

ESTADISTICA GENERAL

Apuntes

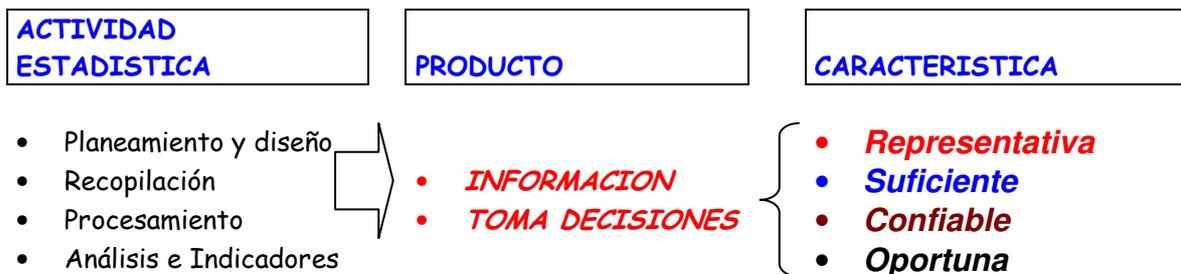


*Elaborados por: **Arturo Rubio Donet***

INTRODUCCIÓN

1. Estadística

Es la ciencia mediante la cual se desarrolla y aplica técnicas específicas para recopilar, consolidar y analizar información representativa, suficiente, confiable y oportuna con el fin de evaluar la incertidumbre en el proceso de la toma de decisiones. El proceso de un análisis estadístico comprende a dos campos complementarios: La estadística descriptiva y la estadística inferencial.



1.1 Estadística Descriptiva

Comprende a los procesos de consolidación, resumen y descripción de los datos recopilados. Estos resúmenes pueden tratarse de tablas, gráficos o de índices que permiten un análisis referido exclusivamente a los datos coleccionados y no se trata de generalizar ni de extrapolar para la totalidad de elementos (población) de donde provienen los datos recabados.

1.2 Estadística Inferencial

Incluye procedimientos que permiten la extrapolación y generalización sobre características que tipifican a todos los elementos de la población. Es decir, la inferencia estadística es el proceso de hacer afirmaciones o predicciones sobre toda la población tomando como base sólo a la información recabada a través de una muestra representativa. Asimismo, la estadística inferencial comprende a la teoría de estimación y a las pruebas de hipótesis.

2. Población o Universo Objetivo.

Es el conjunto (finito o infinito) de unidades de análisis que conforman a la población que se desea describir simplemente o sobre la cual se desea hacer inferencias y conocer sus parámetros característicos (promedio, totales proporciones, etc.).

2.1 Población Finita

Es el conjunto finito de unidades de análisis donde se puede identificar a un elemento inicial y/o a un elemento final. Por ejemplo, la población de hoteles de Lima Metropolitana, población de agencias de viajes existentes en la ciudad de Cajamarca. Turistas de nacionalidad alemana que ingresaron al Perú en el año 2000.

2.2 Población Infinita.

Conjunto infinito de elementos donde no se podría identificar a una unidad inicial ni a la unidad final. Ejemplo, la población de los peces del mar, lo árboles de la selva peruana, etc.

3. Muestra

Es un subconjunto de unidades de análisis extraída de la población objetivo buscando que las unidades seleccionadas sean representativas con el fin de permitir que a través de la información recabada con la muestra se realicen inferencias para toda la población.

3.1 Muestra No Probabilística

Corresponde a un subconjunto de observaciones elegidas siguiendo aquellos criterios de representatividad que fueran establecidos arbitrariamente por el investigador.

3.2 Muestra Probabilística

Comprende a las observaciones realizadas en unidades que han sido elegidas siguiendo un criterio probabilístico, esto es, a cada unidad de la población se asigna probabilidad conocida (puede ser igual) para estar incluida como parte de la muestra, por tanto, las unidades de la muestra son elegidas de la población respetando estrictamente estas probabilidades que les han sido asignadas. Las muestra probabilísticas permiten aplicar los métodos de la estadística inferencial y construir límites de confianza para las estimaciones de los parámetros que se desean estudiar. La representatividad de estas muestras se sustenta en el hecho que son las leyes de las probabilidades las que determinan si una unidad será incluida o no en la muestra.

