# Trabajo Práctico 3 - Inferencia Estadística

María Eugenia Fontecha

09 octubre, 2020

## Parte 1: Variables continuas

Se plantean 4 problemas en los cuales debe analizar tanto por métodos gráfico como analíticos cuál es el método de análisis más apropiado y justificar su respuesta. A continuación, debe realizar el test seleccionado, describir su hipótesis nula y alternativa, y sacar conclusiones sobre el resultado.

```
setwd(path)
```

1. **En una clínica pediátrica se llevó a cabo un estudio para ver la efectividad de la aspirina en la reducción de la temperatura. Se consideraron 12 niños de 5 años que sufrían gripe y se midió su temperatura antes y 1 hora después de tomar aspirina. Los datos se presentan en el archivo *aspirina.csv***

```
aspirina <- read.csv2('aspirina.csv')
summary(aspirina)
```

```
##      Antes            Desp.
##  Min.   :38.20   Min.   :37.60
##  1st Qu.:38.75   1st Qu.:37.80
##  Median :39.10   Median :37.90
##  Mean   :39.00   Mean   :38.03
##  3rd Qu.:39.33   3rd Qu.:38.33
##  Max.   :39.60   Max.   :38.50
```

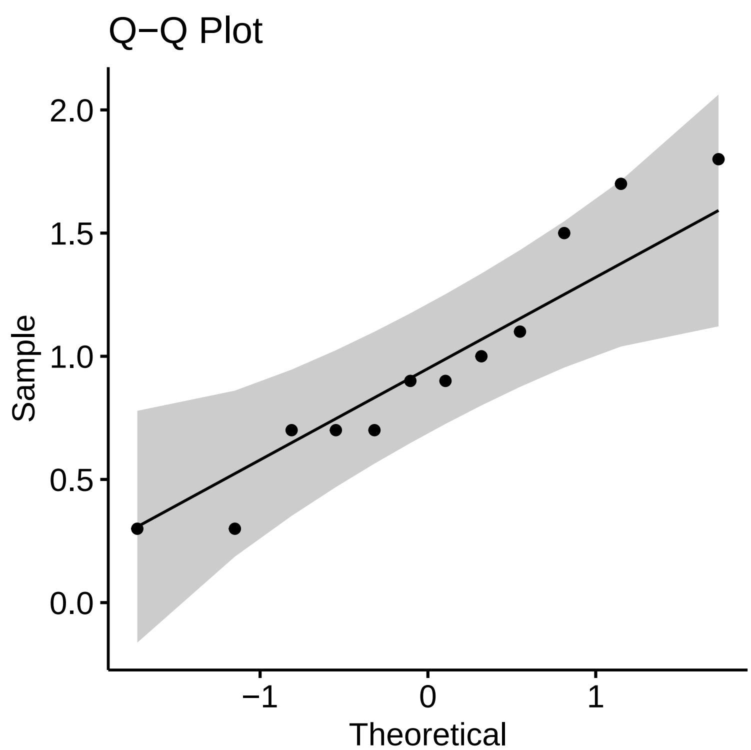
Se tienen dos muestras apareadas ya que se recogieron datos de temperatura de 12 niños antes y después de tomar aspirina. Se trata de un estudio longitudinal y nos interesa estudiar la diferencia $d$ entre las dos muestras.

Genero una variable $d$ que será la diferencia entre *Antes* y *Después* y analizo si tiene distribución normal.

```
aspirina$d = aspirina$Antes - aspirina$Desp.
print(shapiro.test(aspirina$d))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  aspirina$d
## W = 0.92618, p-value = 0.3414
```

```r
ggqqplot(aspirina$d, main = paste("Q-Q Plot"))
```




Para el test de Shapiro-Wilk obtuve un p-valor mayor a 0.05, por lo que no se rechaza la hipótesis nula que dice que la variable presenta distribución normal, y además el Q-Q Plot muestra también una distribución normal.

#### Test T de Student

Como tenemos 12 individuos, es decir menos de 30, se debe utilizar el test T.

