17. Tipos de distribuciones

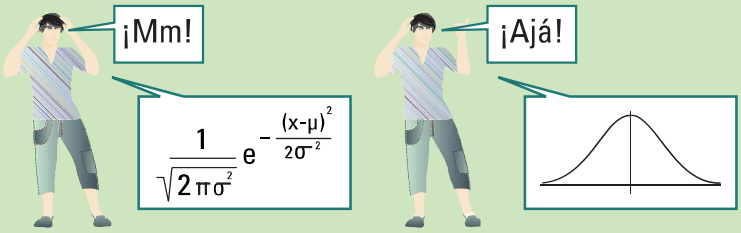
Las distribuciones Normales pueden ser las raras.

17.1. Distribución Normal

Los histogramas y los diagramas tallo-hoja permiten visualizar cómo se distribuyen los valores de una variable numérica. **Muchas veces estos gráficos tienen la forma de una campana,** con una zona central en la cual los valores de la variable son más frecuentes. A medida que nos alejamos de esa zona central las frecuencias disminuyen simétricamente.

Figura 17.1. *Conjunto de datos con distribución en forma de campana, denominada Distribución Normal*

Esta forma de campana también es llamada **campana de Gauss.** Fue descubierta por Abraham de Moivre en 1720. En 1809, Carl Friedrich Gauss, la utilizó para describir los errores de observación cometidos por los astrónomos, al tomar medidas en forma repetida. Fue denominada **curva de error.**



Su fórmula es:

$$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

En esta expresión matemática, aparecen e (un número aproximadamente igual a 2,718), π (el ya famoso 3,1415), μ (el centro de la campana) y σ (que permite variar su ancho).

Johann Carl Friedrich Gauss (1777–1855), físico y matemático alemán. Fue un niño prodigio.

Cuentan que en la escuela, para que los alumnos se queden tranquilos por un rato, el maestro les dio la tarea de sumar los números del 1 al 100. Inmediatamente Gauss respondió 5.050. Se había dado cuenta que la suma de los extremos, y a medida que avanzaba, siempre daba 101:

1+ 100 = 101
2+ 99 = 101
3+ 98 = 101

....hasta llegar a la mitad, 50. Sumar todos es 50 veces 101, o sea 50 x 101 = 5.050

La campana de Gauss se obtiene graficando los pares (x, f (x)) en el plano:

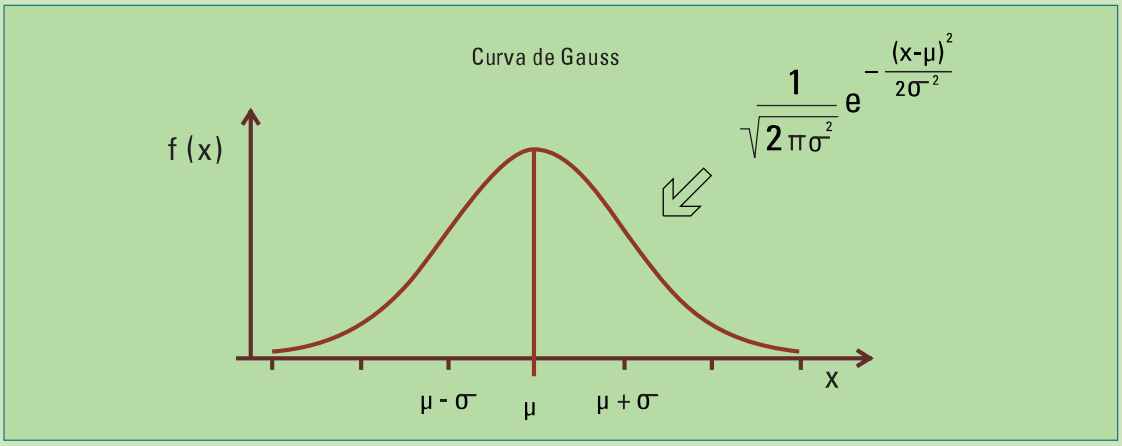


Figura 17.2. *Curva de Gauss junto con su expresión matemática.*

Sólo utilizaremos la forma de la curva y no su expresión.

En 1836 el astrónomo, meteorólogo, estadístico y sociólogo belga Adolphe **Quetelet** extendió la aplicación de la curva, y la utilizó para describir las variaciones de ciertas variables antropomórficas (medidas del cuerpo humano: peso, altura, etc.) entre individuos.