4. Unidad de Análisis

Es el elemento que conforma a la población objetivo y de la cual se extrae la información que se desea estudiar.

Ejemplo: Se desea estudiar la capacidad hotelera en la ciudad de Lima y se define la unidad de análisis "hotel", en cada uno de los hoteles de la ciudad, se solicita la siguiente información:

- Número de habitaciones del hotel.
- Número de empleados del hotel.
- Total de clientes atendidos durante el mes de Julio.
- Ingresos totales en el mes de Julio.
- Tiene servicio de agencia de viaje (Sí =1; No = 0)
- Tiene servicio de restaurante (Sí =1; No = 0)

Recopilando la información especificada en los N hoteles de la ciudad de Lima se formaría la base de datos siguiente:

Unidad (Hotel) (i)	Numero habitaciones X_i	Numero Empleados Y_i	Clientes En Julio Z_i	Ingresos en Julio V_i	Agencia Viaje W_i	Servicio Restaurante R_i
1	55	12	200	14000	1	1
2	30	8	250	16000	0	1
3	20	6	100	7000	0	0
....						
....						
N	60	20	309	25000	0	1

5. Variable

Es todo factor o característica que en forma conjunta e integral conforma la información que se solicita a cada unidad de análisis. Cada variable tomará valores diferentes entre las unidades de análisis que se están investigando.

5.1 Variables Cualitativas

Son aquellas variables que expresan categorías o atributos y que por tanto su medida no tiene un carácter numérico, por su naturaleza, estas variables pueden ser:

NOMINALES: Sus valores sólo representan un atributo a manera de etiqueta y no contiene información sobre ordenamiento. Ejemplo: Sexo del cliente; Nacionalidad del entrevistado.

ORDINALES: Sus valores si representan un ordenamiento del atributo. Ejemplo: Grado de educación del entrevistado, Grado de satisfacción sobre la atención recibida por el cliente..

5.2 Variables Cuantitativas

Comprender aquellos conceptos que sí pueden ser expresados en forma numérica por que corresponde a criterios de cantidad. Estas son de tipo Discreto o de tipo Continuo

5.2.1 Variables Cuantitativa Discretas

Son variables que toman valores que se expresan en números enteros. Ejemplo:

- Número de habitaciones.
- Número de empleados.
- Total de clientes atendidos durante el mes de Julio.

Se aprecia que estas variables no pueden tomar valores fraccionarios.

5.2.2 Variables Cuantitativas Continuas

Son aquellas variables que sus cantidades se expresan con números reales, es decir, tienen parte fraccionaria. Ejemplo:

- Ingresos totales en el mes de Julio.
- Costo de servicio diario del hotel
- Monto pagado por el servicio de electricidad de un hotel en el mes de Julio

6. Observación:

Es un valor particular que toma la variable estudiada en la unidad de análisis i-ésima

7. Parámetro:

Es el indicador resultante de consolidar toda la información referida a la población que se está estudiando, por tanto corresponde a un resultado de una enumeración completa

$$\mu = \frac{1}{n} \sum_{i=1}^N X_i$$

Donde: X_i es el valor de la variable y N es el tamaño de la población

Ejemplo:

- Numero de habitaciones promedio en los hoteles de la ciudad de Lima.
- Número de empleados promedio de los hoteles de la ciudad de Lima.
- Promedio de clientes atendidos por los hoteles de Lima durante el mes de Julio.
- Ingresos promedio de los hoteles de la ciudad de Lima en el mes de Julio.

8. Estadístico:

Es el indicador resultante de consolidar la información obtenida sólo de una muestra representativa de unidades de análisis en la población. Es un valor que estima al parámetro

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n X_i$$

Donde: X_i es el valor de la variable y n es el tamaño de la muestra

Ejemplo: Se elige una muestra de los hoteles existente en la ciudad de Lima

- Numero de habitaciones promedio en la muestra de hoteles.
- Número de empleados promedio en la muestra de hoteles.
- Promedio de clientes atendidos en la muestra de hoteles.
- Ingresos promedio de los hoteles elegidos en la muestra.

