PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE

FACULTAD DE LETRAS

Métodos y técnicas de investigación cuantitativa

César Antonio Aguilar

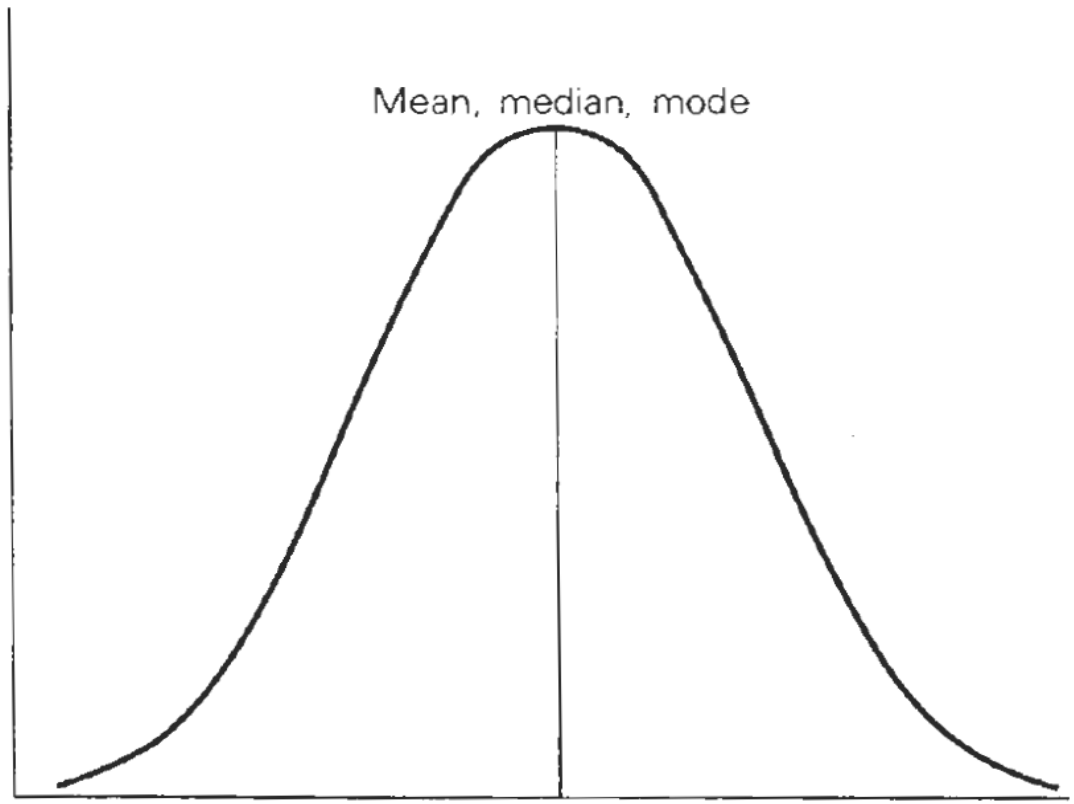
Facultad de Lenguas y Letras

03/06/2013

Cesar.Aguilar72@gmail.com

La distribución normal (1)

En esta clase, vamos a agrupar las tres medidas que hemos estudiado en una **distribución normal**, esto es, una secuencia probabilística de sucesos, en donde cada uno de estos tiene una posibilidad casi simétrica de que ocurra (o no) dentro de una escala dada.



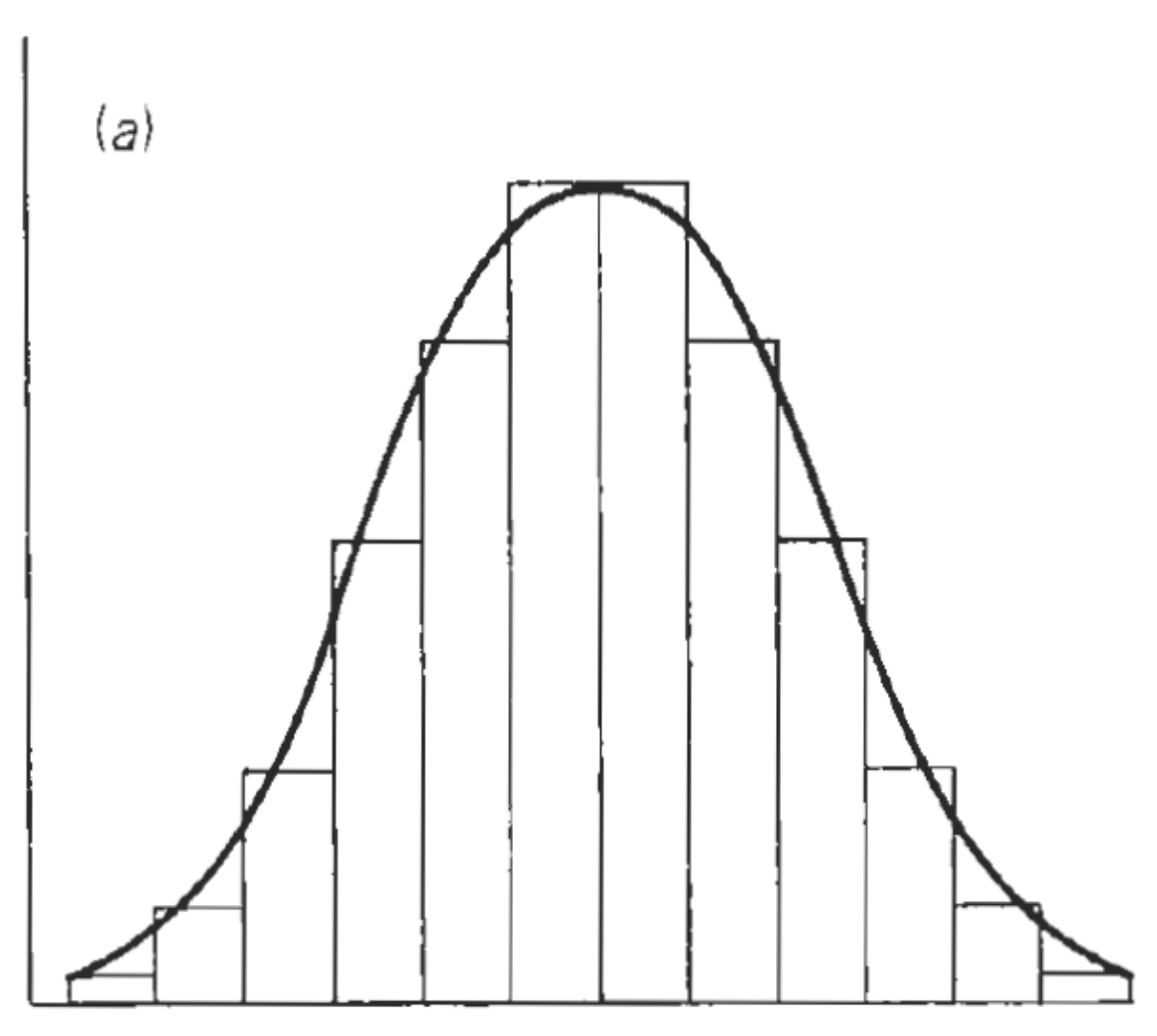
The normal distribution curve

Digamos entonces que, como lo comentamos en la clase pasada, tenemos un punto en donde coinciden la media, la mediana y la moda, y esto nos ayuda a dividir una *campana* de tamaño específico en dos partes exactas, como se ve en la figura.

