Chapter 1

Pruebas de Bondade Ajuste

Cuando recibimos datos y decidimos hacer inferencias sobre una población, lo primero que debemos hacer es tratar de verificar si nuestros datos se ajustan a un modelo paramétrico conocido (Ej. Normal, Poisson, Gamma, etc.).
Al procedimiento estadístico que utilizamos para ajustar un modelo paramétrico se le conoce como "Bondad de ajuste". En dicho procedimiento se pretende contrastar las siguientes hipótesis:

$$H_0 : F_X(x) = F_X^*(x)$$

$$H_1 : F_X(x) \neq F_X^*(x)$$

Donde F_X^* es una distribución que puede o no estar completamente especificada y F_X es la distribución de donde provienen los datos.
Existen varias pruebas en la literatura para hacer bondad de ajuste, las que veremos en este curso son:

- χ^2 de bondad de ajuste.
- Kolmogorov Smirnov.
- Lilliefors.
- Pruebas basadas en la QEDF (Cramer von Misses , Anderson Dorling).

1.1 χ^2 para bondad de ajuste

Se plantea la hipótesis:

$$H_0 : F_X(x) = F_X^*(x)$$

$$H_1 : F_X(x) \neq F_X^*(x)$$

Datos: $X_1, ..., X_n$ m.a. de F_X.
Supuestos: F_X^* está completamente especificada.

1.1.1 Método

Primero se divide el rango de las observaciones en k clases y se construye una tabla de contigencia que cuente el número de observaciones en cada clase.

Clase 1	Clase 2	...	Clase k
O_1	O_2	...	O_k

Recordemos que F_X es la verdadera pero desconocida distribución y F_X^* es alguna distribución completamente especificada y conocida (cuando F_X^* es conocida, excepto por sus parámetros, la prueba se puede ajustar).

Estadístico de prueba

Sea p_j la probabilidad de que una observación de la muestra se encuentre en la clase j bajo la H_0; es decir, bajo el supuesto que $F_X(x) = F_X^*(x)$. Definamos los valores esperado en cada clase como:

$$E_j = \mathbb{E}(O_j) = p_j n \quad j = \{1, 2, ..., k\}$$

donde entonces E_j representa el valor esperado de las observaciones en la clase j, bajo H_0.
Sea

$$T = \sum_{j=1}^{k} \frac{(O_j - E_j)^2}{E_j}$$

Bajo H_0 entonces se esperaría que T tome valores pequeños pues se espera que O_j tome un valor cercano a E_j.

El Teorema de Pearson garantiza que:

$$T \overset{H_0}{\sim} \chi^2_{(k-1)}$$

Luego entonces se rechaza H_0 si T es grande.

Regla de decisión
Rechazar si $T > \chi^{2^{(1-\alpha)}}_{(k-1)}$.

Consideraciones para hacer la prueba:

- Si algunos E_j son pequeñas, la aproximación hacia la χ^2 no es buena.
- Clases con E_j pequeñas deben de ser combinadas con otras de tal manera que sólo el 20% de las E_j sean menores a 5 y ninguna menor que 1.
- El número de clases k es arbitrario, lo que puede quitarle credibilad a la prueba.

Comentarios de la pruebas de bondad de ajuste:

- En la práctica casi siempre pasará que la verdadera distribución de los datos no es exactamente la distribución con la que contrastamos bajo H_0. Sin embargo, las pruebas de bondad de ajuste nos sirven como una medida de qué tan buena es la aproximación y justifica el uso F^*_X para posibles inferencias a las poblaciones estudiadas.
- Siempre esperamos que la prueba arroje $p-values$ altos pues esto se traduce en potencias altas de la prueba y por lo tanto minimiza la probabilidad del error tipo II.
- En estadística clásica es común decir *No se Rechaza* H_0 o *No tenemos evidencia sufuciente para Rechazar* H_0, en el caso de las pruebas de bondad de ajuste existe algo similar, en este caso no podemos afirmar que la verdadera distribución de los datos es la que se constrató, es por eso que en caso de no rechazar la prueba se acostumbra decir:
  Parece ser que el modelo F^*_X es una buena aproximación para la verdadera distribución de los datos.

1.1.2 Distribución exacta de T

Como mencionamos anteriormente la estadística de prueba T sigue una distribución aproximada a la χ^2, sin embargo surge la pregunta de saber cuál es su distribución exacta. Obtener de forma explicita la distribución de T es complicado debido a las operaciones que se realizan en su cálculo, sin embargo hoy en día con ayuda de las computadoras, pueden llevarse a cabo simulaciones de la variable aleatoria T y por tanto aproximar las distintas características de la verdadera distribución por medio de estas simulaciones.

Si F^*_X está completaente especificada y una vez que se definen las k clases se pueden determinar las p_j asociados a cada clase. Luego entonces bajo H_0 el vector aleatorio $(O_1, O_2 \dots O_k)$ sigue una distribución conocida como **multinomial** de parámetros $(n, p_1, \dots, p_k)$ y tiene por densidad:

$$\mathbb{P}\left(O_1 = o_1, O_2 = o_2 \ldots, O_k = o_k\right) = \frac{n!}{o_1! o_1! \ldots o_k!} p_1^{o_1} \ldots p_k^{o_k}$$

La distribución multinomial y sus propiedades pueden ser consultadas en los libros básicos de probabilidad, la gran ventaja que tenemos de esto es que el paquete R tiene funciones específicas para poder simular de dicha distribución.

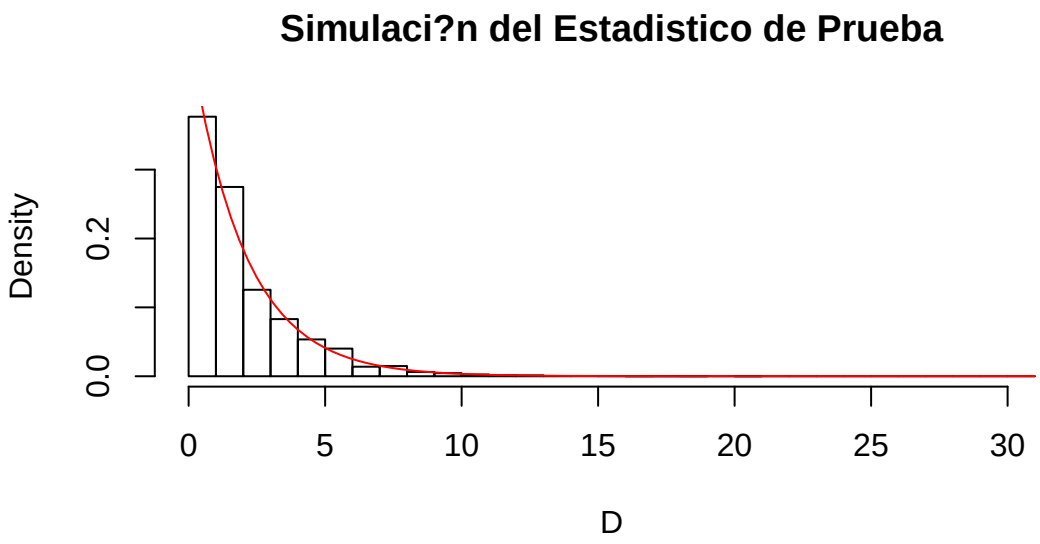
Entonces a través de las simulaciones del vector $(O_1, \ldots, O_k)$ pueden obtenerse simulaciones de la estadística T y, si el número de simulaciones es grande, generar cuantiles de la distribución exacta de T para realizar las pruebas.

Simulación de T

Por medio de un programa en R se puede simular la distribución T, usando la función **rmultinom**.

A continuación se presenta el código para simular la distribución exacta de T bajo H_0, cuando se tiene 3 clases ($k = 3$).

```
#m = numero de simulaciones
 m = 1000000
#n = tamano de la muestra
 n = 40
#k = numero de clases
 k = 3
#p = vector con las probabilidades de X_i este en cada una de las k clases
 p = c(0.3,0.3,0.4)

 #Simulacion de los observado
 O = t(rmultinom(m,n,p))

 #Calculamos  (observados - los esparados)^2/esperados
 S = ( O-rep(1,m)%*%t(n*p) )^2/(n*p)

 #Sumamos
 D=rowSums(S)

 #Obtenemos histograma
 hist(D,freq=FALSE,breaks=30,main="Simulaci?n del Estadistico de Prueba")

 #sobreponemos lo distribucion aproximada de chi.cuadrado
 curve(dchisq(x,k-1),add=TRUE,col=2)
```




```
#comparacion de cuantiles
alpha = c(0.2,0.1,0.05,0.025,0.01)
a <- round(quantile(D,1-alpha),2)
#cuantiles aproximados
b <- as.data.frame(round(t(qchisq(1-alpha,k-1)),2))
colnames(b) <- 1-alpha
a
```

```
##   80%   90%   95% 97.5%   99%
##  3.17  4.58  5.98  7.65  9.48
```

```
b
```

```
##    0.8  0.9 0.95 0.975 0.99
## 1 3.22 4.61 5.99  7.38 9.21
```

En esta última parte debe de obsérvarse como los cuantiles exactos (obtenidos vía simulación), son muy parecidos a los cuantiles propuestos por la aproximación a la χ^2

1.1.3 Teorema de Pearson

La pregunta que surge de manera natural es saber por qué bajo H_0 se tiene que $T \sim \chi^2_{(k-1)}$. Veamos un ejemplo, suponiendo que tenemos sólo 2 clases, $k = 2$

Si $k = 2$ entonces bajo H_0

$$\mathbb{P}\left(X_i \in C_1\right) = p_1 \quad y \quad \mathbb{P}\left(X_i \in C_2\right) = 1 - p_1$$

definamos

$$y_i^1 = \begin{cases} 1 & \text{si } X_i \in C_1 \\ 0 & \text{si } X_i \notin C_1 \end{cases} \tag{1.1}$$

$$y_i^2 = \begin{cases} 1 & \text{si } X_i \in C_2 \\ 0 & \text{si } X_i \notin C_2 \end{cases} \tag{1.2}$$

observaciones

$$\sum_{i=1}^{n} y_i^1 = O_1$$

O_1 es el número de observaciones en C_1 ,

$$\sum_{i=1}^{n} y_i^2 = O_2$$

O_2 es el número de observaciones en C_2 Entonces

$$O_1 \sim Bin(n, p_1)$$

$$O_2 \sim Bin(n, p_2)$$

La prueba χ^2 nos dice que:

$$T = \left(\frac{(O_1 - np_1)^2}{np_1}\right) + \left(\frac{(O_2 - np_2)^2}{np_2}\right) \overset{aprox}{\sim} \chi^2_{(2-1)}$$

Veamos el por qué de esto:

Por un lado sabemos que existe una aproximación de la normal hacia la distribución binomial (Por el Teorema del Límite Central). Entonces

$$O_1 \sim Bin(n, p_1)$$

$$\Rightarrow \frac{(O_1 - np_1)}{\sqrt{np_1(1 - p_1)}} \overset{aprrox}{\sim} N(0, 1)$$

$$\Rightarrow \frac{(O_1 - np_1)^2}{np_1(1-p_1)} \overset{aprrox}{\sim} \chi^2_{(1)}$$

Si O_1 y O_2 fueran independientes entonces se podría probar que:

$$\sum_{j=1}^{2} \left(\frac{(O_j - np_j)^2}{np_j(1-p_j)} \right) \overset{aprrox}{\sim} \chi^2_{(2)}$$

Sin embargo sabemos que O_1 y O_2 sí son dependientes, de hecho $O_2 = n - O_1$ Entonces veamos otro camino, por un lado ya sabemos que:

$$\frac{(O_1 - np_1)^2}{np_1(1-p_1)} \overset{approx}{\sim} \chi^2_{(1)}$$

Expresemos a la cantidad $\frac{(O_1 - np_1)^2}{np_1(1-p_1)}$ de otra forma:

$$\frac{(O_1 - np_1)^2}{np_1(1-p_1)} = \frac{(O_1 - np_1)^2(1-p_1) + (O_1 - np_1)^2 p_1}{np_1(1-p_1)}$$

$$= \frac{(O_1 - np_1)^2(1-p_1)}{np_1(1-p_1)} + \frac{(O_1 - np_1)^2(p_1)}{np_1(1-p_1)}$$

$$= \frac{(O_1 - np_1)^2}{np_1} + \frac{(O_1 - np_1)^2}{np_2} = \frac{(O_1 + np_1)^2}{np_1} + \frac{(n - O_2 - n(1-p_2)^2}{np_2}$$

$$= \frac{(O_1 - np_1)^2}{np_1} + \frac{(np_2 - O_2)^2}{np_2} = \frac{(O_1 - np_1)^2}{np_1} + \frac{(O_2 - np_2)^2}{np_2}$$

$$\Rightarrow \frac{(O_1 - np_1)^2}{np_1} + \frac{(O_2 - np_2)^2}{np_2} = \frac{(O_1 - np_1)^2}{\sqrt{np_1(1-p_1)}} \overset{approx}{\sim} \chi^2_{(1)}$$

La demostración par el caso general la damos a continuación:

Considerense k cajas $B_1, \ldots, B_k$ y supongamos que tiramos n bolas $X_1, \ldots, X_n$ en estas cajas al azar, independiente una de la otra, con probabilidades:

$$\mathbb{P}(X_i \in B_1) = p_1, \ldots, \mathbb{P}(X_i \in B_k) = p_k$$

donde:

$$\sum_{j=1}^{k} p_j = 1$$

Sea v_j el número de bolas en la j ésima caja:

$$v_j = \{\#\ \textbf{bolas en la caja}\ B_j\} = \sum_{i=1}^{n} \mathbf{1}_{(X_l \in B_j)}$$

En promedio, el número de bolas en la j-ésima caja será np_j, por lo que la variable aleatoria debería estar cerca de np_j. También se puede utilizar el Teorema del Límite Central para describir qué tan cerca esta v_j de np_j. El siguiente resultado nos dice cómo podemos describir en cierto sentido la cercanía de v_j a np_j simultáneamente para todas las $j \leq k$. El principal problema en este Teorema viene del hecho de que las variables aleatorias v_j para $j \leq k$ no son independientes, pues sabemos que el número total de bolas es igual a n,

$$\sum_{j=1}^{k} v_j = n$$

es decir, si conocemos estos números en $k-1$ cajas sabremos automáticamente su número en la última caja.

