FEIR 20: Introducción a los contrastes

Apuntes del curso FEIR3, curso 2014/15 actualizados. `Última actualización:` jueves 04 abril 2019, 17:59:38

Laura del Río Alonso y Antonio Maurandi López

Índice

1. Población y Muestra

La meta más simple a la hora de analizar datos es *"sacar las conclusiones más fuertes con la cantidad limitada de datos de los que disponemos"*(Motulsky, 1995).

Se nos planetan dos problemas a la hora de abordar esta tarea:

- Diferencias importantes pueden ser oscurecidas por *variabilidad biológica* e *imprecisión experimental*. Se hace complicado distinguir entre *diferencias reales y variabilidad aleatoria.*
- El ser humano es capaz de encontrar modelos. Nuestra inclinación natural (especialmente con nuestros propios datos) es concluir que las diferencias observadas son ***REALES***: *tendemos a minimizar los efectos de la variabilidad aleatoria.* Como se dicen Martin Krzywinski y Naomi Altman en la columna de Nature Methods *"Points of significance"*: *somos de forma natural buscadores de patrones y debemos reconocer los límites de nuestra intuición* (Krzywinski & Altman, 2013).
- El rigor estadístico nos previene de cometer estos errores. (*"si lo empleamos bien"*).

Un ejemplo: Los estudiantes de Bioestadística reciben diferentes calificaciones en la asignatura (variabilidad). *¿A qué puede deberse?*

Podría deberse a *diferencias individuales en el conocimiento de la materia*, ¿no?, ¿podría haber otras razones? (*otras fuentes de variabilidad*).

Por ejemplo supongamos que todos los alumnos poseen el mismo nivel de conocimiento. ¿Las notas serían las mismas en todos?: Seguramente No.

- Dormir poco el día del examen.
- Diferencias individuales en la habilidad para hacer un examen.
- El examen no es una medida perfecta del conocimiento.
- Variabilidad por error de medida.
- En alguna pregunta difícil, se duda entre varias opciones, y al azar se elige la mala.
- Variabilidad por azar, aleatoriedad.

```
La idea básica de las estadística es extrapolar, desde los datos recogidos,
para llegar a conclusiones más generales sobre la población de la que se han recogido los datos.
```

Los estadísticos han desarrollado métodos basados en un modelo simple:

- Si razonablemente asumimos que los datos han sido obtenidos mediante un muestreo aleatorio de una población infinita. Analizamos estos datos y hacemos inferencias sobre la población.

Pero no siempre es *tan ideal.* En un experimento típico no siempre tomamos una muestra de una población, pero queremos extrapolar desde nuestra muestra a una situación más general. En esta situación aún podemos usar el concepto de población y muestra si definimos la **muestra** como los *"datos recogidos"* y la **población** como *"los datos que habríamos recogido si repitiéramos el experimento un número infinito de veces"* (Motulsky, 1999).

No sólo es necesario que los datos provengan de una población. También es necesario que cada sujeto, (cada observación) sea 'escogido' independientemente del resto.

Ejemplos:

- Si realizas un experimento biomédico 3 veces, y cada vez por triplicado, no tenemos 9 valores independientes. Si promediamos los triplicados, entonces tenemos 3 valores medios independientes.
- Si en un estudio clínico muestreamos 10 pacientes de una clínica y otros 10 de un Hospital: no hemos muestreado 20 individuos independientes de la población. Probablemente hemos muestreado dos poblaciones distintas.

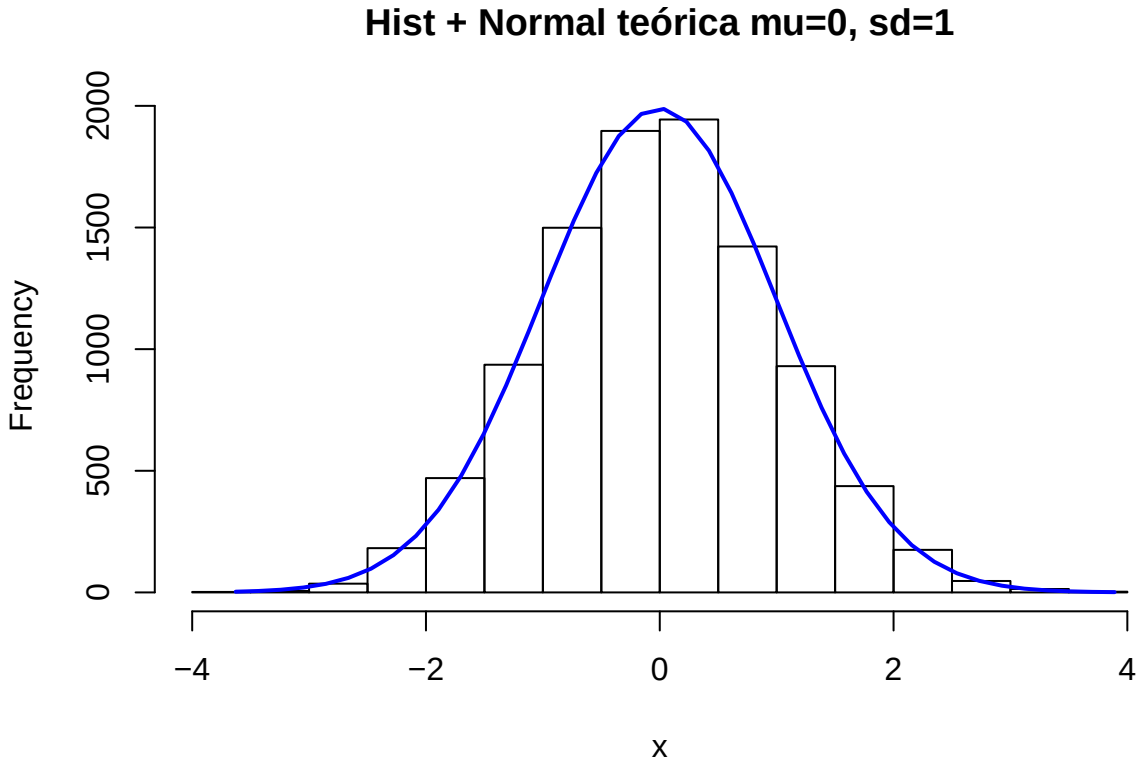
Hay tres enfoques básicos.

- El primer método consiste en asumir que la población sigue un distribución especial conocida como por ejemplo la Normal o Gaussiana (campana de Gauss).
  - Los tests te permiten hacer inferencias sobre la media (y también sobre otras propiedades).
  - Los tests más conocidos pertenecen a este enfoque.
  - También se conoce como enfoque **paramétrico**.
- El segundo enfoque consiste en ordenar los valores y ordenarlos de mayor a menor (rangos) y comparar distribuciones de rangos.
  - Es el principio básico de los **tests no-paramétricos.**
- El tercer enfoque es conocido como **'Resampling'** (se escapa de los objetivos de este curso, aquí también incluiríamos los métodos conocidos como *bootstraping*).
- Incluso hay un cuarto enfoque: **Métodos Bayesianos** (que también se escapan de los objetivos de este curso).

Ya hemos dicho que la idea básica de las estadística es extrapolar desde los datos recogidos para llegar a conclusiones más generales sobre la población de la que proceden los datos. El problema es que sólo podemos aplicar las inferencias a la población de la que hemos obtenido las muestras y *'generalmente'* queremos ir más lejos. Desafortunadamente la estadística no nos puede ayudar con estas extrapolaciones. En este caso se necesita **juicio científico y sentido común** para hacer inferencias más allá de las limitaciones de la estadística. Así *el análisis estadístico es sólo parte de la interpretación de nuestros datos* (Motulsky, 1999).

2. La distribución Normal

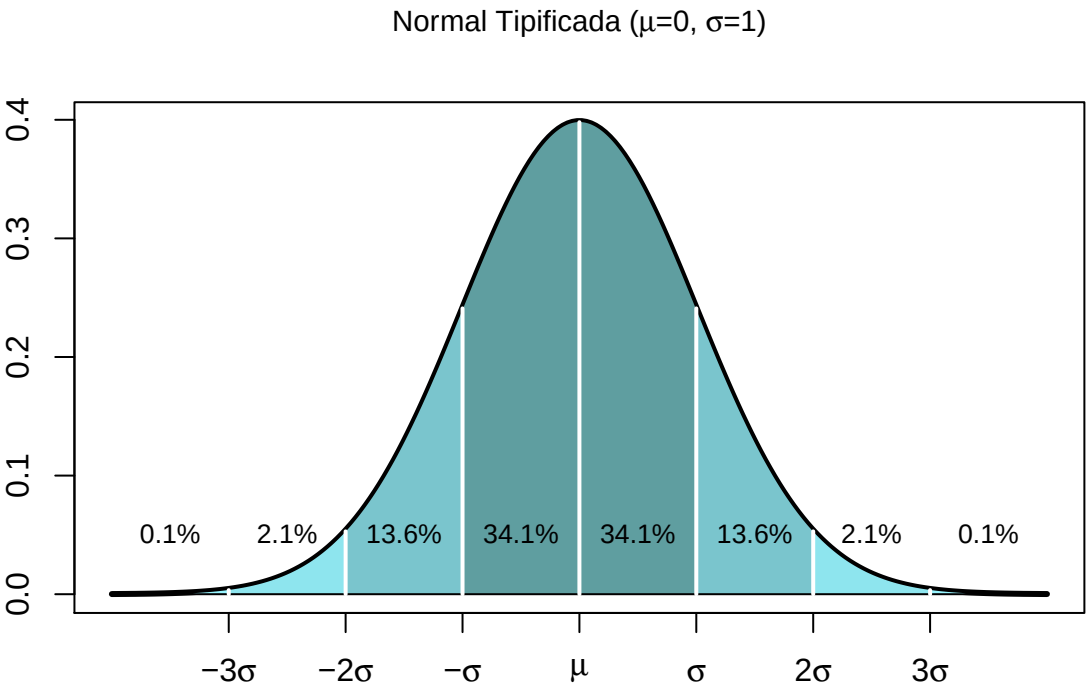
La distribución normal tiene características matemáticas especiales que la hacen la base de la mayoría de los test estadísticos. La razón: el Teorema Central del Límite (TCL) que veremos más adelante ("Distribución normal- wikipedia, la enciclopedia libre," n.d.).



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

¿Qué cosas observamos en esta distribución?

- La distribución de probabilidad normal es simétrica alrededor de su media. **(Coef simetría=0)**
- **media=mediana=moda** (todas las medidas de tendencia central coinciden).
- La curva normal desciende suavemente en ambas direcciones a partir del valor central.
- Es **asintótica**, lo que quiere decir que la curva se acerca cada vez más al eje X pero jamás llega a tocarlo. Es decir, las "colas" de la curva se extienden de manera indefinida en ambas direcciones.
- **Coef. Kurtosis = 0**
- El área total encerrada por f(x) vale 1: $\int_{-\infty}^{+\infty} f(x)dx = P\left(-\infty < X < +\infty\right) = 1$
- Si X es una v.a. que se distribuye según una normal de media μ y sd σ, $Z = \frac{x-\mu}{\sigma}$ se distribuye como una normal tipificada, es decir, de media 0 y sd 1.