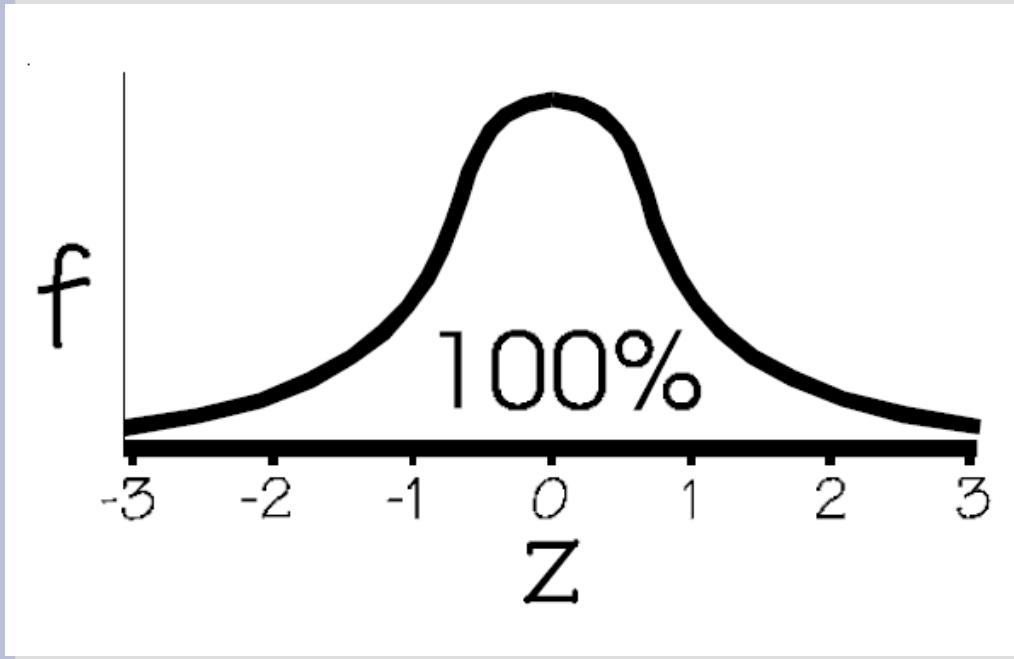
La distribución normal (2)

Cuando vemos una curva como esta, podemos suponer que la distribución de los datos analizados tiende a un patrón *regular*, esto es: la media, la mediana y la moda reflejan cuál es el punto en donde se concentran la mayoría de nuestros datos, así como sus extremos.

Si pensamos en un histograma de frecuencias, tendríamos algo como:



La distribución normal (3)

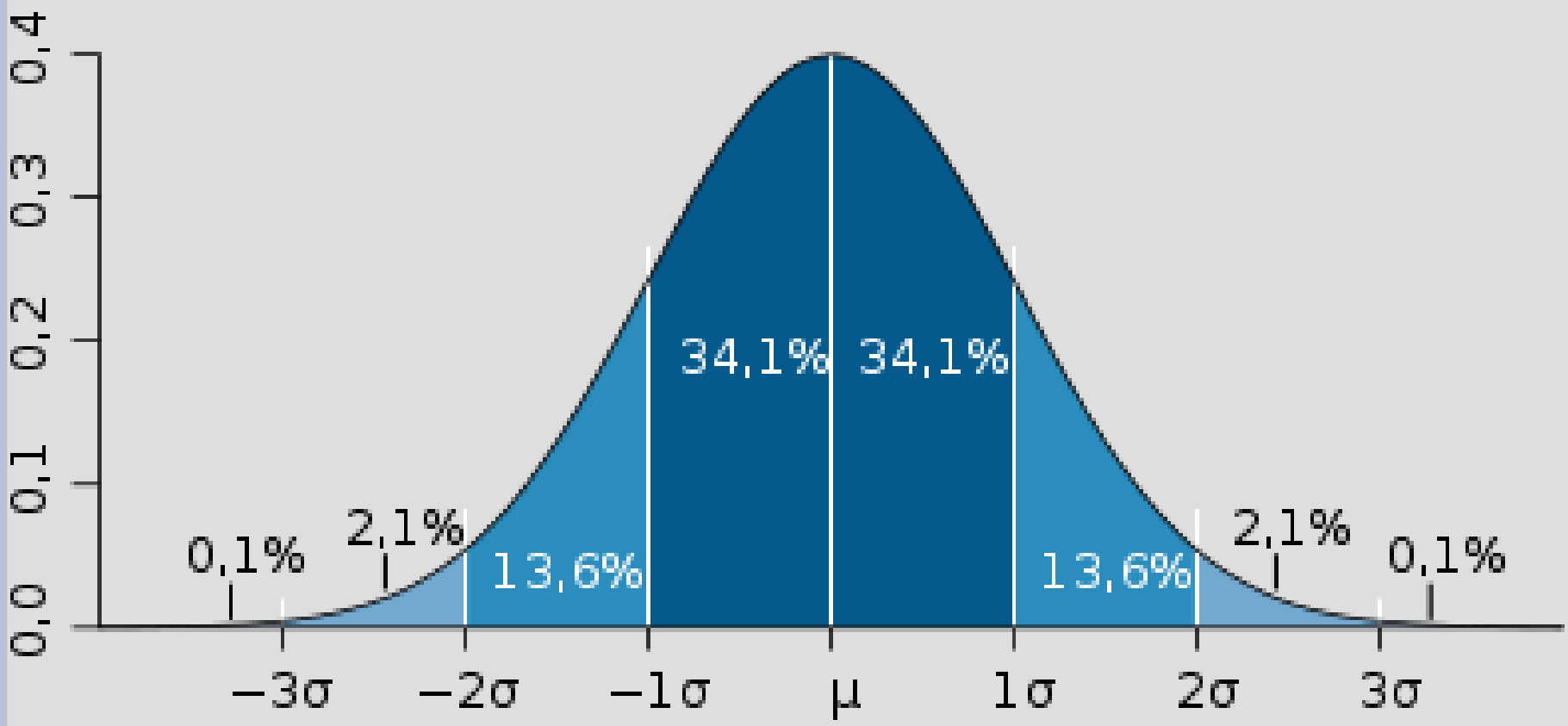


Este gráfico en forma de campana describe una función, la llamada **función de densidad**, la cual caracteriza el comportamiento probable de una población.

El objetivo de esta función es establecer un límite que permita acomodar cada uno de los datos dentro de los espacios que conforman la campana.

