INSTITUTO TECNOLÓGICO SUPERIOR "JAPÓN"

Guía Metodológica De Estadística Aplicada




Compilado por:

Mgs. Daniel Shauri Romero

Carrera: Administración de Empresas

2018

INSTITUTO TECNOLÓGICO SUPUERIOR JAPÓN
GUIA DE APRENDIZAJE

1. IDENTIFICACIÓN DE

Nombre de la Asignatura: **ESTADISTICA APLICADA**	**Componentes del Aprendizaje**	
Resultado del Aprendizaje: **COMPETENCIAS Y OBJETIVOS** • Conocer los conceptos de las teorías básicas de la estadística. • Utilizar fórmulas de medidas, tendencia central en ejercicios prácticos. • Aplicar fórmulas de medidas de dispersión determinando coeficientes de variación, error, estándar.		
Docente de Implementación:		
MsC. JOSÉ DANIEL SHAURI ROMERO	**Duración:** 20 horas	

Unidades	Competencia	Resultados de Aprendizaje	Actividades	Tiempo de Ejecución

1. Generalidades.	Conoce conceptos de las teorías básicas de la estadística.	**COGNITIVO:** Comprender la utilidad, clasificación y los conceptos básicos de la Estadística, Población, Parámetro, Muestra, Estadígrafos, Tipos de Variable (cualitativa y cuantitativa, discreta y continua), Escalas de Medición de las variables; nominal, ordinal, de intervalo y de razón. Proponer lecturas para hacer en casa sobre Investigación Estadística **PROCEDIMENTAL:** Dar ejemplos para identificar cada uno de los términos básicos de la estadística, su aplicación en las diferentes áreas de formación, para	• Dinámicas grupales • Lecturas reflexivas del material proporcionado • Exposiciones orales sobre el tema de investigación asignado. • Intervención de los señores estudiantes con criterios sobre el tema en un foro abierto. • Investigaciones sobre el tema para fortalecer los conocimientos • Ejercicios de Aplicación	**2.5**

		distinguir las diferentes escalas de medición con cada tipo de variable. Identificar los pasos de una investigación estadística. **ACTITUDINAL:** Comprender el proceso de estudio aplicando procedimientos la comprensión de la información, con actitud positiva al trabajo académico.		

2. Organización de información para datos agrupados puntuales.	Aplica la distribución de frecuencias con ejemplos del quehacer educativo.	**COGNITIVO:** Organizar y representar datos puntuales mediante Tablas de Distribución de Frecuencias y gráficas: diagramas de barras y pastel. **PROCEDIMENTAL:** Comprender la necesidad de organizar la información recolectada durante la realización de un estudio, por medio de tablas de distribución de frecuencias y gráficas. Interpretar las tablas de distribución de frecuencias y las gráficas. Proponer trabajo extra-clase en el que se indague como hacer lo visto en clase con Excel.	• Dinámicas grupales • Lecturas reflexivas del material proporcionado • Exposiciones orales sobre el tema de investigación asignado. • Intervención de los señores estudiantes con criterios sobre el tema en un foro abierto. • Investigaciones sobre el tema para fortalecer los conocimientos • Ejercicios de Aplicación	**2.5**

		ACTITUDINAL: Aplicar los procesos de matemáticos al ámbito educativo propio y de los educandos.		

3. Organización de información para datos agrupados por intervalos.	Aplica la distribución de frecuencias para datos agrupados con intervalos con ejemplos del quehacer educativo.	**COGNITIVO:** Organizar y representar datos por intervalos mediante Tablas de Distribución de Frecuencias y gráficas: Histogramas, polígonos, ojivas, diagrama de pastel, diagrama de puntos, diagrama de tallo y hoja. **PROCEDIMENTAL:** Comprender la necesidad de organizar la información recolectada durante la realización de un estudio, por medio de tablas de distribución de frecuencias y gráficas. Interpretar las tablas de distribución de	• Dinámicas grupales • Lecturas reflexivas del material proporcionado • Exposiciones orales sobre el tema de investigación asignado. • Intervención de los señores estudiantes con criterios sobre el tema en un foro abierto. • Investigaciones • Ejercicios de Aplicación	**5**

		frecuencias y las gráficas. Proponer trabajo extra-clase en el que se indague como hacer lo visto en clase con Excel. **ACTITUDINAL:** Aplicar procesos de formación y educativo		

4. Medidas de tendencia central para datos agrupados y no agrupados.	Utiliza fórmulas de medidas de tendencia central en ejercicios prácticos.	**COGNITIVO:** Aplicar las formulas e interpretar cálculos de las medidas de tendencia central para datos agrupados y no agrupados; Media Aritmética, Mediana y Moda; Relación entre Media, Mediana y Moda. **PROCEDIMENTAL:** Hacer talleres aplicando los anteriores conceptos: medidas de tendencia central e Interpretar resultados. Proponer trabajo extra-clase en el que se indague como hacer lo visto en clase con Excel. **ACTITUDINAL:**	• Dinámicas grupales • Lecturas reflexivas del material proporcionado • Exposiciones orales sobre el tema de investigación asignado. • Trabajo cooperativo para la aplicación de talleres sobre el tema de estudio. • Investigaciones sobre el tema para fortalecer los conocimientos • Ejercicios de aplicación	**5**

		Aplicar esfuerzo actitud y aptitud al proceso de formación y educativo		

5. Medidas de Dispersión, para datos agrupados y no agrupados	Aplica fórmulas de medidas de dispersión determinando coeficientes de variación, varianza, error estándar, en ejercicios prácticos	**<u>COGNITIVO:</u>** Aplicar las formulas e interpretar cálculos de las Medidas de Dispersión para datos agrupados y no agrupados; varianza, desviación estándar y coeficiente de variación. **<u>PROCEDIMENTAL:</u>** Plantear situaciones que lleven al cálculo e interpretación de resultados relacionados con las medidas de dispersión para datos agrupados y no agrupados. Proponer trabajo extra-clase en el que se indague como hacer lo visto en clase con Excel. **<u>ACTITUDINAL</u>:**	• Dinámicas grupales • Lecturas reflexivas del material proporcionado • Exposiciones orales sobre el tema de investigación asignado. • Trabajo cooperativo para la aplicación de talleres sobre el tema de estudio. • Investigaciones sobre el tema para fortalecer los conocimientos. • Ejercicios de aplicación (Prácticos)	**5**

		Comprender el proceso de estudio aplicando técnicas para la comprensión de la información, con actitud positiva al trabajo académico.		

2. CONOCIMIENTOS PREVIOS Y RELACIONAD

Co-requisitos
3. UNIDADES TEÓRICAS

- **Desarrollo de las Unidades de Aprendizaje (contenidos)**
  - **A. Base Teórica**

DESCRIPCIÓN DE UNA MUESTRA

<u>1. INTRODUCCIÓN</u>

1.1 DEFINICIÓN DE ESTADÍSTICA

1.2 MODELO ESTADÍSTICO

1.3 ESTADÍSTICA DESCRIPTIVA

1.4 CONCEPTOS BÁSICOS

POBLACIÓN

VARIABLE: **Cualitativas o Categóricas y Cuantitativas (Discretas y Continuas)**

MUESTRA

TAMAÑO MUESTRAL

DATO

2. DISTRIBUCIONES DE FRECUENCIAS

2.1 FRECUENCIA ABSOLUTA

2.2 FRECUENCIA RELATIVA

2.3 FRECUENCIA ACUMULADA

2.4 FRECUENCIA RELATIVA ACUMULADA

2.5 TABLA DE FRECUENCIAS

2.6 DISTRIBUCIONES DE FRECUENCIAS AGRUPADAS

3. MÉTODOS GRÁFICOS

3.1 FRECUENCIAS NO ACUMULADAS

DIAGRAMA DE BARRAS

DIAGRAMA DE SECTORES O DE PASTEL

PICTOGRAMA

HISTOGRAMA

3.2 FRECUENCIAS ACUMULADAS

POLÍGONO DE FRECUENCIAS

4. MEDIDAS DESCRIPTIVAS

4.1 MEDIDAS DE POSICIÓN

4.1.1 MEDIDAS DE TENDENCIA CENTRAL

MEDIA ARITMÉTICA

MEDIANA

MODA

MEDIA GEOMÉTRICA

MEDIA ARMÓNICA

4.1.2 MEDIDAS DE POSICIÓN NO CENTRALES: CUANTILES

PERCENTILES

CUARTILES

DECILES

4.1.3 MOMENTOS

MOMENTOS RESPECTO AL ORIGEN

MOMENTOS CENTRALES O RESPECTO A LA MEDIA

4.2 MEDIDAS DE DISPERSIÓN

4.2.1 MEDIDAS DE DISPERSIÓN ABSOLUTAS

VARIANZA

DESVIACIÓN TÍPICA

CUASI-VARIANZA

DESVIACIÓN MEDIA RESPECTO A LA MEDIA

DESVIACIÓN MEDIA RESPECTO A LA MEDIANA

RECORRIDO O RANGO MUESTRAL

RECORRIDO INTERCUARTÍLICO

4.2.2 MEDIDAS DE DISPERSIÓN RELATIVAS

COEFICIENTE DE VARIACIÓN DE PEARSON

4.3 OTRAS MEDIDAS DESCRIPTIVAS

4.3.1 TIPIFICACIÓN DE UNA DISTRIBUCIÓN DE FRECUENCIAS

4.3.2 MEDIDAS DE FORMA

A: Medidas de *ASIMETRÍA*

COEFICIENTE DE ASIMETRÍA DE FISHER

COEFICIENTE DE ASIMETRÍA DE PEARSON

B: Medidas de *APUNTAMIENTO O CURTOSIS*

COEFICIENTE DE APUNTAMIENTO DE FISHER

4.3.3 MEDIDAS DE CONCENTRACIÓN

ÍNDICE DE CONCENTRACIÓN DE GINI

CURVA DE LORENZ

5. TRANSFORMACIONES LINEALES

5.1 EN LA MEDIA

5.2 EN LA MEDIANA

5.3 EN LA VARIANZA

5.4 EN LA DESVIACIÓN TÍPICA

GENERALIDADES
DESCRIPCIÓN DE UNA MUESTRA

1. INTRODUCCIÓN

Ejemplo 1

El gobierno desea averiguar si el número medio de hijos por familia ha descendido respecto a la década anterior. Para ello ha encuestado a 50 familias respecto al número de hijos y ha obtenido los siguientes datos:

2 4 2 3 1 2 4 2 3 0 2 2 2 3 2 6 2 3 2 2 3 2 3 3 4
3 3 4 5 2 0 3 2 1 2 3 2 2 3 1 4 2 3 2 4 3 3 2 2 1

Ejemplo 2

Un nuevo hotel va abrir sus puertas en una cierta ciudad. Antes de decidir el precio de sus habitaciones, el gerente investiga los precios por habitación de 40 hoteles de la misma categoría de esta ciudad. Los datos obtenidos (en miles de pesetas) fueron:

3.9	4.7	3.7	5.6	4.3	4.9	5.0	6.1	5.1	4.5
5.3	3.9	4.3	5.0	6.0	4.7	5.1	4.2	4.4	5.8
3.3	4.3	4.1	5.8	4.4	3.8	6.1	4.3	5.3	4.5
4.0	5.4	3.9	4.7	3.3	4.5	4.7	4.2	4.5	4.8

1.1 DEFINICIÓN DE ESTADÍSTICA: es la ciencia que se encarga de la recopilación, representación y el uso de datos sobre una o varias características de interés para, a partir de ellos, tomar decisiones o extraer conclusiones generales.

1.2 MODELO ESTADÍSTICO:

- PASO 0: Planteamiento del problema en términos precisos: ámbito de aplicación (población) y característica(s) a estudio (variable(s))
- PASO 1: Recogida de datos de la población de interés (MUESTREO)
- PASO 2: Organización, Presentación y Resumen de los datos (o de la muestra). (ESTADÍSTICA DESCRIPTIVA).
- PASO 3: Confección de modelos matemáticos. (TEORÍA DE LA PROBABILIDAD).
- PASO 4: Obtener conclusiones generales o verificar hipótesis (INFERENCIA ESTADÍSTICA).

1.3 ESTADÍSTICA DESCRIPTIVA: es la parte de la estadística que se encarga de organizar, resumir y dar una primera descripción (sin conclusiones generales) de los datos.

1.4 CONCEPTOS BÁSICOS:

POBLACIÓN**:** Es el conjunto de individuos o entes sujetos a estudio (En nuestro caso las poblaciones serían: en el ejemplo primero el conjunto de todas las familias españolas y en el segundo ejemplo el conjunto de todos los hoteles de esta categoría de esta ciudad.). Algunas poblaciones son finitas y pueden conocerse; otras pueden ser infinitas y abstractas: Ej.: el conjunto de todos los hoteles o el conjunto de todas las piezas fabricadas por una máquina.

VARIABLE: Característica que estamos midiendo (Ej. 1: número de hijos, Ej. 2: precio de la habitación) **Las variables** se suelen denotar por letras mayúsculas: X, *Y,...*

Tipos de variables:

1. **Cualitativas o Categóricas:** aquellas que no son medibles, es decir, aquellas cuyas observaciones no tienen carácter numérico. Expresan cualidades o categorías. Ej.: estado civil, sexo o profesión. (A las variables cualitativas también se les llama atributos).

2. **Cuantitativas:** aquellas que son medibles, es decir sus observaciones tienen carácter numérico. Estas se dividen a su vez en:
   * **Discretas:** toman valores en un conjunto numerable. Ej.: Número de habitaciones de un hotel, número de hijos de una familia, número de obreros de una fábrica.
   * **Continuas:** toman valores en un conjunto no numerable (los números reales o un intervalo). Ej.: peso, estatura.

NOTA: La distinción entre variables discretas y continuas es más teórica que práctica, puesto que la limitación de los aparatos de medida hace que todas las variables se comporten como discretas cuando se pretende observarlas. De momento haremos más flexible el concepto de variable continua considerando continua a aquella variable que toma un gran número de valores diferentes, en este sentido podemos considerar la variable precio de la habitación como continua.

MUESTRA: Es un conjunto finito de elementos seleccionados de la población. (Las 50 familias, los 40 hoteles)

TAMAÑO MUESTRAL: número de observaciones en la muestra. Habitualmente se denotará por *n.*

DATO: cada valor observado de la variable. Si representamos por X a la variable, representaremos por x_i cada dato diferente observado en la muestra, el subíndice i indica el lugar que ocupa si los ordenamos de menor a mayor.

Ej1: $x_1 = 0, x_2 = 1$

Ej2: $x_1 = 3.3, x_2 = 3.7$

Denotaremos por *k* al número de valores distintos.

2. DISTRIBUCIONES DE FRECUENCIAS

Observando los datos del ejemplo es fácil adivinar cuál será el primer paso en la organización de los datos; consistirá en agrupar aquellos datos que se repiten varias veces. Tenemos las siguientes definiciones:

2.1 FRECUENCIA ABSOLUTA (n_i): Es el número de veces que se repite un determinado valor (x_i) de la variable. Ej1: para el dato x_1=0 n_1=2, para el dato x_4=3 n_4=15.

PROPIEDAD*:* la suma de todas las frecuencias absolutas es igual al tamaño muestral.

Este tipo de frecuencias no son comparables con las obtenidas en otras muestras de distinto tamaño.

2.2 FRECUENCIA RELATIVA (fi): Es igual a la frecuencia absoluta dividida por el número total de datos, es decir por el tamaño muestral $f_i=n_i/n$**.** *Ei1.:* f_1=2/50=0.04, f_4=15/50=0.3

PROPIEDAD: la suma de todas las frecuencias relativas es igual a la unidad.

2.3 FRECUENCIA ACUMULADA (N_i): Nos dice el número de datos que hay igual o inferiores a uno determinado. Se calcula: $N_i = \sum_{j=1}^{i} n_j = N_{i-1} + n_i$

Ej1: N_1=2, N_4=42.

PROPIEDAD: La última frecuencia acumulada absoluta es el tamaño muestral.

2.4 FRECUENCIA RELATIVA ACUMULADA (F_i): Es el resultado de dividir cada frecuencia acumulada por el número total de datos $F_i = \frac{N_i}{n} = \sum_{j=1}^{i} f_j$

Ej1: F_1=0.04, F_4=42/50=0.84.

PROPIEDAD: La última frecuencia relativa acumulada es la unidad.

2.5 TABLA DE FRECUENCIAS:

Llamamos así a una tabla conteniendo el conjunto de diferentes valores que ha tomado una variable (los datos sin repetir) ordenados de menor a mayor con sus correspondientes frecuencias.

Ejemplo 1:

x_i	n_i	f_i	N_i	F_i
0	2	0.04	2	0.04
1	4	0.08	6	0.12
2	21	0.42	27	0.54
3	15	0.3	42	0.84
4	6	0.12	48	0.96
5	1	0.02	49	0.98
6	1	0.02	50	1

¿Cuál es el número de familias que tiene como máximo dos hijos?

En la columna de las n_i: 2+4+21=27 o en la columna de las N_i: N2= 27

¿Cuántas familias tienen más de 1 hijo pero como máximo 3?

En la columna de las n_i: 21+15=36 o en la columna de las N_i: 42-6=36

¿Qué porcentaje de familias tiene más de 3 hijos?

En la columna de las f_j: 0.12+0.02+0.02=0.16, que supone un 16% o en la columna de las F_i: 1-0.84=0.16, 16%

Ejemplo 2:

x	n_i	f_j	N_i	F_j
3.6	2	0.05	2	0.05
3.7	1	0.025	3	0.075
3.8	1	0.025	4	0.1
3.9	3	0.075	7	0.175
4	1	0.025	8	0.2
4.1	1	0.025	9	0.225
4.2	2	0.05	11	0.275
4.3	4	0.1	15	0.375
4.4	2	0.05	17	0.425
4.5	4	0.1	21	0.525
4.7	4	0.1	25	0.625
4.8	1	0.025	26	0.650
4.9	1	0.025	27	0.675
5	2	0.05	9	0.725
5.1	2	0.05	31	0.775
5.3	2			
5.4	1			
5.6	1			
5.8	2			
6	1			
6.1	2			

2.6 DISTRIBUCIONES DE FRECUENCIAS AGRUPADAS

Hemos visto en el caso anterior que los valores distintos que tomaba la variable eran muchos, es decir *k* era grande y eso hacía que la tabla obtenida fuera muy poco manejable y por tanto poco clarificadora. Esto nos va a ocurrir frecuentemente en el caso en que la variable a estudiar sea continua. La solución **es agrupar los diferentes valores de la variable en intervalos o intervalos de clase**. Teniendo en cuenta que lo que ganamos en manejabilidad lo perdemos en información, con lo que los resultados serán aproximados.