A partir de Quetelet, Francis **Galton** (primo de Charles Darwin y pionero en estudios de genética y de los mecanismos de la herencia) se enteró de la existencia de la curva y **se enamoró de ella.** Dicen que exclamó: "¡Si los griegos la hubieran conocido la habrían deificado!". Galton la llamó curva Normal por primera vez en 1889.

Cuando los datos se distribuyen en forma de campana decimos que tienen distribución Normal o Gaussiana. En la práctica, **los datos rara vez serán "perfectamente Normales"** pero muchas veces la **campana de Gauss** es una muy buena **aproximación al histograma** de un conjunto de datos.

17.1.1. Curva Normal estándar.

Si $= 0$ y $=1$ la curva Normal es:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

Consideremos un conjunto de datos cuyo histograma puede aproximarse por la curva Normal con parámetros μ y σ . Si a cada dato se le resta μ y se lo divide por σ, entonces el histograma del nuevo conjunto de datos podrá aproximarse por la curva Normal Estándar (figura 17.3).

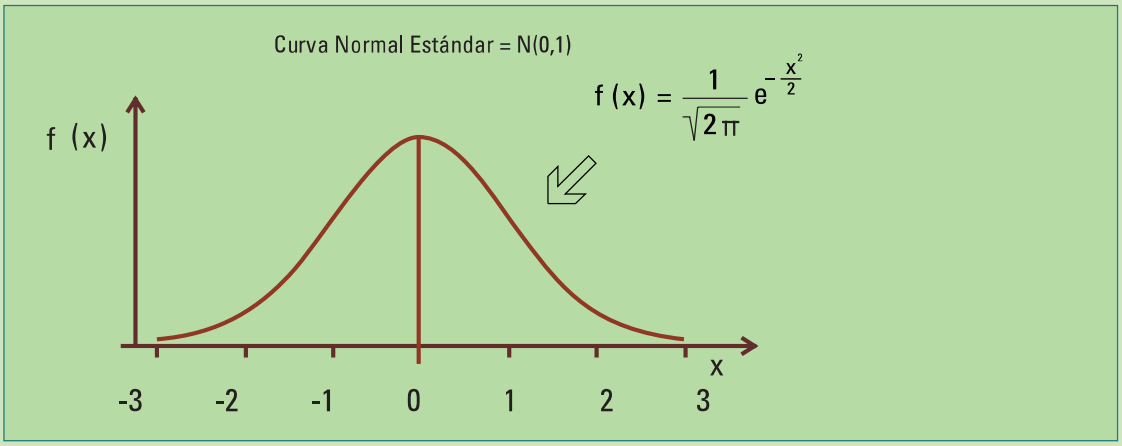


Figura 17.3. *Curva Normal Estándar junto con su expresión matemática.*

17.1.2. ¿Cuándo se obtienen datos con variabilidad Normal y cuando no?

Insistimos, un conjunto de datos rara vez podrá tener una distribución que se ajuste perfectamente a la curva Normal. Sin embargo, en muchas situaciones, esta curva provee una excelente aproximación a los histogramas de los datos. A continuación, presentamos algunos ejemplos en los cuales la aproximación puede ser buena y algunos de cuando no puede serlo.

Ejemplo 17.1. Variabilidad Normal entre unidades muestrales.

Piezas metálicas producidas con la misma máquina, por el mismo operador y en el mismo turno, podrán parecer iguales, pero al medir su **dureza** con cuidado se encontrarán **diferencias.** Cuando estas variaciones se producen en **condiciones normales** (ahora con ene minúscula) – con esto queremos decir: la máquina está funcionando como habitualmente, la materia prima es la de siempre, las herramientas están como todos los días, los operarios descansados y con el ánimo de siempre - entonces las piezas serán parecidas. Las variaciones, respecto de alguna variable (dureza, longitud, peso, elasticidad), darán muchos valores en el centro y pocos en los extremos. A esto lo denominamos "variabilidad Normal" (aquí con ene mayúscula). En cambio, si el producto se fabrica con materias primas defectuosas o los operarios estaban distraídos y siguieron operando cuando la herramienta estaba dañada, la distribución de las variables examinemos ya no será Normal.

