\mu+\sigma]$, es aproximadamente de 0.6826



4. El área bajo la curva, correspondiente al intervalo $[\mu-2\sigma, \mu+2\sigma]$, es aproximadamente de 0.9544
5. El área bajo bajo la curva, correspondiente al intervalo $[\mu-3\sigma, \mu+3\sigma]$, es aproximadamente 1.
6. El eje de las abscisas es una asíntota horizontal.

Puesto que la distribución normal se utiliza para modelar variables aleatorias continuas y éstas, toman valores, solamente en intervalos, resulta necesario tener una *unidad de medida* y un *punto de origen*.

El punto de origen es la media μ de la variable bajo estudio y la unidad de medida es la desviación estándar σ .

Cuando se usa el modelo normal, en el estudio de una variable numérica continua, generalmente se debe de responder a las siguientes preguntas:

a)¿Los valores de la variable están a la izquierda o a la derecha del origen μ ?
b)¿Qué tan alejados están los valores de la variable, del origen?

Cuando los valores de la variable continua se alejan (a la derecha o a la izquierda), 1, 2 o 3 unidades (σ) del origen (μ), se forman intervalos y la probabilidad de que la variable tome valores entre ambos extremos de un intervalo, se determinan usando las propiedades de la distribución normal.

La distribución normal estándar

Al utilizar la distribución normal estándar, debemos de considerar que:

a) El punto de origen es $\mu = 0$
b) La unidad de medida es $\sigma = 1$
c) La variable estandarizada se simboliza con *Z*

Cuando una variable X es normal (pero no estándar), sus valores se pueden transformar en unidades estandarizadas o tipicadas.
Esta transformación se puede lograr con el siguiente procedimiento

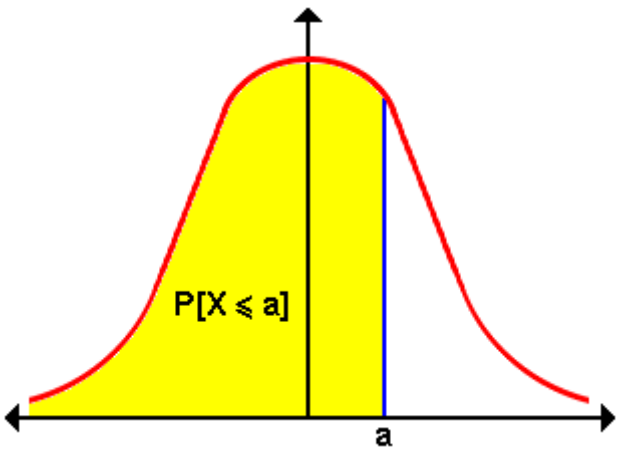
$$z = \frac{x - \mu}{\sigma}$$

El procedimiento anterior permite transformar cualquier valor de la variable X con distribución normal a valores estandarizados *Z*. Por ejemplo, si una variable con distribución normal tiene $\mu = 250$ y $\sigma = 75$, los valores X_1=200, X_2=375, X_3=460, se transforman, respectivamente en Z_1=*-0.66,* Z_2=*1.66 y* Z_3=*2.8*

Uso de las tablas de la distribución normal estándar

Después de estandarizar cualquier valor de la variable bajo estudio (X), se consultan las tablas de probabilidad para la distribución normal estándar: N(0,1). Estas tablas están organizadas en función de desviaciones estándar o unidades estandarizadas *z* que contienen valores de solamente la mitad del área (o probabilidad) bajo la curva normal.

Dado que, la distribución es simétrica, los valores de una mitad de la curva, también lo son para la otra mitad y de acuerdo con esto, podemos utilizar la tabla en situaciones que incluyan ambos lados de la curva normal para calcular la probabilidad de ocurrencia de una variable continua X en un intervalo determinado. Por ejemplo, para calcular la probabilidad indicada en la figura siguiente, se deben de sumar dos probabilidades: la probabilidad desde la cola izquierda de la curva hasta el eje de simetría (0.5), con la probabilidad desde el eje de simetria hasta el valor de X=**a**



Utilizando adecuadamente la tabla podemos calcular cualquier tipo de probabilidad bajo una curva normal estándar (0,1).

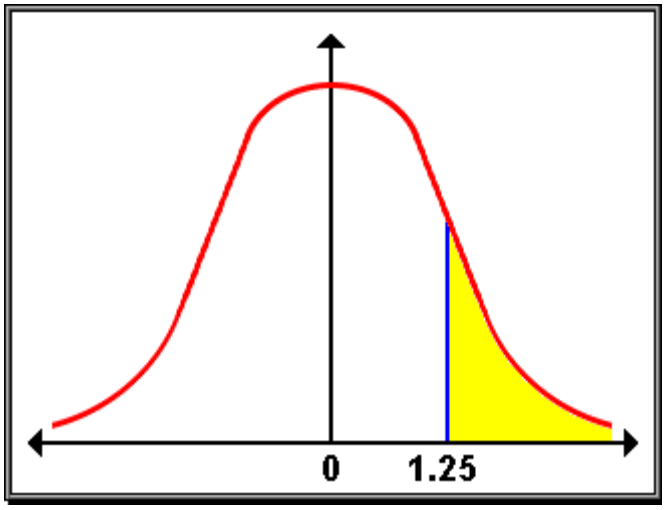
A continuación se proporciona una tabal de la Distribución Normal Estándar y algunos ejemplos con valores específicos para valores tipificados o estandarizados z.

