

Introducción

Contrastar si unos datos provienen de una distribución normal o no es esencial. Y es esencial para poder utilizar unos análisis y estadísticos u otros. Para **contrastar la normalidad**, es necesario aplicar un test estadístico a cada población o muestra por separado. Veamos cuáles son los test más utilizados:

- **Test de Shapiro-Wilk**. Se aplica principalmente a una población / muestra con un tamaño muestral comprendido entre 3 y 50.
- **Test de Kolmogorov-Smirnov** (con corrección **Lilliefors**). Se aplica principalmente a una población / muestra con un tamaño muestral situado por encima de 50.

Ejemplo de medida de bifaces coreanos

Vamos a comenzar este anexo con un ejemplo sobre medidas de bifaces encontrados en tres yacimientos de Corea (Norton et al., 2006). En ese trabajo se presentan las medidas de altura, anchura y grosor de bifaces encontrados en la península de Corea, provenientes de 4 yacimientos. En los ejemplos que realizaremos aquí utilizaremos 3 de ellos: Chongokni, Chuwoli/Kawoli y Kumpari (Figura 1).

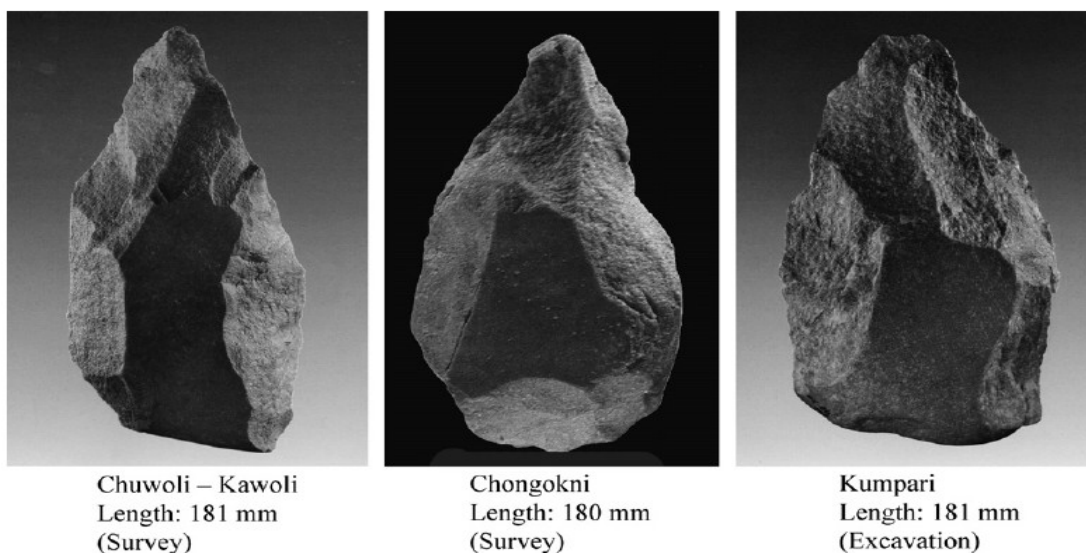
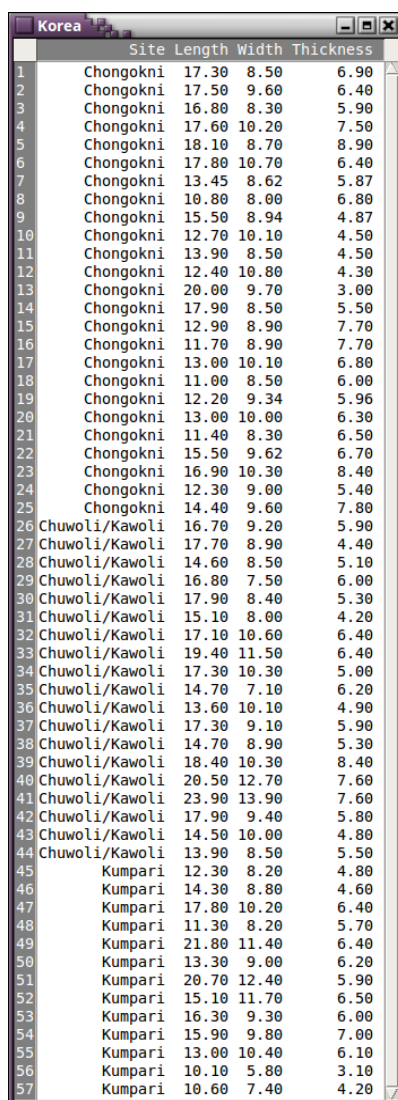


Figura 1: Bifaces provenientes de la Península de Corea y de los tres yacimientos que utilizaremos en el presente módulo. Imagen obtenida del artículo (Norton et al., 2006).

