En Transp4.pdf

se introducen vallas para detectar datos atípicos que permiten construir un gráfico (box-plot) que describe la distribución de los datos utilizando sólo las medidas resumen: mediana, cuarto inferior y cuarto superior.

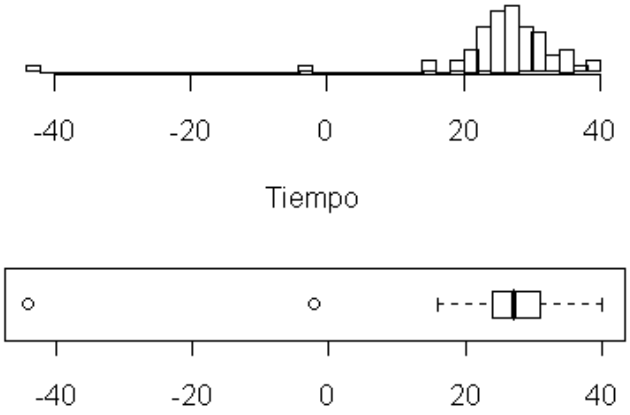
Presentaremos a continuación datos reales, su histograma y boxplot, para luego analizar con detalle la definición de las vallas.

Ejemplo: Mediciones obtenidas por Newcomb entre Julio y Septiembre de 1882.

28	22	36	26	26	28
26	24	32	30	27	24
33	21	36	32	31	25
24	25	28	36	27	32
34	30	25	26	26	25
-44	23	21	30	33	29
27	29	28	22	26	27
16	31	29	36	32	28
40	19	37	23	32	29
-2	24	25	27	24	16
29	20	28	27	39	23

Newcomb midió cuánto tardó la luz en llegar, desde su laboratorio sobre el río Potomac a la base del monumento a Washington y volver, una distancia total de 7400 metros.

Codificación: corremos al punto decimal nueve lugares a la derecha, obteniendo 24828 y luego registramos únicamente **el desvío respecto de 24800**. Luego 28 es la versión corta de 0.000024828 y -2 se corresponde con 0.000024798.



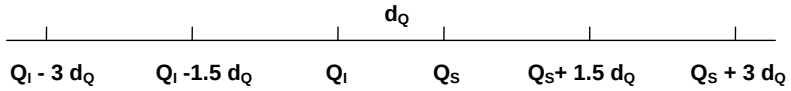
Se han identificado dos valores atípicos que también aparecen en el histograma. Los brazos del boxplot llegan hasta el último dato que se enuentra dentro de las vallas internas.

LAS VALLAS NO SE GRAFICAN

Hemos visto (pág 70 clases teóricas) Identificación de valores atípicos

Para identificar valores atípicos se definen vallas internas y externas de acuerdo con los siguientes criterios:

$$\text{Valla Interna Inferior} = Q_I - 1.5\, d_Q$$
$$\text{Valla Interna Superior} = Q_S + 1.5\, d_Q$$
$$\text{Valla Externa Inferior} = Q_I - 3\, d_Q$$
$$\text{Valla Externa Superior} = Q_S + 3\, d_Q$$



Un dato será considerado atípico si se encuentra a una distancia mayor que 1.5 d_Q del cuarto más cercano a dicho dato.

¿Por qué?

Para responder a esa pregunta es necesario ver

- qué entendemos por datos gaussianos y que
- para datos gaussianos, una proporción muy baja de ellos estará fuera de los puntos de corte dados por las vallas internas y externas.

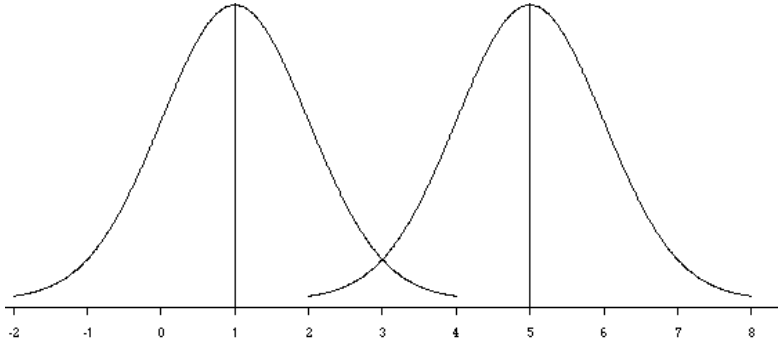
Curvas de Densidad Normales o Gaussianas