Planteo las hipótesis del test, siendo $\mu$ la media de $d$:

- $H_0$: $\mu = 0$
- $H_1$: $\mu > 0$

Como se trata de un test a una cola, voy a rechazar la hipótesis nula si el P-valor es menor a 0.025.

```r
t.test(aspirina$d, mu = 0, alternative = 'greater')
```

```
##
##  One Sample t-test
##
## data:  aspirina$d
## t = 6.8268, df = 11, p-value = 1.425e-05
## alternative hypothesis: true mean is greater than 0
## 95 percent confidence interval:
```

```
##  0.7123697       Inf
## sample estimates:
## mean of x
## 0.9666667
```

Se comprueba que la aspirina tiene un efecto de reducción de la temperatura en 0.9666667 °C en niños de 5 años con gripe, con un intervalo de confianza del 95% entre 0.7123697 e infinito, es decir que en el 95% de la población, el descenso de la temperatura será de al menos 0,71234697 °C. Este descenso resultó estadísticamente significativo (P-Valor menor a 0.025).

2. **Se realizó un estudio con 10 pacientes que padecen insuficiencia renal provocada por diabetes y fueron tratados con captopril durante un periodo de 8 semanas. La proteína urinaria fue medida antes y después de aplicar la droga. Los datos se encuentran en el archivo *proteína.csv*.**

```r
proteina <- read.csv2('proteina.csv')
summary(proteina)
```

```
##      antes          despues
##  Min.   : 4.70   Min.   : 0.700
##  1st Qu.: 5.50   1st Qu.: 3.025
##  Median : 8.05   Median : 4.900
##  Mean   :10.61   Mean   : 5.270
##  3rd Qu.:14.60   3rd Qu.: 6.000
##  Max.   :25.60   Max.   :15.600
```

Se trata de un estudio longitudinal, donde se midió la proteína urinaria antes y después de aplicar la droga, por lo que las muestras están apareadas. Calculo la diferencia de ambas muestras.

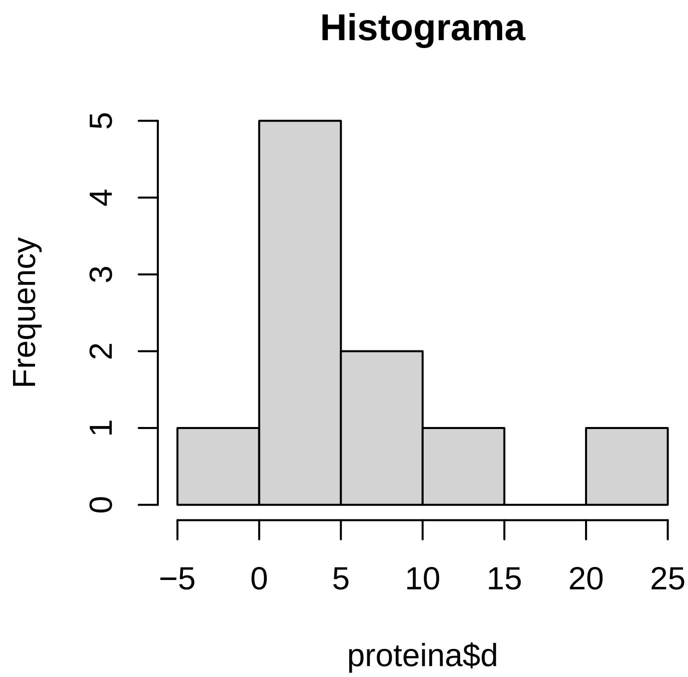
```r
proteina$d <- proteina$antes - proteina$despues
```

Analizo normalidad de la nueva variable.

```r
print(shapiro.test(proteina$d))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  proteina$d
## W = 0.81634, p-value = 0.02289
```

```r
hist(proteina$d, main = 'Histograma')
```




Con el test de Shapiro-Wilk se obtiene un P-Valor menor a 0.05, por lo que se rechaza la hipótesis nula y se concluye que la variable $d$ no presenta distribución normal. También se ve en el método gráfico que el histograma no presenta la forma de campana característica de una distribución normal. Por lo tanto, la variable $d$ no es normal.