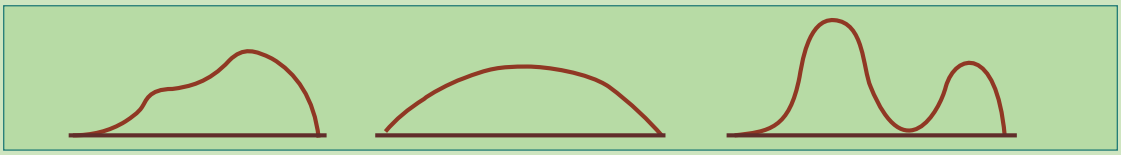


Figura 17.4. *Variabilidad no normal.*

Una aclaración: "normal" con ene minúscula es sinónimo de "habitual"; "Normal" con ene mayúscula se refiere a una distribución de datos en forma de campana

Ejemplo 17.2. Variabilidad debido al error de medición.

Si evaluamos varias veces la **dureza de un mismo** producto, los valores no serán idénticos, aunque la dureza sea siempre la misma. Habrá muchos valores cercanos al **valor verdadero** de la dureza y la frecuencia de los valores disminuirá al alejarse. Si estas mediciones son realizadas con mucho cuidado, por el mismo operador, realizando desde el comienzo los mismos pasos hasta obtener el resultado, su histograma tendrá la forma de la curva Normal. En cambio, si hubo algún descuido al realizar las mediciones, cambió el operador o alguna condición del proceso de medición, estas no tendrán un histograma que pueda ser representado mediante la curva de Gauss.

Ejemplo 17.3. Variabilidad biológica normal.

Si registramos las estaturas de las niñas de una división, encontraremos unas pocas son muy bajitas, otras pocas muy altas y la mayoría con alturas intermedias. Los datos, las mediciones, tendrán una distribución aproximadamente Normal. En cambio, si consideramos las alturas de todos los alumnos (varones y mujeres), los datos provienen de muestras no homogéneas y no tendremos una "variabilidad Normal". Pero no todas las variables antropomórficas tendrán una buena aproximación por la distribución gaussiana como creía Quetelet, por ejemplo, el peso de las personas de una edad y género determinados no tiene distribución simétrica, tampoco los niveles de los triglicéridos en sangre.

La **Normalidad estadística** no implica la normalidad biológica, social o económica. Muchas veces las distribuciones Normales son las raras.

La distribución de los salarios de una población, el caudal de un río de montaña, la precipitación diaria en cierta ciudad, son ejemplos de distribuciones asimétricas.

17.2. Formas que describen diferentes tipos de distribuciones. Curvas de densidad.

La figura 17.5 muestra un histograma de 400 datos de una variable continua y una curva que describe la forma con la que se distribuyen los datos a lo largo de sus valores. Vemos que las mayores frecuencias se encuentran entre cero y cuatro. Para valores mayores que cuatro se reduce constantemente la frecuencia. Podríamos pensar que la curva se obtiene en dos pasos:

- dibujando el borde superior de cada rectángulo de clase y luego.
- suavizando los escalones.

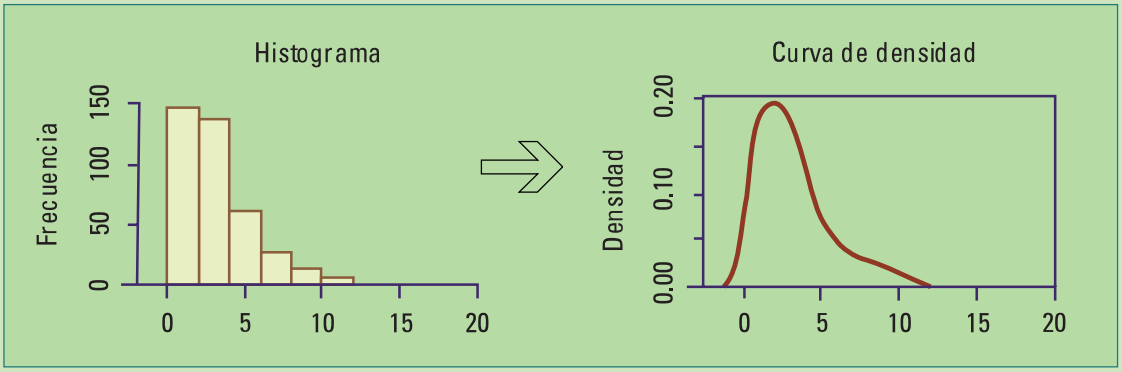
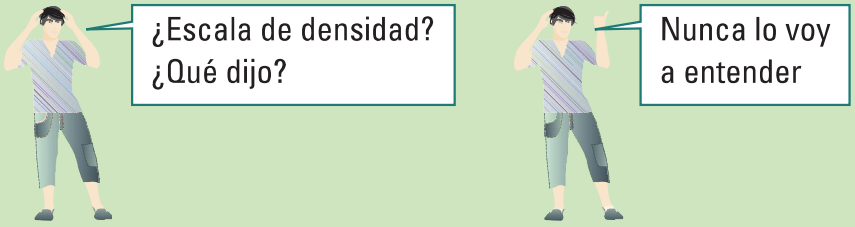