ORGANIZACION Y PRESENTACION DE DATOS: **Tablas de Información Estadística**

Comprende a resúmenes y consolidaciones de informaciones dispersas que se efectúan con fines de realizar una presentación ordenada y clasificada de resultados de modo que se facilite su análisis e interpretación para la toma de decisiones. Las tablas de información resultan de identificar las necesidades de información por tanto es necesario que para su elaboración se consideren las características que puedan cumplir su objetivo final: Informar. La estructura general de una tabla de información estadística es la siguiente:

1. Número de cuadro

Es un código de identificación que permite citarlo cuantas veces sea necesario a lo largo de los comentarios de un informe.

2. Título

Corresponde a una descripción resumida de la información estadística que contiene, en general debería responder a las clásicas preguntas: *QUE, DONDE, COMO Y CUANDO*

- **Qué:** ¿ Qué información contiene el cuadro ?
- **Dónde:** ¿ A dónde corresponde la información tabulada ?
- **Cómo:** ¿ Cómo está arreglada o clasificada la información ?
- **Cuándo:** ¿ A qué período de tiempo se refiere la información ?

3. Encabezamiento

Es la descripción resumida sobre contenido de las filas y columnas que contiene el cuadro de información.

4. Cuerpo del cuadro

Comprende al contenido numérico de la información del cuadro

5. Notas de pié o llamadas

Anote las aclaraciones sobre términos, siglas, abreviaturas usadas. También precisar las unidades de medida y equivalencias que se están utilizando para expresar cantidades.

6. Fuente

Se debe anotar el nombre de la institución o de la investigación de donde se han tomado los datos individuales.

7. Elaboración

Se anotará el nombre de la unidad responsable de la formulación y presentación del tabulado de información estadística.

8. Fecha

Registre la fecha en que se realizó el cuadro para dejar constancia del evento y poder actualizar de ser pertinente

Ejemplo: Según información preliminar de la Oficina de Admisión del hotel CONFORT, en el mes de Febrero de 2006 el hotel atendió a los turistas extranjeros siguientes:

Turista	Nacionalidad	Sexo
1	Argentina (*)	Masculino
2	Argentina (*)	Masculino
3	Argentina	Masculino
4	Argentina	Femenino
5	Argentina	Femenino
6	Argentina	Femenino
7	Boliviana	Masculino
8	Boliviana	Masculino
9	Boliviana	Femenino
10	Chilena	Masculino
11	Chilena	Femenino

(*) El cliente no acreditó su nacionalidad

Elabore una tabla de resultados que consolide la información dispersa en esta base de datos construida por la Oficina de Admisión del hotel CONFORT

NUMERO Cuadro No 01

TITULO

¿Qué? Numero de turistas extranjeros que fueron atendidos

¿Dónde? en el hotel CONFORT

¿Cómo? según nacionalidad por sexo (1)

¿Cuándo? (Periodo: Febrero 2006)

Nacionalidad	Numero de Turistas Extranjeros		Sexo			
			Masculino		Femenino	
	Numero (fi)	% (hi)	Numero (fi)	% (hi)	Numero (fi)	% (hi)
Argentina (2)	6	54.5	3	50.0	3	60.0
Boliviana	3	27.3	2	33.3	1	20.0
Chilena	2	18.2	1	16.7	1	20.0
Total General	11	100.0	6	100.0	5	100.0

NOTAS DE PIE

(1) Cifras preliminares registradas al 01.03.2006

(2) Comprende a ciudadanos con nacionalización no acreditada

FUENTE

Oficina de Admisión del hotel CONFORT

RESPONSABLE

Elaborado por: Asesores y Consultores S.A.