Los datos con los que trabajaremos en R Commander los podréis descargar del Campus Virtual, con el nombre de *Norton et al., 2006. Bifaces Korea. Base de datos general.txt*. Aunque he de comunicaros que, dependiendo del test que realicemos y sobre qué elementos sean ejecutados, habrá que reestructurar dicha base de datos para

adaptarla a las necesidades de R Commander. En cualquier caso, os lo comentaremos cuando llegue el momento.

Los datos de ese archivo de texto tienen la siguiente estructura (Figura 2): 4 columnas (con Yacimiento, Altura, Anchura y Grosor de los bifaces, en centímetros), y 57 filas que se corresponden con los 57 bifaces.



	Site	Length	Width	Thickness
1	Chongokni	17.30	8.50	6.90
2	Chongokni	17.50	9.60	6.40
3	Chongokni	16.80	8.30	5.90
4	Chongokni	17.60	10.20	7.50
5	Chongokni	18.10	8.70	8.90
6	Chongokni	17.80	10.70	6.40
7	Chongokni	13.45	8.62	5.87
8	Chongokni	10.80	8.00	6.80
9	Chongokni	15.50	8.94	4.87
10	Chongokni	12.70	10.10	4.50
11	Chongokni	13.90	8.50	4.50
12	Chongokni	12.40	10.80	4.30
13	Chongokni	20.00	9.70	3.00
14	Chongokni	17.90	8.50	5.50
15	Chongokni	12.90	8.90	7.70
16	Chongokni	11.70	8.90	7.70
17	Chongokni	13.00	10.10	6.80
18	Chongokni	11.00	8.50	6.00
19	Chongokni	12.20	9.34	5.96
20	Chongokni	13.00	10.00	6.30
21	Chongokni	11.40	8.30	6.50
22	Chongokni	15.50	9.62	6.70
23	Chongokni	16.90	10.30	8.40
24	Chongokni	12.30	9.00	5.40
25	Chongokni	14.40	9.60	7.80
26	Chuwoli/Kawoli	16.70	9.20	5.90
27	Chuwoli/Kawoli	17.70	8.90	4.40
28	Chuwoli/Kawoli	14.60	8.50	5.10
29	Chuwoli/Kawoli	16.80	7.50	6.00
30	Chuwoli/Kawoli	17.90	8.40	5.30
31	Chuwoli/Kawoli	15.10	8.00	4.20
32	Chuwoli/Kawoli	17.10	10.60	6.40
33	Chuwoli/Kawoli	19.40	11.50	6.40
34	Chuwoli/Kawoli	17.30	10.30	5.00
35	Chuwoli/Kawoli	14.70	7.10	6.20
36	Chuwoli/Kawoli	13.60	10.10	4.90
37	Chuwoli/Kawoli	17.30	9.10	5.90
38	Chuwoli/Kawoli	14.70	8.90	5.30
39	Chuwoli/Kawoli	18.40	10.30	8.40
40	Chuwoli/Kawoli	20.50	12.70	7.60
41	Chuwoli/Kawoli	23.90	13.90	7.60
42	Chuwoli/Kawoli	17.90	9.40	5.80
43	Chuwoli/Kawoli	14.50	10.00	4.80
44	Chuwoli/Kawoli	13.90	8.50	5.50
45	Kumpari	12.30	8.20	4.80
46	Kumpari	14.30	8.80	4.60
47	Kumpari	17.80	10.20	6.40
48	Kumpari	11.30	8.20	5.70
49	Kumpari	21.80	11.40	6.40
50	Kumpari	13.30	9.00	6.20
51	Kumpari	20.70	12.40	5.90
52	Kumpari	15.10	11.70	6.50
53	Kumpari	16.30	9.30	6.00
54	Kumpari	15.90	9.80	7.00
55	Kumpari	13.00	10.40	6.10
56	Kumpari	10.10	5.80	3.10
57	Kumpari	10.60	7.40	4.20

Figura 2: Datos generales de las medidas de bifaces coreanos.

