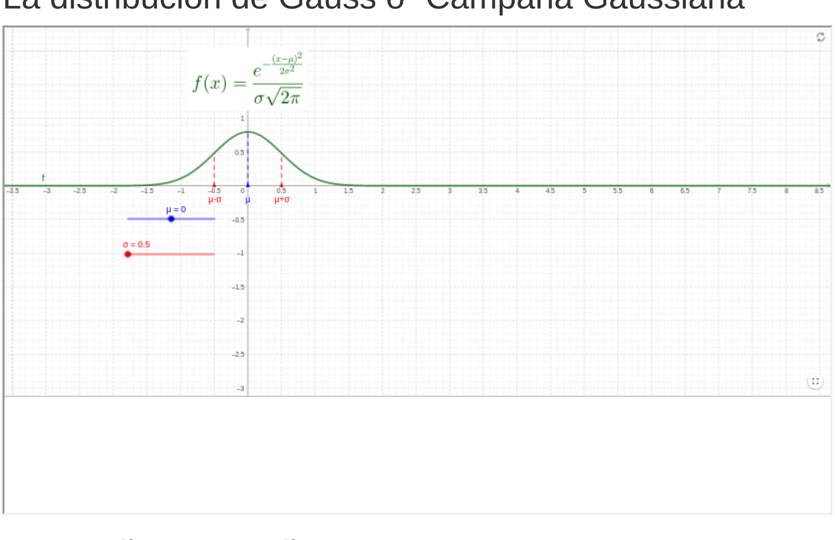
Función de distribución

Probabilidad Aplicada 3-602

La distribución de Gauss o “Campana Gaussiana”



Percentiles y cuartiles

Los **percentiles** son una medida de posición dentro de una distribución, que no necesariamente describen la tendencia central (como la media, mediana o moda que sí lo hacen). Un **percentil** es el valor que si ordenamos los datos de la distribución de menor a mayor, deja un determinado porcentaje de registros por debajo. Por ejemplo, el percentil 15 (P_{15}) de una variable, deja por debajo, el 15% de los registros. Por otro lado, el percentil 50 (P_{50}) representa la **mediana** y coincide con ella.

Hay tres percentiles que reciben un nombre específico: **cuartiles**. El percentil 25 (P_{25}) es el **primer cuartil** $ (Q_{1})$. El percentil 50 (P_{50}) es el **segundo cuartil** $ (Q_{2})$ y mediana también. Y finalmente, el percentil 75 (P_{75}) es el **tercer cuartil** $ (Q_{3})$. De esta manera, los tres cuartiles dividen la distribución en cuatro segmentos con el 25% de las observaciones en cada uno.

En R, se pueden calcular los percentiles y cuartiles con la función **quantile()**

```
quantile(insurance$edad)
```

```
##   0%  25%  50%  75% 100%
##   18   27   39   51   64
```

Esto se interpreta de la siguiente manera: El 25% de los asegurados tienen 27 años o menos y que por ejemplo, el 75% de los asegurados tiene 51 años o menos.

También se pueden especificar los percentiles así:

```
quantile(insurance$edad, c(0.15, 0.85))
```

```
## 15% 85%
##  22  56
```

De esta forma, tenemos los percentiles 15 y 85. Entonces el 15% de los asegurados tiene 22 años o menos y el 85% tiene 56.

