

Estadística Descriptiva Métodos descriptivos visuales y medidas resumen

Las técnicas de la estadística descriptiva pueden aplicarse tanto a datos muestrales como a datos poblacionales.

Tipos de datos.

Los datos son los **valores** de la variable en estudio.

Éstos pueden ser *cuantitativos* ó *categoricos*.

Ejemplo. Consideremos el voltaje de salida de cierto tipo de generadores:

5.10 5.25 5.80 4.92 6.10 5.71 5.94.

los datos son una *medida*, son numéricos.

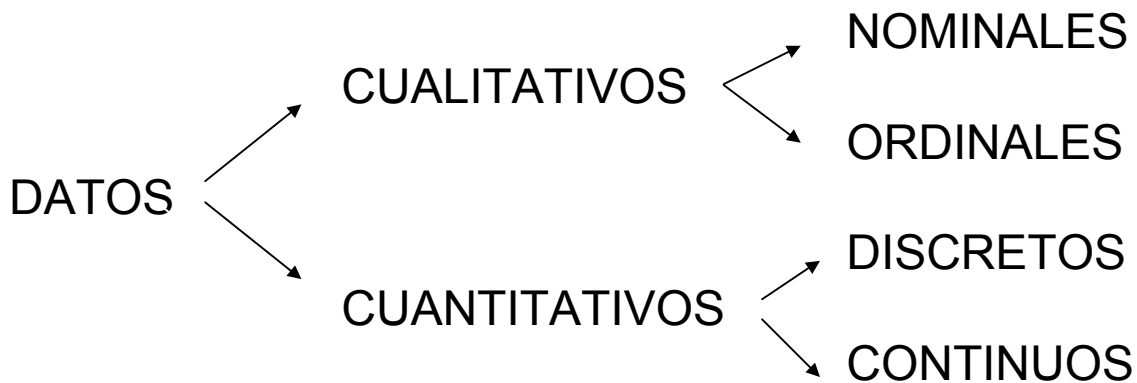
Cada generador cumple ó no con las normas

1 1 1 0 0 1 0

Datos categoricos: se asigna 1 cuando el generador cumple con los requerimientos y 0 si nó.

Se trata de *datos muestrales*

La población en estudio es



El tratamiento estadístico difiere según de qué tipo de datos se trate:

- datos cuantitativos tiene sentido calcular su promedio.
- datos cualitativos puede calcularse la cantidad (ó la proporción ó el porcentaje) de individuos que pertenecen a cada clase.

Tomar los números sin pensar en la variable y el proceso que llevó a medirlos puede producir resultados equivocados.

Ejemplo: En 1985, en USA, 4472 conductores de 65 años ó más estuvieron involucrados en accidentes fatales, 2790 conductores con edades entre 16 y 17 años tuvieron accidentes fatales.

La variable indicada es la proporción (tasa) de accidentes en cada grupo:

66 de cada 100 000 de los conductores adolescentes

24 de cada 100 000 de los conductores mayores

El uso de *totales* cuando es más adecuado utilizar *proporciones* es un mal uso de la estadística

Distribución

La naturaleza y la magnitud de la variación pueden ser tan importantes como el resultado promedio.

El patrón de variación de una variable es llamado distribución. La distribución registra los valores numéricos de la variable y cuán frecuentemente ocurre cada valor.

distribución muestral
distribución poblacional
de una variable

Métodos descriptivos visuales:

Diagramas de tallo-hoja.(Tukey, 1977)

Ejemplo. White, Riethof y Kushnir (1960) realizaron un estudio con el objetivo de detectar la presencia de ceras sintéticas adicionadas a las ceras naturales de abeja.

El agregado de cera microcristalina eleva el punto de fusión de la cera de abeja:

- el agregado de 5% aumenta el punto de fusión en $.85^{\circ}\text{C}$
- el agregado de 10% aumenta el punto de fusión en 2.22°C

Si todos los tipos de cera de abeja tuviesen el mismo punto de fusión, su determinación sería un procedimiento razonable para detectar diluciones.

Sin embargo, el punto de fusión y otras propiedades químicas de la cera de abeja varían de una colmena a otra.