Tabla de probabilidades de la Distribuciòn Normal Estándar

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990

Ejemplos

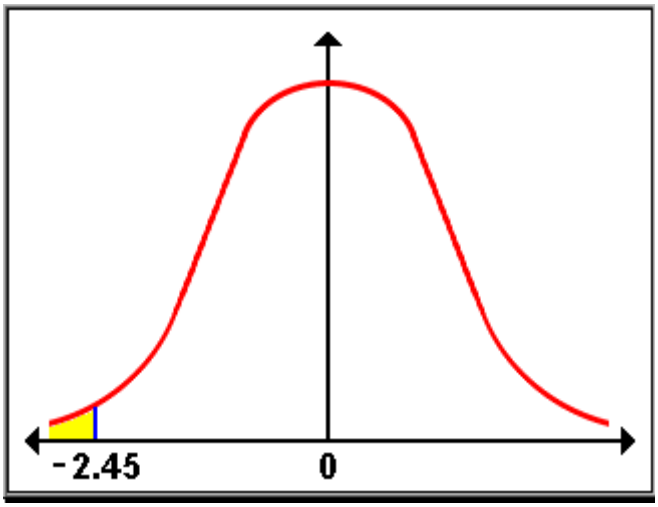
1. Calcular la probabilidad de que la variable *z* sea mayor o igual a 1.25: $P(z \geq 1.25)$



Buscamos en la tabla $P(0 \leq z \leq 1.25) = 0.3944$

Usando solamente la mitad de la curva: $P(z \geq 1.25) = 0.5 - 0.3944 = 0.1056$

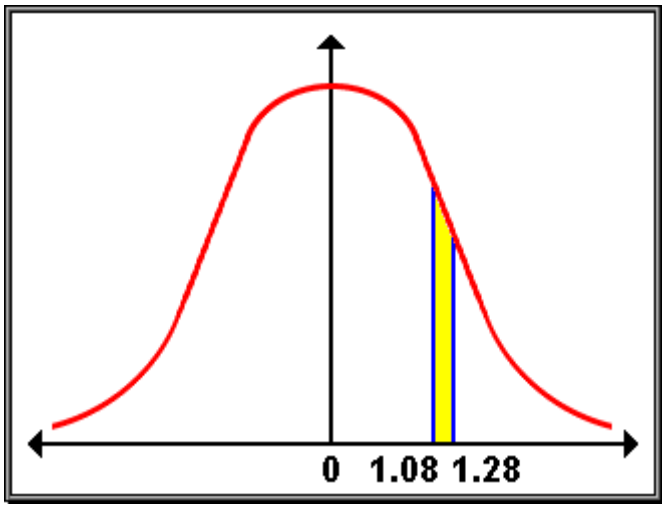
1. Calcular la probabilidad de que *z* sea menor o igual a -2.45: $P(z \leq -2.45)$



Los números negativos no vienen en la tabla, pero podemos aprovechar que la curva normal es simétrica:

$P(z \leq -2.45) = P(z \geq 2.45) = 0.5 - 0.4929 = 0.0071$

2. Calcular la probabilidad de que *z* se encuentre entre 1.08 y 1.28: $P(1.08 \leq z \leq 1.28)$



La probabilidad desde 0 hasta 1.28, es: 0.3997; la probabilidad desde 0 hasta 1.08, es 0.3599, por lo tanto

$$P(1.08 \leq z \leq 1.28) = 0.3997 - 0.3599 = 0.0398$$

TIpificación o estandarización de una variable normal

Consideremos el siguiente problema: Supongamos que los coeficientes de inteligencia (CI) están distribuidos aproximadamente de forma normal con media μ = 100 y desviación estándar σ = 10. Si una persona es seleccionada aleatoriamente ¿Cuál es la probabilidad de que su CI esté entre 100 y 115?

Calculemos primero los valores de Z para CI = 100 y CI = 115.

$Z_1 = \frac{100-100}{10} = 0$ y $Z_2 = \frac{115-100}{10} = 1.5$

Ahora calculemos el área entre 0 y 1.5. Para ello utilizaremos una tabla de valores de probabilidad normal estándar de valores positivos.

La probabilidad buscada de Z_1 = 0 a Z_2 = 1.5 es:

P (0 < Z ≤ 1.5) = 0.4332

Por lo tanto la probabilidad de que al seleccionar una persona de forma aleatoria, este presente en CI entre 100 y 115 es:

P (110 $\leq CI \leq 115$) = 0.4332

Como un segundo ejemplo, calcularemos la probabilidad de que una persona seleccionada al azar (aleatoriamente) tenga CI entre 90 y 105.

Calculamos los valores de Z

$Z_1 = \frac{90-100}{10} = -1$ $Z_2 = \frac{105-100}{10} = 0.5$

Como la gráfica es simétrica, los valores de Z negativos se calculan tomando Z positivo, con ayuda de la tabla, esto es:

P (0 < Z < 1) = 0.3413 y P (0< Z < 0.5) = 0.1915

por lo tanto la probabilidad de que la persona tenga un CI entre 90 y 105 es 0.5328.