La siguiente ecuación describe las curvas Normales (también llamadas Campanas de Gauss)

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

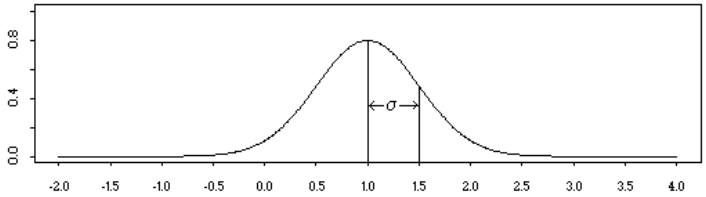
Observación: la curva queda completamente especificada cuando se conocen los valores de μ y σ. Se la indica por $N(\mu,\sigma^2)$

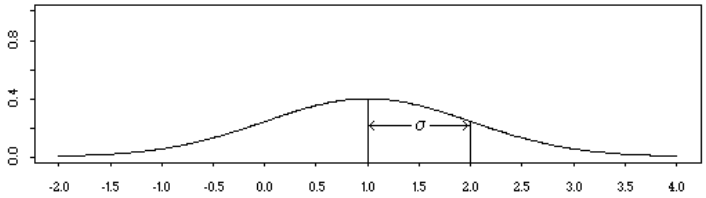
Todas las curvas tienen la misma forma.



Dos curvas de densidad $N(\mu,\sigma^2)$

con diferente μ y el mismo σ (μ= 1 y μ=5 y σ=1)





Dos curvas de densidad $N(\mu,\sigma^2)$

con el mismo μ y diferente σ (μ= 1 y σ=0.5 y σ=1)

Podemos localizar σ a ojo en una curva de densidad Normal. 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) en que 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.

Algunas consideraciones

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

Existen otras distribuciones no gaussianas con forma de campana.

Las *distribuciones Normales* proveen buenos modelos para

- puntajes de pruebas tomadas en poblaciones grandes (pruebas habilidades escolares y muchas pruebas psicológicas),

- mediciones cuidadosamente replicadas y de la misma calidad, sin outliers,
- características de una población biológicamente homogénea (longitudes de las cucarachas, rendimiento de la soja y pérdida de humedad en carne de pollo envasada).

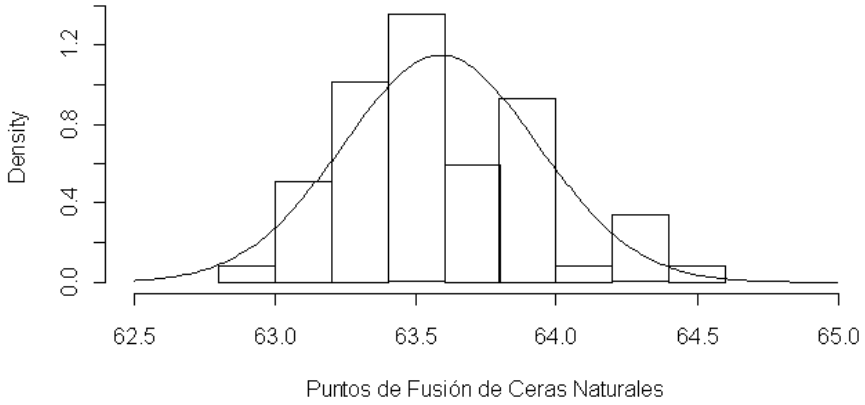
Las distribuciones de las siguientes variables, en cambio, son generalmente *asimétricas*:

- variables económicas (ingreso personal, ventas en firmas comerciales),
- tiempos de supervivencia (de pacientes de cáncer luego de realizado un tratamiento),
- tiempo de vida (de componentes mecánicos o electrónicas).

Es muy riesgoso suponer la normalidad de los datos sin inspeccionarlos aunque la experiencia sugiera que un modelo gaussiano es factible en un caso particular.

- El **desvío estándar no significa nada** si los datos **no son Normales** o aproximadamente Normales
- La **media no describe el centro** si los datos **no son simétricos**
- **La mediana y la distancia intercuartil** pueden fallar si los datos forman grupos
- El significado de las medidas resumen está atado a la forma de la distribución de los datos.