2.1. TCL: Teorema Central del Límite

El Teorema central del limite se puede enunciar de la siguiente manera ("Teorema del límite central- wikipedia, la enciclopedia libre," n.d.):

`Si tus muestras son suficientemente grandes, la distribución de las medias seguirá una distribución Normal con la misma media que la distribución original.`

y varianza σ^2/n

De una forma más sencilla...

1. Dada una población con una distribución cualquiera.
2. Aleatoriamente obtenemos varias muestras de esa población. Calculamos sus medias.
3. Construimos un histograma de la distribución de frecuencias de las medias (¡ojo! *de las medias*)
4. Esta distribución de medias sigue una distribución 'Normal'.

¿Qué significa suficientemente grandes?: Depende de cuán distinta sea la distribución de una normal.

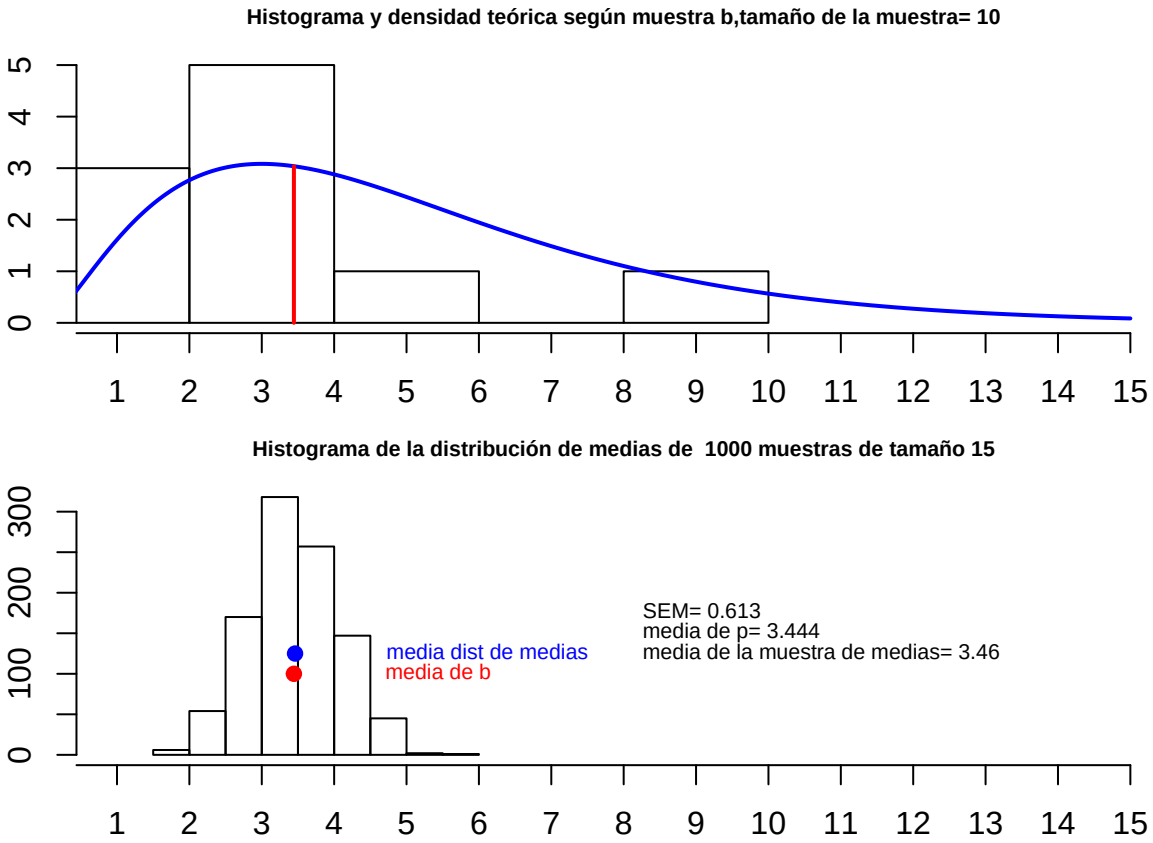
Así pues, el teorema del límite central *garantiza* una distribución normal cuando `n` es suficientemente grande.

Existen diferentes versiones del teorema, en función de las condiciones utilizadas para asegurar la convergencia. Una de las más simples establece que es suficiente que las variables que se suman sean independientes, idénticamente distribuidas, con valor esperado y varianza finitas.

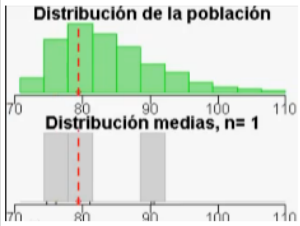
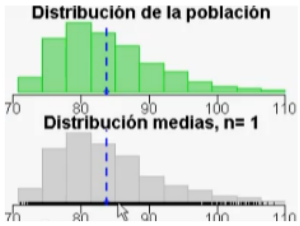
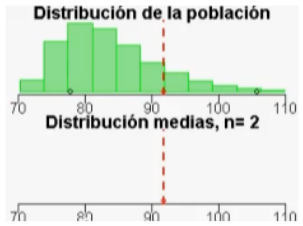
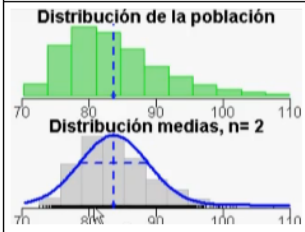
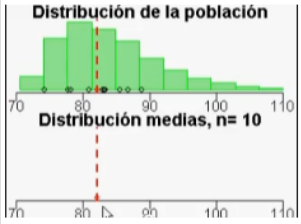
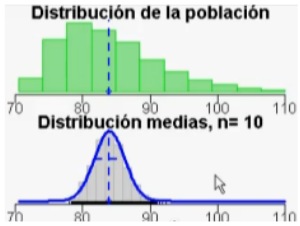
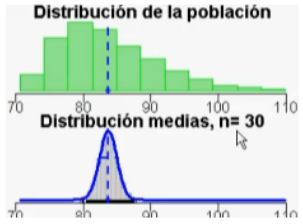
Nota: La aproximación entre las dos distribuciones es, en general, mayor en el centro de las mismas que en sus extremos o colas, motivo por el cual es preferible el nombre *"teorema del límite central"*, ya que *"central"* califica al límite, más que al teorema.

```
# simulación del TCL (aml!)
# b<-rbinom(1000,10,0.25);hist(b)
# b<-rnorm(1000)
# b<-runif(100,0,1);hist(b)
# b<-rexp(100,1/10);hist(b)
b<-rchisq(100,5);hist(b)
```

```
plot(density(b))

N1<-1000        # el número de repeticiones debe ser alto
N2<-15          # tamaño muestras. Probar con 2, 5, 15, 30, 50
vectordemedias<-NA
for (i in 1:N1){
    vectordemedias[i]=mean(sample(b,N2, replace=T))
  }

#vectordemedias
hist(vectordemedias,breaks=20) # observamos si se parece a una normal
SEM<-sd(vectordemedias);SEM  # indicador de la calidad de la estimación de la media
mean(vectordemedias)
mean(b)  #observad que difieren poco
```



Ver el vídeo del profesor F. J. Barón López de la Universidad de Málaga: [http://www.youtube.com/watch?v=FcDcJnw00hk] (http://www.youtube.com/watch?v=FcDcJnw00hk)

			
Distribución cualquiera, muy asimétrica	*Tomamos muchas (rep=2000) sub-muestras con **n=1***	*La dist de las frec se parece mucho ala dist original*	*Tomamos muchas (2000) sub-muestras con **n=2***
			
La dist de las medidas de las sub-muestras de n=2 se agrupan con una dist	*Tomamos muchas (2000) sub-muestras con **n=10***	*Mejoramos la estimación de la media de la dist original. La dist de las medias se parece a una N(μ, σ)*	*Cuando las sub-muestras son suficientemente grandes, n>30 el SEM se hace chico y la estimación es mejor.*

El TCL nos da una idea de lo importante que es en estadística la distribución normal o gaussiana, ya que muestreando, con muestras suficientemente grandes, podemos estimar la media *bastante bien* (dependerá del error típico de la media, es decir la precisión de la estimación de la media, hecha con una muestra, depende del tamaño de la muestra y de la desviación típica de la misma, a mayor muestras menor error típico de la media, *mayor precisión*), como sabemos que la distribución de las medias sigue una distribución normal, esto nos puede dar una idea de lo importante que es esta distribución (Braón-López, 2011).

¿Cómo de grande ha de ser la muestra?, pues tan grande como quiera que sea de pequeño el error típico de la media, es decir va a depender de la desviación típica de los valores de a muestra. Así que no hay una *receta mágica*, esta pregunta enlaza con el el concepto de potencia estadística que trataremos mas adelante.

veremos como el TCL es el que nos permite calcular intervalos de confianza, p-valores (contrastes) y tamaños muestrales.

¿Cómo evaluamos la normalidad de una distribución?: Hay muchas formas de abordar este problema, pero básicamente hay dos estrategias:

- **Gráficamente:** mediante histogramas, Q-Q plots...
- **Analíticamente:** con contrastes de hipótesis sobre normalidad: test de Shapiro-Wilk, de Kolmogorov–Smirnov, de Lilliefors, de Jaques-Bera... (los veremos luego)

Lo veremos en detalle más adelante.

3. Intervalos de confianza

Ya hemos visto que la media que calculamos de una muestra probablemente no sea igual a la media de la población. El tamaño de la diferencia dependerá del tamaño y de la variabilidad de la muestra (n, σ). Basándonos en el TCL, combinamos estos dos factores: **tamaño de la muestra** (n) y **variabilidad** (σ), para calcular intervalos de confianza a un determinado nivel de confianza; generalmente *95 %*.

IC(95 %) ~ n, σ

Si asumimos que tenemos una muestra aleatoria de una población. Podemos estar seguros al 95 % de que el IC (95 %) contiene a la media poblacional. Dicho de otra manera: si generamos 100 intervalos de confianza al 95 % de una población, se espera que contengan la media poblacional en 95 casos y no lo haga en 5. No sabemos cuál es la media poblacional, así que no sabemos cuándo ocurre esto.