Test de Shapiro-Wilk

El **test de Shapiro-Wilk** se usa para contrastar si un conjunto de datos **sigue una distribución normal o no**. Recordad la distribución normal en el Módulo 3. Este hecho es de vital importancia porque otros muchos análisis estadísticos requieren de la normalidad de los datos para poder llevarlos a cabo. Se suele utilizar en muestras cuyo tamaño está

comprendido entre 3 y 50 observaciones. Veamos cuál es el contraste de hipótesis específico del test de Shapiro-Wilk.

Contraste de hipótesis del test de Shapiro-Wilk

- H_0 : los datos **proviene**n de una distribución normal
- H_1 : los datos **no proviene**n de una distribución normal

Estructura de los datos en R Commander

Para hacer un test de Shapiro-Wilk es necesario que aparezca en cada columna los datos de la variable que queremos contrastar si se distribuyen siguiendo una normal.

Por ejemplo, en la Figura 2 se podrían contrastar la normalidad de las variables *Length*, *Width* y *Thickness* de todos los yacimientos en conjunto. Sin embargo, si queremos testear la normalidad de la variable *Length* del yacimiento Kumpari, esa estructura no sería viable. Tendríamos que reestructurar los datos para que solo aparecieran en una columna los datos de *Length* de Kumpari. Veámoslo a continuación.

R Commander: test de Shapiro Wilk

Importamos los datos en R Commander (se encuentran en el Campus Virtual, con el nombre *Norton et al., 2006. Bifaces Korea. Base de datos general.txt*). Le damos el nombre de *Korea*. Vamos a realizar primero el test de Shapiro-Wilk sin distinguir entre yacimientos para sus tres variables, y posteriormente teniendo en cuenta los yacimientos.

Pero antes de realizar los análisis, creo que es conveniente realizar histogramas para evaluar *a priori* si nuestras muestras (tanto en conjunto como separadas por yacimiento) se podrían comportar como una normal o no (Figura 3). Revisar el Módulo 2 para realizar histogramas.

