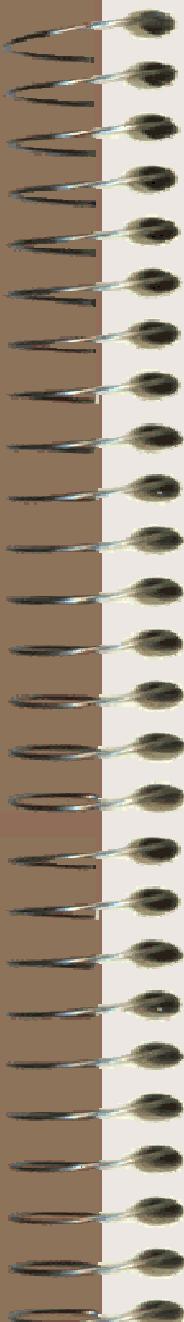


Métodos gráficos y numéricos para la descripción de un conjunto de datos



Estadística y Análisis de Datos

Estadística es la ciencia de recolectar, analizar y sacar conclusiones a partir de un conjunto de datos

Recolección / Fuente de datos apropiada



Organizar y resumir la información

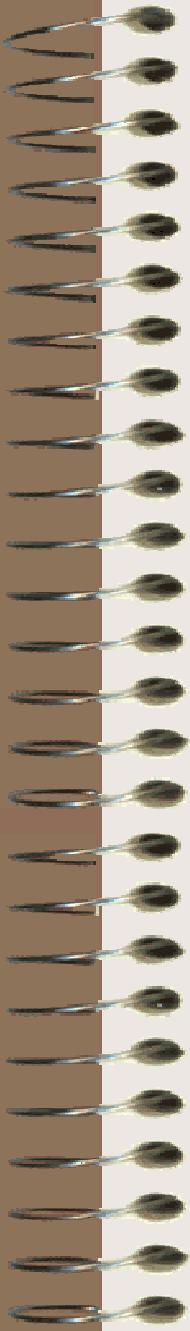
- ★ Tablas
- ★ Gráficos
- ★ Medidas resumen

Estadística
Descriptiva



Sacar conclusiones o tomar decisiones
a partir de los datos: generalizar los
resultados

Estadística
Inferencial



Recordemos:

Población: conjunto de individuos u objetos con una característica común observable. Es decir **población** de interés al conjunto completo de individuos u objetos acerca del cual se desea obtener información.

Una **muestra** es un subconjunto de la población seleccionada de una determinada manera.

Cuando se generalizan los resultados de una muestra a una población se corre el riesgo de realizar una conclusión incorrecta debido a que dicha conclusión se basará en información incompleta.

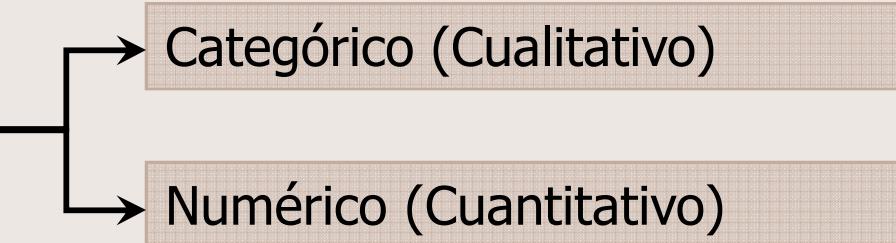
Un aspecto importante en el desarrollo de técnicas inferenciales es la cuantificación de la probabilidad de realizar una conclusión incorrecta.

Tipos de Datos

Una **variable** es una característica cuyo valor puede cambiar de un individuo u objeto a otro.

Un conjunto de datos que consiste de observaciones de una única variable constituye un conjunto de datos **univariado**

Conjunto de Datos
Univariado

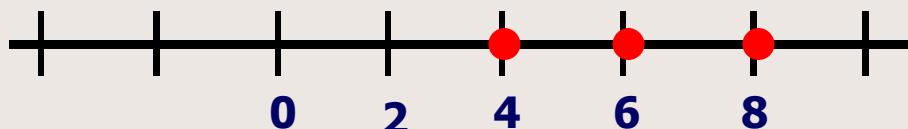


Conjunto de datos **bivariado**: cuando se estudian simultáneamente dos atributos. Ejemplo: peso y altura.

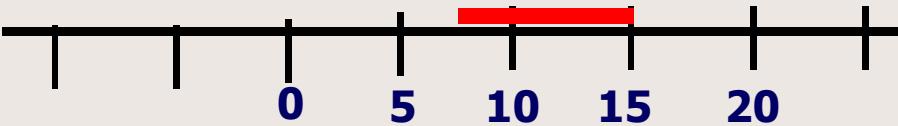
Datos **multivariados**: más de dos variables simultáneamente: peso, altura, frecuencia cardíaca y estado civil.

Variables cuantitativas

discretas: los posibles valores de la variable son puntos separados sobre la recta de números Reales.



continuas: el conjunto de posible valores forma un intervalo sobre la recta de números Reales.



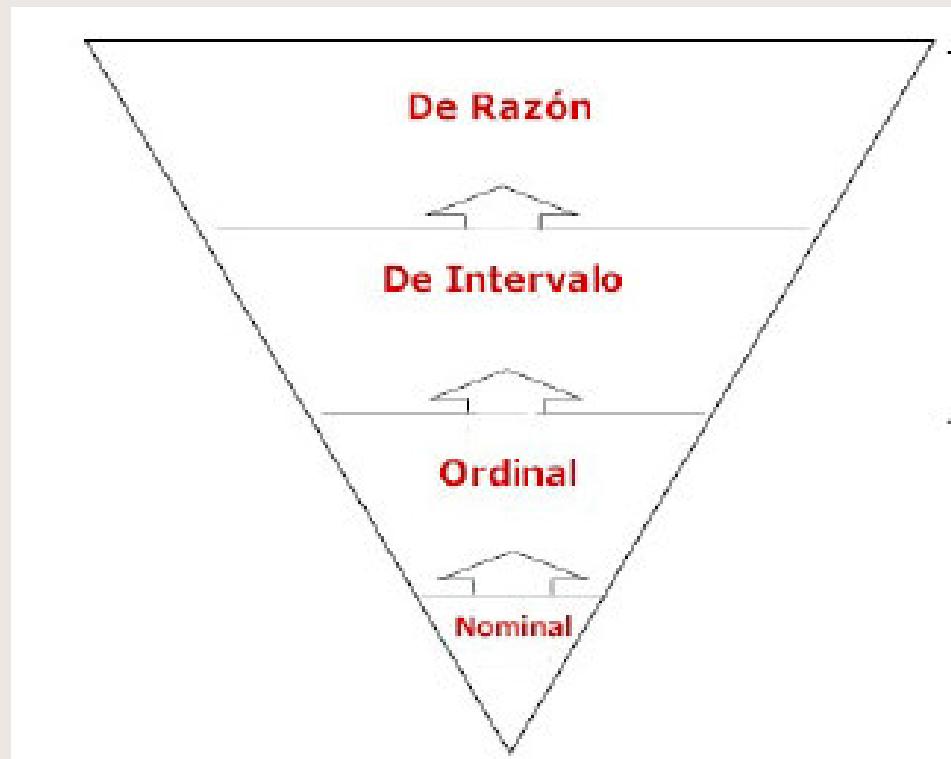
Las variables discretas generalmente aparecen cuando cada observación se determina a través de un proceso de conteo.

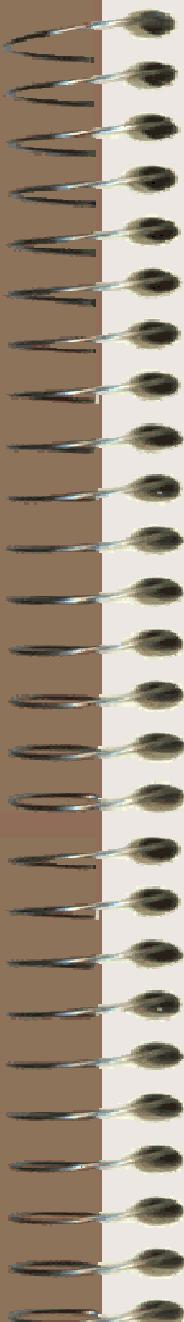
Ej. número de hijos, número de pétalos de una flor, etc.

Las variables continuas generalmente aparecen cuando cada observación se determina a través de un proceso de medición.

Ej. peso, altura, diámetro, temperatura, etc.

Escalas de medición





El Proceso del Análisis de Datos



- ✓ Comprender la naturaleza del problema
- ✓ Decidir qué medir y cómo medirlo
- ✓ Recolección de los datos
- ✓ Resumen de los datos y análisis preliminar
- ✓ Análisis de datos formal
- ✓ Interpretación de los resultados

Información a partir de los datos

Una característica de un libro es la cantidad de páginas que contiene.

Inclusive dentro de los distintos géneros literarios, la cantidad de páginas puede ser muy variable.

Se obtuvo una muestra de 40 libros de misterio de una librería y se registró el número de páginas.

229	247	347	246	307	181	198	214	234	340
314	260	202	320	360	320	200	414	262	248
376	211	214	218	276	628	255	352	197	308
203	371	203	406	261	378	223	181	284	196

....
Algunas de las preguntas que se podrían plantear acerca de los datos podrían ser:

Cúal es el número de páginas más común o representativo?

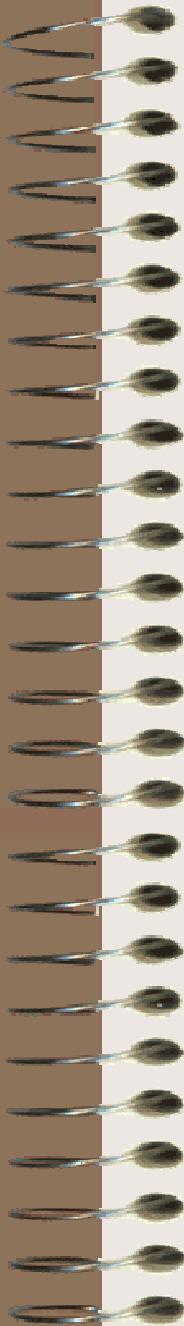
Se concentran las observaciones alrededor del valor más común o están dispersas?

Predominan los libros cortos o los libros largos, o se observaron aproximadamente la misma cantidad de ambos tipos de libros?

Existe algún libro cuya cantidad de páginas es inusual con respecto al resto de los libros?

Que proporción de los libros observados tienen al menos 500 páginas o menos de 200 páginas?

Para poder responder estas preguntas es necesario organizar los datos de una manera adecuada mediante presentaciones tabulares y gráficas.



Presentación de Datos Categóricos: Distribución de Frecuencias

Cuando el conjunto de datos es categórico, una forma muy común de presentar los datos es a través de una tabla llamada **distribución de frecuencias**.

Una distribución de frecuencias para datos categóricos es una tabla que muestra las posibles categorías junto con las frecuencias asociadas o las frecuencias relativas.

La frecuencia para una categoría en particular es el número de veces que la categoría aparece en el conjunto de datos.

La frecuencia relativa para una categoría en particular es la fracción o proporción de veces que la categoría aparece en el conjunto de datos, se calcula como el cociente entre la frecuencia y el número de observaciones en el conjunto de datos.

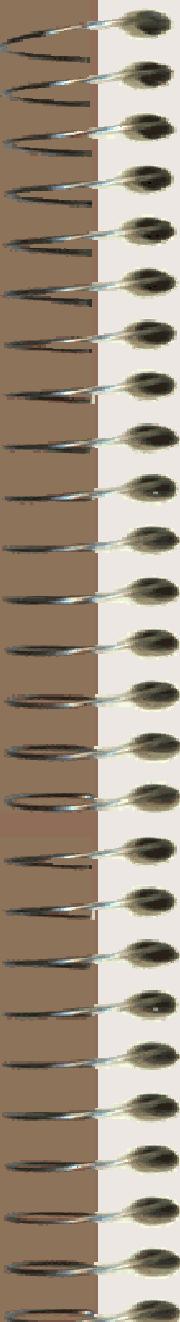
Distribución de Frecuencias

Muchos esfuerzos en salud pública están dirigidos al aumento de los niveles de la actividad física. Un informe sobre actividad física en mujeres de origen mejicano residentes en zonas urbanas reporta los siguientes patrones de actividad al aire libre preferida:

W	T	A	W	G	T	W	W	C	W
T	W	A	T	T	W	G	W	W	C
A	W	A	W	W	W	T	W	W	T

W: caminar, T: trotar, C: ciclismo, G: jardinería, A: aerobics

Categoría	Frecuencia	Frecuencia Relativa
Caminar	15	0.500
Trotar	7	0.233
Ciclismo	2	0.067
Jardinería	2	0.067
Aerobics	4	0.133
Total	30	1.000



Presentación de Datos Categóricos: Gráfico de Barras

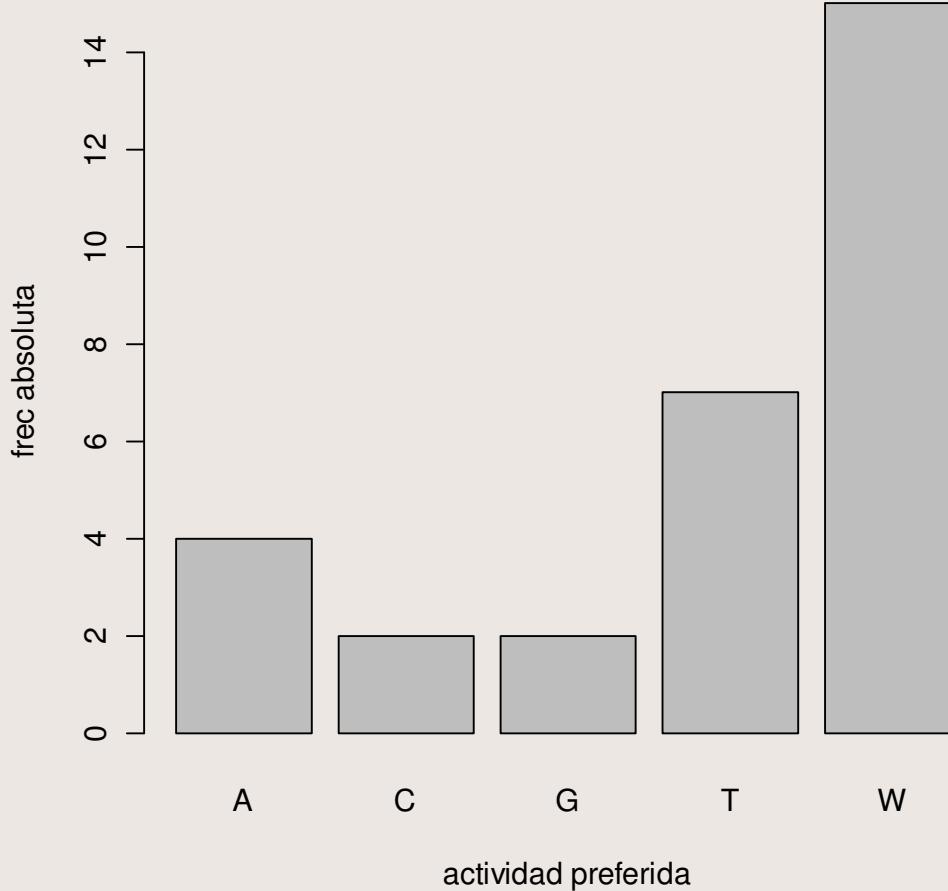
Un gráfico de barras presenta una distribución de frecuencia para datos categóricos. Cada categoría en la distribución de frecuencias se representa por una barra o rectángulo y el gráfico se construye de forma tal que el « largo » de la barra es proporcional a la frecuencia o frecuencia relativa correspondiente y además todas las barras deben tener el mismo « ancho », el cual debe ser mayor que la separación entre las barras. Evitar usar colores diferentes para cada barra (puede engañar visualmente)

El gráfico de barras provee una representación visual de la información en una distribución de frecuencias.

En R...

```
> actividad<-  
  c ("W", "T", "A", "W", "G", "T", "W", "W", "C", "W", "T"  
  , "W", "A", "T", "T", "W", "G", "W", "W", "C", "A", "W",  
  "A", "W", "W", "W", "T", "W", "W", "W", "T")  
  
> table (actividad)  
actividad  
  A   C   G   T   W  
  4   2   2   7  15  
> barplot (table (actividad), xlab="actividad  
preferida", ylab="frec absoluta")  
>
```

Presentación de Datos Categóricos o variables cualitativas: Gráfico de Barras



A partir del gráfico es fácil observar que "caminar" fue la actividad que se reportó más frecuentemente, seguida por "trotar".