Cómo los calculamos:

$$(\overline{x} - z_{\alpha/2}\frac{S}{\sqrt{n}}, \overline{x} + z_{\alpha/2}\frac{S}{\sqrt{n}})$$

$z_{\sigma/2}$ para la normal es muy próximo a 1.96, en R lo calculamos con la función: `qnorm(0.975)`.

```
# Intervalos de confianza para una Gaussiana
media <- 5
sd    <- 2
n     <- 20

inf<-NA; sup<-NA
ic    <- data.frame(inf,sup)

error <- qnorm(0.975) * sd / sqrt(n)

ic[1,1] <- media - error
ic[1,2] <- media + error
ic
```

```
##    inf  sup
## 1 4.12 5.88
```

$$\left.\begin{array}{r}(x_1, x_2, \cdots, x_n)\\ X \equiv D(\mu, \sigma)\end{array}\right\} \quad \Rightarrow \begin{cases} n \gg;\ \bar{X} \equiv N(\mu, \sigma^2/n) \\ n \ll;\ \bar{X} \equiv T_{n-1} \end{cases} \tag{1}$$

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \equiv t_{n-1}$$

```
# Intervalos de confianza para una T
media <- 5
sd    <- 2
n     <- 20

inf<-NA; sup<-NA
ic    <- data.frame(inf,sup)

error <- qt(0.975, df=n-1) * sd / sqrt(n)

ic[1,1] <- media - error
ic[1,2] <- media + error
ic
```

```
##    inf  sup
## 1 4.06 5.94
```

Podemos también hacer uso de la función `t.test()` para calcular IC.

```
 # IC para una media de una muestra
 x <- rnorm( n=20 , mean=5 , sd=2 )
 t.test( x )
```

```
##
##  One Sample t-test
##
## data:  x
## t = 10, df = 20, p-value = 7e-09
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  4.12 6.35
## sample estimates:
## mean of x
##      5.24
```

En R para crear intervalos de confianza para proporciones empleamos la función `prop.test()`.

```
# IC para una proporción
# 42 de 100 dijeron que sí,
prop.test( 42, 100 )
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  42 out of 100, null probability 0.5
## X-squared = 2, df = 1, p-value = 0.1
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.323 0.523
## sample estimates:
##    p
## 0.42
```

Para la mediana podemos calcularlo empleando la función `wilcox.test()`.

```
x <- c( 110, 12, 2.5, 98, 1017, 540, 54, 4.2, 150, 432)
mean( x )
```

```
## [1] 242
```

```
median( x )
```

```
## [1] 104
```

```
wilcox.test( x, conf.int=T )
```

```
##
##  Wilcoxon signed rank test
##
## data:  x
## V = 60, p-value = 0.002
## alternative hypothesis: true location is not equal to 0
## 95 percent confidence interval:
##   33 514
```

```
## sample estimates:
## (pseudo)median
##            150
```

¿Qué asumimos cuando interpretamos un IC para una media?:

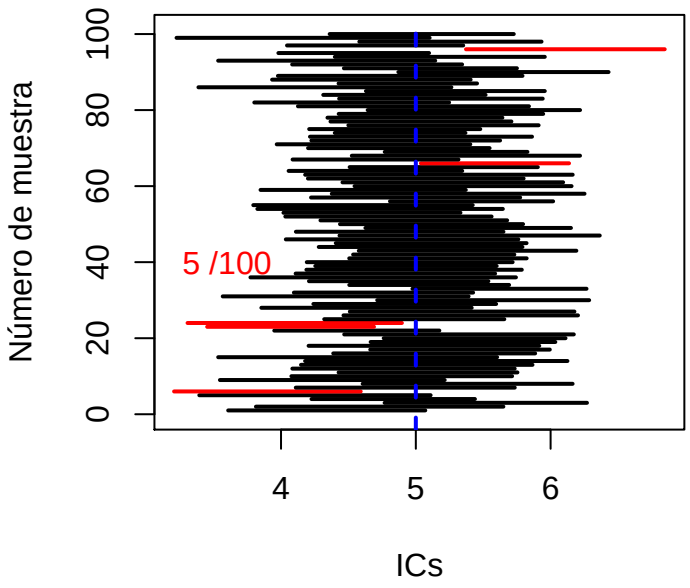
- Muestreo aleatorio de una población.
- La población se distribuye, aproximadamente, como una Normal.
  - Si n es grande no es tan importante está condición.
- Existe independencia de las observaciones.

Nota: Si se viola alguna de estas condiciones, el IC real es más 'ancho' que el calculado.

Así pues un intervalo de confianza es un rango de valores (calculado en una muestra) en el cual se encuentra el verdadero valor del parámetro, con una probabilidad determinada.

- Nivel de confianza $1-\alpha$: probabilidad de que el verdadero valor del parámetro se encuentre en el intervalo.
- Nivel de significación α: probabilidad de equivocarnos.
- Normalmente $1-\alpha = 95\,\%$ ($\alpha = 5\,\%$)

```
set.seed(30)
n  <- 30
mu <- 5
s  <- 2
Ns <- 100               # tomamos Ns samples
muestras = matrix( rnorm(Ns * n, mu, s),  n )

fci <- function(x) {
    t.test(x)$conf.int
}
conf.int = apply(muestras, 2, fci)

plot( range(conf.int)
      , c(0, 1 + Ns)
      , type = "n"
      , xlab = "ICs"
      , ylab = "Número de muestra"
      )
for (i in 1:Ns) {
    if ((conf.int[1, i]<=mu && conf.int[2, i]<=mu)|(conf.int[1, i]>=mu && conf.int[2, i]>=mu)) {
        lines(conf.int[, i], rep(i, 2), lwd = 2, col="red")
    } else lines(conf.int[, i], rep(i, 2), lwd = 2)

}
abline(v = mu, lwd = 2, lty = 2, col="blue")
cumplen <- sum( conf.int[1, ] <= mu & conf.int[2, ] >= mu)
text(mu-0.7*s, 40, paste(100-cumplen,"/100"), col="red")
```


Número de muestra
0
20
40
60
80
100
5 /100
4
5
6
ICs

4. P-valor. Contrastes de hipótesis

Observar medias muestrales diferentes no es suficiente evidencia de que sean diferentes las medias poblacionales. Es posible que sean iguales y que la diferencia observada se deba a una coincidencia, nunca se puede estar seguro. Lo único que se puede hacer es **calcular las probabilidades de equivocarnos**.

"Statistics does not tell us wheter we are rigth. It tells us the chances of beeing wrong"(Krzywinski & Altman, 2013)

Si las poblaciones tienen la misma media realmente: ¿cuál es la probabilidad de observar una diferencia tan grande o mayor entre las medias muestrales?

- El p-valor es una probabilidad de 0 a 1.
- Si p es pequeño, podemos concluir que la diferencia entre muestras ('probablemente') no es debida al azar.
- Concluiríamos que tiene distintas medias.

Otra forma de ver lo mismo.

- Los estadísticos hablan de hipótesis nula (H_0).
- La hipótesis nula dice que "no hay diferencia entre las medias".
- La hipótesis nula es lo contrario que la hipótesis experimental.
- Así podemos definir **p-valor** como la ***probabilidad de observar una diferencia tan grande o mayor que la observada si la hipótesis nula fuera cierta***.

Así pues:

- Se dice que rechazamos la hipótesis nula si $p < 0{,}05$ y la diferencia es estadísticamente significativa (ss).
- Si $p > 0{,}05$, **no rechazamos la hipótesis nula** y decimos que la diferencia no es estadísticamente significativa (ns).
- No podemos decir que la H_0 sea verdad, simplemente "no la rechazamos", es decir, no tenemos suficiente evidencia para rechazar la hipótesis de igualdad (la H_0).

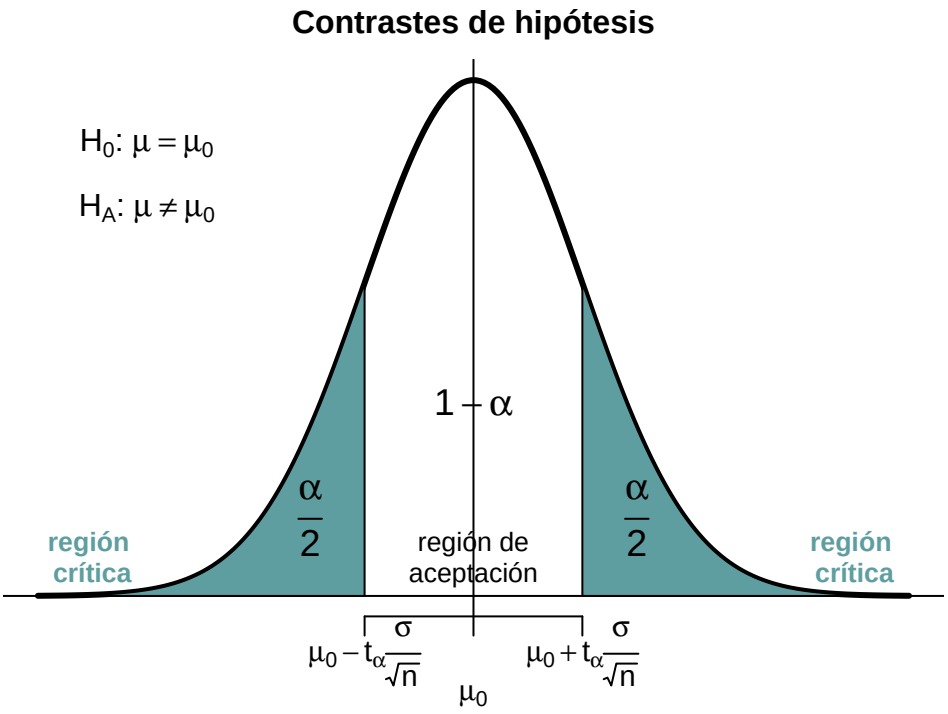
¿En qué se basa el cálculo del p-valor? Pues en el TCL y en las propiedades de la distribución normal, la cual tenemos tabulada.

Se puede también abordar el tema de los contrastes de hipótesis y cálculo del p-valor desde una perspectiva integradora con el concepto de intervalo de confianza (Bland, 2000).

Supongamos que deseamos hacer un contraste acerca de **un parámetro q, de la población**. Para llevarlo a cabo consideraremos la distribución de algún estadístico muestral que de alguna manera se corresponda y se relacione con el parámetro q. Designemos en general a este estadístico como T. **Si con los datos muestrales obtenemos un valor concreto para T tal que pertenezca a una determinada región del campo de variación de T optaremos por no rechazar la hipótesis y en caso contrario por rechazarla**. Obviamente la clave del problema será delimitar la región del campo de variación de T que consideraremos como zona de aceptación de la hipótesis. **Esto se resolverá por un criterio probabílistico partiendo de la distribución muestral de ese estadístico T.**

Región crítica: Región del campo de variación del estadístico tal que si contiene al valor evaluado del mismo con los datos muestrales nos llevará a rechazar la hipótesis. La designaremos por R_1.

Región de aceptación: Es la región complementaria a la región crítica. Si el valor evaluado del estadístico pertenece a ella: no rechazamos la hipótesis. La designaremos por R_0. Ambos conjuntos/regiones son disjuntos.



NOTA: La región de aceptación no es más que un intervalo de confianza al 95 % (si la significación es 0.05) del estadístico T.

Un ejemplo intuitivo: *Supongamos que un fabricante de coches dice que un determinado vehículo recorre 25m por galón de gasolina, 25mpg. El consumidor pregunta a 10 usuarios su consumo y en media le responden que gasta 22mpg con una sd de 1,5. ¿Nos engaña el fabricante?*

$$H_0 : \mu = 25$$
$$H_a : \mu < 25$$

(la función `t.test()` no funciona porque los datos vienen resumidos, ¡no tenemos los datos crudos!, así que construimos nosotros el estimador y el p-valor) (Verzani, 2002).