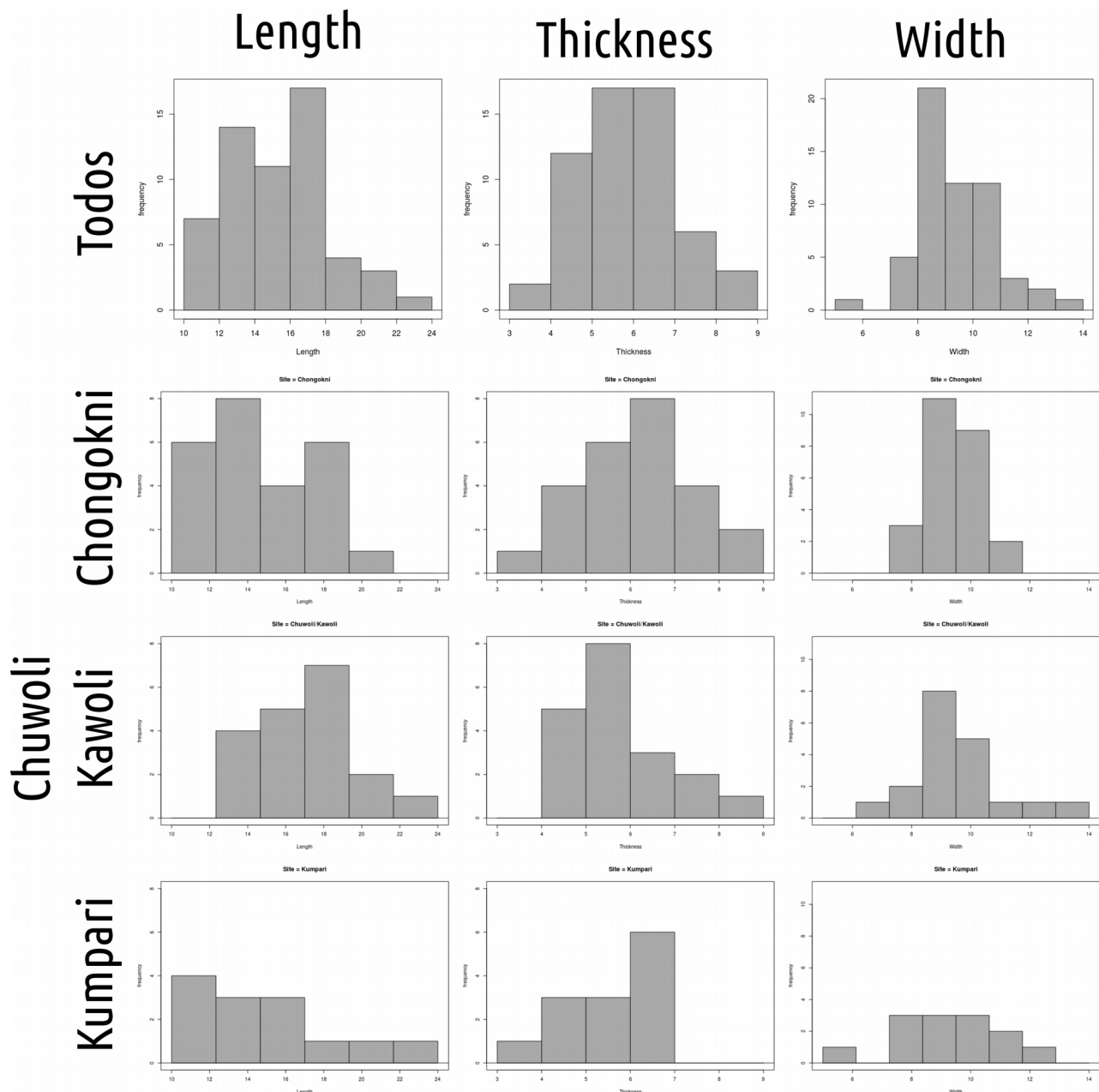


Figura 3: Histogramas de frecuencias de las medidas de bifaces coreanos (Length, Thickness y Width), en conjunto (Todos) y separados por yacimiento.

De *visu* podemos observar que prácticamente la totalidad de los histogramas de la Figura 3 podrían asemejarse a una campana, morfología propia de una distribución normal. Sin embargo, los histogramas para *Length* y *Thickness* del yacimiento de Kumpari no parece que se vea claramente esa campana. Comprobemos con los test de normalidad cuáles son los que siguen una distribución normal y cuáles no.

Test con los yacimientos juntos

En este caso trabajamos con los datos tal y como están organizados y estructurados en la Figura 2, en la que los datos de las 3 variables para todos los bifaces están escritos de

forma continua en cada columna. En R Commander se realiza siguiendo la siguiente ruta (Figura 4):

Statistics ► Summaries ► Shapiro-Wilk test of normality...