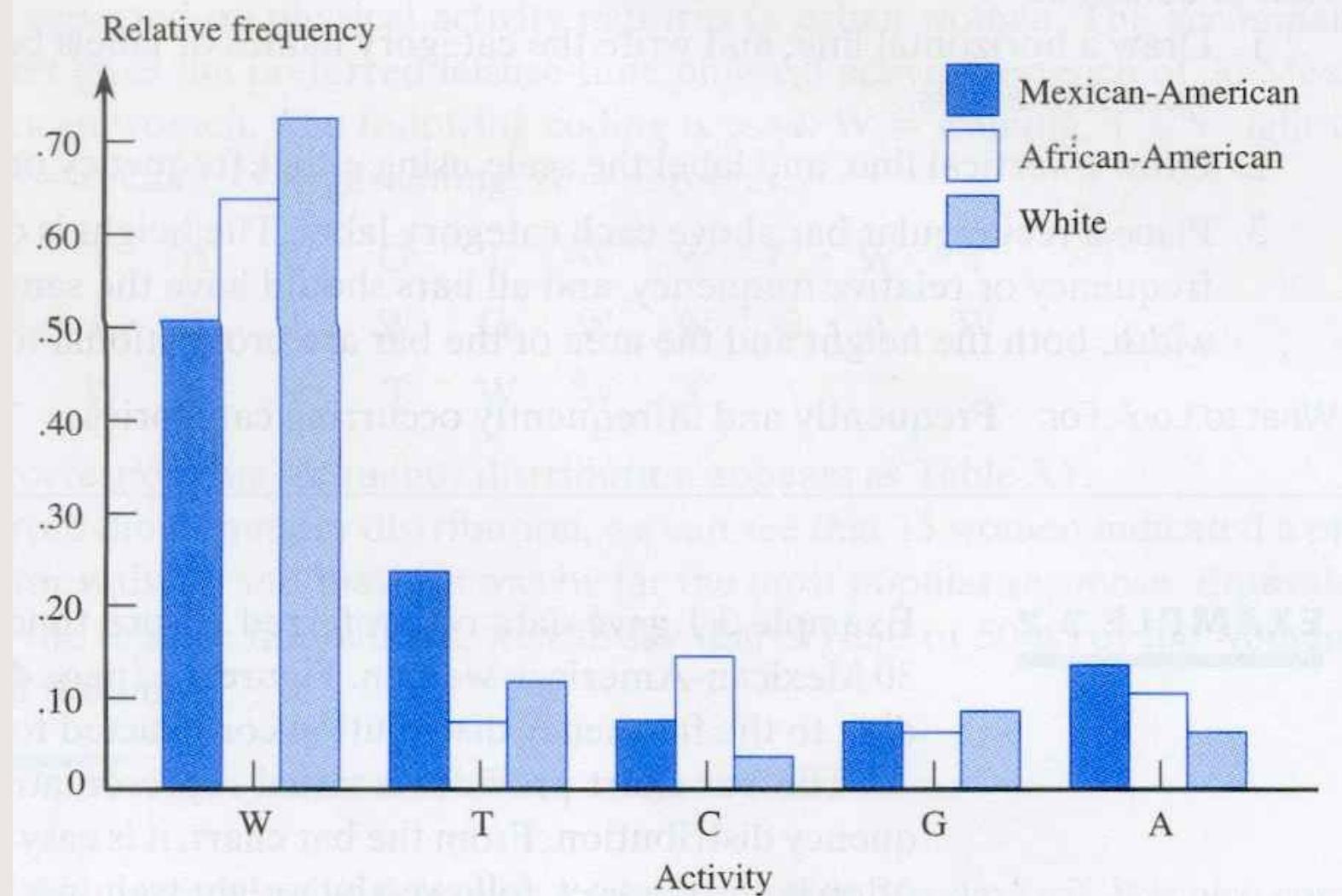


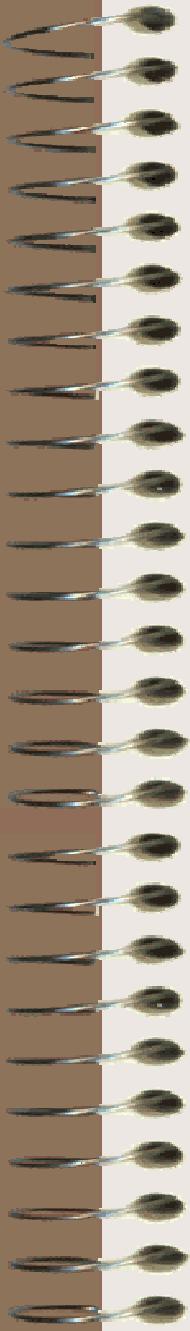
Gráfico de Barras: Comparación de Distribuciones

Los gráficos de barras también pueden utilizarse para presentar una comparación visual entre 2 o más grupos. Ésto se realiza construyendo 2 o más gráficos de barras usando los mismos ejes verticales y horizontales.

El artículo citado en ejemplo anterior sobre actividad física al aire libre también presenta resultados para mujeres de raza blanca y afro-americana.

Los tamaños muestrales para los tres grupos fueron diferentes por cual es aconsejable utilizar las frecuencias relativas en el gráfico de barras para poder realizar comparaciones entre los 3 grupos.





Presentación de Datos Categóricos: Gráfico de Sectores

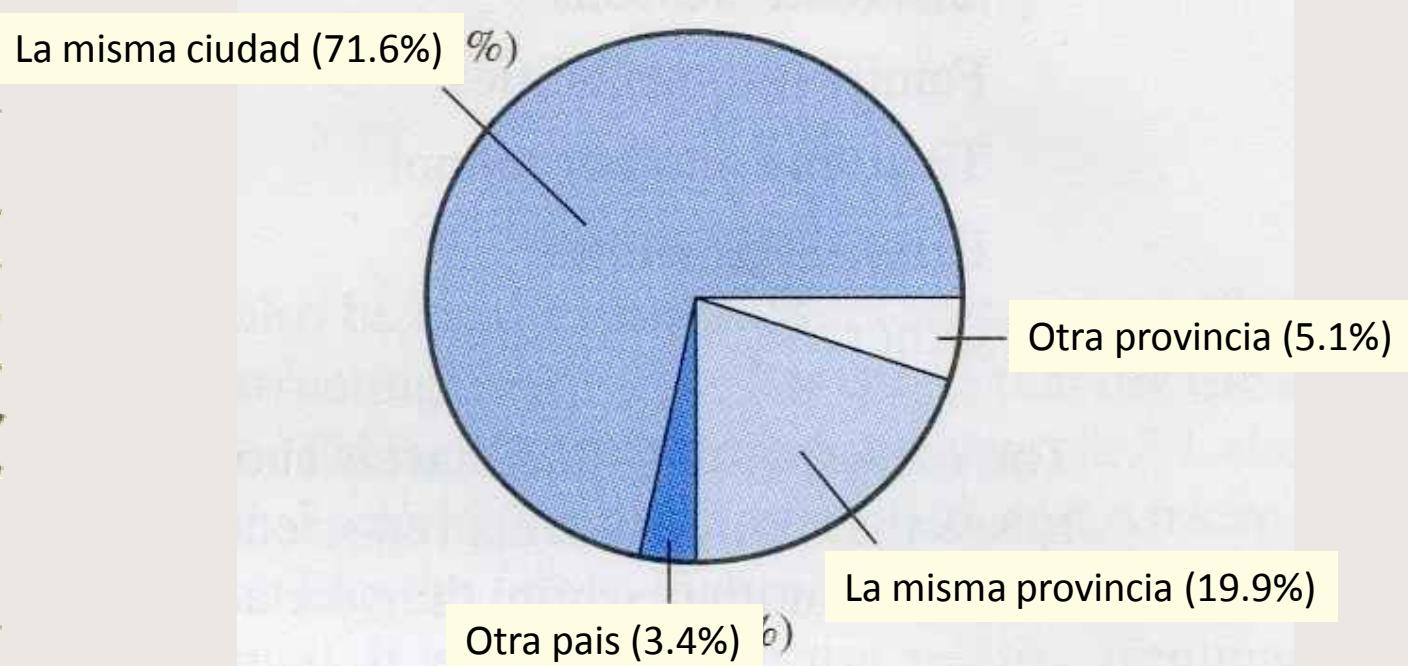
Un conjunto de datos categóricos tambien pueden ser resumido utilizando un gráfico de sectores.

En un gráfico de sectores, se utiliza un círculo para representar todo el conjunto de datos, mientras que "sectores" del círculo representan las posibles categorías.

El tamaño de un sector para una categoría en particular es proporcional a la correspondiente frecuencia o frecuencia relativa.

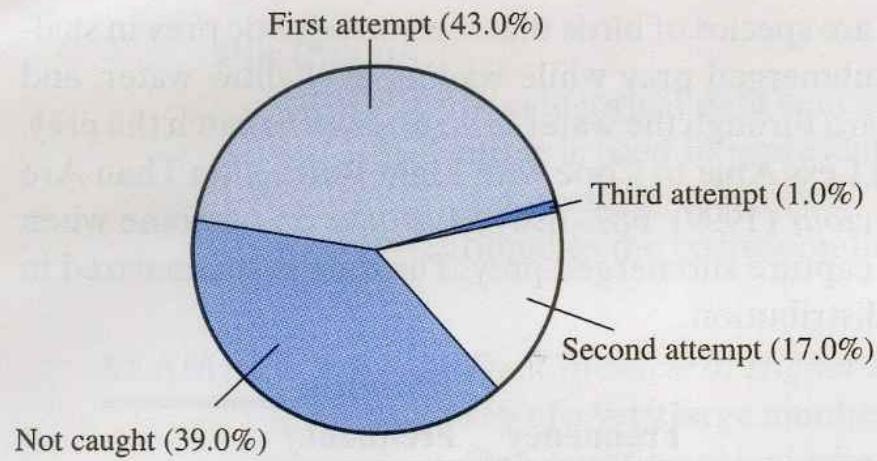
Un diario publicó los resultados de una encuesta realizada a un gran número de estudiantes que ingresaron a la universidad. Una de las preguntas investigaba el lugar de origen de los estudiantes, con las siguientes categorias: la misma ciudad donde se encontraba la universidad, la misma provincia, otra provincia y otro país.

Gráfico de sectores (torta o Pie)

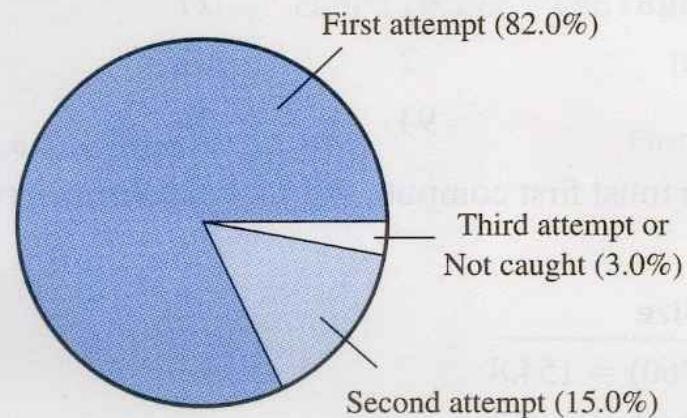


Los gráficos de sectores son útiles para presentar resultados de variables categóricas con pocas categorías.

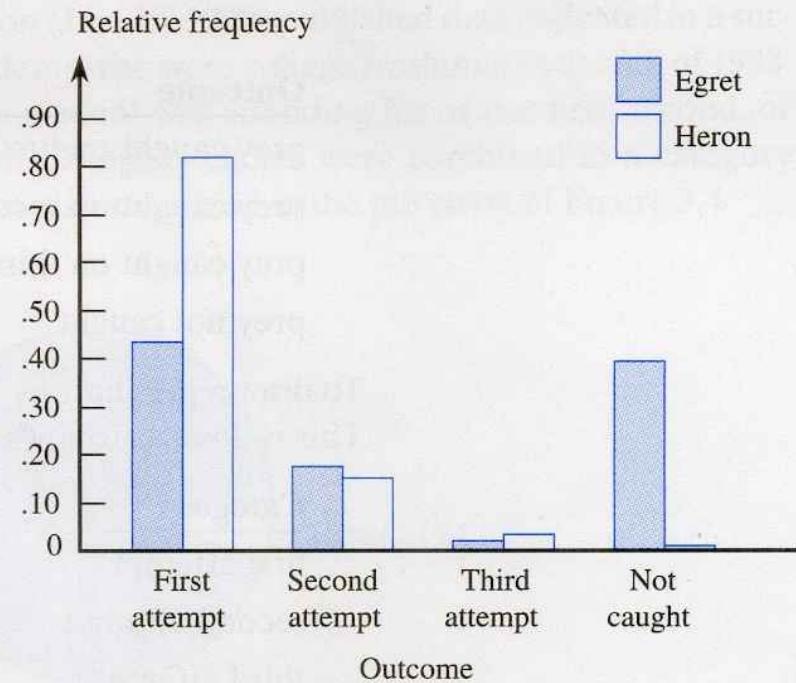
Comparativos: barras adyacentes o varios gráficos de sectores



(a)

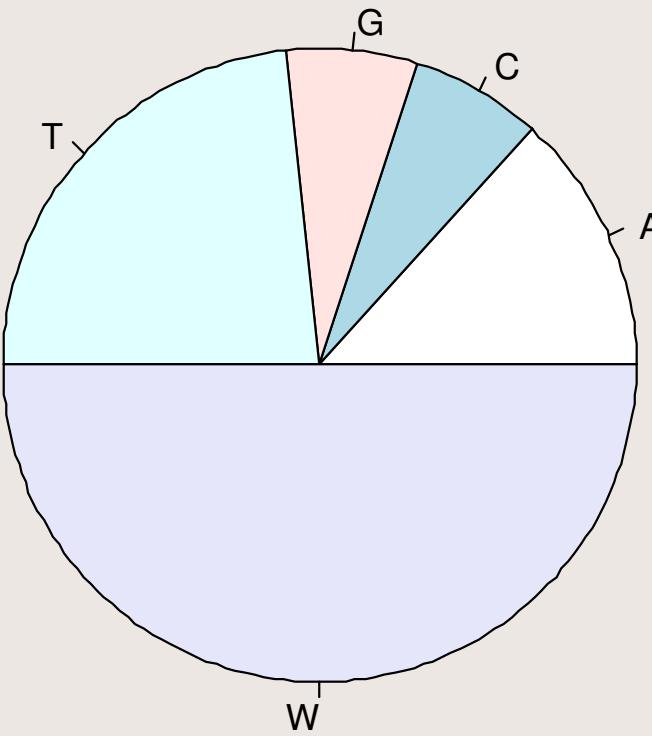


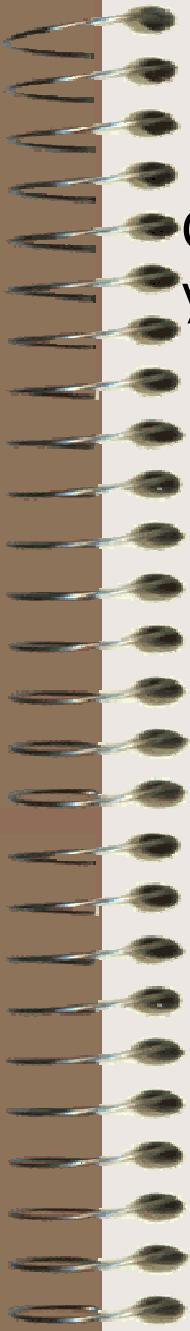
(b)



En R... ejemplo de actividad preferida

```
> pie(table(actividad))
```





Otros Usos de Gráficos de Barras y Gráfico de Sectores

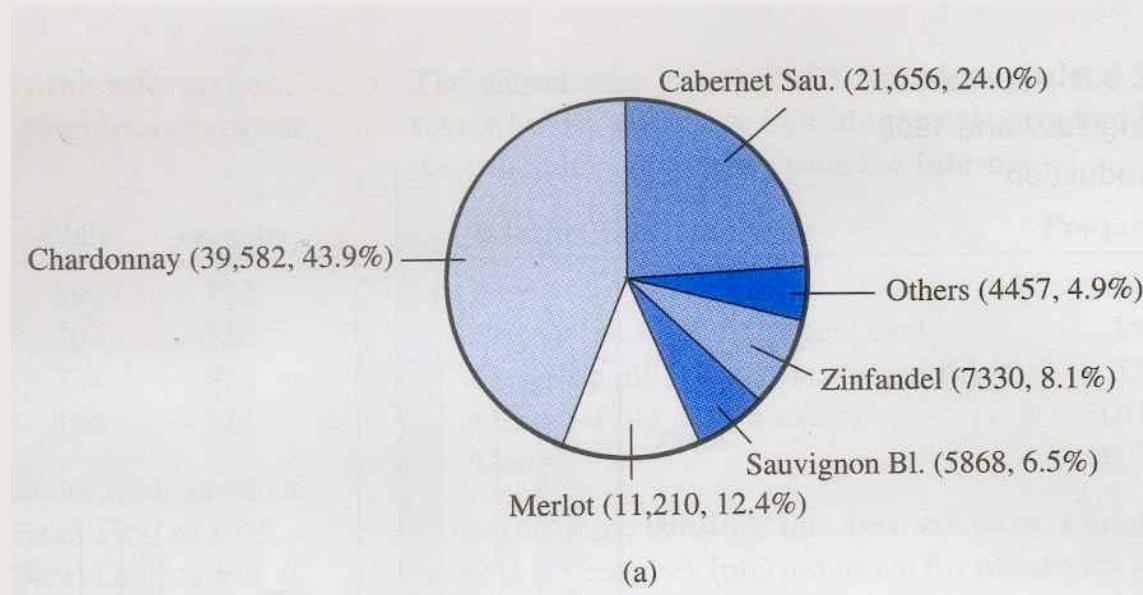
Como se vio en los ejemplos anteriores, los gráficos de barras y sectores pueden ser utilizados para resumir datos categóricos.

Sin embargo, ocasionalmente este tipo de gráfico se utiliza con otros propósitos.