Agrupar en intervalos de clase consiste en agrupar los datos en un número relativamente pequeño de intervalos que cumplan:

No se superpongan entre sí, de forma que no exista ambigüedad con respecto a la clase a que pertenece una observación particular.

Cubran todo el rango de valores que tenemos en la muestra.

Llamaremos:

- A las fronteras del intervalo, **límites inferior y superior** de la clase y los denotaremos por $\mathbf{L_{i-1}, L_i}$.
- **Marca de clase (c_i)** al punto medio del intervalo, es decir, al promedio aritmético entre el límite inferior y superior: $c_i = \frac{L_i + L_{i-1}}{2}$ .Es el valor que tomamos como representativo.
- **Amplitud (a_i)** a la diferencia entre el extremo superior e inferior: $a_i = L_i - L_{i-1}$.
- Al número de observaciones de una clase se le llama **frecuencia de clase (n_i),** si dividimos esta frecuencia por el número total de observaciones, se llama **frecuencia relativa de clase (f_i)**, y del mismo modo que lo hacíamos para datos sin agrupar definiríamos $\mathbf{N_i}$**, y** $\mathbf{F_i}$**.**

NOTA: <u>COMO CONSTRUIR UNA DISTRIBUCIÓN DE FRECUENCIAS AGRUPADA EN INTERVALOS</u>

1. Empezamos determinando el **recorrido de la variable o rango** de valores que tenemos en la muestra. Se define como la diferencia entre el mayor y el menor valor de la variable. $\mathbf{Re = x_k - x_1}$

2. Número de clases: depende del tamaño de la muestra. Para muestras de tamaño moderado, *n* <50, se suele elegir un número de clases igual a $\sqrt{n}$, o bien se usa la fórmula de Sturtges, (se toma el resultado de calcular el logaritmo de *n,* dividir por el logaritmo de 2 y sumar 1: $m = \frac{\log(n)}{\log(2)} + 1$); en general el número de clases no debe sobrepasar de 15 o 20, en casos de muestras muy grandes.

3. Determinamos la amplitud de lo intervalos. Es más cómodo que la amplitud de todas las clases sea la misma (siempre que sea posible), si es $a_i = \frac{Re}{n^a \, de \, int \, ervalos}$

NOTA: Tomaremos como regla, a no ser que se indique lo contrario, coger el intervalo cerrado por la izquierda y abierto por la derecha.

<u>Ejemplo 2</u>:

El menor valor es 3.3 y el mayor 6.1, la diferencia es 2.8 y por tanto Re=2.8.

N=40, cogemos 6 clases.

a=2.8/6=0.467.

Como la amplitud es un número con muchos decimales, los intervalos quedarán poco clarificadores, podemos hacer lo siguiente :Para que los intervalos nos queden con amplitud 0.5 cogeremos como primer valor 3.25 en vez de 3.3 y como último 6.25 en vez de 6.1 de esta manera: Re=6.25-3.25=3 y amplitud= 3/6=0.5.

Así pues una posible tabla sería:

$[L_{i-1}, L_i)$	**c_i**	**n_i**	**f_i**	**N_i**	**F_i**
[3.25, 3.75)	3.5	3	0.075	3	0.075
[3.75, 4.25)	4	8	0.2	11	0.275
[4.25, 4.75)	4.5	14	0.35	25	0.625
[4.75, 5.25)	5	6	0.15	31	0.775
[5.25, 5.75)	5.5	4	0.1	35	0.875
[5.75, 6.25)	6	5	0.125	40	1

¿Cuantos hoteles tienen un precio entre 3.25 y 3.75? 3

¿Cuantos hoteles tienen un precio superior a 4.75? 15

¿Qué porcentaje de hoteles cuestan como mucho 4.25? 27.5 %

3. MÉTODOS GRÁFICOS

La forma de la distribución de frecuencias se percibe más rápidamente y quizás se retiene durante más tiempo en la memoria si la representamos gráficamente.

3.1 FRECUENCIAS NO ACUMULADAS

DIAGRAMA DE BARRAS: Es la representación gráfica usual para variables cuantitativas sin agrupar o para variables cualitativas. En el eje de ordenadas representamos los diferentes valores de la variable (x_i). Sobre cada valor levantamos una barra de altura igual a la frecuencia (absoluta o relativa).

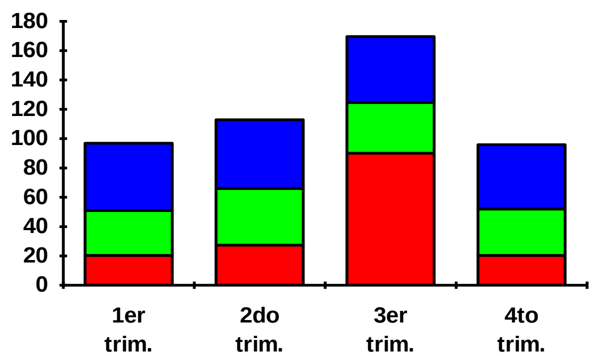


DIAGRAMA DE SECTORES O DE PASTEL: Es el más usual en variables cualitativas. Se representan mediante círculos. A cada valor de la variable se le asocia el sector circular proporcional a su frecuencia.

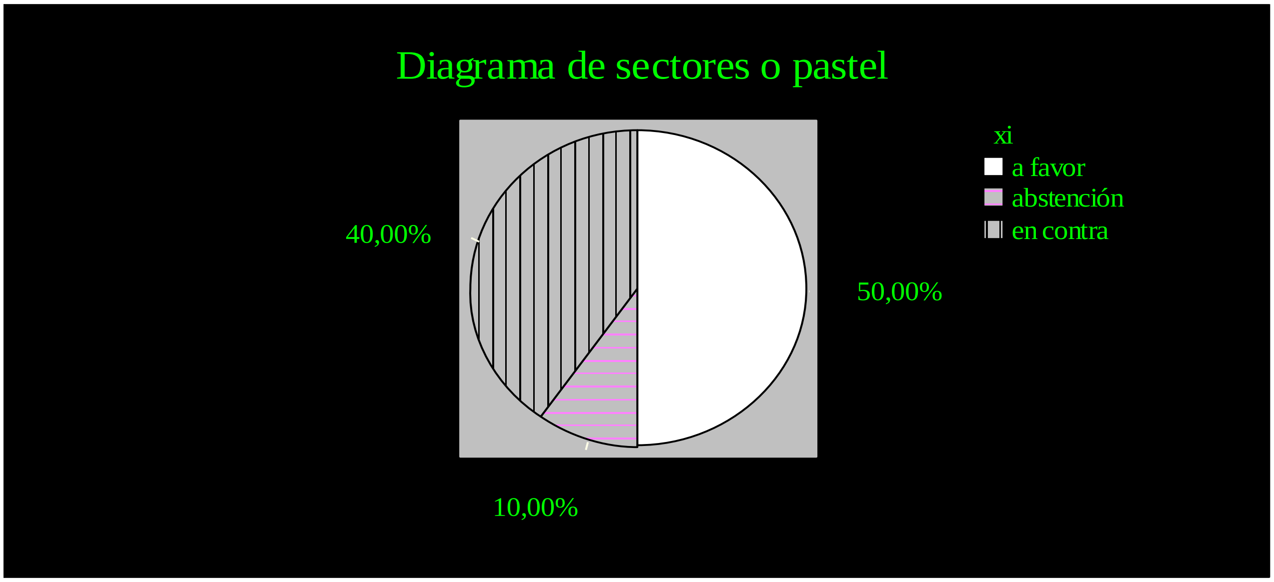
Para hallar el ángulo usamos la siguiente proporción: al tener una circunferencia 360º , el cociente entre la frecuencia absoluta (o relativa) total y la frecuencia absoluta (o relativa) que queramos representar será igual al cociente entre los 360º de la circunferencia y el ángulo a determinar, así:

$$\frac{n}{n_i} = \frac{360^{o}}{\alpha} \qquad\qquad \frac{1}{f_i} = \frac{360^{o}}{\alpha}$$

donde α es el ángulo a determinar.

<u>Ejemplo 3:</u> Los siguientes datos corresponden a una encuesta referente a elecciones locales de un partido político:

x_i	f_i
favor	0.5
en contra 0.4	
0.4	
abstención	0.1

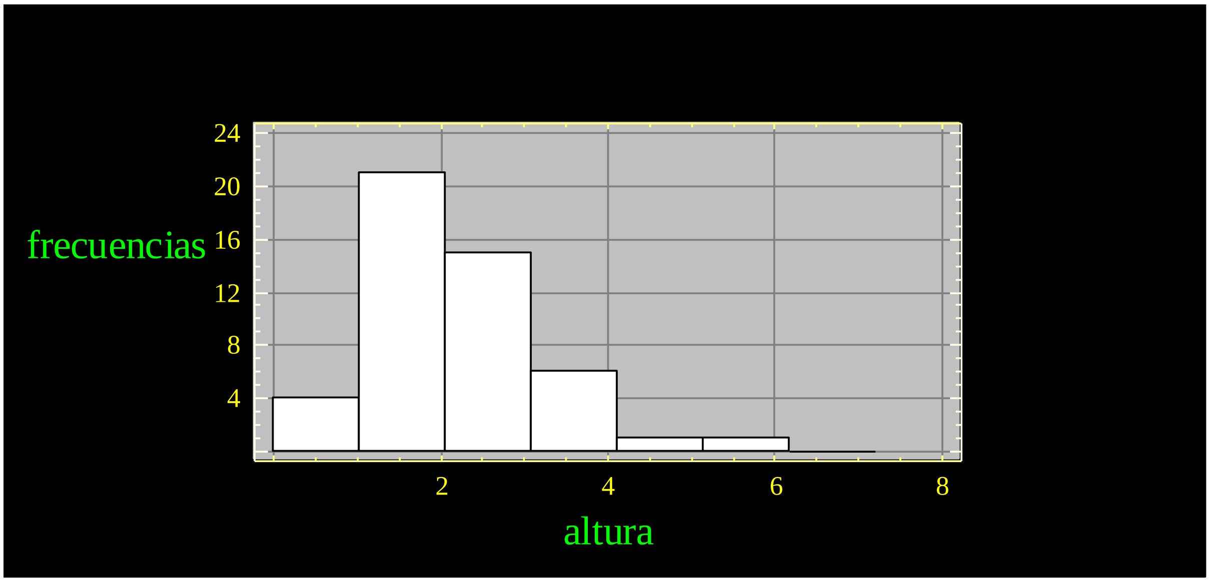


PICTOGRAMA: Se usa también para variables cualitativas, expresan con dibujos alusivos al tema de estudio las frecuencias de las modalidades de la variable. Estos gráficos se hacen representando en diferentes escalas el mismo dibujo. La escala de los dibujos tiene que ser tal que el área de cada uno de ellos sea proporcional a la frecuencia de la modalidad que representa.