Teorema 1.1.1. *Tenemos que la variable aleatoria*

$$\sum_{j=1}^{k} \frac{(v_j - np_j)^2}{np_j} \to \chi^2_{(k-1)}$$

converge en distribución a una distribución $\chi^2_{(k-1)}$ con $k-1$ grados de libertad

Proof. Fijemos una caja de B_j. Luego definamos las variables aleatorias:

$$\mathbf{1}_{(X_i \in B_j)}, \ldots, \mathbf{1}_{(X_n \in B_j)}$$

que indican si cada observación X_i esta en la caja B_j o no. Observe que estas v.a. son i.i.d con una distribución Bernoulli con probabilidad de éxito

$$\mathbb{E}\big(\mathbf{1}_{(X_1 \in B_j)}\big) = \mathbb{P}\left(X_1 \in B_j\right) = p_j$$

y varianza

$$Var(\mathbf{1}_{(X_1 \in B_j)}) = p_j(1-p_j)$$

Por lo tanto, por el teorema del límite central sabemos que la variable aleatoria

$$\frac{v_j - np_j}{\sqrt{np_j(1-p_j)}}) = \frac{\sum_{i=1}^{n} \mathbf{1}_{(X_l \in B_j)} - np_j}{\sqrt{np_j(1-p_j)}}$$

$$= \frac{\sum_{i=1}^{n} \mathbf{1}_{(X_l \in B_j)} - n\mathbb{E}\big(\mathbf{1}_{(X_l \in B_j)}\big)}{\sqrt{n\mathrm{Var}\big(\mathbf{1}_{(X_l \in B_j)}\big)}} \to N(0,1)$$

converge a una distribución normal estándar. Por lo tanto, la variable aleatoria

$$\frac{v_j - np_j}{\sqrt{np_j}} \to \sqrt{1-p_j}N(0,1) = N(0, 1-p_j)$$

converge a una distribución normal con varianza $1-p_j$. Siendo un poco informal, podemos decir que

$$\frac{v_j - np_j}{\sqrt{np_j}} \to Z_j$$

donde $Z_j \sim N(0, 1-p_j)$. Sabemos que cada Z_j tiene una distribución $N(0, 1-p_j)$ pero, desafortunadamente esto no nos dice lo que $\sum Z_j^2$ será, ya que como hemos mencionado arriba v_j no son independientes y su estructura de correlación jugará un papel importante. Calculemos la covarianza entre $\frac{v_i - np_i}{\sqrt{np_i}}$ y $\frac{v_j - np_j}{\sqrt{np_j}}$

$$Cov\left(\frac{v_i - np_i}{\sqrt{np_i}}, \frac{v_j - np_j}{\sqrt{np_j}}\right) = \mathbb{E}\left(\left(\frac{v_i - np_i}{\sqrt{np_i}}\right)\left(\frac{v_j - np_j}{\sqrt{np_j}}\right)\right)$$

$$= \frac{1}{n\sqrt{p_i p_j}}(\mathbb{E}(v_i v_j) - \mathbb{E}(v_i np_j) - \mathbb{E}(v_j np_i) + n^2 p_i p_j)$$

$$= \frac{1}{n\sqrt{p_i p_j}}(\mathbb{E}(v_i v_j) - np_i np_j - np_j np_i + n^2 p_i p_j) = \frac{1}{n\sqrt{p_i p_j}}(\mathbb{E}(v_i v_j) - n^2 p_i p_j)$$

Para calcular $\mathbb{E}(v_i v_j)$ vamos a utilizar el hecho de que una pelota no puede estar dentro de dos diferentes cajas de forma simultánea lo que significa que

$$\mathbf{1}_{(X_l \in B_i)}\mathbf{1}_{(X_l \in B_j)} = 0$$

Por lo tanto,

$$\mathbb{E}(v_i v_j) = \mathbb{E}\left(\sum_{l=1}^{n} \mathbf{1}_{(X_l \in B_i)} \sum_{l'=1}^{n} \mathbf{1}_{(X_{l'} \in B_j)}\right) = \mathbb{E}\left(\sum_{l,l'} \mathbf{1}_{(X_l \in B_i)}\mathbf{1}_{(X_{l'} \in B_j)}\right)$$

$$= \mathbb{E}\left(\sum_{l=l'} \mathbf{1}_{(X_l \in B_i)}\mathbf{1}_{(X_{l'} \in B_j)}\right) + \mathbb{E}\left(\sum_{l \neq l'} \mathbf{1}_{(X_l \in B_i)}\mathbf{1}_{(X_{l'} \in B_j)}\right)$$

$$= n(n-1)\mathbb{E}\big(\mathbf{1}_{(X_l \in B_i)}\big)\,\mathbb{E}\big(\mathbf{1}_{(X_{l'} \in B_j)}\big) = n(n-1)p_i p_j$$

Por lo tanto, la covarianza de arriba es igual a

$$\frac{1}{n\sqrt{p_i p_j}}(n(n-1)p_i p_j - n^2 p_i p_j) = -\sqrt{p_i p_j}$$

En resumen, hasta ahora hemos demostrado que la variable aleatoria

$$\sum_{j=1}^{k} \frac{(v_j - np_j)^2}{np_j} \to \sum_{j=1}^{k} Z_j^2$$

donde las variables aleatorias $Z_1, \ldots, Z_k$ satisfacen que son normales y que $\mathbb{E}(Z_i^2) = 1 - p_i$ con covarianza $\mathbb{E}(Z_i Z_j) = -\sqrt{p_i p_j}$
Es decir, viendo de forma vectorial $\underline{Z}$ se tiene que:

$$\begin{aligned} \mathrm{Var}(\underline{Z}) &= \begin{pmatrix} 1-p_1 & -\sqrt{p_1p_2} & -\sqrt{p_1p_3} & \ldots & -\sqrt{p_1p_k} \\ -\sqrt{p_2p_1} & 1-p_2 & -\sqrt{p_2p_3} & \ldots & -\sqrt{p_2p_k} \\ -\sqrt{p_3p_1} & -\sqrt{p_3p_2} & 1-p_3 & \ldots & -\sqrt{p_3p_k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -\sqrt{p_kp_1} & -\sqrt{p_kp_2} & -\sqrt{p_kp_3} & \ldots & 1-p_k \end{pmatrix} = \left(\mathbf{I}_{k\times k} - \underline{p}\underline{p}^T\right) \\ \mathbb{E}(\underline{Z}) &= \underline{0} \end{aligned}$$

Donde:

$$\underline{p} = (\sqrt{p_1}, \ldots, \sqrt{p_k})^T \Rightarrow \underline{p}^T\underline{p} = \sum_{j=1}^{k} p_j = 1$$

Para demostrar el teorema queda mostrar que la estructura de covarianzas de la secuencia de $Z_i's$ implicará que su suma de cuadrados tiene distribución $\chi^2_{(k-1)}$. Para mostrar esto, encontraremos una representación diferente para $\sum_{j=1}^{k} Z_j^2$.

Sea $G_1, \ldots, G_k$ una secuencia de normales estándar i.i.d., es decir:

$$\underline{G} = (G_1, \ldots, G_k)^T \sim N_k\left(\underline{0}, \mathbf{I}_{k\times k}\right)$$

Con lo anterior construyamos al vector aleatorio $\underline{H}$ como:

$$\underline{H} = \underline{G} - \underline{p}(\underline{p}^T\underline{G}) = \underline{G} - (\underline{p}\underline{p}^T)\underline{G} = \left(\mathbf{I}_{k\times k} - \underline{p}\underline{p}^T\right)\underline{G}$$

Mostraremos que $\underline{H}$ y $\underline{Z}$ tienen la misma distribución. Para mostrar esto notemos que:

$$\begin{aligned}
\mathbb{E}(H) &= \mathbb{E}\left(\left(\mathbf{I}_{k\times k} - \underline{pp}^T\right)\underline{G}\right) = \left(\mathbf{I}_{k\times k} - \underline{pp}^T\right)\mathbb{E}(\underline{G}) = \underline{0} = \mathbb{E}(\underline{Z}) \\
\mathrm{Var}(H) &= \mathrm{Var}\left(\left(\mathbf{I}_{k\times k} - \underline{pp}^T\right)\underline{G}\right) = \left(\mathbf{I}_{k\times k} - \underline{pp}^T\right)\mathrm{Var}(\underline{G})\left(\mathbf{I}_{k\times k} - \underline{pp}^T\right)^T \\
&= \left(\mathbf{I}_{k\times k} - \underline{pp}^T\right)\left(\mathbf{I}_{k\times k} - \underline{pp}^T\right)^T = \left(\mathbf{I}_{k\times k} - \underline{pp}^T\right)\left(\mathbf{I}_{k\times k} - \underline{pp}^T\right) \\
&= \left(\mathbf{I}_{k\times k} - \underline{pp}^T\right) = \mathrm{Var}(\underline{Z})
\end{aligned}$$

Observe entonces que la matriz $\mathbf{I}_{k\times k} - \underline{pp}^T$ es idempotente.

Por lo tanto $\underline{H}$ (Combinación lineal de $\underline{G} \sim N_k\left(\underline{0}, \mathbf{I}\right)$) y $\underline{Z}$ son vectores aleatorios con entradas normales y con la misma esperanza y misma matriz de varianzas covarianzas, por lo tanto los vectores $\underline{H}$ y $\underline{G}$ tienen la misma distribución. Entonces:

$$\sum_{j=1}^{k} Z_j^2 \overset{d}{=} \sum_{j=1}^{k} H_j^2$$

Donde $\overset{d}{=}$ indica igualdad entre distribución.

Pero

$$\begin{aligned}
\sum_{j=1}^{k} H_j^2 &= \underline{H}^T\underline{H} = \left(\left(\mathbf{I}_{k\times k} - \underline{pp}^T\right)\underline{G}\right)^T\left(\mathbf{I}_{k\times k} - \underline{pp}^T\right)\underline{G} \\
&= \underline{G}^T\left(\mathbf{I}_{k\times k} - \underline{pp}^T\right)\left(\mathbf{I}_{k\times k} - \underline{pp}^T\right)\underline{G} \\
&= \underline{G}^T\left(\mathbf{I}_{k\times k} - \underline{pp}^T\right)\underline{G}
\end{aligned}$$

Se tiene entonces que $\sum_{j=1}^{k} Z_j^2$ es igual en distribución a una forma cuadrática $\underline{G}^T\left(\mathbf{I}_{k\times k} - \underline{pp}^T\right)\underline{G}$ donde $\underline{G} \sim N_r\left(\underline{0}, \mathbf{I}_{k\times k}\right)$. Luego por teorema de formas cuadráticas tenemos que:

$$\underline{G}^T\left(\mathbf{I}_{k\times k} - \underline{pp}^T\right)\underline{G} \sim \chi^2_{\left(Rank\left(\mathbf{I}_{k\times k} - \underline{pp}^T\right)\right)}$$

Pero por la idempotencia de la matriz $\mathbf{I}_{k\times k} - \underline{pp}^T$ se tiene que:

$$Rank\left(\mathbf{I}_{k\times k} - \underline{pp}^T\right) = Traza\left(\mathbf{I}_{k\times k} - \underline{pp}^T\right) = k - 1$$

Por lo tanto:

$$\underline{G}^T\left(\mathbf{I}_{k\times k} - \underline{pp}^T\right)\underline{G} \sim \chi^2_{(k-1)}$$

□

Uno de los puntos importantes de esta prueba es que se requiere que bajo H_0 se tenga una distribución complemente especificada (simple), surge entonces de manera natural la pregunta ¿cómo adaptar la prueba al caso compuesto?, es decir cuando bajo H_0 la distribución no está completamente especificada.

Supongamos entonces que ahora F_X^* está completamente especificada excepto por q parámetros desconocidos. Ejemplo

$$H_0 : F_X(x) = N(\mu, \sigma^2)$$

$$H_1 : F_X(x) \neq N(\mu, \sigma^2)$$

$$H_0 : F_X(x) = N(\mu, 1)$$

$$H_1 : F_X(x) \neq N(\mu, 1)$$

La solución que se propone para adaptar la prueba consiste en reducir los grados de libertad asociados a la distribución χ^2 asociada.

Recordemos que en estadística inferencial una característica que se observa con la distribución χ^2 es que va perdiendo grados de libertad en función de los parámetros estimados dentro de la estadística utilizada, en este caso aplicaremos esta misma regla y ajustaremos la prueba de la siguiente manera:

- Dada la muestra $X_1, \ldots, X_n$ de F_X estimar los q parámetros faltantes.
- **Enchufar** las estimaciones en la distribución F_X y hacer la prueba χ^2, es decir calcular el estadístico:

$$T = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

- Rechazar H_0 a un nivel de significancia α si:

$$T > \chi^{2^{(1-\alpha)}}_{(k-1-q)}$$

Es decir, ahora la prueba supone que el estadístico T sigue una distribución $\chi^2_{(k-1-q)}$, en donde se restan los grados de libertad en función de los parámetros estimados en la distribución.

Observe que siempre se debe garantizar que los grados de libertad cumplan con $k-1-q \geq 1$, es decir $k \geq 2+q$ por lo que el número de clases se debe incrementar para ganranizarlo. Lo

anterior nos obliga a tener más muestra para garantizar que los números esperados en cada clase no sean menores a 5.

1.2 Kolmogorov-Smirnov test

Suponga nuevamente que tenemos una muestra $X_1, \ldots, X_n$ i.i.d. con alguna distribución desconocida F_X y nos gustaría probar si la verdadera distribución es F_X^*, entonces planteamos la siguiente prueba de hipótesis:

$$H_0 : F_X(x) = F_X^*(x)$$

$$H_1 : F_X(x) \neq F_X^*(x)$$

En la sección anterior vimos como probar esta hipótesis por medio de la prueba χ^2. Consideraremos ahora una prueba diferente para H_0 basada en una idea diferente que evita la construcción de clases arbitrarias. Para ello definamos la función de distribución empírica.