Actividades

1. Un reporte de la SEP, publicado en La Gaceta de la UNAM (Nov. del 2000) señala que el sueldo promedio de un maestro (a) de primaria, en las escuelas oficiales, tiene una media de $4800 mensuales con una desviación estándar de $230. Si se supone una distribución normal para la variable en estudio,

    a) Definir la variable bajo estudio y su tipo.
    b) Calcular la probabilidad de que al seleccionar un profesor (a) gane entre $4300 y $4550 mensuales.
    c) Calcular la probabilidad de que al seleccionar un profesor (a) gane más de $4500 mensuales
    d) Si se considera una población de 15000 maestros, ¿cuántos de estos ganan menos de $4500 mensuales?
    e) Determine el valor del sueldo a partir del cual está el 5% de los maestros que más ganan.
    f) Determine el valor del sueldo a partir del cual está el 5% de los maestros que ganan menos.
    g) ¿Entre cuáles valores del sueldo (simétricos con respecto a la media) se encuentra el 90% de los maestros?

2. Una universidad lleva a cabo una prueba para seleccionar nuevos profesores. Por la experiencia de pruebas anteriores, se sabe que las puntuaciones siguen una distribución normal de media 80 y desviación típica (estándar) 25. En esta ocasión se han presentado 145 candidatos.

a) ¿Qué porcentaje de candidatos obtendrá entre 90 y 125 puntos?

b) ¿Qué porcentaje de candidatos obtendrá más de 45 puntos?

c) ¿Qué porcentaje de candidatos obtendrá menos de 145 puntos?

d) ¿Cuántos candidatos obtendrán entre 55 y 145 puntos?

IV. DESCRIPCIÓN CONJUNTA DE DOS VARIABLES

Tema 1. Variables numéricas.

Relación o dependencia entre dos variables numéricas.

Batanero y Diaz Godino (2001) manifiestan que, cuando se realiza un estudio estadístico, se está interesado en más de un carácter de los individuos de la población. Una de las preguntas a las que se trata de dar respuesta es si existe alguna relación entre dos variables X e Y. Para algunos fenómenos, es posible encontrar una fórmula que exprese exactamente los valores de una variable en función de la otra: son los fenómenos llamados deterministas (o matemáticos).

Por ejemplo si la relación entre las variables X e Y, es de tipo lineal, a partir de una tabla de valores se puede obtener el modelo lineal correspondiente $y = mx+b$. De manera reversible, partiendo del modelo lineal es posible obtener los valores tabulados.

En contraparte, en una relación estadística con datos bivariados, no es posible obtener los valores tabulados a partir del modelo estimado, debido a la variabilidad presente en los datos bivariados.

Se puede estar interesado, por ejemplo, en la relación entre el número de respuestas irrelevantes de niños (Y) y la edad, en años (X). Cuando se dice que existe una relación funcional entre estas dos variables, significa que existe una función que describe esta relación de manera ”aproximada”.

De acuerdo con las ideas anteriores, *dos variables tienen una dependencia estadística o correlativa si existe una relación más o menos fuerte entre ambas variables y no existe una fórmula matemática que nos permita calcular el valor de una variable en función del valor de la otra.*

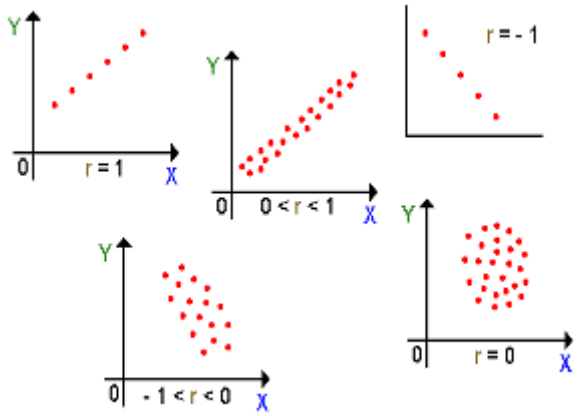
Diagrama de Dispersión

Batanero y Díaz Godino (2001) mencionan que existen muchos fenómenos en los que, al observar pares de valores (X, Y) correspondientes a variables estadísticas, representados en un plano cartesiano, los puntos, en general, no se ajustan de un modo preciso a una recta, sino que se obtiene un conjunto de puntos más o menos dispersos. Una representación de este tipo recibe el nombre de Nube de Puntos o Diagrama de Dispersión.

La dependencia o correlación es positiva cuando a mayor valor de una de ellas le corresponde mayor valor de la otra y, será una correlación negativa si a mayor valor de una le corresponde menor valor de la otra. Dos variables estadísticas son independientes (no están correlacionadas) cuando no existe correlación ni positiva ni negativa (su correlación es 0).

El grado de Correlación Lineal entre dos variables numéricas

El grado de correlación lineal entre dos variables se mide objetivamente con el Coeficiente de Correlación Lineal de Pearson. Este coeficiente (simbolizado con r) tiene las siguientes propiedades (observe las gráficas):



1. Su valor varía desde -1 hasta 1
2. Si r=-1 o r=1, la correlación lineal entre las variables es perfecta (funcional). En los dos casos en función de una variable podemos encontrar la otra.
3. Si -1<r<0, la correlación lineal será más fuerte a medida que r se aproxima a -1, y más débil a medida que se aproxima a cero. En este caso la correlación lineal es negativa o inversa.
4. Si r=0 entonces no existe ningún tipo de correlación lineal entre las dos variables. En este caso se dice que las variables X e Y son aleatoriamente independientes.
5. Si 0<r<1, la correlación lineal será más fuerte a medida que r se aproxima a 1, y tanto más débil a medida que r se aproxima a cero. En este caso la correlación es positiva o directa.