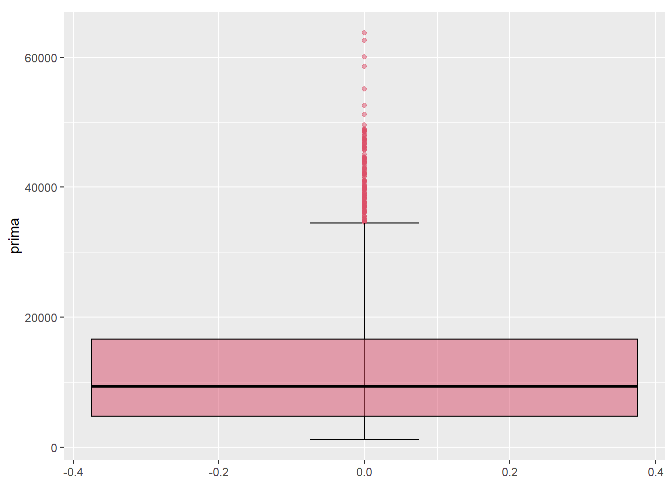
Representación gráfica de los cuartiles

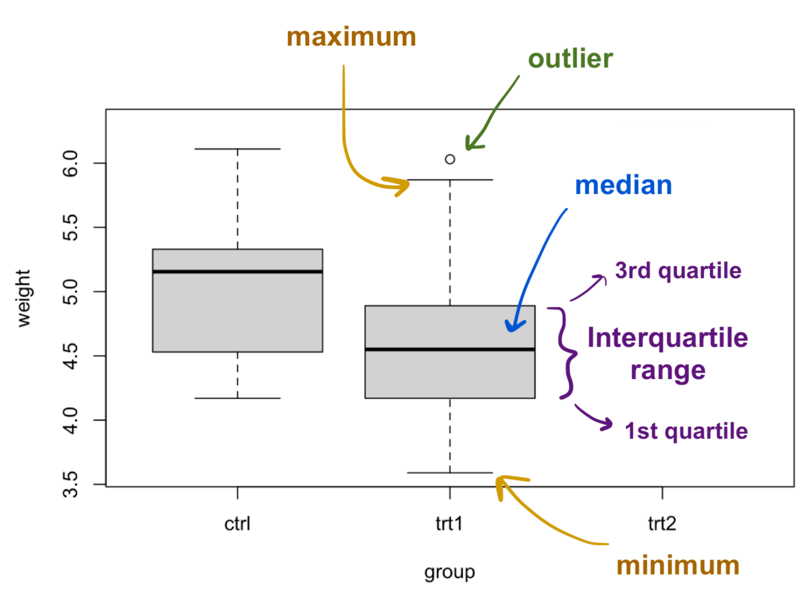
Una forma de visualizar los cuartiles es a través de un **boxplot**. Un *boxplot* o diagrama de caja y bigotes, es un gráfico estandarizado que se utiliza comúnmente para representar datos. Es más completo que el gráfico de barras y más complejo. Está compuesto por:

- La caja central, que representa el rango desde Q_1 a Q_3 o intervalo intercuartílico
- La mediana, coincidente con Q_2, mostrada como una línea dentro de la caja.
- Los bigotes, que cubren la distancia hasta el valor más bajo y más alto, pero sin sobrepasar 1.5 veces el valor de Q_1 y Q_3 respectivamente. Cualquier punto más allá, se considera un valor atípico.
- Valores atípicos que se dibujan como puntos mas allá de los bigotes. Pueden ser candidatos a *outliers* (valores extremos).

En R la geometría que utilizamos para lograr este gráfico es **geom_boxplot()**. Veamos un ejemplo:

```
g1 <- ggplot(data = insurance, aes(x = prima)) + geom_boxplot()
```





Lectura e interpretación del boxplot

Medidas de variabilidad

Una vez caracterizada la tendencia central de una variable, el siguiente paso es entender la **variabilidad** de los datos en torno a la medida central. Por ejemplo, en un grupo de 5000 personas que tienen una edad media de 40 años, no es lo mismo que tengan entre 35 y 45 años, o que por el contrario los valores varíen de 10 a 70 años.

Rango

Mide la distancia entre el menor y el mayor de los datos.

```
rango <- max(insurance$edad) - min(insurance$edad)
rango
```

```
## [1] 46
```

Desvación estandar y varianza

Varianza: $s^2 = \frac{(x_1-\overline{x})^2 + (x_2-\overline{x})^2 + (x_3-\overline{x})^2 + \dots}{n-1}$

Desvío estandar: $s = \sqrt{\frac{(x_1-\overline{x})^2 + (x_2-\overline{x})^2 + (x_3-\overline{x})^2 + \dots}{n-1}}$

En R el desvío se calcula con **sd()**:

```
sd(insurance$IMC)
```

```
## [1] 6.098187
```

Para calcular la varianza, elevamos al cuadrado:

```
sd(insurance$IMC)^2
```

```
## [1] 37.18788
```

En este caso, la desviación estandar de la variable IMC es $6.1 kg/m^2$.

