UNIVERSIDAD DE EL SALVADOR
FACULTAD MULTIDISCIPLINARIA DE OCCIDENTE
DEPARTAMENTO DE MATEMÁTICA

Licenciatura en Estadística

Control Estadistico del Paquete R

"UNIDAD CINCO"
Práctica 21 - Prueba de hipótesis estadísticas y prueba de normalidad.

Alumna:
Martha Yoana Medina Sánchez

Fecha de elaboración
Santa Ana - 27 de noviembre de 2015

1. FORMULACIÓN Y PRUEBA DE HIPÓTESIS

Los pasos del método científico se pueden resumir de la siguiente forma:

1. Plantear el problema a resolver.
2. Efectuar las observaciones.
3. Formular una o más hipótesis.
4. Probar dichas hipótesis, y
5. Proclamar las conclusiones.

La Estadística nos puede ayudar en los pasos 2) (diseño de las observaciones) y 4) (prueba de hipótesis). Una definición de hipótesis es la siguiente: ünа explicación tentativa que cuenta con un conjunto de hechos y puede ser probada con una investigación posterior". La formulación de una hipótesis se logra examinando cuidadosamente las observaciones,para luego proponer un resultado posible.

La formulación formal de una hipótesis en el método científico se realiza definiendo la hipótesis nula (H_0) y la hipótesis alternativa (H_1). La hipótesis alternativa H_1, por otra parte, suele indicarse como el complemento de la H_0.

A la hora de tomar una decisión respecto de la hipótesis nula, surgen situaciones que nos pueden llevar a cometer diferentes errores. En los casos que H_0 se acepte y sea verdadera, así como también en el caso que H_0 se rechace y sea falsa, la decisión habrá sido la correcta. Pero en los otros dos casos se producen los denominados errores tipo I y tipo II.

El error tipo I (o de primera especie), se produce cuando se rechazó H_0 y es verdadera, $alfa$ quien representa la probabilidad de haber cometido este tipo de error y que se conoce como el nivel de significancia, suele fijarse antes de realizar la prueba. En el caso que H_0 sea aceptada siendo falsa, se cometerá el error denominado de tipo II, *beta* representa la probabilidad de cometer tal error. La potencia de un método estadístico en una determinada situación se calcula como $(1 - beta)$, lo que se corresponde con la situación de haber rechazado correctamente H_0.

Una hipótesis no se acepta, simplemente la evidencia no alcanza para rechazarla, y se mantiene como cierta mientras no se rechace.

En cualquier caso rechazar H_0 es lo mismo que aceptar la H_1 y viceversa. El resultado final de un método estadístico para la prueba de una hipótesis es el valor p, que indica la probabilidad de obtener un valor más extremo que el observado si H_0 es verdadera. Cuando p es menor que $alfa$ se procede a rechazar H_0.

Por ejemplo, un problema a resolver podría ser la importancia del estado nutricional en pacientes diabéticos con complicaciones; ya tenemos el paso 1) del método científico; luego efectuamos observaciones en dos grupos de sujetos, uno de control (saludables, denominados de aquen adelante como controles) y otro de diabéticos con complicaciones (denominados de aquí en adelante como pacientes); el tamaño de dichas muestras se basa en estudios similares ya publicados y/o

experiencia de los investigadores sobre y/o cálculos sobre tamaño de las muestras.

Uno de los indicadores más comunes del estado nutricional de una persona se puede cuantificar con el denominado índice de masa corporal (IMC), el cual se define con la siguiente ecuación:

$$IMC = Peso[Kg]/(Altura[m]^2)(1)$$

Los valores normales (y por lo tanto saludables) del IMC van de 20 a 25 Kg/m^2, valores superiores a 25 Kg/m^2 y menores de 30 Kg/m^2 se consideran como sobrepeso, finalmente IMC iguales o superiores a 30 Kg/m^2 se consideran como indicativos de obesidad. Valores altos de IMC son predictores de muerte en algunas patologías como enfermedades cardiovasculares, diabetes, cáncer, hipertensión arterial y osteoartritis. La obesidad por sí sola es un factor de riesgo de muerte prematura.