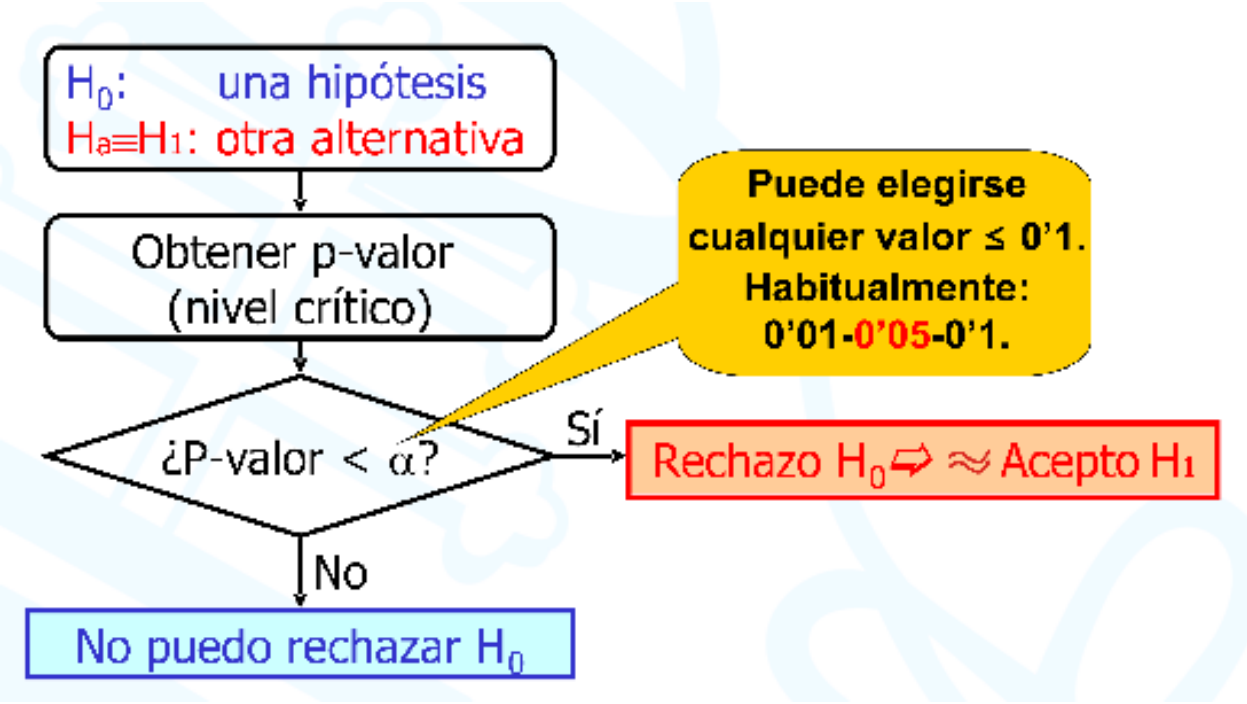
```
 # t test a mano (aml!)
 # asumimos mu=25 (HO)
 mua<-22
 sda<-1.5
 n<-10
 t<-(mua-25)/(sda/(sqrt(n)))
 t
```

```
## [1] -6.32
```

```
# ahora con la función 'pt', porque
# asumimos que el estadístico 't' sigue una dist t (TCL)
#calculamos la probabilidad de que ese valor se dé
pt(t,df=n-1)
```

```
## [1] 0.0000685
```

```
 # ¡Ojo!¡Hay que mosquearse...!
```


H0: una hipótesis
Ha≡H1: otra alternativa
Obtener p-valor
(nivel crítico)
¿P-valor < α?
Sí
No
Rechazo H0 ⇨ ≈ Acepto H1
No puedo rechazar H0
Puede elegirse
cualquier valor ≤ 0'1.
Habitualmente:
0'01-0'05-0'1.

5. Potencia Estadística

Cuando un experimento concluye diciendo que no se ha encontrado una *'diferencia significativa'*, no implica que no haya diferencia; simplemente no la hemos encontrado. Decimos entonces que no tenemos *suficiente evidencia para rechazar la hipótesis nula (la igualdad), así que la asumimos.*

Posibles razones para un p-valor no significativo ($p > 0{,}05$):

- En verdad no hay diferencia, es decir, es correcto nuestro test (nuestro p-valor) y no hay diferencias reales poblacionales.
- Sí hay diferencias poblacionales (no lo podemos saber con seguridad porque no conocemos el parámetro, sólo el estadístico/estimador).
  Algunas causas posibles:
- Tamaño de la muestra.
- Alta variabilidad.

Si éstas son las razones estamos cometiendo **error de tipo II**.

5.1. Error de tipo I y error de tipo II.

- Cometemos error de tipo II o β, cuando 'afirmamos' que no hay diferencias ($p > 0{,}05$) y en verdad sí las hay.
- Cometemos error de tipo I o α, cuando 'afirmamos' que sí hay diferencias ($p < 0{,}05$) y en verdad no las hay.

Llamamos **potencia estadística** de un test a la capacidad de un test para revelar diferencias que realmente existen.

¿Cuánta potencia tiene nuestro análisis para encontrar hipotéticas diferencias en el caso de existir éstas? La potencia depende del tamaño (n) y de la cantidad de variación de la muestra (d, desviación estándar o típica). También influye el tamaño de lo que es para nosotros una diferencia, o ¿cuánto han de diferir para considerarlos distintos?

$$n = (\frac{\sigma \times 1{,}96}{d})^2 \ , \ z_{1-\alpha/2} = 1{,}96 \ , \ \alpha = 0{,}05$$

Cuadro 1: Condición de la Población

Acción	H_0 Verdadera	H_A Falsa
Aceptar	Conclusión Correcta	Error de Tipo II
Rechazar	Error de Tipo I	Conclusión Correcta

5.2. Múltiples comparaciones.

Interpretar un p-valor es sencillo, *si la hipótesis nula (igualdad) es cierta hay un 5 % de posibilidades de que una selección aleatoria de sujetos muestre una diferencia tan grande (o mayor) como la que muestran"*. Pero interpretar muchos p-valores a la vez puede no ser tan sencillo. Si testeamos diferentes hipótesis nulas (independientes) a la vez, con un nivel de significación del `0.05` hay más de un `5%` de probabilidades obtener un resultado significativo por azar (Motulsky, 1995).

Ejemplo: Si hacemos 3 tests con $\alpha = 0{,}05$. La probabilidad de no cometer error de tipo I es **0.95 (95 %) para cada test**. Suponemos que los tests son independientes.

- La probabilidad general de **NO** cometer error de tipo I es: $(0{,}95)^3 = 0{,}875$.

- Así la probabilidad general de cometer error de tipo I es: $1-(0{,}95)^3 = 1-0{,}875 = 0{,}143$, es decir, un **14,3 %**. ¡¡Se ha incrementado de 5 % a 14,3 %!! con sólo 3 comparaciones.

Cuantas más comparaciones hagamos más crece el error de tipo I.

Cuadro 2: Comparaciónes múltiples

Nº hip nulas	Prob de al menos un p< 0,05	α para mantener Error de tipo I=0.05
1	5 %	0.05
2	10 %	0.0253
3	14 %	0.0170
4	19 %	0.0127
...	...	...
100	99 %	0.0005
N	$100(1-0{,}95^N)$	$(1-0{,}95^N)$

Veremos más sobre este problema y de cómo resolverlo (correcciones de la significación y comparaciones planificadas) más adelante.

6. Contrastes de normalidad

Cuando contrastamos lo primero que debemos tener en mente es cuál es la hipótesis nula. En los contrastes de normalidad la hipótesis nula es también conocida como la *hipótesis de normalidad*, es decir, *"no hay diferencias entre nuestra distribución y una distribución normal con esa media y esa sd"*.

Recordemos que hay muchas maneras de evaluar la normalidad:

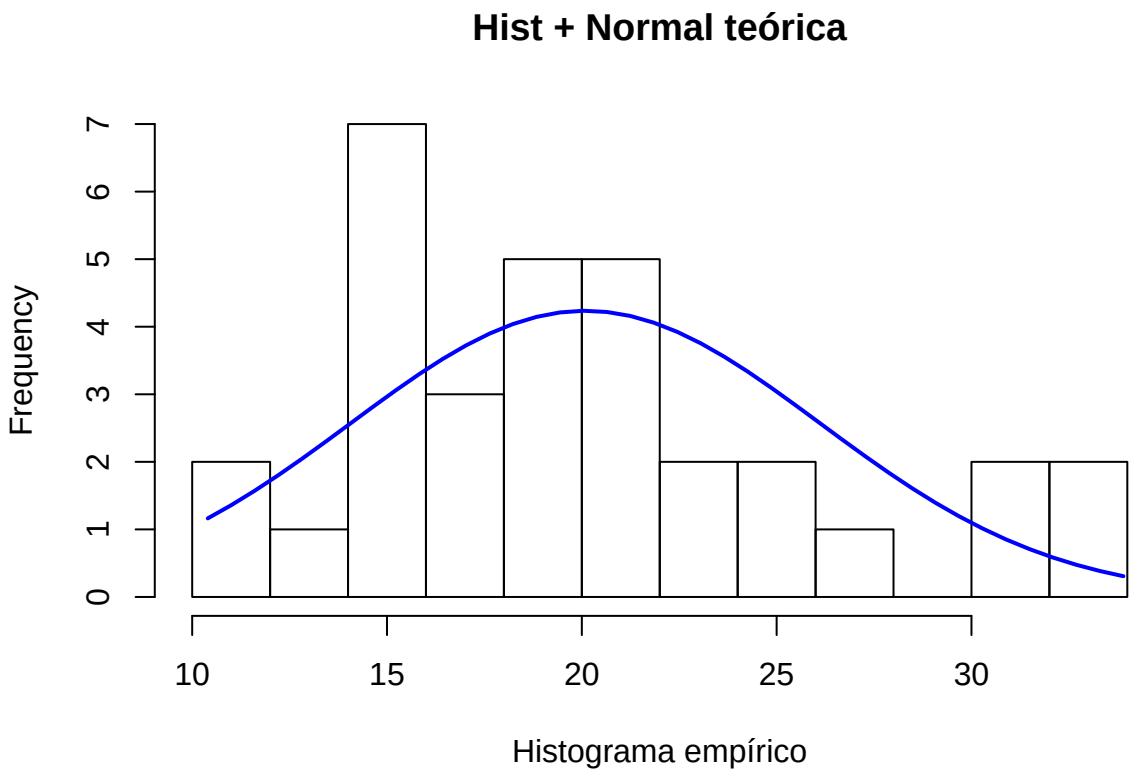
- **Gráficamente:** histogramas, Q-Q plots...
- **Analíticamente:** test de Shapiro-Wilk, de Kolmogorov–Smirnov, de Lilliefors, de Jarque-Bera... (los veremos luego)