Intervalo intercuartilico

El intevalo intercuartílico es adecuado cuando los datos no se distribuyen normalmente. En este caso, la mediana es la medida de tendencia central. Se define como la diferencia entre el tercer cuartil Q_3 y el primer cuartil Q_1. Por lo tanto, el IQR describe el rango de valores donde se encuentran el 50% de los registros centrales de la variable a estudiar.

```
IQR(insurance$IMC)
```

```
## [1] 8.3975
```

La distribución Normal II- Representación en R

¿Qué representan los parámetros μ y σ en la función gaussiana?

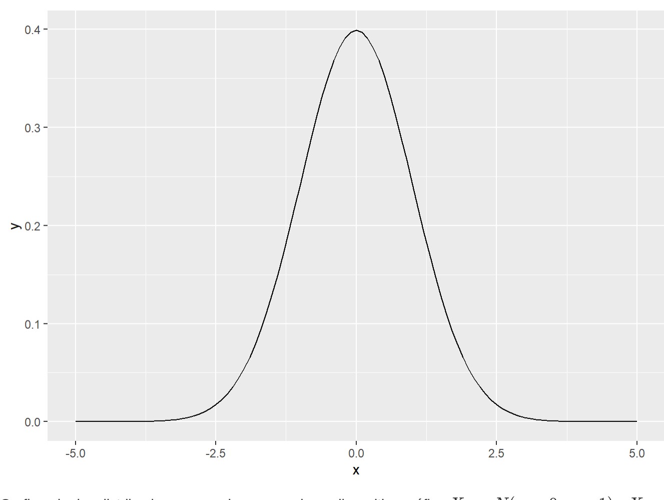
μ: media de la distribución.

σ: desvío de la distribución.

Si X es una variable aleatoria continua que se distribuye normalmente con media μ y desvío σ, lo representamos así: $X \sim N(\mu; \sigma)$.

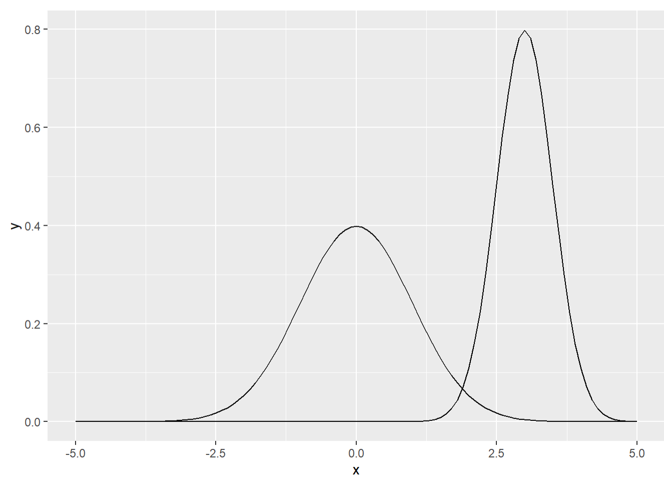
En R la función **geom_function** permite graficar funciones y la función **dnorm** permite graficar la normal…

```
g2 <- ggplot(data = data.frame(x = c(-5, 5)), aes(x = x)) + geom_function(fun = dnorm)  # si no se espec
```



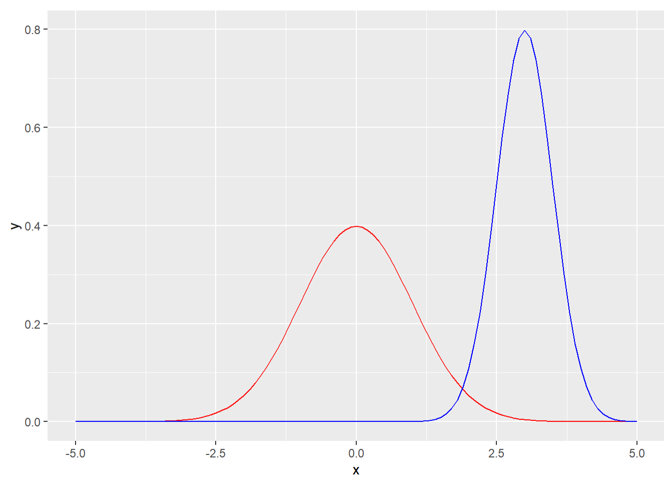
Graficando dos distribuciones normales en un mismo dispositivo gráfico $X_1 \sim N(\mu = 0; \sigma = 1)$ y $X_2 \sim N(\mu = 3; \sigma = 0.5)$

```
g3 <- ggplot(data = data.frame(x=c(-5,5)),aes(x=x))+
  geom_function(fun = dnorm) + # si no se especifican argumentos, se asume mu=0 y sigma=1
  geom_function(fun = dnorm, args = list(3,0.5)) # se especifican los argumentos con mu=3 y sigma =0.5
```