Ejemplo 4: Ante un estudio sobre un tema concreto, buscaríamos un dibujo, (como el siguiente), decidiríamos el tamaño del área correspondiente a un valor y a partir de él, y proporcionalmente, asignaríamos al mismo dibujo el tamaño de área que explicara su frecuencia.

HISTOGRAMA: Es la representación gráfica equivalente al diagrama de barras para datos agrupados, en el eje de ordenadas representarnos las clases y levantarnos sobre cada clase rectángulos unidos entre sí de altura igual a la frecuencia de la clase (absolutas o relativas)

Ejemplo:

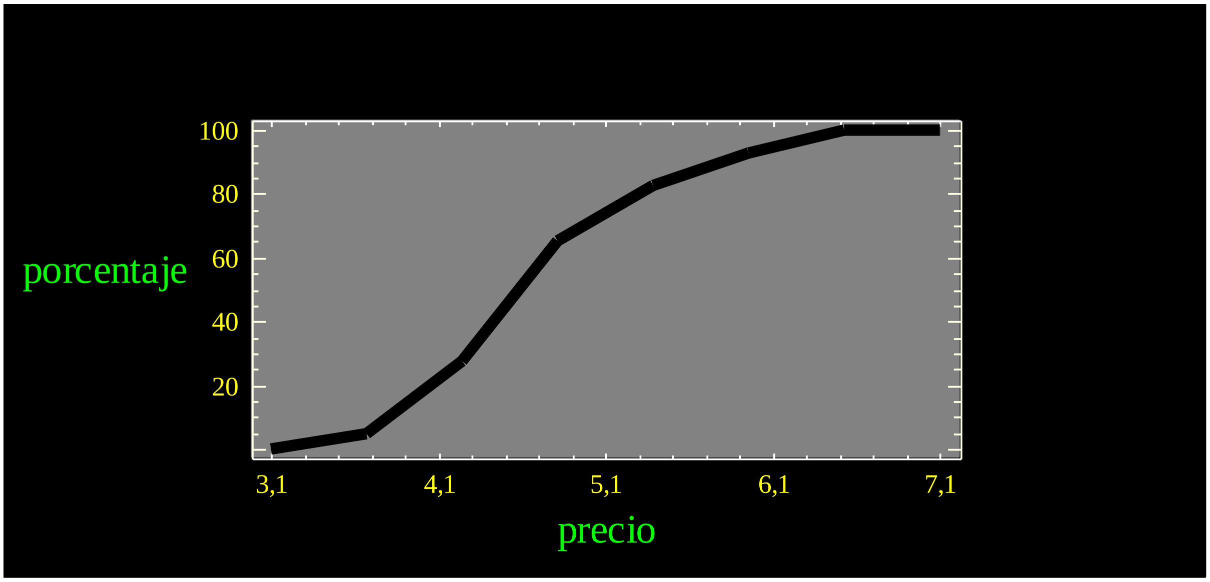


El histograma o diagrama de barras proporcionan mucha información respecto a la estructura de los datos (y si la muestra es representativa de la población, respecto a la estructura de la población): el valor central de la distribución, su dispersión y la forma de la distribución. Cuando nos encontramos en distribuciones donde los intervalos no tienen la misma amplitud, las barras del histograma tienen que tener un área proporcional a la frecuencia que queramos representar

3.2. FRECUENCIAS ACUMULADAS

POLÍGONO DE FRECUENCIAS: Es la representación habitual para datos cuantitativos agrupados de las frecuencias acumuladas (absolutas o relativas), mediante puntos se representan las frecuencias en el eje de ordenadas y la marca de clase en el de abscisas. Después se unen estos puntos por trozos de rectas.

Ejemplo 2:



4 MEDIDAS DESCRIPTIVAS

Para datos cualitativos, la distribución de frecuencias proporciona un resumen conciso y completo de la muestra, pero para variables cuantitativas puede complementarse este resumen utilizando medidas descriptivas numéricas extraídas de los datos.

Las medidas descriptivas son valores numéricos calculados a partir de la muestra y que nos resumen la información contenida en ella. En la parte de inferencia estadística les llamaremos estadísticos.

4.1 MEDIDAS DE POSICIÓN

Nos dan el valor que ocupa una determinada 'posición" respecto al resto de la muestra.

4.1.1 MEDIDAS DE TENDENCIA CENTRAL

Nos dan un centro de la distribución de frecuencias, es un valor que se puede tomar como representativo de todos los datos. Hay diferentes caminos para definir el "centro" de las observaciones en un conjunto de datos. Por orden de importancia, son:

MEDIA ARITMÉTICA: (o simplemente media). Es el promedio aritmético de las observaciones, es decir, el cociente entre la suma de todos los datos y el número de ellos (Teniendo en cuenta que si un valor se repite hay que considerar estas repeticiones)

$$\bar{x} = \frac{\sum_i x_i n_i}{n}$$

Si los datos están agrupados utilizamos las marcas de clase, es decir c_i en vez de x_i.

Es la medida de centralización más importante.

Ejemplo 1: $x = \frac{0*2+1*4+...+6*1}{50}$ =2.52

Ejemplo 2: 4.6875

PROPIEDADES

1. La suma de las diferencias de los valores de la variable y la media es cero.

$$\sum_i (x_i - \bar{x}) n_i = 0$$

2. La suma de las desviaciones al cuadrado de los valores de la variable respecto a una constante k cualquiera, se hace mínima cuando esa constante es la media. Es decir:

$$\sum_i (x_i - \bar{x})^2 n_i \le \sum_i (x_i - k)^2 n_i$$, para cualquier constante k.

MEDIANA (Me): es el valor que separa por la mitad las observaciones ordenadas de menor a mayor, de tal forma que el 50% de estas son menores que la mediana y el otro 50% son mayores. Si el número de datos es impar la mediana será el valor central, si es par tomaremos como mediana la media aritmética de los dos valores centrales.

Distinguiremos entre distribuciones no agrupadas y distribuciones agrupadas:

DISTRIBUCIONES NO AGRUPADAS:

- Calculamos n/2.
- Se busca en la tabla $N_{i-1} < n/2 < N_i$ (es decir aquel valor cuya frecuencia acumulada más se acerca a n/2 por arriba).

-Si $n/2 < N_i$ la mediana es aquel valor de la variable cuya frecuencia cumulada es N_i es decir: $Me = x_i$ tal que $n/2 < N_i$

-Si $n/2 = N_i$ la mediana será la media aritmética de aquellos valores cuya frecuencia acumulada es N_i y N_{i+1} respectivamente, es decir: $Me = (x_i + x_{i+1})/2$ tal que $N_i = n/2$

<u>Ejemplo 1</u>:

n=50

N/2=25

$N_2 = 6 < 25 < 27 = N_3$

Como *25*$< N_3 = 27$ entonces $Me = x_3 = 2$

DISTRIBUCIONES AGRUPADAS

- Se calcula n/2.
- Se busca en la tabla el intervalo, $[L_{i-1}, L_i)$, que cumple $N_{i-1} < n/2 < N_i$ (a este intervalo lo llamamos intervalo mediano).
- A continuación para encontrar la mediana, aplicaremos la siguiente fórmula:

$$Me = L_{i-1} + \frac{\left(\frac{n}{2} - N_{i-1}\right)a_i}{n_i}$$

El razonamiento es el siguiente: La frecuencia acumulada hasta el intervalo anterior al mediano es N_{i-1}; para llegar a la mitad de los datos, es decir, n/2 necesitamos tomar $n/2 - N_{i-1}$ del intervalo mediano, el cual tiene n_i datos repartidos en una amplitud a_i; como a cada dato le corresponde una longitud a_i / n_i, a los $n/2 - N_{i-1}$ datos les corresponderá

$$\frac{\left(\frac{n}{2} - N_{i-1}\right)a_i}{n_i}$$

Ejemplo 2:

n=40

N/2=20

$N_2=11<20<25=N_3$

El intervalo mediano es el intervalo $[L_{i-1}, L_i)= [4.25, 4.75)$ con lo que

$$Me = 4.25 + \frac{\left(\frac{40}{2} - 11\right)0.5}{14} = 4.57$$

PROPIEDAD: La mediana hace mínima la suma de todas las desviaciones absolutas de los valores de la variable respecto a una constante k cualquiera. Es decir,

$$\sum_i |x_i - Me| n_i \leq \sum_i |x_i - k|\, n_i$$

Para cualquier constante *k.*

MODA (M_0) es el valor de la variable que más veces se repite, es decir, aquella cuya frecuencia absoluta es mayor. No tiene por qué ser única. Distinguiremos:

DISTRIBUCIONES NO AGRUPADAS

Simplemente observamos en la columna de las frecuencias absolutas y aquel o aquellos valores (no tiene por qué ser única) de la variable a los que corresponde la mayor frecuencia será la moda. Cuando encontramos dos modas decimos que es una distribución bimodal, tres, trimodal, etc.