### Test de Wilcoxon

Dado que la variable bajo estudio no presenta distribución normal, utilizo el Test de Wilcoxon. Siendo $\theta$ la mediana, las hipótesis del test son:

- $H_0$: $\theta_{antes} - \theta_{desp} = 0$
- $H_1$: $\theta_{antes} - \theta_{desp} \neq 0$

```
wilcox.test(proteina$antes,proteina$despues, paired = TRUE, conf.int = TRUE)
```

```
##
##  Wilcoxon signed rank exact test
##
## data:  proteina$antes and proteina$despues
## V = 54, p-value = 0.003906
## alternative hypothesis: true location shift is not equal to 0
## 95 percent confidence interval:
##   1.05 10.60
## sample estimates:
## (pseudo)median
##           4.35
```

Se comprueba que el captopril tiene un efecto sobre la concentración de la proteína urinaria, produciendo la disminución de dicha concentración con una mediana de 4.35 mg/ml y un intervalo de confianza al 95% de 1.05 a 10.60. Es decir, que en el 95% de la población de pacientes con insuficiencia renal provocada por diabetes, el captopril producirá una disminución de entre 1.05 y 10.60 mg/ml en la concentración de proteína en orina.

3. **El codo de tenista afecta a los jugadores de tenis frecuentemente. Se condujo un estudio comparando la efectividad del ibuprofeno por 3 semanas y el placebo. La mitad de los participantes recibió un placebo. Se midió el grado de dolor en una escala del 1-6. Los datos se encuentran en el archivo *dolor.csv*.**

```
dolor <- read.csv2('dolor.csv')
summary(dolor)
```

```
##      Grupo      Grado.de.dolor
##  Min.   :1.0   Min.   :1.000
##  1st Qu.:1.0   1st Qu.:2.000
##  Median :1.5   Median :3.000
##  Mean   :1.5   Mean   :3.409
##  3rd Qu.:2.0   3rd Qu.:5.000
##  Max.   :2.0   Max.   :9.000
```

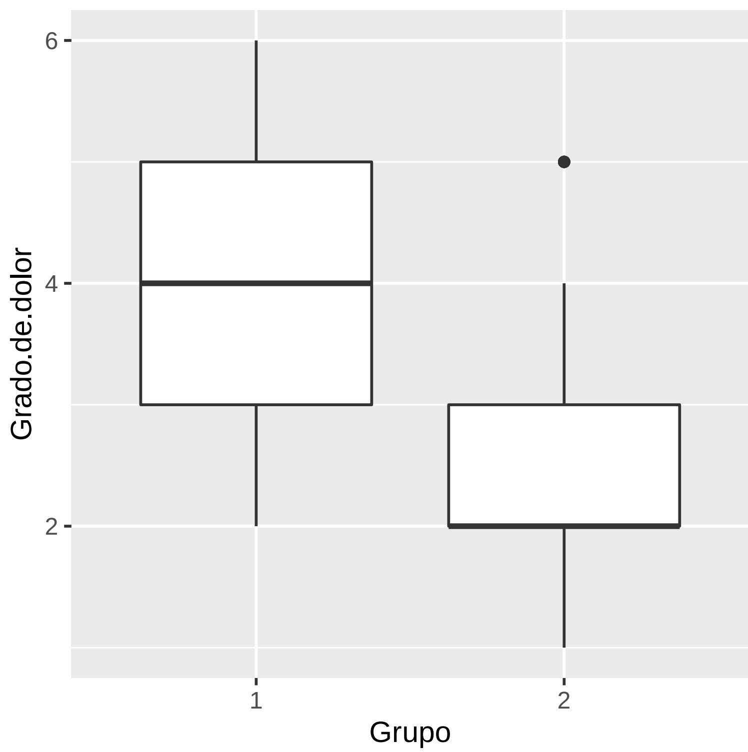
La variable Grupo indica si al individuo se le administró el placebo o el ibuprofeno, por lo que es una variable categórica. Modifico el tipo de variable a *factor*. Además, elimino los registros que tengan grado de dolor fuera de la escala del 1-6.

```
dolor$Grupo <- as.factor(dolor$Grupo)

dolor <- dolor[dolor$Grado.de.dolor >= 1 & dolor$Grado.de.dolor <= 6,]
summary(dolor)
```

```
##  Grupo  Grado.de.dolor
##  1:43   Min.   :1.000
##  2:44   1st Qu.:2.000
##         Median :3.000
##         Mean   :3.345
##         3rd Qu.:5.000
##         Max.   :6.000
```

```
plt <- ggplot(dolor, aes(x=Grupo ,y=Grado.de.dolor)) +
geom_boxplot()
plt
```