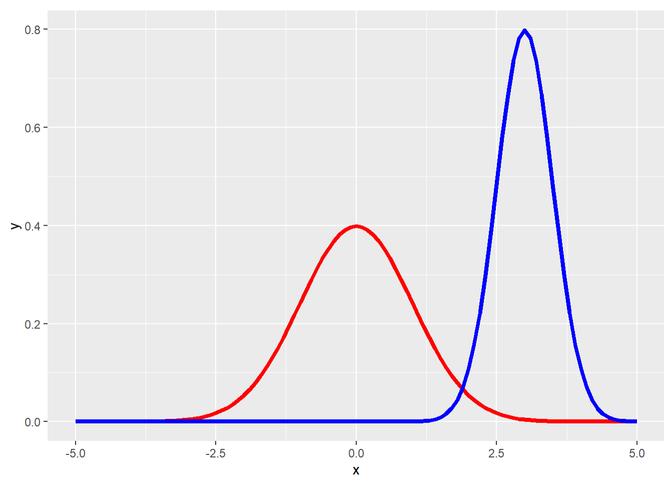
Mejorando la visualización: Agregando color a las curvas

```
g4 <- ggplot(data = data.frame(x=c(-5,5)),aes(x=x))+
  geom_function(fun = dnorm, colour="red") + #especificamos el color
  geom_function(fun = dnorm, args = list(3,0.5),colour="blue") #especificamos el color
```



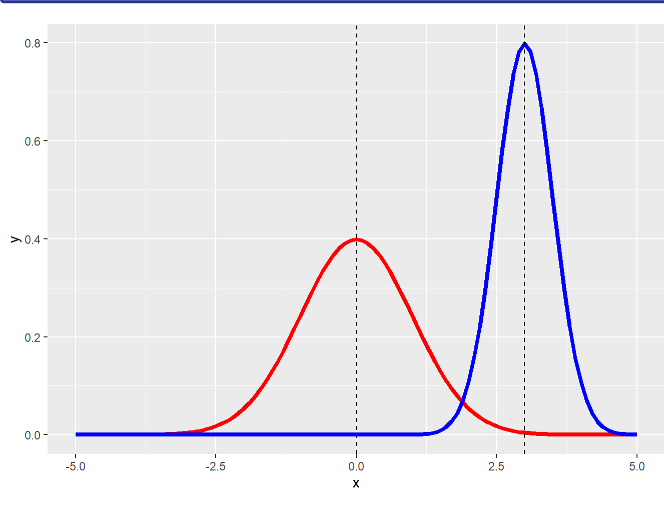
Mejorando la visualización: Engrosando lineas

```
g5 <- ggplot(data = data.frame(x=c(-5,5)),aes(x=x))+
  geom_function(fun = dnorm, colour="red", linewidth=1.5) + #especificamos el tamaño con linewidth=
  geom_function(fun = dnorm, args = list(3,0.5),colour="blue", linewidth=1.5) #especificamos el tamaño co
```



