# 6. INFERENCIA

Generalmente, cuando en una investigación se analiza una población es casi imposible tomarla en su conjunto, individuo por individuo, ya sea por cuestiones económicas o de accesibilidad, entre otras; por tanto, se hace necesario seleccionar una muestra representativa, de un tamaño manejable, la cual es utilizada para sacar conclusiones de la población de interés. Al realizar el proceso anterior se está utilizando *estadística inferencial*. Dentro de este contexto, la estadística inferencial involucra el uso de un estadístico para obtener una conclusión o inferencia sobre su parámetro correspondiente; por ejemplo: se puede utilizar el estadístico (media muestral) como estimador de (media poblacional).

Las muestras tienen un impacto directo en las decisiones que se toman, por tanto, se hace necesario realizar estas inferencias de manera correcta; en muchas aplicaciones prácticas se trabaja con estadísticos que contemplan dentro de sus supuestos básicos que la distribución muestral de los datos se ajuste a una Normal. Teniendo en cuenta este supuesto, se presentan algunas gráficas y pruebas para verificar el ajuste de un conjunto de datos.

## 6.1 GRÁFICO CUANTIL-CUANTIL (Q-Q PLOT)

Este tipo de gráfico compara los cuartiles de un conjunto de datos dado contra los cuantiles de la distribución Normal; mediante esta comparación es posible identificar si el conjunto de datos se aproxima a la distribución normal. En R, el comando que permite realizar este tipo de gráfico es **qqnorm(vector)**.

Ejemplo: Verificar que el siguiente conjunto de datos sigue una distribución aproximadamente normal.

```
>
> # vector de datos
> Y=c(9,5,8,10,7,5,6,4,1,1,4,4,5,0,10)
>
> # gráfico Q-Q
> qqnorm(Y)
>
```

*Imagen 106. Salida R gráfico Q-Q*

A través de estas instrucciones se produce un gráfico de la forma:

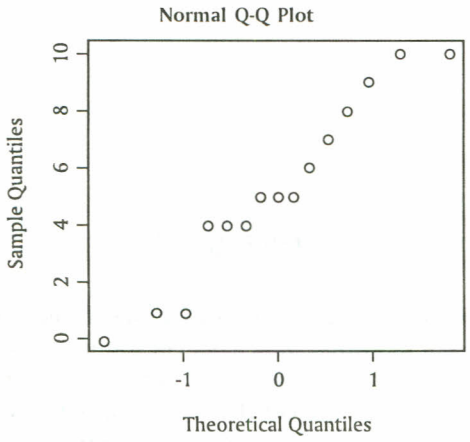
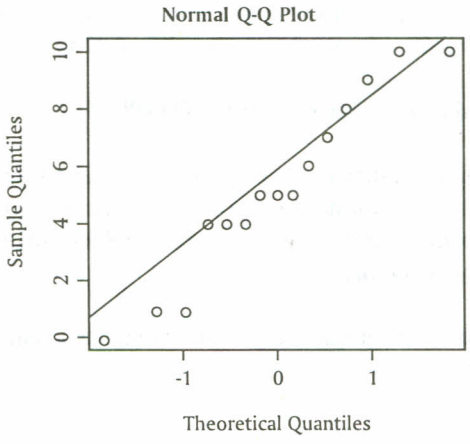



*Imagen 107. Salida R gráfico Q-Q*

Si se escribe el comando **qqline()**, este añade una recta que pasa por los cuantiles de la distribución y de los datos.

```
>
> # gráfico Q-Q
> qqnorm(Y)
> qqline(Y)
>
```




*Imágenes 108 y 109. Salida R gráfico Q-Q*

Se espera que, si el conjunto de datos pertenece o se aproxima a la distribución normal, los puntos estén sobre la línea. Como se puede apreciar, los datos del ejemplo pertenecen a una distribución normal.

Además, el comando **qqplot(vector1, vector2)** representa los cuantiles del vector1 sobre los cuantiles del vector2 para comparar sus distribuciones respectivas.