Puntos de fusión de distintas ceras naturales

63.78	63.45	63.58	63.08	63.40	64.42	63.27	63.10
63.34	63.50	63.83	63.63	63.27	63.30	63.83	63.50
63.36	63.86	63.34	63.92	63.88	63.36	63.36	63.51
63.51	63.84	64.27	63.50	63.56	63.39	63.78	63.92
63.92	63.56	63.43	64.21	64.24	64.12	63.92	63.53
63.50	63.30	63.86	63.93	63.43	64.40	63.61	63.03
63.68	63.13	63.41	63.60	63.13	63.69	63.05	62.85
63.31	63.66	63.60					

Diagrama tallo-hoja

El punto decimal está un lugar a la izquierda de los dos puntos

TALLO HOJAS

628 : 5
629 :
630 : 358
631 : 033
632 : 77
633 : 001446669
634 : 01335
635 : 0000113668
636 : 0013689
637 : 88
638 : 334668
639 : 22223
640 :
641 : 2
642 : 147
643 :
644 : 02

El 95% (56/59) de los puntos de fusión de la cera natural de abeja se encuentra entre 62.9 y 64.3 (una diferencia de 1.4 °C). Por lo tanto el agregado de 5% de cera microcristalina (aumento de 0.85 °C) puede ser muy difícil de detectar especialmente si fue realizado en ceras con bajo punto de fusión, pero el agregado de 10% sería detectable (aumento de 2.22 °C).

A continuación veamos el diagrama tallo hoja obtenido con el Statistix

STEM AND LEAF PLOT OF FUSION

LEAF DIGIT UNIT = 0.01

628 5 REPRESENTS 62.85

MINIMUM 62.850

MEDIAN 63.530

MAXIMUM 64.420

	STEM	LEAVES
1	628	5
1	629	
4	630	358
7	631	033
9	632	77
18	633	001446669
23	634	01335
(10)	635	0000113668
26	636	0013689
19	637	88
17	638	334668
11	639	22223
6	640	
6	641	2
5	642	147
2	643	
2	644	02

La primera columna tiene la profundidad, salvo en la fila que contiene a la observación central (mediana). Allí aparece, entre paréntesis, la cantidad de hojas. ¿Qué es la profundidad?

La mediana es el valor que deja la misma cantidad de datos a cada lado. La profundidad de la mediana es $(59+1)/2 = 30$. Obtenemos así el valor 63.53 para la mediana.

En un diagrama tallo-hoja podemos ver:

- Cuán aproximadamente simétricos son los datos.

- Cuán dispersos están los valores.
- La aparición de valores inesperadamente más frecuentes.
- Si algunos valores están alejados del resto.
- Si hay concentraciones de valores.
- Si hay grupos separados.
- La distribución de los datos dentro de cada intervalo como patrones dentro de los datos.

Otro ejemplo. Gasto en \$ que realizaron 50 clientes sucesivos en un almacén

2.3	6.65	6.9	8.2	9.5	20.9	20.9	21.15	23.9	26.1
10.2	11.35	11.6	12.6	12.9	27.1	28.7	29.15	30.5	31.9
13	13.5	14.5	14.9	14.8	32.8	33.4	33.8	34.8	36.2
15.1	15.35	16.6	17.15	18.2	36.5	39.2	40.8	43.9	45.5
18.3	18.7	19.5	19.5	20.5	52.4	61.5	63.8	64.3	69.5

Para construir el diagrama tallo-hoja eliminamos los centavos.

LEAF DIGIT UNIT = 1
0 2 REPRESENTS 2.

Tallo Hoja			
1	0	2	0 2 representa al 2
5	0	6689	
15	1	0112233444	
24	1	556788899	
(5)	2	00013	
21	2	6789	
17	3	012334	
11	3	669	
8	4	03	
6	4	5	
5	5	2	
4	5		
4	6	134	

1	6	9	
---	---	---	--

Si giramos el diagrama de manera que los valores mayores queden del lado derecho, vemos que la distribución del gasto está fuertemente sesgada a la derecha. El centro de la distribución es 20, contando 25 lugares desde el valor menor. Pero el 10% de los consumidores gastan más de 50\$. Si se pudiera estudiar a los gastadores mayores en más detalle podrían tomarse medidas para atraerlos.

Una virtud de los diagramas tallo-hoja es que permiten examinar los valores de las observaciones.

Esto es muy inconveniente para conjuntos de datos muy grandes.

Histogramas.

Los histogramas muestran solamente la cantidad ó proporción de observaciones que pertenecen a cada intervalo.

La elección de la cantidad de clases, en un histograma o en un diagrama tallo-hoja puede ser un problema cuando se realizan a mano. Entre 5 y 20 suelen ser adecuadas, pero los programas estadísticos los eligen en forma automática.