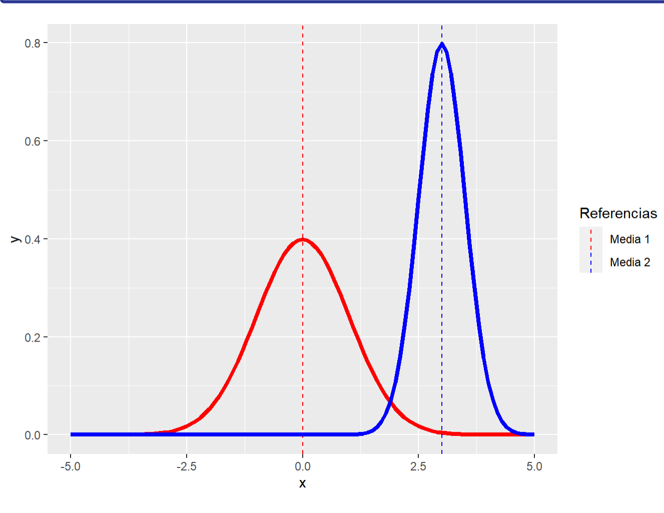
Mejorando la visualización: Agregando linea de media

```
g5 <- ggplot(data = data.frame(x=c(-5,5)),aes(x=x))+
  geom_function(fun = dnorm, colour="red", linewidth=1.5) +
  geom_vline(aes(xintercept = 0), linetype = "dashed")+#Con xintercept=0 indicamos dónde queremos trazar
  geom_function(fun = dnorm, args = list(3,0.5),colour="blue", linewidth=1.5) +
  geom_vline(aes(xintercept = 3), linetype = "dashed") #Con xintercept=3 indicamos dónde queremos trazar
```



Mejorando la visualización: Agregando color y referencias a las lineas

```
g6 <- ggplot(data = data.frame(x = c(-5, 5)), aes(x = x)) + geom_function(fun = dnorm,
    colour = "red", linewidth = 1.5) + geom_vline(aes(xintercept = 0, color = "Media 1"),
    linetype = "dashed") + geom_function(fun = dnorm, args = list(3, 0.5), colour = "blue",
    linewidth = 1.5) + geom_vline(aes(xintercept = 3, color = "Media 2"), linetype = "dashed") +
    scale_color_manual(name = "Referencias", breaks = c("Media 1", "Media 2"), values = c(`Media 1` = "re
        `Media 2` = "blue"))
```



Probabilidad acumulada en la Normal.

Representación de las regiones de probabilidad acumulada

¿Cómo se calcula la probabilidad acumulada para un valor de la variable aleatoria?

Utilizaremos la función **pnorm**

Supongamos que deseamos calcular $P(X \leq 0)$ donde $x \sim N(\mu = 0, \sigma = 1)$ es decir, la probabilidad acumulada hasta $x = 0$. Para ello…

```
P <- pnorm(q = 0, mean = 0, sd = 1, lower.tail = TRUE)
P
```

```
## [1] 0.5
```

Entonces: $P(X \leq 0) = 0.5$. De esta forma, la variable $x = 0$ deja acumulada un 50% de probabilidad.

Para utilizar la función **pnorm** es necesario ingresa:

q valor de la variable x.

mean valor de μ

sd valor de σ

lower.tail que puede tomar TRUE o FALSE. Con TRUE le indicamos que la probabilidad es acumulada a izquierda o por debajo del valor x.

Veamos otro ejemplo. Supongamos que deseamos calcular $P(X \leq 1.78)$ donde $x \sim N(\mu = 0, \sigma = 1)$ es decir, la probabilidad acumulada hasta $x = 1.78$.

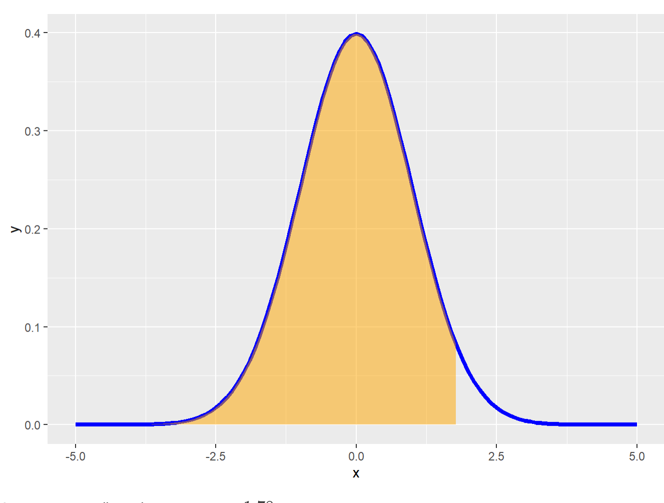
```
P <- pnorm(q = 1.78, mean = 0, sd = 1, lower.tail = TRUE)
P
```

```
## [1] 0.962462
```

Entonces: $P(X \leq 1.78) = 0.96$. De esta forma, la variable $x = 1.78$ deja acumulada un 96% de probabilidad.

