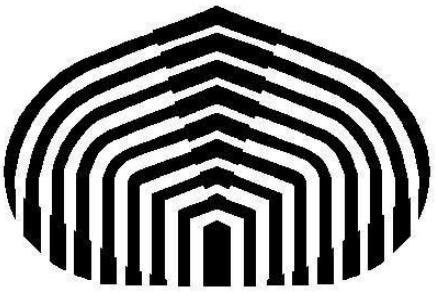


Universidad Simón Bolívar

Departamento de Cómputo científico y Estadística

Estadística para Ingenieros – CO3321

Enero-Marzo 2019

Resultados
Laboratorio 3

Preparador: Andy Guevara

Arturo Yépez. Carnet: 15-11551

Antonella Requena. Carnet: 15-11196

1. **Se desea conocer si existe alguna relación entre el nivel socioeconómico (pobre, medio y rico) y la zona donde vive (rural y urbano) 505 grupos de familias:**

Como queremos conocer si existe una relación/independencia sobre el nivel socioeconómico de nuestra población, evaluamos los datos como una tabla de contingencia. De esta forma, primero proponemos nuestras hipótesis:

H0: Hay relación sobre el nivel socioeconómico y la zona donde viven las familias.

Ha: No existe tal relación.

Ahora, guardamos los respectivos datos en variables, creamos nuestra matriz y evaluamos con la función de R `chisq.test()` y obtenemos:

```
        Pearson's Chi-squared test

data:  matriz

X-squared = 31.347, df = 2, p-value = 1.56e-07
```

Podemos ver que nuestro estadístico X^2 = 31.347, y nuestro p-valor = 1.56e-07. Como no nos proporcionaron nivel de significancia, utilizamos el p-valor y nos fijamos es en lo pequeño que resulta y de eso podemos de una concluir directamente que para la mayoría de niveles de significancia podemos observar que no hay ninguna relación entre el nivel socioeconómico y la zona donde viven las familias.

2. **Se efectuó un estudio para determinar si los conductores preferı´an los carriles centrales de una vı´a rápida con cuatro carriles para cada sentido. Se observó un total de mil autos a las horas picos de la mañana y se registraron los carriles por los que éstos circulaban. La siguiente tabla contiene los resultados. Proporcionan los datos suficiente evidencia para concluir que los conductores tienen preferencia por algunos carriles? (Pruebe la hipótesis que afirma que p1 = p2 = p3 = p4 = ¼ utilizando un nivel de significancia de α = 0,05). Establezca lı´mites para el p-valor asociado.**

El ejercicio nos pide verificar si de los datos obtenidos en el estudio podemos confirmar si hay evidencia suficiente para concluir si los conductores tienen preferencia por algún carril en específico. Para ello, nuestra hipótesis nula es:

H0: p1 = p2 = p3 = p4 = ¼

Por la forma en la que están agrupados nuestros datos, nos indica que debemos hacer una prueba para proporciones. Para aplicar eso, los respectivos datos los guardamos en variables y usamos la función de R `prop.test()` que luego los arroja los siguientes resultados:

```
data:  fi out of c(1000, 1000, 1000, 1000), null probabilities p

X-squared = 32.64, df = 4, p-value = 1.415e-06

alternative hypothesis: two.sided

null values:
```

```
prop 1 prop 2 prop 3 prop 4
  0.25   0.25   0.25   0.25
sample estimates:
prop 1 prop 2 prop 3 prop 4
 0.294  0.276  0.238  0.192
```

Inmediatamente observamos que nuestro estadístico X^2 = 32.64, y nuestro p-valor = 1.415e-06. De nuevo, con el p-valor podemos observar que como es lo suficientemente bajo de una podemos decir que hay evidencia más que suficiente para concluir que hay preferencia por sobre los carriles centrales de la vía.

3. **Realizar una prueba Chi-cuadrado de bondad de ajuste para decidir si puede aceptarse que las edades sigan una distribución normal.**

En este ejercicio se nos pide determinar la bondad de ajuste para saber si podemos decir que los datos se aproximan a una distribución normal.