Ejemplo Una prueba de vocabulario para alumnos de 7mo. grado tiene asignado puntaje de 2.0 a 12.9 (correspondiente al vocabulario de un niño de del comienzo del 2do. grado hasta 12.9). Los puntajes de los primeros 947 alumnos evaluados en 1985 en Capital Federal son:

5.4 6.8 2.1 7.6 6.6 7.8 8.1 5.6 6.0

Para construir un histograma de la distribución de los puntajes, proceda como sigue:

1. *Divida el rango de los datos en clases de igual longitud.*

$2.0 \leq \text{puntaje} < 3.0$

$3.0 \leq \text{puntaje} < 4.0$
...
$12.0 \leq \text{puntaje} < 13.0$

2. *Cuente la cantidad de observaciones que pertenecen a cada clase.* esas cantidades de observaciones se llaman *frecuencias*.

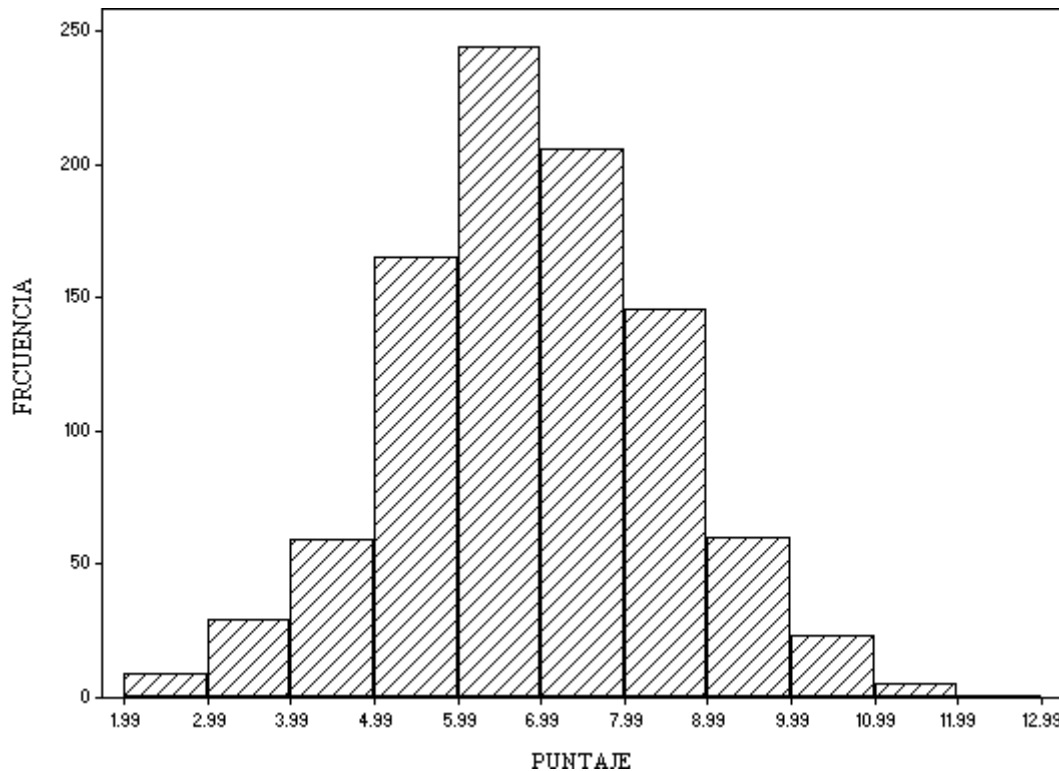
Tabla de frecuencias para el puntaje

Clase	Frecuencia # Estudiantes	Porcentaje	Porcentaje Acumulado
2.0 - 2.9	9	0.95	0.95
3.0 - 3.9	28	2.96	3.91
4.0 - 4.9	59	6.23	10.14
5.0 - 5.9	165	17.42	27.56
6.0 - 6.9	244	25.77	53.33
7.0 - 7.9	206	21.75	75.08
8.0 - 8.9	146	15.42	90.50
9.0 - 9.9	60	6.34	96.84
10.0 - 10.9	24	2.53	99.37
11.0 - 11.9	5	0.53	99.90
12.0 - 12.9	1	0.11	100.01
Total	947	100.01	

Observación: los porcentajes no suman 100 por errores de redondeo

3. *Dibuje el histograma.* En el eje horizontal se grafica la escala del puntaje y se lo divide de acuerdo con las clases definidas. Sobre cada intervalo de clase se dibuja un rectángulo cuya altura corresponde a la frecuencia de dicha clase. El eje vertical indica la frecuencia.