Definición 1.2.1 (Función de distribución empírica)**.** *Sea $X_1, \ldots, X_n$ una muestra aleatoria de cierta distribución $F_X(x)$, se define la función de distribución empírica como:*

$$F_n(x) = \frac{\textbf{\textit{Número de Observaciones}} \leq x}{n} = \frac{\sum_{i=1}^{n} \mathbf{1}_{(X_i \leq x)}}{n}$$

Dado que suponemos que la muestra aun no ha sido observada entonces X_i es v.a. y por tanto $F_n(x)$ es una v.a. también. Lo primero que mostraremos es que esta v.a. es un estimador insesgado y consistente para $F_X(x)$.

Obtengamos la esperanza:

$$\mathbb{E}(F_n(x)) = \mathbb{E}\left(\frac{\sum_{i=1}^{n} \mathbf{1}_{(X_i \leq x)}}{n}\right) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left(\mathbf{1}_{(X_i \leq x)}\right)$$

Pero observe que $\mathbf{1}_{(X_i \leq x)} \sim Bernoulli\left(\mathbb{P}(X_i \leq x)\right) = Bernoulli\left(F_X(x)\right)$ por lo tanto $\mathbb{E}\left(\mathbf{1}_{(X_i \leq x)}\right) = F_X(x)$ para toda i, lo que prueba entonces que:

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left(\mathbf{1}_{(X_i \leq x)}\right) = \frac{1}{n} n \mathbb{E}\left(\mathbf{1}_{(X_i \leq x)}\right) = F_X(x)$$

Lo que prueba el insesgamiento de $F_n(x)$.

Por otro lado, para probar la consistencia de $F_n(x)$, basta probar que la varianza del estimador

tiende a 0 conforme la muestra tiende a infinito, para ello calculemos la varianza:

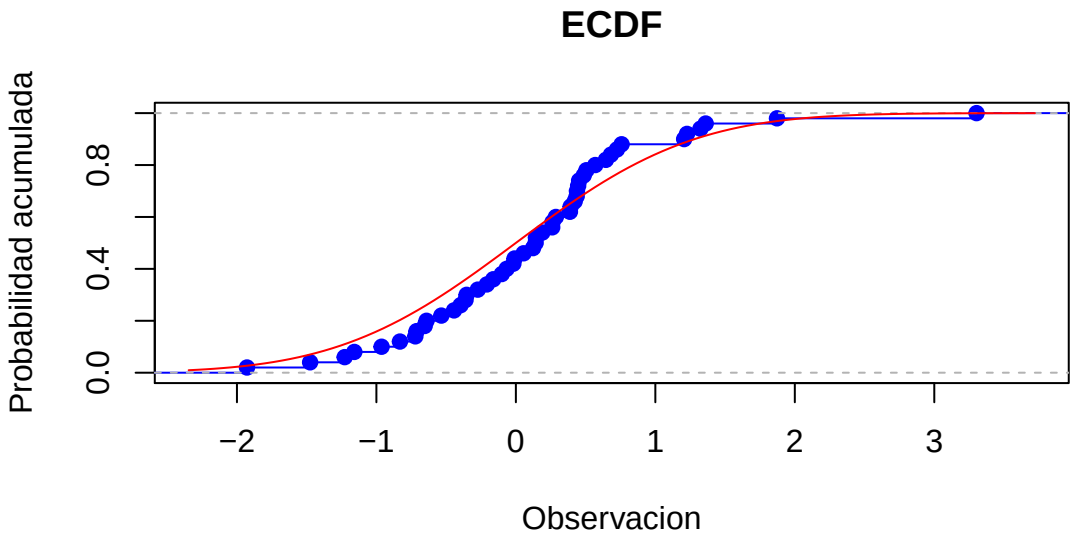
$$\mathrm{Var}(F_n(x)) = \mathrm{Var}\left(\frac{\sum_{i=1}^{n} \mathbf{1}_{(X_i \leq x)}}{n}\right) = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}\left(\mathbf{1}_{(X_i \leq x)}\right) = \frac{F_X(x)(1 - F_X(x))}{n}$$

De donde se observa que en efecto, conforme aumenta la muestra la varianza del estimador se hace cada vez más pequeña.

Cuando suponemos observada la muestra, entonces F_n es una función que puede ser graficada, la cual esperamos aproxime bastante bien a la verdadera distribución de donde proviene la muestra.

A continuación se escribe el código en R de una simulación de una muestra normal, se construye la distribución empírica y se observa el ajuste que hace esta función con la verdadera distribución de los datos.

```
x<-rnorm(50,0,1);
f<-ecdf(x)
plot(f,xlab="Observacion",main="ECDF",ylab="Probabilidad acumulada",col=4);
curve(pnorm(x,0,1),add=TRUE,col=2);
```



Suena lógico entonces que la función de distribución empírica tiene la información necesaria como para ser utilizada en la construcción de un estadístico de prueba para verificar la bondad de ajuste.

Con ayuda del insesgamiento y consitencia del estimandor F_n se puede probar que $F_n(x)$ converge casi seguramente a $F_X(x)$ de forma puntual, es decir, para cada x, sin embargo hay un

resultado aun mejor que nos indica que la convergencia se hace de forma uniforme.

Teorema 1.2.1 (Teorema de Glivenko-Cantelli)**.** *Sea $X_1, \ldots, X_n$ m.a. de $F_X(x)$ y sea $F_n(x)$ la respectiva función de distribución empírica. Entonces:*

$$\sup_{x \in R} |F_n(x) - F_X(x)| \to 0$$

[1]

Es decir, que conforme tenemos más muestra entones F_n prácticamente reproduce a la verdader función de distribución.

La idea entonces que utiliza la prueba de Kolmogorov es definir por estadística precisamente al supremo de las diferencias absolutas entre la distribución empírica y la que se propone en el contraste de hipótesis bajo H_0.

La observación clave en la prueba de Kolmogorov-Smirnov es que la distribución de este supremo no depende de la distribución desconocida F_X^* de la muestra, siempre y cuando F_X^* sea una distribución continua.

Teorema 1.2.2. *Si $F_X(x)$ es continua entonces la distribución de*

$$\sup_{x \in R} |F_n(x) - F_X(x)|$$

no depende de F_X.

Proof. Como $F_X(x)$ es continua, entonces sabemos que existe la inversa $F_X^{-1}(y)$. Entonces ha-

[1]Una de las aplicaciones de este Teorema se encuentra en las gráficas QQPlot la cuales verifican visualmente si una distribución propuesta se está ajustando a los datos proporcionados. La idea de las gráficas QQPlot es verficar si la distribución empirica F_n y la distribución propuesta F_X se parecen, para ello considera el siguiente resultado teórico basa en el Teorema de Glivenko-Cantelli

$$F_X^{-1}\left(F_n(x)\right) \approx x$$

Luego, en el eje X se coloca la cantidad $F_X^{-1}\left(F_n(x)\right)$ mientras que en el eje Y el valor x, si en realidad hay una buena aproximación de la distribución empírica con la distribución F_X, se espera que los puntos se ajusten bastante bien a la recta identidad, en caso contrario entonces se asume que no hay ajuste de los datos al modelo F_X

Debido a que $F_n(x_{(n)} = 1$ entonces $F_X^{-1}\left(F_n(x_{(n)})\right)$ tomará el valor de infinito en el caso de distribuciones como la Normal, para evitar esto se lleva a cabo ua ajuste definiendo a la distribución empírica como:

$$F_n(x_{(i)}) = \frac{i - 0.5}{n}$$

ciendo el cambio de variable $y = F_X(x)$ o $x = F_X^{-1}(y)$ podemos escribir

$$\mathbb{P}\left(\sup_{x\in R}|F_n(x) - F_X(x)| \leq t\right) = \mathbb{P}\left(\sup_{0\leq y\leq 1}|F_n(F_X^{-1}(y)) - y| \leq t\right)$$

Usando la definición de F_n podemos escribir

$$F_n(F_X^{-1}(y)) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{(X_i\leq F^{-1}(y))} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{(F_X(X_i)\leq y)}$$

Por lo tanto,

$$\mathbb{P}\left(\sup_{0\leq y\leq 1}\left|F_n(F_X^{-1}(y)) - y\right| \leq t\right) = \mathbb{P}\left(\sup_{0\leq y\leq 1}\left|\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{(F_X(x_i)\leq y)} - y\right| \leq t\right).$$

Sin embargo sabemos que la distribución de $F_X(x_i)$ es uniforme en el intervalo $[0, 1]$ pues

$$\mathbb{P}\left(F_X(x_1) \leq t\right) = \mathbb{P}\left(x_1 \leq F_X^{-1}(t)\right) = F_X\left(F_X^{-1}(t)\right) = t$$

Por lo que las variables aleatorias

$$U_i = F_X(x_i) \quad i \in \{1, \ldots, n\}$$

son independientes y tinen una distribución uniforme entre $[0, 1]$, así hemos probado que

$$\mathbb{P}\left(\sup_{x\in R}|F_n(x) - F_X(x)| \leq t\right) = \mathbb{P}\left(\sup_{0\leq y\leq 1}\left|\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{(U_i\leq y)} - y\right| \leq t\right)$$

es independiente de F_X. □

Para motivar la prueba KS, necesitaremos un resultado más, el cual formularemos sin probarlo. Primero, notemos que para un punto fijo x el Teorema del Límite Central nos dice que:

$$\sqrt{n}(F_n(x) - F_X(x)) \to N(0, F_X(x)(1 - F_X(x)))$$

porque $F_X(x)(1 - F_X(x))$ es la varianza de $\mathbf{1}_{(x_1\leq x)}$. Resulta que si consideramos

$$\sqrt{n}\sup_{x\in R}|F_n(x) - F_X(x)|$$

también convergerá en distribución.

Teorema 1.2.3. *Tenemos*

$$\mathbb{P}\left(\sqrt{n}\sup_{x\in R}|F_n(x)-F_X(x)|\leq t\right)\to H(t)=1-2\sum_{i=1}^{\infty}(-1)^{i-1}e^{-2i^2t}$$

donde H(t) es la función de la distribución de Kolmogorov-Smirnov.

Reformulemos la hipótesis en términos de la función de distribución acumulada:

$$H_0: F_X(x)=F_X^*(x) \quad \textbf{vs.} \quad H_1: F_X(x)\neq F_X^*(x)$$

donde F_X^* es la distribución que queremos ajustar a los datos. Consideremos el siguiente estadístico:

$$D_n=\sup_{x\in R}|F_n(x)-F_X^*(x)|=^2\max_{1\leq i\leq n}\left\{\max\left\{F_n\left(x_{(i-1)}\right)-F_X^*(x_{(i)})\right\},\max\left\{F_n\left(x_{(i)}\right)-F_X^*(x_{(i)})\right\}\right\}$$

Si la hipótesis nula es verdadera entonces, por el Teorema 1.2.2, podemos tabular la distribución de D_n (dependerá sólo de n). Además, si n es lo suficientemente grande entonces la distribución de $\sqrt{n}D_n$ es aproximada por la distribución de Kolmogorov-Smirnov del Teorema 1.2.3. Por otra parte, si suponemos que la hipótesis nula es falsa; es decir, $F_X\neq F_X^*$, entonces al ser F_X la verdadera distribución de los datos, entonces la distribución empírica F_n convergerá a F_X y por lo tanto no se aproximaría a F_X^*; es decir, para una n grande tendremos que existe una $\delta>0$ tal que:

$$\sup_x|F_n(x)-F_X^*(x)|>\delta$$

Concluimos entonces que si H_0 falla entonces $D_n>\delta$. Por lo tanto, la prueba H_0 considerará una regla de decisión como sigue:

$$\delta=\begin{cases}H_0 & : D_n\leq c\\ H_1 & : D_n>c\end{cases} \tag{1.3}$$

El límite de c depende del nivel de significancia α el cual debe de cumplir con la siguiente condición:

$$\alpha=\mathbb{P}(\ \textbf{Rechazar}\ H_0|H_0\textbf{Cierta})=\mathbb{P}(D_n\geq c|H_0\ \textbf{Cierta})$$

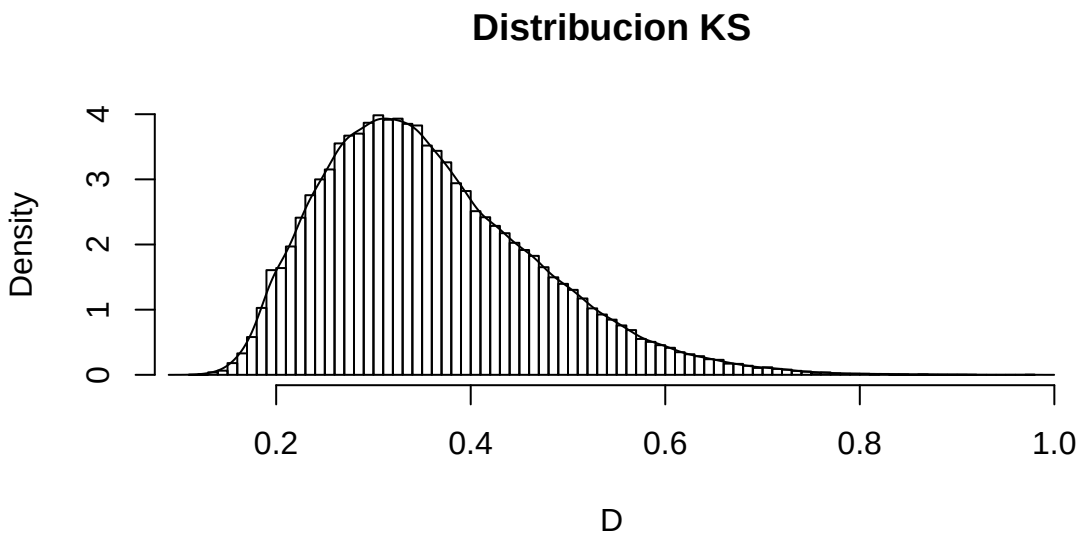
Debido a que bajo H_0 la distribución de D_n puede ser tabulada para cada n, podemos encontrar el valor c en tablas o por medio de simulación. De hecho, la mayoría de los libros de estadística tienen esas distribuciones desde $n=1$ hasta $n=40$.