Superposición de una curva normal a un histograma




```
media.muestral <- mean(Cera$CERA)
desvío.muestral <- sd(Cera$CERA)
ejex<-seq(62.5,65,0.01)

hist (Cera$CERA,probability=T,
      main="", xlab ="Puntos de Fusión de Ceras
Naturales", xlim = c(62.5,65))

lines(ejex,dnorm(ejex, media.muestral,
desvío.muestral))
```

A mano

- curva simétrica de altura = $\frac{1}{DS\sqrt{2\pi}}$ y puntos de inflexión en $\bar{x} \pm DS$.
- la escala en el eje vertical del histograma es la frecuencia relativa, siempre que la longitud de la base de los rectángulos de clase sea 1. En cualquier otro caso, en el eje vertical se

grafica (la frecuencia relativa de cada clase) / (longitud de la clase) de manera que el

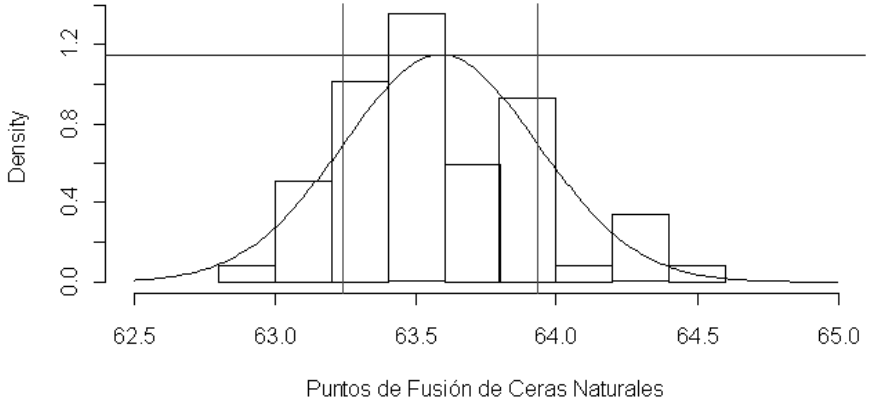
área de un rectángulo = (longitud de la base)*(altura del rectángulo)= frecuencia relativa

Verifiquemos este procedimiento para la superposición que muestra la figura sabiendo que:

```
> media.muestral
[1] 63.58881

> desvío.muestral
[1] 0.3472209

> 1/(desvío.muestral*sqrt(2*pi))
[1] 1.148958
```




```
hist (Cera$CERA,probability=T,
      main="", xlab ="Puntos de Fusión de Ceras
```

```
      Naturales", xlim = c(62.5,65))

lines(ejex,dnorm(ejex,
media.muestral,desvío.muestral))

abline (
  h= 1/(desvío.muestral*sqrt(2*pi)),col="blue")

abline (
v= c(media.muestral- desvío.muestral,
media.muestral+ desvío.muestral ), col="red")
```

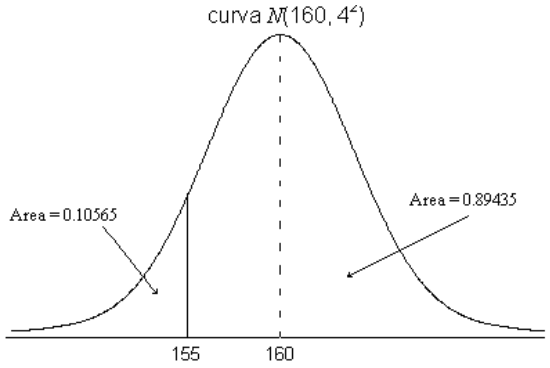
Datos con Distribución Gaussiana

Si se sabe que los datos tienen distribución gaussiana, conocer la media y el desvío, permite calcular la proporción de datos que se encuentran dentro de un intervalo de valores cualquiera.

Ejemplo. Supongamos el histograma de las alturas de las mujeres jóvenes argentinas puede construirse con intervalos de clase muy pequeños y que una curva Gaussiana con μ = 160 cm σ = 4 cm sea una muy buena aproximación de dicho histograma.

¿Qué proporción de mujeres miden menos de 155 cm? Podemos aproximar esta frecuencia relativa por el área bajo la curva de densidad $N(160, 4^2)$ a la izquierda del punto 155.

```
> pnorm(155,160,4)
[1] 0.1056498
```

El área que deja a la izquierda del 155 bajo la curva de densidad $N(160, 4^2)$ es 0.1056498 corresponde a la proporción de mujeres que tiene una altura menor que 155

Recíprocamente podemos preguntar
¿Cuál es la altura tal que el 25% de las mujeres tiene una altura menor? o equivalentemente
¿Cuál es la altura tal que una proporción de 0.25 de las mujeres tiene una altura menor?

```
> qnorm( 0.25,160, 4 )
[1] 157.302
```

157.302 es el **cuarto inferior** de la distribución $N(160, 4^2)$

Propiedades de la Distribución Normal

1) Si un conjunto de datos (*x1*, ..., *xn*) se distribuye de acuerdo con una curva de densidad $N(\mu,\sigma^2)$, es decir que su histograma puede ser aproximado por dicha curva, entonces el conjunto de datos estandarizados *z1*= (x1-μ)/σ, ..., *zn1*= (xn-μ)/σ tienen una

distribución **normal estándar** *N*(0,1).