Producción de uvas utilizadas para hacer vino en 2007

Tipo de Uva	Toneladas
Chardonnay	39.582
Chenin Blanc	1.601
Sauvignon Blanc	5.868
Cabernet Sauvignon	21.656
Merlot	11.210
Pinot Noir	2.856
Zinfandel	7.330
Total	90.103

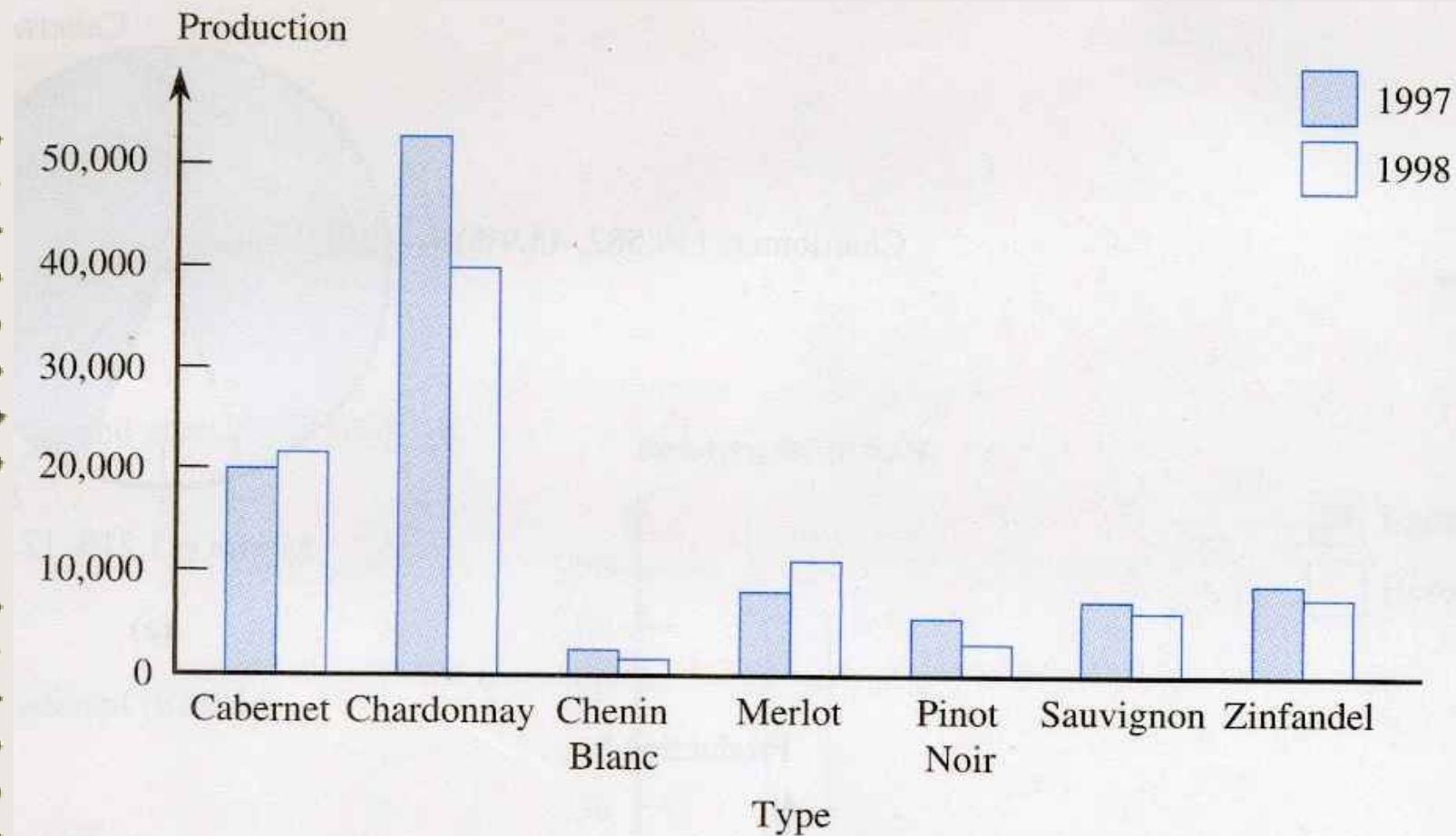
Otros Usos de Gráficos de Barras y Gráfico de Sectores

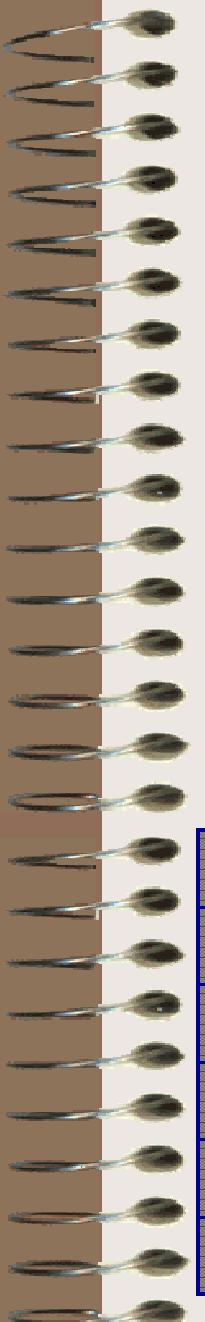


Aunque los datos anteriores no representan una distribución de frecuencias para un conjunto de datos categóricos, es común representar información de este tipo graficamente utilizando un gráfico de sectores o de barras.

El círculo completo representa la producción total de uvas y cada sector muestra la proporción del total para cada uno de 7 diferentes tipos de uvas.

Todo gráfico debe tener un título que responda las siguientes preguntas: qué? cómo? dónde? y cuándo?





Presentación de Datos Numéricos (variables cuantitativas): Gráficos de Puntos (Dotplots)

Un gráfico de puntos es una manera simple de presentar datos numéricos cuando el conjunto de datos es azonablemente pequeño.

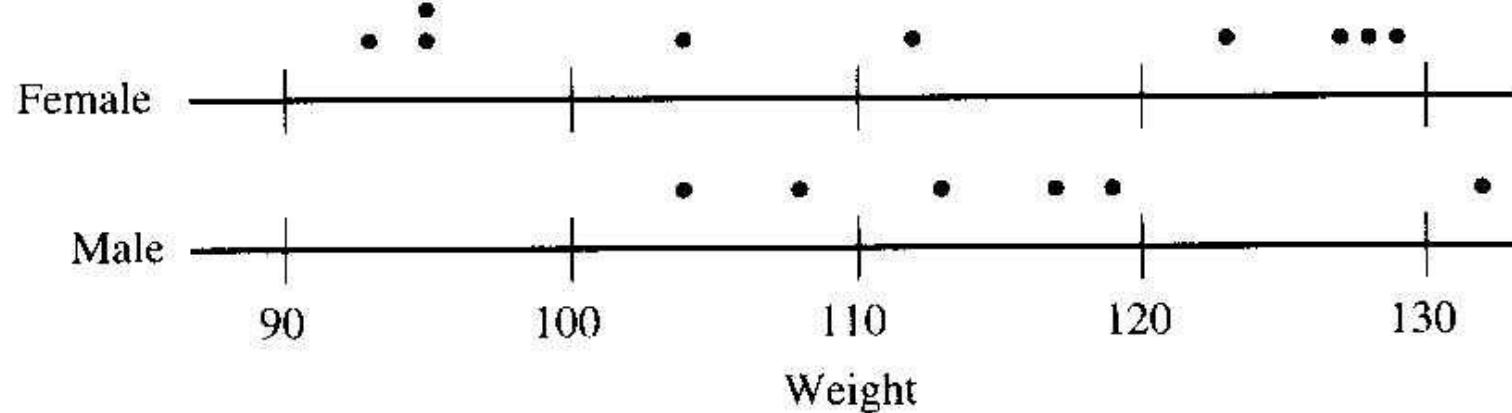
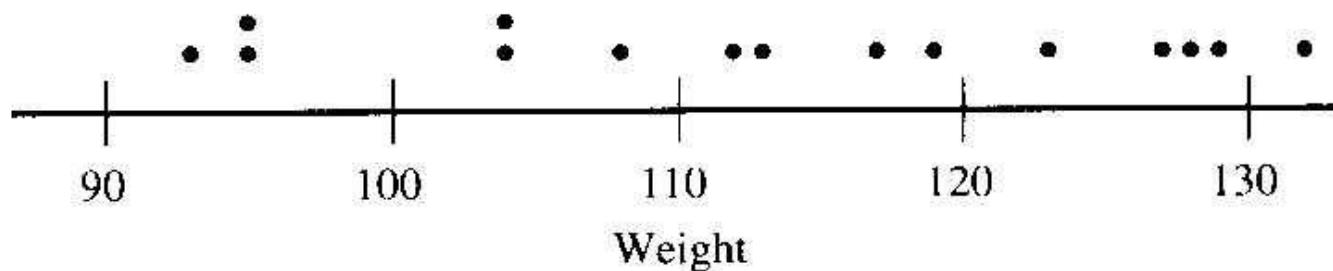
Cada observación se representa por un punto sobre la ubicación correspondiente a su valor en una escala horizontal.

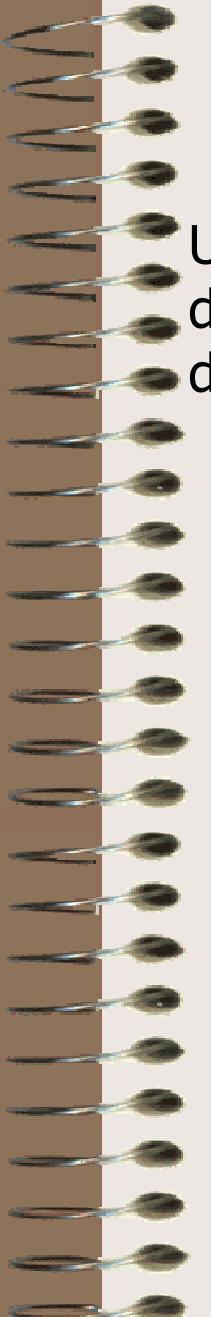
Cuando un valor se presenta en más de una ocasión, los puntos se apilan verticalmente.

Peso y Género de 15 Potrillos recién nacidos

Potrillo	Género	Peso	Potrillo	Género	Peso	Potrillo	Género	Peso
1	F	129	6	M	113	11	M	108
2	M	119	7	F	95	12	F	95
3	M	132	8	F	104	13	M	117
4	F	123	9	M	104	14	F	128
5	F	112	10	F	93	15	F	127

Presentación de Datos Numéricos: Gráficos de Puntos (Dotplots)





Presentación de Datos Numéricos: Diagramas de Tallo y Hoja

Un diagrama de tallo y hoja es una forma efectiva y compacta de resumir información numérica. Cada número en el conjunto de datos se divide en dos partes llamadas el tallo y la hoja.

El tallo es la primer parte del número y consiste del primer dígito (o primeros dígitos).

La hoja es la última parte del número y consiste de el o los dígitos finales.

Por ejemplo el número 213 puede ser dividido en un tallo igual a 2 y una hoja igual a 13 o un tallo igual a 21 y una hoja igual a 3.

Por último se utilizan los tallos y hojas resultantes para construir el diagrama.

En R...

> potrillos<-

c(129,119,132,123,112,113,95,104,93,108,95,117,128,127)

> stem(potrillos)

The decimal point is 1 digit(s) to the right of the |

9 | 355

10 | 48

11 | 2379

12 | 3789

13 | 2

Diagramas de Tallo y Hoja: Ejemplo

Variable: porcentaje de estudiantes universitarios que trabajan.

0	4
1	1345678889
2	1223456666777889999
3	011223334455566667777888899999
4	1112222334444556666677788888999
5	00111222233455666667777888899
6	01111244455666778

Stem: Tens digit

Leaf: Ones digit



Diagramas de Tallo y Hoja

En gráfico anterior cada linea ha sido ordenada de menor a mayor.

Aun cuando no se aplique dicho ordenamiento es posible obtener un diagrama informativo que muestra muchas características importantes de un conjunto de datos tales como la forma y la dispersión.

Los diagramas de tallo y hoja pueden ser muy útiles para obtener una idea de los valores más comunes en un conjunto de datos y cuan dispersos están los datos.

También es posible detectar valores que se encuentran muy alejados del resto de los valores en conjunto de datos. Dichos valores se denominan "extremos" o "outliers".

Diagramas de Tallo y Hoja: Ejemplo

Utilizando los datos del ejemplo del número de páginas en libros de misterio.

229	247	347	246	307	181	198	214	234	340
314	260	202	320	360	320	200	414	262	248
376	211	214	218	276	628	255	352	197	308
203	371	203	406	261	378	223	181	284	196

Una elección natural para el tallo es utilizar las centenas. Esto resultaría entre 6 tallos (1,2,3,4,5,6). Si se utilizan los dos primeros dígitos de un número como tallo se obtendrían 45 tallos (18, 19, ..., 62).

Un diagrama de tallo y hojas con 45 tallos no sería un resumen muy efectivo de los datos. En general, los diagramas de tallo y hoja deben contener entre 5 y 20 tallos.

Diagramas de Tallo y Hoja

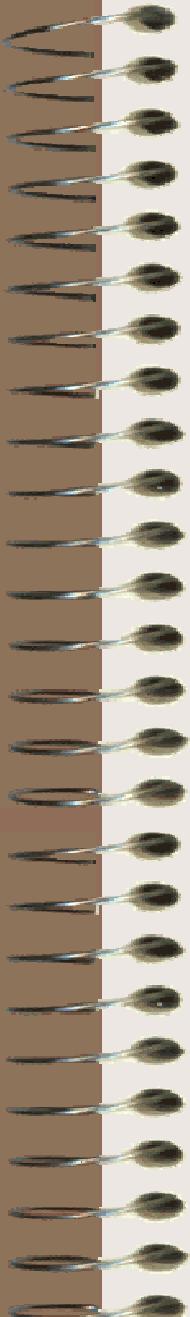
Ejemplo:

Si se elige el dígito de las centenas como tallo, los dos dígitos restantes conforman la hoja. Por ejemplo:

229:	tallo = 2,	hoja = 29
247:	tallo = 2,	hoja = 47
347:	tallo = 3,	hoja = 47

1	81,98,97,81,96
2	29,47,46,14,34,60,02,00,62,48,11,14,18,76,55,03,03,61,23,84
3	47,07,40,14,20,60,20,76,52,08,71,78
4	14,06
5	
6	28

Stem: Hundreds
Leaf: Tens and ones



Diagramas de Tallo y Hoja: Ejemplo

Se utilizan comas para separar las hojas cuando cada hoja tiene 2 o más dígitos.

El diagrama muestra que la mayoría de los libros de misterio tenían entre 200 y 400 páginas siendo 300 un número de páginas típico.

Existe un "outlier" o valor extremo, 638 páginas que resulta inusualmente grande cuando se lo compara con los otros libros en la muestra.

Una presentación alternativa se obtiene si se eliminan todos excepto el último dígito de las hojas. Este truncamiento no produce una gran pérdida de información con respecto a la forma o dispersión de los datos.

1	89989
2	24413600641117500628
3	404126275077
4	10
5	
6	2

Stem: Hundreds
Leaf: Tens

Diagrams de Tallo y Hoja: Variante

En ciertas ocasiones la elección natural de los tallos produce un diagrama en el cual muchas observaciones se concentran en unos pocos tallos.

Se puede obtener un diagrama más informativo dividiendo las hojas en un tallo determinado en dos grupos: las que comienzan con 0,1,2,3 o 4 ("low"), y las que comienzan con 5,6,7,8 o 9 ("high").

Luego, cada tallo se lista dos veces cuando se construye el diagrama.

Ejemplo: ingesta diaria de proteínas (en gramos de proteínas por kilogramo de peso) para 20 atletas.

1.4	2.2	2.7	1.5	2.3	1.7	2.3	1.5	1.8	2.8
1.8	1.9	2.0	2.3	1.5	1.9	1.7	1.8	1.6	3.0

1 | 457588959786

2 | 2733803

Stem: Ones

3 | 0

Leaf: Tenths

1L | 4

1H | 57588959786

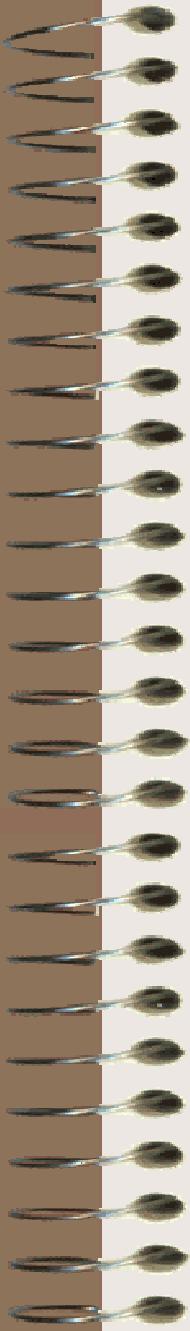
2L | 23303

2H | 78

Stem: Ones

3L | 0

Leaf: Tenths



Presentación de Datos Numéricos: Distribuciones de Frecuencias e Histogramas

Un diagrama de tallo y hoja puede no ser adecuado cuando el conjunto de datos contiene muchas observaciones. Generalmente, las distribuciones de frecuencias y los histogramas son muy útiles en tales ocasiones.

Datos Numéricos Discretos:

Generalmente se originan en procesos de conteo. En tales casos, cada observación es un número entero. Como en el caso de datos categóricos, una distribución de frecuencias lista cada valor posible (en forma individual o agrupados en intervalos), la frecuencia asociada y la frecuencia relativa.

Presentación de Datos Numéricos: Distribuciones de Frecuencias

Ejemplo:

Número de crías nacidos a 29 mapaches hembras observadas durante un periodo de 4 años.