FECHA

01.03.2006

ORGANIZACIÓN Y PRESENTACIÓN GRÁFICA DE DATOS

La presentación de datos a través de tablas estadísticas es una actividad importante dentro de los sistemas de información gerencial, estas se fortalecen significativamente cuando se la acompañan con gráficos descriptivos e ilustrativos. En el contexto de los sistemas de información, en más de una oportunidad se encontrará que un buen gráfico resume y expresa mucho más que párrafos completos de comentarios e interpretaciones literales.

1. Organización de Variables Cualitativas

Comprende la presentación gráfica de conceptos cualitativos y/o atributos que se registran para las unidades de análisis. Por ejemplo, el número de turistas que registraron su ingreso por el aeropuerto de la ciudad de MODERNA durante el pasado mes de Enero se clasifica según su nacionalidad y se obtiene la tabla siguiente:

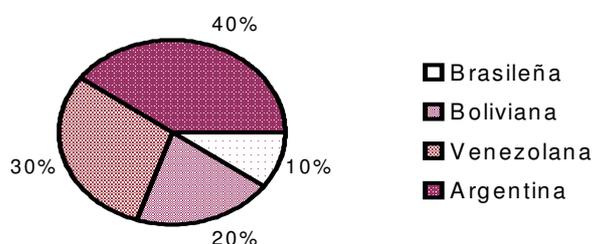
Número de turistas que registraron su ingreso a la ciudad MODERNA durante el mes de Enero según nacionalidad

NACIONALIDAD	NUMERO DE TURISTAS (fi)	FRECUENCIA RELATIVA (hi)
Argentina	20	0.40
Boliviana	10	0.20
Brasileña	5	0.10
Venezolana	15	0.30
TOTAL	50	1.00

Fuente: Oficina de Estadística del Municipio

Considerando las frecuencias relativas como una estructura porcentual, el área total de un círculo puede fraccionarse en estos componentes, con el fin de representar la composición estructural del número de turistas que visitan la ciudad, por tanto, considerando que un círculo comprende a 360 grados:

Utilizando un gráfico de círculo la representación gráfica de estos resultados podría ser:



Estos gráficos pueden ser mejorados de acuerdo con la imaginación y necesidades de presentación que cada sistema de información requiera.

2. Organización de Variables Discretas.

Comprende clasificaciones de variables que sólo toman valores enteros, por tanto, las unidades de análisis se ordenan de acuerdo con sus propios valores. Por ejemplo, el hotel "PANAMONTE" durante los últimos 50 días el número de clientes que atendió fue:

{ 7, 2, 4, 3, 6, 3, 3, 3, 4, 8, 7, 3, 7, 4, 6, 2, 5, 2, 3, 7, 6, 2, 6, 4, 6, 4, 6, 8, 2, 7, 5, 3, 5, 5, 5, 5, 4, 5, 5, 8, 2, 4, 5, 5, 4, 5, 3, 4, 6 y 8 }

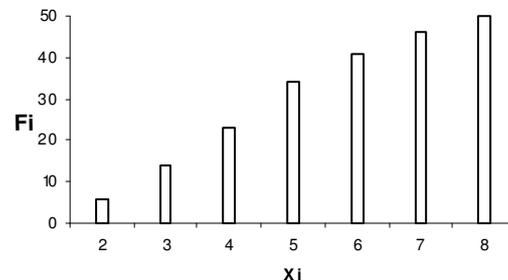
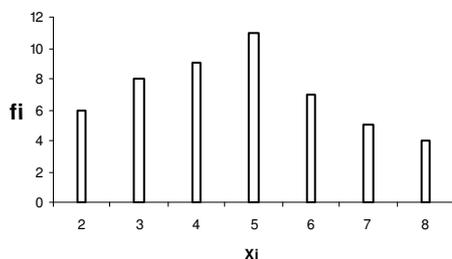
Esta información diaria y dispersa no permite analizar su comportamiento por tanto es importante tratar de resumirla en tablas de frecuencia que consoliden la información en números absolutos y acompañarla con números relativos (porcentajes).

Tabla de frecuencia correspondiente al número de clientes atendidos por el hotel "PANAMONTE" durante los últimos 50 días.