El Coeficiente de Correlación Lineal de Pearson se calcula con el siguiente procedimiento

$$r = \frac{n\left(\sum XY\right) - \left(\sum X\right)\left(\sum Y\right)}{\sqrt{n\left(\sum X^2\right) - \left(\sum X\right)^2} * \sqrt{n\left(\sum Y^2\right) - \left(\sum Y\right)^2}}$$

El coeficiente de correlación lineal toma valores entre -1 y 1. A continuación se presenta una tabla y algunas gráficas para diferentes valores de r, así como su interpretación correspondiente

Valor de r	Interpretación
r=-1	Correlación negativa perfecta
-1<r<-0.75	Correlación negativa alta
-0.75<r<-0.50	Correlación negativa moderada
-0.50<r<0	Correlación baja negativa
r=0	No existe correlación
0<r<0.50	Correlación baja positiva
0.50<r<0.75	Correlación positiva moderada
0.75<r<1	Correlación positiva alta
r=1	Correlación positiva perfecta

La recta de Regresión Lineal

Para ilustrar el concepto de regresión, consideremos los siguientes datos bivariados referidos a las variables edad (en años) y estatura promedio de 9 alumnos de una pequeña escuela

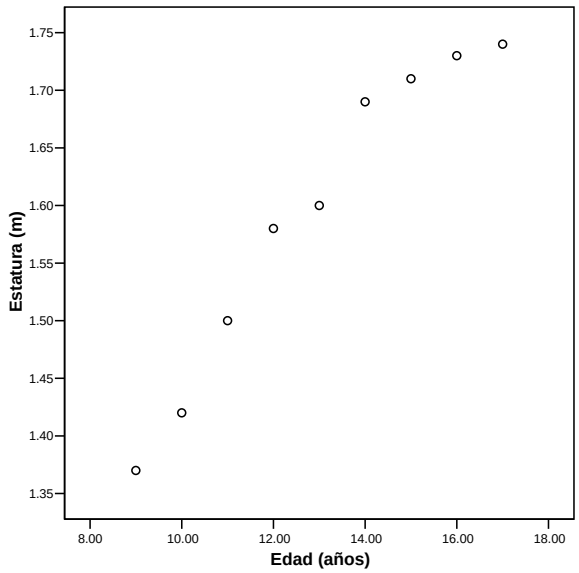
Alumno	Edad (años) (X)	Estatura promedio (m) (Y)	X^2	XY
1	9	1.37	81	12.33
2	10	1.42	100	14.2
3	11	1.50	121	16.5
4	12	1.58	144	18.96
5	13	1.60	169	20.8
6	14	1.69	196	23.66
7	15	1.71	225	25.65
8	16	1.73	256	27.68
9	17	1.74	289	29.58
Suma	117	14.34	1581	189.36

$$r = \frac{(9)(189.36)-(117)(14.34)}{\sqrt{(9)(1581-(117^2}\sqrt{(9)(23)-(14.34^2)}} = \frac{1704.24-1677.78}{(\sqrt{540})(\sqrt{1.3644})} = \frac{26.46}{27.14}=0.97$$

La interpretación es: hay correlación positiva alta entre las variables Edad y Estatura promedio.

La recta de Regresión Lineal

Consideremos la Gráfica siguiente que corresponde a la Edad y Estatura Promedio de los 9 alumnos de la tabla anterior.



La tabla y la gráfica nos confirman que existe una relación entre ambas variables y que ésta es positiva (a mayor edad mayor estatura). Podemos calcular el coeficiente de correlación lineal para medir el grado de relación entre ambas y utilizarla para estudiar la variable altura que llamaremos *dependiente o explicada* y para su estudio contamos con la variable edad (*llamada independiente o explicativa*).

Si en un momento dado elegimos un alumno y queremos conocer su altura, en primer lugar nos preguntaríamos por su edad, por ejemplo si necesitamos estimar la altura de un alumno de 13:5 años, intuitivamente contestaríamos que su estatura se encuentra entre: 1.60 y 1.69 m

Si además de la edad conociéramos más características que tienen que ver con la edad como el número de calzado, el peso y el género, podríamos dar una contestación más aproximada.

El ejemplo anterior nos permite ilustrar el funcionamiento de la regresión.

Cuando tenemos una variable independiente (o explicativa) hablamos de *regresión simple*. En el caso anterior (altura y edad de los alumnos) nos encontramos con que no existe una función que de manera exacta nos permita relacionar a la *variable dependiente (altura)* con la *variable independiente (edad)*. Es decir, la representación gráfica de los datos *(diagrama de dispersión)* no coincide con la gráfica de ninguna función matemática. Por esta razón al hablar de la relación entre ambas variables, y al relacionar la altura de un alumno cualesquiera (Y_i) con su edad (X_i), lo denotaremos como $Y_i=f(X_i)+d_i$, donde d_i significa la inexactitud de $f(X_i)$ con respecto a Y_i.