## 6.2 PRUEBA EXACTA DE NORMALIDAD

Una prueba exacta para probar la normalidad de un conjunto de datos es la **Shapiro-wilks**; en R es posible realizarla mediante el comando **shapiro.test(vector)**. La prueba arroja el estadístico W (Shapiro – Wilk) y el p valor, es decir, el valor más bajo de significancia al cual se puede rechazar la hipótesis nula. La hipótesis nula de la prueba es:

$H_0$: *Los datos siguen una distribución normal*

Ejemplo: Los valores sobre las longitudes en micras de 50 filamentos de la producción de una máquina son los siguientes:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 102 | 98 | 93 | 100 | 98 | 105 | 115 | 110 | 99 | 120 |
| 115 | 130 | 100 | 86 | 95 | 103 | 105 | 92 | 99 | 134 |
| 116 | 118 | 89 | 102 | 128 | 99 | 119 | 128 | 110 | 130 |
| 112 | 114 | 106 | 114 | 100 | 116 | 108 | 113 | 106 | 105 |
| 120 | 106 | 110 | 100 | 106 | 117 | 109 | 108 | 105 | 106 |

*Imagen 110. Datos de 50 filamentos*

Determinar si los datos de la muestra provienen de una distribución aproximadamente normal. Para lo cual se procede de la siguiente manera:

```
>
> # Vector de datos
> X=c(102,115,116,112,120,98,130,118,114,106
+       93,100,89,106,110,100,86,102,114,100,
+       98,95,128,100,106,105,103,99,116,117,
+       115,105,119,108,109,110,92,128,113,108,
+     99,99,110,106,105,120,134,130,105,106)
>
> # prueba Shapiro - Wilks
> shapiro.test(X)

        Shapiro-Wilk normality test

data:  X
W = 0.9759, p-value = 0.3954
```

*Imagen 111. Salida R Prueba Shapiro-Wilks*

## 6.3 PRUEBA KOLGOMOROV-SMIRNOV

Si se sospecha que la distribución de un conjunto de datos pertenece a una distribución diferente a la distribución normal y se desea comprobar a qué distribución se ajusta, se usa la prueba de **Kolmogorov-Smirnov**, que permite contrastar un conjunto de datos contra cualquier distribución; el comando por utilizar es:

**ks.test(vector,"distribución",parámetros,alternative=c("two.sided","less", greater"))**; los argumentos utilizados son: **vector**, se refiere al vector con el conjunto de datos; **"distribución"**, hace referencia a la función de distribución con la cual se desee contrastar (esta distribución debe estar escrita como se vio en la sección de distribuciones); **parámetros**, son los parámetros de cada distribución, y **alternative**, hace referencia al tipo de prueba que se desee calcular, es decir, prueba a dos colas "two.sided": prueba a cola izquierda, "less", y prueba a cola derecha, "greater".

Ejemplo: Se cree que las fallas de la máquina empacadora en la empresa Empacar Ltda. sigue una distribución de Poisson; comprobar este supuesto con los datos de los últimos 15 días de producción de la empresa; por experiencia del año anterior, se tiene que en promedio una máquina de la empresa presenta 5 fallas diariamente.

```
>
> # Fallas
> X=c(3,3,2,4,1,6,7,5,6,8,5,6,8,7,5)
>
> # Prueba Kolgomorov - smirnov
> ks.test(X,"ppois",lambda=5,alternative="two.sided")

        One-sample Kolmogorov-Smirnov test

data:  X
D = 0.2826, p-value = 0.1820
alternative hypothesis: two-sided
```

*Imagen 112. Salida R Prueba Kolmogorov-Smirnov*

El contraste Kolgomorov-Smirnov también se puede utilizar para comparar dos conjuntos de datos a fin de verificar si pertenecen aproximadamente a una misma distribución.

## 6.4 HIPÓTESIS ESTADÍSTICA

Una hipótesis estadística es un enunciado acerca de la distribución de probabilidad de una variable aleatoria y de sus parámetros. Las hipótesis estadísticas involucran a menudo una o más características de la distribución, como, por ejemplo, forma o independencia de la variable aleatoria. A continuación se muestran los pasos por seguir para realizar una hipótesis estadística:

1. Expresar la hipótesis nula
2. Expresar la hipótesis alternativa
3. Especificar el nivel de significancia
4. Determinar el tamaño de la muestra
5. Establecer los valores críticos que determinan las regiones de rechazo y las de no rechazo.
6. Determinar la prueba estadística.
7. Recolectar los datos y calcular el valor del estadístico de la muestra para la prueba apropiada.

8. Determinar si la prueba estadística ha sido en la zona de rechazo o en una de no rechazo.
9. Determinar la decisión estadística.
10. Expresar la decisión estadística en términos del problema.

**6.4.1 Pruebas de hipótesis para la media.** Cuando se van a realizar pruebas de hipótesis relativas a la media poblacional se debe saber si la varianza poblacional $\sigma^2$ es conocida o desconocida, ya que la distribución subyacente al estadístico de prueba será la normal estándar, si la varianza es conocida, y t-student, si la varianza es desconocida.