Figura 17.5. *Un histograma y su curva de densidad.*

Existe una diferencia para resaltar entre histogramas y curvas de densidad. Las curvas de densidad se grafican en escala de densidad y los histogramas en escala de frecuencias o frecuencias relativas.

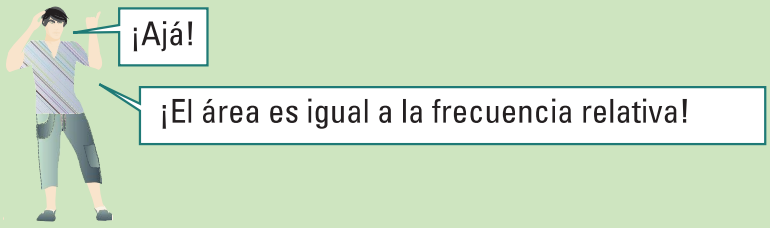


La mayoría de los histogramas muestran la cantidad (frecuencia) o proporción (frecuencia relativa) de observaciones de cada intervalo de clase mediante la altura del rectángulo. De esta manera, el **área de cada rectángulo es proporcional a la frecuencia relativa.** En una **escala de densidad,** el **área de cada rectángulo** es **IGUAL a la frecuencia relativa,** y se obtiene graficando en el eje vertical la frecuencia relativa dividida la longitud del intervalo de clase. En escala de densidad, el área total de los rectángulos del histograma es 1.

Para los datos de la figura 17.5 la longitud de los intervalos es 2 y tenemos:

Escala	Frecuencias		Frecuencias relativas		Densidad	
	Altura	Área	Altura	Área	Altura	Área
	147	294	0,3675	0,735	0,18375	0,3675
	138	276	0,3450	0,690	0,17250	0,3450
	62	124	0,1550	0,100	0,07750	0,1550
	29	58	0,0725	0,145	0,03625	0,0725
	16	32	0,0400	0,080	0,02000	0,0400
	6	12	0,0150	0,030	0,00750	0,0150
	1	2	0,0025	0,005	0,00125	0,0025
	0	0	0,0000	0,000	0,00000	0,0000
	1	2	0,0025	0,005	0,00125	0,0025
Total	400	800	1,0000	2,000	0,5000	1,0000

En **escala de densidad** el **área del rectángulo** de clase es igual a la **frecuencia relativa** y la suma de las áreas es 1.



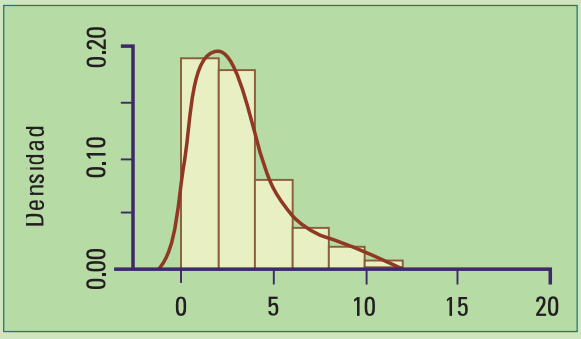
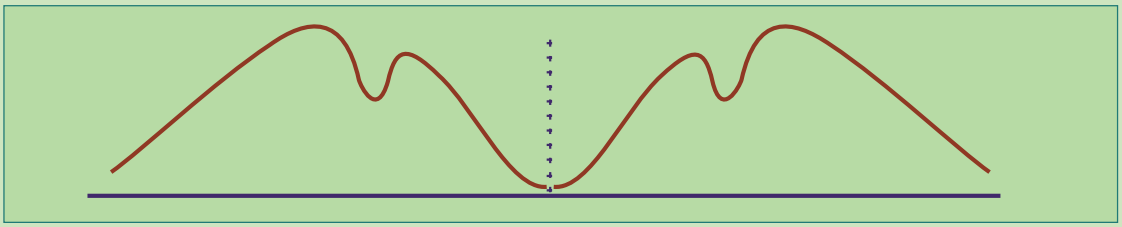