La distribución normal (4)

¿Qué distribución pueden cubrir nuestros datos dentro de la campana? Veamos:



La distribución normal (5)

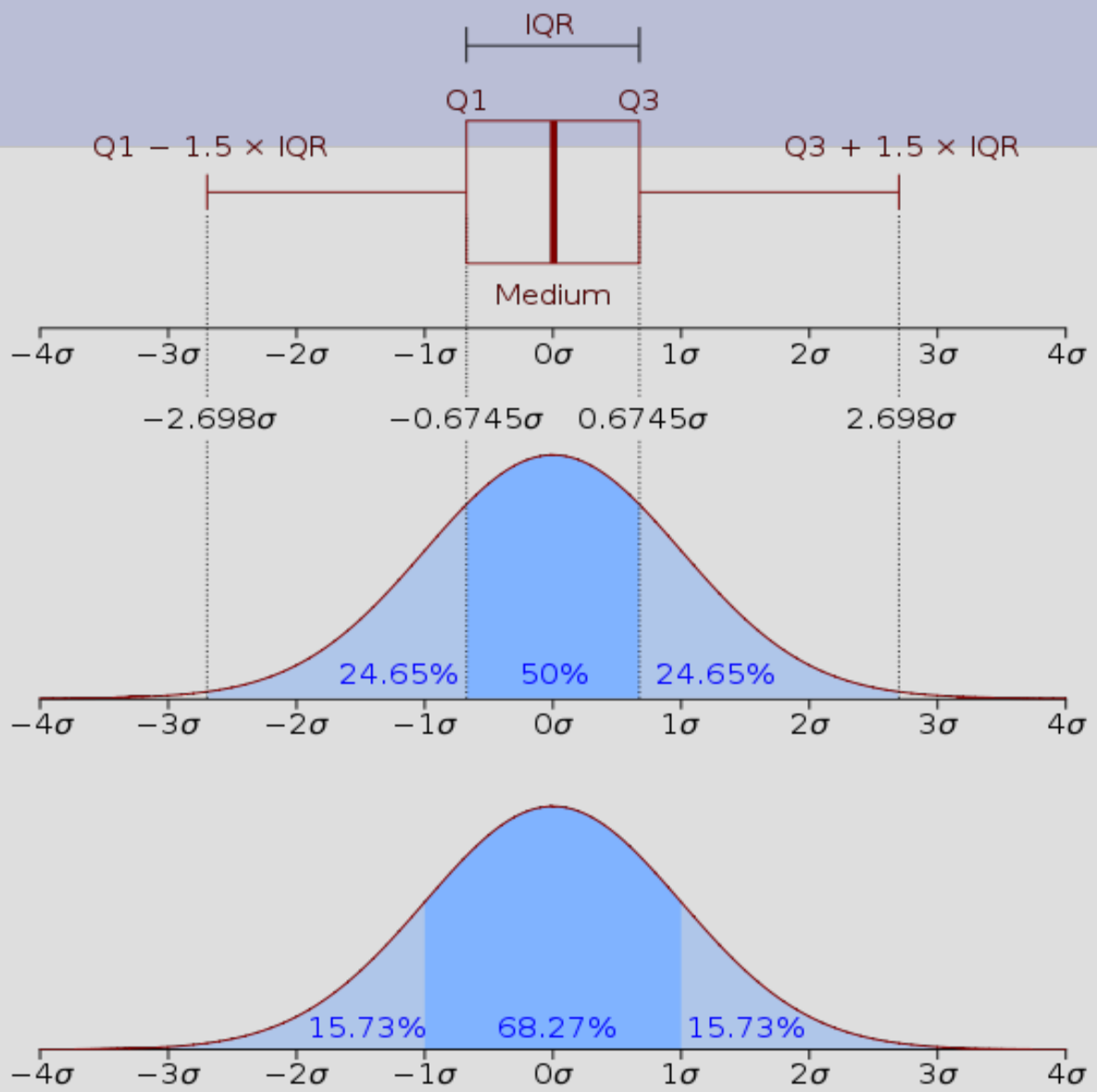
De acuerdo con la gráfica anterior, lo que tenemos es lo siguiente: bajo el supuesto de que nuestros datos siguen una distribución regular (esto es, la mayoría tiende a agruparse en torno a la media (o el promedio), ¿qué tan cercanos o alejados de este punto se muestra el resto de nuestros datos?

Si recordamos las líneas imaginarias que particionan nuestra curva (los **cuantiles**), entonces lo que hacemos es calcular qué tan distantes se concentran los demás datos respecto a la línea central de la campana (ya saben, el segundo cuartil), esperando que tal concentración se vaya acomodando de manera tal que conforma la base de nuestra campana.

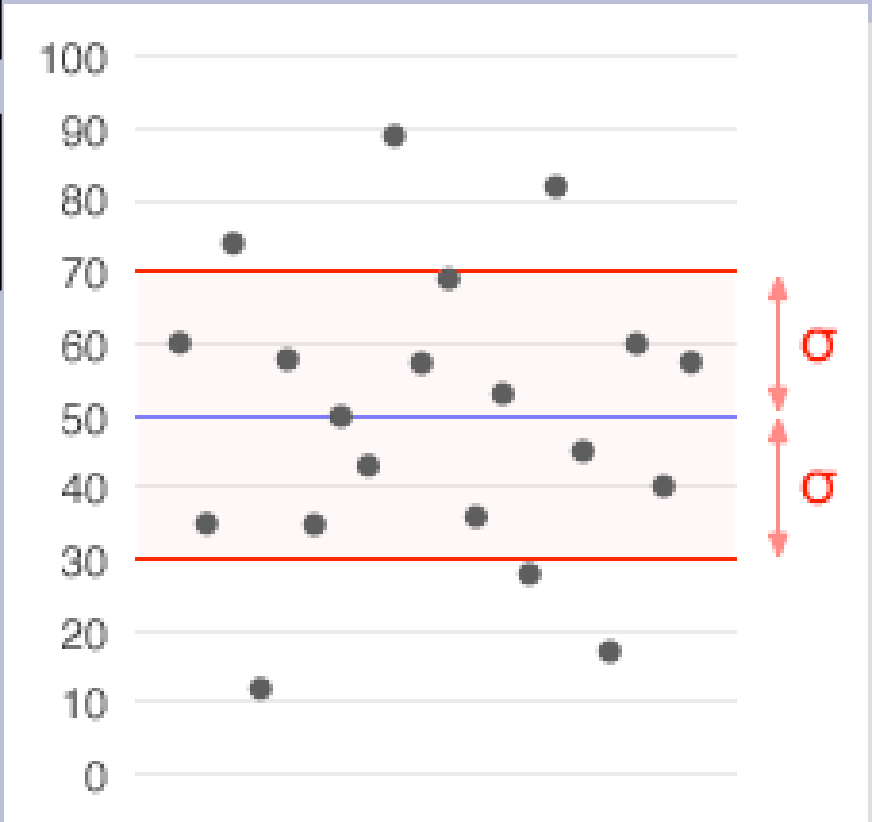
Esto lo podemos ver mejor en la siguiente lámina.

La distribución normal (6)

A este comportamiento regular que muestran los datos respecto a su alejamiento de la media se le conoce como **desviación estándar**, y la vamos midiendo a partir precisamente de la distancia que emerge conforme se aleja (o se acerca) al centro de la campana.



Desviación estándar (1)



La desviación estándar es una medida que muestra la variabilidad de una distribución. Por medio de un valor numérico, indica si las fluctuaciones de una variable están alejadas respecto de la media. Cuanto mayor sea ese valor, mayor será la variabilidad; cuanto menor sea, se observará menos dispersión respecto a la media.

En resumen, la idea es saber si todos los datos que analizamos tienden a mostrar semejanzas, o encontramos un comportamiento dispar entre unos y otros.

Desviación estándar (2)

Para delimitar esta desviación, supongamos lo siguiente: cuando sacamos el promedio de un conjunto de datos, éste lo consideramos como el punto central de nuestra campana, al cual le vamos a dar un valor cero (0).

