20. Curvas de densidad

Poderosa herramienta para describir la distribución de los datos.

Hemos desarrollado un conjunto de herramientas para describir la distribución de los datos: tablas de frecuencias, histogramas, diagramas tallo-hoja, cálculo de medidas resumen (media, mediana, desvío estándar, distancia intercuartil, percentiles), gráfico de caja y brazos. Algunas veces estas herramientas tienen inconvenientes:

- un diagrama tallo-hoja no es práctico para conjuntos con muchos datos.
- las tablas de frecuencias, así como sus representaciones gráficas (los histogramas), eliminan los detalles y dependen de la longitud de los intervalos de clase.
- las medidas resumen (media, mediana, desvío estándar, distancia intercuartil, percentiles) muestran aspectos parciales de los datos.

¿Es posible describir la distribución de los datos en forma completa mediante una única expresión?

La respuesta es: ¡Depende!

¿De qué depende?

Si estamos dispuestos a **describir el patrón general de los datos,** omitiendo los atípicos, **la respuesta es sí.**

Esa respuesta la provee la expresión de una curva - **un modelo matemático, curva de densidad** - para la distribución de los datos.

En la sección 17.2 presentamos algunos patrones especiales que pueden presentar los histogramas mediante curvas. Las expresiones de dichas curvas son precisamente los modelos que necesitamos. Se trata de **descripciones matemáticas idealizadas;** constituyen poderosas herramientas para describir la distribución de los datos. Son especialmente útiles cuando de trata de describir una cantidad muy grande de observaciones.

Podemos establecer un paralelo con la física del movimiento de los cuerpos. La ecuación de la recta describe un movimiento rectilíneo uniforme; pero, ningún desplazamiento real será perfectamente rectilíneo y uniforme. Si graficamos distancia en función del tiempo, con valores medidos de un desplazamiento real, los puntos no caerán exactamente sobre una recta, pero la recta es una buena descripción del movimiento cuando la velocidad es pareja y el desplazamiento es en una única dirección.

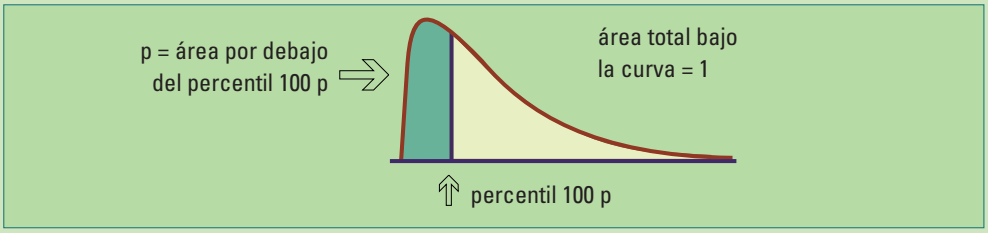
De la misma manera, como la recta es una de las muchas curvas requeridas para describir el desplazamiento de un objeto en función del tiempo, la Curva de Gauss o curva Normal es uno de los tipos de curvas que pueden utilizarse para describir los diferentes tipos de variabilidad de los datos.

□ 20.1. Medias resumen en curvas de densidad

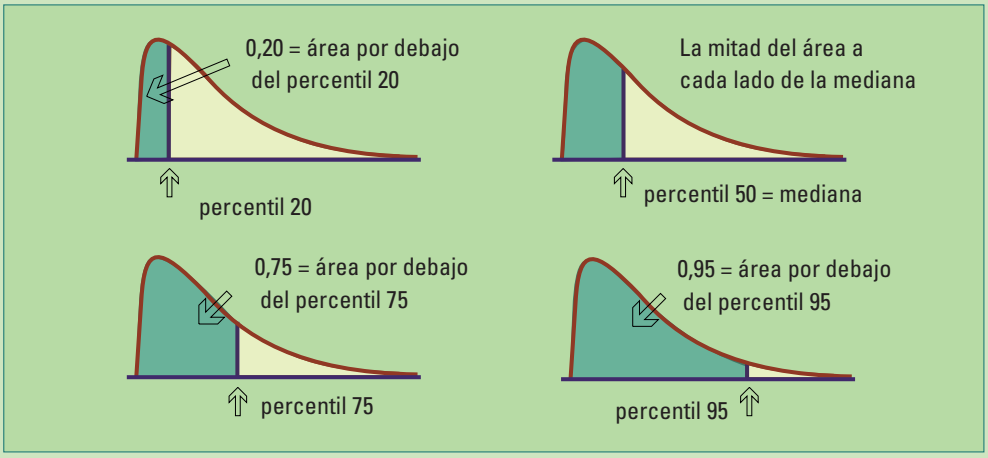
Las curvas de densidad son histogramas idealizados. Las medidas de posición y dispersión se aplican tanto a curvas de densidad como a conjuntos de datos.