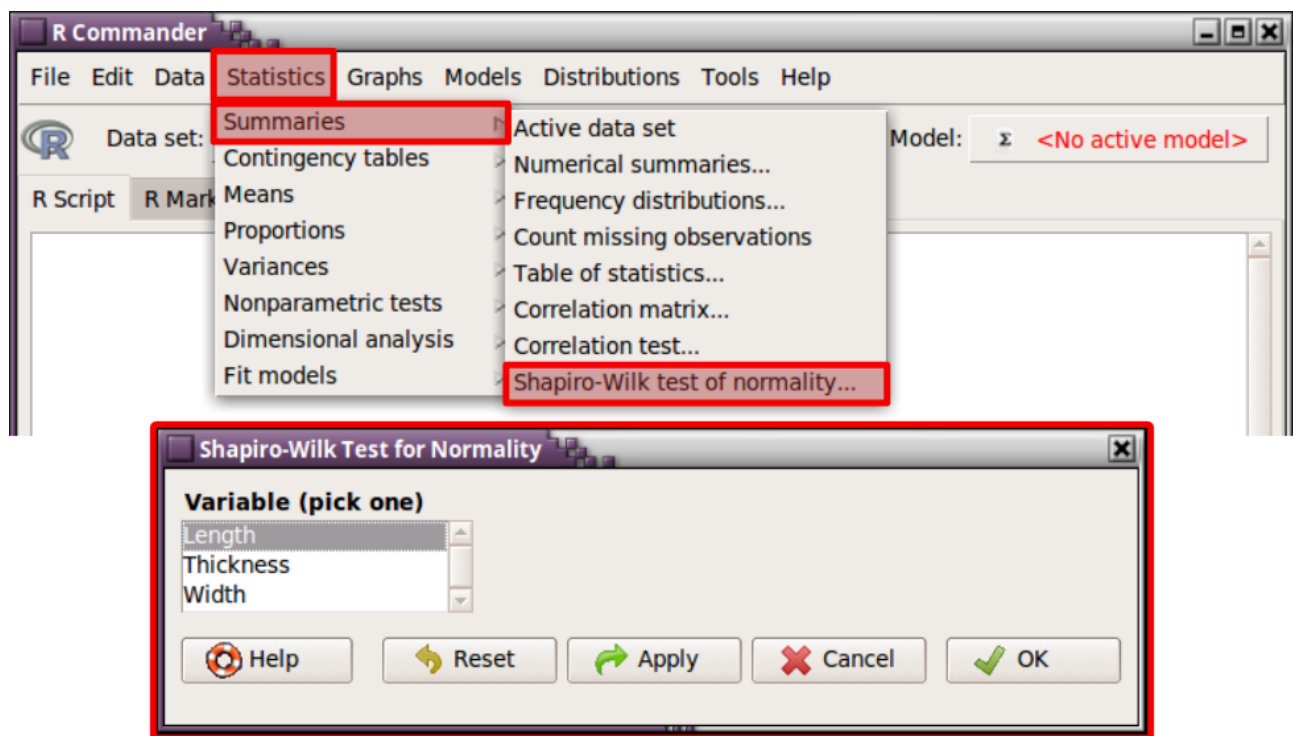


Figura 4: Ruta en R Commander para realizar un test de normalidad Shapiro-Wilk. En la ventana que se abre, hay que seleccionar una sola variable.

Los resultados que nos da la ejecución del test de Shapiro-Wilk para las tres variables son los siguientes:

```
> shapiro.test(Korea$Length)
Shapiro-Wilk normality test
data:  Korea$Length
W = 0.97326, p-value = 0.2371

> shapiro.test(Korea$Thickness)
Shapiro-Wilk normality test
data:  Korea$Thickness
W = 0.98428, p-value = 0.6648

> shapiro.test(Korea$Width)
Shapiro-Wilk normality test
data:  Korea$Width
W = 0.96011, p-value = 0.05762
```

Los resultados indican que **las tres variables** (*Length*, *Thickness* y *Width*) **se comportan siguiendo una distribución normal**, ya que sus p-valores se encuentran situados por encima de 0.05. Esto nos lleva a aceptar sus hipótesis nulas (H_0) indicando que se distribuyen normalmente, tal y como aparece resumido en la Tabla 1.

Variable	p-valor	Conclusión
Length	0.2371	Aceptamos H_0
Thickness	0.6648	Aceptamos H_0
Width	0.0576	Aceptamos H_0

Tabla 1: P-valores de las tres variables del estudio de los bifaces coreanos sin distinción por yacimiento usando el test de Shapiro-Wilk.

Test con los yacimientos separados

Los datos tienen que reestructurarse para poder analizar su normalidad por separado. Os lo podéis descargar del Campus Virtual, presentando el nombre de *Norton et al., 2006. Bifaces Korea. Base de datos general separada por yacimiento.txt*. Es importante destacar que las celdas vacías deben tener el nombre de **NA**. De otro modo, dará un error al importar el conjunto de datos. El conjunto de datos importado en R Commander lo hemos llamado **KoreanSites** (Figura 5).