¿Por qué le damos un valor cero? Porque necesitamos estandarizar nuestros datos para ubicarlos en una escala de desviación, esto es, una escala *Z*. Esto lo deducimos usando la siguiente fórmula:

$$z = \frac{x - \bar{x}}{S}$$

En donde:

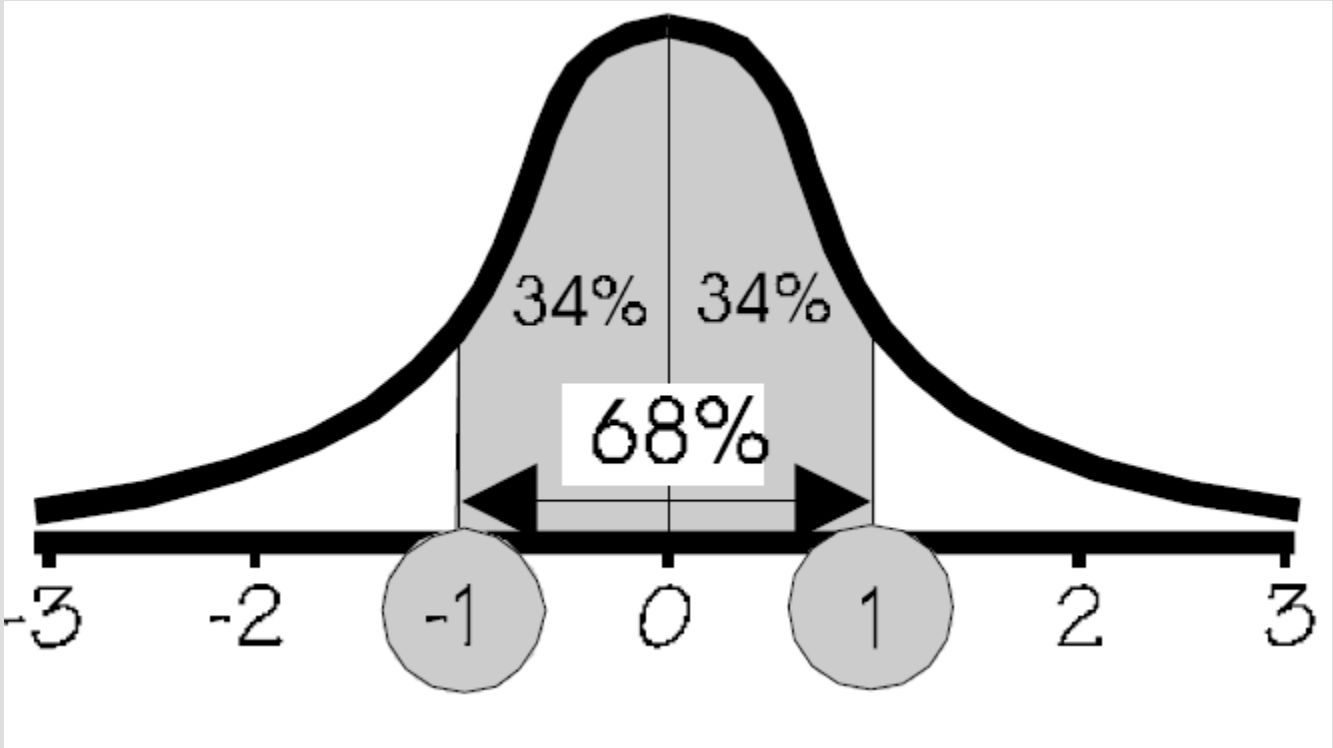
X es un valor de la variable cualquiera.

ẋ Es el valor de la media o promedio.

S (o también *δ*) representa la desviación estándar.

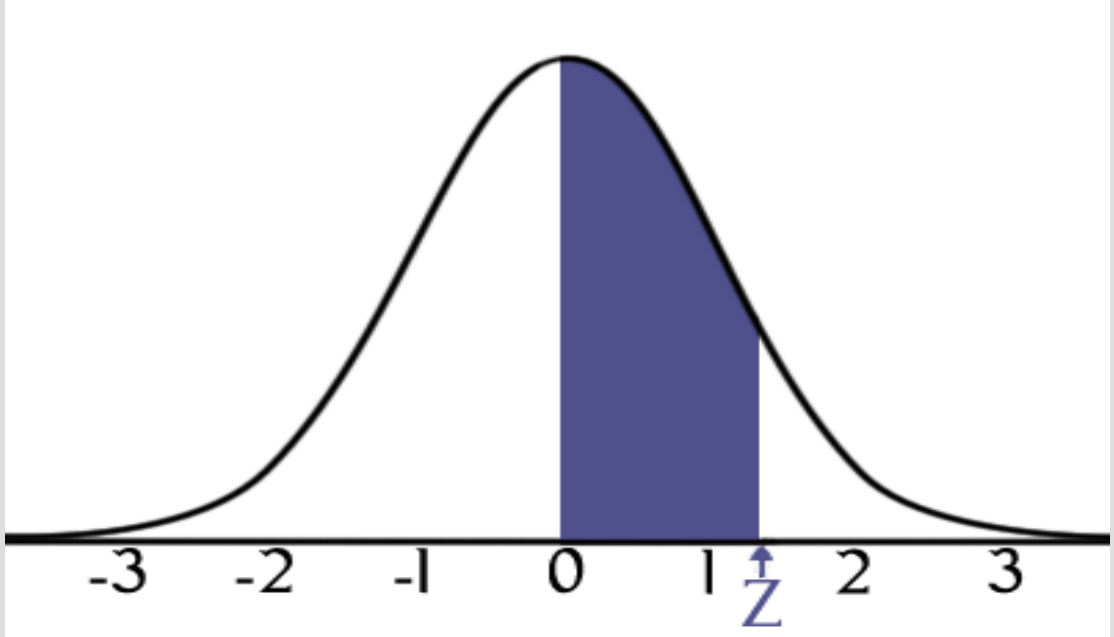
Desviación estándar (3)

Esta fórmula nos ayuda a indicar lo siguiente: si tenemos un conjunto de datos que se mueven en una escala de 0 a 1, y en donde 0 es equivalente al promedio, ¿qué tan cercanos a ese 0 debieran estar los datos para considerar se ubican dentro de sus “márgenes normales”? ¿Qué les parecería un margen de un 34%?:



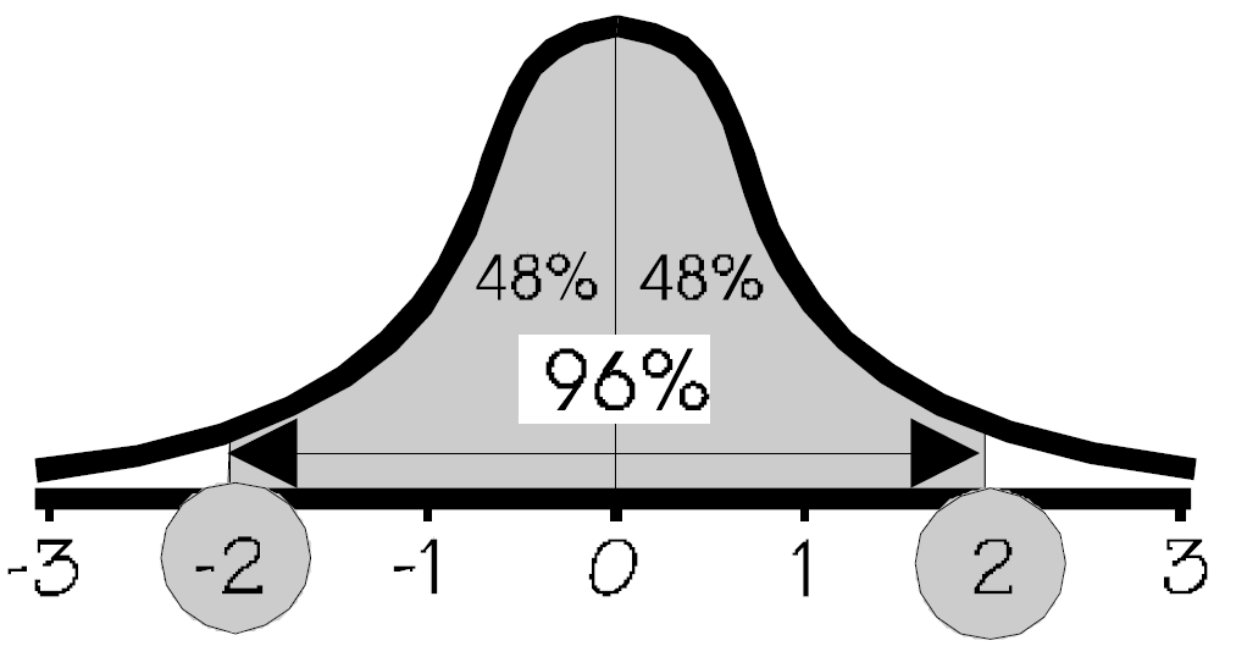
Desviación estándar (4)