Recíprocamente

2) Si un conjunto de datos (z*1*, ..., *zn*) se distribuye de acuerdo con una curva de densidad **normal estándar** entonces el conjunto de datos (*x1* = μ + z*1* σ , ..., *x n* = μ + z*1* σ) se distribuye de acuerdo con una curva de densidad *N*(μ,σ^2)

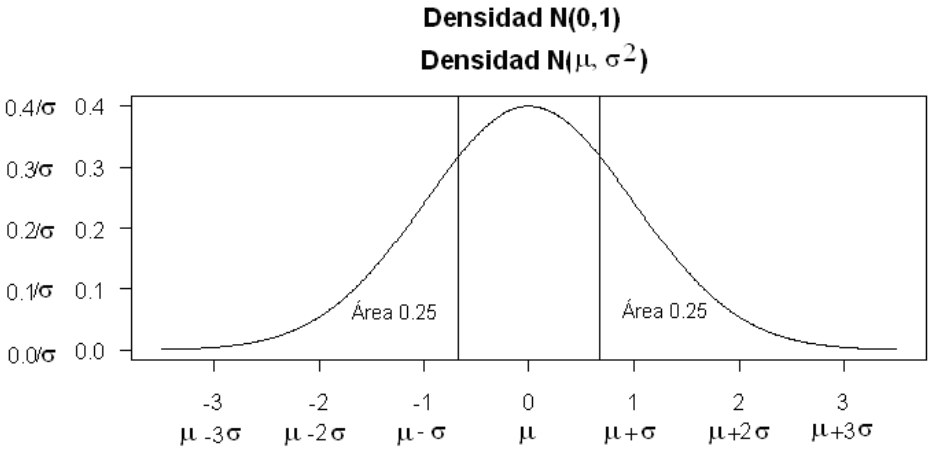
Proporción de datos que caen fuera de las vallas internas

En una distribución *N*(0,1) el cuarto inferior es **-0.6744898**

```
> qnorm(0.25)
[1] -0.6744898
```

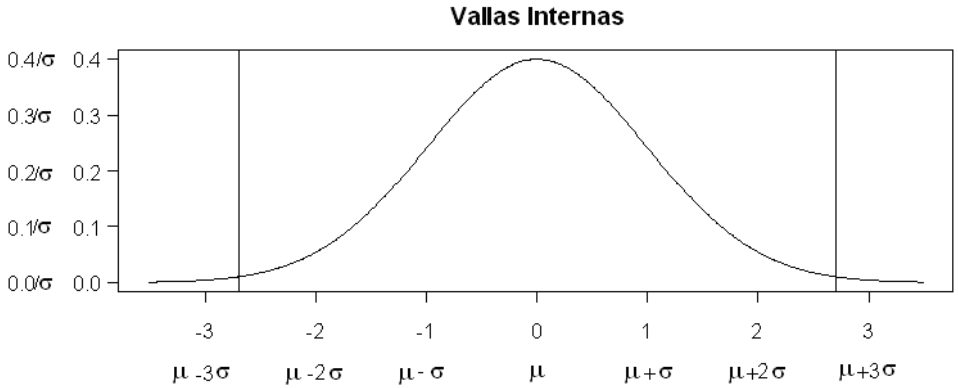
de manera que en una *N*(μ,σ^2) los cuartos inferior y superior son: μ - 0.6745σ y μ + 0.6745σ, dando

una distancia intercuartos d_Q= 1.348σ.



Esto hace que las vallas internas estén en

μ - 0.6745σ - 1.5 * 1.348σ = μ - 2.698σ
y
μ + 0.6745σ - 1.5 * 1.348σ = μ + 2.698σ,



dando un área de .00349 en cada cola.

```
> 2*pnorm(-2.698)
[1] 0.006975744
```

Por lo tanto la proporción de la distribución que cae debajo de μ - 2.698σ ó encima de μ + 2.698σ es .00698.