6.1. Métodos Gráficos

6.1.1. Histogramas y curvas de densidad

La primera aproximación al problema de evaluar la normalidad nos la dan los histogramas cruzados con la curva de densidad teórica. De esta forma se representan los datos *empíricos* (*reales*) con la distribución teórica y se comparan *visualmente*. Es decir la distribución que deberían tener los datos empirícos de seguir una distribución normal con esa media y desrviación tipica y nuestros datos)

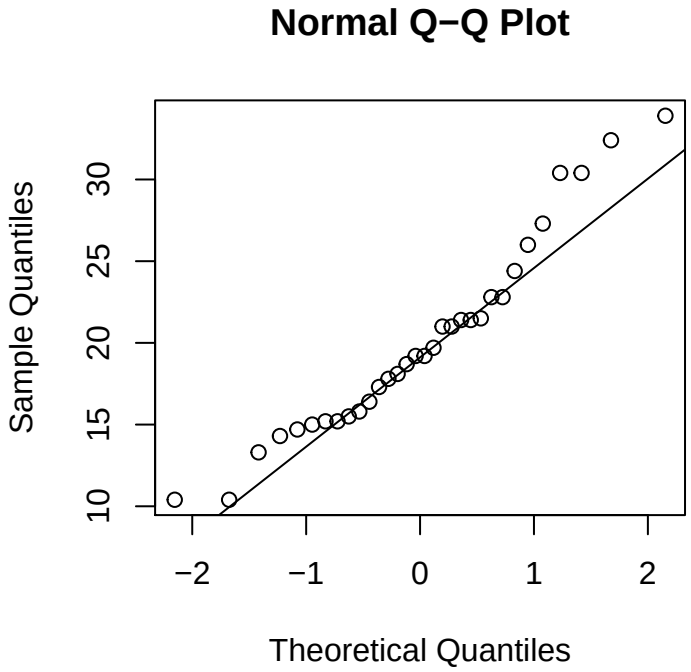
```
x <- mtcars$mpg
h<-hist(x, breaks=10, xlab="Histograma empírico" ,main= "Hist + Normal teórica")

xfit <- seq( min(x), max(x), length=40)
yfit <- dnorm( xfit,mean=mean(x), sd=sd(x))
yfit <- yfit * diff( h$mids[1:2]) * length(x)

lines( xfit, yfit, col="blue", lwd=2)
```



6.1.2. Gráficos Q-Q.

La manera de evaluarla mediante histogramas es muy *'rustica'* pues simplemente lo que se hace es comparar el histograma frente a la densidad teórica de una normal con esos parámetros. Los Q-Q plots son más informativos, *"Q"* viene de cuantil (*quantile* en inglés). El método se emplea para el diagnóstico de diferencias entre la distribución de probabilidad de una población de la que se ha extraído una muestra aleatoria y una distribución usada para la comparación, es decir, no sólo para contrastar normalidad.

Cuando queremos contrastar normalidad para una muestra de tamaño `n`, se dibujan n puntos con los $(n+1) - cuantiles$ de la distribución de comparación, la distribución normal, en el eje horizontal y el estadístico de k-ésimo orden (para k = 1,2,...n) de la muestra en el eje vertical. Si la distribución de la variable es la misma que la distribución de comparación se obtendrá, aproximadamente, una línea recta, especialmente cerca de su centro.

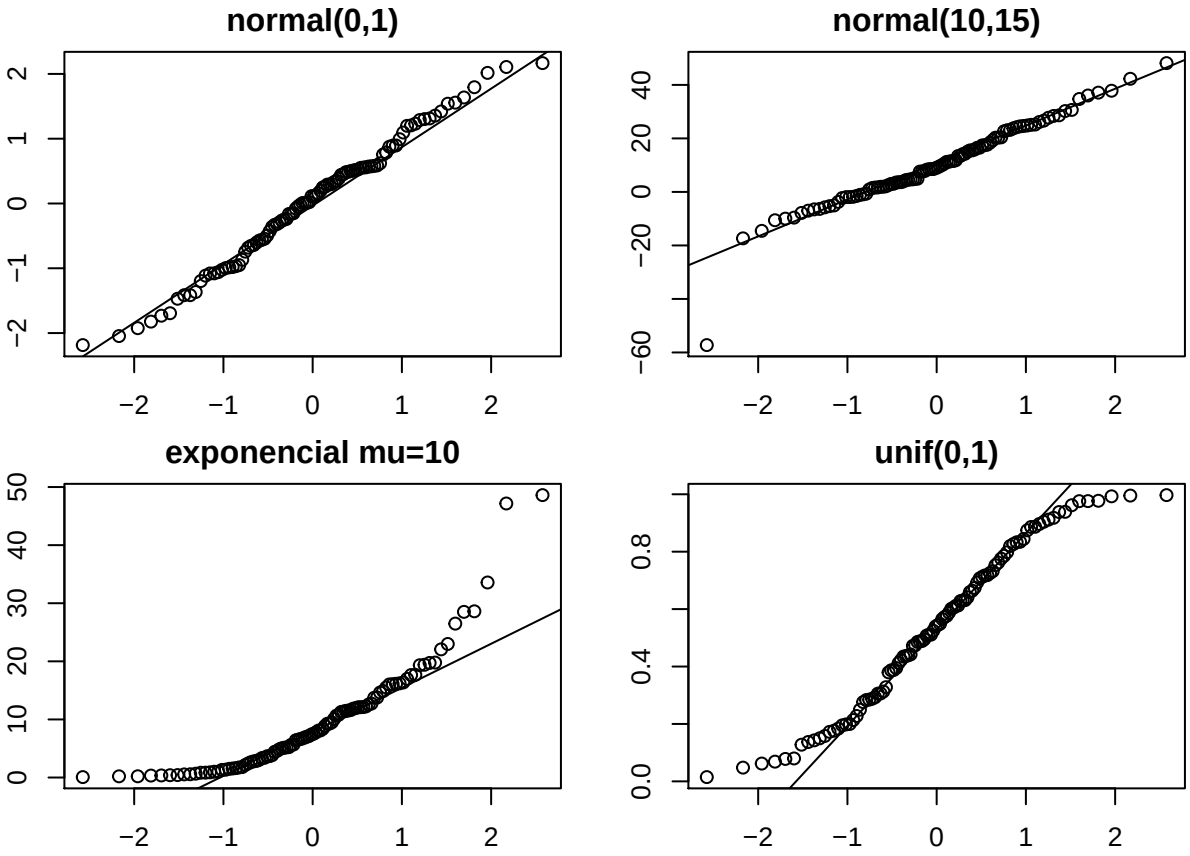


En R los podemos crear con la función `qqnorm()`.

```
qqnorm( mtcars$mpg)
qqline( mtcars$mpg)    # añadimos laa línea diagonal
```

Ejercicio: Compara los siguientes Q-Q plots.

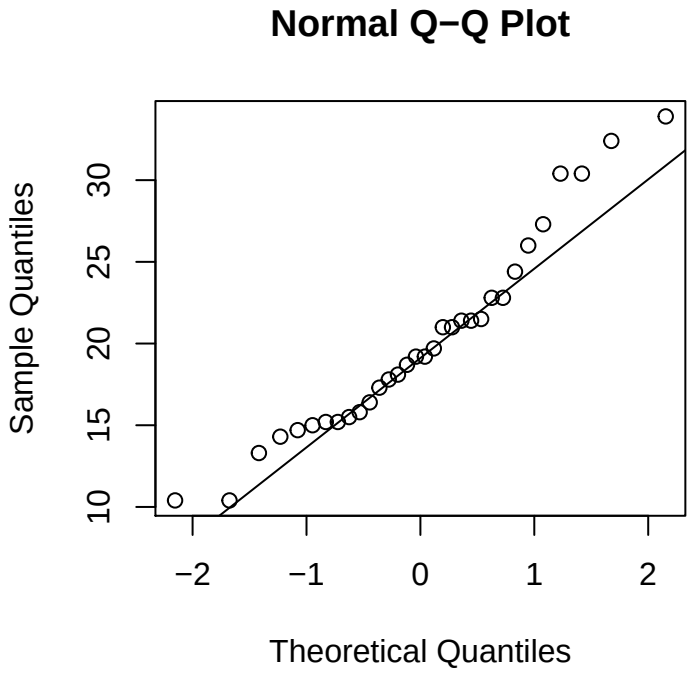
```
par( mfrow = c(2,2) )
par( mar = c(2,2,2,2) )

x<-rnorm( 100,0,1)  ;qqnorm( x, main="normal(0,1)")        ; qqline(x)
x<-rnorm( 100,10,15);qqnorm( x, main="normal(10,15)")      ; qqline(x)
x<-rexp(  100,1/10) ;qqnorm( x, main="exponencial mu=10") ; qqline(x)
x<-runif( 100,0,1)  ;qqnorm( x, main="unif(0,1)")          ; qqline(x)
```




```
par(mfrow=c(1,1))
```

En R los podemos crear con la función `qqnorm()`.

```
qqnorm( mtcars$mpg)
qqline( mtcars$mpg)
```



6.2. Métodos analíticos

6.2.1. Regla aproximativa

Regla aproximativa de evaluación de normalidad según la asimetría y la curtosis.

Curtosis y/o coeficiente de asimetría entre -1 y 1, es generalmente considerada una muy ligera desviación de la normalidad ((Bulmer, 1979), (Brown, n.d.)).

Entre -2 y 2 tampoco es malo del todo, según el caso.

```
# cálculo de la curtosis y el coef. de asimetría en R

g3 <- mean( (x-mean(x))^3) / sqrt(sd(x)^3 )
g4 <- mean( (x-mean(x))^4) / sqrt(sd(x)^4 )
```

6.2.2. Contraste sobre asimetría y curtosis.

Se trata siemplemente de ver si son atípicos. Primero normalizamos los coeficientes:

$$Z_s = \frac{S}{S \times E_s}$$

$$Z_k = \frac{K}{S \times E_k}$$

¡¡Ojo, para ambas la media es 0!!

Si alguno de los valores es mayor en valor absoluto que 1.96 se considera significativo ($|Z_s| > 1{,}96 =$ significativo). Y consideramos que hay demasiada asimetría (en su caso curtosis).

Nota: En muestras grandes es más interesante mirar a la distribución de forma gráfica (histograma).

6.2.3. Contraste de normalidad de Shapiro-Wilk

El test de Shapiro-Wilk funciona bien con muestras pequeñas (menores de 50), para muestras grandes es equivalente al test de kolmogorov-Smirnov que veremos luego. Se ejecuta con la función `shapiro.test()`.

```
shapiro.test(x)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  x
## W = 1, p-value = 0.007
```

6.2.4. Contraste de normalidad de Kolmogorov-Smirnov

El test de Kolmogorov-Smirnov puede testar si nuestra muestra proviene de una población con distribución cualquiera, no sólo de la normal. Se ejecuta con la función `ks.test()`. Es preferible a Shapiro-Wilk si las muestras son grandes (mayores a 100).

```
ks.test(x, y, ...,
        alternative = c("two.sided", "less", "greater"),
        exact = NULL)
```

```
# testamos si 'x' sigue una normal (0,1) (pnorm)
ks.test(x,"pnorm", 0, 1)
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  x
## D = 0.5, p-value <2e-16
## alternative hypothesis: two-sided
```

En este ejemplo rechazariamos la hipótesis nula (*rechazariamos la normalidad*) pues $p < 0{,}05$,

```
# testamos si 'x' sigue una normal de los parametros de x (pnorm)
 ks.test(x,"pnorm", mean(x), sd(x) )
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  x
## D = 0.07, p-value = 0.7
## alternative hypothesis: two-sided
```

En este ejemplo no rechazariamos la hipótesis nula ($p > 0{,}05$), así pues aceptamos que la muestra pueda provenir de una distribución normal con esos parametros (media y sd).