Esta fórmula nos ayuda a indicar lo siguiente: si tenemos un conjunto de datos que se mueven en una escala de 0 a 1, y en donde 0 es equivalente al promedio, ¿qué tan cercanos a ese 0 debieran estar los datos para considerar se ubican dentro de sus “márgenes normales”? ¿Qué les parecería un margen simétrico de un 34% a la izquierda y la derecha de 0? La suma de estos promedios nos da un 68%, esto es, un porcentaje posible de 0.68 de cercanía respecto al punto medio:



Desviación estándar (5)

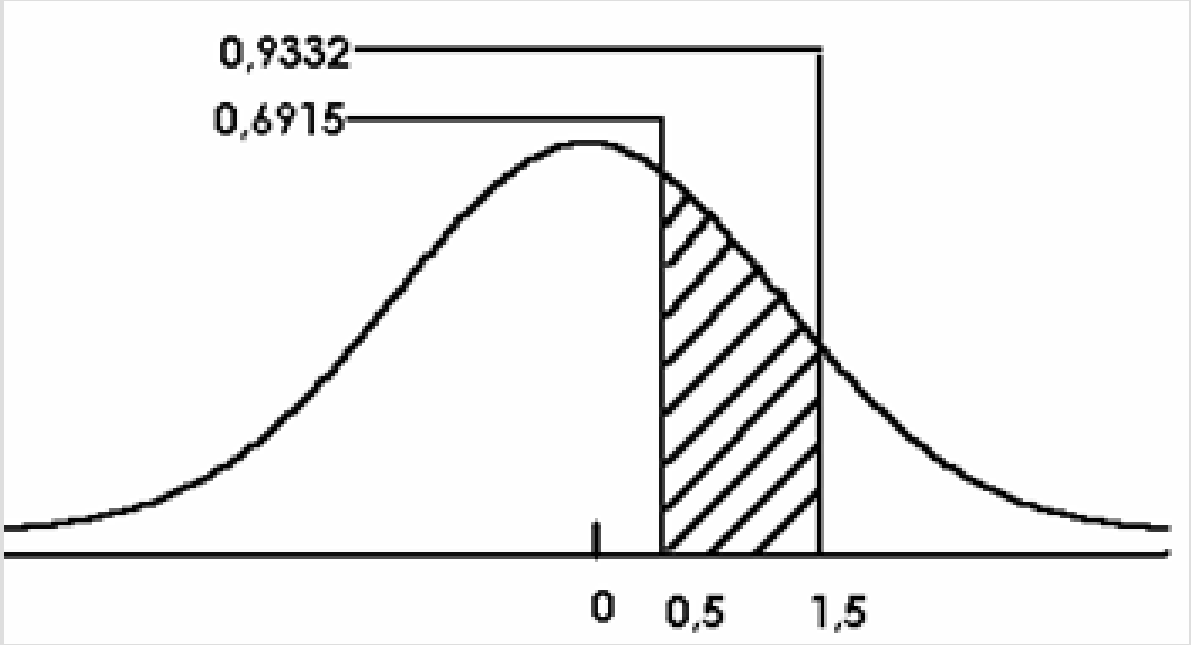
Ahora, supongamos que estos márgenes en torno al 68% no son suficientes para ubicar a todos los sujetos de nuestra población, de tal suerte que empezamos a ver variaciones entre tales sujetos que, dada esta distribución, nos resultan significativas. Lo que hacemos entonces es establecer un segundo puntaje Z, con valores de 2/-2, a la derecha e izquierda del punto medio, esto es:



Desviación estándar (6)

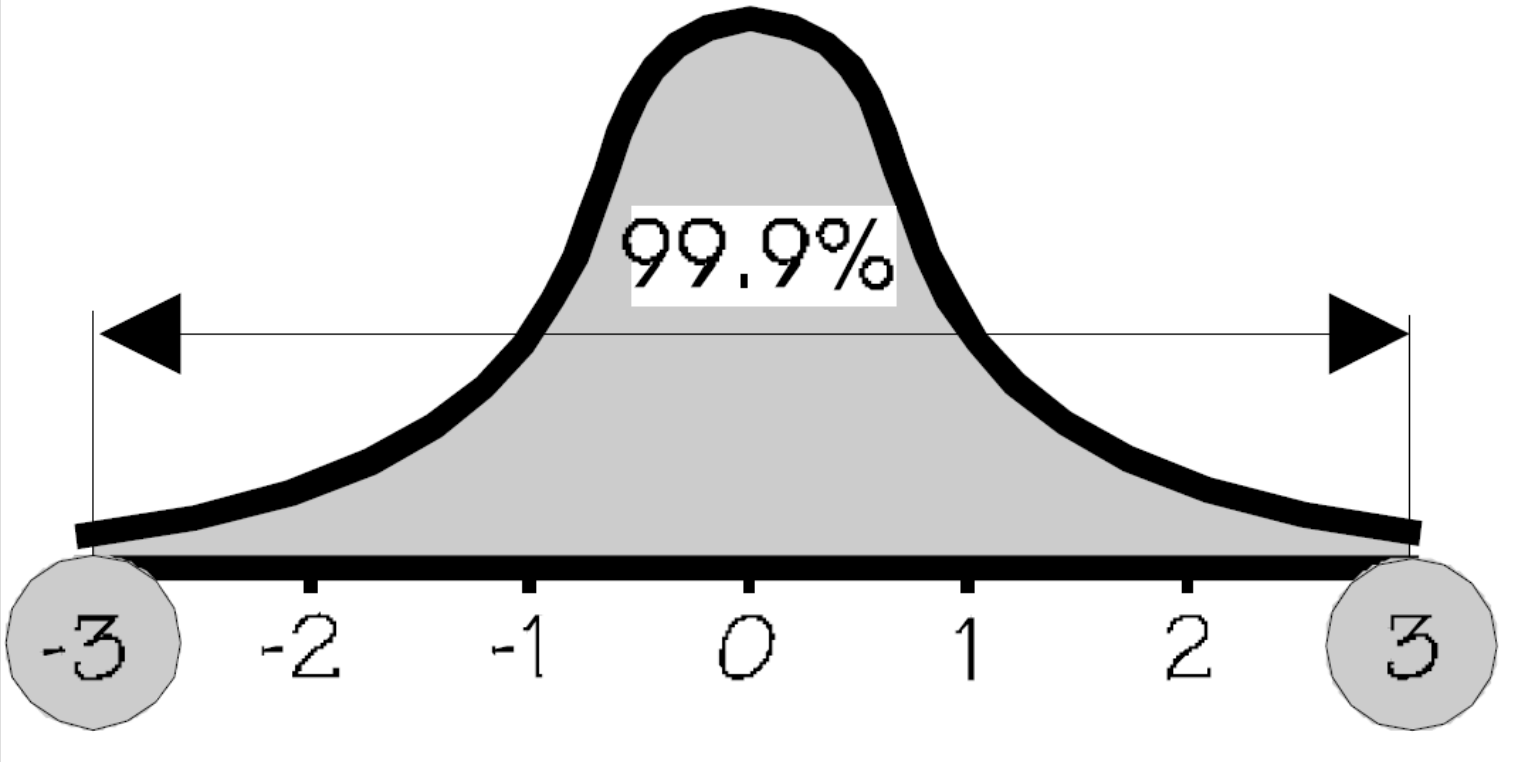
Lo que estamos haciendo entonces es fijar nuevos márgenes, y entonces determinamos cuán alejados se ubican estos sujetos respecto a 0. Considerando que esta área es simétrica a la derecha y a la izquierda, entonces fijaríamos una distancia de un 48% (o de un promedio de 0.48).