A partir del boxplot voy a suponer que el Grupo 1 es el grupo al que se le administró un placebo y el Grupo 2 es el grupo al que se le administró el ibuprofeno, ya que no se especifica.

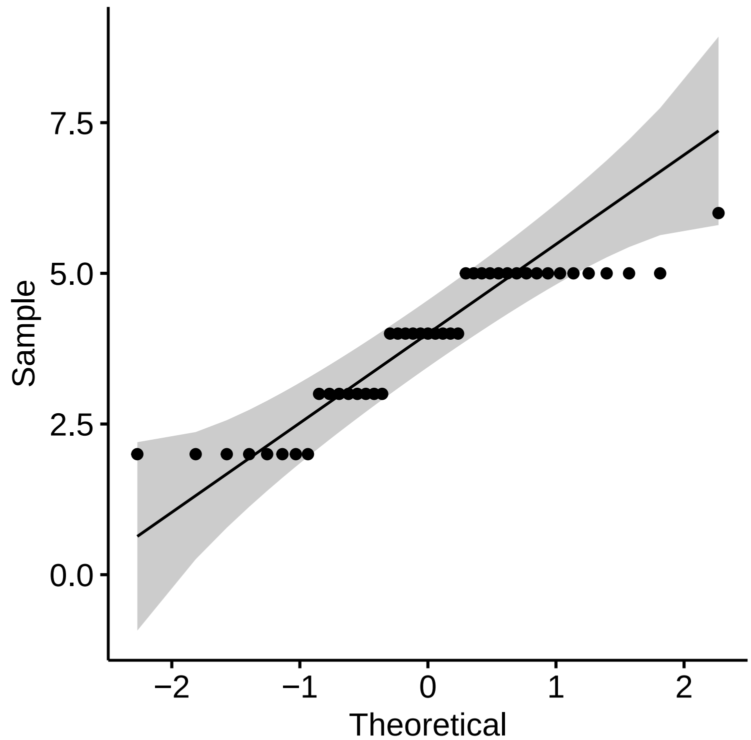
Analizo normalidad para ambos grupos.

```
dolor1 = dolor$Grado.de.dolor[dolor$Grupo == 1]

lillie.test(dolor1)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  dolor1
## D = 0.22685, p-value = 7.087e-06
```

```
ggqqplot(dolor1)
```