Puesto que desconocemos el tipo de función $f(X_i)$ que relaciona ambas variables el primer paso consiste en establecer este tipo de función, esto es, si la relación entre las variables es lineal curvilínea o de cualquier otra forma. A continuación pasaríamos a concretar la ecuación estimando los parámetros de dicha función.

La linealidad de la relación

Etxeberría (1999**)** afirma que existen diversas razones que "justifican" la suposición de la linealidad de la relación:

1. De forma empírica se ha comprobado la linealidad de la relación en numerosas ocasiones.
2. En la mayoría de las veces, no disponemos de teorías o conocimientos que nos permitan asegurar que la relación no es lineal.
3. En la mayor parte de los casos, la representación gráfica de los datos tampoco aporta una alternativa válida a la linealidad de la función (nos encontramos con una nube dispersa de puntos).
4. Por último, la especificación lineal es, generalmente, la más sencilla.

Adecuación del modelo lineal: la recta que mejor se ajusta

Una vez decidido el modelo, en nuestro caso una recta $y = mx + b$, debemos de calcular su ecuación. Esta debe ser la que mejor se ajusta a nuestro conjunto de datos. Pero surge una pregunta, qué quiere decir ¿recta que mejor se ajusta? El concepto de "ajuste" es el criterio que utilizaremos para poder hablar de una buena o mala aproximación de la recta a nuestro conjunto de datos. Dentro de los criterios existentes y debido a sus excelentes propiedades estadísticas, el más utilizado es el *Criterio de los Mínimos Cuadrados*. Siguiendo este criterio podemos afirmar *que los valores de la pendiente de la recta (m) y de la ordenada al origen (b) que definen la ecuación de regresión, serán aquellos que minimicen la suma de las distancias (o diferencias) al cuadrado entre los valores Y_i observados y los valores estimados con la ecuación y=mx+b*

De manera general la recta que mejor se ajusta a los datos se obtiene resolviendo el siguiente sistema de dos ecuaciones con dos incógnitas. El sistema a resolver es

$$\bar{x}m + b = \bar{y} \quad \text{----------} \quad (1)$$
$$\left(\sum x^2\right)m + \left(\sum x\right)b = \sum xy \quad \text{-------} \quad (2)$$

Sustituyendo en las ecuaciones anteriores:

$$13m+b=1.59$$
$$1581m+117b=189.36$$

Resolviendo el sistema se obtiene *m*=0.05 y *b*=0.94 y la ecuación de la recta de regresión que mejor se ajusta a los datos es: y=0.05x+0.94.
Significado e interpretación de *m* y *b*.

En este modelo el coeficiente *b* indica el valor de Y (estatura) cuando la variable (X) toma el valor cero y representa la influencia de otras variables que no hemos tenido en cuenta al analizar la variable Y. También es el punto donde la recta de regresión corta al eje Y.

El valor *m* representa lo que aumenta el valor de Y cuando la variable X aumenta en una unidad. También es la pendiente de la recta de regresión.

De acuerdo con lo anterior, esta ecuación se interpreta de la siguiente manera: *a partir de una altura de 0.94 m la estatura se incrementa 0.05 m por cada año de edad transcurrido.*

Predicción estadística

Este modelo se puede utilizar para predecir un valor de Y, a partir de un valor de X; por ejemplo si un alumno tiene 13.5 años, tendrá un estatura estimada de y=(0.05)(13.5)+0.94=1.615 m

Es importante señalar que las predicciones para la variable dependiente (Y) solamente son válidas dentro del rango de valores de la variable dependiente (X); en este caso las predicciones de la altura sólo son válidas desde (incluso) 9 hasta (incluso) 17.

En este sentido, Flores y Lozano (1998) afirman que las extrapolaciones fuera de los valores no tienen sustento alguno, por eso no es recomendable extrapolar más allá de los valores límites del rango; mientras más alejados estén los valores asignados a X, menos confiables van a ser las predicciones.

Los valores de m y b de la recta de regresión lineal también se pueden calcular con los siguientes procedimientos

$$m = \frac{n\sum X - \sum X \sum Y}{n\sum X^2 - (\sum X)^2} \qquad b = \frac{\sum Y - m\sum X}{n}$$

La cuantificación de la variación explicada por el modelo lineal $y = mx + b$

Otro objetivo importante en la descripción numérica de dos variables relacionadas consiste en cuantificar la proporción de variación (o variabilidad) de la variable dependiente (Y) explicada por la variable independiente (X). A la proporción de la variación no explicada se le llama error. Este error es debido a la variabilidad presente en los datos bivariados.

Para ejemplificar esta idea, consideremos el siguiente caso referido a 10 datos bivariados obtenidos de alumnos de Quinto Semestre del CCH-Sur de la

UNAM. Las variables involucradas son: Horas dedicadas al estudio por día (X) y promedio académico (Y).

X	10	6	7	15	12	10	9	7	9	5
Y	8.2	7.3	8.0	9.4	9.2	9.0	9.0	7.8	8.1	6.2

El modelo lineal estadístico es y=0.28x+5.64, que se interpreta como: a partir de un promedio de 5.64, por cada hora de estudio este promedio aumenta en 0.28

La variación total de la variable Y (promedio académico)

Vamos a considerar que la variación total (VT) es la suma de la variación explicada (VE) por el modelo lineal más la variación no explicada (Error).