Consideremos los percentiles en primer término.

Sabemos que una proporción p de observaciones está por debajo del percentil 100 p.



El percentil 100 x p de una curva de densidad es el punto sobre el eje horizontal para el cual queda a su izquierda el 100 x p % del área bajo la curva, o una proporción p.



En una curva de **densidad simétrica** es fácil **ver "a ojo" donde se encuentra la mediana,** el punto que divide al área en dos partes iguales.

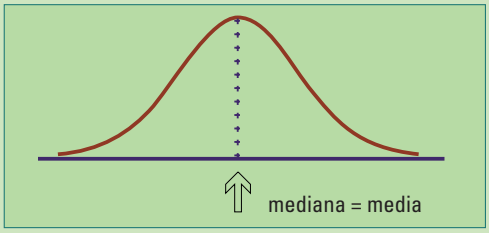


Figura 20.3. *Media y mediana en una curva de densidad simétrica.*

Para una **curva simétrica, la media, el punto de equilibrio coincide con la mediana** que divide el área en dos partes iguales (figura 20.3.).

Como parte de la idealización inherente a un modelo matemático, las curvas de densidad simétricas son "perfectamente simétricas" aunque los datos reales rara vez presenten una simetría perfecta.

Para cualquier curva general **no es fácil hallar a ojo** la mediana, la media y los percentiles. Pero es posible utilizar integrales para obtenerlos. Las integrales son herramientas de análisis matemático que permiten obtener el área por debajo de una curva cuando se conoce la expresión de la misma. No lo haremos aquí.

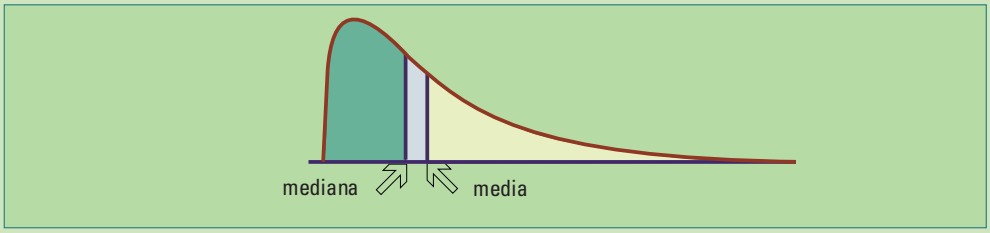


Figura 20.4. *Media y mediana en una curva de densidad asimétrica a derecha.*

La media es el punto de equilibrio de una vara sin peso sobre la que se colocan en cada punto correspondiente al valor de cada dato, pesos idénticos (sección 18.1.1.). La **vara no queda en equilibro** si se apoya en cualquier otro punto. Esta interpretación se extiende a curvas de densidad.

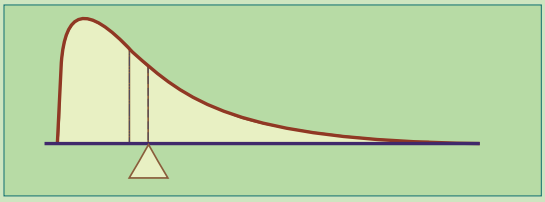


Figura 20.5. *La media es el punto donde la curva de densidad quedaría en equilibrio.*

En una curva asimétrica la media (el punto de equilibrio) es arrastrado hacia la cola larga de la distribución más que la mediana (figuras 20.4 y 20.5). Hallar a ojo la media en una curva asimétrica es más difícil que la mediana, pero la podemos obtener mediante integrales (no lo haremos aquí).