Figura 17.6. *Superposición de un histograma y una curva de densidad.*

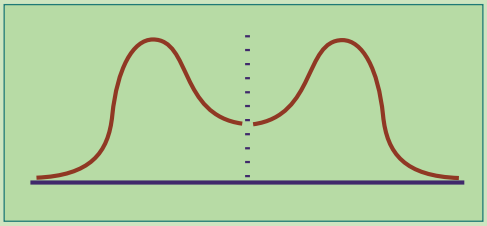
Los histogramas pueden tener distintas formas. Mostraremos algunos patrones especiales en forma simplificada mediante **curvas,** también llamadas **curvas de densidad.** La campana de Gauss es una de ellas.

17.2.1. Simétrica

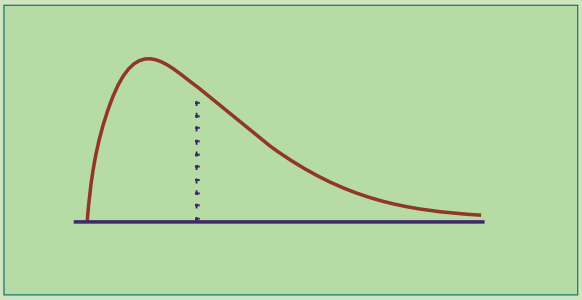
Una **distribución es simétrica** cuando sus dos mitades son imágenes especulares una de la otra.



Por ejemplo, un histograma de las alturas de los mayores de 18 años de un pueblo tendrá dos zonas más altas en espejo, una para los varones y otro para las mujeres, mientras haya la misma cantidad de varones y mujeres. Esto se debe a la superposición de dos curvas simétricas con distinto centro e igual ancho.



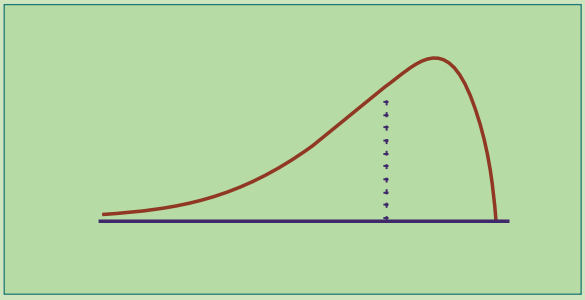
17.2.2. Asimétrica a derecha



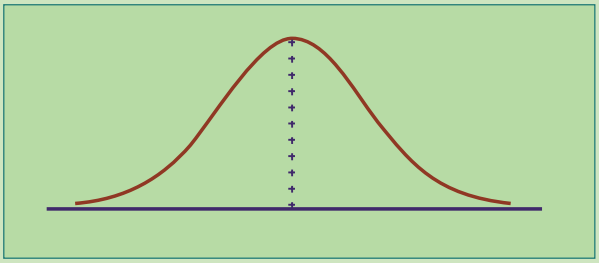
Una **distribución es asimétrica a derecha** cuando la mitad derecha es más finita y más larga. Por ejemplo, la distancia de los domicilios de los alumnos a la escuela mostrará muchos valores pequeños, en la mitad izquierda del histograma, son las de los alumnos que viven cerca y habrá pocos valores grandes de los alumnos que viven lejos.

17.2.3. Asimétrica a izquierda

Una **distribución es asimétrica a izquierda** cuando la mitad izquierda es más finita y más larga. En un **examen fácil,** la mayoría de las notas serán altas y estarán amontonadas del lado derecho, con unas pocas notas bajas (las del lado izquierdo).