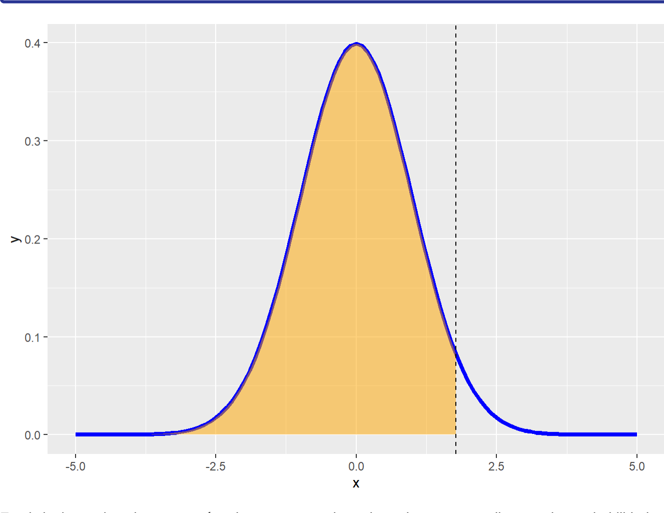
¿Cómo lo representamos gráficamente?

```
g7 <- ggplot(data = data.frame(x = c(-5, 5)), aes(x = x)) + geom_function(fun = dnorm,
    colour = "blue", linewidth = 1.5) + geom_area(stat = "function", fun = dnorm,
    fill = "orange", alpha = 0.5, xlim = c(-5, 1.78))
```



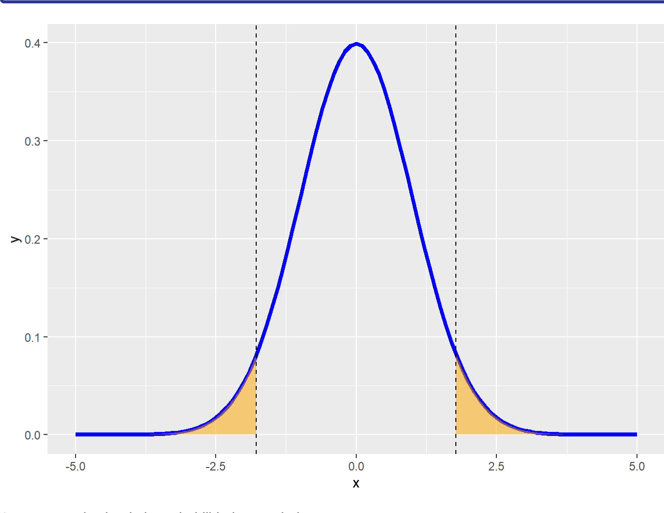
Agregamos una linea de corte en $x = 1.78$

```
g8 <- ggplot(data = data.frame(x = c(-5, 5)), aes(x = x)) + geom_function(fun = dnorm,
    colour = "blue", linewidth = 1.5) + geom_area(stat = "function", fun = dnorm,
    fill = "orange", alpha = 0.5, xlim = c(-5, 1.78)) + geom_vline(aes(xintercept = 1.78),
    linetype = "dashed")
```



En el siguiente ejemplo, vemos cómo hacer para sombrear las colas correspondientes a las probabilidades acumuladas.