También puede ser usado par acontrastar si dos muestras siguen la misma ley, en el siguiente ejemp0lo contrastariamos si podemos aceptar que una muestra que sigue una normal sigue la misma ley que una muestra que sigue una exponencial.

```
ks.test(rnorm(100), rexp(80))
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  rnorm(100) and rexp(80)
## D = 0.6, p-value = 2e-13
## alternative hypothesis: two-sided
```

El resultado es que no, el test resulta significativo y rechazamos la hipótesis de igualdad, es decir que ambas variables tienen la misma distribución.

6.2.5. Test de normalidad de Jarque-Bera.

El test de Jarque-Bera no requiere estimaciones de los parámetros que caracterizan la normal.

```
#install.packages("tseries")
library(tseries)
x <- mtcars$mpg
jarque.bera.test(x)
```

```
##
##  Jarque Bera Test
##
## data:  x
## X-squared = 2, df = 2, p-value = 0.3
```

Cuando no podemos asumir la normalidad esto influye en los modelo en:

- Los estimadores mínimo-cuadráticos no son eficientes (de mínima varianza).

- Los intervalos de confianza de los parámetros del modelo y los contrastes de significación son solamente aproximados y no exactos.

En otras palabras: Se pierde precisión.
EL TCL, nos permite *'saltarnos'* esta hipótesis para muestras suficientemente grandes.

A menudo es interesante locaalizar las causas de la falta de normalidad de nuestras muestras.

Causas comunes que originan falta de normalidad:

- Existen observaciones heterogéneas .
- Errores en la recogida de datos.
- El modelo especificado no es correcto (por ejemplo, no se ha tenido en cuenta una variable de clasificación cuando las observaciones proceden de diferentes poblaciones: *existen factores que no se han tenido en cuenta*).
- Hay observaciones atípicas (Se puede recurrir a estimadores robustos).
- Existe asimetría en la distribución (Se puede recurrir a transformaciónes de los datos: familia de transformaciones de Box-Cox).

7. Contrastes de homogeneidad de varianza

El supuesto de homogeneidad de varianzas también se conoce como *supuesto de homocedasticidad*, nosotros habitualmente lo abreviamos como HOV, y dice que: *"la varianza es constante (no varía) en los diferentes niveles del factor"*. La falta de homocedasticidad se denomina heterocedasticidad (HEV).

Nota: Si el diseño es balanceado ($n_i = n_j$, para todo $i, j \in \{1, .., I\}$), la heterocedasticidad no afecta tanto a la calidad de los contrastes, a no ser que la varianza de la respuesta para algún grupo particular sea considerablemente mayor que para otros.

Reglilla: Balanceadas y HEV: ok si $\frac{\widehat{S}^2_{Max}}{\widehat{S}^2_{Min}} < 3$ *(Si no hay balanceo esta regla puede usarse con un 2)*

Si los tamaños muestrales son muy distintos se verifica que: si los grupos **con tamaños muestrales pequeños tienen mayor varianza** la probabilidad de cometer un error de tipo I en las pruebas de hipótesis será menor de lo que se obtiene y los niveles de confianza de los intervalos serán inferiores a lo que se cree (**conservador**).

Si por el contrario son los tratamientos con **tamaños muestrales grandes tienen mayor varianza** entonces se tendrá el efecto contrario y las pruebas serán más **liberales**.

Así pues es deseable que se verifique el supuesto de homocedasticidad.

7.0.1. Test de Levene

El test de Levene del paquete `car`, función `levene.test()`. El test de Levene se resuelve con un ANOVA (ya veremos lo que es) de los valores absolutos de las desviaciones de los valores muestrales respecto a un estadístico de centralidad (media, mediana o media truncada) para cada grupo. La elección del estadístico de centralidad de los grupos determina la robustez y la potencia del test. Por *robustez* se entiende la habilidad del test para no detectar erróneamente varianzas distintas, cuando la distribución no es normal y las varianzas son realmente iguales. La potencia significa la habilidad del test para señalar varianzas distintas, cuando efectivamente lo son. El test de Levene permite comparar más de dos grupos a la vez.

```
#install.packages("car");
library(car)

levene.test(df$q1,df$gender, center="mean")
  Levene's Test for Homogeneity of Variance (center = "mean")
        Df F value  Pr(>F)
  group  1      15 0.03047 *
         3
```

Con la mediana:

```
leveneTest(df$q1,df$gender, center="median")
  Levene's Test for Homogeneity of Variance (center = "median")
        Df F value Pr(>F)
  group  1     2.4 0.2191
         3
```

7.0.2. F test para la evaluación de homocedasticidad

También podemos empelar la función `var.test()` para hacer un *F test* y comparar las varianzas de dos poblaciones.

Ejemplo para dos variables:

```
var.test(df$q1,df$q2)
```

```
            F test to compare two variances

    data:  df$q1 and df$q2
    F = 0.7059, num df = 4, denom df = 4, p-value = 0.7439
    alternative hypothesis: true ratio of variances is not equal to 1
    95 percent confidence interval:
     0.07349473 6.77966815
    sample estimates:
    ratio of variances
             0.7058824
```

Ejemplo cuando tenemos un factor que me determina dos grupos en una variabe:

```
var.test(df$q1~df$gender)
```

```
            F test to compare two variances

    data:  df$q1 by df$gender
    F = 0.1667, num df = 2, denom df = 1, p-value = 0.2679
    alternative hypothesis: true ratio of variances is not equal to 1
    95 percent confidence interval:
     0.0002084636 6.4177215190
    sample estimates:
    ratio of variances
             0.1666667
```

7.0.3. Test de Bartlett para testar más de dos grupos.

Empleamos la función `bartlett.test()`para testar la homocedasticidad de más de dos muestras. El test de Levene es menos sensible a la falta de normalidad que el de Bartlett. Sin embargo, si estamos seguros de que los datos provienen de una distribución normal, entonces el test de Bartlett es el mejor.

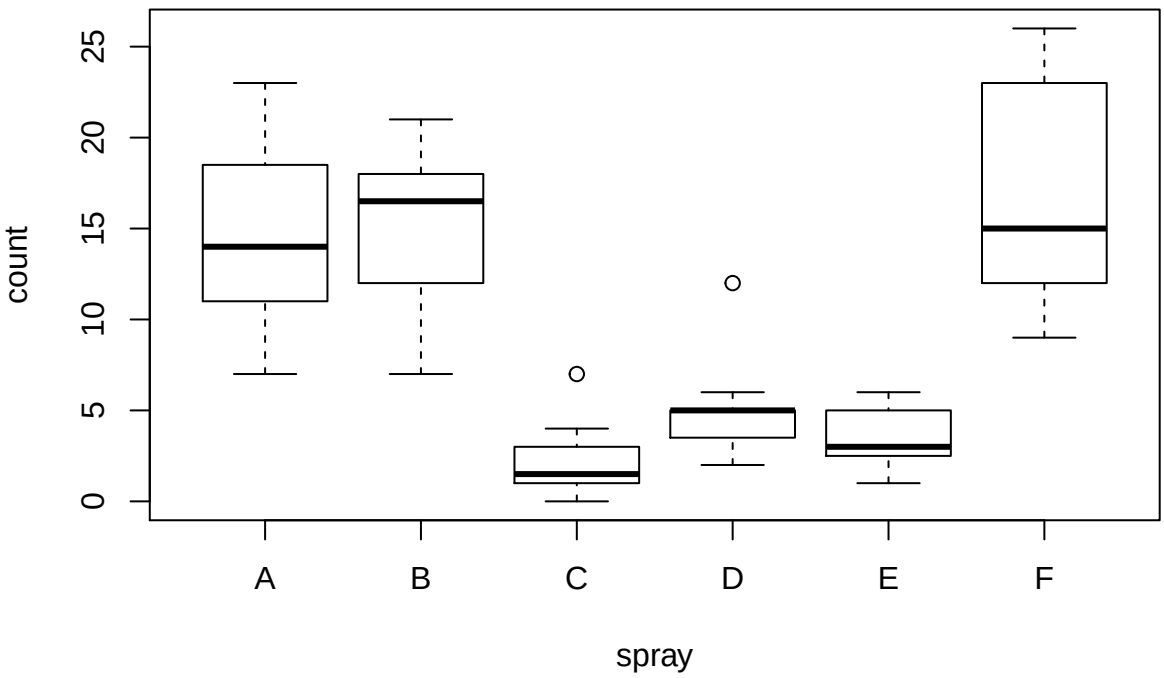
```
require(graphics)
str(InsectSprays)
```

```
## 'data.frame':    72 obs. of  2 variables:
##  $ count: num  10 7 20 14 14 12 10 23 17 20 ...
##  $ spray: Factor w/ 6 levels "A","B","C","D",..: 1 1 1 1 1 1 1 1 1 1 ...
```

```
head(InsectSprays)
```

```
##   count spray
## 1    10     A
## 2     7     A
## 3    20     A
## 4    14     A
## 5    14     A
## 6    12     A
```

```
plot(count ~ spray, data = InsectSprays)
```




```
fit<-bartlett.test(InsectSprays$count, InsectSprays$spray);fit
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  InsectSprays$count and InsectSprays$spray
## Bartlett's K-squared = 30, df = 5, p-value = 0.00009
```

y efectivamente se comprueba que no debemos aceptar la hipótesis de homogeneiadad de varianzas ya que $p < 0{,}05$