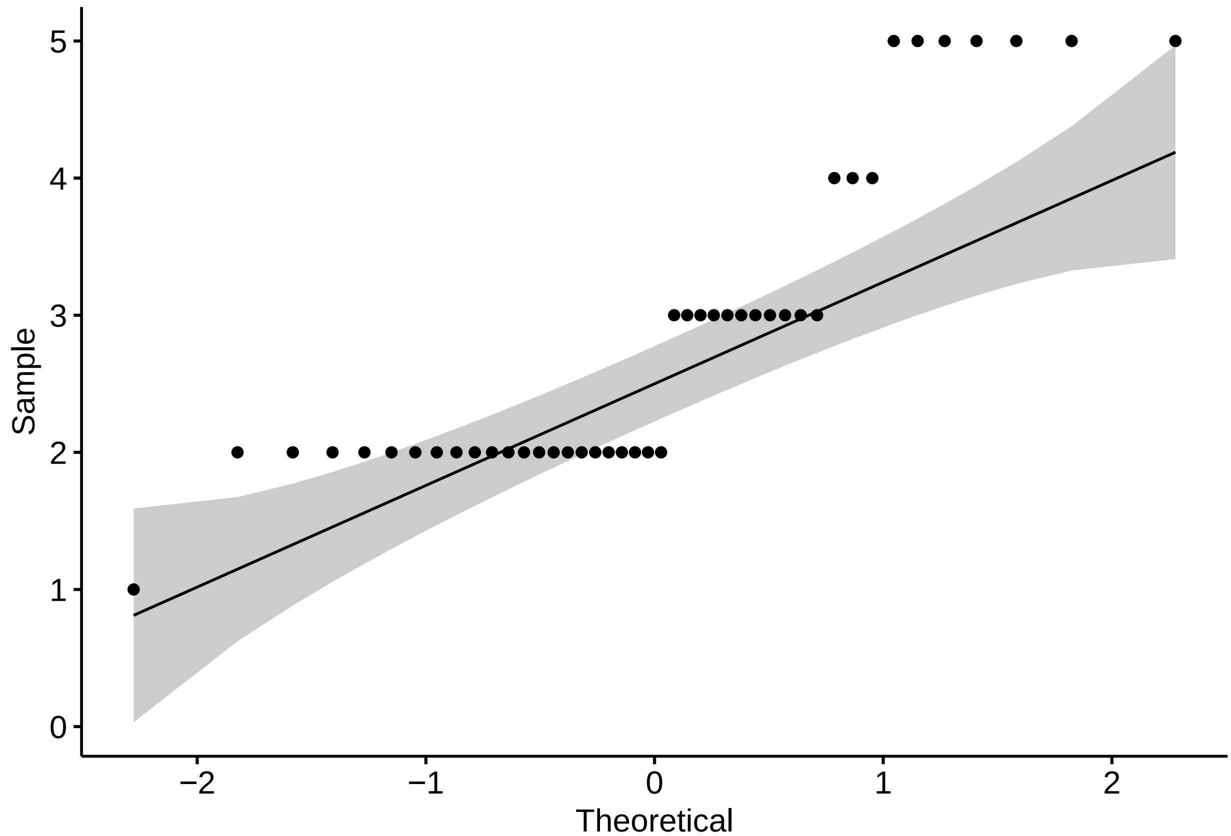



```r
dolor2 = dolor$Grado.de.dolor[dolor$Grupo == 2]

lillie.test(dolor2)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  dolor2
## D = 0.29235, p-value = 2.832e-10
```

```r
ggqqplot(dolor2)
```




Para los dos grupos se puede concluir que la variable Grado de dolor no tiene distribución normal. En los gráficos de QQ Plot se puede ver que hay datos que se alejan de la recta de normalidad y además los tests de Lilliefors arrojaron un P-Valor menor a 0.05.

**Test de Wilcoxon-Mann-Whitney**

Las hipótesis del test son:

- $H_0$: $\theta_{placebo} - \theta_{ibu} = 0$
- $H_1$: $\theta_{placebo} - \theta_{ibu} \neq 0$

```
wilcox.test(dolor1, dolor2, paired = FALSE, exact = FALSE)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  dolor1 and dolor2
## W = 1371, p-value = 0.0001816
## alternative hypothesis: true location shift is not equal to 0
```

```
print('placebo')
```

```
## [1] "placebo"
```

```r
summary(dolor1)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.00    3.00    4.00    3.86    5.00    6.00
```

```r
print('ibuprofeno')
```

```
## [1] "ibuprofeno"
```

```r
summary(dolor2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   2.000   2.841   3.000   5.000
```

El ibuprofeno tiene un efecto en el grado de dolor causado por el codo de tenista, con una mediana de 4 en la escala de dolor del grupo placebo y un rango intercuatílico de 2 y una mediana de 2 y un rango intercuartílico de 1 en el grupo al que se le suministró ibuprofeno. Por lo tanto, se concluye que el ibuprofeno disminuye el dolor causado por el codo de tenista. Los resultados tienen significancia estadísica ya que el P-Valor es menor a 0.05.