Ejemplo1 $M_0=2$

DISTRIBUCIONES AGRUPADAS

Es importante distinguir aquí también entre intervalos de igual amplitud, o distribuciones de frecuencias donde los intervalos no tengan la misma amplitud.

Intervalos de igual amplitud.

Observando las frecuencias absolutas, determinamos el intervalo con mayor frecuencia $[L_{i-1},L_i)$, a este intervalo le llamaremos intervalo modal.

A continuación para encontrar la moda aplicamos la siguiente fórmula:

$$Mo = L_{i-1} + \frac{n_{i+1}}{n_{i-1} + n_{i+1}} a_i$$

El razonamiento es el siguiente: Consideramos los intervalos anterior y posterior al modal, con frecuencias n_i y n_{i-1}. Si estas frecuencias son iguales, la moda sería el centro del intervalo modal, en caso contrario, la moda estaría más cerca de aquel intervalo contiguo cuya frecuencia es mayor, es decir, las distancias de la moda al intervalo contiguo son inversamente proporcionales a las frecuencias de dichos intervalos. Como consecuencia $M_0=L_{i-1}+m$ con:

$$\frac{m}{a_i - m} = \frac{n_{i+1}}{n_{i-1}}$$

Despejando *m* y sustituyendo obtenemos la fórmula anterior.

Ejemplo 2: El intervalo modal es $[L_{i-1}, li)= [4.25, 4.75)$, la moda será:

$$Mo = 4.25 + \frac{6}{8+6} 0.5 = 4.46$$

Intervalos de distinta amplitud.

Tendremos que hallar en primer lugar la densidad de frecuencia de cada intervalo que se define como: $d_i = n_i / a_i$.

El intervalo modal $[L_{i-1}, L_i)$ será ahora el intervalo con mayor densidad de frecuencia y para hallar la moda de nuevo aplicamos la fórmula anterior pero sustituyendo las frecuencias por las densidades de frecuencia:

$$Mo = L_{i-1} + \frac{d_{i+1}}{d_{i-1} + d_{i+1}} a_i$$

NOTA: COMPARACIÓN ENTRE MEDIA, MODA Y MEDIANA

Estas tres medidas de tendencia central son las más importantes y las más usuales. ¿Cuándo utilizamos una u otra?

- La media es la mejor por que utiliza toda la información, es decir, tiene en consideración todos los valores de la distribución, tiene también como ventaja que es única. Como desventaja más importante está el hecho de que es muy sensible a la presentación de datos anómalos o atípicos que hacen que la media se desplace hacia ellos y como consecuencia no es recomendable usar la media en estos casos. Otra desventaja es que puede no coincidir con uno de los valores de la variable.

- La mediana utiliza menos información que la media puesto que no depende de los valores de la variable sino del orden que ocupa. Por este motivo tiene la ventaja de no estar afectada por observaciones extremas. La mediana la utilizaremos cuando la media falle. Otra ventaja frente a la media es que es un valor de la variable.

- La moda es la que menos información maneja y por tanto la peor. Tiene la ventaja de que puede calcularse incluso para datos cualitativos. Otra desventaja es que no es única.

Si la distribución es simétrica y campaniforme coinciden. En el caso de distribuciones campaniformes, la mediana está con frecuencia entre la media y la moda (algo más cerca de la media). La siguiente relación nos permite calcular una de estas medidas de centralización en función de las otras:

$$M_O \approx 3Me - 2\bar{x}$$

Las siguientes medidas de centralización tienen un significado estadístico menos intuitivo y se utilizan en situaciones más específicas:

MEDIA GEOMÉTRICA (G) Se define como la raíz n-ésima del producto de los *n* datos. Así:

$$G = \sqrt[n]{\prod_i x_i^{n_i}}$$

PROPIEDAD: El logaritmo de la media geométrica es igual a la media aritmética de los logaritmos de los valores de la variable.

La media geométrica se suele emplear para promediar porcentajes, tasas y números índices.

MEDIA ARMÓNICA (H) Se define como el recíproco de la media aritmética de los recíprocos de los datos:

$$H = \frac{n}{\sum_i \frac{1}{x_i} n_i}$$

Se suele utilizar para promediar velocidades, rendimientos y en general magnitudes expresadas en términos relativos.

NOTA: Si los datos están agrupados, para calcular las medidas anteriores utilizamos las marcas de clase, es decir x_i indicará el punto medio del intervalo.

La relación existente entre la media, la media geométrica, y la media armónica sería:

$$H \leq G \leq \overline{x}$$

4.1.2 MEDIDAS DE POSICIÓN NO CENTRALES: CUANTILES

Los cuantiles son valores de la distribución que la dividen en partes iguales, es decir, en intervalos, que comprenden el mismo número de valores. Los más usados son los cuartiles, los deciles y los percentiles

PERCENTILES. Son 99 valores que dividen en cien partes iguales el conjunto de datos ordenados.

El percentil de orden p (P_p) es el menor valor superior al p% de los datos (ordenados de menor a mayor los datos, deja el p% de datos por delante). La forma más cómoda de calcularlos es a partir de las frecuencias acumuladas:

DISTRIBUCIONES NO AGRUPADAS: El percentil p es aquel valor cuya frecuencia acumulada más se acerca por arriba al p% de n, es decir:

$P_P = X_i$ tal que $N_{i-1} < pn/100 \leq N_i$

DISTRIBUCIONES AGRUPADAS: Usamos la misma idea que cuando calculábamos la mediana, buscamos en primer lugar el intervalo $[L_{i-1}, L_i)$ cuya frecuencia acumulada sea $N_{i-1} < pn/100 \leq N_i$, a continuación para hallar el percentil aplicamos la siguiente fórmula:

$$P_p = L_{i-1} + \frac{\left(\frac{pn}{100} - N_{i-1}\right)a_i}{n_i}$$

CUARTILES *(C_1)* son los tres valores que dividen al conjunto de datos ordenados en cuatro partes iguales, son un caso particular de los percentiles:

$C_1=P_{25}$

$C_2=P_{50}$

$C_3=P_{75}$.

<u>*Ejemplo 1:*</u>

$$C_1 = P_{25} \qquad N_i = E\left(\frac{25.50}{100}\right) = 10 \qquad C_1 = 2$$

$$C_2 = P_{50} \qquad N_i = E\left(\frac{50.50}{100}\right) = 20 \qquad C_2 = 2$$

$$C_3 = P_{75} \qquad N_i = E\left(\frac{75.50}{100}\right) = 30 \qquad C_3 = 3$$

DECILES (D_i): Son los nueve valores que dividen al conjunto de datos ordenados en diez partes iguales, son también un caso particular de los percentiles.

$D_1=P_{10}$

$D_2=P_{20}$

....

$D_9=P_{90}$

NOTA: La Mediana también es un caso particular de percentil: $Me=P_{50}$

4.1.3 MOMENTOS

Los momentos de una distribución se definen como una generalización de la media. Como veremos serán la base para describir algunas características importantes de la distribución de frecuencias. Pero lo más importante de ellos, es que caracterizan a la distribución de frecuencias, es decir, dos distribuciones son iguales si tienen todos sus momentos iguales, y son tanto más parecidas cuanto mayor sea el número de momentos iguales que tengan.

MOMENTOS RESPECTO AL ORIGEN: Se define el momento de orden r (a_r) (r=0,1,2) respecto al origen como la media aritmética de las potencias r-ésimas de los datos:

$$a_r = \frac{\sum_i x_i^r n_i}{n}$$

CASOS PARTICULARES:

$$a_0 = \frac{\sum_i x_i^0 n_i}{n} = \frac{n}{n} = 1$$

$$a_1 = \frac{\sum_i x_i n_i}{n} = \overline{x}$$

MOMENTOS CENTRALES O RESPECTO A LA MEDIA: Se define el momento de orden r (m_r) (r=0,1,2) respecto a la media como:

$$m_r = \frac{\sum_i (x_i - \overline{x})^r n_i}{n}$$

CASOS PARTICULARES:

$$m_O = \frac{\sum_I (x_I - \bar{x})^0 n_I}{n} = \frac{n}{n} = 1$$

$$m_1 = \frac{\sum_I (x_I - \bar{x})\ n_I}{n} = \bar{x} - \bar{x} = 0$$

4.2 MEDIDAS DE DISPERSIÓN

Las medidas de tendencia central tenían como objetivo el sintetizar los datos en un valor representativo, las medidas de dispersión nos dirán hasta qué punto estas medidas de tendencia central son representativas como síntesis de la información. Las medidas de dispersión cuantifican la separación, la dispersión, la variabilidad de los valores de la distribución respecto al valor central.