Tabla 1: IMC para cada sujeto, Grupo de Control

```
sujecto <- c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17,
                      18);
sujecto
```

```
##  [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18
```

```
imc <- c(23.6, 22.7, 21.2, 21.7, 20.7, 22.0, 21.8, 24.4,
                   20.1, 21.3, 20.5, 21.1, 21.4, 22.2, 22.6, 20.4, 23.3, 24.8);
imc
```

```
##  [1] 23.6 22.7 21.2 21.7 20.7 22.0 21.8 24.4 20.1 21.3 20.5 21.1 21.4 22.2
## [15] 22.6 20.4 23.3 24.8
```

```
hoja1 <- data.frame(Sujecto=sujecto, IMC=imc); hoja1
```

```
##    Sujecto  IMC
## 1        1 23.6
## 2        2 22.7
## 3        3 21.2
## 4        4 21.7
## 5        5 20.7
## 6        6 22.0
## 7        7 21.8
## 8        8 24.4
## 9        9 20.1
## 10      10 21.3
## 11      11 20.5
## 12      12 21.1
## 13      13 21.4
## 14      14 22.2
## 15      15 22.6
## 16      16 20.4
## 17      17 23.3
## 18      18 24.8
```

Tabla 2: IMC para cada sujeto, Grupo de Pacientes

```
sujecto <- c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14);
sujecto

##  [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14

imc <- c(25.6, 22.7, 25.9, 24.3, 25.2, 29.6, 21.3, 25.5, 27.4, 22.3, 24.4, 23.7,
         20.6, 22.8);
imc

##  [1] 25.6 22.7 25.9 24.3 25.2 29.6 21.3 25.5 27.4 22.3 24.4 23.7 20.6 22.8

hoja1 <- data.frame(Sujecto=sujecto, IMC=imc); hoja1

##    Sujecto  IMC
## 1        1 25.6
## 2        2 22.7
## 3        3 25.9
## 4        4 24.3
## 5        5 25.2
## 6        6 29.6
## 7        7 21.3
## 8        8 25.5
## 9        9 27.4
## 10      10 22.3
## 11      11 24.4
## 12      12 23.7
## 13      13 20.6
## 14      14 22.8
```

2. PRUEBAS DE NORMALIDAD DE UNA MUESTRA.

Antes de proceder a la prueba de una hipótesis debemos determinar la distribución de las variables consideradas en la muestra. En los métodos convencionales se trabaja con la distribución normal de dichas variables. El paso inicial entonces, es determinar si las variables en estudio pueden ser representadas por una distribución "normal". En otras palabras necesitamos verificar esta primera hipótesis.

La importancia de verificar la normalidad de las muestras en estudio es fundamental en estadística porque si las muestras son normales se pueden aplicar métodos estadísticos paramétricos convencionales, en caso contrario se deben o bien transformar los datos, o bien utilizar métodos como los no paramétricos u otros métodos estadísticos más sofisticados.

Los métodos de la estadística descriptiva nos pueden ayudar a verificar la normalidad de las variables, un histograma y un gráfico de cajas nos muestra en dos formas distintas la distribución