4. **En algunos estudios se observó una asociación entre consumo de beta-caroteno y un decrecimiento en la incidencia de cáncer. En un ensayo clínico se compararon 2 preparados de beta-caroteno en cápsulas a fin de elevar los niveles plasmáticos de esta vitamina. Los sujetos fueron asignados en forma aleatoria a ambos grupos. Los datos se encuentran en *cancer.csv*.**

```r
cancer <- read.csv2('cancer.csv')
summary(cancer)
```

```
##   Betacaroteno     Preparado
##  Min.   :104.0   Min.   :1.000
##  1st Qu.:218.0   1st Qu.:1.000
##  Median :332.0   Median :1.000
##  Mean   :312.1   Mean   :1.478
##  3rd Qu.:381.0   3rd Qu.:2.000
##  Max.   :566.0   Max.   :2.000
```

Modifico la variable Preparado para que sea de tipo factor.

```r
cancer$Preparado = as.factor(cancer$Preparado)
summary(cancer)
```

```
##   Betacaroteno   Preparado
##  Min.   :104.0   1:12
##  1st Qu.:218.0   2:11
##  Median :332.0
##  Mean   :312.1
##  3rd Qu.:381.0
##  Max.   :566.0
```

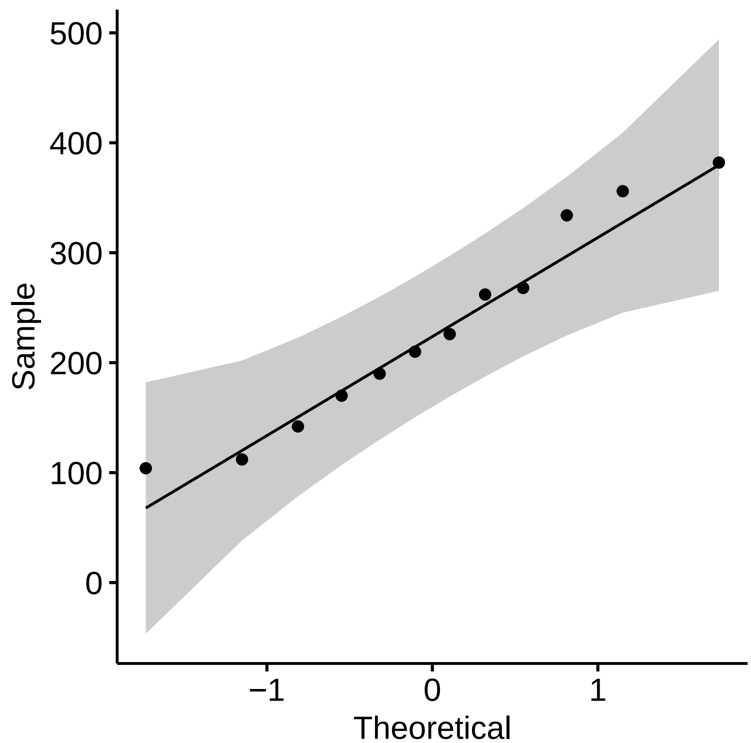
Analizo normalidad para cada grupo.

```r
prep_1 <- cancer$Betacaroteno[cancer$Preparado == 1]

shapiro.test(prep_1)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  prep_1
## W = 0.94995, p-value = 0.6362
```

```r
ggqqplot(prep_1)
```