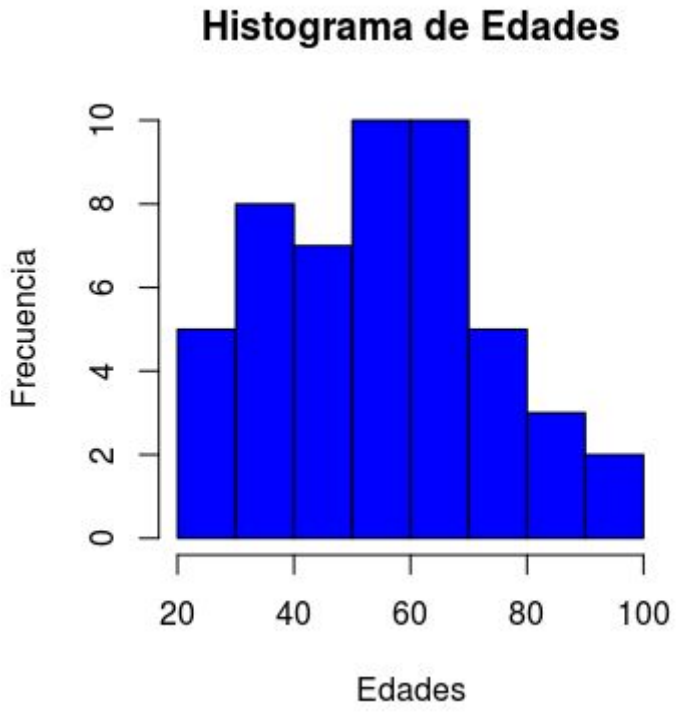
Para ello, nuestra hipótesis nula es:

Ho: X ~ Nor(mu,sigma^2) donde X es la variable aleatoria de la edad de las personas que asisten a un bingo.

Así que, para calcular pi asumimos cierto que X se distribuye de forma normal.

Como además, los datos proporcionados no están agrupados, procedemos a agruparlos en clases con ayuda de R generando un histograma.

```
> xi = c(32, 23, 64, 31, 74, 44, 61, 33, 66, 73,
+        27, 65, 40, 54, 23, 43, 58, 87, 58, 62,
+        68, 89, 93, 24, 73, 42, 33, 63, 36, 48,
+        77, 75, 37, 59, 70, 61, 43, 68, 54, 29,
+        48, 81, 57, 97, 35, 58, 56, 58, 57, 45)
>
> hist(xi, main = "Histograma de Edades", ylab = "Frecuencia", xlab = "Edades", col = "blue")
```

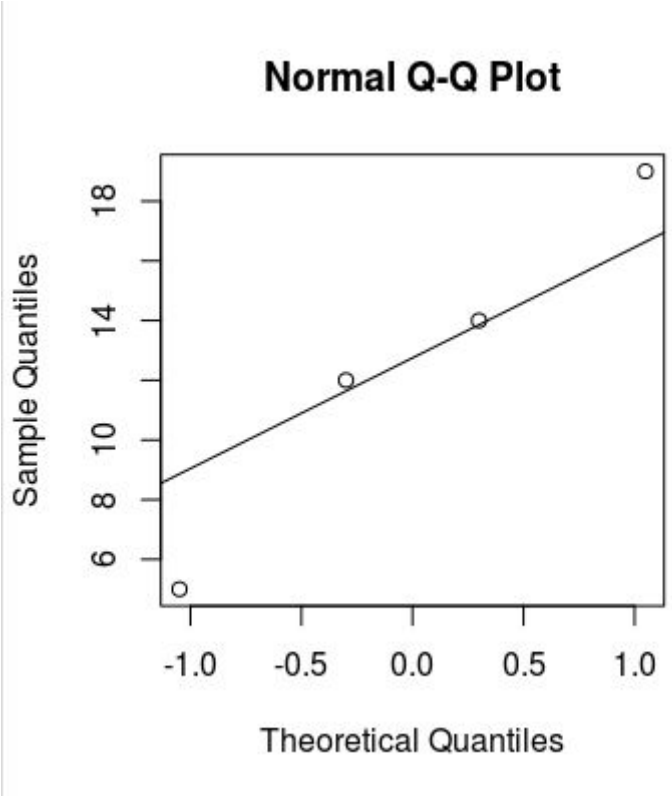


De la figura podemos ver que tenemos 4 clases: [20,40),[40,60),[60,80),[80,100], cuyas marcas correspondientes son: mi = 30,50,70,90 (i=1...4). Asi, k = 4.