Para encontrar la variación total, se le resta a cada valor observado de la variable Y, la media de esta variable, la diferencia se eleva al cuadrado:

($Y_i - \bar{y}$)^2 y se suma el total de estos cuadrados.

($Y_i - \bar{y}$)	($Y_i - \bar{y}$)^2
8.2-8.22=-0.02	0.0004
7.3-8.22=-0.42	0.8464
8.0-8.22=-0.22	0.0484
9.4-8.22=1.18	1.3924
9.2-8.22=0.98	0.9604
9.0-8.22=0.78	0.6084
9.0-8.22=0.78	0.6084
7.8-8.22=-0.42	0.1764
8.1-8.22=-0.12	0.0144
6.2-8.22=-2.02	4.0804

La suma de cuadrados que expresa la variación total (VT), es

0.0004+0.8464+……+4.0804=8.736

La variación no explicada por el modelo lineal (Error)

Para encontrar la variación no explicada (Error), a los valores observados de Y_i se le restan los valores estimados por el modelo y=0.28x+5.64, se elevan las diferencias al cuadrado y se suman estaos cuadrados, como se observa en la siguiente tabla:

$(8.2-[0.28(10)+5.64])^2 =0.0576$
$(8.2-[0.28(6)+5.64])^2 =0.0004$
$(8.2-[0.28(7)+5.64])^2 =0.16$
$(8.2-[0.28(15)+5.64])^2 =0.1936$
$(8.2-[0.2812)+5.64])^2 =0.04$
$(8.2-[0.2810)+5.64])^2 =0.3136$
$(8.2-[0.28(9)+5.64])^2 =0.7056$
$(8.2-[0.28(7)+5.64])^2 =0.04$
$(8.2-[0.28(9)+5.64])^2 =0.0036$
$(8.2-[0.28(5)+5.64])^2 =0.7056$

La suma de cuadrados que expresa el Error es:

0.0576+0.0004+……+0.7056=2.22

La variación explicada por el modelo lineal (VE)

Como se menciona anteriormente, VT=VE+Error, sustituyendo la suma de cuadrados de VE y la suma de cuadrados del Error

8.736=VE+2.22, por lo que VE=8.736-2.22=6.516 que se puede expresar como: 1=0.7458+0.2541, o de manera equivalente, 100%=74.58%+25.41%

A la proporción de VE por el modelo también se le llama Coeficiente de Determinación y se simboliza como r^2. En este caso r^2=74.58%, lo que significa que el 74.58% de la variación en el promedio esta explicada por las horas dedicadas al estudio diariamente. El 25% se debe a otras variables como pueden ser; capacidad de aprendizaje, asistencia a clases, etc.

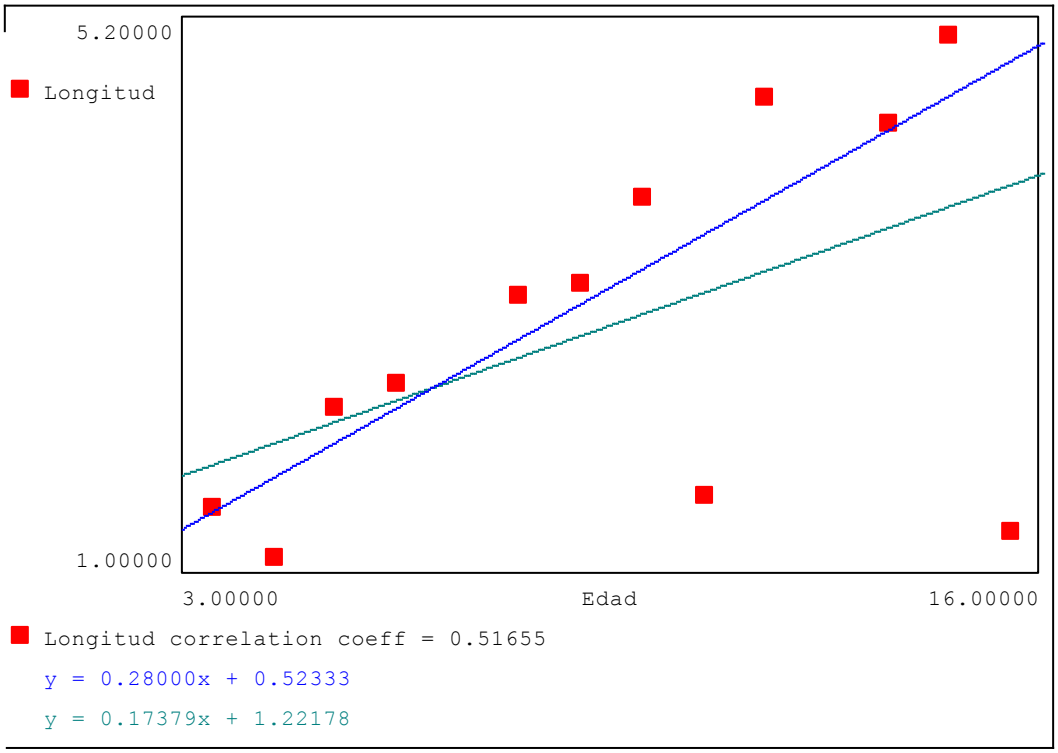
Cuando se extrae la raíz cuadrada del Coeficiente de Determinación, se obtiene el Coeficiente de Correlación Lineal de Pearson (r). En este caso

r=$\sqrt{0.7458}$=0.8636, lo que se interpreta como una relación positiva alta entre las variables horas de estudio y promedio.