	Length.Chongokni	Width.Chongokni	Thickness.Chongokni	Length.Chuwoli.Kawoli	Width.Chuwoli.Kawoli	Thickness.Chuwoli.Kawoli	Length.Kumpari	Width.Kumpari	Thickness.Kumpari
1	17.30	8.50	6.90	16.7	9.2	5.9	12.3	8.2	4.8
2	17.50	9.60	6.40	17.7	8.9	4.4	14.3	8.8	4.6
3	16.80	8.30	5.90	14.6	8.5	5.1	17.8	10.2	6.4
4	17.60	10.20	7.50	16.8	7.5	6.0	11.3	8.2	5.7
5	18.10	8.70	8.90	17.9	8.4	5.3	21.8	11.4	6.4
6	17.80	10.70	6.40	15.1	8.0	4.2	13.3	9.0	6.2
7	13.45	8.62	5.87	17.1	10.6	6.4	20.7	12.4	5.9
8	10.80	8.00	6.80	19.4	11.5	6.4	15.1	11.7	6.5
9	15.50	8.94	4.87	17.3	10.3	5.0	16.3	9.3	6.0
10	12.70	10.10	4.50	14.7	7.1	6.2	15.9	9.8	7.0
11	13.90	8.50	4.50	13.6	10.1	4.9	13.0	10.4	6.1
12	12.40	10.80	4.30	17.3	9.1	5.9	10.1	5.8	3.1
13	20.00	9.70	3.00	14.7	8.9	5.3	10.6	7.4	4.2
14	17.90	8.50	5.50	18.4	10.3	8.4	NA	NA	NA
15	12.90	8.90	7.70	20.5	12.7	7.6	NA	NA	NA
16	11.70	8.90	7.70	23.9	13.9	7.6	NA	NA	NA
17	13.00	10.10	6.80	17.9	9.4	5.8	NA	NA	NA
18	11.00	8.50	6.00	14.5	10.0	4.8	NA	NA	NA
19	12.20	9.34	5.96	13.9	8.5	5.5	NA	NA	NA
20	13.00	10.00	6.30	NA	NA	NA	NA	NA	NA
21	11.40	8.30	6.50	NA	NA	NA	NA	NA	NA
22	15.50	9.62	6.70	NA	NA	NA	NA	NA	NA
23	16.90	10.30	8.40	NA	NA	NA	NA	NA	NA
24	12.30	9.00	5.40	NA	NA	NA	NA	NA	NA
25	14.40	9.60	7.80	NA	NA	NA	NA	NA	NA

Figura 5: Estructura de los datos para testear la normalidad para cada variable y yacimiento coreano.

La ejecución del test de Shapiro-Wilk es el mismo que el mostrado en la Figura 4 (página 5). Los resultados aparecen a continuación:

```
> shapiro.test(KoreanSites$Length.Chongokni)
Shapiro-Wilk normality test
data:  KoreanSites$Length.Chongokni
W = 0.92353, p-value = 0.0617

> shapiro.test(KoreanSites$Length.Chuwoli.Kawoli)
Shapiro-Wilk normality test
data:  KoreanSites$Length.Chuwoli.Kawoli
W = 0.91279, p-value = 0.08331

> shapiro.test(KoreanSites$Length.Kumpari)
Shapiro-Wilk normality test
data:  KoreanSites$Length.Kumpari
W = 0.94276, p-value = 0.4936

> shapiro.test(KoreanSites$Thickness.Chongokni)
Shapiro-Wilk normality test
data:  KoreanSites$Thickness.Chongokni
W = 0.98291, p-value = 0.936
```

```

> shapiro.test(KoreanSites$Thickness.Chuwoli.Kawoli)
Shapiro-Wilk normality test
data:  KoreanSites$Thickness.Chuwoli.Kawoli
W = 0.93708, p-value = 0.2334

> shapiro.test(KoreanSites$Thickness.Kumpari)
Shapiro-Wilk normality test
data:  KoreanSites$Thickness.Kumpari
W = 0.89099, p-value = 0.1006

> shapiro.test(KoreanSites$Width.Chongokni)
Shapiro-Wilk normality test
data:  KoreanSites$Width.Chongokni
W = 0.94043, p-value = 0.1515

> shapiro.test(KoreanSites$Width.Chuwoli.Kawoli)
Shapiro-Wilk normality test
data:  KoreanSites$Width.Chuwoli.Kawoli
W = 0.93484, p-value = 0.2125

> shapiro.test(KoreanSites$Width.Kumpari)
Shapiro-Wilk normality test
data:  KoreanSites$Width.Kumpari
W = 0.98463, p-value = 0.9948

```

Estos resultados son más comprensibles si los agrupamos en la Tabla 2. Como podemos observar, todas las variables separadas por yacimiento siguen una distribución normal, ya que sus p-valores están situados por encima de 0.05.

Variable	Yacimiento	p-valor	Conclusión
Length	Chongokni	0.0617	Aceptamos H_0
Length	Chuwoli/Kawoli	0.0833	Aceptamos H_0
Length	Kumpari	0.4936	Aceptamos H_0
Thickness	Chongokni	0.9360	Aceptamos H_0
Thickness	Chuwoli/Kawoli	0.2334	Aceptamos H_0
Thickness	Kumpari	0.1006	Aceptamos H_0
Width	Chongokni	0.1515	Aceptamos H_0
Width	Chuwoli/Kawoli	0.2125	Aceptamos H_0
Width	Kumpari	0.9948	Aceptamos H_0

Tabla 2: Resultados del test de Shapiro-Wilk para testear la normalidad para las diferentes variables del estudio de bifaces coreanos separados por yacimiento.