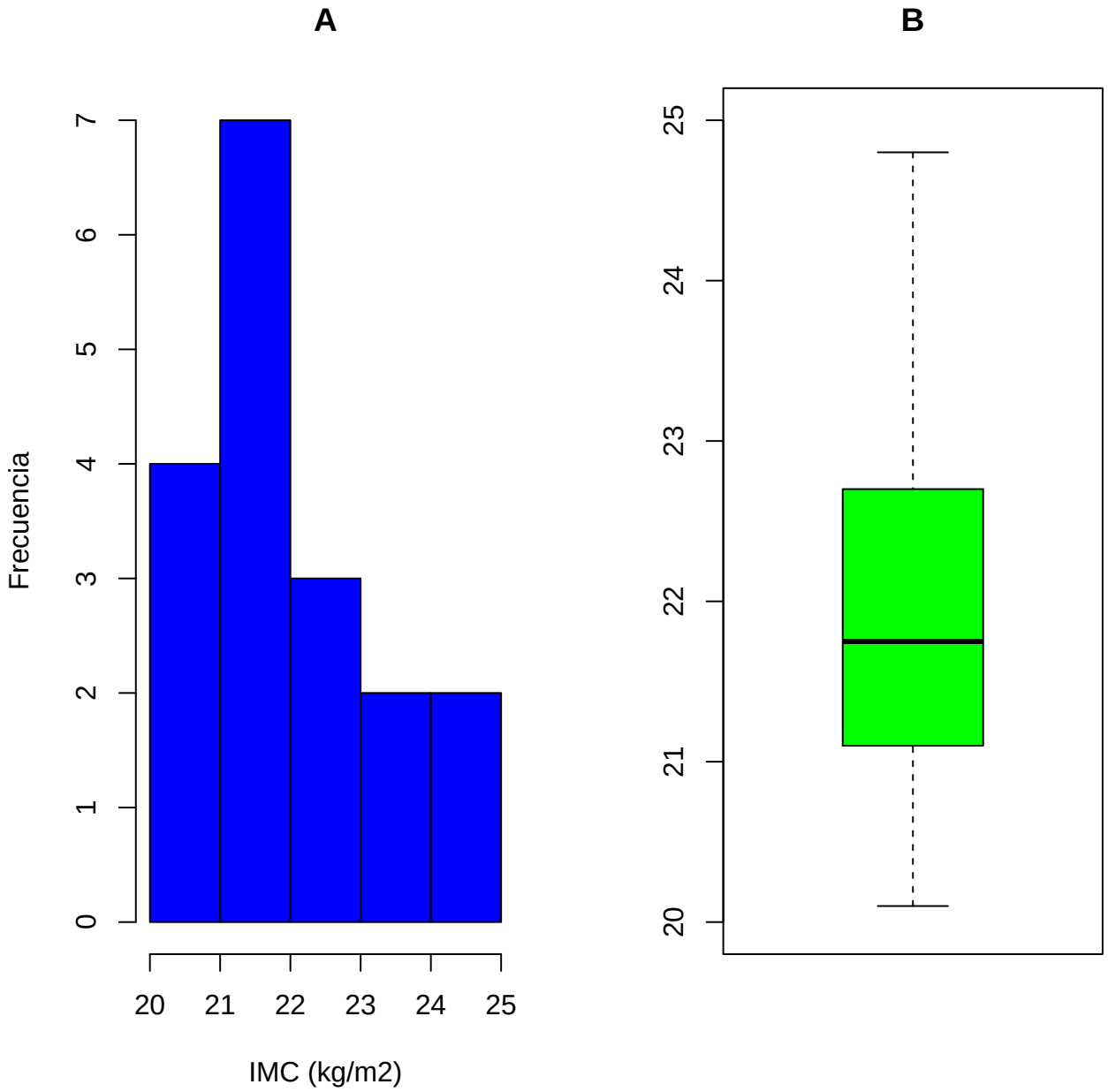
de los datos. Pruebas de normalidad más formales, no paramétricas, muy recomendables para verificar la normalidad de una variable son las pruebas de Shapiro-Wilk, y de Kolmogorov-Smirnov. También existen los gráficos PP y QQ.

Contrariamente a lo que se desea en la mayoría de los casos, en las pruebas de normalidad se busca aceptar H_0, dado que en la mayoría de los métodos estadísticos convencionales es necesaria la distribución normal de la variable de interés, siendo posible conocer los parámetros que la describen tales como su media $(\hat{A}\mu)$ y su desviación estándar (s) . Un p - valor mayor a 0.10 en los tests de normalidad indicaría que no hay prueba suficiente para rechazar la normalidad de la variable. Por el contrario un p - valor menor a 0.01 indicaría que nuestros datos no siguen una distribución normal.

A continuación procedemos a contrastar normalidad para los datos del IMC en los grupos de Control y de Pacientes. Observe que la característica de interés debe ser normal en ambos grupos, es decir, la normalidad se estudia en cada uno de ellos y noen la información combinada de los grupos.

El siguiente código en lenguaje R podría ser utilizado para dichos fines:

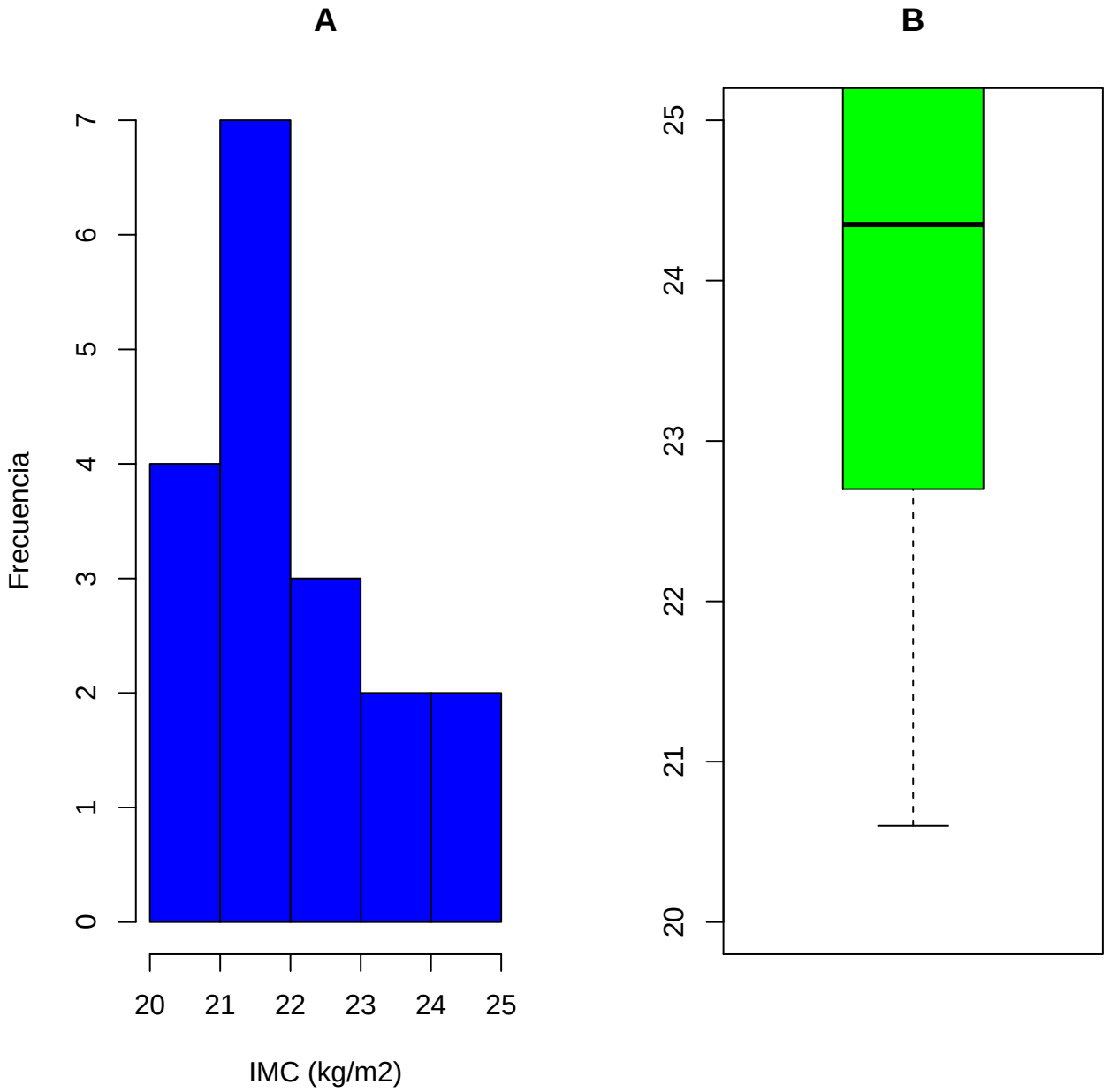
```
#se digitan los datos del grupo de control

IMC_Control<-c(23.6, 22.7, 21.2, 21.7, 20.7, 22.0, 21.8, 24.2, 20.1, 21.3, 20.5,
                21.1, 21.4, 22.2, 22.6, 20.4, 23.3, 24.8)
par(mfrow=c(1,2))

#se genera el histograma de la variables de inter\'es

hist(IMC_Control,main="A",xlab="IMC (kg/m2)",ylab="Frecuencia",col="blue")

# se genera el diagrama de caja de la variable de inter\'es y se muestra en
# la misma ventana

boxplot(IMC_Control,main="B", lab="IMC (kg/m2)",ylim=c(20,25), col="green")
```




```
# los commandos para contrastar normalidad son los siguientes

sw <- shapiro.test(IMC_Control)
sw

##
##  Shapiro-Wilk normality test
##
## data:  IMC_Control
## W = 0.95321, p-value = 0.4776

# note que en la prueba de Shapiro solo es necesario especificar la variable que
```

```
# se est\'a contrastado. ESTA PRUEBA SOLAMENTE SEUTILIZA PARA VERIFICAR NORMALIDAD.

ks <- ks.test(IMC_Control,"pnorm",mean=mean(IMC_Control),sd=sd(IMC_Control))
ks

##
##  One-sample Kolmogorov-Smirnov test
##
## data:  IMC_Control
## D = 0.11172, p-value = 0.9595
## alternative hypothesis: two-sided

# note que la prueba de Kolmogorov es m\'as general, permite contrastar cualquier
# tipo de distribuci\'on, en "pnorm" se indicaque la distribuci\'on que se desea
# contrastar es la normal; sin embargo, es necesario especificar los par\'ametros
# de la distribuci\'on media (mean) y desviaci\'on (sd) estimados a partir de los
# datos.

# Luego se digitan los datos para pacientes y se ejecutan las mismas instrucciones

IMC_Pacientes <- c(25.6, 22.7, 25.9, 24.3, 25.2, 29.6, 21.3, 25.5, 27.4, 22.3, 24.4,
                   23.7, 20.6, 22.8)
```

```
IMC_Pacientes <- c(25.6, 22.7, 25.9, 24.3, 25.2, 29.6, 21.3, 25.5, 27.4, 22.3, 24.4,
                   23.7, 20.6, 22.8)
par(mfrow=c(1,2))

#se genera el histograma de la variables de inter\'es

hist(IMC_Control,main="A",xlab="IMC (kg/m2)",ylab="Frecuencia",col="blue")

# se genera el diagrama de caja de la variable de inter\'es y se muestra en
# la misma ventana

boxplot(IMC_Pacientes,main="B", lab="IMC (kg/m2)",ylim=c(20,25), col="green")
```




```
# los commandos para contrastar normalidad son los siguientes

sw <- shapiro.test(IMC_Pacientes)