**6.4.1.1** ***Prueba de hipótesis para la media, varianza conocida.*** Cuando la varianza $\sigma^2$ es conocida, las pruebas de hipótesis se basan en el estadístico de prueba:

$$Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}}$$

*Ecuación 13. Estadístico de prueba*

Existen tres tipos de pruebas que se pueden plantear con respecto a la media:

1. Prueba cola a derecha $H_o : \mu = \mu_o$ $H_1 : \mu > \mu_o$
2. Prueba cola a izquierda $H_o : \mu = \mu_o$ $H_1 : \mu < \mu_o$
3. Prueba a dos colas $H_o : \mu = \mu_o$ $H_1 : \mu \neq \mu_o$

Para este caso, el valor o los valores críticos se obtienen de la distribución normal estándar, es decir, de la distribución normal con media igual a cero y varianza uno. Mediante el siguiente ejemplo se muestra cómo se puede realizar esta prueba paso a paso.

Ejemplo: Suponga una variable aleatoria P que se designa para el peso de un pasajero de avión; el interés está en conocer el peso promedio de todos los pasajeros. Como hay limitaciones de tiempo y dinero para pesarlos a todos, se toma una muestra de 36 pasajeros, de la cual se obtiene una media muestral $\bar{x} = 160$ libras. Suponga además que la distribución de los pasajeros es aproximadamente normal con desviación estándar ó = 30 libras, con un nivel de significancia de 0.05. ¿Se puede concluir que el peso promedio de todos los pasajeros es menor que 170 libras?

Las hipótesis planteadas para este ejercicio son:

$$H_o : \mu \geq 170 \qquad \text{vs} \qquad H_1 : \mu < 170$$

```
>
> # media muestral
> X = 160
>
> # Mu
> M = 170
>
> # desviación estándar
> d = 30
>
> # Tamaño de la muestra
> n = 36
>
```

```
>
> # Zona de rechazo
> # Una Z con (alfa = 0.05)
> Zcritico = qnorm(0.05,mean=0,sd=1)
> Zcritico
[1] -1.644854
>
> # Estadístico de prueba
> Zcalculado = (X - M) / (d / sqrt(n))
> Zcalculado
[1] -2
>
> # Decisión rechazar Hipótesis nula si
> Zcalculado < Zcritico
[1] TRUE
```

*Imágenes 113 y 114. Salida R Prueba de hipótesis para la media*

Los puntos críticos se determinan de acuerdo con la hipótesis alternativa $H_1$; es así que para la hipótesis $H_1 : \mu > \mu_o$ el punto crítico es una $Z_{(1-\alpha)}$, para la hipótesis $H_1 : \mu < \mu_o$ el punto crítico es una $Z_{(\alpha)}$ y para la hipótesis $H_1 : \mu \neq \mu_o$ se tiene que los puntos críticos son $Z_{(1-\alpha/2)}$ y $Z_{(\alpha/2)}$.

Esta misma prueba es posible realizarla en R mediante un comando directo; para poder realizarla es necesario cargar los paquetes PASWR, e1071, class y MASS; el comando utilizado para esto es:

z.test(x, y = NULL, alternative = "two.sided", mu = 0, sigma.x = NULL,
sigma.y = NULL, conf.level = 0.95)

Los argumentos utilizados son: **x**, vector numérico; **y**, vector numérico en caso de realizar una prueba de diferencias de medias; **alternative**, se selecciona el tipo de prueba deseado; **mu**, especifica el valor de la media o la diferencia de medias; **sigma.x** y **sigma.y**, número que representa la desviación estándar de x o y; **conf.level**, es el nivel de significancia al cual se realiza la prueba.

Ejemplo: Los siguientes conjuntos de datos representan la producción en toneladas de dos fincas; probar si los promedios de producción no difieren en más de dos toneladas; las producciones en toneladas se distribuyen de manera normal con desviaciones estándar 0.5.

```
>
> # Datos de producción en toneladas
> x = c(7.8, 6.6, 6.5, 7.4, 7.3, 7, 6.4, 7.1, 6.7, 7.6, 6.8)
> y = c(4.5, 5.4, 6.1, 6.1, 5.4, 5, 4.1, 5.5)
>
> # Prueba para diferencias de medias
> z.test(x, sigma.x=0.5, y, sigma.y=0.5, mu=2)

        Two-sample z-Test

data:  x and y
z = -1.0516, p-value = 0.293
alternative hypothesis: true difference in means is not equal to 2
95 percent confidence interval:
 1.300323 2.211040
sample estimates:
mean of x mean of y
 7.018182 5.262500
```