1	3	2	1	1	4	2	4	1	1
1	3	1	1	1	1	2	2	1	1
4	1	1	2	1	1	1	1	3	

Número de Crías	Frecuencia	Frecuencia Relativa
1	18	0.621
2	5	0.172
3	3	0.103
4	3	0.103
Total	29	0.999



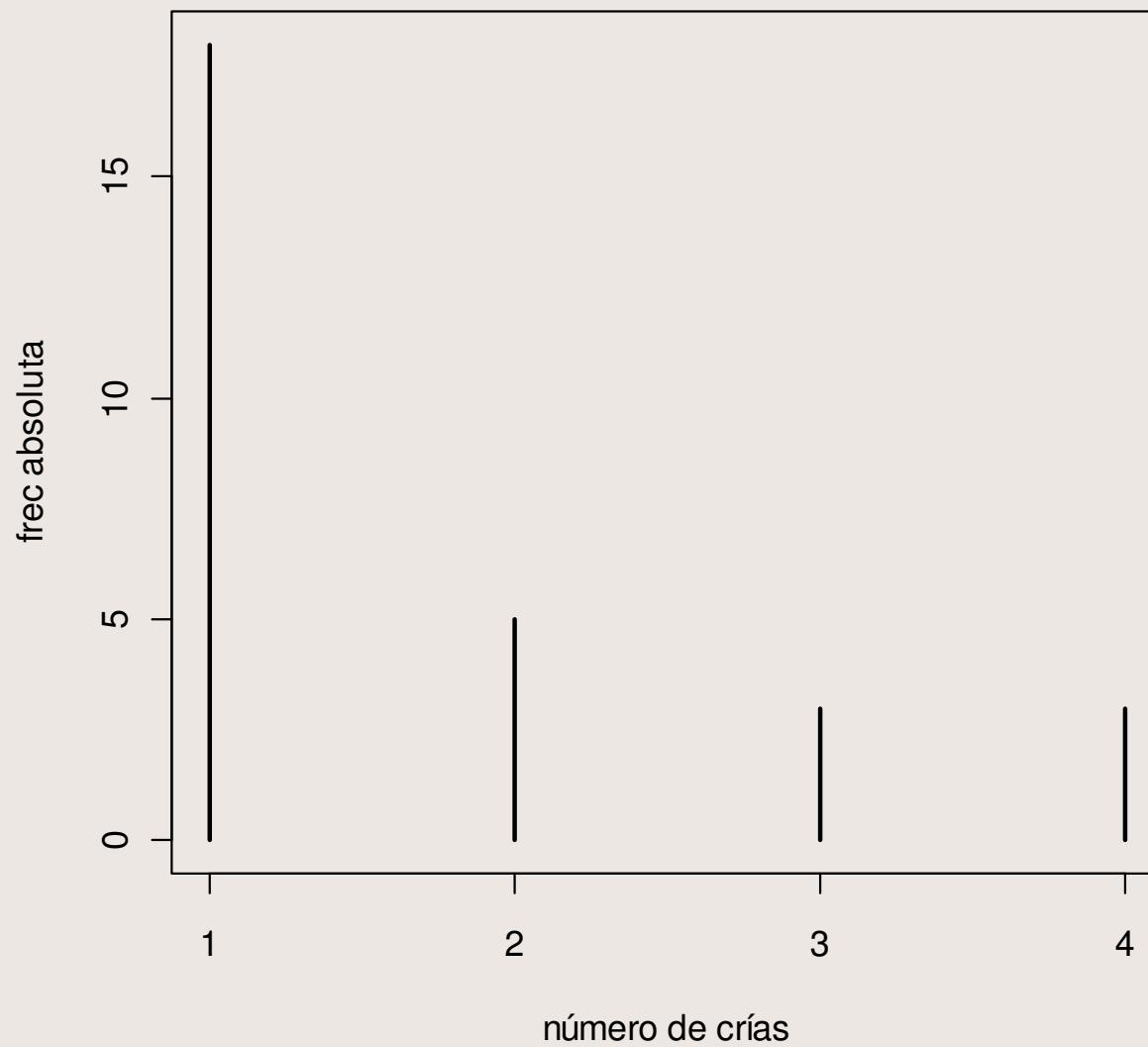
Presentación de Datos Numéricos:

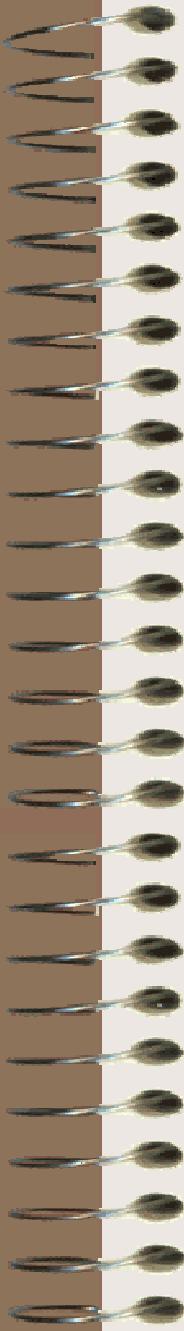
En un sistema de coordenadas cartesianas se representan en el eje de las abscisas u horizontal los distintos valores que asume la variable discreta en estudio y en el eje de las ordenadas o vertical se construye una escala adecuada para representar la frecuencia correspondiente a cada uno de esos valores. Sobre cada valor de la variable, se levanta una línea o bastón igual a la frecuencia de la categoría en cuestión.

La representación Gráfica se denomina

"Gráfico de bastones".

Presentación de Datos Numéricos:





Presentación de Datos Numéricos

En algunos casos un conjunto de datos numéricos discretos contiene un gran número de valores posibles y también pueden existir algunos valores muy grandes o muy pequeños que están muy alejados del resto de los datos.

En este caso, en lugar de construir una distribución de frecuencias con una larga lista de valores posibles, es común agrupar los valores observados en intervalos o rangos.

Ejemplo:

uso de alcohol en una muestra de 176 estudiantes universitarios.

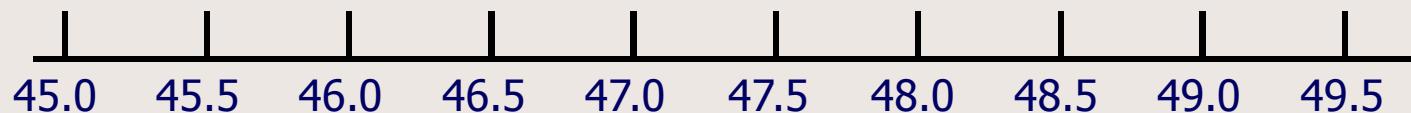
Tragos por Semana	Frecuencia
0 a 1	52
2 a 5	38
6 a 9	17
10 a 15	35
16 o más	34

Presentación de Datos Numéricos Continuos: Distribución de Frecuencias e Histogramas

La dificultad para construir distribuciones de frecuencias e histogramas para datos numéricos continuos aparece del hecho de que no existen categorías naturales.

Para variables tales como tiempo de reacción (segundos) o rendimiento de combustible (kilometros por litro), será necesario definir categorías arbitrarias.

Intervalos para rendimiento de combustible:



Los intervalos se denominan: intervalos de clase o simplemente "clases".

Por convención cuando un valor cae exactamente en uno de los puntos que dividen a las clases se lo ubica en la clase superior.

Ejemplo

El zinc es un elemento importante en la constitución de la dieta debido a que ayuda al mantenimiento del sistema inmunológico. A continuación se presentan datos correspondientes a una muestra de 40 pacientes con artritis reumatoidea.

8.0	12.9	13.0	8.9	10.1	7.3	11.1	10.9	6.2	8.1
8.8	10.4	15.7	13.6	19.3	9.9	8.5	11.1	10.7	8.8
10.7	6.8	7.4	4.8	11.8	13.0	9.5	8.1	6.9	11.5
11.2	13.6	4.9	18.8	15.7	10.8	10.7	11.5	16.1	9.9

La menor observación es 4.8 y la mayor es 19.3. Resulta razonable comenzar el primer intervalo de clase en 3.0 y asignar un ancho de 3.0.

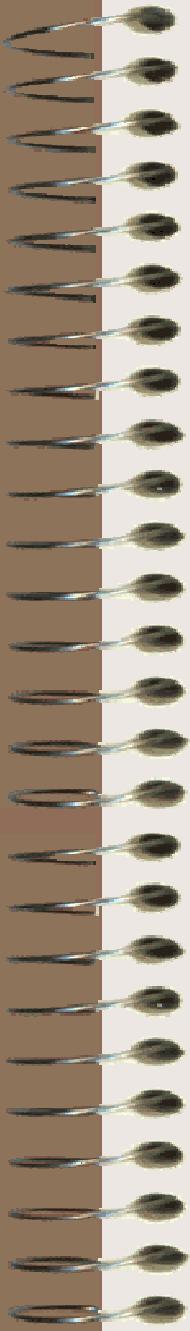
Distribución de Frecuencias: Ejemplo

Distribución de Frecuencias para Ingesta de Zinc (mg)

Intervalo de Clase	Frecuencia	Frecuencia Relativa
[3 , 6)	2	0.050
[6 , 9)	12	0.300
[9 , 12)	16	0.400
[12 , 15)	5	0.125
[15 , 18)	3	0.075
[18 , 21)	2	0.050
Total	40	1.000

Es posible sumar varias frecuencias relativas para obtener resultados como:

proporción de individuos con ingestas menores que 12 mg :
 $(0.050 + 0.300 + 0.400) = 0.750 (75\%)$



Distribución de Frecuencias

No existen reglas para seleccionar el número o la longitud de los intervalos.

Utilizar muy pocos intervalos "anchos" amontonará los datos, mientras que intervalos "angostos" puede distribuir los datos sobre demasiados intervalos con muchos intervalos con pocas observaciones.

Ninguno de los dos casos anteriores va a brindar una idea de cómo se distribuyen los datos y se pueden perder algunas características especiales de la distribución de los datos.

Una regla comunmente usada es:

$$\sqrt{\text{numero de observaciones}}$$

Frecuencias Relativas Acumuladas

Se obtienen acumulando las frecuencias relativas.

Ejemplo:

fuerza necesaria para romper una soldadura en aviones
(medida en libras).

Intervalo de Clase	Frecuencia	Frecuencia Relativa	Frecuencia Relativa Acumulada
4000 a <4200	1	0.01	0.01
4200 a <4400	2	0.02	0.03
4400 a <4600	9	0.09	0.12
4600 a <4800	14	0.14	0.26
4800 a <5000	17	0.17	0.43
5000 a <5200	22	0.22	0.65
5200 a <5400	20	0.20	0.85
5400 a <5600	7	0.07	0.92
5600 a <5800	7	0.07	0.99
5800 a <6000	1	0.01	1.00
	100	1.00	

Histogramas

Cuando los intervalos de clase de las distribuciones de frecuencias tienen la misma longitud, es relativamente fácil construir histogramas.

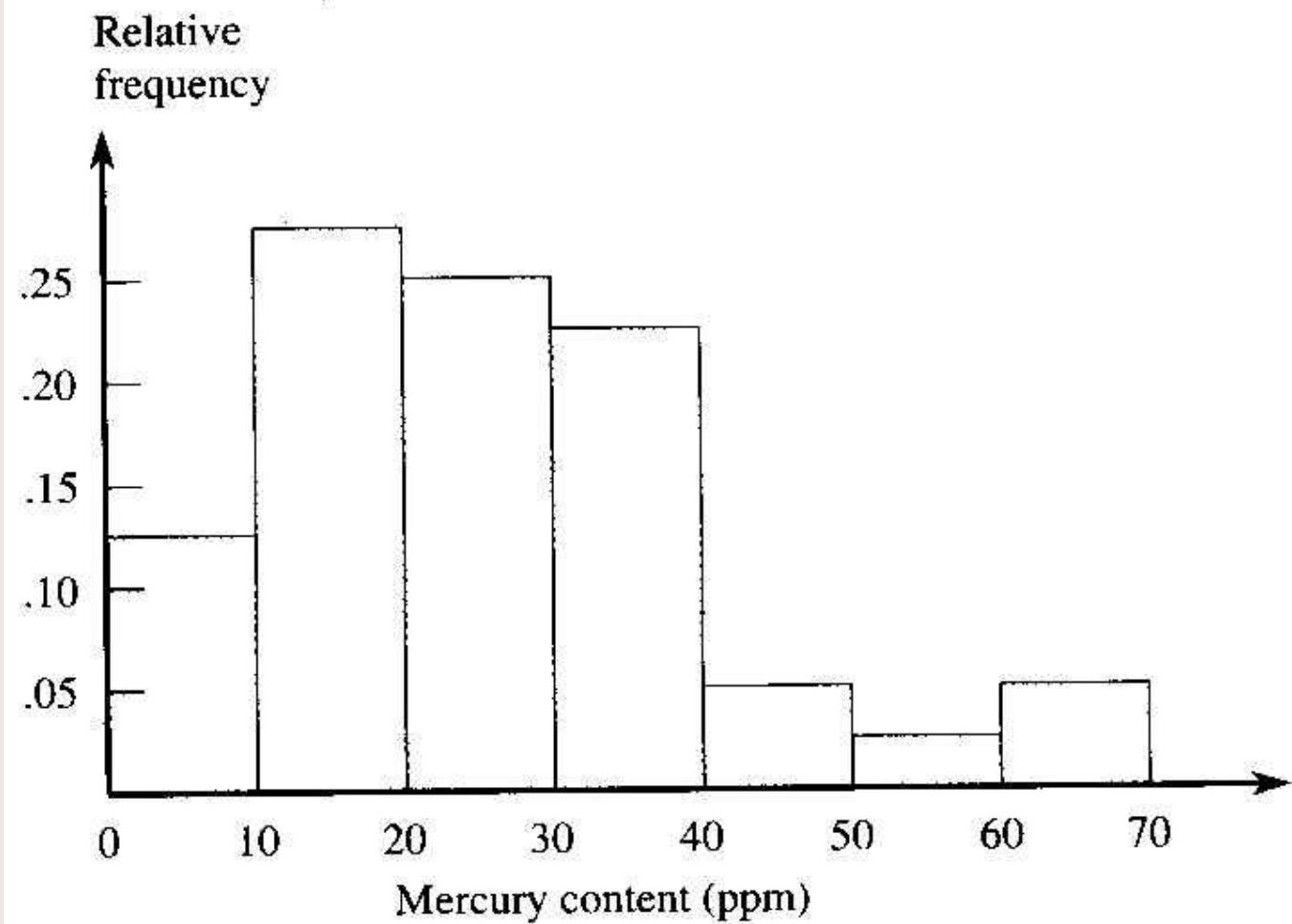
Ejemplo: La contaminación con mercurio es una preocupación ambiental muy importante. Los niveles de mercurio pueden ser particularmente altos en ciertos tipos de peces. Los habitantes de las islas Seychelles en el Océano Índico están entre los mayores consumidores de pescado en el mundo. A continuación se presenta el contenido de mercurio en una muestra de 40 pescadores de las islas Seychelles:

13.26	32.43	18.10	58.23	64.00	68.20	35.35	33.92	23.94	18.28
22.05	39.14	31.43	18.51	21.03	5.50	6.96	5.19	28.66	26.29
13.89	25.87	9.84	26.88	16.81	37.65	19.63	21.82	31.58	30.13
42.42	16.51	21.16	32.97	9.84	10.64	29.56	40.69	12.86	13.80

Histogramas: Ejemplo

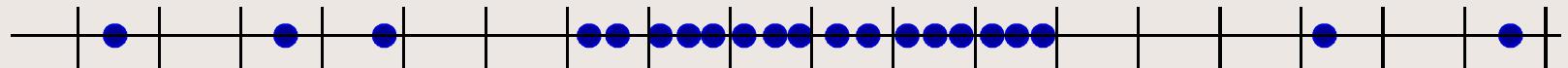
Intervalo de Clase	Frecuencia	Frecuencia Relativa
0 a <10	5	0.125
10 a <20	11	0.275
20 a <30	10	0.250
30 a <40	9	0.225
40 a <50	2	0.050
50 a <60	1	0.025
60 a <70	2	0.050
	40	1.000

Histogramas: Ejemplo

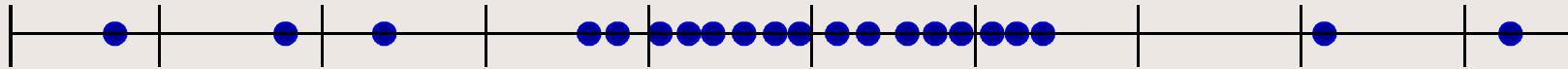


Intervalos de Clase de Ancho Desigual

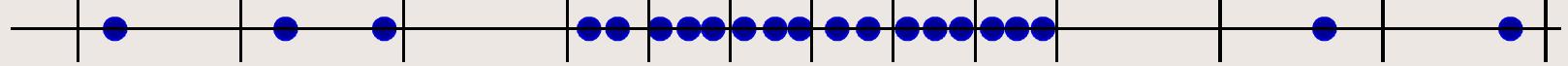
(a)



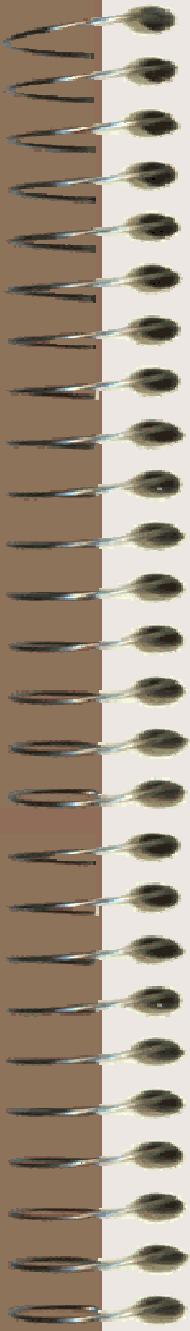
(b)



(c)



- a) Muchos intervalos angostos: unos pocos intervalos incluirán todas las observaciones, y muchos contendrán 0 observaciones.
- b) Pocos intervalos anchos: unos pocos intervalos incluirán la mayoría de las observaciones.
- c) La mejor alternativa es utilizar pocas clases (relativamente anchas) en los extremos e intervalos angostos en el centro de la distribución



Intervalos de Clase de Ancho Desigual

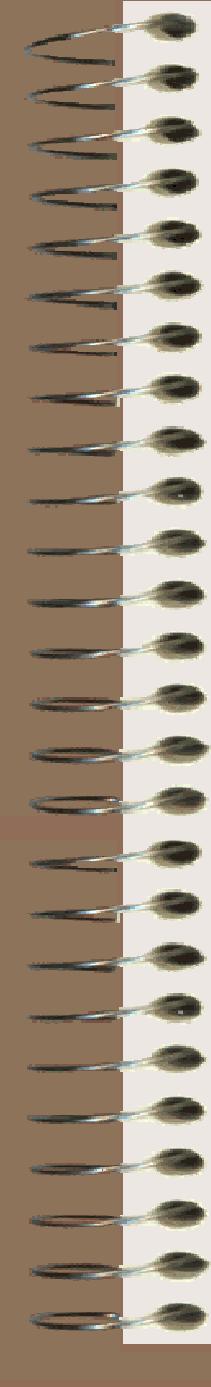
En este caso, las frecuencias o frecuencias relativas no deben usarse en los ejes verticales. En su lugar se utiliza la altura del rectángulo, llamado la "**densidad**" de la clase.

densidad = altura del rectángulo =

$$\frac{\text{frecuencia relativa}}{\text{ancho de la clase}}$$

El eje vertical se denomina la "escala de densidad".

El uso de la escala de densidad al construir el histograma asegura que el área de cada rectángulo en el histograma es proporcional a la frecuencia relativa correspondiente.



Intervalos de Clase de Ancho Desigual

La tabla siguiente presenta la distribución de la diferencia entre los notas promedio reportadas por una muestra de estudiantes y las verdaderas notas. El objetivo fue evaluar la confiabilidad de utilizar notas auto-reportadas en una investigación sobre técnicas educativas.