[2]Donde $x_{(0)}:=0$

A continuación presentamos el código en R para simular la distribución para $n = 5$.

```
#######################################################
# Programa que simula lo Distribucion de Kolmogorov #
#######################################################

#Seleccione el tama\~no de muestra
n=5
#################################################################
# Libro de Conover Tabla 14     (2 colas)                       #
#Cuantil         0.80    0.90    0.95    0.98    0.99          #
#n=5             0.447   0.509   0.563   0.627   0.669         #
#################################################################

#############
m=100000    #Numero de simulaciones
#############

#Inicializamos variable que guardara las simulaciones
D=rep(0,m)
for (j in 1:m){
 #Simulacion de las variables uniformes
 x=runif(n,0,1)
 #Se ordena la muestra
 x=sort(x)
 #Calculamos la funcion de distribucion empirica
 Fn=ecdf(x)
 #A la muestra ordenada le agregamos el 0 al principio
 #Sirve para el caso F_n(x(0))=0
 y=c(0,x)
 #Inicializamos busqueda de supremo
 D1=0
 D2=0
 for (i in 2:(n+1)){
   D1[i]=abs(Fn(y[i])-y[i])
   D2[i]=abs(Fn(y[i-1])-y[i])
 }
 #Obtenemos maximo de maximos
 D[j]= max(D1,D2)

 if( j %% 10000 == 0) print(j)
}
```

```
## [1] 10000
## [1] 20000
## [1] 30000
## [1] 40000
## [1] 50000
## [1] 60000
## [1] 70000
## [1] 80000
## [1] 90000
## [1] 100000

#Grafica el  histograma que aproxima la distribucin KS
hist(D,freq=FALSE,breaks=100, main="Distribucion KS")
lines(density(D))
```



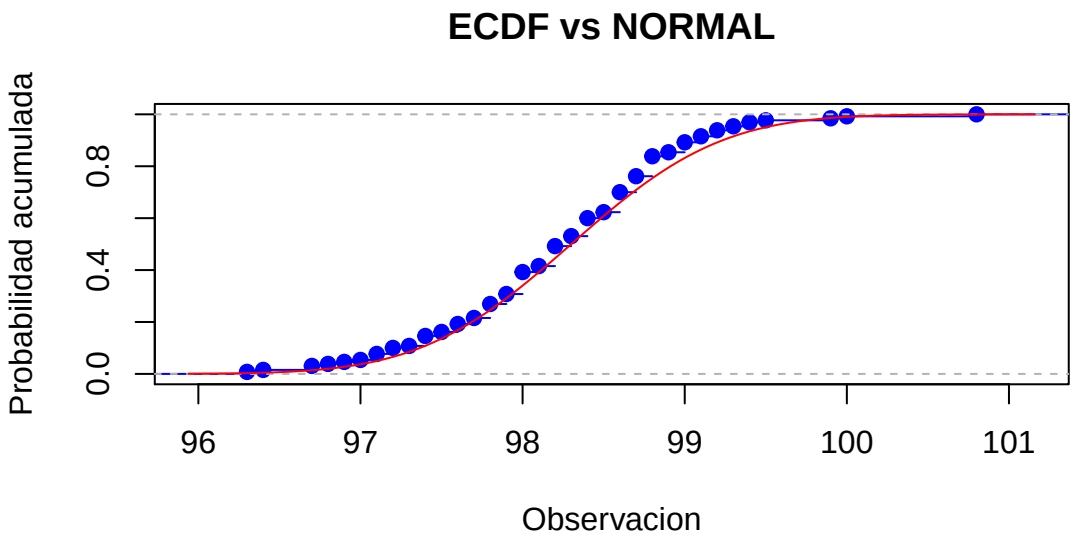

```
#Cuantiles, comparar con los cuantiles obtenidos en Conover
round(quantile(D,c(0.80,0.90,0.95,0.98,0.99)),3)

##   80%   90%   95%   98%   99%
## 0.448 0.510 0.563 0.628 0.671
```

Veamos un ejemplo de como correr la prueba dentro del ambiente R, para ello analizaremos unos datos referentes a la temperatura del cuerpo humano almacenadas en la siguiente dirección (http://venus.unive.it/romanaz/statistics/data/bodytemp.txt). Suponga que se desea ajustar una distribución normal de parámetros $\mu = 98.3$ y $\sigma^2 = 0.53$ al conjunto de datos.

```
temp <- read.table("http://venus.unive.it/romanaz/statistics/data/bodytemp.txt",header = TRUE)
mu <- 98.3
sigma <- 0.53
```

Ahora grafiquemos la distribución empírica.

```
f <- ecdf(temp[,1])
plot(f,xlab="Observacion",main="ECDF vs NORMAL", ylab="Probabilidad acumulada",col=4);
curve(pnorm(x,mu,sqrt(sigma)),add=TRUE,col=2);
```



Como se observa en la figura, la distribución normal propuesta se ajusta muy bien. Realicemos el test KS para probar si los datos realmente provienen de la Normal con los parámetros estimados, es decir nos planteamos la hipótesis

$$H_0 : F_X(x) = N\left(98.3, 0.53\right)$$

$$H_1 : F_X(x) \neq N\left(98.3, 0.53\right)$$

Para correr la prueba, primero, obtenemos el estadístico D_n asociado:

```
#Leemos los datos
x=temp[,1]
#Obtenemos el tamano de muestra
n=length(x)
#Ordenamos la muestra
x=sort(x)
```

```
  #Calculamos la funcion de distribucion empirica
  Fn=ecdf(x)
  #A la muestra ordenada le agregamos el 0 al principio
  #Sirve para el caso F_n(x(0))=0
  y=c(min(x)-1,x)
  #Inicializamos busqueda de supremo
  D1=0
  D2=0
  for (i in 2:(n+1)){
    D1[i]=abs(Fn(y[i])-pnorm(y[i],mu,sqrt(sigma)))
    D2[i]=abs(Fn(y[i-1])-pnorm(y[i],mu,sqrt(sigma)))
  }
  #Obtenemos el estadistico de prueba
  D.n= max(D1,D2)
  D.n

## [1] 0.08456503
```

En este caso el estadístico de prueba para n =130 toma el valor de D_n =0.084565 el cual debe de ser comparado con el cuantil correspondiente de la distribución KS para el caso n =130.

Para no ir a tablas simularemos la distribución y obtendremos el cuantil asociado.

```
#########################################################
# Programa que simula lo Distribucion de Kolmogorov #
#########################################################

#Tama\~no de muestra en nuestro ejemplo
n=130
##############
m=100000    #Numero de simulaciones
##############
#Inicializamos variable que guardara las simulaciones
D=rep(0,m)
for (j in 1:m){
 #Simulacion de las variables uniformes
 x=runif(n,0,1)
 #Se ordena la muestra
 x=sort(x)
 #Calculamos la funcion de distribucion empirica
 Fn=ecdf(x)
 #A la muestra ordenada le agregamos el 0 al principio
 #Sirve para el caso F_n(x(0))=0
```

```
 y=c(0,x)
 #Inicializamos busqueda de supremo
 D1=0
 D2=0
 for (i in 2:(n+1)){
   D1[i]=abs(Fn(y[i])-y[i])
   D2[i]=abs(Fn(y[i-1])-y[i])
 }
 #Obtenemos maximo de maximos
 D[j]= max(D1,D2)
}
#Cuantiles, comparar con los cuantiles obtenidos en Conover
round(quantile(D,c(0.80,0.90,0.95,0.98,0.99)),3)

##   80%   90%   95%   98%   99%
## 0.093 0.106 0.118 0.131 0.141
```

El cuantil al 95% de la distribución para el caso $n = 130$ es $w_{0.95}$ =0.1177524, entonces, como D_n =0.084565<0.1177524 se decide no Rechazar H_0 y por tanto podemos decir que el modelo N(98.3,0.53) es una buena aproximación para la verdadera distribución de estos datos.

La forma mas comoda de correr la prueba en R es mediante la función *ks.test*:

```
ks.test(temp[,1],pnorm,mean=mu,sd=sqrt(sigma), exact = TRUE)

## Warning in ks.test(temp[, 1], pnorm, mean = mu, sd = sqrt(sigma), exact = TRUE): ties should
not be present for the Kolmogorov-Smirnov test

##
##  One-sample Kolmogorov-Smirnov test
##
## data:  temp[, 1]
## D = 0.084565, p-value = 0.2936
## alternative hypothesis: two-sided
```

En este caso el p-value que nos reporta es mayor al 0.05 lo que nos lleva a la misma conclusión que ya sabíamos de no rechazar H_0.

1.2.1 Prueba KS de una cola

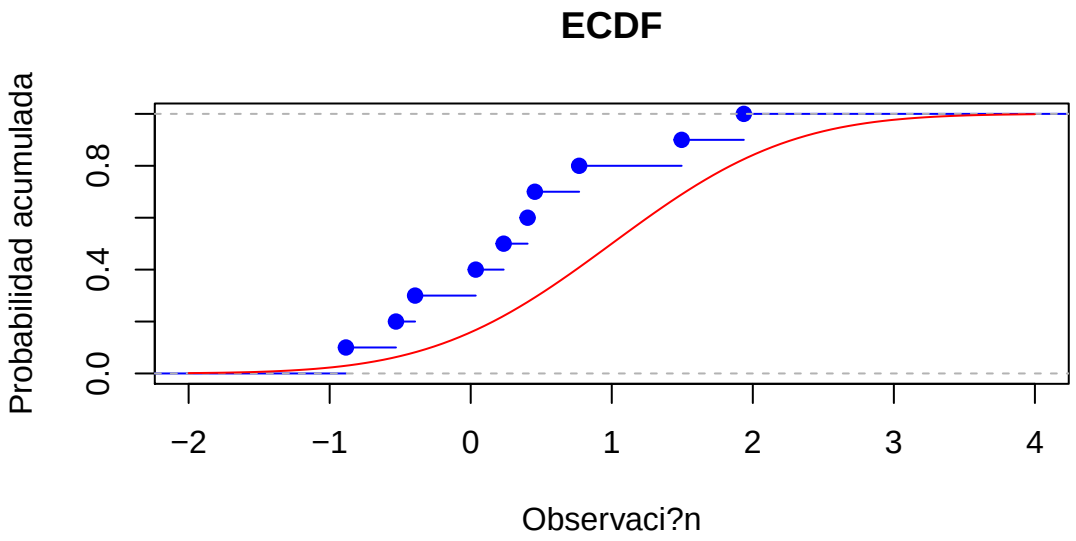
Una de las ventajas que tiene la prueba KS es que se puede definir pruebas de una sola cola, es decir plantarse la hipótesis de la siguiente forma:

$$H_0 : F_X(x) \leq F_X^*(x)$$

$$H_1 : F_X(x) > F_X^*(x)$$

En cuyo caso la estadística de prueba cambia de tal forma que sólo toma en cuenta la diferencia en un sentido sin tomar valor absoluto. En este caso rechazar H_0 quere decir que la distribución verdader esta siempre por arriba de la distribución propuesta F_X^*. Por ejemplo:

```
x<-rnorm(10,0,1);
f<-ecdf(x)
plot(f,xlab="Observaci?n",xlim=c(-2,4),main="ECDF",ylab="Probabilidad acumulada",col=4);
curve(pnorm(x,1,1),add=TRUE,col=2)
```



En este caso se observa que la empírica siempre esta por arriba de la distribución propuesta por lo que se debería rechazar H_0.

Suena entonces de forma natural proponer como estadística de prueba a:

$$D_n^+ = \sup_{x \in \mathbb{R}} \left(F_n(x) - F_X^*(x) \right)$$

En cuyo caso nuevamente hay que obtener la distribución de D_n^+ bajo H_0 por medio de simula-

ciones y asi encontrar el cuantil asociado y rechazar H_0 cuando $D_n^+ > w_{1-\alpha}$

De forma analoga, si nos planteamos la prueba

$$H_0 : F_X(x) \geq F_X^*(x)$$

$$H_1 : F_X(x) < F_X^*(x)$$

Ahora definiremos como estadística de prueba a:

$$D_n^- = \sup_{x \in \mathbb{R}} \left(F_X^*(x) - F_n(x) \right)$$

En cuyo caso se recahzara H_0 si $D_n^- > w_{1-\alpha}$ con $w_{1-\alpha}$ el cuantil $1-\alpha$ asociado a la distribución D_n^- bajo H_0.