Distinguiremos entre medidas de dispersión absolutas, que no son comparables entre diferentes muestras y las relativas que nos permitirán comparar varias muestras.

4.2.1 MEDIDAS DE DISPERSIÓN ABSOLUTAS

Por orden de importancia tenemos:

VARIANZA (s^2) es el promedio del cuadrado de las distancias entre cada observación y la media aritmética del conjunto de observaciones

$$s^2 = \frac{\sum_i (x_i - \bar{x})^2 n_i}{n}$$ Si los datos están agrupados utilizamos las marcas de clase, es decir C_i en vez de X_i.

En el caso extremo en que todas las observaciones fueran iguales, la media coincidiría con ese valor común y la varianza sería cero. En general, cuanto más dispersas sean las observaciones, mayores serán las diferencias dentro de los cuadrados y por tanto mayor será el valor de s^2.

NOTA: La varianza es el momento de orden 2 respecto a la media: $s^2 = m_2$.

PROPIEDADES:

1. La varianza nunca puede ser negativa, $s^2 > 0$.

2. Otra forma más sencilla de calcular la varianza es:

$$s^2 = \frac{\sum_i x_i^2 n_i}{n} - \overline{x}^2 = a_2 - (a_1)^2$$

Demostración:

$$s^2 = \frac{\sum_i (x_i - \overline{x})^2}{n} = \frac{\sum_i (x_i^2 - 2\overline{x}x_i + \overline{x}^2) n_i}{n} = \frac{\sum_i x_i^2 n_i}{n} - 2\overline{x}\frac{\sum_i x_i n_i}{n} + \frac{\overline{x}^2 \sum_i n_i}{n} =$$

$$= \frac{\sum_i x_i^2 n_i}{n} - 2\overline{x}\overline{x} + \frac{\overline{x}^2 n}{n} = \frac{\sum_i x_i^2 n_i}{n} - \overline{x}^2$$

<u>Ejemplo 1:</u>

Usaremos la propiedad 2

x_i	n_i	x_i^2	$n_i x_i^2$
0	2	0	0
1	4	1	4
2	21	4	84
3	15	9	135
4	6	16	96
5	1	25	25
6	1	36	36

50 380

s^2 = (380/50)-6.35 = 1.25

o directamente:

$s^2 = (0^2 *2 + 1^2 *4+........+6^2 *1)/50 \ -2.52^2 = (380/50\,)-6.35 = 1.25$

Otras medidas de dispersión directamente relacionadas con la variaza son las dos siguientes.

DESVIACIÓN TÍPICA (S). La varianza vendría dada por las mismas unidades que la variable pero al cuadrado, para evitar este problema podemos usar como medida de dispersión la desviación típica que se define como la raíz cuadrada positiva de la varianza

$$s = \sqrt{s^2}$$

PROPIEDAD: Se observa a partir de la definición que $s \geq 0$

<u>Ejemplo 1</u>: s=1.12

CUASI-VARIANZA (s^{*2}) Se define de forma muy parecida a la varianza pero dividiendo por n-1.

$$s^{*2} = \frac{\sum_i (x_i - \bar{x})^2 n_i}{n-1} = \frac{n}{n-1} s^2$$

<u>Ejemplo 1</u>: s^{*2}= 1.2

DESVIACIÓN MEDIA RESPECTO A LA MEDIA ($D_{\bar{x}}$) Se define como el promedio de las desviaciones en valor absoluto respecto a la media aritmética:

$$D_{\bar{x}} = \frac{\sum_i |x_i - \bar{x}| \, n_i}{n}$$

Si toma valores grandes significa que los valores de la variable se distribuirán en valores alejados de la media.

Ejemplo 1:

x_i	n_i	$\lvert x_i - \bar{x} \rvert$	$n_i \lvert x_i - \bar{x} \rvert$
0	2	2.52	5.04
1	4	1.52	6.04
2	21	0.52	10.92
3	15	0.48	7.2
4	6	1.48	8.88
5	1	2.48	2.48
6	1	3.48	3.48
			44.38

$D_{\bar{x}} = 44.38/50 = 1.77$

DESVIACIÓN MEDIA RESPECTO A LA MEDIANA (D_e) Se define como el promedio de las desviaciones en valor absoluto respecto a la mediana:

$$D_{Me} = \frac{\sum_i |x_i - Me| n_i}{n}$$

Si D_{Me} *es* grande los valores están dispersos respecto de la mediana.

Ejemplo 1:

x_i	n_i	$\lvert x_i - \overline{M}e \rvert$	$n_i \lvert x_i - \overline{M}e \rvert$
0	2	2	4
1	4	1	4
2	21	0	0
3	15	1	15
4	6	2	12
5	1	3	3
6	1	4	4
			42

D_{Me} = 42/50 = 0.84

RECORRIDO O RANGO MUESTRAL (R_e). Es la diferencia entre el valor de las observaciones mayor y el menor. $R_e = x_{max} - x_{min}$

<u>Ejemplo 1</u>: R_e = 6-1 = 5

RECORRIDO INTERCUARTíLICO (RQ). *Es* la diferencia entre el primer y el tercer cuartil.

$RQ = C_3 - C_1$

<u>Ejemplo 1:</u> RQ = 3-2 =1

4.2.2 MEDIDAS DE DISPERSIÓN RELATIVAS

COEFICIENTE DE VARIACIÓN DE PEARSON: Cuando se quiere comparar el grado de dispersión de dos distribuciones que no vienen dadas en las mismas unidades o que las medias no son iguales se utiliza el coeficiente de variación de Pearson que se define como el cociente entre la desviación típica y el valor absoluto de la media aritmética

$$CV = \frac{s}{|\overline{x}|}$$

Al hacer el cociente eliminamos las unidades.

CV representa el número de veces que la desviación típica contiene a la media aritmética y por lo tanto cuanto mayor es CV mayor es la dispersión y menor la representatividad de la media.

Ejemplo 1: CV=l.12/2.52=0.44

4.3 OTRAS MEDIDAS DESCRIPTIVAS

4.3.1 TIPIFICACIÓN DE UNA DISTRIBUCIÓN DE FRECUENCIAS

Supongamos que hacemos la siguiente transformación a los datos:

$$z_i = \frac{x_i - \bar{x}}{s_x}$$

es decir, a cada valor de la variable le restamos la media y lo dividirnos por la desviación típica.

Se trata de una transformación lineal $z_i = a + bx_i$ con $a = \frac{-\bar{x}}{s_x}$ y $b = \frac{1}{s_x}$.

Usando las propiedades de la media y de la desviación típica que aparecen en el apartado 5 del tema es fácil demostrar que la nueva distribución de frecuencias tiene media aritmética cero y desviación típica 1. Diremos entonces que la muestra o la distribución de frecuencias está tipificada y a la transformación anterior se le llama tipificación.

4.3.2 MEDIDAS DE FORMA

Comparan la forma que tiene la representación gráfica, bien sea el histograma o el diagrama de barras de la distribución, con la distribución normal.

A: Medidas de *ASIMETRÍA*

Nos miden la simetría de la distribución. Supongamos que hemos representado gráficamente una distribución de frecuencias: tracemos una perpendicular al eje de las x por $\bar{x}$. Diremos que la distribución es simétrica si existe a ambos lados el mismo número de valores, equidistantes dos a dos y cada par de puntos equidistantes con la misma frecuencia.

COEFICIENTE DE ASIMETRÍA DE FISHER:

$$g_1 = \frac{\sum_i (x_i - \bar{x})^3 n_i}{ns^3} = \frac{m_3}{s^3}$$

Sí la distribución es simétrica en el denominador tendremos el mismo número de desviaciones positivas como negativas y por tanto $g_1 = 0$.

Si $g_1 > 0$ la distribución es asimétrica positiva o asimétrica a derechas.

Si $g_1 < 0$ la distribución es asimétrica negativa o asimétrica a izquierdas.

Elemplo 1:

x_i	n_i	$x_i - \bar{x}$	$(x_i - \bar{x})^3$	$n_i(x_i - \bar{x})^3$
0	2	-2.52	-16.003	-32.006
l	4	-1.52	-3.512	-14.047
2	21	-0.52	-0.141	-2.953
3	15	0.48	0.11	1.658
4	6	1.48	3.242	19.451
5	1	2.48	15.253	15.253
6	1	3.48	42.144	42.144
				29.5

$$g_1 = \frac{\sum_i (x_i - \bar{x})^3 n_i}{ns^3} = 0.42 > 0$$ luego asimétrica positiva.

COEFICIENTE DE ASIMETRÍA DE PEARSON: Es mucho más fácil de calcular que el anterior pero sólo es aplicable a aquellas distribuciones que tienen una sola moda y cuya distribución tiene forma de campana. Se define:

$$A_s = \frac{\bar{x} - M_o}{s}$$

Si la distribución es simétrica $\bar{x} = M_e$ y por tanto $A_s=0$. Si $A_s>0$ la distribución es asimétrica positiva. Si $A_s<0$ la distribución es asimétrica negativa.

Ejemplo 1:

$A_s = (2.52-2)/1.12=0.46$

B: Medidas de *APUNTAMIENTO O CURTOSIS*

Miden la mayor o menor cantidad de datos que se agrupan en torno a la moda. Solo tienen sentido en distribuciones campaniformes, es decir, unimodales simétricas o ligeramente asimétricas.

Si para valores próximos a la moda las frecuencias son más altas que en la distribución normal, la gráfica será muy apuntada en esa zona, y se dice que es de tipo leptocúrtico. Cuando son más bajas que en la distribución normal se dice que es de tipo platicúrtico. Cuando la distribución de frecuencias es igual de apuntada que la normal se dice que es mesocúrtica.