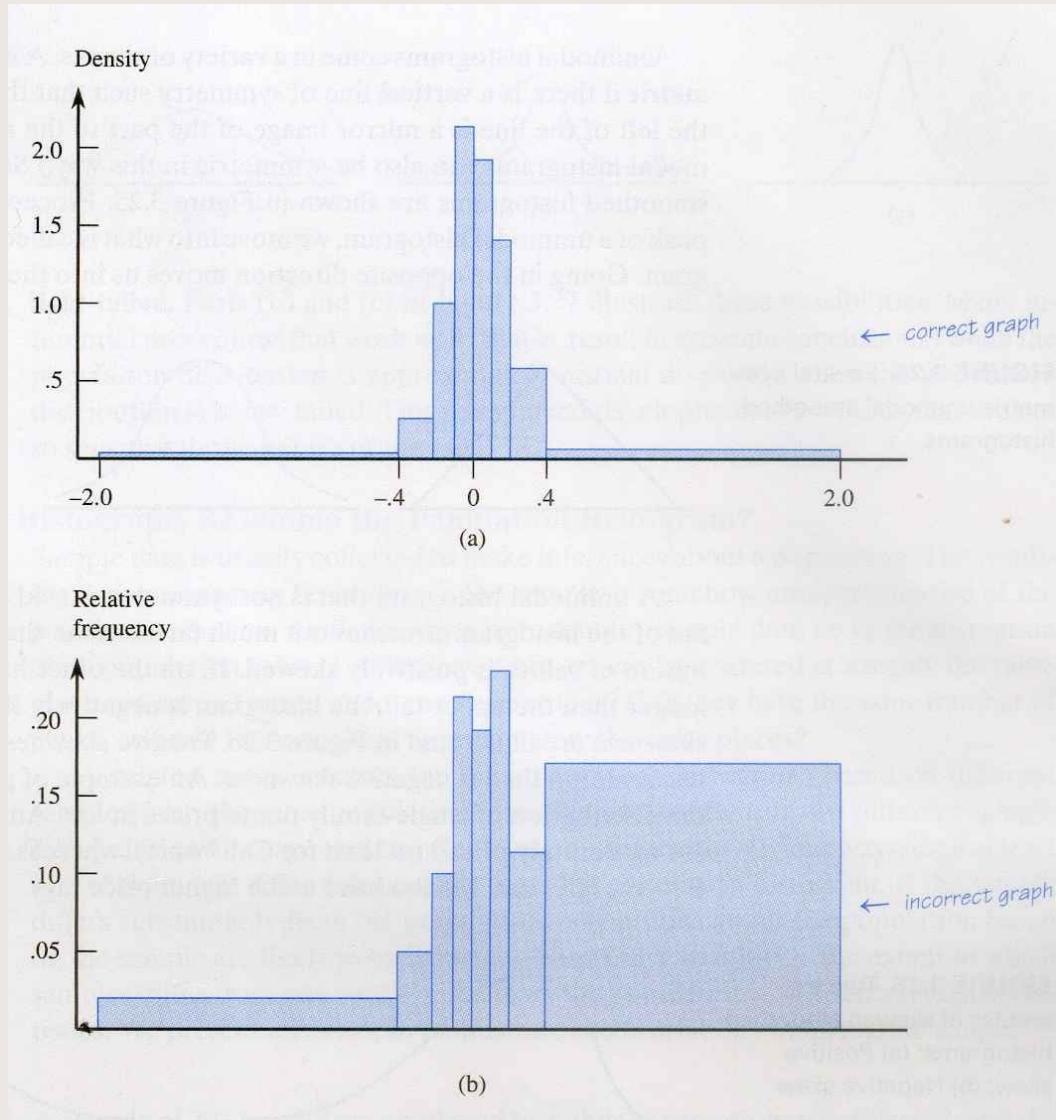
Valores positivos resultan de individuos que reportan promedios mayores a los valores correctos. La mayoría de las diferencias son cercanas a 0, pero se observaron algunas diferencias bastante grandes.

Como consecuencia de esto, una distribución con intervalos desiguales brindará un resumen conciso y a la vez informativo.

Intervalos de Clase de Ancho Desigual

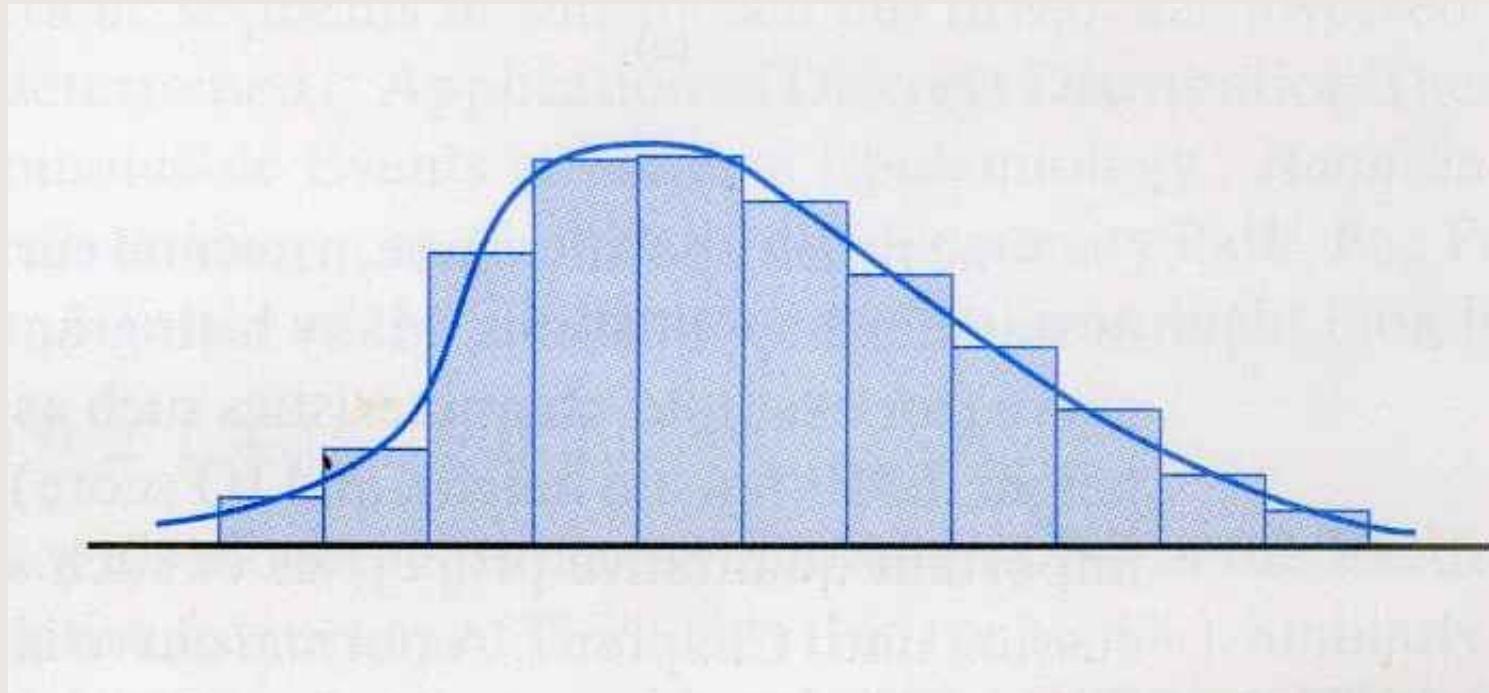
Intervalo de Clase	Frecuencia Relativa	Ancho	Densidad
-2.0 a < -0.4	0.023	1.6	0.014
-0.4 a < -0.2	0.055	0.2	0.275
-0.2 a < -0.1	0.097	0.1	0.970
-0.1 a < 0.0	0.210	0.1	2.100
0.0 a < 0.1	0.189	0.1	1.890
0.1 a < 0.2	0.139	0.1	1.390
0.2 a < 0.4	0.116	0.2	0.580
0.4 a < 2.0	0.171	1.6	0.107

Intervalos de Clase de Ancho Desigual



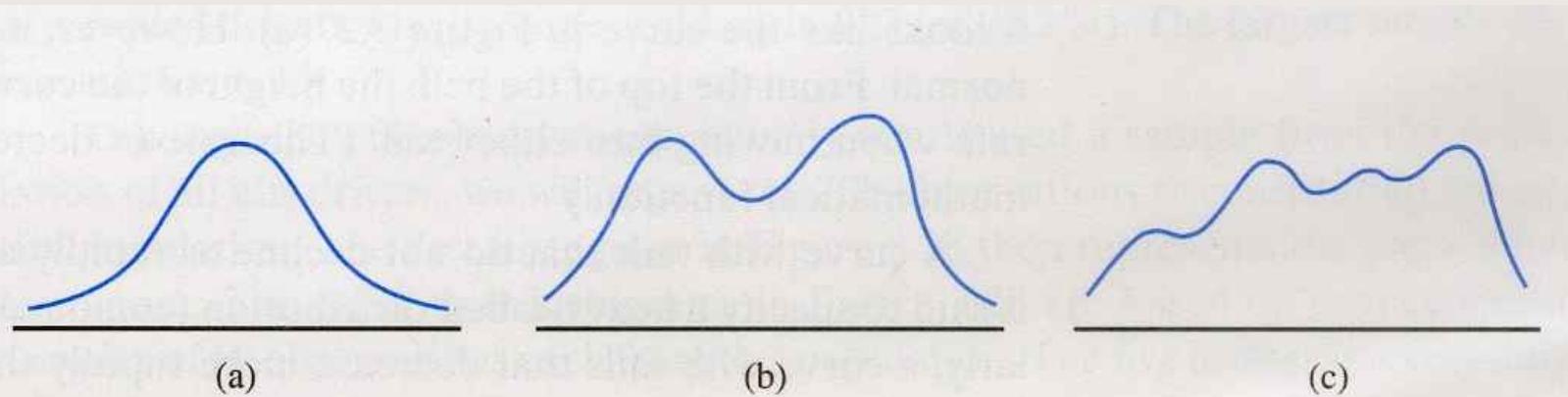
Forma de los Histogramas

La forma general de un histograma es una característica importante. Al describir las diferentes formas, es conveniente aproximar el histograma con una curva suavizada.



Forma de los Histogramas

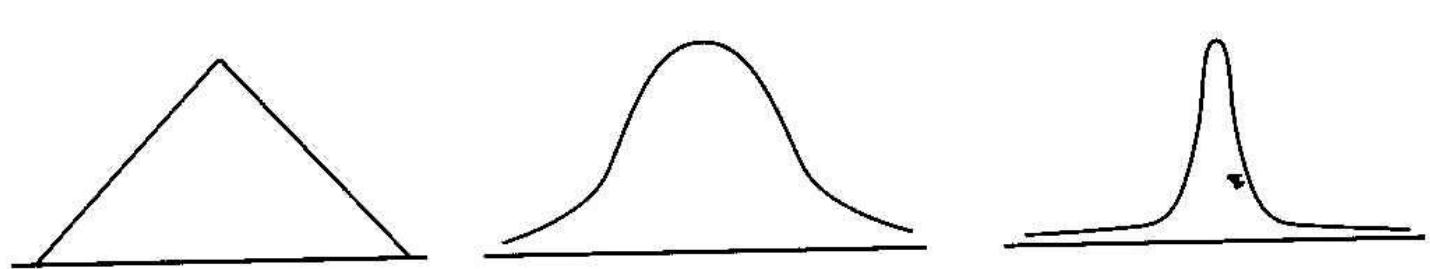
Una caracterización de la forma general se relaciona con el número de "picos" o "modos". Se dice que un histograma es "**unimodal**" si tiene un único pico, "**bimodal**" si tiene dos picos y "**multimodal**" si tiene más de dos picos.



Distribuciones bimodales se presentan cuando los datos son observaciones realizadas en dos grupos diferentes de individuos u objetos.

Forma de los Histogramas

Los histogramas unimodales son los más comunes. Un histograma es simétrico si existe una linea vertical de simetría tal que la parte del histograma a la izquierda de dicha línea es un reflejo de la parte ubicada a la derecha.

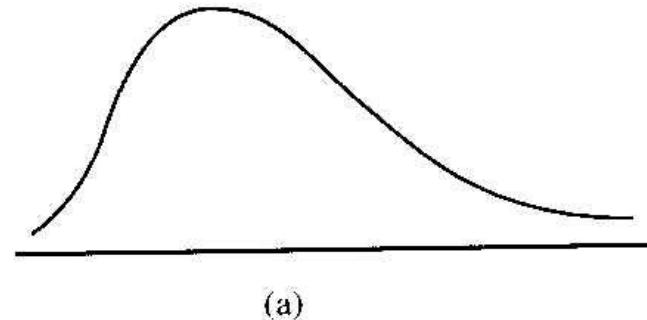


Forma de los Histogramas

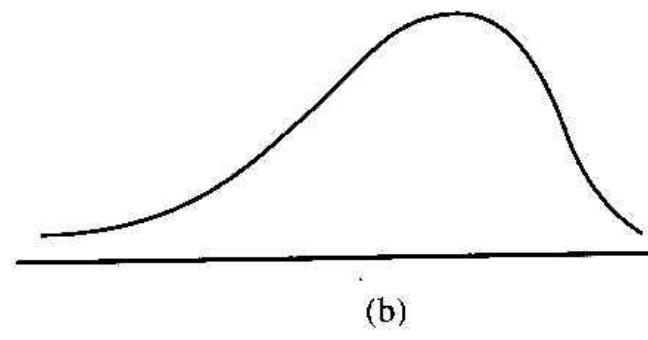
Un histograma unimodal que no es simétrico es asimétrico.

Si la cola superior del histograma se prolonga mucho más que la cola inferior, la distribución de los valores es **asimétrica positiva**.

Si la cola inferior es mucho mayor que la cola superior la distribución es **asimétrica negativa**.



(a)

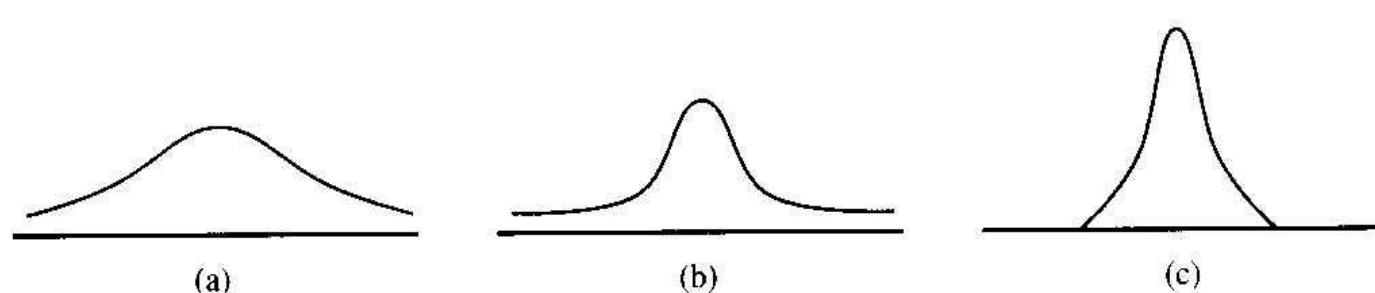


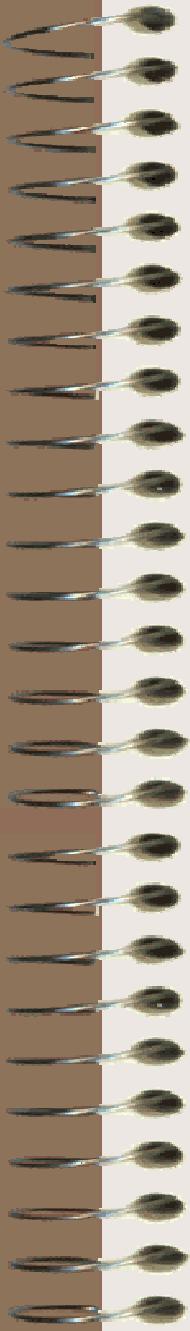
(b)

Forma de los Histogramas

Una forma específica, la curva normal, aparece muy frecuentemente. Muchos histogramas pueden ser aproximados por una curva normal. Por ejemplo, presión sanguínea, altura de hombres, altura de mujeres, resultados de un test de inteligencia).

Una curva normal no solo es simétrica sino que tiene una forma de campana.





Forma de los Histogramas

No todas las curvas con forma de campana son normales. Desde la parte superior de la campana, la curva decrece hacia ambas direcciones a una tasa definida por una función matemática.

Una curva con colas que no disminuyen tan rápidamente como las colas de una distribución normal se dice que caracterizan una distribución con "colas pesadas".

Similarmente, una curva con colas que disminuyen más rápidamente que las colas de una distribución normal se dice que proviene de una distribución con "colas livianas"

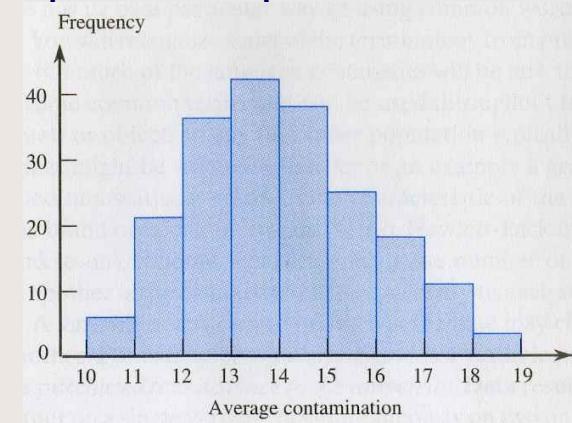
Variabilidad

Poblaciones sin variabilidad: no presentan ningún desafío desde el punto de vista estadístico. (Situación no realista)

La presencia de variabilidad en los datos condiciona la forma en la cual se recolectan, analizan y se sacan conclusiones a partir de los mismos.

Ejemplo:

Como parte de los esfuerzos para el monitoreo de la calidad del agua, un organismo de control ambiental selecciona 5 muestras de agua de distintos lugares todos los días. Se mide la concentración de contaminantes en partes por millón (ppm) en cada parte y luego se las promedia. El siguiente histograma resume la contaminación promedio para 200 días.



.....

Supongamos que ocurrió un accidente que involucró el derramamiento de elementos químicos en una planta distante a un kilómetro del lugar donde se toman las muestras.

No se sabe si un derramamiento de este tipo contaminará el agua y si dada la distancia se producirá contaminación.