Número de Clientes X_i	Frecuencia Absoluta f_i	Frecuencia Relativa h_i	Frecuencia Acumulada	
			Absoluta F_i	Relativa H_i
2	6	0.12	6	0.12
3	8	0.16	14	0.28
4	9	0.18	23	0.46
5	11	0.22	34	0.68
6	7	0.14	41	0.82
7	5	0.10	46	0.92
8	4	0.08	50	1.00
TOTAL	50	1.00		

El gráfico que corresponde a esta tabla de frecuencia se denomina: **Histograma:**

Histograma de frecuencias absolutas Histograma frecuencias absolutas acumuladas



3. Organización de Variables Continuas

Comprende clasificaciones de unidades de análisis que toman valores con decimales, por tanto, las unidades de análisis se ordenarán de acuerdo con valores centrales que serían representativos de los valores que toman los elementos que se están analizando.

Ejemplo, el hotel "PANAMONTE" durante los últimos 32 días el valor de las compras en revistas y periódicos para la sala de recepción fueron:

5.2	10.2	7.0	7.1	10.2	8.3	9.4	9.2	6.5	7.1	6.6	7.8	6.8	7.2	8.4	9.6
8.5	5.7	6.4	10.1	8.2	9.0	7.8	8.2	5.3	6.2	9.1	8.6	7.0	7.7	8.3	7.5

Esta información diaria y dispersa no permitirá analizar su comportamiento por tanto es importante tratar de resumirla en una tabla de frecuencia que consolide la información en números absolutos y acompañarla con números relativos (porcentajes).

Para organizar una tabla de frecuencia se deberá seguir el procedimiento siguiente:

- *Elegir un número de intervalos de clases (k).*

Puede utilizar la regla **Sturges**:

$$k = 1 + 3.3 \text{ Log } [n]$$

Donde:

k = Número de intervalos

n = Número de datos

En el ejemplo: $k = 1 + 3.3 \text{ Log}[32] = 1 + 3.3(1.505) = 5.967 = \text{Aprox. } 6 \text{ intervalos}$

- *Determinar el Tamaño del Intervalo de Clase (C)*

$$c = A / k$$

A = Amplitud de los datos = (Observación Máxima - Observación Mínima) = 10.2 - 5.2 = 5.0

$$k = 6$$

Por tanto: $C = 5.0 / 6 = 0.8333 = \text{Aproximadamente } = 0.9$

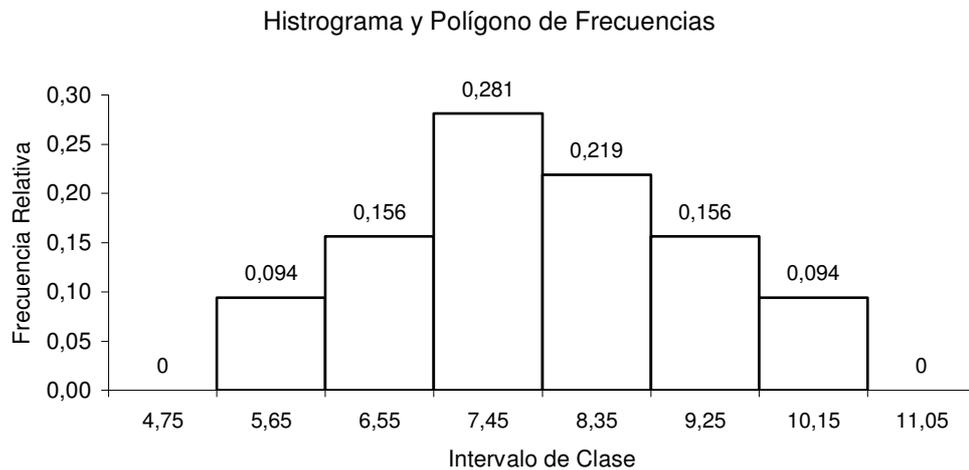
- *Realizar la clasificación y el conteo de datos en cada clase construida:*