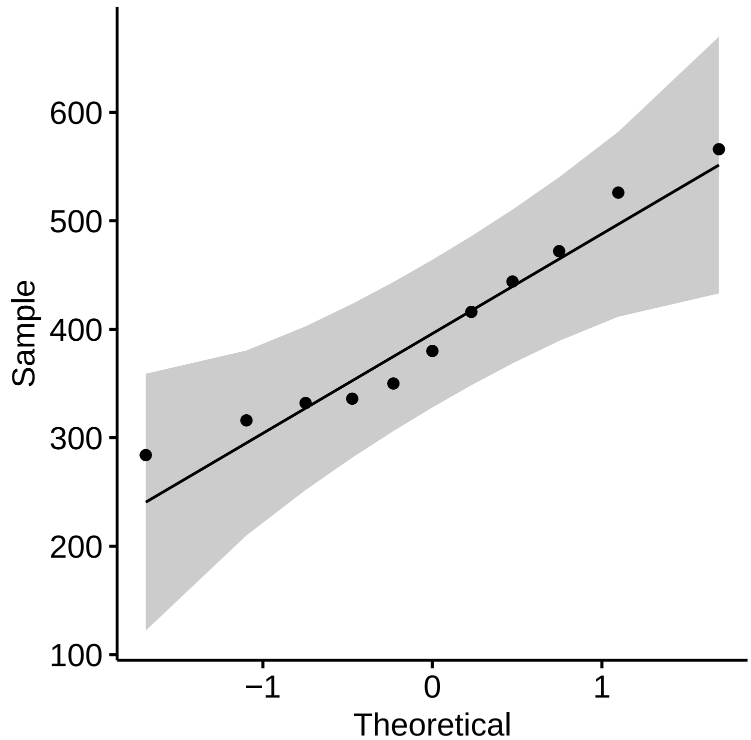



```r
prep_2 <- cancer$Betacaroteno[cancer$Preparado == 2]

shapiro.test(prep_2)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  prep_2
## W = 0.94079, p-value = 0.5299
```

```r
ggqqplot(prep_2)
```




Para ambos preparados se cumple que la variable Beta-Caroteno presenta distribución normal, como se pude verificar con el test de Shapiro, que dio P-Valores menores a 0.05, y con el QQ Plot donde todos los datos presentaban una tendencia normal.

Analizo homocedasticidad para poder realizar el test T.

```
leveneTest(y = cancer$Betacaroteno, group = cancer$Preparado)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  1  0.0061 0.9384
##       21
```

La hipótesis nula del test de Levene dice que las varianzas de los dos grupos son homogéneas y la hipótesis alternativa dice que las varianzas de los dos grupos son heterogéneas. Como se obtuvo un P-valor de 0.9384 se acepta la hipótesis nula, por lo que se cumple la condición de homocedasticidad del test T.

**Test T para muestras independientes**

Como se cumple normalidad de las variables y homocedasticidad y dado que la cantidad de muestras es pequeña, se realiza el test T.

Las hipótesis del test T son:

- $H_0$: $\mu_{prep\_1} - \mu_{prep\_2} = 0$
- $H_1$: $\mu_{prep\_1} - \mu_{prep\_2} \neq 0$.

```r
t.test(prep_1, prep_2, paired = FALSE, int.conf =TRUE)
```

```
##
##  Welch Two Sample t-test
##
## data:  prep_1 and prep_2
## t = -4.4849, df = 20.902, p-value = 0.0002059
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -252.26582  -92.40084
## sample estimates:
## mean of x mean of y
##  229.6667  402.0000
```

```r
print('preparado 1')
```

```
## [1] "preparado 1"
```

```r
CI(prep_1, ci = 0.95)
```

```
##    upper     mean    lower
## 288.8441 229.6667 170.4893
```

```r
print('preparado 2')
```

```
## [1] "preparado 2"
```

```r
CI(prep_2, ci = 0.95)
```

```
##    upper     mean    lower
## 463.1664 402.0000 340.8336
```

El nivel de beta-carotenos plasmático es diferente entre los dos tipos de preparados con alta significancia estadística (P-valor menor a 0.05), siendo mayor en el caso del preparado 2. El preparado 1 tiene una media de 229.667 con un intervalo de confianza al 95% de 170.4893 a 288.8441, mientras que el preparado 2 tiene una media del nivel de beta-carotenos de 402.000 y un intervalo de confianza al 95% de 340.8336 a 463.1664. Se observa que no existe solapamiento entre los intervalos de confianza del nivel de beta-carotenos de ambos grupos, lo que refuerza la conclusión sacada en cuanto a la diferencia de niveles de beta-carotenos que se generan con los distintos preparados.