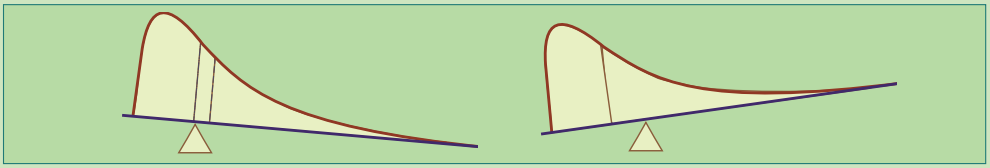


Figura 20.6. *La curva no queda en equilibrio cuando se apoya en un punto diferente de la media.*

La parte inferior de la figura 20.6 ilustra que **la curva no queda en equilibrio cuando se apoya en la mediana.** El área bajo la curva del lado derecho de la mediana "pesa más". Decimos que la distribución tiene **cola pesada a derecha.**

Media y mediana de una curva de densidad: La mediana es el punto que divide el área bajo la curva en dos partes iguales. La media es el punto de equilibrio o centro de gravedad, sobre el cual quedaría en equilibrio si se construyera con un material sólido.

Para calcular la mediana a ojo tratamos de dividir el área en dos partes iguales. Para hallar los **cuartiles,** tratamos de dividir el área por debajo de la curva de densidad en 4 partes iguales (figura 20.7).

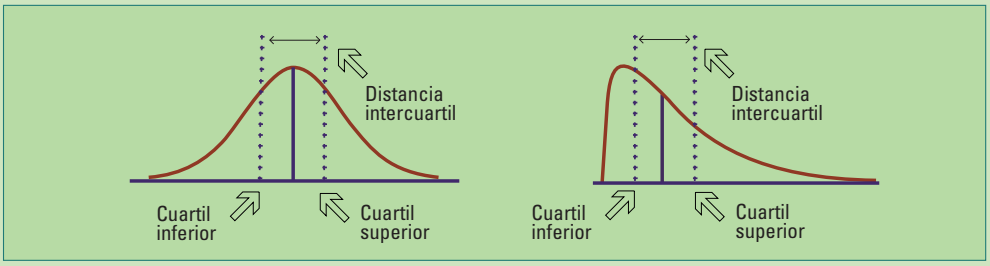


Figura 20.7. *Los cuartiles, la mediana y la distancia intercuartil en una curva simétrica y en una curva asimétrica a derecha.*

La distancia intercuartil es la diferencia entre el cuartil superior y el inferior (también llamados tercer y primer cuartil).

Los cuartiles, por lo tanto, la mediana y la distancia intercuartil, pueden calcularse en forma aproximada a ojo para cualquier curva de densidad. Esto no ocurre con el desvío estándar (18.2.3), que no es una medida natural para la mayoría de las distribuciones. Cuando es necesario, el desvío estándar correspondiente a una curva de densidad, también (como dijimos para los percentiles), puede calcularse utilizando integrales. No se desarrollará esta forma de calcular en este libro.

La curva de densidad es una descripción idealizada de la distribución de los datos, por eso distinguimos la **media** y el **desvío estándar** de una **curva de densidad** de los números

y s (media muestral y desvío estándar muestral respectivamente) y se obtienen a partir de un conjunto de datos. La forma habitual de indicar la media de una distribución idealizada es mediante la letra griega "mu": **μ.** El desvío estándar se indica por **σ**, la letra griega "sigma".

□ 20.2. Ventajas de la curva Normal

¿Para qué sirve tener un conjunto de datos cuyo histograma es aproximadamente Normal? ¿Por qué se habrá enamorado Galton de la curva gaussiana? (sección 17.1)

Hemos visto que es muy útil reemplazar un conjunto de datos por unos pocos valores, las medidas resumen, para describir sus características generales.

> Cuando los datos tienen una **distribución Normal** la distribución de los mismos se puede reducir a **dos números:** la media y el desvío.