Test de Kolmogorov-Smirnov (corrección Lilliefors)

El **test de Kolmogorov-Smirnov** (con la **corrección Lilliefors**) se utiliza para contrastar si un conjunto de datos se ajustan o no a una **distribución normal**. Es similar en este caso al test de Shapiro-Wilk, pero la principal diferencia con éste radica en el tamaño muestral. Mientras que el test de Shapiro-Wilk se puede utilizar con hasta 50 datos, el test de Kolmogorov-Smirnov es recomendable utilizarlo con más de 50 observaciones.

A pesar de que continuamente se alude al test Kolmogorov-Smirnov como un test válido para contrastar la normalidad, en verdad **esto no es del todo cierto**. El test Kolmogorov-Smirnov asume conocida la media y varianza poblacional, lo que, en la mayoría de los casos, es imposible conocer. Esto hace que el test sea **muy conservador y poco potente**. Para solventar este problema, se desarrolló una modificación del Kolmogorov-Smirnov conocida como **test Lilliefors**. **El test Lilliefors asume que la media y la varianza son desconocidas, estando espacialmente desarrollado para testear la normalidad.**

Antes de realizar el test de Kolmogorov-Smirnov (con la corrección Lilliefors), es necesario conocer cuál es el contraste de hipótesis que se va a realizar.

Contraste de hipótesis del test de Kolmogorov-Smirnov (Lilliefors)

- H_0 : los datos **provienen** de una distribución normal
- H_1 : los datos **no provienen** de una distribución normal

R Commander: test de Kolmogorov-Smirnov (Lilliefors)

La estructura de los datos es exactamente la misma que la mostrada en la Figura 2 (página 2) y Figura 5 (página 6), dependiendo de lo que queramos comprobar. De hecho, trabajamos con los mismos datos.

Como este test está especialmente desarrollado para tamaños muestrales superiores a 50, lo ejecutaremos con las variables sin separar por yacimiento, ya que son $n=57$.

Sin embargo, no existe un modo gráfico de hacerlo en R Commander, por lo que tendremos que recurrir a comandos y códigos.

Instalación y carga del paquete *nortest*

Ahora bien, es muy importante lo siguiente. Hay que instalar un paquete adicional (**nortest**) en vuestro R Commander.

Tenéis que ejecutar el código siguiente:

```
install.packages("nortest")
```

Cargáis el paquete a continuación.

Ejecución del test Lilliefors

La ejecución del test de Kolmogorov-Smirnov con corrección Lilliefors (para abreviarlo test de Lilliefors) se realiza, como se comentó previamente, de un modo manual, con código. En nuestro caso tenemos que tener en cuenta **el nombre que le hemos dado al conjunto de datos** en R Commander (*Korea*) y al **nombre de las variables** (*Length*, *Width*, *Thickness*), ya que en los códigos hay que introducirlos separados por el símbolo de dólar (\$).

Para contrastar el test de Lilliefors para nuestras tres variables, introducimos los siguientes comandos:

```
lillie.test(Korea$Length)
lillie.test(Korea$Width)
lillie.test(Korea$Thickness)
```

Los resultados que obtenemos son los siguientes:

```
> lillie.test(Korea$Length)
      Lilliefors (Kolmogorov-Smirnov) normality test
data:  Korea$Length
D = 0.080975, p-value = 0.4643

> lillie.test(Korea$Width)
      Lilliefors (Kolmogorov-Smirnov) normality test
data:  Korea$Width
D = 0.092678, p-value = 0.2582

> lillie.test(Korea$Thickness)
      Lilliefors (Kolmogorov-Smirnov) normality test
data:  Korea$Thickness
D = 0.088805, p-value = 0.3188
```

Resumiendo los resultados en la Tabla 3, observamos que las tres variables presentan una distribución normal, ya que sus p-valores están situados

Variable	p-valor	Conclusión
Length	0.4643	Aceptamos H_0
Thickness	0.3188	Aceptamos H_0
Width	0.2582	Aceptamos H_0

Tabla 3: P-valores de las tres variables del estudio de los bifaces coreanos sin distinción por yacimiento usando el test de Lilliefors por ser $n > 50$.