1.2.2 Bandas de Confianza para $F_X(x)$

Una de las grandes ventajas que tenemos al conocer la distribución de $D_n = \sup_{x \in \mathbb{R}} |F_n(x) - F_X(x)|$ es que nos puede ayudar a encontrar bandas de confianza para la distribución verdadera $F_X(x)$, para ello denotemos a $w_{1-\alpha}$ como el cuantil $1-\alpha$ de la distribución D_n. Entonces:

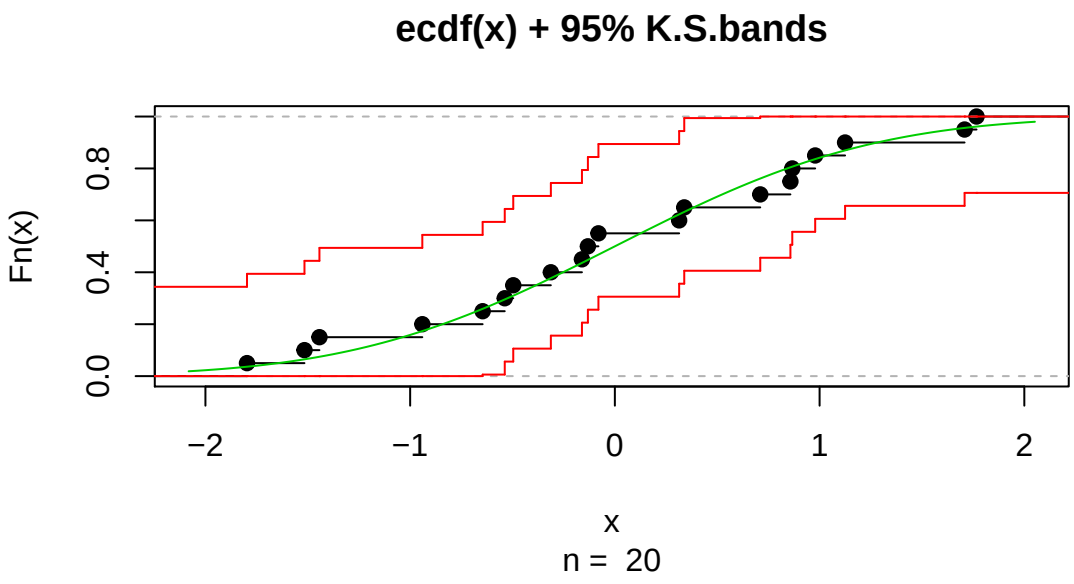
$$\begin{aligned}
&\mathbb{P}\left(D_n \leq w_{1-\alpha}\right) = 1-\alpha \\
&\mathbb{P}\left(\sup_{x \in \mathbb{R}} |F_n(x) - F_X(x)| \leq w_{1-\alpha}\right) = 1-\alpha \\
&\mathbb{P}\left(|F_n(x) - F_X(x)| \leq w_{1-\alpha} \quad \forall x \in \mathbb{R}\right) = 1-\alpha \\
&\mathbb{P}\left(-w_{1-\alpha} \leq F_n(x) - F_X(x) \leq w_{1-\alpha} \quad \forall x \in \mathbb{R}\right) = 1-\alpha \\
&\mathbb{P}\left(F_n(x) - w_{1-\alpha} \leq F_X(x) \leq F_n(x) + w_{1-\alpha} \quad \forall x \in \mathbb{R}\right) = 1-\alpha
\end{aligned}$$

Por lo tanto $F_n(x) - w_{1-\alpha}$ y $F_n(x) + w_{1-\alpha}$ forman una banda de confianza para $F_X(x)$.

Afortunadamente este calculo ya está programado en R dentro de la paqueteria *NSM3* en la función *ecdf.ks.CI* , como ejemplo tenemos lo siguiente:

```
library(NSM3)

## Warning: package 'NSM3' was built under R version 3.2.5
## Loading required package: combinat
##
## Attaching package: 'combinat'
```

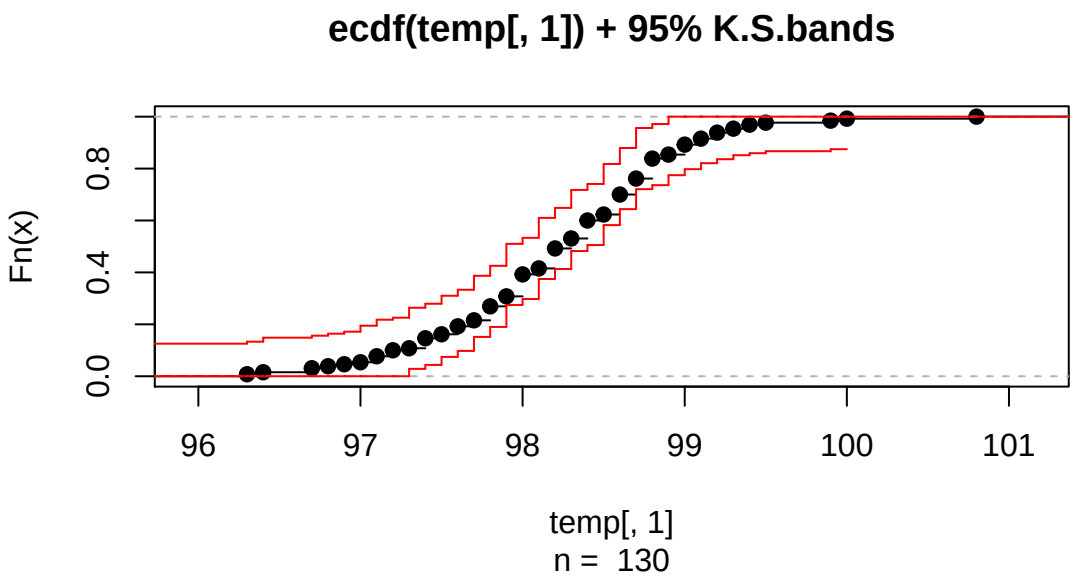
```
## The following object is masked from 'package:utils':
##
##     combn
## Loading required package:  MASS
## Loading required package:  partitions
## Loading required package:  survival
## fANCOVA 0.5-1 loaded

#Simulamos 20 observaciones una normal estandar
x<-rnorm(20,0,1)
#Consturimos las bandas de confianza
ecdf.ks.CI(x)

## $lower
##  [1] 0.00000 0.00000 0.00000 0.00000 0.00000 0.00592 0.05592 0.10592
##  [9] 0.15592 0.20592 0.25592 0.30592 0.35592 0.40592 0.45592 0.50592
## [17] 0.55592 0.60592 0.65592 0.70592
##
## $upper
##  [1] 0.34408 0.39408 0.44408 0.49408 0.54408 0.59408 0.64408 0.69408
##  [9] 0.74408 0.79408 0.84408 0.89408 0.94408 0.99408 1.00000 1.00000
## [17] 1.00000 1.00000 1.00000 1.00000

#Agregamos la curva teorica de donde vino la muestra
curve(pnorm(x,0,1),add=TRUE,col=3)
```



Para el caso del ejemplo de temperaturas del cuerpo humano almacenadas en la siguiente di-

rección (http://venus.unive.it/romanaz/statistics/data/bodytemp.txt). Las bandas de confianza son:

```
temp <- read.table("http://venus.unive.it/romanaz/statistics/data/bodytemp.txt",header = TRUE)
ecdf.ks.CI(temp[,1])
```




```
## $lower
##  [1] 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000 0.00000000
##  [7] 0.00000000 0.00000000 0.00000000 0.02838846 0.04377308 0.07454231
## [13] 0.09761923 0.15146538 0.18992692 0.27454231 0.29761923 0.37454231
## [19] 0.41300384 0.48223461 0.50531154 0.58223461 0.64377308 0.72069615
## [25] 0.73608077 0.77454231 0.79761923 0.82069615 0.83608077 0.85146538
## [31] 0.85915769 0.86685000 0.87454231 0.88223461
##
## $upper
##  [1] 0.1254577 0.1331500 0.1485346 0.1562269 0.1639192 0.1716115 0.1946885
##  [8] 0.2177654 0.2254577 0.2639192 0.2793038 0.3100731 0.3331500 0.3869962
## [15] 0.4254577 0.5100731 0.5331500 0.6100731 0.6485346 0.7177654 0.7408423
## [22] 0.8177654 0.8793038 0.9562269 0.9716115 1.0000000 1.0000000 1.0000000
## [29] 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
```

1.3 Prueba Lilliefors

Una de las principales limitantes de la prueba KS es que requiere tener completamente especificada la distribución bajo H_0 pues esto se requiere para calcular el estadístico de prueba corre-

spondiente, es decir la prueba KS no puede llevarse a cabo cuando se desconoce algún parámetro de la distribución bajo la hipótesis nula.

Por ejemplo, supongamos que tenemos $X_1, \ldots, X_n$ m.a. de $F_X(x)$ desconocida y queremos verificar si la muestra proviene del modelo normal $N(\mu, \sigma^2)$ con ambos parámetros desconocidos, en terminos de hipótesis plantearíamos el problema:

$$H_0 : F_X(x) = N\left(\mu, \sigma^2\right)$$

$$H_1 : F_X(x) \neq N\left(\mu, \sigma^2\right)$$

Hemos visto ya que la prueba χ^2 puede adaptarse para atacar este problema por medio de la resta de los grados de libertad asociados, en este caso ahora se presenta una modificación a la prueba KS para poder atacar este problema (sólo para el caso normal).

Lilliefors modifica la prueba KS para poder resolver el problema de parámetros desconocidos en el caso normal, él método que el propone es el siguiente:

Método Lilliefors para Normalidad (ambos parámetros deconocidos)

1. Dada $x_1, \ldots, x_n$ de $F_X(x)$, estimar los parámetros μ y σ^2 es decir:

$$\begin{aligned} \hat{\mu} &= \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \\ \hat{\sigma}^2 &= \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \overline{x})^2 \end{aligned}$$

2. Transformar la muestra $x_1, \ldots, x_n$ por medio de la siguiente relación:

$$z_i = \frac{x_i - \hat{\mu}}{\hat{\sigma}}$$

3. Con la muestra transformada $z_1, \ldots, z_n$, llevar a cabo la prueba KS para contrastar la hipótesis

$$H_0 : F_Z(z) = N(0, 1)$$

$$H_1 : F_Z(z) \neq N(0, 1)$$

Observe que en este caso ya la distribución está complentamente especificada bajo H_0, por lo que la prueba KS puede llevarse a cabo sin ningún problema y se puede calcular entonces el estadístico de prueba:

$$D_n = \sup_{z \in \mathbb{R}} |F_n(z) - F_Z^*(z)|$$

Donde en este caso $F_Z^*(z) = \Phi(z)$ corresponde a la distribución normal estándar.

4. Nuevamente, la idea es rechazar H_0 si D_n es grande por lo que se debe de comparar contra los cuantiles de la distribución teórica de D_n bajo H_0, es decir, se rechaza H_0 a un nivel de significancia α si

$$D_n > w_{1-\alpha}$$

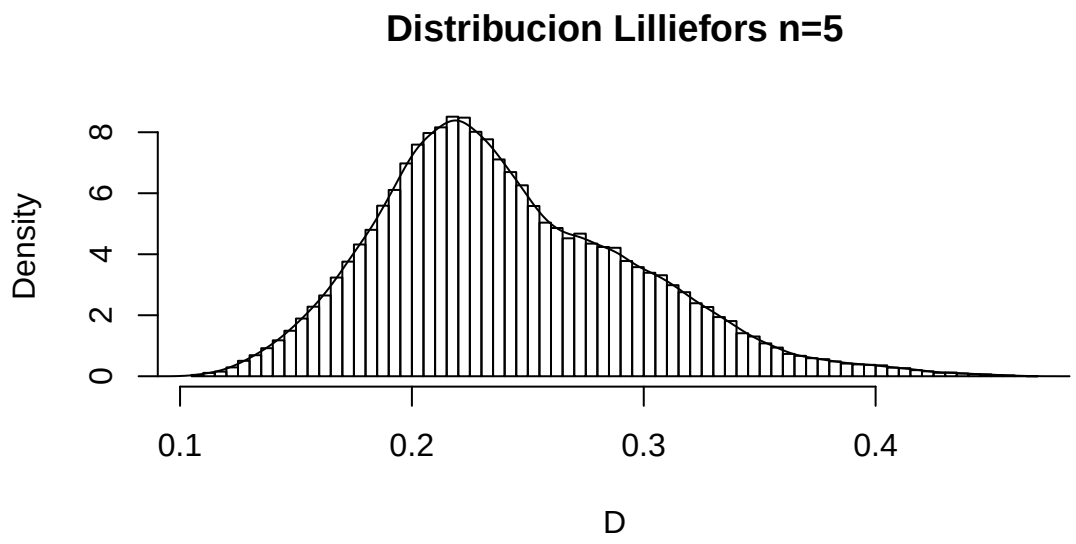
Donde $w_{1-\alpha}$ es un cuantil buscado en tablas de la distribución Lilliefors (Tabla 15 del Libro de Conover), la cual puede ser simulada dentro del paquete R

El gran aporte que hace Lilliefors, es que prueba que la distribución D_n no depende de los parámetros desconocidos μ y σ^2 y sólo depende de n.

A continuación se presenta un código en R para simular la distribución de Lilliefors para una determinada n, observe que en este caso se simula de una normal de parametros $\mu = 0$, $\sigma^2 = 1$, un ejercicio interesante es volver a correr este mismo programa simulando de otra Normal (con otros parámetros) y notar que la distribución D_n sigue siendo la misma.

```
#######################################################
#Programa que simula lo Distribucion de Lilliefors  #
#######################################################

#Seleccione el tama\~no de muestra
n=5
#################################################################
# Libro de Conover Tabla 15   1971                             #
#Cuantil        0.80    0.85    0.90    0.95    0.99          #
#n=5            0.285   0.299   0.315   0.337   0.405         #
#################################################################

#################################################################
# Tabla Revisada en 1987                                       #
#Cuantil        0.80    0.85    0.90    0.95    0.99          #
#n=5            0.289   0.303   0.319   0.343   0.396         #
#################################################################



#############
m=100000    #Numero de simulaciones
#############

#Inicializamos variable que guardara las simulaciones
```

```r
D=rep(0,m)
for(j in 1:m){
 #Simulamos de la distribucion normal 0 1
 x=rnorm(n,0,1)
 #Transformamos la muestra y aplicamos la prueba KS
 mu=mean(x)
 sigma.2=var(x)
 z=(x-mu)/sqrt(sigma.2)
 #Ordenamos la muestra
 z=sort(z)
 #Calculamos la funcion de distribucion empirica
 Fn=ecdf(z)
 #A la muestra ordenada le agregamos un nuevo minimo al principio
 #Sirve para tener definido el  caso F_n(z(0))=0
 y=c(min(z)-1,z)
 #Inicializamos busqueda de supremo
 D1=0
 D2=0
 for (i in 2:(n+1)){
   D1[i]=abs(Fn(y[i])-pnorm(y[i],0,1))
   D2[i]=abs(Fn(y[i-1])-pnorm(y[i],0,1))
 }

 D[j]=max(D1,D2)

}
#Grafica el  histograma que aproxima la distribucin Lilliefors
hist(D,freq=FALSE,breaks=100, main="Distribucion Lilliefors n=5")
lines(density(D))
```




```
#Cuantiles, comparar con los cuantiles obtenidos en Conover en 1971 y en 1987
round(quantile(D,c(0.80,0.85,0.90,0.95,0.99)),3)

##   80%   85%   90%   95%   99%
## 0.289 0.302 0.318 0.343 0.397
```

Vemos un ejemplo con los datos de la temperatura del cuerpo, ahora nos interesa probar si los datos son normales (sin especificar sus parámetros).