Luego de un mes del incidente se extraen 5 muestras de agua,

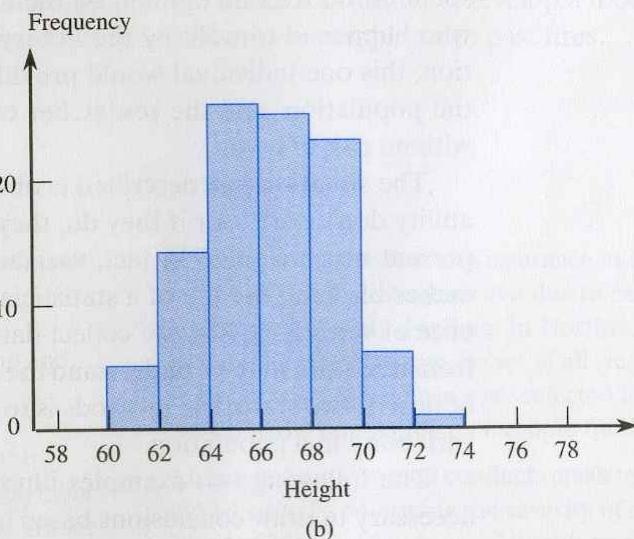
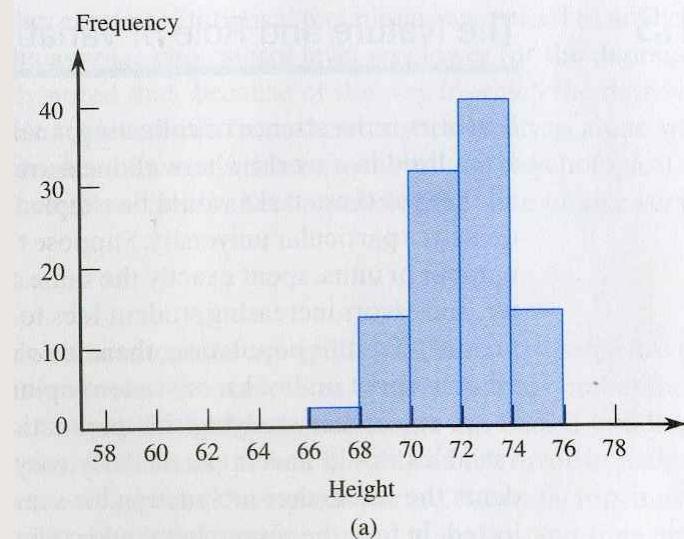
✓ Si la contaminación promedio es 16 ppm, tomaría esto como evidencia de que el incidente ha afectado la calidad del agua?

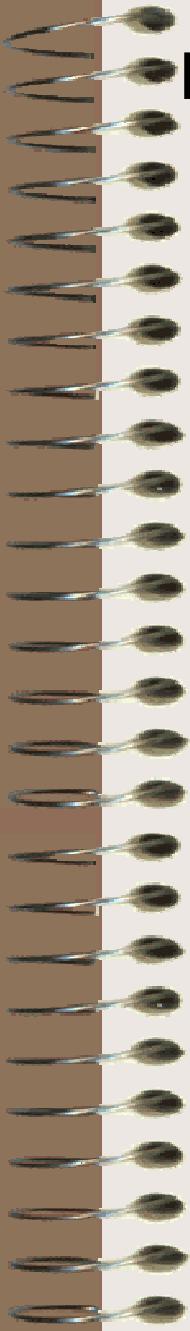
✓ ¿Qué pasaría si el promedio fuese 18 ppm o 22 ppm?

Ayuda: Observe el histograma correspondiente a las muestras tomadas antes del accidente.

Ejemplo:
se registra (en pulgadas) las alturas de los alumnos que concurren a una Universidad y se confecciona un histograma para hombres y las mujeres.

- ✓ ¿qué forma aproximada presentaría un histograma realizado sin tener en cuenta el sexo?
- ✓ ¿qué sucede en ese caso con la variabilidad presentada por los datos?





Histogramas Muestrales y Poblacionales

Las conclusiones obtenidas de datos muestrales puede ser incorrecta cuando la muestra no es representativa de la población.

Cuán diferentes o similares pueden ser los histogramas basados en una muestra del histograma poblacional?

Centrados en el mismo lugar

Igual dispersión

Igual número de picos y en los mismos lugares?

Un tema relacionado es: en qué grado histogramas basados en diferentes muestras de la misma población serán parecidos?

Si una muestra difiere substancialmente de la población, las conclusiones acerca de la población basadas en dicha muestra van a ser incorrectas.

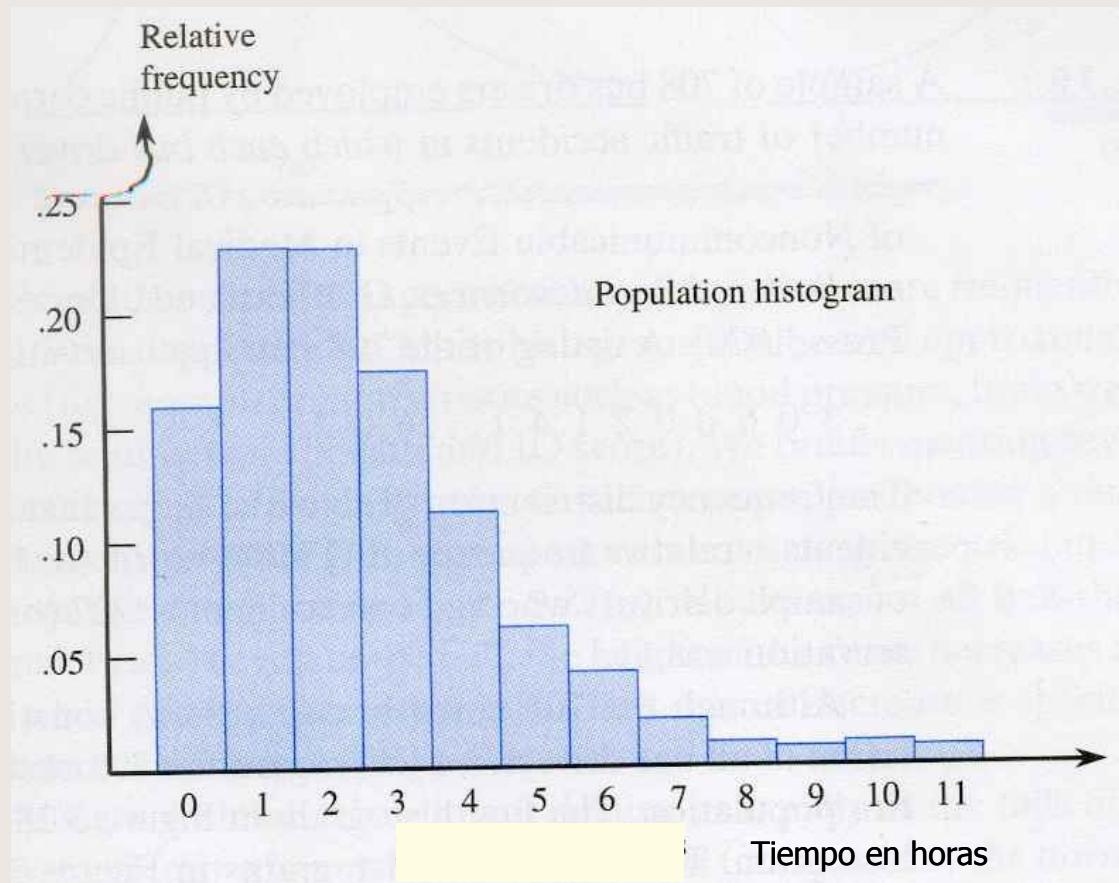
Histogramas Muestrales y Poblacionales: Ejemplo

Una muestra de 708 alumnos universitarios fue seleccionada y se registró el tiempo en horas que habían dedicado al estudio en los últimos 3 días (excluyendo las horas de asistencia a las clases)

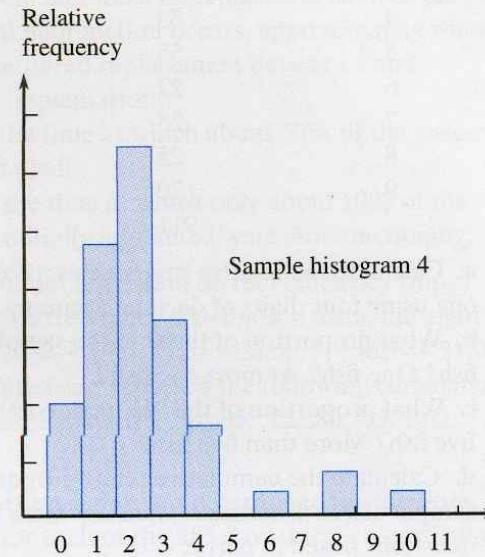
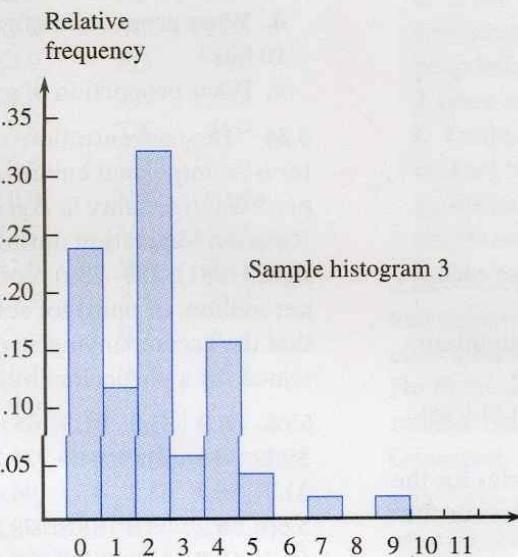
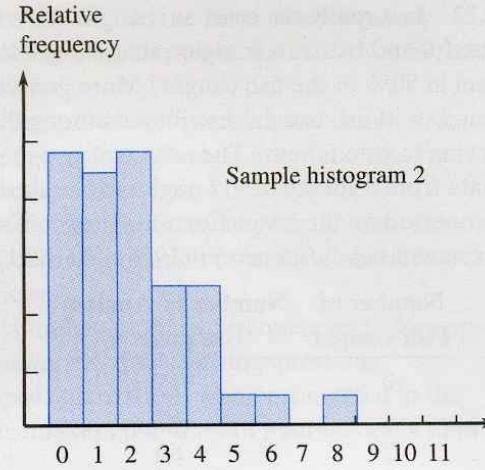
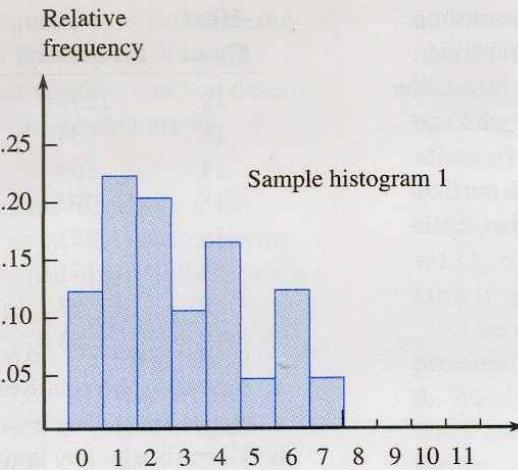
Tiempo en horas destinadas al estudio	Frecuencia	Frecuencia Relativa
[0,1)	117	0.165
[1,2)	157	0.222
[2,3)	158	0.223
[3,4)	115	0.162
[4,5)	78	0.110
[5,6)	44	0.062
[6,7)	21	0.030
[7,8)	7	0.010
[8,9)	6	0.008
[9,10)	1	0.001
[10,11)	3	0.004
[11,12)	1	0.001
Total	708	0.998

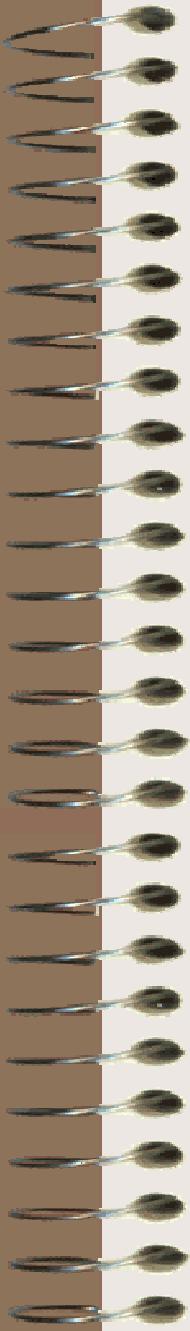
Histogramas Muestrales y Poblacionales: Ejemplo

Aunque las 708 observaciones constituyeron realmente una muestra de la población de todos los universitarios, se considerará que las 708 observaciones constituyen toda la población.



Histogramas Muestrales y Poblacionales: Ejemplo





Histogramas Muestrales y Poblacionales: Ejemplo

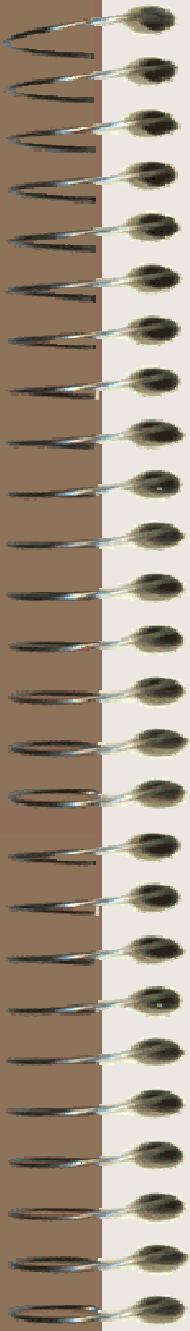
Los cuatro histogramas anteriores se basan en cuatro muestras diferentes de 50 observaciones.

Los cinco histogramas se parecen entre si en formas generales, pero también presentan ciertas diferencias obvias.

El histograma poblacional crece hasta alcanzar un pico y luego disminuye suavemente, mientras que los histogramas muestrales tienden a tener más picos, valles y lagunas (gaps).

Aunque el conjunto de datos poblacional contiene una observación igual a 11, ninguna de las 4 muestras contiene un valor de 11.

En clases siguientes se verá como la variabilidad muestral puede ser descripta e incorporada en las conclusiones basadas en métodos inferenciales.



Interpretación de los Resultados de Análisis Estadísticos

Algunos puntos a tener en cuenta cuando se analiza un gráfico:

- ✓ El gráfico elegido es apropiado para el tipo de datos recolectados?
- ✓ Cómo describiría la forma de la distribución, y qué se podría decir acerca de la variable analizada?
- ✓ Existen outliers en el conjunto de datos? Existe una explicación convincente de porqué pueden diferir del resto de los datos?
- ✓ Dónde caen la mayoría de los datos? Cuál es un valor típico para la variable?
- ✓ Existe mucha variabilidad en los datos?

Para pensar!!!!!!

La exposición a determinadas sustancias químicas puede representar un riesgo para la salud. Evidencia recolectada en los últimos años sugiere que una alta concentración de gas radón en el interior de hogares podría estar relacionada con el desarrollo de cáncer en niños.

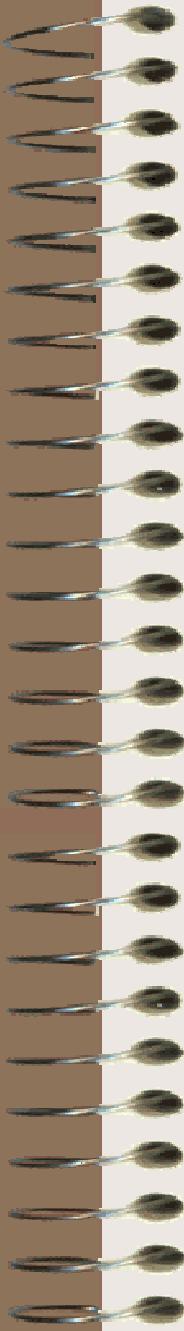
En un articulo aparecido en una publicación científica se presentan los siguientes datos de concentración de radón (Bq/m^3) en dos muestras diferentes de hogares.

**Hogares de Niños
con Cancer**

10	21	5	23	15	11	9
13	27	13	39	22	7	20
45	12	15	3	8	11	18
16	23	16	34	10	15	11
18	210	22	11	6	17	33
10	9	57	16	21	18	38

**Hogares de Niños
sin Cancer**

9	38	11	12	29	5	7
6	8	29	24	12	17	11
3	9	33	17	55	11	29
13	24	7	11	21	6	39
29	7	8	55	9	21	9
3	85	11	14			



Preguntas frecuentes:

Cuál es un concentración de radón típica o representativa para cada muestra?

Presentan ambas muestras la misma variabilidad o resultan más dispersas las observaciones de una de las muestras?

Los métodos gráficos como los diagramas de tallo y hoja y los histogramas dan una idea acerca del centro de la distribución y cómo se distribuyen los datos alrededor del centro.

En esta clase se presentarán medidas resumen numéricas que describen más precisamente el centro y el grado de dispersión.

Descripción del Centro de un Conjunto de Datos

Las dos medidas más populares para describir el centro de una distribución son la media (o promedio) y la mediana.

x = variable para la cual se cuenta con datos muestrales

n = número de observaciones muestrales (tamaño muestral)

x_1 = primera observación muestral

x_2 = segunda observación muestral

x_n = n -ésima (última) observación muestral

Media Muestral

La media muestral de una muestra numérica x_1, x_2, \dots, x_n es:

$$\bar{x} = \frac{\text{suma de todas las observaciones en la muestra}}{\text{números de observaciones en la muestra}} = \frac{\sum_{i=1}^n x_i}{n}$$

Una medida de recuperación luego de una operación para reparar ligamentos rotos en la rodilla es el rango de movimiento, medido en grados. Los siguientes son los datos pertenecientes a 13 pacientes:

$x_1 = 154$	$x_2 = 142$	$x_3 = 137$	$x_4 = 133$	$x_5 = 122$	$x_6 = 126$	$x_7 = 135$
$x_8 = 135$	$x_9 = 108$	$x_{10} = 120$	$x_{11} = 127$	$x_{12} = 134$	$x_{13} = 122$	

La media muestral para el rango de movimiento es:

$$\bar{x} = \frac{\sum x}{n} = \frac{1695}{13} = 130.38$$

Media Muestral

Los valores en la muestra eran todos enteros, sin embargo la media se reportó como: 130.38. Es común utilizar más dígitos de precisión decimal cuando se reporta la media.

La media muestral se calcula a partir de observaciones muestrales y por lo tanto es una característica de la muestra.

Generalmente se utilizan caracteres romanos para representar características de la muestra. Para características de la población se utilizan letras griegas.

La media poblacional, simbolizada con μ , es el promedio de todos los valores de x en la población.

El valor de la media muestral varía de muestra en muestra mientras que solo existe un único valor de μ .