Clase	Intervalos Clase	Clasificación y conteo de datos	Frecuencia
1	5.2 a menos de 6.1	5.2 5.7 5.3	3
2	6.1 a menos de 7.0	6.5 6.6 6.8 6.4 6.2	5
3	7.0 a menos de 7.9	7.0 7.1 7.1 7.8 7.2 7.8 7.0 7.7 7.5	9
4	7.9 a menos de 8.8	8.3 8.4 8.5 8.2 8.2 8.6 8.3	7
5	8.8 a menos de 9.7	9.4 9.2 9.6 9.0 9.1	5
6	9.7 a 10.6	10.2 10.2 10.1	3
	TOTAL		32

- *Construir la Tabla de Frecuencia*

Intervalo De Clase (Escala de Gasto)	Marca de Clase X_i	Frecuencia Absoluta f_i	Frecuencia Relativa h_i	Frec.Acum. Absoluta F_i	Frec. Acum. Relativa H_i
5.2 a menos de 6.1	5.65	3	0.094	3	0.094
6.1 a menos de 7.0	6.55	5	0.156	8	0.250
7.0 a menos de 7.9	7.45	9	0.281	17	0.531
7.9 a menos de 8.8	8.35	7	0.219	24	0.750
8.8 a menos de 9.7	9.25	5	0.156	29	0.906
9.7 a 10.6	10.15	3	0.094	32	1.000
TOTAL		32	1.000		

- *Construir el Histograma y el Polígono de Frecuencias*



Marca de Clase:
$$X_i = \frac{LI_i + LS_s}{2}$$

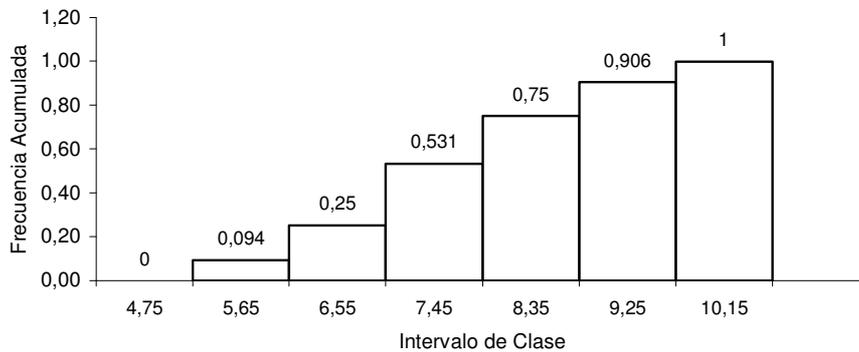
Se verifica que:
$$X_{i+1} = X_i + c$$

Gasto Total estimado:
$$\sum f_i X_i = 3(5.65) + 5(6.55) + 9(7.45) + 7(8.35) + 5(9.25) + 3(10.15) = 251.9$$

Gasto Total verdadero
$$\sum_{i=1}^{32} X_i = 250.2$$

- *Construir una tabla de frecuencias "Menor que".*

Histograma y Polígono de Frecuencias Relativas Acumuladas



<i>Gasto Diario</i>	Frecuencia Acumulada	Frecuencia Relativa
<i>Menos de 5.2</i>	0	0.000
<i>Menos de 6.1</i>	3	0.094
<i>Menos de 7.0</i>	8	0.250
<i>Menos de 7.9</i>	17	0.531
<i>Menos de 8.8</i>	24	0.750
<i>Menos de 9.7</i>	29	0.906
<i>Menos de 10.6</i>	32	1.000

- *Construir una tabla de frecuencias "O más que".*

<i>Gasto Diario</i>	Frecuencia Acumulada	Frecuencia Relativa
<i>5.2 ó más</i>	32	1.000
<i>6.1 ó más</i>	29	0.906
<i>7.0 ó más</i>	24	0.750
<i>7.9 ó más</i>	15	0.469
<i>8.8 ó más</i>	8	0.250
<i>9.7 ó más</i>	3	0.094
<i>10.6 ó más</i>	0	0.000