Leemos los datos y estimamos parámetros

```
temp <- read.table("http://venus.unive.it/romanaz/statistics/data/bodytemp.txt",header = TRUE)
x <- temp[,1]
n <-  length(x)
mu <- mean(x)
sigma.2 <- var(x)
```

Tranformamos la muestra y estimamos D_n:

```
  z=(x-mu)/sqrt(sigma.2)
  #Ordenamos la muestra
  z=sort(z)
  #Calculamos la funcion de distribucion empirica
  Fn=ecdf(z)
  #A la muestra ordenada le agregamos un nuevo minimo al principio
  #Sirve para tener definido el  caso F_n(z(0))=0
  y=c(min(z)-1,z)
```

```r
  #Inicializamos busqueda de supremo
  D1=0
  D2=0
  for (i in 2:(n+1)){
    D1[i]=abs(Fn(y[i])-pnorm(y[i],0,1))
    D2[i]=abs(Fn(y[i-1])-pnorm(y[i],0,1))
  }

  D=max(D1,D2)
  D
```

```
## [1] 0.06472685
```

El estadístico de Lilliefors asociado a esa muestra es D_n=0.0647269 el cual debe ser comparado con el cuantil de distribución Lilliefors para el caso $n = 130$. En tablas de conover se observa que el cuantil asociado al 95% es $w_{0.95} = 0.07595322$, y dado que $0.0647269 < w_{0.95}$, entonces **No se rechaza** H_0, se concluye entonces que el modelo Normal es una buena aproximación para la verdadera distribución de los datos.

Esta misma pueba se puede correr por medio de la función *lilllie.test* de la librearia *nortest*

```r
 library(nortest)
 lillie.test(x)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  x
## D = 0.064727, p-value = 0.2009
```

Donde el p-value mayor al 0.05 nos garantiza que no se rechaza H_0.

Una observación interesante es que esta idea de Lilliefors puede extenderse al caso en donde sólo uno de los dos parámetros es desconocido, por ejemplo, supongamos que estamos interesados ahora en probar lo siguiente:

$$H_0 : F_X(x) = N\left(0, \sigma^2\right)$$

$$H_1 : F_X(x) \neq N\left(0, \sigma^2\right)$$

Donde se observa que la media es conocida, en este caso la prueba de Lilliefors se adapta de la siguiente manera:

Método Lilliefors para Normalidad (varianza deconocida y media conocida)

1. Dada $x_1, \ldots, x_n$ de $F_X(x)$, estimar el parámetro σ^2 es decir:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_0)^2$$

2. Transformar la muestra $x_1, \ldots, x_n$ por medio de la siguiente relación:

$$z_i = \frac{x_i - \mu_0}{\hat{\sigma}}$$

   Donde μ_0 es la media conocida bajo H_0.

3. Con la muestra transformada $z_1, \ldots, z_n$, llevar a cabo la prueba KS para contrastar la hipótesis

$$H_0 : F_Z(z) = N(0,1)$$

$$H_1 : F_Z(z) \neq N(0,1)$$

   Observe que en este caso ya la distribución está complentamente especificada bajo H_0, por lo que la prueba KS puede llevarse a cabo sin ningún problema y se puede calcular entonces el estadístico de prueba:

$$D_n = \sup_{z\in\mathbb{R}} |F_n(z) - F_Z^*(z)|$$

   En este caso $F_Z^*(z) = \Phi(z)$ corresponde a la distribución normal estándar.

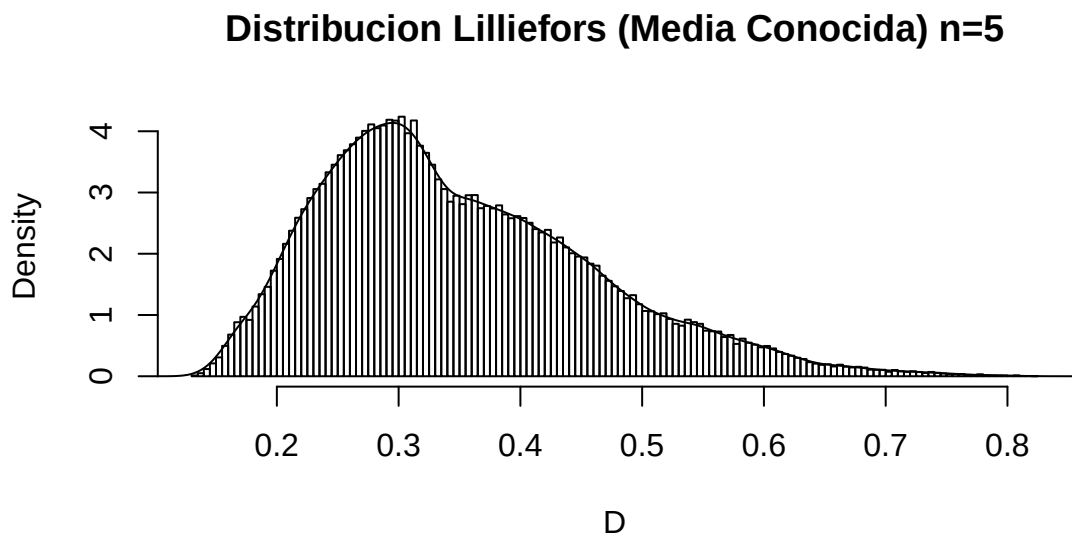
4. Nuevamente, la idea es rechazar H_0 si D_n es grande por lo que se debe de comparar contra los cuantiles de la distribución teórica de D_n bajo H_0, es decir, se rechaza H_0 a un nivel de significancia α si

$$D_n > w_{1-\alpha}$$

   Donde $w_{1-\alpha}$ es el cuantil de la distribución Lilliefors para el caso de Media Conocida y Varianza Desconocida, dicha distribución casi no se encuentra tabulada en los libros sin embargo puede simularse por medio de un programa en R el cual se presenta a continuación para el caso $n = 5$.

```
#########################################################################
#Programa que simula lo Distribucion de Lilliefors (Media Conocida)  #
#########################################################################

#Seleccione el tama\~no de muestra
n=5
```

```r
#############
m=100000    #Numero de simulaciones
#############

#Inicializamos variable que guardara las simulaciones
D=rep(0,m)
for(j in 1:m){
 #Simulamos de la distribucion normal 0 1
 x=rnorm(n,0,1)
 #Transformamos la muestra y aplicamos la prueba KS
 mu=0  #Media conocida
 sigma.2=1/n*sum((x-mu)^2) # Estimacion de la varianza con media conocida
 z=(x-mu)/sqrt(sigma.2)
 #Ordenamos la muestra
 z=sort(z)
 #Calculamos la funcion de distribucion empirica
 Fn=ecdf(z)
 #A la muestra ordenada le agregamos un nuevo minimo al principio
 #Sirve para tener definido el  caso F_n(z(0))=0
 y=c(min(z)-1,z)
 #Inicializamos busqueda de supremo
 D1=0
 D2=0
 for (i in 2:(n+1)){
   D1[i]=abs(Fn(y[i])-pnorm(y[i],0,1))
   D2[i]=abs(Fn(y[i-1])-pnorm(y[i],0,1))
 }

 D[j]=max(D1,D2)

}
#Grafica el  histograma que aproxima la distribucion Lilliefors Media Conocida
hist(D,freq=FALSE,breaks=100, main="Distribucion Lilliefors (Media Conocida) n=5")
lines(density(D))
```




```
#Cuantiles
round(quantile(D,c(0.80,0.85,0.90,0.95,0.99)),3)

##   80%   85%   90%   95%   99%
## 0.442 0.469 0.506 0.563 0.662
```

Método Lilliefors para Normalidad (varianza conocida y media desconocida)

1. Dada $x_1, \ldots, x_n$ de $F_X(x)$, estimar el parámetro μ es decir:

$$\hat{\mu} \quad = \quad \overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

2. Transformar la muestra $x_1, \ldots, x_n$ por medio de la siguiente relación:

$$z_i = \frac{x_i - \hat{\mu}}{\sigma_0}$$

   Donde σ_0^2 es la varianza conocida bajo H_0, entonces σ_0 es la desviación estándar.

3. Con la muestra transformada $z_1, \ldots, z_n$, llevar a cabo la prueba KS para contrastar la hipótesis

$$H_0 : F_Z(z) = N(0,1)$$

$$H_1 : F_Z(z) \neq N(0,1)$$

Observe que en este caso ya la distribución está complentamente especificada bajo H_0, por lo que la prueba KS puede llevarse a cabo sin ningún problema y se puede calcular entonces el estadístico de prueba:

$$D_n = \sup_{z \in \mathbb{R}} |F_n(z) - F_Z^*(z)|$$

Donde en este caso $F_Z^*(z) = \Phi(z)$ corresponde a la distribución normal estándar.

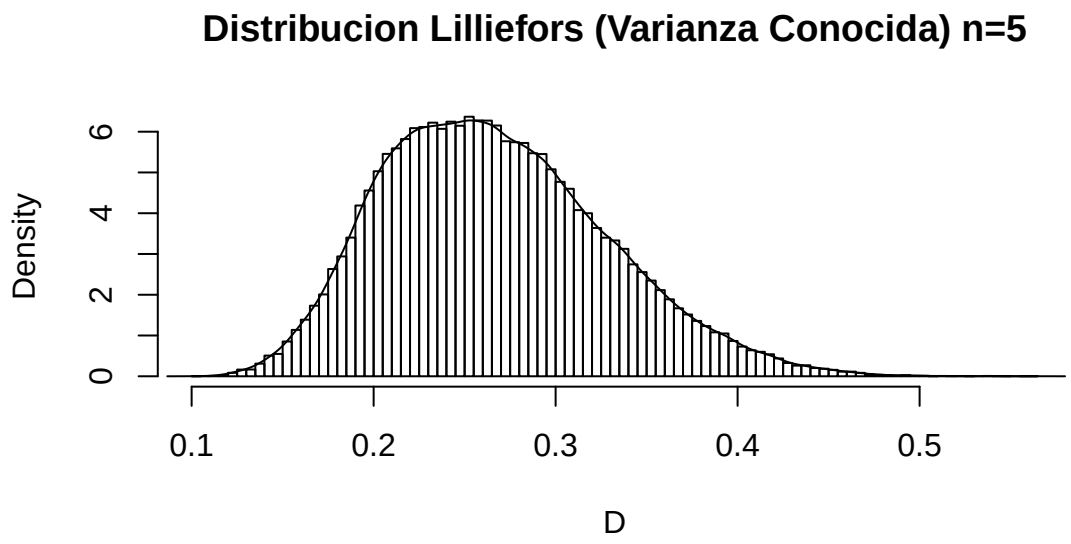
4. Nuevamente, la idea es rechazar H_0 si D_n es grande por lo que se debe de comparar contra los cuantiles de la distribución teórica de D_n bajo H_0, es decir, se rechaza H_0 a un nivel de significancia α si

$$D_n > w_{1-\alpha}$$

Donde $w_{1-\alpha}$ es el cuantil de la distribución Lilliefors para el caso de Varianza Conocida y Media Desconocida, dicha distribución puede simularse en R, a continuación presentamos el programa correspondiente para el caso $n = 5$

```
#########################################################################
#Programa que simula lo Distribucion de Lilliefors (Media Conocida)  #
#########################################################################

#Seleccione el tama\~no de muestra
n=5
#############
m=100000    #Numero de simulaciones
#############

#Inicializamos variable que guardara las simulaciones
D=rep(0,m)
for(j in 1:m){
 #Simulamos de la distribucion normal 0 1
 x=rnorm(n,0,1)
 #Transformamos la muestra y aplicamos la prueba KS
 mu=mean(x)  #Estimacion de la media
 sigma.2=1   #Varianza Conocida
 z=(x-mu)/sqrt(sigma.2)
 #Ordenamos la muestra
 z=sort(z)
 #Calculamos la funcion de distribucion empirica
 Fn=ecdf(z)
 #A la muestra ordenada le agregamos un nuevo minimo al principio
 #Sirve para tener definido el  caso F_n(z(0))=0
```

```r
 y=c(min(z)-1,z)
 #Inicializamos busqueda de supremo
 D1=0
 D2=0
 for (i in 2:(n+1)){
   D1[i]=abs(Fn(y[i])-pnorm(y[i],0,1))
   D2[i]=abs(Fn(y[i-1])-pnorm(y[i],0,1))
 }

 D[j]=max(D1,D2)

}
#Grafica el  histograma que aproxima la distribucion Lilliefors Media Conocida
hist(D,freq=FALSE,breaks=100, main="Distribucion Lilliefors (Varianza Conocida) n=5")
lines(density(D))
```




```r
#Cuantiles
round(quantile(D,c(0.80,0.85,0.90,0.95,0.99)),3)

##   80%   85%   90%   95%   99%
## 0.318 0.333 0.350 0.377 0.424
```

1.4 Pruebas QEDF

La pruebas **QEDF** (Quadratic Empirical Distribution Function) se basan en la misma idea que la prueba KS, es decir, cuantificar una distancia entre la distribución empírica obtenida con la muestra y la distribución $F_X^*(x)$ propuesta bajo la hipótesis nula.

En este caso, supongamos nuevamente que deseamos contrastar las hipótesis

$$H_0 : F_X(x) = F_X^*(x)$$

$$H_1 : F_X(x) \neq F_X^*(x)$$

Donde $F_X^*(x)$ supondremos es una distribución completamente conocida.

La idea de las estadística QEDF es cuantificar la discrepancia al cuadrado y luego integrar respecto a la distribución propuesta bajo H_0.

Definición 1.4.1 (QEDF Statistic)**.** *Definimos la estadística QEDF para probar bondad de ajuste como:*

$$Q_n = n \int_{-\infty}^{\infty} \left(F_n(x) - F_X^*(x)\right)^2 \psi(x) dF_X^*$$

Donde $F_n(x)$ es la distribución empírica obtenida con la muestra y $F_X^*(x)$ la distribución (completamente especificada) bajo H_0. Por otro lado $\psi(x)$ es una función ponderadora que nos ayuda a darle peso a la diferencia cuadrática en cada x.