COEFICIENTE DE APUNTAMIENTO DE FISHER. Se define como:

$$g_2 = \frac{\sum_i (x_i - \bar{x})^4 n_i}{ns^4} - 3 = \frac{m_4}{s^4} - 3$$

- si $g_2>0$ leptocúrtica.
- si $g_2<0$ platicúrtica.
- si $g_2=0$ mesocúrtica o normal.

Ejemplo 1:

x_i	n_i	$x_i- \bar{x}$	$(x_i- \bar{x})^4$	$n_i(x_i- \bar{x})^4$
0	2	-2.52	40.327	80.655
1	4	-1.52	3.512	14.047
2	21	-0.52	0.141	2.953
3	15	0.48	0.11	1.658
4	6	1.48	3.242	19.451
5	1	2.48	15.253	15.253
6	1	3.48	42.144	42.144
				127.512

$$g_2 = \frac{\sum_i (x_i - \bar{x})^4 n_i}{ns^4} - 3 = 1.815 > 0 \quad \text{leptocúrtica.}$$

4.3.3 MEDIDAS DE CONCENTRACIÓN

Las medidas de concentración tratan de poner de manifiesto el mayor o menor grado de igualdad en el reparto total de los valores de la variable. Son por tanto, indicadores del grado de equidistribución de la variable. Estas medidas tienen especial aplicación a variables económicas (rentas, salarios, etc.).

Supongamos que tengamos n sujetos cuyos valores de la variable (rentas, salarios, etc.) son:

$$x_1 \leq x_2 \leq x_3 < ... < x_n$$

Nos interesa estudiar hasta qué punto la suma total de valores (rentas, salarios, etc.) esta equitativamente repartida.

Las dos situaciones extremas serian:

1. Concentración máxima: de los n sujetos, sólo uno percibe el total y los demás nada:

$$x_1 = x_2 = x_3 = ... = x_{n-1} = 0 \qquad y\ x_n \neq 0$$

2. Concentración mínima o equidistribución: todos tienen el mismo valor

$$x_1 = x_2 = x_3 = ... = x_{n-1} = x_n$$

NOTA: Hay que tener en cuenta que desde el punto de vista estadístico los términos dispersión y concentración no son opuestos, recordemos que el primero hacía referencia a la variabilidad de los datos con respecto al promedio, mientras que el segundo, como acabamos de definir, a la no equidad en el reparto de la suma total de la variable.

ÍNDICE DE CONCENTRACIÓN DE GINI (I_G) El índice de concentración de Gini se construye a partir de las siguientes cantidades:

1. Los productos $x_i n_i$ que nos indicarán el total percibido (renta total, ganancia total, etc.) por los n_i sujetos con valor (renta,...) x_i. A este producto le llamaremos riqueza del grupo y.

2. Las riquezas acumuladas de la variable (u_i), se calculan de la siguiente forma:

$u_1=x_1n_1$

$u_2=x_1n_1+x_2n_2$

$u_3=x_1n_1+x_2n_2+x_3n_3$

....

$u_k=x_1n_1+x_2n_2+...+u_kn_k$

3. Las riquezas acumuladas (u_i) las expresamos en tanto por ciento del total u_k.

$$q_i = \frac{u_i}{u_k} x100$$

4. Las frecuencias relativas acumuladas, expresadas en tanto por ciento:

$$p_i = \frac{N_i}{n} x100 = F_i \ x\ 100$$

A partir de todo esto se define el índice de concentración de Gini mediante la fórmula:

$$I_G = \frac{\sum_{i=1}^{k-1}(p_i - q_i)}{\sum_{i=1}^{k-1} p_i}$$

Podemos observar que:

a) Si $q_i = 0$, para i=1,2,...,k-1, y $q_k \neq 0$ entonces $I_G = \frac{\sum_{i=1}^{k-1} p_i}{\sum_{i=1}^{k-1} p_i} = 1$ y la concentración es máxima.

b) Si para cada i es $p_i=q_i$, $I_G=0$ y el reparto es equitativo, ya que cada porcentaje de individuos posee el mismo porcentaje de riqueza.

CURVA DE LORENZ Una forma de estudiar gráficamente la concentración es mediante la curva de Lorenz que se construye representado en el eje de abcisas el porcentaje de frecuencias acumuladas (p_i) y en el eje de ordenadas los porcentajes acumulados del total de la variable (q_i). Al unir estos puntos obtenemos la curva de Lorenz.

Como para $p_i = 0$, la gráfica pasa por el punto (0,0), y para $p_i = 100\%$ es $q_i = 100\%$, la gráfica pasa por los puntos O= (0,0) y P(100,100). Por otra parte, al ser $p_i \leq q_i$, por estar ordenados los datos en sucesión creciente, la gráfica está siempre situada por debajo de la diagonal del cuadrado o coincidente con ella. En el caso de existir reparto equitativo, es decir concentración mínima, la curva coincide con la diagonal (OB), pues en ese caso $p_i=q_i$. Si la concentración es máxima la curva de Lorenz estaría formada por los lados OA y OB.

Se demuestra que aproximadamente:

$$I_G = \frac{\text{Area entre la curva y la diagonal OB}}{\text{Area del triangulo OAB}}$$

NOTA: *COMPARACIÓN ENTRE LAS DOS MEDIDAS:*

Si bien el índice de Gini tiene la ventaja de resumir la información en una sola cifra y por tanto comparar más fácilmente que la curva de Lorenz, esta ventaja tiene como contrapartida el que dos distribuciones con aspecto muy diferente pueden tener el mismo índice de Gini.

Ejemplo 1:

x_i	n_i	$x_i n_i$	u_i	q_i	F_i	p_i	$p_i - q_i$
0	2	0	0	0	0.04	4	4
1	4	4	4	3.17	0.12	12	8.83
2	21	42	46	36.51	0.54	54	17.49
3	15	45	91	72.22	0.84	84	11.78
4	6	24	115	91.27	0.96	96	4.73
5	1	5	120	95.24	0.98	98	2.76
6	1	6	126	100	1	100	

I_G = 49.59 / 348 = 0.142 Lo que nos indica poca concentración.

5. TRANSFORMACIONES LINEALES

En este apartado veremos cómo quedan afectadas algunas de las medidas de una variable cuando le sumamos o multiplicamos alguna cantidad. Es decir, calculamos una transformación lineal de la variable original, y de la que obtenemos queremos saber cuánto vale su media, mediana, varianza y desviación típica.

5.1 EN LA MEDIA

1. Si a todos los valores de una variable les sumamos una constante *k,* la media aritmética queda aumentada en esa constante. (La media aritmética queda afectada por los cambios de origen).

Es decir, si $y_i = k + x_i$ entonces $\bar{y} = k + \bar{x}$

Dem:

$$\bar{y} = \frac{\sum_i y_i n_i}{n} = \frac{\sum_i (k + x_i)}{n} = \frac{k\sum_i n_i + \sum_i x_i n_i}{n} = \frac{kn}{n} + \frac{\sum_i x_i n_i}{n} = k + \bar{x}$$

2. Si todos los valores de una variable los multiplicamos por una constante *k,* su media aritmética queda multiplicada por la misma constante (La media aritmética queda afectada por los cambios de escala).

Es decir, si $y_i = k\,x_i$ entonces $\bar{y} = k\bar{x}$

3. Como corolario de las anteriores, si consideramos la transformación lineal $y_i=a+bx_i$ siendo a y b dos constantes cualesquiera, la nueva media aritmética quedaría:

$$\bar{y} = a + b\bar{x}$$

5.2 EN LA MEDIANA

1. Si a todos los valores de una variable les sumamos una constante *k,* la mediana queda aumentada en esa constante. Es decir, la mediana queda afectada por los cambios de origen.

Es decir, si $y_i=k+x_i$ entonces: $Me_y=k+Me_x$

2. Si todos los valores de una variable los multiplicamos por una constante k, su mediana queda multiplicada por la misma constante. Es decir, la mediana queda afectada por los cambios de escala.

Es decir, si $y_i = k\,x_i$ entonces $Me_y=kMe_x$

3. Como corolario de las anteriores, si consideramos la transformación lineal $y_i=a+bx_i$ siendo a y b dos constantes cualesquiera, la nueva mediana quedaría $Me_y=a+bMe_x$

5.3 EN LA VARIANZA

1. Si a todos los valores de una variable les sumamos una constante *k,* la varianza no varia. Es decir:

Si $y_i = k + x_i$ entonces $s_y^2 = s_x^2$

2. Si todos los valores de una variable los multiplicamos por una constante *k,* su varianza queda multiplicada por el cuadrado de la constante.