sw

##
##  Shapiro-Wilk normality test
##
## data:  IMC_Pacientes
## W = 0.97437, p-value = 0.929

# note que en la prueba de Shapiro solo es necesario especificar la variable que
```

```r
# se est\'a contrastado. ESTA PRUEBA SOLAMENTE SEUTILIZA PARA VERIFICAR NORMALIDAD.

ks <- ks.test(IMC_Pacientes,"pnorm",mean=mean(IMC_Pacientes),sd=sd(IMC_Pacientes))
ks

##
##  One-sample Kolmogorov-Smirnov test
##
## data:  IMC_Pacientes
## D = 0.12172, p-value = 0.9695
## alternative hypothesis: two-sided

# note que la prueba de Kolmogorov es m\'as general, permite contrastar cualquier
# tipo de distribuci\'on, en "pnorm" se indicaque la distribuci\'on que se desea
# contrastar es la normal; sin embargo, es necesario especificar los par\'ametros
# de la distribuci\'on media (mean) y desviaci\'on (sd) estimados a partir de los
# datos.
```

3. PRUEBAS SOBRE MUESTRAS NO NORMALES.

Hasta el momento en el ejemplo anterior la distribución de los datos es normal, por lo cual la aplicación de pruebas paramétricas normales es totalmente válido. ¿Qué pasa si estamos ante muestras no normales? la respuesta obvia es que nos olvidamos de las pruebas paramétricas y buscamos la equivalente no paramétrica, pero siempre que se pueda es aconsejable transformar la muestra para que tenga distribución normal y así poder aplicar los métodos clásicos.

La transformación de la cual estamos hablando es numérica, puede ser simplemente calcular el logaritmo natural de cada observación, y luego verificar la normalidad de la muestra transformada.

Por lo tanto el test medirá si los *logaritmos* de las variables difieren o no, en este caso se debería considerar si esto tiene interpretación biológica.

Un comentario especial merecen las pruebas de normalidad, a veces omitidas por algunos investigadores, pero que se consideran como fundamentales para poder verificar la normalidad de las muestras, y de esta forma poder aplicar apropiadamente las pruebas estadísticas paramétricas. La prueba de normalidad de Shapiro-Wilk está considerada como la más poderosa para verificar la normalidad de una muestra, por lo cual algunos estadísticos consideran que por sí sola es suficiente.