Lo interesante de esta estadística de prueba es que bajo H_0 y suponiendo que $\psi(x)$ es una función que depende de $F_X^*(x)$, es decir $\psi(x) = g\left(F_X^*(x)\right)$, entonces se tiene una distribución que sólo depende de n, para probar lo anterior basta con suponer que $F_X^*(x)$ es continua y aplicar el siguiente cambio de variable a la integral:

$$u = F_X^*(x) \quad \Rightarrow \quad x = F_X^{*^{-1}}(u) \quad \Rightarrow \quad du = dF_X^*$$

Llevando a cabo el cambio de variable obtenemos que:

$$Q_n = n \int_{-\infty}^{\infty} \left(F_n(x) - F_X^*(x)\right)^2 \psi(x) dF_X^* = n \int_0^1 \left(F_n(F_X^{*^{-1}}(u)) - u\right)^2 \psi(F_X^{*^{-1}}(u)) du$$

Pero por la condición que hemos impuesto $\psi(x) = g\left(F_X^*(x)\right)$ se tiene que $\psi(F_X^{*^{-1}}(u)) = g\left(F_X^*(F_X^{*^{-1}}(u))\right) = g(u)$ por lo tanto la integral queda como:

$$Q_n = n \int_0^1 \left(F_n(F_X^{*^{-1}}(u)) - u\right)^2 g(u) du$$

En este punto pareciera que Q_n sigue dependiente de $F_X^*(x)$,sin embargo veremos $F_n(F_X^{*^{-1}}(u))$ ya no depende de $F_X^*(x)$ pues no es mas que la distribución empírica de una muestra transformada, para ver esto recordemos que por definición de distribución empírica se tiene que:

$$F_n\left(F_X^{*^{-1}}(u)\right) = \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{1}_{\left(X_i \leq F_X^{*^{-1}}(u)\right)}\right) = \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{1}_{(F_X^*(X_i)\leq u)}\right) = \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{1}_{(U_i\leq u)}\right)$$

Donde $U_i = F_X^*(X_i)$ ya no depende de $F_X^*(x)$ pues tiene distribución uniforme(0,1). Es decir se prueba entonces que $F_n\left(F_X^{*^{-1}}(u)\right)$ no es mas que la distribución empírica de la muestra transformada $U_1, \ldots, U_n$. Por lo anterior se prueba entonces que Q_n no dependerá de $F^*(x)$ y sólo ependerá de n.

1.4.1 Prueba Anderson Darling

Una de las pruebas más importantes que se obtienen de esta idea es la denominada *Prueba Anderson Darling* la cual se obtiene haciendo que la función de ponderación tome la siguiente forma:

$$\psi(x) = \frac{1}{F_X^*(x)\left(1 - F_X^*(x)\right)}$$

La idea detrás de esta ponderación es que hace que las colas de la distribución tengan pesos más grandes, de esta forma, la prueba es capaz de detectar diferencias en las colas de la distribución más facilmente, es por lo anterior que esta es una de las pruebas más potentes para detectar normalidad en especial cuando la muestra es generada de distribuciones t-student.

En este caso el estadístico de prueba toma la siguiente forma:

$$A_n^2 = Q_n^2 = n\int_{-\infty}^{\infty}\left(F_n(x) - F_X^*(x)\right)^2 \frac{1}{F_X^*(x)\left(1 - F_X^*(x)\right)} dF_X^*$$

El cual, al hacer el cambio de variable correspondiente ($u = F_X^*(x)$) se transforma en la siguiente expresión:

$$A_n^2 = n\int_0^1 \left(F_n(u) - u\right)^2 \frac{1}{u\left(1-u\right)} du$$

Donde hay que recordar que $F_n(u)$ corresponde a la distribución empírica generada por la muestra tranformada $u_1, \ldots, u_n$ Dicha integral puede ser resuelta integrando por pedazos ya que $F_n(u)$ es constante en el intervalo $\left(u_{(i)}, u_{(i+1)}\right)$. (Recordemos que $u_{(1)} \ldots, u_{(n)}$ es la muestra ordenada). Entonces:

$$A_n^2 = n\left(\int_0^{u_{(1)}} \frac{u}{1-u} du + \sum_{i=1}^{n-1}\int_{u_{(i)}}^{u_{(i+1)}} \frac{\left(\frac{i}{n} - u\right)^2}{u(1-u)} du + \int_{u_{(n)}}^{1} \frac{(1-u)}{u} du\right) \tag{1.4}$$

Estas integrales se resuelven recordando que:

$$\begin{aligned}\int \frac{u}{1-u}du &= -u - \log(1-u) \\ \int \frac{(a-u)^2}{u(1-u)}du &= a^2\log(u) - (a-1)^2\log(1-u) - u \\ \int \frac{(1-u)}{u}du &= \log(u) - u\end{aligned}$$

Con lo anterior se prueba que (1.4) se tranforma en:

$$A_n^2 = -n - \sum_{i=1}^{n}\left(\frac{2i-1}{n}\right)\left(\log(u_{(i)}) + \log(1-u_{(n-i+1)})\right)$$

La cual recordando el cambio de variable $u = F_X^*(x)$ el estadístico de prueba toma la forma:

$$A_n^2 = -n - \sum_{i=1}^{n}\left(\frac{2i-1}{n}\right)\left(\log(F_X^*(x_{(i)})) + \log(1-F_X^*(x_{(n-i+1)}))\right)$$

Ya sabemos que dicho estadístico tiene una distribución que no depende de $F_X^*(x)$ y sólo depende de n, se puede mostrar además que $A_n^2 \to A^2$ donde A^2 se le conoce como la distribución asintótica de Anderson la cual esta tabulada en la mayoría de los libros.

10%	5%	2.5%	1%
1.933	2.492	3.070	3.857

Nuevamente la idea será rechazar H_0 si $A_n^2 > w_{1-\alpha}$, donde $w_{1-\alpha}$ es el cuantil asociado a la distribución de A_n^2 bajo H_0 la cual puede simularse por medio de un programa en R.

A manera de resumen entonces el método propuesto por Anderson-Darling es el siguiente:

1. Se plantea la hipótesis.

$$H_0 : F_X(x) = F_X^*(x)$$

$$H_1 : F_X(x) \neq F_X^*(x)$$

2. Dada $x_1, \ldots, x_n$ m.a. de $F_X(X)$ se transforma la muestra mediante el cambio $u_i = F_X^*(x_i)$, de tal forma que se obtiene la muestra $u_1, \ldots, u_n$

3. Ordenar la muestra de menor a mayor obteniendo la muestra ordenada

$$u_{(1)}, \ldots, u_{(n)}$$

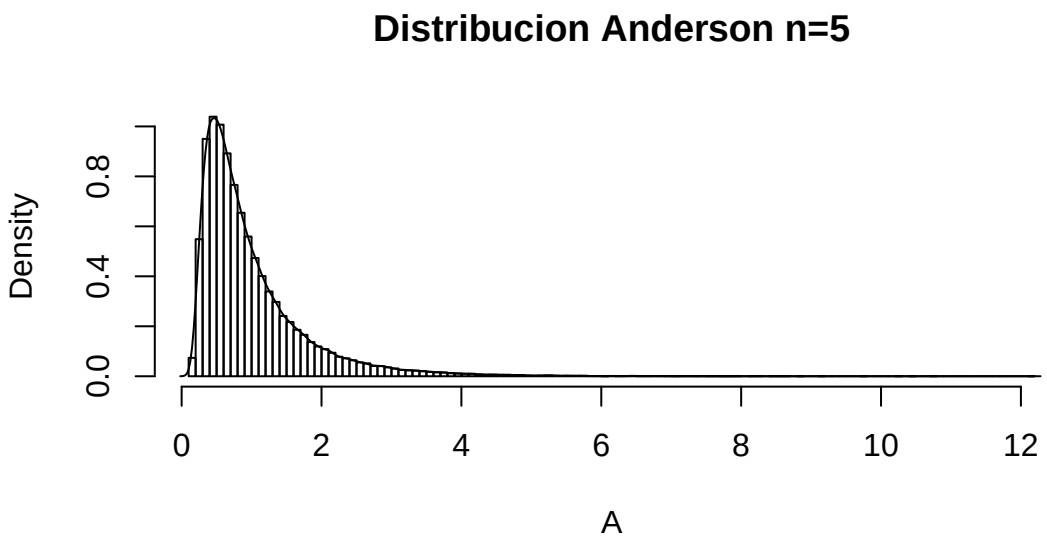
4. Calcular el estadístico de prueba:

$$A_n^2 = -n - \sum_{i=1}^{n} \left(\frac{2i-1}{n} \right) \left(\log\left(u_{(i)}\right) + \log\left(1 - u_{(n-i+1)}\right) \right)$$

5. Rechazar H_0 si $A_n^2 > w_{1-\alpha}$, donde $w_{1-\alpha}$ es el cuantil asociado a la distribución de A_n^2 bajo H_0 la cual puede simularse por medio de un programa en R.

A continuación presentamos el código en R para simular la distribución bajo H_0 para el caso $n = 5$

```
#############################################################
# Programa que simula lo Distribucion de Anderson Darling #
#############################################################
#Numero de simulaciones
m=100000
#Tama\~no de muestra
n=5
#Vector donde guardaremos las simulaciones
A=rep(0,m)
for(i in 1:m){
 #Simulamos de una uniforme(0,1)
 u=runif(n,0,1)
 #Ordenamos la muestra
 ur=sort(u)
 #Construimos estadistico de prueba
 s=0
 for (j in 1:n){
   s=s+(2*j-1)/n*(log(ur[j])+log(1-ur[n-j+1]))
 }
 A[i]=-n-s
}
hist(A,freq=FALSE, breaks=100, main="Distribucion Anderson n=5")
lines(density(A))
```



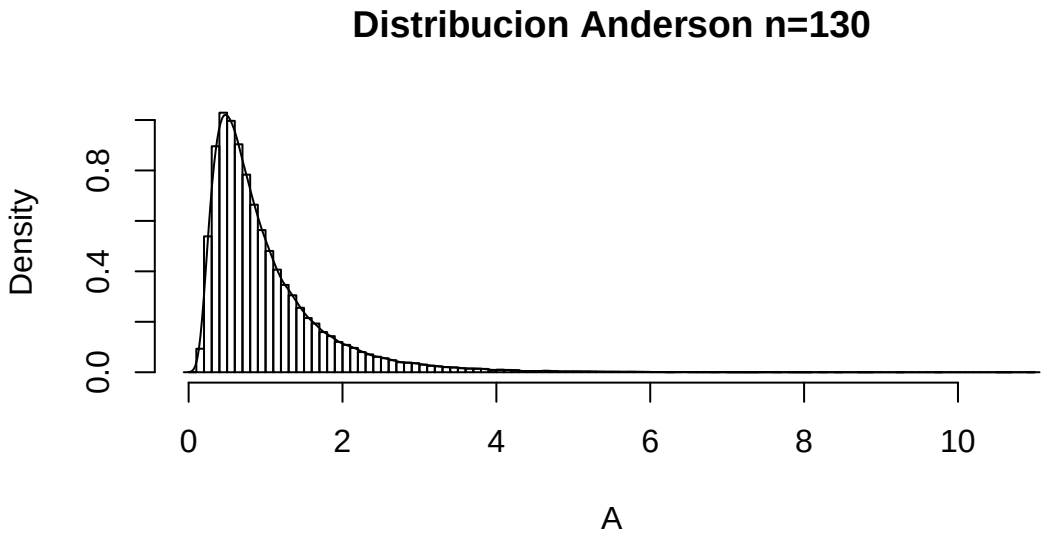

```
round(quantile(A,c(0.900,0.950,0.975,0.990)),3)

##   90%   95% 97.5%   99%
## 1.940 2.525 3.136 3.946
```

Como ejemplo, supongamos nuevamente que queremos probar normalidad con parámetros conocidos $\mu = 98.3$ y $\sigma^2 = 0.53$ a los datos de la temperatura corporal ahora haciendo uso de la prueba Anderson Darling.

```
temp <- read.table("http://venus.unive.it/romanaz/statistics/data/bodytemp.txt",header = TRUE)
x<-temp[,1]
n<-length(x)
mu<-98.3
sigma<-0.53
#Transformamos la muestra
u<-pnorm(x,mu,sqrt(sigma))
ur=sort(u)
#Construimos estadistico de prueba
 s=0
 for (j in 1:n){
   s=s+(2*j-1)/n*(log(ur[j])+log(1-ur[n-j+1]))
 }
 A.t<- -n-s
```

En este caso el estadístico de prueba toma el valor de $A^2 = 0.7082857$, para ver si rechazamos o no la hipótesis tendremos que simular la estadistica de prueba asociada para el caso $n = 130$, a continuación se presenta el histograma y cuantiles de dicha simulación:

```r
############################################################
# Programa que simula lo Distribucion de Anderson Darling #
############################################################
#Numero de simulaciones
m=100000
#Tama\~no de muestra
n=130
#Vector donde guardaremos las simulaciones
A=rep(0,m)
for(i in 1:m){
 #Simulamos de una uniforme(0,1)
 u=runif(n,0,1)
 #Ordenamos la muestra
 ur=sort(u)
 #Construimos estadistico de prueba
 s=0
 for (j in 1:n){
   s=s+(2*j-1)/n*(log(ur[j])+log(1-ur[n-j+1]))
 }
 A[i]=-n-s
}
hist(A,freq=FALSE, breaks=100, main="Distribucion Anderson n=130")
lines(density(A))
```



```r
round(quantile(A,c(0.900,0.950,0.975,0.990)),3)

##   90%   95% 97.5%   99%
## 1.921 2.482 3.072 3.855
```

Con $\alpha = 0.05$ se tiene que el cuantil asociado es 2.482, luego como $A^2 = 0.7082857 < 2.482$ se concluye que no se puede rechazar H_0. Una ventaja que tenemos de haber hecho las simulaciones es que también prodremos aproximar el p-value asociado a la prueba mediante el siguiente código:

```r
p.value<-mean(A>A.t)
p.value

## [1] 0.5476
```

Como el p.value es mayor a 0.05 se concluye también que no se debe rechazar H_0.