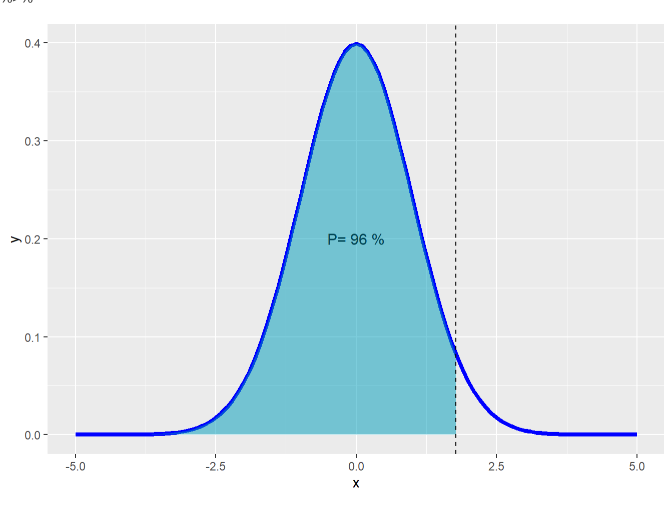
```
g8 <- ggplot(data = data.frame(x = c(-5, 5)), aes(x = x)) + geom_function(fun = dnorm,
    colour = "blue", linewidth = 1.5) + geom_area(stat = "function", fun = dnorm,
    fill = "orange", alpha = 0.5, xlim = c(-5, -1.78)) + geom_area(stat = "function",
    fun = dnorm, fill = "orange", alpha = 0.5, xlim = c(1.78, 5)) + geom_vline(aes(xintercept = -1.78),
    linetype = "dashed") + geom_vline(aes(xintercept = 1.78), linetype = "dashed")
```



Agregamos el valor de la probabilidad acumulada…

```
g9 <- ggplot(data = data.frame(x = c(-5, 5)), aes(x = x)) + geom_function(fun = dnorm,
    colour = "blue", linewidth = 1.5) + geom_area(stat = "function", fun = dnorm,
    fill = "#049DBF", alpha = 0.5, xlim = c(-5, 1.78)) + geom_vline(aes(xintercept = 1.78),
    linetype = "dashed") + geom_text(aes(x = 0, y = 0.2, label = paste("P=", round(P,
    2) * 100, "%")), stat = "unique", size = 4, color = "#044A59")
```

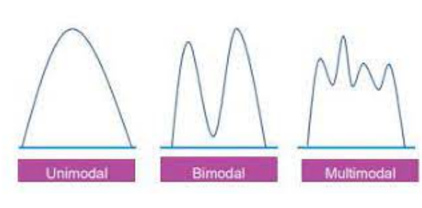
%>%



Medidas de forma

Modalidad

Cuando las varibles siguen una distribución normal, tienen un único pico o máximo, en este caso, se llaman **unimodales**. Otras distribuciones, que tienen dos o más picos, se llaman **bimodales** o **multimodales**.



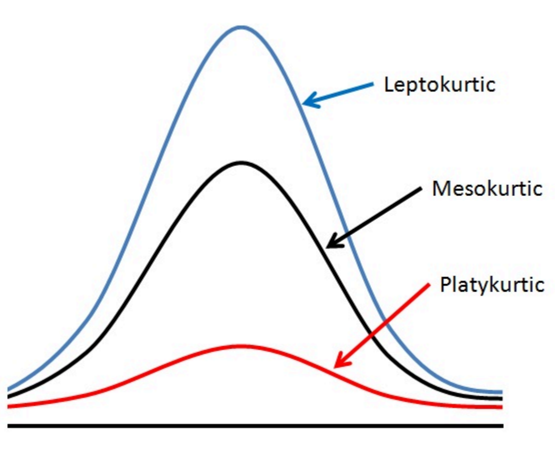
Asimetría

Simétrica: Cuando media $\approx$ mediana $\approx$ moda.

Asimetría positiva: Cuando media $>$ mediana $>$ moda.

Asimetría negativa: Cuando media $<$ mediana $<$ moda.

Kurtosis



Cuantiles y la probabilidad

Un cuantil es aquel punto que divide a la distribución de una variable aleatoria en intervalos regulares. ya vimos algunos cuantiles importantes como los percentiles, deciles y cuartiles.

Surge la siguiente cuestión, dada una probabilidad, ¿cuál es el valor de la varibale (cuantil) que deja esa probabilidad por debajo?

Es decir, ¿cuál es x_0 tal que $P(X \leq x_0) = p$?

Supongamos que $X \sim N(\mu = 0, \sigma = 1)$ entonces ¿cuál es x_0 tal que $P(X \leq x_0) = 0.9$? es decir, ¿cuál es el valor de x_0 (cuantil) que deja por debajo el 90% de los datos por debajo de él?

Esta pregunta se resuelve con la función inversa de **pnorm** que es **qnorm** la letra **q** hace referencia a “quantil”.

```
x_0 <- qnorm(0.9, mean = 0, sd = 1)
x_0
```

```
## [1] 1.281552
```