*Imagen 115. Salida R prueba Z diferencias de medias*

6.4.1.2 ***Prueba de hipótesis para la media con varianza desconocida y tamaño de muestra pequeño.*** Cuando la varianza $\sigma^2$ no es conocida, las pruebas de hipótesis se basan en el siguiente estadístico de prueba que sigue una distribución t - student con n - 1 grados de libertad.

$$t = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

*Ecuación 14. Estadístico de prueba*

El comando utilizado en R para realizar esta prueba es **t.test(x,y=NULL,alternative="", mu=,paired=,var.equal=,conf.level=)**, donde **x** se refiere al vector de datos, **alternative** selecciona el tipo de prueba deseada, **mu** es el valor de la media y **conf.level** es el nivel de significancia al cual se realiza la prueba. Los siguientes parámetros son utilizados dentro de la prueba si se desea comparar los promedios de dos muestras independientes: **mu** especifica la diferencia entre los promedios; si las muestras son pareadas, entonces **paired=TRUE**; si se asumen varianzas iguales en las muestras, entonces **var.equal=TRUE**.

Ejemplo: En una muestra de 15 bolsas de arroz de un kilo de la molinera "El Bello Arroz" se encontró la siguiente información sobre el peso de ellas:

987, 997, 1006, 965, 1009, 968, 1007, 999, 1006, 1099, 1000, 997, 985, 1002, 1069

Si un cliente demanda a la compañía por no presentar el peso de la referencia, realice una prueba de hipótesis y resuelva el conflicto con una confiabilidad del 95%. La hipótesis alternativa para este caso sería $H_1 : \mu \neq 1000$ gramos, con lo que se tiene el siguiente desarrollo:

```
>
> # Pesajes
> X=c(987,997,1006,965,1009,968,1007,
+ 999,1006,1099,1000,997,985,1002,1069)
>
> # prueba de Hipótesis
> t.test(X,alternative="two.sided",
+         mu=1000,conf.level=0.95)

        One Sample t-test

data: X
t = 0.7152, df = 14, p-value = 0.4862
alternative hypothesis: true mean is not equal to 1000
95 percent confidence interval:
  987.2075 1025.5925
sample estimates:
mean of x
   1006.4
```

*Imagen 116. Salida R prueba t*

**6.4.2 Prueba de hipótesis para la homogeneidad de varianzas.** Si se requiere contrastar la igualdad de varianzas se puede utilizar el comando

var.test(x,y,ratio=1,alternative=c("two.sided","less","greater"), conf.level=),

que desarrolla una prueba F para comparar las varianzas de dos muestras provenientes de distribuciones normales. El parámetro ratio se refiere al cociente hipotético de las varianzas de las poblaciones.

```
>
> # Muestras de la distribución normal
> x = rnorm(50, mean = 0, sd = 2)
> y = rnorm(30, mean = 1, sd = 1)
>
> # Prueba F
> var.test(x, y, ratio=1)

        F test to compare two variances

data:  x and y
F = 3.1209, num df = 49, denom df = 2, p-value = 0.001583
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 1.568018 5.871735
sample estimates:
ratio of variances
          3.120912
```

*Imagen 117. Salida R prueba F*

**6.4.3 Prueba de hipótesis para la Correlación/Asociación entre muestras pareadas.** Si se quiere verificar el grado de correlación/asociación entre muestras se puede utilizar la función:

```
cor.test( x, y, alternative =c("two.sided", "less", "greater",
    method= c("pearson", "kendall","spearman",
    exact = NULL, conf.level = 0.95)
```

Donde **method** se refiere al tipo de método por el cual se calcula el coeficiente de correlación utilizado en la prueba; las opciones son "pearson" (coeficiente de correlación de Pearson), "kendall" (tao de kendall) y "spearman" (rho de Sperman).

En R también es posible realizar pruebas de hipótesis para el Sesgo y Curtosis de un conjunto de datos; para esto se hace necesario que previamente a la prueba se cargue el paquete **moments**.

**6.4.4 Prueba de hipótesis para el Sesgo.** Bajo supuesto de normalidad, un conjunto de datos debe tener una distribución simétrica, para lo cual el sesgo debe ser igual a cero; esta última afirmación es la hipótesis nula para la prueba que realiza R. La instrucción que permite realizar esta prueba es la siguiente:

**agostino.test(x, alternative=c("two.sided", "less", "greater"))**

Ejemplo: Generar 1000 datos de la distribución Normal estándar y verificar si el sesgo es igual o aproximadamente igual a cero (hipótesis nula).

```
>
> # Aleatorios de Normal estándar
> x = rnorm(1000)
>
> skewness(x)
[1] 0.07151397
>
> agostino.test(x)

        D'agostino skewness test

data:   x
skew = 0.0715, z = 0.6121, p-value = 0.5404
alternative hypothesis: data have a skewness
```

*Imagen 118. Salida R prueba para el sesgo*

En la prueba anterior, cuando no se especifica el tipo de prueba que se desee, R por defecto realiza una prueba a dos colas. Los resultados muestran la no existencia de evidencia estadística para rechazar la hipótesis nula.