Prueba Anderson-Lilliefors

Al igual que en el caso de la prueba KS, la prueba Anderson Darling puede adaptarse al caso de bondad de ajuste para una normal de parámetros desconocidos, la idea es nuevamente estimar los parámetros asociados y llevar a acabo la transformación correspondiente para estandarlizar la normal y luego llevar a cabo la prueba de hipótesis ya con los parámetros conocidos $\mu = 0$ y $\sigma^2 = 1$. El algoritmo quedaría como sigue:

1. Se plantea la hipótesis.

$$H_0 : F_X(x) = N\left(\mu, \sigma^2\right)$$
$$H_1 : F_X(x) \neq N\left(\mu, \sigma^2\right)$$

2. Dada $x_1, \ldots, x_n$ m.a. de $F_X(x)$, estimar parámetros:

$$\begin{aligned} \hat{\mu} &= \frac{1}{n}\sum_{i=1}^{n} x_i \\ \hat{\sigma}^2 &= \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \overline{x}) \end{aligned}$$

3. Transformar la muestra $x_1, \ldots, x_n$ por medio de la siguiente relación:

$$z_i = \frac{x_i - \hat{\mu}}{\hat{\sigma}}$$

4. Transformar la muestra $z_1, \ldots, z_n$ mediante el cambio $u_i = \Phi(z_i)$, donde $\Phi(x)$ es la distribución de una variable aleatoria $N(0,1)$. Con la transformación se genera entonces la muestra $u_1, \ldots, u_n$

5. Ordenar la muestra de menor a mayor obteniendo la muestra ordenada

$$u_{(1)}, \ldots, u_{(n)}$$

6. Calcular el estadístico de prueba:

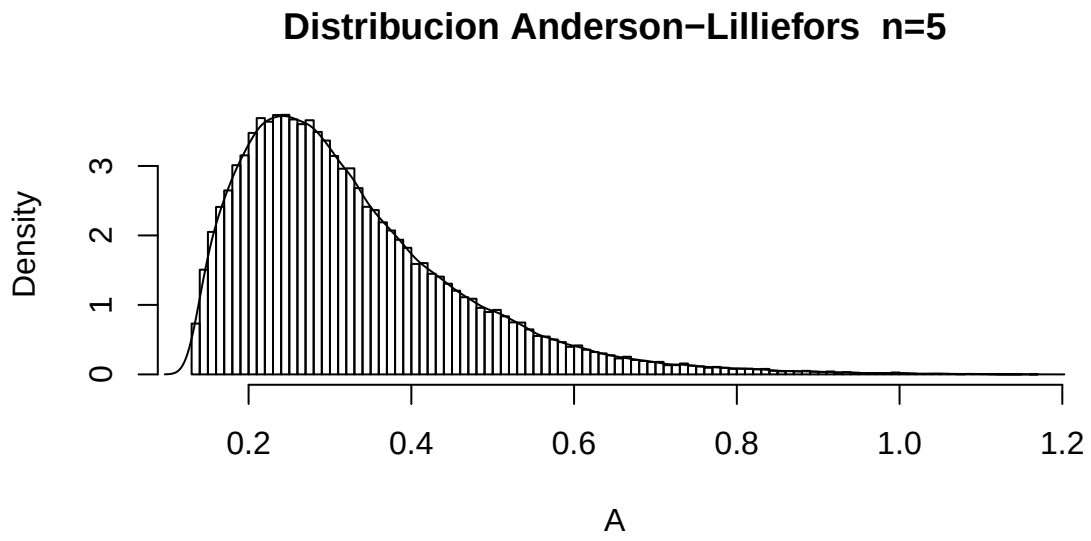
$$A_n^2 = -n - \sum_{i=1}^{n} \left(\frac{2i-1}{n} \right) \left(\log\left(u_{(i)}\right) + \log\left(1 - u_{(n-i+1)}\right) \right)$$

7. Rechazar H_0 si $A_n^2 > w_{1-\alpha}$, donde $w_{1-\alpha}$ es el cuantil asociado a la distribución de A_n^2 bajo H_0 la cual puede simularse por medio de un programa en R.

A continuación presentamos el código en R para simular la distribución bajo H_0 para el caso $n = 5$

```
#############################################################
# Programa que simula lo Distribucion de Anderson Darling #
#############################################################
#Numero de simulaciones
m=100000
#Tama\~no de muestra
n=5
#Vector donde guardaremos las simulaciones
A=rep(0,m)
for(i in 1:m){
 #Simulamos de una uniforme(0,1)
 x=rnorm(n,0,1)
 #Estimamos parametros
 mu<-mean(x)
 sigma<-var(x)
 #Transformamos muestra
 z<-(x-mu)/sqrt(sigma)

 #Transformamo nuevamente
 u<-pnorm(z,0,1)
 #Ordenamos la muestra
 ur=sort(u)
```

```
 #Construimos estadistico de prueba
 s=0
 for (j in 1:n){
   s=s+(2*j-1)/n*(log(ur[j])+log(1-ur[n-j+1]))
 }
 A[i]=-n-s
}
hist(A,freq=FALSE, breaks=100, main="Distribucion Anderson-Lilliefors  n=5")
lines(density(A))
```




```
round(quantile(A,c(0.900,0.950,0.975,0.990)),3)

##   90%   95% 97.5%   99%
## 0.514 0.598 0.682 0.793
```

Como ejemplo, supongamos nuevamente que queremos probar normalidad a los datos de la temperatura corporal ahora haciendo uso de la prueba Anderson Darling - Lilliefors

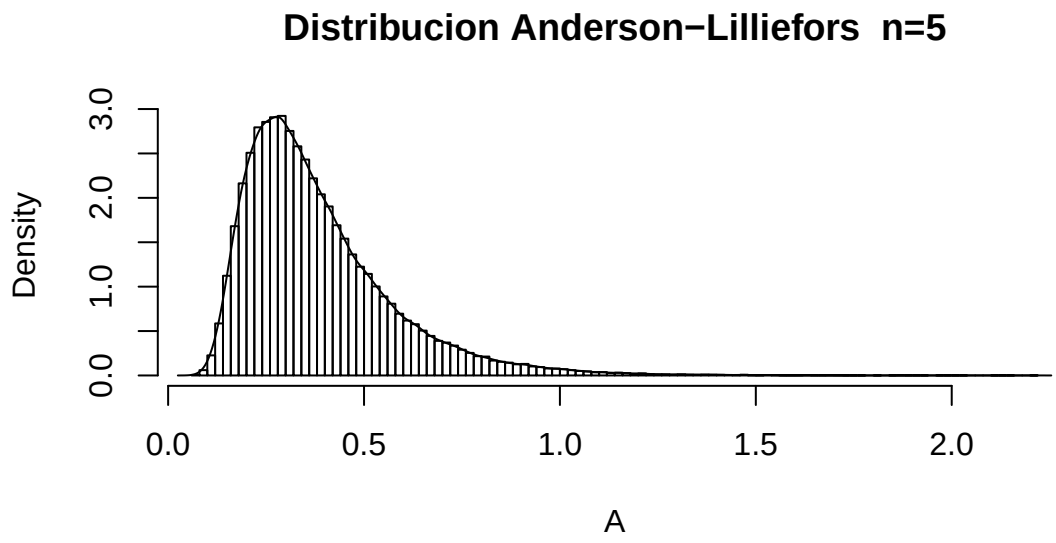
```
temp <- read.table("http://venus.unive.it/romanaz/statistics/data/bodytemp.txt",header = TRUE)
x<-temp[,1]
n<-length(x)
mu<-mean(x)
sigma<-var(x)
#Transformamos la muestra
z<-(x-mu)/sqrt(sigma)
```

```
#Transformamos nuevamente la muestra
u<-pnorm(z,0,1)
ur=sort(u)
#Construimos estadistico de prueba
 s=0
 for (j in 1:n){
   s=s+(2*j-1)/n*(log(ur[j])+log(1-ur[n-j+1]))
 }
 A.t<- -n-s
```

En este caso el estadístico de prueba toma el valor de $A^2 = 0.5201039$, para ver si rechazamos o no la hipótesis tendremos que simular la estadistica de prueba asociada para el caso $n = 130$, a continuación se presenta el histograma y cuantiles de dicha simulación:

```
################################################################
# Programa que simula lo Distribucion de Anderson Darling #
################################################################
#Numero de simulaciones
m=100000
#Tama\~no de muestra
n=130
#Vector donde guardaremos las simulaciones
A=rep(0,m)
for(i in 1:m){
 #Simulamos de una uniforme(0,1)
 x=rnorm(n,0,1)
 #Estimamos parametros
 mu<-mean(x)
 sigma<-var(x)
 #Transformamos muestra
 z<-(x-mu)/sqrt(sigma)

 #Transformamo nuevamente
 u<-pnorm(z,0,1)
 #Ordenamos la muestra
 ur=sort(u)
 #Construimos estadistico de prueba
 s=0
 for (j in 1:n){
   s=s+(2*j-1)/n*(log(ur[j])+log(1-ur[n-j+1]))
 }
 A[i]=-n-s
}
```

```
hist(A,freq=FALSE, breaks=100, main="Distribucion Anderson-Lilliefors  n=5")
lines(density(A))
```




```
round(quantile(A,c(0.900,0.950,0.975,0.990)),3)

##   90%   95% 97.5%   99%
## 0.629 0.749 0.872 1.025
```

Con $\alpha = 0.05$ se tiene que el cuantil asociado es 0.749, luego como $A^2 = 0.5201039 < 0.749$ se concluye que no se puede rechazar H_0. Una ventaja que tenemos de haber hecho las simulaciones es que también prodremos aproximar el p-value asociado a la prueba mediante el siguiente código:

```
p.value<-mean(A>A.t)
p.value

## [1] 0.18534
```

Como el p.value es mayor a 0.05 se concluye también que no se debe rechazar H_0.

Otra forma de correr la prueba es utilizando la función *ad.test* dentro de la libreria *nortest*

```
library(nortest)
temp <- read.table("http://venus.unive.it/romanaz/statistics/data/bodytemp.txt",header = TRUE)
x<-temp[,1]
ad.test(x)
```

```
##
##  Anderson-Darling normality test
##
## data:  x
## A = 0.5201, p-value = 0.1829
```

1.4.2 Prueba Cramer-Von-Mises

Otra de las puebas más conocidas que se obtiene a partir de las $QEDF$ es la denominada *Prueba Cramer-Von-Mises* la cual se obtiene haciendo que la función de ponderación tome la siguiente forma:

$$\psi(x) = 1$$

La idea detrás de esta ponderación es que da pesos iguales a toda la distribución es decir, pondera de igual forma tanto a las colas como al centro de la densidad.

En este caso el estadístico de prueba toma la siguiente forma:

$$V_n^2 = Q_n^2 = n \int_{-\infty}^{\infty} \left(F_n(x) - F_X^*(x)\right)^2 dF_X^*$$

El cual, al hacer el cambio de variable correspondiente ($u = F_X^*(x)$) se transforma en la siguiente expresión:

$$V_n^2 = n \int_0^1 \left(F_n(u) - u\right)^2 du$$

Donde hay que recordar que $F_n(u)$ corresponde a la distribución empírica generada por la muestra tranformada $u_1, \ldots, u_n$ Dicha integral puede ser resuelta integrando por pedazos ya que $F_n(u)$ es constante en el intervalo (u_i, u_{i+1}). Entonces:

$$V_n^2 = n \left(\sum_{i=0}^{n} \int_{u_{(i)}}^{u_{(i+1)}} \left(\frac{i}{n} - u \right)^2 \right) du \tag{1.5}$$

Donde tenemos que definimos $u_{(0)} = 0$ y $u_{(n+1)} = 1$. Estas integrales se resuelven recordando que:

$$\int \left(a - u\right)^2 du \;\; = \;\; \frac{1}{3} u \left(3a^2 - 3au + u^2\right)$$

Con lo anterior se prueba que (1.5) se tranforma en:

$$V_n^2 = \frac{1}{12n} + \sum_{i=1}^{n} \left(\frac{2i-1}{2n} - u_{(i)} \right)^2$$

La cual recordando el cambio de variable $u = F_X^*(x)$ el estadístico de prueba toma la forma:

$$V_n^2 = \frac{1}{12n} + \sum_{i=1}^{n} \left(\frac{2i-1}{2n} - F_X(x_{(i)}) \right)^2$$

Este estadístico tiene una distribución que ya no depende de $F_X^*(x)$ y sólo depende de n por lo cual se puede simular facilmente mediante simulaciones de muestras uniformes.

Nuevamente la idea será rechazar H_0 si $V_n^2 > w_{1-\alpha}$, donde $w_{1-\alpha}$ es el cuantil asociado a la distribución de V_n^2 bajo H_0 la cual puede simularse.

A manera de resumen entonces el método propuesto por Cramer-Von-Mises es el siguiente:

1. Se plantea la hipótesis.

$$H_0 : F_X(x) = F_X^*(x)$$

$$H_1 : F_X(x) \neq F_X^*(x)$$

2. Dada $x_1, \ldots, x_n$ m.a. de $F_X(X)$ se transforma la muestra mediante el cambio $u_i = F_X^*(x_i)$, de tal forma que se obtiene la muestra $u_1, \ldots, u_n$

3. Ordernar la muestra de menor a mayor obteniendo la muestra ordenada

$$u_{(1)}, \ldots, u_{(n)}$$

4. Calcular el estadístico de prueba:

$$V_n^2 = \frac{1}{12n} + \sum_{i=1}^{n} \left(\frac{2i-1}{2n} - u_{(i)} \right)^2$$

   Rechazar H_0 si $V_n^2 > w_{1-\alpha}$, donde $w_{1-\alpha}$ es el cuantil asociado a la distribución de V_n^2 bajo H_0 la cual puede simularse por medio de un programa en R.