Media Muestral

Un problema potencial de la media como medida de posición central es que su valor puede estar muy influido por la presencia de observaciones extremas (outliers).

Tasa de interés para 15 instituciones bancarias.

2.15	2.20	2.20	2.22	2.22	2.23	2.25	2.25	2.27	2.28
2.28	2.28	2.29	2.30	2.30	2.32	2.32	2.33	2.33	2.33
2.33	2.33	2.33	2.38	2.43	2.55	2.79	3.05	3.68	4.35

El valor de la media muestral es 2.45.

Este valor es mayor que 25 de las 30 observaciones en la muestra. Por lo tanto se podría pensar que no es un valor demasiado representativo de toda la muestra.

Mediana

Luego de que los datos son ordenados de menor a mayor, la mediana es el valor central de la lista y divide la lista en dos partes iguales.

La mediana se calcula en forma diferente dependiendo si n es par o impar.

Cuando n es impar, la mediana muestral es un único valor central. Cuando n es par, existen dos valores centrales en la lista ordenada, y la mediana muestral resulta el promedio de los dos valores centrales.

mediana
muestral

{ el valor central si n es impar
el promedio de los dos valores centrales si n es par

Mediana

Para los datos de tasas de interés, el tamaño muestral fue $n = 30$. La mediana es el promedio de las observaciones 15 y 16 en el conjunto de datos ordenados.

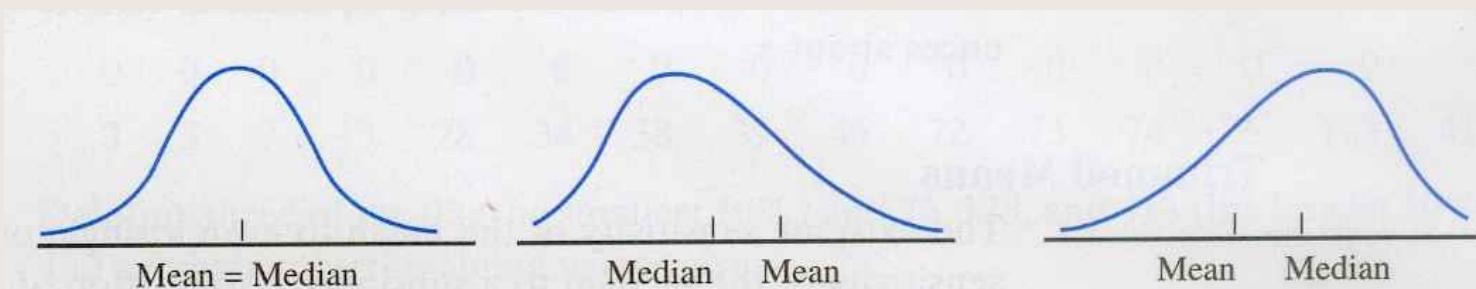
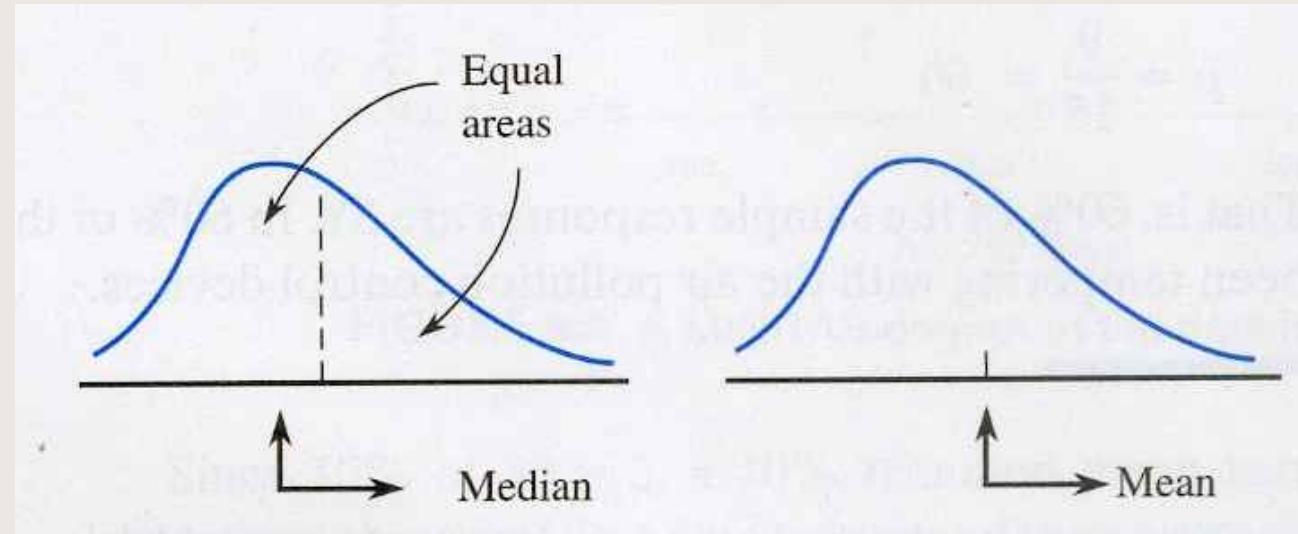
$$\text{mediana muestral} = \frac{2.30 + 2.32}{2} = 2.31$$

Este valor resulta ser más típico del conjunto de datos que el valor de la media (2.45).

La mediana poblacional cumple el mismo rol para la población que la mediana muestral en la muestra.

Si existen valores extremos, la media se desplaza hacia dichos valores mientras que la mediana es insensible a tales valores.

Comparación de la Media y la Mediana



Datos Categóricos

Las medidas resumen para un conjunto de datos categórico son las frecuencias relativas para las distintas categorías.

Cada frecuencia relativa es la proporción (fracción) de respuestas que caen en la categoría correspondiente.

Cuando solo existen dos categorías (variable dicotómica), es común asignar a una de las categorías el rótulo "éxito" (o "si") y reportar solamente la proporción de éxitos en la muestra (proporción muestral).

La proporción muestral de éxitos, simbolizada con \hat{p} , es

$$\hat{p} = \frac{\text{número de éxitos en la muestra}}{n}$$

donde "éxitos" representa una de las dos categorías de la variable dicotómica.

Datos Categóricos: Ejemplo

El uso de equipamiento contra la contaminación en automóviles ha mejorado la calidad del aire en ciertas áreas. Desafortunadamente, muchos propietarios de automóviles modifican los mecanismos contra la contaminación para obtener mejor rendimiento en sus automóviles. Suponer que se selecciona una muestra de 15 automóviles y se clasifica como E (éxito), si los mecanismos han sido modificados.

E	F	E	E	E	F	F	E	E	F	E	E	E	F	F
---	---	---	---	---	---	---	---	---	---	---	---	---	---	---

La muestra contiene 9 E's, por lo tanto

$$\hat{p} = 9 / 15 = 0.60$$

Para representar la proporción poblacional se utiliza la letra griega π .

Describiendo la Variabilidad

Una medida de posición central solo brinda información parcial. Es importante describir también la dispersión de los valores alrededor del centro.

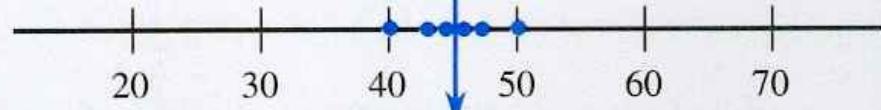
1. 20, 40, 50, 30, 60, 70



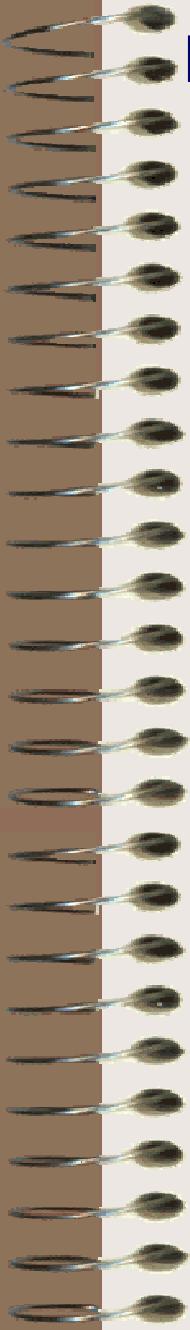
2. 47, 43, 44, 46, 20, 70



3. 44, 43, 40, 50, 47, 46

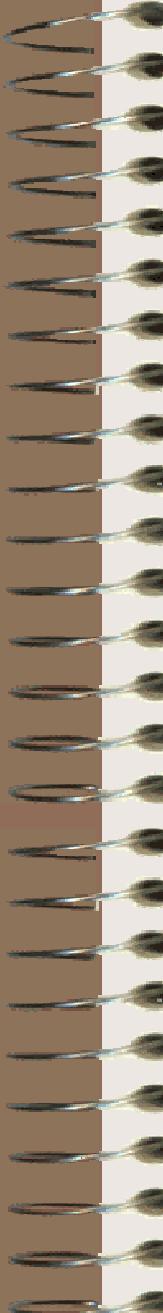


Mean = Median



Describiendo la Variabilidad

- La medida de variabilidad más simple es el rango, definido como la diferencia entre el valor más grande y el valor más pequeño.
- En general, mayor variabilidad se reflejará en un mayor valor del rango.
- Sin embargo, la variabilidad no solo depende de los valores extremos. La variabilidad es una característica del conjunto de datos completo y cada observación contribuye a la variabilidad.
- En la figura anterior, las dos primeras muestras tienen el mismo rango pero existe menos dispersión en la segunda muestra.



Desviaciones de la Media

Las medidas más comunes de variabilidad describen el grado en el cual las observaciones muestrales se alejan de la media muestral.

Substrayendo la media muestral de cada observación se obtiene un conjunto de desviaciones de la media.

Una desviación es positiva si los valores de x exceden \bar{x} y es negativa si los valores de x son menores que \bar{x} .

En el ejemplo de las tasas de interés, $\bar{x} = 2.452$ y solamente 5 desviaciones resultan positivas, ya que solo 5 observaciones exceden \bar{x} .

Generalmente, a mayor variabilidad en la muestra, mayor es la magnitud (ignorando el signo) de las desviaciones.



Desviaciones de la Media

Una forma de resumir las desviaciones de la media en una única medida podría ser promediarlas a través de toda la muestra.

Sin embargo, salvo por efectos de redondeo, la suma de las desviaciones de la media siempre es 0 y por lo tanto su promedio es 0, y no puede ser utilizada como medida de variabilidad.

La forma de prevenir que desviaciones negativas y positivas se compensen es elevándolas al cuadrado antes de combinarlas.

Variancia Muestral y Desvio Estándar Muestral

La variancia muestral, simbolizada por s^2 , es la suma de las desviaciones de la media al cuadrado dividida por $n - 1$. Es decir,

$$s^2 = \frac{\sum(x - \bar{x})^2}{n-1} = \frac{S_{xx}}{n-1}$$

El desvio estándar muestral es la raiz cuadrada positiva de la variancia muestral, simbolizada por s .

Una gran variabilidad en la muestra se reflejará en un valor relativamente grande de s^2 o s .

Notar que el desvio estándar se expresa en la misma unidad de medida que la variable original.

Ejemplo: Rango de Movimiento

Observación	Desviación	Desviación al Cuadrado
154	23.62	557.9044
142	11.62	135.0244
137	6.62	43.8244
133	2.62	6.8644
122	-8.38	70.2244
126	-4.38	19.1844
135	4.62	21.3444
135	4.62	21.3444
108	-22.38	500.8644
120	-10.38	107.7444
127	-3.38	11.4244
134	3.62	13.1044
122	-8.38	70.2244
Suma = 1579.0772		

Ejemplo: Rango de Movimiento

Por lo tanto,

$$\sum(x - \bar{x})^2 = 1579.0772$$

$$s^2 = \frac{\sum(x - \bar{x})^2}{n-1} = \frac{1579.0722}{12} = 131.5898$$

$$s = \sqrt{131.5898} = 11.4713$$



Interpretación y Propiedades

Un desvío estándar puede ser interpretado informalmente como la magnitud de una desviación "típica" de la media.

Por lo tanto, en el ejemplo anterior, una desviación típica de la media 130.38 es 11.47. Algunas observaciones estarán más cerca de 130.38 y otras más alejadas.

Se calculó el valor de $s = 11.47$ pero no se señalo si dicho valor representa una gran variabilidad o una pequeña variabilidad.

Por ahora, se utilizará s solo con fines comparativos en lugar de utilizarlo como una medida de variabilidad absoluta.

Por ejemplo, si en otra muestra de rangos de movimiento se obtiene $s = 9.1$, se podría concluir que existe más variabilidad en la primera muestra que en la segunda.



Interpretación y Propiedades

- Existen medidas de variabilidad para la población completa que son análogas a s^2 y s para la muestra.
- Estas cantidades se denominan variancia poblacional (σ^2) y desvío estándar poblacional (σ).
- En muchos procedimientos estadísticos vamos a querer utilizar el valor de σ , pero generalmente dicho valor no se conoce.
- En su lugar se utiliza un valor calculado a partir de la muestra el cual se espera que sea cercano a σ (es decir, sea una buena estimación de σ).
- Se utiliza el divisor $(n - 1)$ en s^2 en lugar de n porque, en promedio, tiende a estar un poco más cercano a σ^2 .

Rango Intercuartil

Al igual que la media, el valor de s resulta muy afectado por la presencia de observaciones extremas.

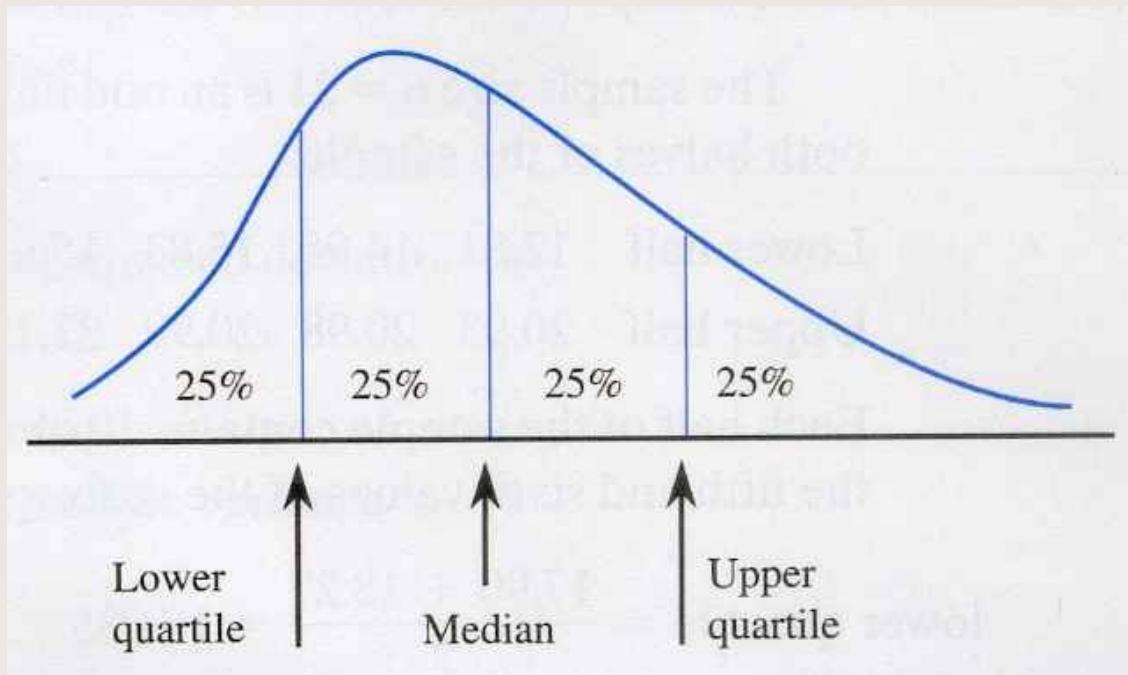
El rango intercuartil es una medida de variabilidad que es resistente a los efectos de los "outliers". Se basa en cantidades denominadas "cuartiles".

El cuartil inferior separa el 25% inferior del conjunto de datos del 75% superior.

El cuartil superior separa el 25% superior del conjunto de datos del 75% inferior.

El cuartil central es la mediana la cual separa el 50% inferior del 50% superior.

Cuartiles



Los cuartiles de una muestra se obtienen dividiendo las n observaciones ordenadas en una mitad inferior y una mitad superior; si n es impar la mediana se excluye de ambas mitades. Los dos cuartiles extremos son las medianas de las dos mitades.



Rango Intercuartil

El rango intercuartil (iqr), se obtiene como:

$$\text{iqr} = \text{cuartil superior} - \text{cuartil inferior}$$

La naturaleza resistente del iqr se deriva del hecho de que hasta el 25% de las observaciones más pequeñas de la muestra y hasta el 25% de las observaciones más grandes de la muestra pueden hacerse más extremas sin que el valor de iqr resulte afectado.

Rango Intercuartil

El rango intercuartil poblacional es la diferencia entre los cuartiles superior e inferior poblacionales.

Si un histograma de una conjunto de datos puede ser aproximado razonablemente bien por una curva normal, entonces la relación entre el desvío estándar y el rango intercuartil es aproximadamente $s = \text{iqr}/1.35$.

Un valor del desvío estándar mucho mayor que $\text{iqr}/1.35$ sugiere un histograma con colas pesadas (o más largas) que una curva normal.



Boxplots

Hasta aqui se presentaron medidas numéricas para describir el centro y la variabilidad de un conjunto de datos.

Un boxplot es un diagrama que permite resumir un conjunto de datos brindando mas detalle que las medidas de posición y dispersión pero menos detalle que un diagrama de tallo y hoja.

El boxplot es una representación compacta que provee información acerca del centro, la dispersión y la simetría o asimetría de los datos.

Un boxplot se basa en la mediana y el rango intercuartil en lugar de la media y el desvio estándar.