Es importante señalar que los datos atípicos aumentan la variación del conjunto de datos y, en consecuencia el error, por lo que el modelo lineal no proporciona un buen ajuste.

Para ilustrar esta situación consideremos el estudio de la relación entre las variables X: Edad (en días) de mariposas y Y: Longitud de ala (cm), se analizan los siguientes datos bivariados (adaptados de Badii et al, 2012). En el Diagrama de Dispersión siguiente se observa que los puntos siguen una tendencia aproximadamente lineal, pero existen 3 datos bivariados (atípicos) que se alejan demasiado de los otros 10.

Usando el Método de Mínimos Cuadrados, se obtiene el modelo lineal y=0.28x+0.5233. La interpretación de este modelo es: A partir de una longitud de ala de 0.5233 cm, por cada día transcurrido la longitud del ala se incrementa en 0.28 cm. Con este modelo lineal se puede predecir la longitud de ala de una mariposa de 7 días de edad: y=(0.28)(7)+0.5233=2.4833. El modelo lineal estima la longitud de ala como 2.4833 cm.



El Método de Mínimos Cuadrados se basa en las medias aritméticas de cada variable ($\bar{x}$=9.4166, $\bar{y}$=2.8583) y, estos valores se ven afectados por los datos atípicos. Un método alternativo para encontrar otro modelo lineal es el Método de Tukey (que se basa en las medianas de los datos). Las medianas tienen la propiedad de que no se ven alteradas por los datos atípicos. El modelo lineal obtenido con este procedimiento es: y=0.1738x+1.2218. Este modelo informa

que la longitud inicial de ala es de 0.2218 cm, que se incrementa en 0.1738 cm por cada día de edad. La predicción estadística para la longitud de ala de una mariposa con edad de 7 días, es y=(0.1738)(7)+1.2218=2.4384 cm.

Los dos modelos lineales tienen un mal ajuste, el error en el modelo y=0.28x+0.5233, es $\frac{21.18}{23.65} = 0.895$, el error en el modelo y=0.1738x+1.2218, es $\frac{16.65}{23.65}$=0.7

Ejemplos de Regresión y Correlación Lineal

Ejemplo 1

Los datos siguientes corresponden a la edad (años) y el tiempo (minutos) que 10 alumnos invierten en realizar alguna actividad física.

Edad (X)	22	25	30	28	31	21	29	26	27	33
Tiempo (Y)	100	75	60	80	50	110	60	95	85	35

El modelo estadístico de regresión lineal es, y=-5.813x+233.115 y su interpretación es: partiendo de 233.115 minutos, por cada año de edad, el tiempo dedicado a realizar alguna actividad física disminuye en 5.813 minutos.

El Coeficiente de Correlación Lineal de Pearson, es r=-0.944 se interpreta como: existe correlación alta negativa entre la Edad y el Tiempo en realizar alguna actividad física.

El Coeficiente de Determinación (r^2) es $(-0.944)^2$=0.891, lo que significa que el 89.1% de la variación en el tiempo para realizar alguna actividad física está explicada por la edad, el 10.9% restante se debe a otras variables no analizadas, por ejemplo condiciones climáticas no favorables.

Ejemplo 2

Encuesta realizada a 10 contadores de una empresa. Los datos se refieren a los años que llevan laborando (X) y la satisfacción en el trabajo (Y), medida en una escala de 0 a10.

X	5	2	6	17	9	12	24	17	25	28
Y	8	7	7.5	9.5	8.5	7.5	3	6	6.5	5

El modelo estadístico lineal es, y=-0.11x+8.44, lo que significa que a partir de 8.44 puntos de satisfacción laboral, por cada año trabajando, esta satisfacción disminuye en 0.11 puntos.

La Suma de Cuadrados de la Variación Total (VT) es 31.025

La Suma de Cuadrados del Error, es 20.36

La suma de cuadrados de la variación explicada por el modelo lineal es 31.025-20.36=10.665

El Coeficiente de Determinación es, $r^2 = \frac{10.665}{31.025} = 0.3437$, que se interpreta como: el 34% de la variación en la satisfacción laboral esta explicada por los años de trabajo.

La proporción de variación no explicada (error), es: 1-0.3437=0.6563, lo que indica que el modelo tiene mal ajuste.

Actividades

1. Los datos bivariados siguientes corresponden a la edad (años) y la conducta agresiva (escala de 0 a 10) en 10 niños

Edad (X)	6	6	6.7	7	7	4.7	9.8	8.2	8.5	8.9
Conducta (Y)	9	6	7	8	7	4	2	3	3	1

   a) Construya el diagrama de dispersión.
   b) Verifique que la ecuación de regresión lineal es y=-2.34x+22.51
   c) Haga el diagrama de dispersión.
   d) Interprete este modelo lineal.
   e) Haga una predicción estadística.
   f) Calcule e interprete el coeficiente de determinación.