En general, es deseable tener **patrones** que representen **la forma** de la distribución de **los datos** y que permitan además representar sus características más importantes mediante **una cantidad pequeña de números.**

20.2.1. Histogramas y la curva Normal

Pensemos primero en un conjunto con **muchísimos datos.** Podemos construir histogramas con intervalos de distinta longitud y superponerle una Curva de Gauss. Como los datos son muchísimos podemos achicar la longitud de los intervalos de clase tanto como queramos.

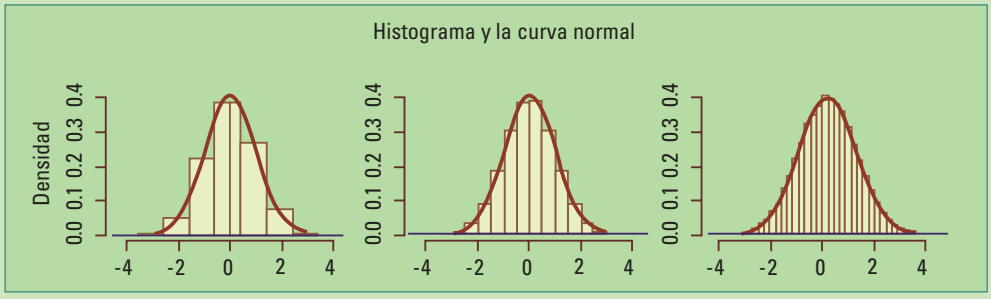


Figura 20.8. *Superposición de la curva de Gauss a histogramas con intervalos de longitud decreciente.*

A medida que se achica la longitud de los intervalos de clase mejora la aproximación de la campana de Gauss.

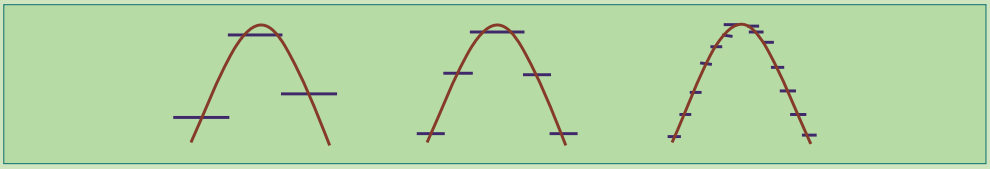


Figura 20.9. *Detalle, en un sector ampliado, de la arista superior de los rectángulos de clase de los histogramas y su aproximación creciente a la curva de Gauss.*

El primero de los histogramas muestra escalones resultantes del agrupamiento de los datos en intervalos de clase, pero estas irregularidades disminuyen al reducir la longitud de los intervalos (figura 20.9). La curva de Gauss en la figura 20.8 describe la distribución de los datos en forma más precisa que los histogramas.

Cuando un histograma se grafica utilizando las frecuencias en el eje vertical, la escala depende de la cantidad de datos. Si se utilizan frecuencias relativas o porcentajes esto es menos arbitrario y el **área del rectángulo es proporcional** a la frecuencia relativa.

Es más natural que el **área del rectángulo** sea **igual a la frecuencia relativa.** Lo logramos si en el **eje vertical** graficamos la **frecuencia relativa dividida la longitud del intervalo.** Esto se llama **escala de densidad** y permite tener la misma escala vertical aunque cambiemos la longitud de los intervalos y el área total de los rectángulos del histograma siempre 1 (ó 100 si las frecuencias relativas están expresadas como porcentajes).

La curva que describe la forma de la distribución se llama **curva de densidad** y tiene área 1. El **área bajo la curva** sobre cualquier **intervalo de valores** del eje horizontal es **la proporción de observaciones** que caen en ese intervalo.

En la figura 20.8 la escala de densidad va de 0 a 0,4 en los 3 histogramas y en la curva. Podemos calcular en forma aproximada el área total pues la figura es aproximadamente un triángulo cuya base tiene longitud aprox. 5 y la altura es aprox. 0,4. El cálculo aproximado resulta:

$$\frac{\text{long de la base x altura}}{2} = \frac{5 \times 0{,}4}{2}$$

$$\frac{\text{long de la base x altura}}{2} = 1$$

20.2.2. Media y desvío de la curva normal

Todas las curvas Normales son simétricas, tienen un único pico y forma de campana.