Sumando ambos valores, nos da un total de 96% (esto es, 0.96, si lo redondeamos en un número fijo).



Desviación estándar (7)

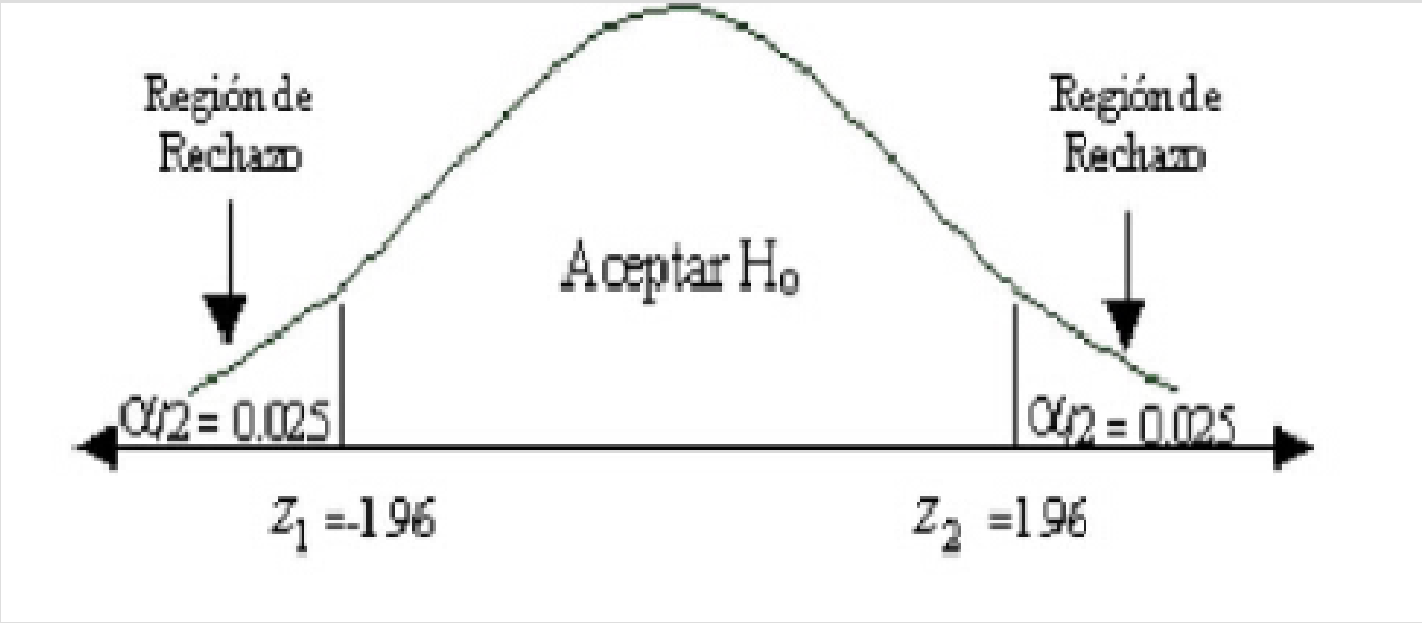
Finalmente, suponiendo que nos quedan todavía sujetos que quedan fuera del alcance de estos dos márgenes, fijamos un par de nuevos puntos Z, los cuales prácticamente se ubican a los extremos de la media. Estos puntos, al igual que los anteriores, son simétricos a la derecha y a la izquierda, de tal suerte que nuestra campana asume la siguiente forma:



Desviación estándar (7)

Los puntos Z sirven para fijar umbrales que nos permitan decir si los sujetos que analizamos manifiestan un comportamiento regular, o lo contrario.

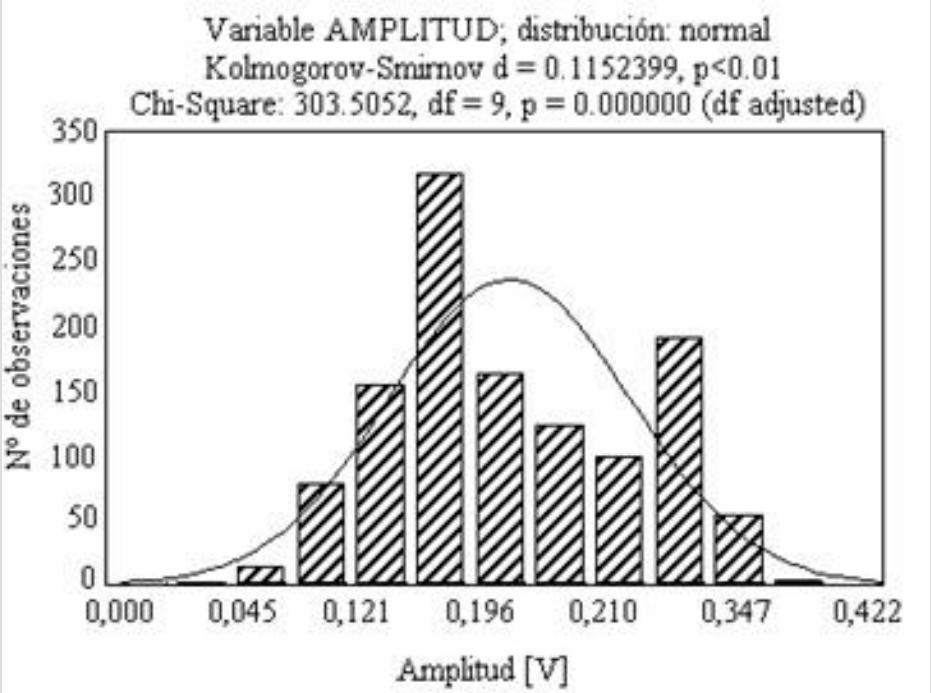
Claro está, siempre habrá excepciones, pero esperamos que éstas se sitúen de forma “organizada” en los extremos. Dependiendo de su grado de lejanía, podemos decidir si las integramos al grupo medio, o de plano las rechazamos (porque son muy pocas y atípicas).