2. En la tabla siguiente se muestra la relación entre el número de artículos y su precio

No. de artículos (X)	1	3	5	10	12	15	24
Costo por unidad (Y)	55	52	48	36	32	30	25

   a) Trace el diagrama de dispersión de los datos.

   b) Verifique que la ecuación de regresión es: y = - 1.407x + 53.793. Interprete esta ecuación.
   c) Grafique la recta de regresión.

   d) Si una persona compra 20 artículos, ¿Cuál será el costo por unidad?

   e) Si una persona paga $ 42.55 por unidad, ¿cuántos artículos debe comprar?
   f) Calcule e interprete el coeficiente de determinación

Tema 2. Variables categóricas

Dependencia estadística entre dos variables categóricas binomiales

La dependencia entre dos variables categóricas binomiales se puede calcular utilizando el Coeficiente Φ:

$$\Phi=\frac{AD-BC}{\sqrt{(A+B)(C+D)(A+C)(B+D)}}$$

Ejemplo

Para estudiar hasta qué punto los trastornos de ansiedad y depresión están relacionados, disponemos de una muestra de 100 personas. Después de pasarles una prueba para evaluar el estado de ánimo depresivo, dividimos a los sujetos por la mediana, siendo considerados depresivos los que se encuentran por encima de la misma y no depresivos los que se encuentran por debajo de ella. El 47% de los sujetos presentaba un trastorno de ansiedad, mientras que el 48% no padecía ni ansiedad ni depresión.

	Depresión (no)	Depresión (si)	
Ansiedad (no)	A=48	B=5	A+B=53
Ansiedad (si)	C=2	D=45	C+D=47
	A+C=50	B+D=50	n=100

$$\Phi = \frac{(48)(45)-(5)(2)}{\sqrt{(53)(47)(50)(50)}} = \frac{2160-10}{\sqrt{6227500}} = \frac{2150}{2495.496} = 0.86$$

A diferencia del coeficiente de correlación linear de Pearson *r*, que toma valores entre –1 y 1; el coeficiente Φ solamente toma valores entre 0 y 1 interpretándose de manera semejante a *r*. En este ejemlo, la interpretación es: existe correlación positiva alta entre las variables “trastornos de ansiedad” y “depresión”.

Actividad 1

Se desea saber si existe relación entre las variables “consumo de alcohol del padre” y “consumo de alcohol de los hijos estudiantes de preparatoria”

Estudiantes	Padre que si consumen alcohol	Padres que no consumen alcohol	Total
Si consumen alcohol	21	9	
No consumen alcohol	30	80	
Total			

Actividad 2

Se ha observado que los estudiantes que inician estudios de maestría presentan dificultad en el primer semestre por lo que algunos abandonan sus estudios. A continuación se presentan los resultados de un seguimiento realizado a 15 estudiantes de una maestría.

Estado Civil	**Permanencia**
Casado (0)	Abandona el Curso (0)
No casado (1)	Permanece en el curso (1)
0	1
0	0
1	1
1	0
0	0
1	1
0	0
0	1
0	0
1	1

0	0
0	0
0	0

¿Existe relación entre las variables estado civil y permanencia en los estudiantes que inician estudios de maestría?

Fuentes consultadas

Badii, M. H. et al (2012). Análisis de Regresión Lineal Simple para Predicción. En Daena: International Journal of Good Conscience. Vol. 7. No. 3. 67-81.

Batanero, C. et al. (2015). La dispersión como elemento estructurador del Currículo de Estadística y Probabilidad. Epsilon. Vol. 32. No. 2. 7-20.

Behar, R. (2009) Búsqueda del conocimiento y pensamiento estadístico. Segundo Encuentro Iberoamericano de Biometría. Veracruz.

Behar, R. y Grima, P. (2013). El histograma como un instrumento para la comprensión de las funciones de densidad de probabilidad. En J. M. Contreras, G. R. Cañadas, M. M. Gea y P. Arteaga (Eds.) Actas de las Jornadas Virtuales en Didáctica de la Estadística, Probabilidad y Combinatoria (pp. 229-235). Granada, Departamento de Didáctica de la Matemática de la Universidad de Granada.

Etxeberría, J. (1999). Regresión Múltiple. Cuadernos de Estadística No. 4. La Muralla.

Flores, R. y Lozano, H. (1998). Estadística aplicada para la administración. Grupo Editorial Iberoamérica. México.

Flores, G. y Díaz, M. A. (2013). México en PISA 2012. INEE.

Jiménez, Ma. R. y Hernández J. D. (2004). Propuesta para desarrollar los contenidos temáticos de Estadística y Probabilidad I. CCH-Sur. UNAM

Orozco Ma. de J. et al (2009). Identificación de factores que obstaculizan la estancia y egreso de estudiantes del programa de Nivelación a Licenciatura en Trabajo Social; Abierto y a Distancia. Humanismo y Trabajo Social. 247-258.

Profeco. (2003). Revista del Consumidor. 24-27

Rodríguez, M. I. (2012). Inferencia informal: del análisis de los datos a la inferencia estadística. X Congreso Latinoamericano de Sociedades de Estadística. Córdoba.

STPS-INEGI. (2017). Encuesta Nacional de Ocupación y. Empleo.

UPN-SEP. Introducción a los Métodos Estadísticos. Vol. 1. 1981

.