Histograma de puntajes de vocabulario en alumnos de 7o. grado



Un histograma de frecuencias relativas tiene el mismo aspecto que un histograma de frecuencias, simplemente se cambia la escala del eje vertical (0% al 100%) para leer los porcentajes. Los histogramas de frecuencias relativas son preferibles para comparar dos distribuciones con distinta cantidad de observaciones.

Importante: Áreas.

Nuestros ojos responden al área de las barras de manera que una figura precisa requiere que

el área del rectángulo sobre cada clase sea exactamente proporcional a la frecuencia de dicha clase.

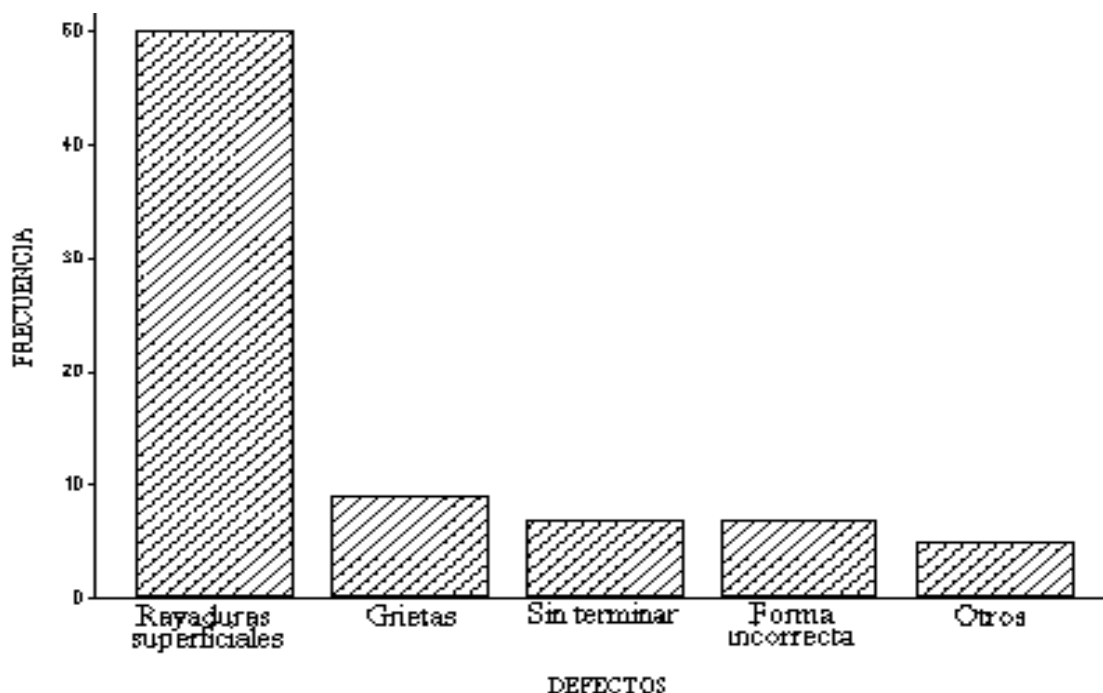
Cuando las clases tienen igual longitud, solamente tenemos que graficar las alturas de las barras proporcionales a las frecuencias.

Cuando se estudian datos que se publican como tablas de frecuencias con clases desiguales deben variarse las alturas de las barras de manera que las áreas (base x altura) sean proporcionales a las frecuencias.

Diagrama de Pareto, para datos categóricos.

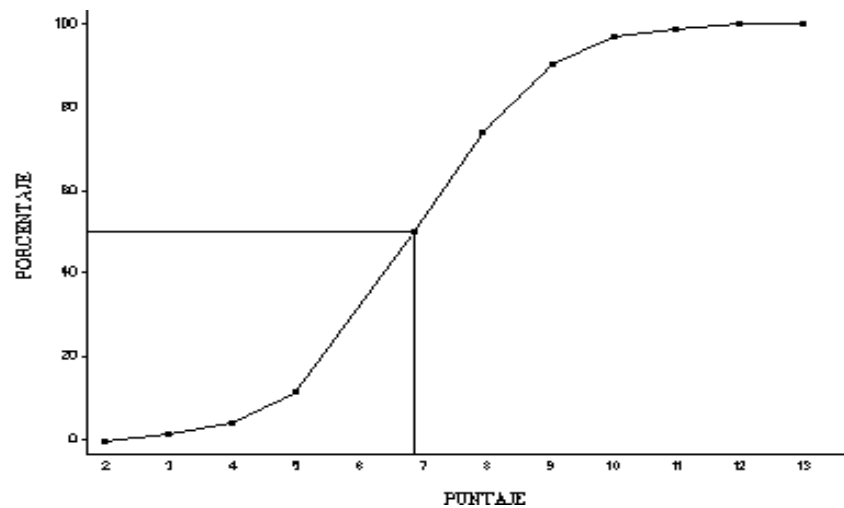
Si las clases en que se dividen los datos no tienen un orden natural, ni el ancho de las barras ni el orden tiene en general demasiada importancia. Sin embargo en el *diagrama de Pareto* las clases están ordenadas de acuerdo con su frecuencia.