Análisis de la distribución de frecuencias

- ¿Cuántos días el hotel gastó "De 7.0 a menos de 7.9 soles"? : 9 días
- ¿Cuántos días el hotel gastó "Menos de 7.9 soles"? : 17 días
- ¿Cuántos días el hotel gastó "Menos de 9.7 soles"? : 29 días
- ¿Qué porcentaje de días el hotel gastó "Menos de 7.9 soles"? : 53.1%
- ¿Cuál es el gasto que comprende al 25% de los días que se gastó menos? : 7.0 soles
- ¿Cuál es el gasto que comprende al 25% de los días que se gastó mas? : 8.8 soles
- ¿Qué porcentaje de días el hotel gastó "Más de 7.9 soles"? : 46.9%

- Estimar el porcentaje de días en los que el hotel gastó "Menos de 9.0 soles" :78.5%

S/ 9.7--- 0.906	(9.7-8.8)---(0.906-0.75)	0.9 --- 0.156	X=0.75 +(0.156)(0.2)/0.9
S/ 9.0 --- X	(9.0-8.8)---(X-0.75)	0.2 --- (X-0.75)	X=0.75+0.035
S/ 8.8 --- 0.75			X=0.785

El 78.5% de los días el hotel gastó menos de 9.0 soles diarios en periódicos

- Estimar el monto del gasto diario en periódicos que comprendería al 50% de los menores gastos realizados por el hotel durante los 32 días analizados :7.8 soles

S/ 7.9--- 0.531	(7.9-7.0)---(0.531-0.25)	0.9 --- 0.281	X=7.0 +(0.9)(0.25)/0.281
X --- 0.50	(X-7.0) --- (0.50-0.25)	(X-7.0)---0.25	X=7.0+0.8
S/ 7.0 --- 0.25			X=7.8 soles

El 50% de los días observados el hotel gastó menos de 7.8 soles diarios en periódicos

- Estimar el monto del gasto diario en periódicos que comprendería al 85% de los menores gastos realizados por el hotel durante los 32 días analizados :9.38 soles

S/ 9.7--- 0.906	(9.7-8.8)---(0.906-0.85)	0.9--- 0.156	X=8.8 +(0.9)(0.10)/0.156
X --- 0.85	(X-8.8) --- (0.85-0.75)	(X-8.8)---0.10	X=8.8+0.58
S/ 8.8 --- 0.75			X=9.38 soles

El 85% de los días observados el hotel gastó menos de 9.38 soles diarios en periódicos

- Cuántos días el hotel gastó "Menos de 9.0 soles" al día en periódicos :25 días

S/ 9.7--- 29	(9.7-8.8)---(29-24)	0.9 --- 5	X=24 +(0.2)(5)/0.9
S/ 9.0--- X	(9.0-8.8) --- (X-24)	0.2 --- (X-24)	X=24+1.11
S/ 8.8 --- 24			X=25 días

En el hotel 25 días aproximadamente se gastó menos de 9.0 soles diarios en periódicos

• Representación Tallo - Hoja

Cuando se requiere conservar la información primaria y no se la pierda con el proceso de consolidarla en intervalos de clase, se recurrirá a los diagramas tallo hoja:

Ejemplo: Representar el valor de las compras de periódicos durante los 32 últimos:

5.2	10.2	7.0	7.1	10.2	8.3	9.4	9.2	6.5	7.1	6.6	7.8	6.8	7.2	8.4	9.6
8.5	5.7	6.4	10.1	8.2	9.0	7.8	8.2	5.3	6.2	9.1	8.6	7.0	7.7	8.3	7.5

Amplitud de hoja: 0.10

	Tallo	Hoja
3	5	2 3 7
8	6	2 4 5 6 8
(9)	7	0 0 1 1 2 5 7 8 8
15	8	2 2 3 3 4 5 6
8	9	0 1 2 4 6
3	10	1 2 2

	Tallo	Hoja
2	5	2 3
3	5	7
5	6	2 4
8	6	5 6 8
13	7	0 0 1 1 2
(4)	7	5 7 8 8
15	8	2 2 3 3 4
10	8	5 6
8	9	0 1 2 4
4	9	6
3	10	1 2 2

Ubicación del punto medio es $(n+1)/2=(32+1)/2=16.5$