**6.4.5 Prueba de hipótesis para la Curtosis.** Bajo supuestos de normalidad, un conjunto de datos debe tener curtosis igual a 3; por tanto, la hipótesis nula para la prueba es que la curtosis es igual a 3. La instrucción que permite realizar esta prueba es la siguiente:

**anscombe.test(x, alternative=c("two.sided", "less", "greater"))**

```
>
> # Aleatorios de Normal estándar
> x = rnorm(1000)
>
> kurtosis(x)
[1] 3.171290
>
> anscombe.test(x,alternative="two.sided")

        Anscombe-Glynn kurtosis test

data:  x
kurt = 3.1713, z = 1.1467, p-value = 0.2515
alternative hypothesis: kurtosis is not equal to 3
```

*Imagen 119. Salida R prueba para la curtosis*

Ejemplo: Generar 1000 datos de la distribución Normal estándar y verificar si la curtosis es igual o aproximadamente igual a tres (hipótesis nula).

Como se ha indicado, una de las condiciones de aplicación de los contrastes anteriores es la normalidad. Si esta falta, es posible utilizar otro tipo de pruebas para contrastar parámetros de una o de dos muestras; entre estos contraste se cuenta con el contraste de Wilcoxon (o de Mann-Whitney), que solo presupone que la distribución común es continua.

**6.4.6 Prueba de Wilcoxon.** Si la prueba se realiza con solo una muestra o se tienen dos conjuntos de datos y las muestras de estos conjuntos son pareadas, la prueba de rangos Wilcoxon tiene como hipótesis nula que la distribución de la muestra (en el caso de una muestra) o de las dos muestras (dos muestras pareadas) es simétrica sobre su media ($\mu$). Por otra parte, si las muestras no son pareadas, la prueba tiene como hipótesis nula que las distribuciones de las muestras difieren por la localización de la media. A continuación se presenta el comando y los argumentos utilizados para realizar esta prueba.

**wilcox.test(x, y=NULL, alternative=c("two.sided","less","greater"),**
**mu = 0, paired=, exact = NULL, correct = TRUE , conf.int = FALSE, conf.level= )**

x, y se refieren a los vectores de datos si **y=NULL**, la prueba es para una sola muestra; **mu** se refiere a un número específico sobre el parámetro usado en la prueba de hipótesis; **exact** es indicación lógica que determina si el p valor es calculado; **correct** es indicación lógica si se aplica una corrección continua en la distribución normal usada para calcular el p valor; **conf.int**, si es igual **TRUE**, calcula un intervalo de confianza para el parámetro de localización; **conf.level** determina el nivel de confianza para el intervalo.

Existen más pruebas de hipótesis que se desarrollan y que no tienen una instrucción especifica en R; pero como se vio en la primera prueba de hipótesis, estas se pueden escribir con facilidad paso a paso, ya que, como se ha mencionado, R es un lenguaje sencillo.

## 6.5 EJERCICIOS

**6.5.1** Los siguientes datos hacen parte de una investigación sobre el crecimiento y desarrollo de niños de cierta comunidad indígena.

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 21 | 45 | 35 | 28 | 42 | 30 | 31 | 32 | 36 | 48 | 41 |
| 22 | 26 | 37 | 40 | 29 | 28 | 27 | 28 | 32 | 33 | 35 |

Verificar que el conjunto de datos sigue una distribución aproximadamente normal, mediante pruebas graficas (Q–Q PLOT) y prueba exacta de normalidad (Shapiro).

**6.5.2** Se registraron los siguientes datos, en segundos, sobre lo que tardan algunos hombres y mujeres en resolver ciertos ejercicios de matemáticas en la universidad, los cuales fueron seleccionados aleatoriamente.

| Hombres | Mujeres |
|---|---|
| $n_1 = 12$ | $n_2 = 15$ |
| $\bar{x}_1 = 19$ | $\bar{x}_2 = 22$ |
| $S_1^2 = 1.8$ | $S_2^2 = 2$ |

Suponga que los tiempos para los dos grupos se distribuyen normalmente y que las varianzas son iguales, aunque desconocidas. Pruebe la hipótesis referente a que no existe diferencia entre el tiempo promedio de solución de los ejercicios entre hombres y mujeres, a un nivel de confianza del 95%.