Sus colas caen rápidamente, por lo tanto no se esperan valores muy alejados (outliers). La media, la mediana y el pico coinciden en el centro de la curva.

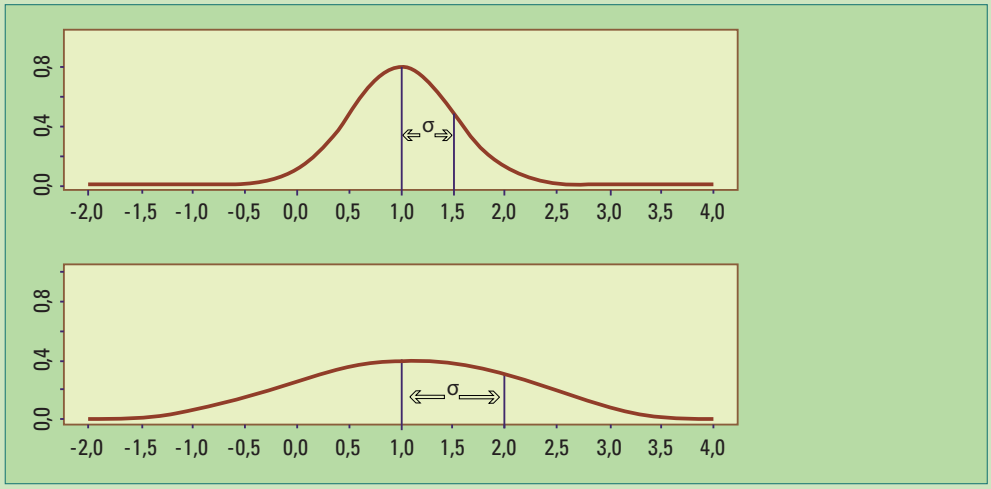
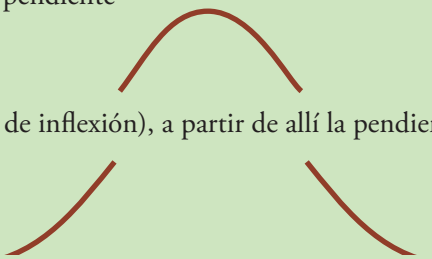


Figura 20.10. *Dos curvas normales con media 1 y distintos desvíos.*

Otra propiedad importante de la curva de densidad Normal es poder localizar el desvío estándar a ojo: a medida que nos movemos en ambas direcciones desde el centro **μ** de la curva, ésta aumenta su pendiente



hasta un punto (punto de inflexión), a partir de allí la pendiente empieza a disminuir

Los dos puntos en los cuales ocurre este cambio de curvatura están localizados a una distancia **σ** a cada lado del centro **μ**.

μ es la media
σ es el desvío

Recuerde, **μ** y **σ** solos no determinan la forma de una distribución en general. Éstas son propiedades de las distribuciones gaussianas.

Pero...

> Cuando los valores de una variable tienen distribución Normal, **sólo dos números** alcanzan para determinar la distribución de todos sus valores. Esos dos números, **μ** y **σ** son **los parámetros** de la distribución Normal.

Pero...

Pequeños alejamientos de la distribución Normal pueden llevar a que **μ** y **σ** no signifiquen nada.

Un detalle extra:

Siempre es más seguro utilizar los percentiles porque tienen el mismo significado en todo tipo de distribuciones. Cuando no hay grupos aislados, las 5 medidas resumen: mínimo, cuartil inferior, mediana, cuartil superior y máximo, son en general una buena representación de los datos.

20.2.3. Otras característcas interesantes

Si un histograma se aproxima por una curva Normal podremos decir algunas cosas más que, simplemente, caracterizar su media y su desvío.