Desviación estándar (8)

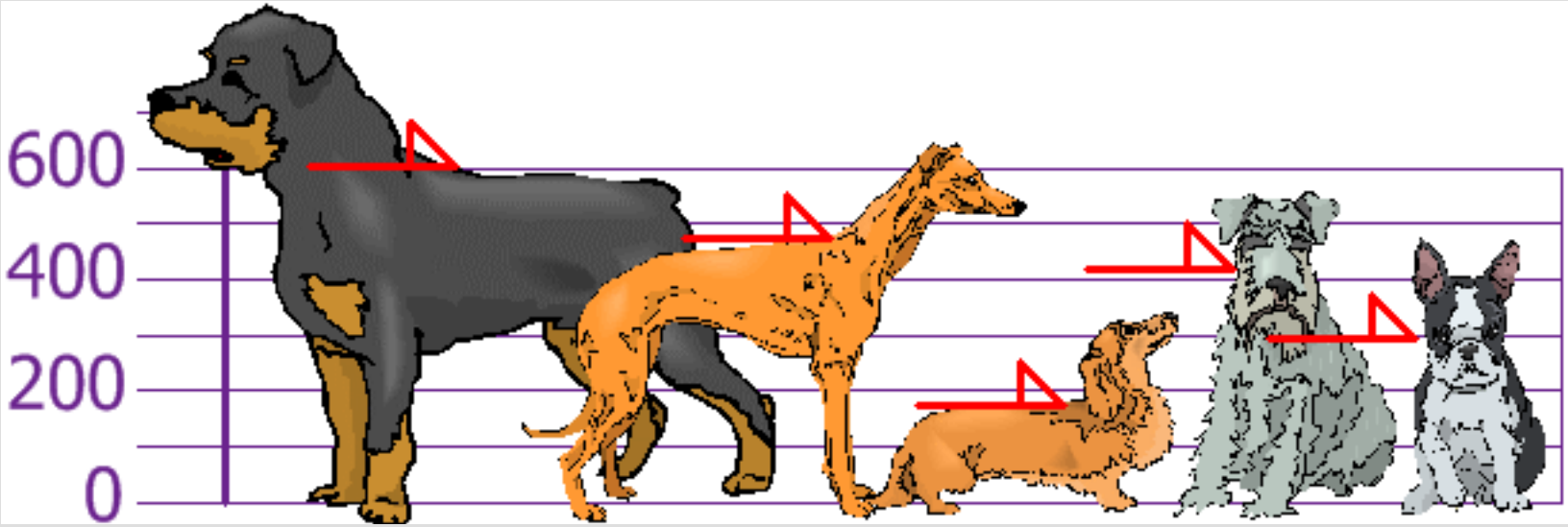
La desviación estándar es una herramienta que nos ayuda a presentar nuestros datos de un modo “resumido” el orden que vamos deduciendo en nuestra población.

Claro está, este orden no se infiere de forma inmediata, por lo que nos conviene ser lo más detallados posibles para observar en qué partes vemos regularidades, y en qué partes observamos dispersión.



Interludio: un ejemplo (1)

Vamos a tratar de hacer nuestro primer cálculo de desviación estándar analizando la estatura de una muestra. Nuestros sujetos de trabajo son:



Nuestra escala de medidas es en milímetros, y hemos decidido que el patrón de medida es de los hombros hacia el suelo (esto puede discutirse, pero para el ejemplo nos sirve muy bien).

Interludio: un ejemplo (2)

De acuerdo con nuestra convención, las medidas que tenemos son:

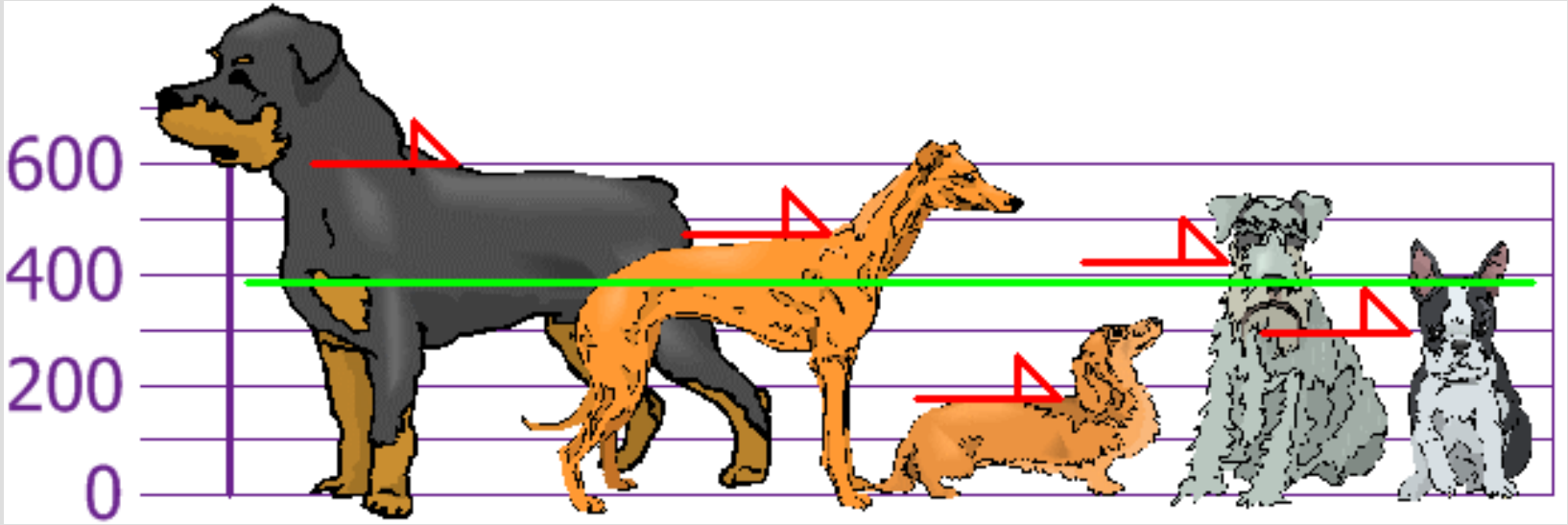
1. Rocky (Rottweiler): 600 mm
2. Alexa (Galgo española): 470 mm
3. Mr. Big (Dachshund): 170 mm
4. Alfie (Terrier escocés): 430 mm
5. “El Rudo” (Pit Bull): 300 mm

Interludio: un ejemplo (3)

¿Cuál es la estatura media de estos sujetos? Saquemos el promedio:

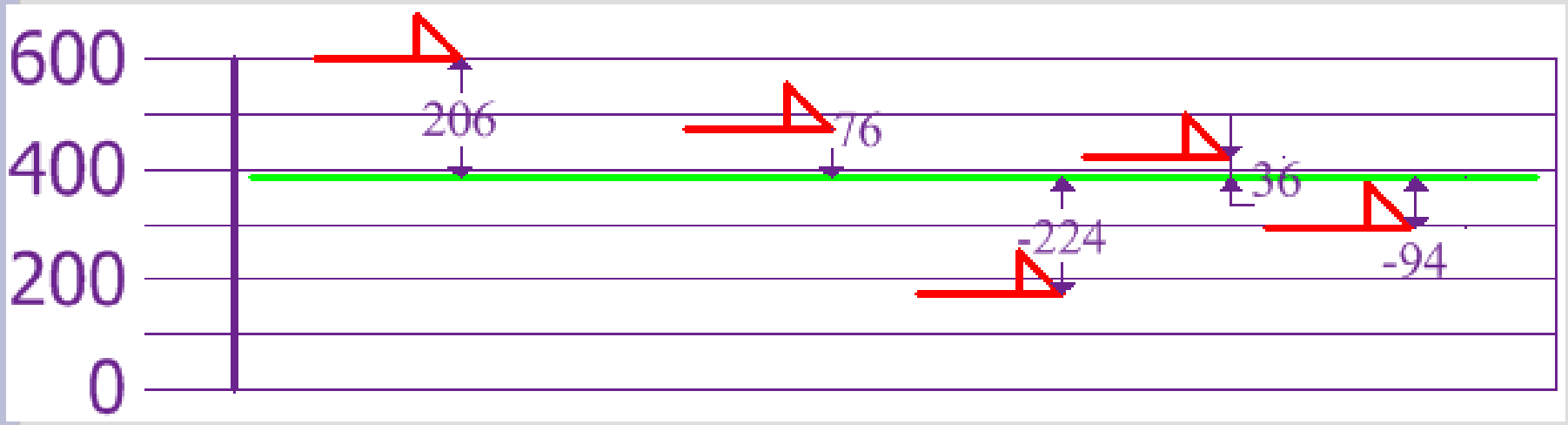
$$\text{Media} = \frac{600 + 470 + 170 + 430 + 300}{5} = \frac{1970}{5} = 394$$