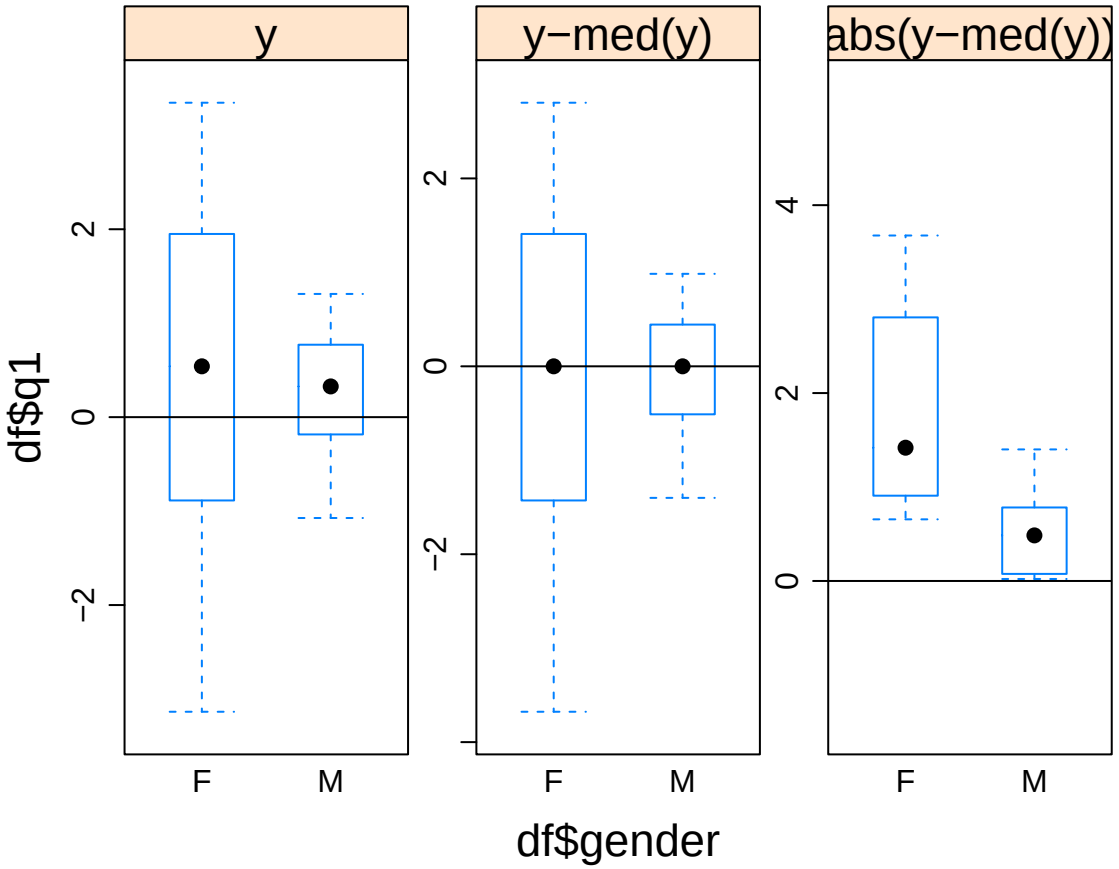
7.0.4. Test de Brown-Forsyth

La función `plot.hov()` (paquete `HH`), nos ofrece un test gráfico de HOV basado en Brown-Forsyth.

```
#install.packages("HH")
gender <- c(rep("M",10), rep("F",10))
q1 <- c(rnorm(10,0,0.5),rnorm(10,0,2))
df <- data.frame(gender, q1,  stringsAsFactors = TRUE)
library(HH)

hov(df$q1~df$gender)
```

```
##
##  hov: Brown-Forsyth
##
## data:  df$q1
## F = 10, df:df$gender = 1, df:Residuals = 20, p-value = 0.003
## alternative hypothesis: variances are not identical
```

```
# plot.hov.bf(df$q1~df$gender), con la versión 3 cambia el nombre de la función a hovPlot()
hovPlot(df$q1~df$gender)
```



7.0.5. Test de Fligner-Killeen

Podemos también comparar varianzas empleando un test no paramétrico con la función `Fligner.test()`que se basa en la mediana.

```
> fligner.test(InsectSprays$count, InsectSprays$spray)

        Fligner-Killeen test of homogeneity of variances

data:  InsectSprays$count and InsectSprays$spray
Fligner-Killeen:med chi-squared = 14.4828, df = 5, p-value = 0.01282
```

Notas generales sobre la elección del test: El artículo original de Levene proponía la media como estadístico de centralidad. Brown y Forsythe (Morton B. Brownab, 1974) extendieron este test al utilizar la mediana e incluso la media truncada al 10 %. Sus estudios de Monte Carlo mostraron que la utilización de la media truncada mejoraba el test cuando los datos seguían una distribución de Cauchy (colas grandes) y la mediana conseguía mejorarlo cuando los datos seguían una (distribución asimétrica). Con la media se consigue el mejor test para distribuciones simétricas y con colas moderadas. Así pues, aunque la elección óptima depende de la distribución de los datos, la definición del test basada en la mediana es la recomendación general, ya que, proporciona una buena robustez para la mayoría de distribuciones no normales y, al mismo tiempo, una aceptable potencia. Si conocemos la distribución de los datos, podemos optar por alguna otra de las opciones.

8. Transformación de datos

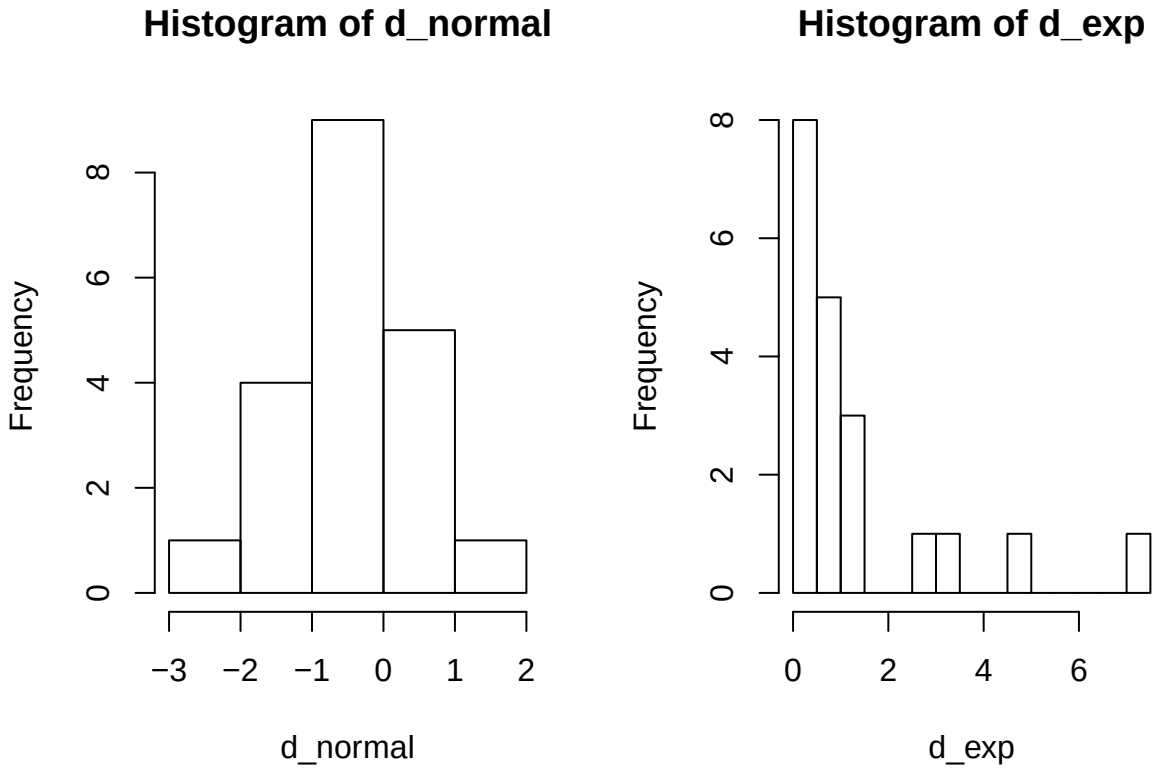
Vamos a comparar dos conjuntos de datos, analizar su normalidad y aplicarles transformaciones.

Generamos datos aleatorios con dos leyes diferentes.

```
set.seed( 111 )
d_normal <- rnorm( 20 )
d_exp    <- rexp(  20 )
```

Representamos las gráficas y vemos que son muy diferentes. La función `layout()` permite decidir cómo van a ser representadas las gráficas (en este caso, una al lado de la otra).

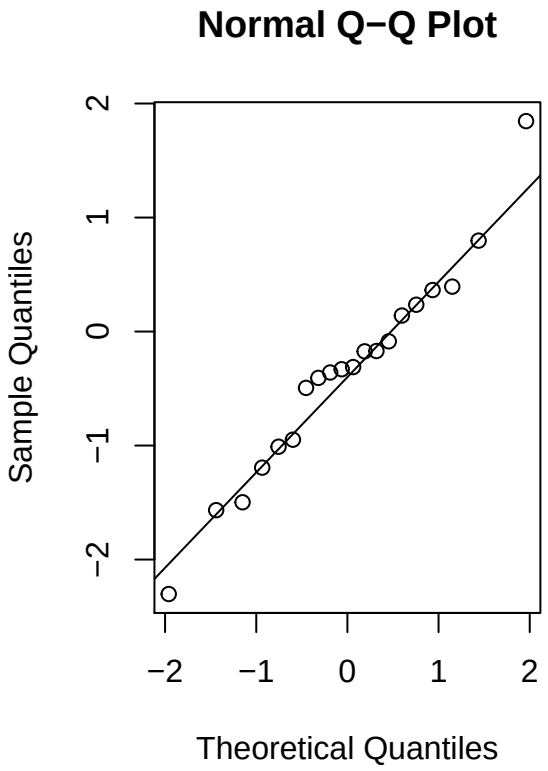
```
layout( matrix( c( 1, 2 ), nrow = 1) )
hist( d_normal, breaks = "FD" )
hist( d_exp,    breaks = "FD")
```

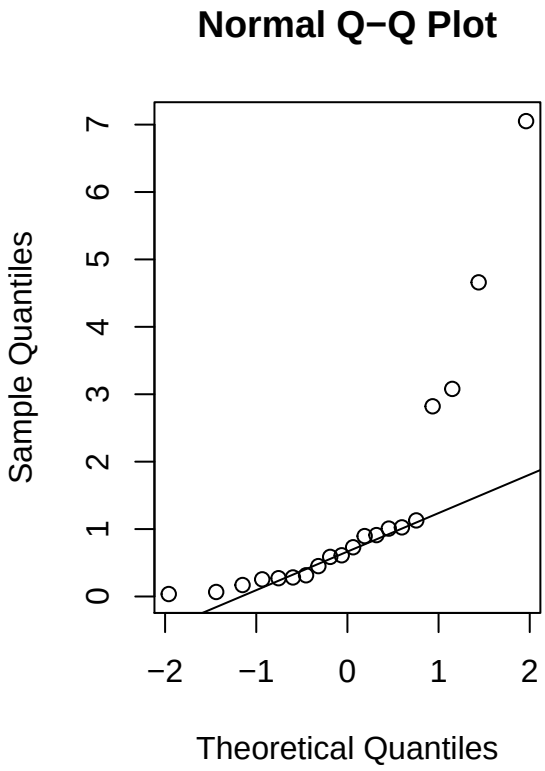


La función `hist()` selecciona de forma automática la anchura de cada marca/caja del histograma, podemos seleccionar diferentes algoritmos para determinar el número de bins y su anchura con la opción "`breaks`". Generalmente el algoritmo "FD" es el mejor en casi todos los casos `breaks="Sturges/"Scott/"FD"`.