Podremos establecer **criterios** sobre **donde se encuentra** la mayoría de los **datos.**

Los **criterios** que veremos a continuación se utilizan cuando **podemos suponer** que los **datos** tienen una distribución **aproximadamente Normal,** por la naturaleza del experimento, con **μ** y **σ** conocidos. Cuando no son conocidos se estiman mediante la media muestral (x) y el desvío estándar muestral (s), respectivamente, tal como vimos en las secciones 18.1.1 y 18.2.3:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad \text{y} \qquad s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}$$

donde los x_1 son los datos y n es la cantidad total ($x_1, x_2, \ldots x_n$)

Utilizaremos estas reglas en el próximo capítulo sobre control de calidad.

Si una distribución tiene una forma gaussiana, entonces vale la siguiente **regla 68-95-99,7 :**

- Aproximadamente el 68% de los valores se encuentran dentro de 1 desvío estándar (**σ**) de la media (**μ**). Es decir que bastante más de la mitad de los valores están comprendidos dentro del intervalo (μ - **σ**, μ +**σ**) ó μ ± **σ** (figura 20.10).
- Cerca del 95% de los valores están entre la media menos 2 veces el desvío estándar y la media más 2 veces el desvío estándar, o sea dentro de **σ**μ ± 2 (figura 20.11).

- El 99,7% (casi todos) de los valores se encuentran en el intervalo (**μ** - 3**σ**, **μ** + 3**σ**), o sea dentro de μ ± 3 **σ** (figura 20.12).

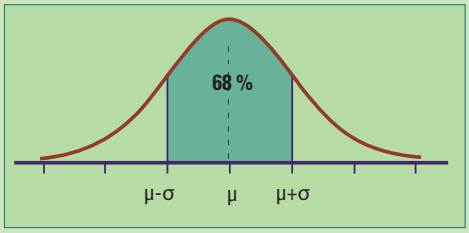


Figura 20.11.

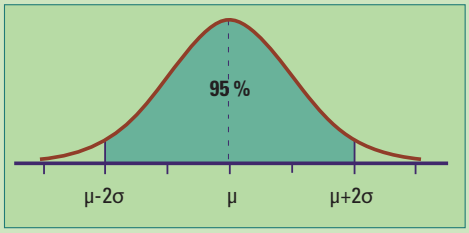


Figura 20.12.

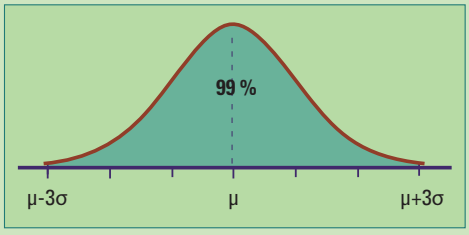


Figura 20.13.

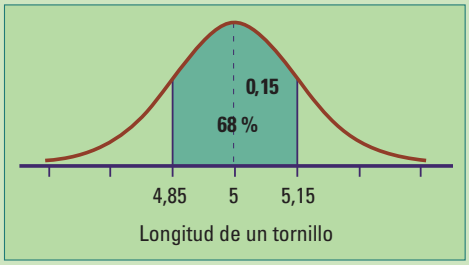


Figura 20.14

Como la mayoría de los valores en una distribución Normal se encuentran en la zona central, alrededor de la media (**μ**), el 68% de los valores están a una distancia no mayor al desvío. Al alejarnos un desvío más de la media, hacia los dos lados, agregamos más valores (un 14% a cada lado); pero, son menos porque se trata de una zona de menor concentración de datos. Obtenemos así el intervalo (**μ**-2**σ**, **μ**+2**σ**) allí se encuentra aproximadamente el 95% de los valores.

Alejándonos otro desvío más, agregamos apenas un 2% de cada lado, llegando a 99.7 % en el intervalo (**μ**-3**σ**, μ+3**σ**).

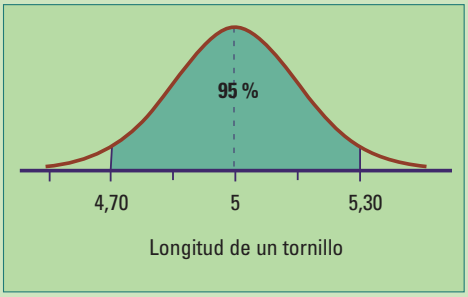
Ejemplo: Un taller metalúrgico produce remaches cuya longitud debe ser de 5 cm con una tolerancia de 0,3 cm (5 ± 0,3 cm). Por lo tanto, las longitudes aceptables están en el intervalo (4,7; 5,3). Interesa evaluar la calidad de la producción teniendo en cuenta este requerimiento.