Organicemos nuestros sujetos en torno a la media (línea verde):



Interludio: un ejemplo (4)

Ahora, calculemos qué tan cercanos o alejados están estos 5 amigos de la media. Veamos:



Así, por sentido común, suponemos que “Rocky” es el más alto, y que “Mr. Big” es el pequeño del grupo, ¿pero dónde se quedan los otros tres? ¿Se alejan demasiado de la media, o son precisamente los que ubican más cerca de esta medida?

Interludio: un ejemplo (4)

Para calcular qué tanto se acercan o desvían de la medida promedio, apliquemos esta fórmula:

$$\sigma^2 = \frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \ldots + (x_n - \mu)^2}{n}$$

Lo que quiere decir que para calcular la varianza total de la muestra (esto es, δ), necesitamos sumar cada diferencia, a la que le restamos el promedio, y luego la elevamos al cuadrado, y finalmente la dividimos entre el número de sujetos, esto es:

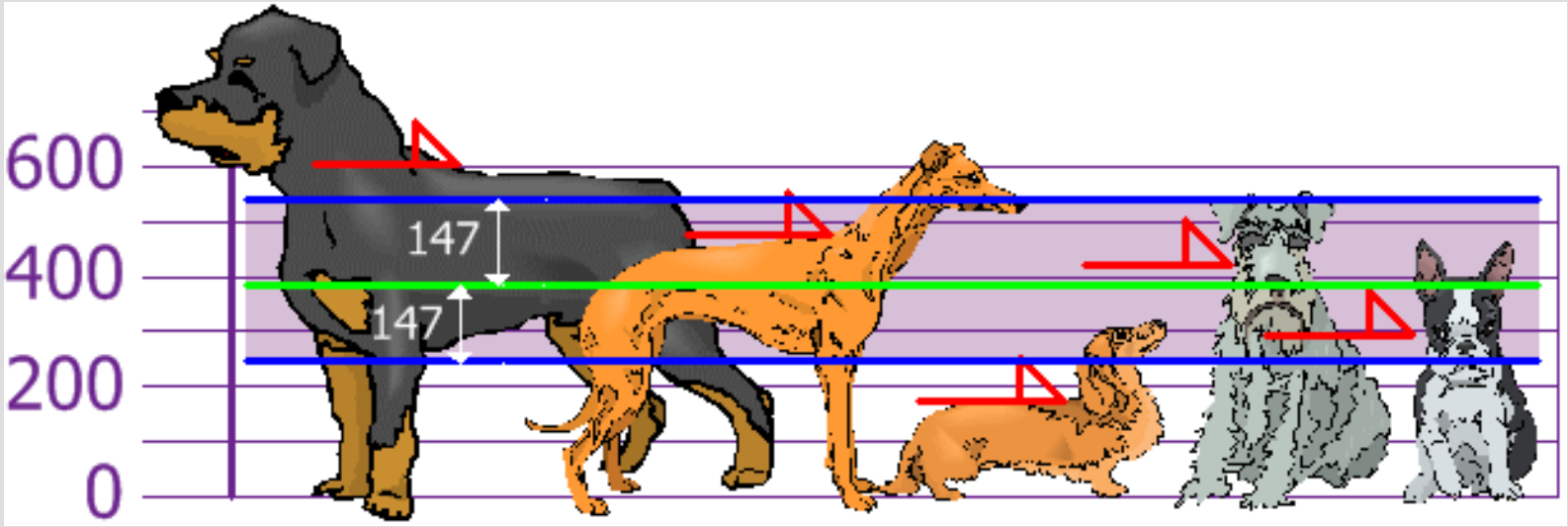
$$\text{Varianza: } \sigma^2 = \frac{206^2 + 76^2 + (-224)^2 + 36^2 + (-94)^2}{5} = \frac{108{,}520}{5} = 21{,}704$$

Interludio: un ejemplo (5)

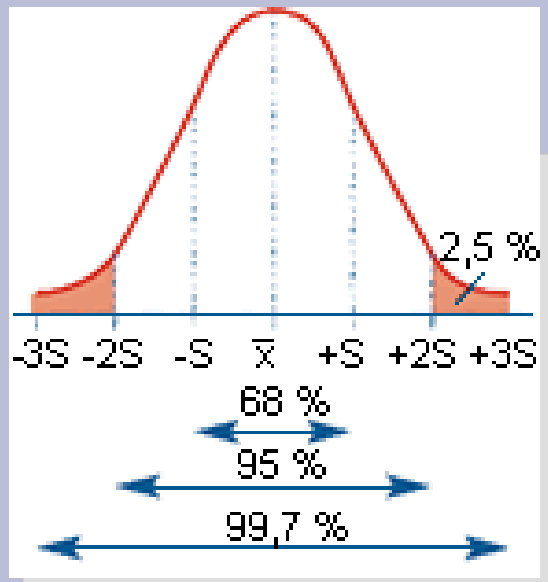
Nuestra varianza es de 21.704 mm. Para deducir la desviación estándar, hay sacar la raíz cuadrada de este número, esto es:

Desviación estándar: $\sigma = \sqrt{21{,}704} = 147$

Tenemos que la desviación estándar respecto a la estatura promedio es de 147 mm, ya sea hacia arriba o hacia abajo, es decir:



Interludio: un ejemplo (6)



Si hiciéramos nuestra curva (o gráfico de campana), y quisiéramos colocar a cada uno de estos sujetos en el sitio que les corresponde, ¿cómo lo haríamos? La solución es:

$$(600 - 394) / 147 = 1.4$$

Decimos entonces que Rocky se sitúa a una desviación estándar de 1.4 respecto a la medida promedio donde una desviación estándar (1s) equivale a 147 mm.

Si lo transformamos a milímetros, lo que tenemos que hacer es multiplicar la desviación estándar (147) por la medida de desviación de Rocky, lo que nos da un total de 205.8 mm extras que tiene nuestro amigo respecto a la población.

Ejercicio

Deduzcan las puntuaciones z del resto de nuestros amigos, aplicando esta fórmula:

$$z = \frac{x - \bar{x}}{s}$$

O lo que es lo mismo:

(tamaño en mm – promedio) /desviación estándar

Gracias por su atención

Blog del curso:
http://cesaraguilar.weebly.com/meacutetodos-y-teacutecnicas-de-investigacioacuten-cuantitativa.html