Si $y_i = kx_i$ entonces $s_y^2 = k^2\,s_x^2$

3. Como corolario de las anteriores, si consideramos la transformación lineal $Y_i=a+bx_i$ siendo a y b dos constantes cualesquiera, la nueva varianza quedaría

$$s_y^2 = b^2\,s_x^2$$

5.4 EN LA DESVIACIÓN TÍPICA

1. Si $y_i = k + x_i$ entonces $s_y = s_x$.

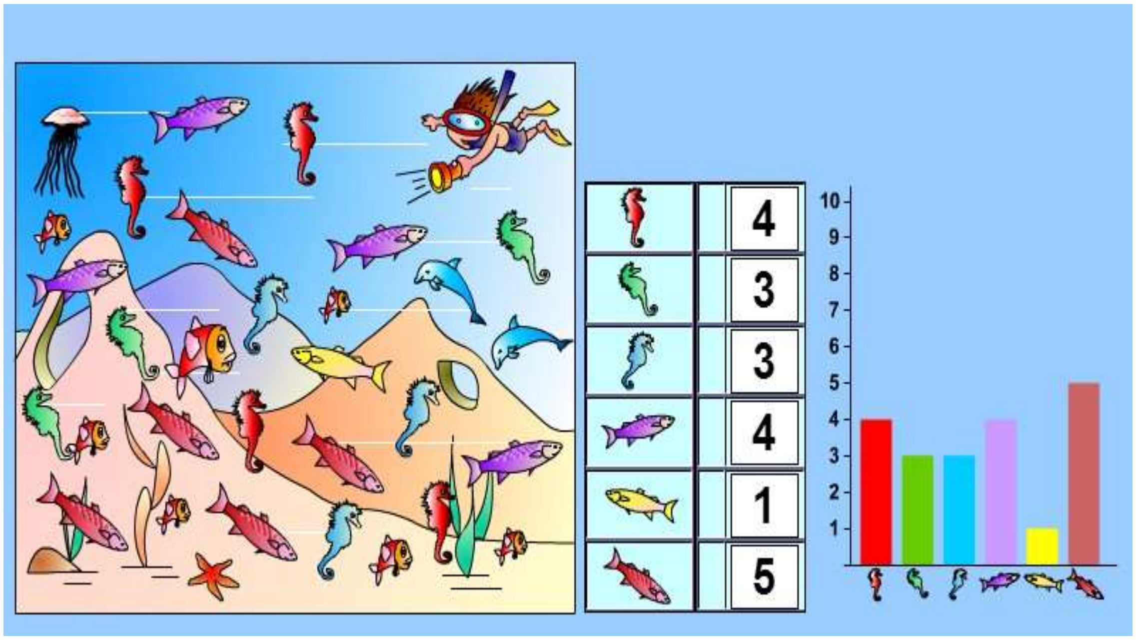
2. Si $y_i = k\, x_i$ entonces $s_y = |k| s_x$.

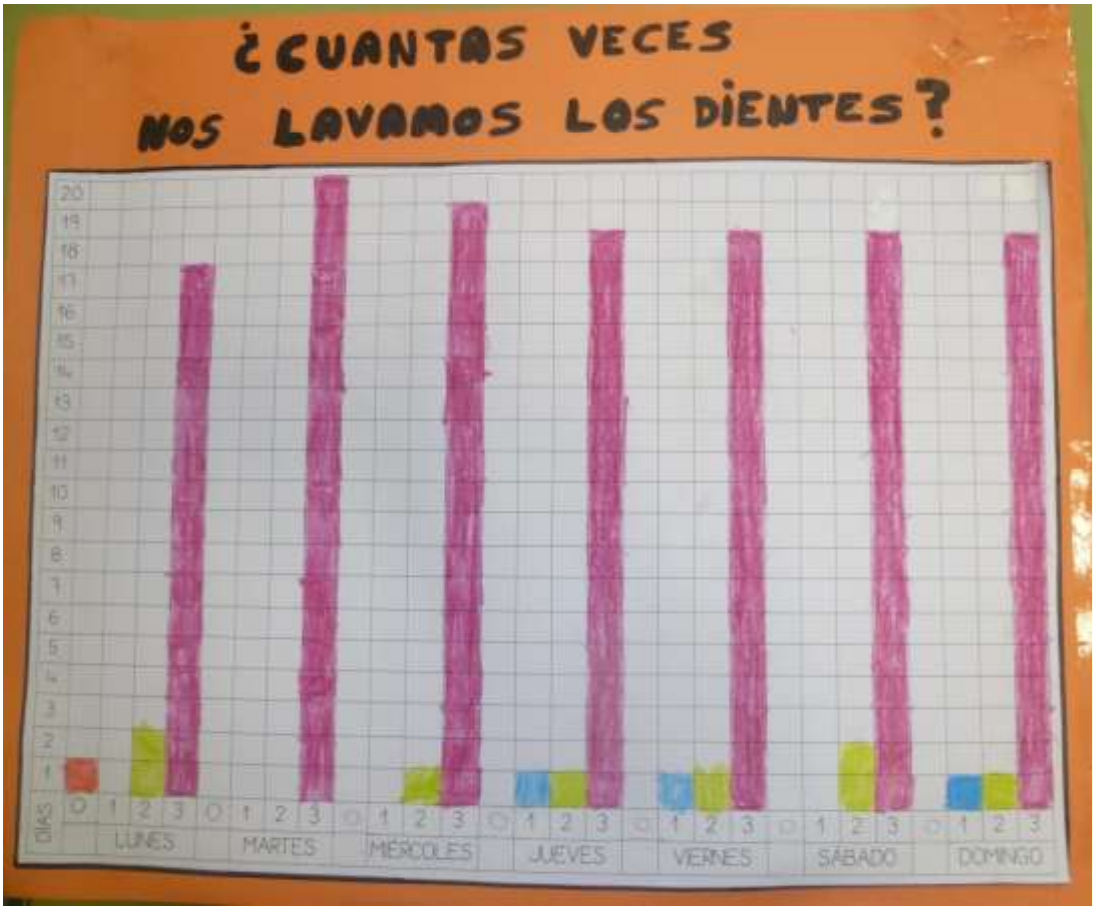
3. Si $y_i = a + bx_i$ entonces $s_y = |b| s_x$.

B. Base de Consulta

TÍTULO	AUTOR	EDICIÓN	AÑO	IDIOMA	EDITORIAL
Estadística y Muestreo	MARTINEZ Ciro	9na Ed. Colombia	1999	Español	Ecoe Ediciones
Estadística	SPIEGEL Murray R	3ra Ed. México	2002	Español	McGraw-Hill Interamericana
Estadística Aplicada a La Administración y Economía	KAZMIER Leonard J,	3ra Ed. México	1998	Español	McGraw-Hill Interamericana
Estadística para Admon y Economía	NEWBOLD Paul,	6ta Ed. España	2008	Español	Pearson Educación S.A
Estadística Descriptiva para las Organizaciones	QUINTERO Ramiro,	9na Ed. Colombia	2008	Español	Media Print Group S.A
Estadística Aplicada a los Negocios y a la Economía	WEBSTER Allen	4ta Ed. México	2000	Español	McGraw- Hill.
Probabilidad y Estadística	WALPOLE Ronald E	4ta Ed. México	1992	Español	McGraw-Hill Interamericana
Probabilidad y Estadística Aplicadas a la Ingeniería	MONTGOMERY Runger	2da. Ed	2002	Español	Editorial Limusa

C. Base práctica con ilustraciones





4. ESTRATEGIAS DE APRENDIZAJE

ESTRATEGIA DE APRENDIZAJE 1: Análisis y Planeación
Descripción: - Lecturas reflexivas del material proporcionado - Investigaciones en bibliotecas, Internet y de campo - Conversatorios mediante el Método Socrático - Liderar clases a cargo de cada uno de los estudiantes - Equipos de Investigación y de resolución de problemas - Dinámicas grupales

- Presentaciones apoyadas en el uso de las TIC's - Hojas de Ejercicios de Aplicación
Ambiente(s) requerido: Aula amplia con buena iluminación.
Material (es) requerido: Proyector, pizarrón, marcadores, materiales de apoyo para los estudiantes,
Docente: Con conocimiento de la materia.

5. ACTIVIDADES

- Controles de lectura
- Exposiciones
- Desarrollo de Talleres y actividades grupales en el aula
- Tareas en Plataforma
- Elaboración de ensayos
- Proyecto Final

6. EVIDENCIAS Y EVALUACIÓN

Tipo de Evidencia	**Descripción (de la evidencia)**
De conocimiento:	Definición del tema de investigación Lecturas que permitan el resumen y aplicación de definiciones en los respectivos ejercicios de aplicación.
Desempeño:	Trabajo colaborativo para aplicar talleres, resúmenes, subrayado, diagramas, esquemas y ejercicios aplicados en las clases y videos relacionados a cada tema socializado.
De Producto:	Proyecto final donde los estudiantes aplican las definiciones y ejercicios correspondientes trabajados en el módulo.

Criterios de Evaluación (Mínimo 5 Actividades por asignatura)	Actividad N.-1 AULA Exposición de la clase Elaboración de plan de estudios considerando aspectos de tiempo, actividades, espacios y recursos. Taller sobre trabajo Generalidades PLATAFORMA Investigación sobre la producción de trabajo intelectual
	Actividad N.- 2 AULA Exposición de la clase Talleres PLATAFORMA Estadística en Parvularia
	Actividad N.- 3 AULA Exposición de temas. Trabajo cooperativo aplicación de problemas y ejercicios de aplicación PLATAFORMA Investigación Medidas de tendencia central
	Actividad N.- 4 AULA Exposición de temas. Ejercicios de aplicación PLATAFORMA Foro de la estadística en los CDI
	Actividad N.- 5 AULA Presentación de Proyecto final

MSc. Daniel Shauri	**Ing. Alexis Benavides**	**Dr. Milton Altamirano**
Elaborado por: **(Docente)**	**Revisado Por:** **(Director)**	**Reportado Por:** **(Vicerrector)**

INSTITUTO TECNOLÓGICO SUPERIOR "JAPÓN"




www.itsjapon.edu.ec

Calle Mariete de Veintimilla y
Cuarta Transversal
2 356 368