El siguiente diagrama muestra un diagrama de Pareto para la frecuencia distintos tipos de defectos en un cierto tipo maquinaria inspeccionada

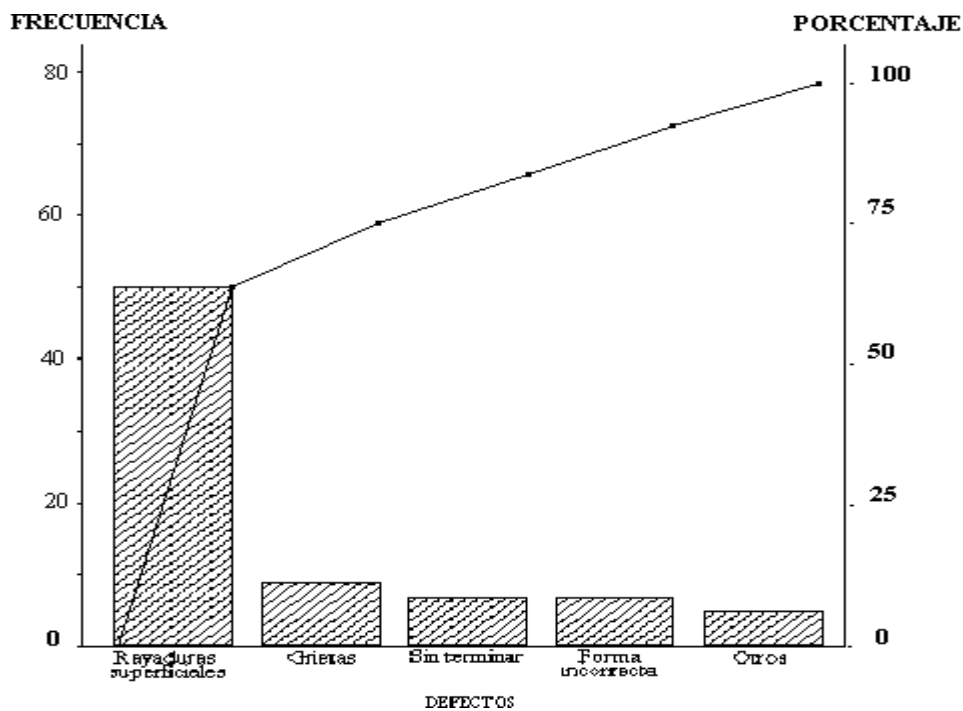


Gráficos de frecuencias relativas acumuladas

Frecuencias relativas acumuladas de los datos del puntaje



Un diagrama de Pareto habitualmente incluye al gráfico de frecuencias acumuladas



El 75% de las fallas se deben a rayaduras o grietas. Si se corrigieran esos dos problemas se mejoraría en gran medida la calidad.

RESUMEN

- **Medición:** es un proceso de representación de una característica mediante números.
- **Variable:** característica que varía cuando las mediciones se realizan sobre diferentes objetos o individuos.
- **Población:** totalidad de objetos de interés en un estudio.
- **Muestra:** subconjunto de la población.
- **Datos:** valores observados o medidos de una variable.
- La **distribución** de una variable es su patrón de variación, descrito por los valores de la variable y sus frecuencias ó por sus **frecuencias relativas acumuladas**.
- La distribución se representa gráficamente mediante un diagrama de **tallo-hoja**, que separa cada observación en tallo y hoja. También mediante un **histograma** que está basado en la frecuencia o frecuencia relativa de intervalos de clases o mediante un gráfico de frecuencias acumuladas.
- Al examinar una distribución, primero se debe localizar su centro. Luego, observar su **forma global** y las desviaciones claras de esa forma.
- La distribución, en un diagrama tallo hoja o un histograma, puede ser aproximadamente simétrica (a cada lado del centro se tiene una imagen en espejo de la otra) o asimétrica (una cola se extiende más lejos que la otra). La cantidad de picos es otro aspecto importante de la forma global.
- Las desviaciones de la forma global incluyen vacíos y valores extremos.