Como suele ocurrir en esta industria, si la producción se realiza en condiciones normales tendremos muchos remaches cuya longitud esté cerca de 5 cm y pocos alejados; las longitudes tendrán una distribución gaussiana.

Supongamos que los registros históricos de la producción de estos remaches, con el mismo equipamiento, muestran que la media de las longitudes es efectivamente 5 cm con un desvío de 0,15 cm (**μ**=5 y **σ**=0,15).

Luego:

- Aproximadamente el 68% de los valores se encuentran dentro de1 desvío estándar (σ) de la media (μ). Es decir que bastante más de la mitad de los valores están comprendidos dentro del intervalo (μ-σ ,μ+σ) ó μ ± σ (figura 20.11).
- Cerca del 95% de los valores están entre la media menos 2 veces el desvío estándar y la media más 2 veces el desvío estándar, o sea dentro de μ ± 2**σ** (figura 20.12).



- El 99,7% (casi todos) de los valores se encuentran en el intervalo (μ-3**σ** ,μ+3**σ**), o sea dentro de μ ± 3**σ** (figura 20.13).

Casi el 5% de los remaches tendrá una longitud por fuera de los límites especificados. El encargado de control sabrá si este porcentaje de remaches a desechar (scrap) es admisible. Si no lo es deberán modificarse los procesos de producción, hasta que se logre un perfil de calidad adecuado.

20.2.3.1. Regla 68 - 95 - 99,7

Supongamos que un conjunto de datos (x_1, x_2 ... x_n) tiene una distribución gaussiana con media **μ** y desvío estándar **σ**. El conjunto **de datos estandarizados** (z_1, z_2 ... z_n), o "puntajes z", que se obtiene restando **μ** y dividiendo por **σ** ($z_i = \frac{x_i - \mu}{\sigma}$), tendrá una distribución Normal Estándar (figura 20.15).

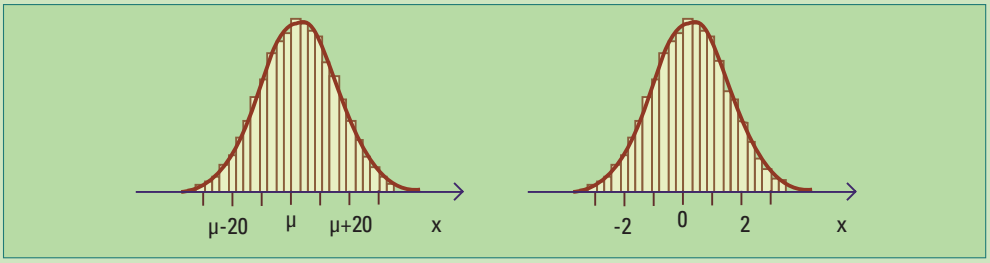


Figura 20.15. *Histogramas de un conjunto de datos en su escala original (x) y transformados en puntaje z.*

Recordemos (sección 17.1.1) que la curva Normal Estándar, también llamada N(0,1), está **dada por** $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ **, y no depende de parámetros desconocidos.**

Las áreas bajo esta curva se pueden calcular.

¡Conocer el área bajo la curva Normal Estándar sobre cualquier intervalo, permite conocerla para todos los intervalos bajo cualquier curva Normal!

En particular, las áreas sobre los intervalos, (**μ-σ**; **μ+σ**); (**μ-2σ**; **μ+2σ**) y (**μ-3σ**; **μ+3σ**) bajo la curva N (**μ**, **σ**) son iguales a las áreas sobre los intervalos (-1,1) ; (-2,2) y (-3;3) bajo la curva N(0,1).

Los valores 68; 95 y 99,7 para los porcentajes de áreas por encima de los intervalos (-1,1); (-2,2) y (-3;3) bajo la curva Normal Estándar son aproximados. Valores más precisos son: 68,27; 95,45 y 99,73 respectivamente.