Si ahora analizamos los Q-Q plots:

```
layout( matrix( c( 1, 2 ), nrow = 1) )
qqnorm( d_normal )
qqline( d_normal )
qqnorm( d_exp )
qqline( d_exp )
```





Podemos aplicar un **test de normalidad** para evaluarla.

```
shapiro.test( d_normal )
```

```
##
##  Shapiro-Wilk normality test
##
## data:  d_normal
## W = 1, p-value = 0.8
```

El p-valor > 0.05, es decir, se acepta la hipótesis nula de normalidad.

```
shapiro.test( d_exp )
```

```
##
##  Shapiro-Wilk normality test
##
## data:  d_exp
## W = 0.7, p-value = 0.00002
```

El p-valor < 0.05, es decir, se rechaza la hipótesis nula de normalidad.

Más generalmente podemos aplicar el test de koplmogorov-Smirnov obteniendo de nuevo los mismos resultados.

```
ks.test( d_normal, "pnorm", mean = mean( d_normal ), sd = sd( d_normal ) )
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  d_normal
## D = 0.1, p-value = 0.8
## alternative hypothesis: two-sided
```

```
ks.test( d_exp, "pnorm", mean = mean( d_exp ), sd = sd( d_exp ) )
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  d_exp
## D = 0.3, p-value = 0.01
## alternative hypothesis: two-sided
```

Este test `ks.test()` nos permite comprobar que nuestra distribución d-exp viene de una exponencial (ya lo sabíamos por cómo la hemos generado).

```
ks.test( d_exp, "pexp" )
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  d_exp
## D = 0.1, p-value = 0.8
## alternative hypothesis: two-sided
```

También nos permite testar si ambas siguen la misma distribución.

```
ks.test( d_normal, d_exp )
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  d_normal and d_exp
## D = 0.7, p-value = 0.00006
## alternative hypothesis: two-sided
```

Vemos que efectivamente **NO siguen la misma**, $p < 0.05$.

Imaginemos ahora que queremos contrastar esas poblaciones, ¿qué test empleamos?. Como para una de ellas no se cumple la hipótesis de normalidad, una opción es aplicar transformaciones a ver si logramos obtenerla. Hay un conjunto de transformaciones que son las usuales y las que mejores resultados proporcionan.

En todo caso después de la transformación podemos aplicar tests paramétricos si hemos alcanzado los supuestos, y cuando comuniquemos análisis descriptivos lo hacemos de los datos sin transformar aunque los test hayan sido llevados a cabo con los datos transformados.

Las transformaciones más usuales son el logaritmo y la raíz cuadrada. **Ojo** cuando transformemos si tenemos ceros o valores negativos, en ese caso habrá que trasladar las muestras.

Transformamos la primera aplicándole la función `log()`:

```
ld_exp <- log( d_exp )
shapiro.test( ld_exp )
```

```
##
##  Shapiro-Wilk normality test
##
## data:  ld_exp
## W = 1, p-value = 0.9
```

Y hacemos lo mismo con la segunda pero, además, le sumaremos uno para evitar hacer el logaritmo de cero:

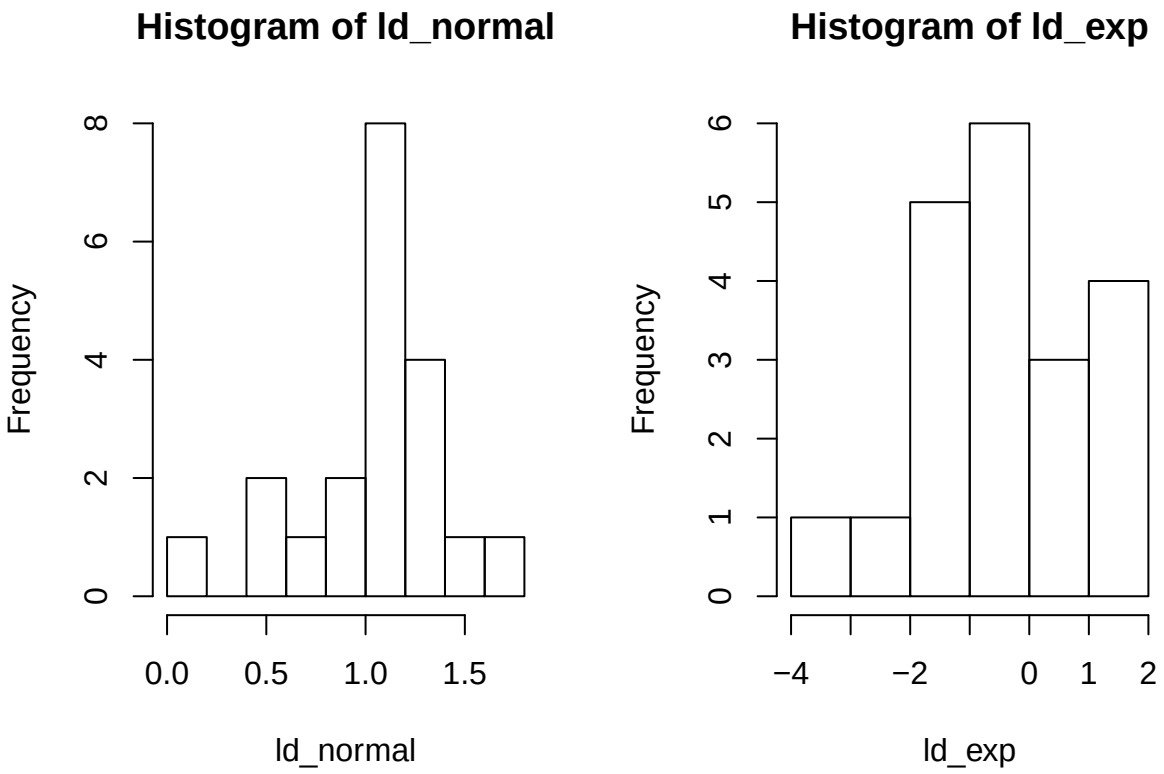
```
k         <- abs( min( d_normal ) ) + 1
ld_normal <- log( d_normal + k )
shapiro.test( ld_normal )
```

```
##
##  Shapiro-Wilk normality test
```

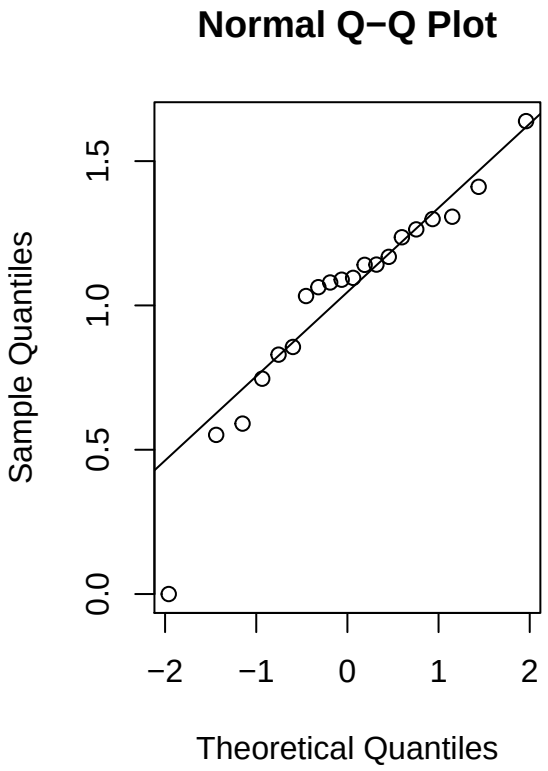
```
##
## data:  ld_normal
## W = 0.9, p-value = 0.08
```

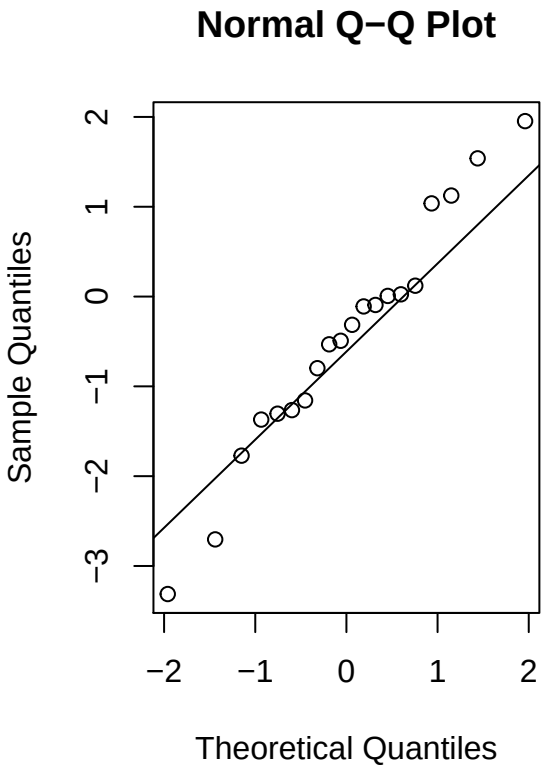
Una vez que se han transformado ambas se pueden comparar de nuevo:

```
layout( matrix( c( 1, 2 ), nrow = 1 ) )
hist( ld_normal, breaks = "FD" )
hist( ld_exp,    breaks = "FD" )
```




```
layout( matrix(c( 1, 2 ), nrow = 1 ) )
qqnorm( ld_normal )
qqline( ld_normal )
qqnorm( ld_exp )
qqline( ld_exp )
```





Volver al índice del curso

9. Referecias y bibliografía

Bland, J. M. (2000). *An introducction to medical statistics*. Oxford Medical Publications.

Braón-López, J. (2011). *Bioestadística. teorema del límite central - youtube*. Retrieved from https://www.youtube.com/watch?v=FcDcJnw00hk

Brown, S. (n.d.). Measures of shape: Skewness and kurtosis. *"Tompkins Cortland Community College, Math Course Pages by Stan Brown"*. Retrieved from http://www.tc3.edu/instruct/sbrown/stat/shape.htm

Bulmer, M. G. (1979). *Principles of statistics*. Dover Books on Mathematics.

Distribución normal- wikipedia, la enciclopedia libre. (n.d.). Retrieved May 9, 2012, from http://es.wikipedia.org/w/index.php?title=Distribuci%C3%B3n_normal&oldid=55624567

Krzywinski, M., & Altman, N. (2013). Points of significance: Importance of being uncertain, *10*(9), 809–810. Retrieved from http://dx.doi.org/10.1038/nmeth.2613

Morton B. Brownab, A. B. F. (1974). Robust tests for the equality of variances, *Volume 69, Issue 346, 1974*(96), 364–367. doi:10.1080/01621459.1974.10482955

Motulsky, H. (1995). *Intuitive biostatistics*. Oxford University Press, Inc.

Motulsky, H. (1999). *Analyzing data with graphpad prism*. GraphPad Software. Retrieved from http://

graphpad.com/manuals/analyzingdata.pdf

Teorema del límite central- wikipedia, la enciclopedia libre. (n.d.). Retrieved October 22, 2014, from http://es.wikipedia.org/wiki/Teorema_del_l%C3%ADmite_central

Verzani, J. (2002). *SimpleR – using r for introductory statistics* (4th ed.). Retrieved from http://www.math.csi.cuny.edu/Statistics/R/simpleR/