Boxplots: Ejemplo

Tasa de Interés

2.15	2.20	2.20	2.22	2.22	2.23	2.25	2.25	2.27	2.28
2.28	2.28	2.29	2.30	2.30	2.32	2.32	2.33	2.33	2.33
2.33	2.33	2.33	2.38	2.43	2.55	2.79	3.05	3.68	4.35

Para construir el boxplot se necesita la siguiente información:

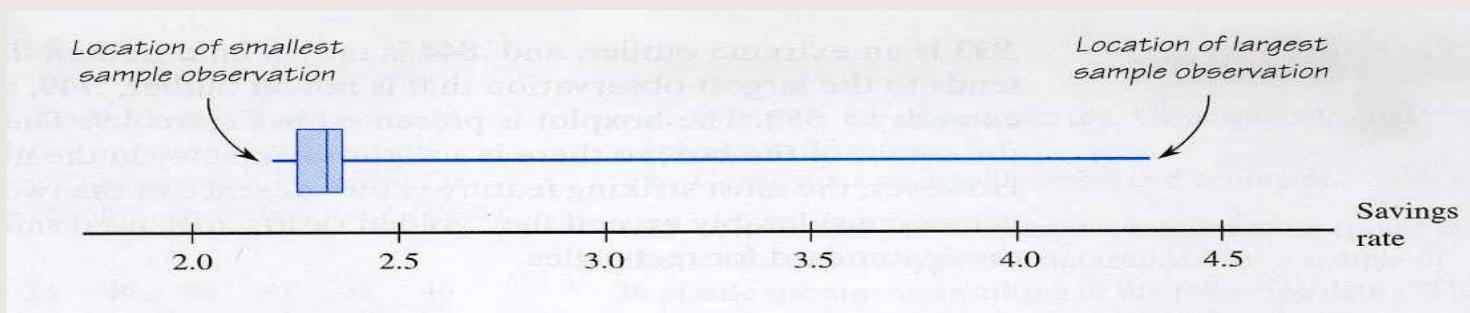
la menor observación = 2.15

cuartil inferior = mediana de la mitad inferior = 2.25

mediana = promedios de las observaciones 15 y 16 en la lista ordenada = 2.31

cuartil superior = mediana de la mitad superior = 2.33

la mayor observación = 4.35



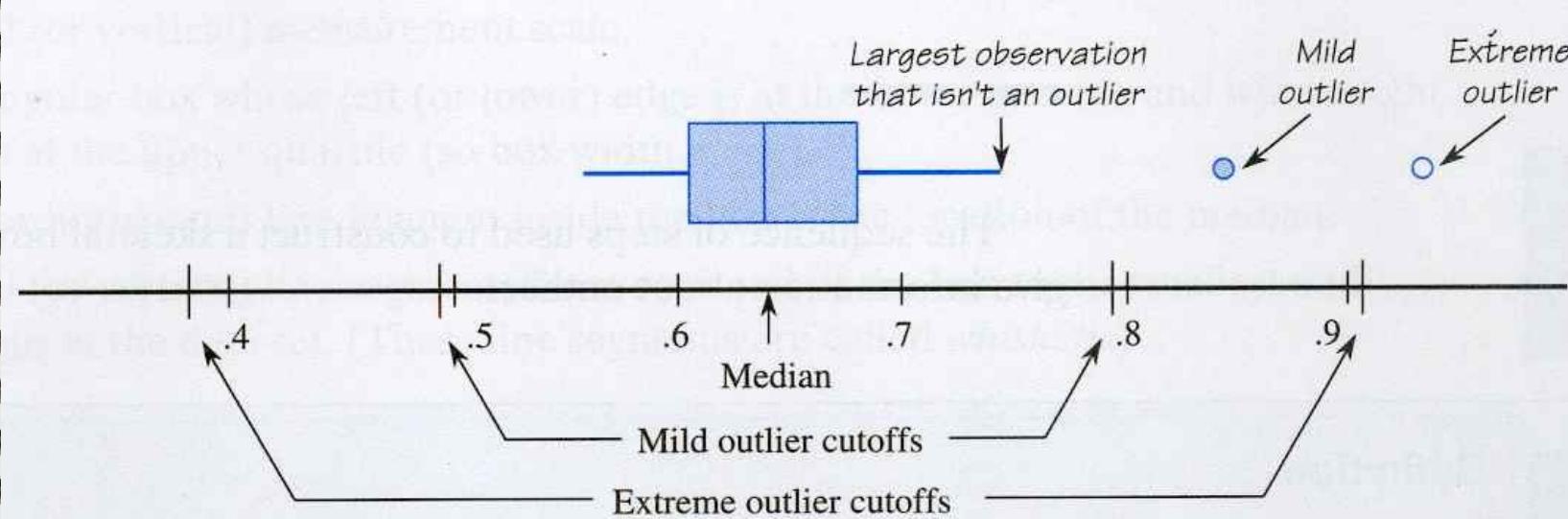
Boxplot Modificado

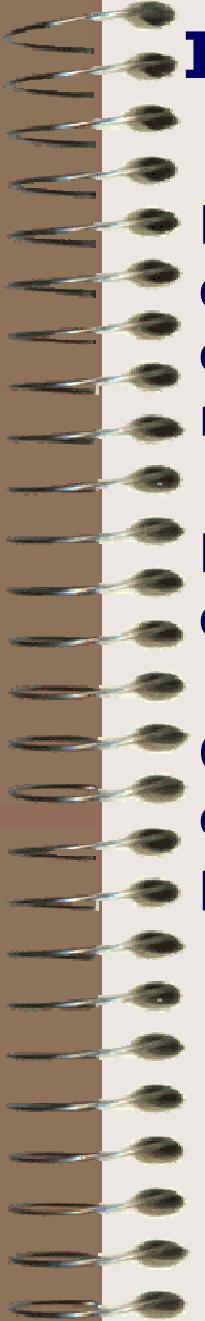
Una observación es un outlier si se encuentra a más de 1.5 veces el rango intercuartil del extremo más cercano de la caja.

Un outlier es extremo si se encuentra a más de 3 veces el rango intercuartil del extremo más cercano de la caja en caso contrario, el outlier se considera moderado.

Un boxplot modificado utiliza líneas que se extienden desde los extremos de la caja hasta las observaciones más extremas que no son outliers. Los outliers se identifican con puntos.

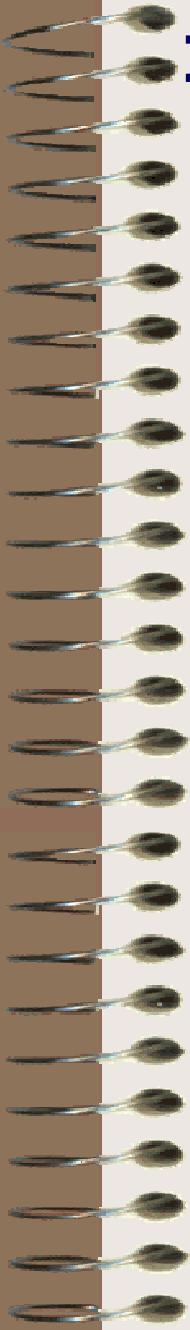
Boxplot Modificado





Interpretación de Centro y Variabilidad

- La media y el desvío estándar pueden combinarse para obtener información acerca de cómo los valores en un conjunto de datos se distribuyen y acerca la posición relativa de un valor particular en un conjunto de datos.
- Esto se logra describiendo cuán alejada se encuentra una observación de la media en términos de desvíos estándares.
- Considerar un conjunto de datos con puntajes de un test estandarizado con media 100 y desvío estándar 15. Se puede afirmar lo siguiente:
 - Cómo $100 - 15 = 85$, se dice que un puntaje de 85 se encuentra un desvío estándar por debajo de la media. Similarmente, un puntaje de 115 se encuentra un desvío estándar por encima de la media.



Interpretación de Centro y Variabilidad

Puntajes entre 70 y 130 se encuentran dentro de dos desvíos estándares de la media.

Puntajes mayores que 145 exceden la media por más de 3 desvíos estándares.

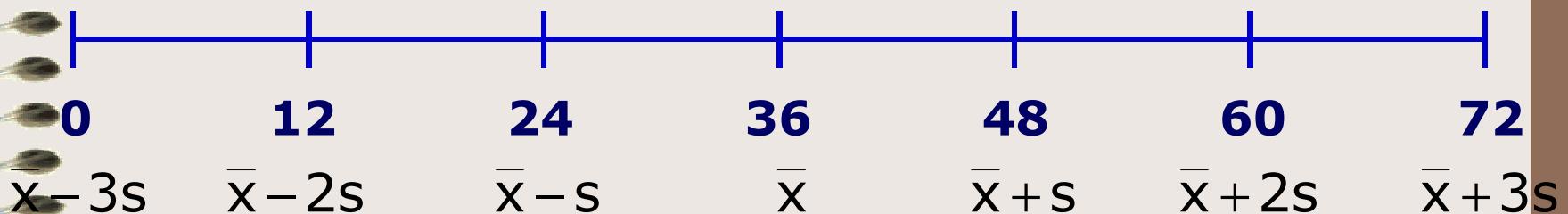
Regla de Chebyshev

Considerar cualquier número k , con $k \geq 1$. Entonces el porcentaje de observaciones que se encuentran dentro de k desvios estándares de la media es al menos $100(1 - 1/k^2)\%$

k	$1 - 1/k^2$	% dentro de k desvios de la media
2	$1 - \frac{1}{4} = 0.75$	al menos 75%
3	$1 - \frac{1}{9} = 0.89$	al menos 89%
4	$1 - \frac{1}{16} = 0.94$	al menos 94%
4.472	$1 - \frac{1}{20} = 0.95$	al menos 95%
5	$1 - \frac{1}{25} = 0.96$	al menos 96%
10	$1 - \frac{1}{100} = 0.99$	al menos 99%

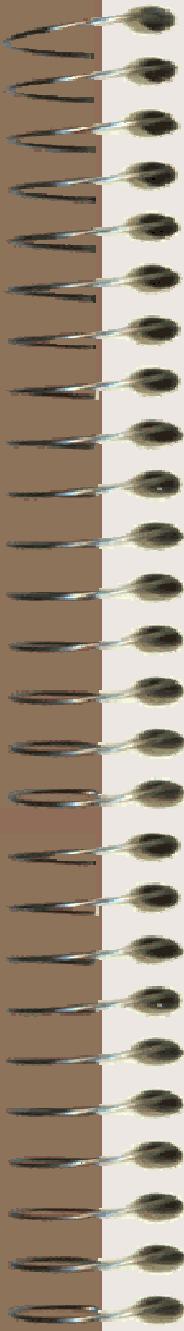
Regla de Chebyshev: Ejemplo

Para una muestra de familias se reportó que los tiempos dedicados al cuidado de niños tenían una media de 36 horas por semana y un desvío estándar de 12.



Al menos 75% de las observaciones muestrales deben estar entre 12 y 60 horas.

Como al menos 89% de las observaciones deben estar entre 0 y 72, a lo sumo 11% estarán afuera de este intervalo y como el tiempo no puede ser negativo, se puede concluir que a lo sumo el 11% de las observaciones excede 72.



Regla de Chebyshev

Como la regla de Chebyshev es aplicable a cualquier conjunto de datos, ya sea simétrica o asimétrica, se debe tener cuidado cuando se describe la proporción de observaciones ubicadas por encima de determinado valor o por debajo de determinado valor.

Los valores 18 y 52 se encuentran a 1.5 desviós a ambos lados de la media, por lo tanto usando $k = 1.5$ en la regla de Chebyshev implica que al menos 55.6% de las observaciones deben ubicarse entre estos dos valores. Por lo tanto, a lo sumo 44.4% de las observaciones son menores que 18 (no 22.2%).

Si bien la regla establece que al menos 75% de las observaciones se encuentran dentro de 2 desviós de la media, en muchos conjuntos de datos una proporción mayor de las observaciones se encontrará dentro de 2 desviós de la media.

Regla de Chebyshev

6	1	
7	25679	
8	0000124555668	
9	0000112333446666778889	
10	0001122222333566677778899999	
11	00001122333344444477899	
12	01111123445669	
13	006	
14	26	Stem: Tens
15	2	Leaf: Ones

112 puntajes de test de inteligencia en niños.

$$\bar{x} = 104.5$$

$$s = 16.3$$

$$2s = 32.6$$

$$3s = 48.9$$

Regla de Chebyshev

k	media \pm k desvio	Chebyshev	Real
2	71.9 a 137.1	al menos 75%	96% (108)
2.5	63.7 a 145.3	al menos 84%	97% (109)
3	55.6 a 153.4	al menos 89%	100% (112)



Regla Empírica

El hecho de que afirmaciones basadas en la regla de Chebyshev sean frecuentemente muy conservadoras sugiere que se debería buscar reglas que sean menos conservadoras y más precisa.

La regla más útil es la llamda regla empírica que puede ser aplicada siempre que la distribución de los datos pueda ser descripta por la distribución normal.

Si el histograma de los valores de un conjunto de datos puede ser razonablemente aproximado por una curva normal, entonces

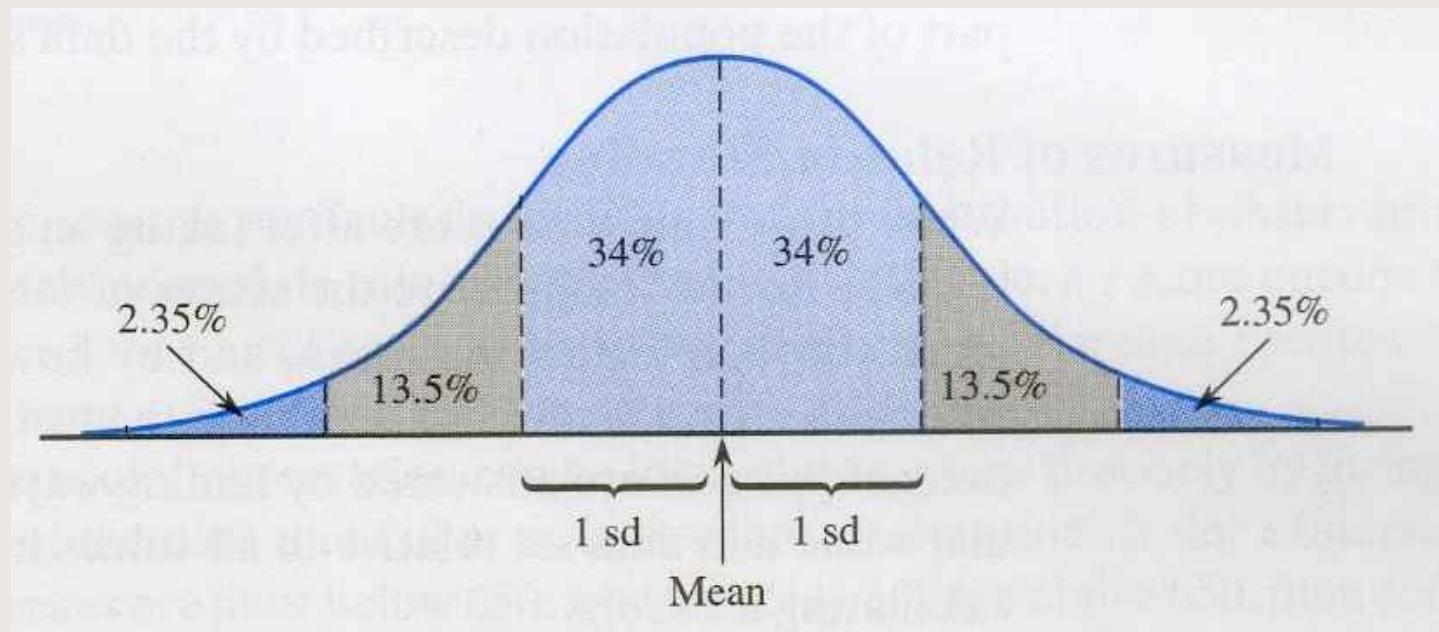
Aproximadamente 68% de las observaciones están dentro de un desvio estándar de la media.

Aproximadamente 95% de las observaciones están dentro de dos desvios estándares de la media.

Aproximadamente 99.7% de las observaciones están dentro de tres desvios estándares de la media.

Regla Empírica

La regla empírica utiliza "aproximadamente" en lugar de "al menos" y los porcentajes para $k = 1, 2$ y 3 desviaciones estándares son mucho mayores que los utilizados en la regla de Chebyshev.



Regla Empírica

A partir de un conjunto de datos correspondiente a 1052 mujeres se obtuvo que la altura promedio es 62.484 pulgadas y el desvío estándar 2.390 pulgadas.

La curva normal resultó un buen ajuste para los datos.

Desviós	Intervalo	Real	Regla Empírica	Regla de Chebyshev
1	60.094 a 64.874	72.1%	68%	al menos 0%
2	57.704 a 67.264	96.2%	95%	al menos 75%
3	55.314 a 69.654	99.2%	99.7%	al menos 89%

Medidas de Ubicación Relativa

Una puntaje z (o score z), correspondiente a una observación particular en un conjunto de datos es

$$\text{score } z = \frac{\text{observación} - \text{media}}{\text{desvio estándar}}$$

El score z indica a cuantos desvios estándares se encuentra la observación de la media. Es positivo o negativo de acuerdo si la observación se encuentra por encima o por debajo de la media.

El proceso de restar la media y luego dividir por el desvio estándar se denomina estandarización.

Z-Score: Ejemplo

Suponer que dos profesionales recien graduados comparan ofertas de trabajo. La profesión del primero es contador y recibió de 35000 dólares por año y el segundo que es licenciado en administración tiene una oferta de 33000 dólares por año.

La información con respecto a la distribución de las ofertas de trabajo es la siguiente:

Contador:	promedio = 36000	desvio = 1500
Administración:	promedio = 32500	desvio = 1000

Los scores z resultan:

$$\text{Contador} = \frac{35000 - 36000}{1500} = -0.67$$

$$\text{Administración} = \frac{33000 - 32500}{1000} = 0.50$$

Z-Score

- El score z es útil cuando la distribución de las observaciones es aproximadamente normal. En este caso, utilizando la regla empírica, un score z fuera del intervalo -2 a + 2 ocurrirá en alrededor del 5% de los casos.
- Un observación en particular puede ser ubicada más precisamente informando el porcentaje de los datos que caen por debajo o en dicha observación.
- Si por ejemplo, 95% de los puntajes de un exámen son menores o iguales a 650, mientras solo el 5% están por encima de 650, entonces 650 se denomina percentil del 95%.

Percentiles

Para cualquier número r entre 0 y 100, el percentil del $r\%$ es un valor tal que el $r\%$ de las observaciones son menores o iguales que dicho valor.