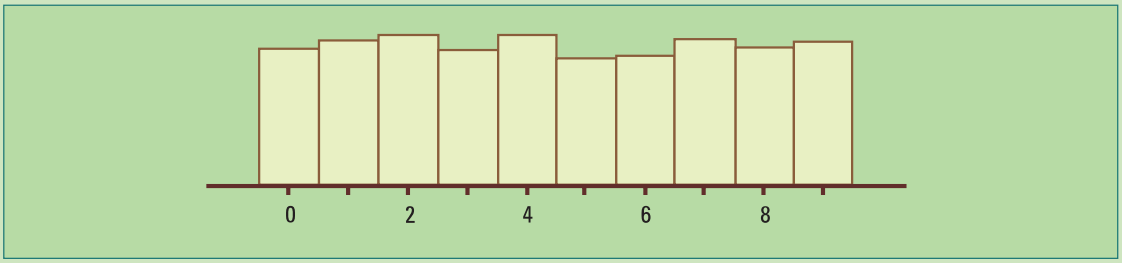


□ 17.2.4. Con forma de campana

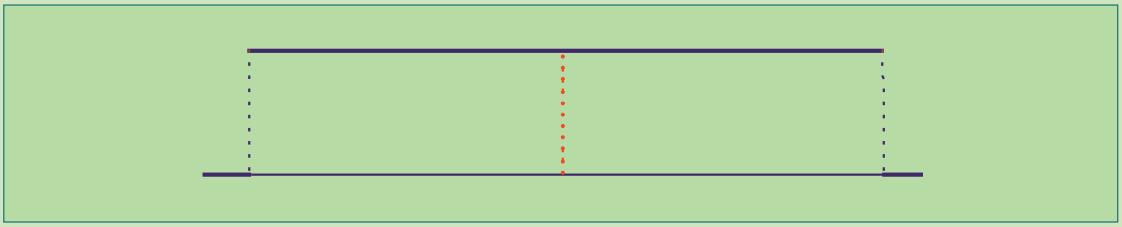


Una **distribución con forma de campana** es **simétrica** con un montículo en el centro y dos caídas como toboganes hacia los costados. Es una de las distribuciones de datos que tal vez aparezca con más frecuencia y es la más estudiada.

□ 17.2.5. Uniforme



Las frecuencias de la última cifra de los resultados de una lotería muestran una distribución pareja sobre todos los dígitos de 0 a 9. Si el mecanismo que genera los números de la lotería funciona correctamente, ninguno de los dígitos tiene más chances de aparecer. Este tipo de distribuciones se llama uniforme y se representa mediante una recta:



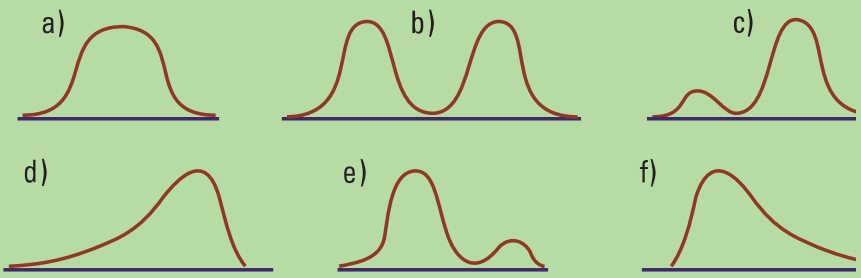
17.3. Actividades y ejercicios

1. Se enumeran las edades de los miembros de 15 familias, casados por 3 años como máximo. En cada fila de cada casilla tenemos las edades de los miembros de una familia.

20, 19	23, 23	23, 25, 2	26, 30, 1, 2	28,31, 2
25, 18, 2, 3	18, 26	38, 34	32, 32	17,19
30, 35, 1	34, 29	21, 19,1	24,26	21,27

Construya un diagrama tallo-hojas y un histograma de las edades. Describa las formas que tienen.

2. ¿Cuál de las siguientes figuras puede representar histogramas de las edades de todos los miembros de familias constituidas a lo sumo hace 2 años?



3. ¿Puede el año de emisión de las monedas decirnos algo más? Para hacer entre todos

- Cada alumno obtiene 10 monedas de 10 centavos y las agrupa en pilas de acuerdo con su año de emisión.
- Cuenta cuantas monedas tiene para cada año.
- Cada alumno/a indica cuantas monedas tiene por cada año y completa la tabla.
- Se obtiene un histograma de la distribución de las fechas.
- ¿Qué forma tiene?
- ¿Qué forma, le parece, debería tener si se perdiera una proporción constante de monedas cada año y a su vez se emitiera una misma cantidad de monedas cada año?
- ¿Puede hallar alguna explicación a la forma del histograma correspondiente a las monedas verdaderas?

Año	Frecuencia
1992	
1993	
.....	
2008	
.....	
	Completar

4. Para estudiar las longitudes de las palabras, seleccione un artículo de una revista de deportes y otro de una de divulgación científica. Para cada uno de los artículos obtenga:

   - la distribución de frecuencias
   - la distribución de frecuencias relativas
   - el histograma

de la variable "cantidad de letras" que tiene cada palabra. Compare las distribuciones obtenidas.

Observación: Diferentes idiomas tienen diferentes distribuciones de las longitudes de las palabras.