Determinamos la frecuencia de cada clase (contando a mano), respectivamente:

fi = 12,19,14,5

Cargamos estos valores en R y hacemos un qqplot



Ahora, una vez teniendo los datos agrupados, para poder hacer nuestra prueba por hipótesis, necesitamos determinar la media y la varianza, que son los parámetros que recibe la distribución normal.

```
> prom = sum(fi*mi)/n
> prom
[1] 54.8
```

La media nos da 54.8

```
> varianza = sum((xi-prom_n)^2 )/(n-1)
>
> sd = sqrt(varianza)
> sd
[1] 19.00634
```

Calculamos la desviación estándar porque es el parámetro que recibe la función qnorm de R, y nos da 19.00634

Ahora, con estos dos valores y los límites superior e inferior de cada clase, podemos calcular la función de probabilidad de la distribución normal.

```
> pi = pnorm(lu, prom, sd) - pnorm (ll, prom, sd)
> pi
[1] 0.18452922 0.38971943 0.29975794 0.08374101
```

Ahora, podemos calcular el estadístico X^2

```
> chi2_est = sum((fi-n*pi)^2/(n*pi))
> chi2_est
[1] 1.068821
```

Y da como resultado 1.0688

Como no nos proveen un nivel de significancia alfa, calculamos el p-valor

```
> p_valor = 1 - pchisq(chi2_est, k-1-r)
> p_valor
[1] 0.3012119
```

y como p-valor = 0.3 > 0.1 tenemos evidencia para decir que no rechazamos la hipótesis nula, es decir, los datos se ajustan a una distribución normal.

CÓDIGO DEL LABORATORIO

```
# Laboratorio 3

# Alumnos: Arturo Yepez 15-11551, Antonella Requena 15-11196


### PREGUNTA 1


# Como nos piden establacer una relación entre nivel socioeconomico y la zona donde vive

# utilizamos Tablas de Contingencia


# Número de columnas

c = 2


# Número de categorias

r = 3


# Datos de los ambientes que tenemos

rural = c(249, 80, 2)

urbano = c(139, 20, 15)


# Agrupación de esos datos

matriz = cbind(rural, urbano)


# Hacemos chisq.test para sacar los resultados de la tabla de contingencia

chisq.test(matriz)


### PREGUNTA 2


# Debemos realizar una prueba para proporciones
```

```
# Nivel de significancia
alpha = 0.05

# Cantidad total de datos
n = 1000

# Número de categorias

k = 4
r = 0

# Probabilidad con la que queremos usar H0
p = c(1/4, 1/4, 1/4, 1/4)

# Datos proporcionados según la categoria
fi = c(294, 276, 238, 192)

# Hacemos prop.test para verificar el resultado
prop.test(fi, c(1000,1000,1000,1000), p)

### PREGUNTA 3

# Como tenemos datos no agrupados, vamos a dividir las clases
# para obtener la frecuencia
n = 50
alpha = 0.05
r = 2

xi = c(32, 23, 64, 31, 74, 44, 61, 33, 66, 73,
      27, 65, 40, 54, 23, 43, 58, 87, 58, 62,
      68, 89, 93, 24, 73, 42, 33, 63, 36, 48,
```

```
      77, 75, 37, 59, 70, 61, 43, 68, 54, 29,
      48, 81, 57, 97, 35, 58, 56, 58, 57, 45)

hist(xi, main = "Histograma de Edades", ylab = "Frecuencia", xlab
= "Edades", col = "blue")

# Ahora, agrupando los datos
k = 4
fi = c(12,19,14,5)
mi= c(30, 50, 70, 90)
lu = c(40, 60, 80, 100)  # limite superior de cada clase
ll = c(20, 40, 60, 80)   # limite inferior de cada clase


# Hacemos qq-plot
qqnorm(fi)
qqline(fi)

# Determinamos el promedio
prom = sum(fi*mi)/n

# Ahora la varianza
prom_n = rep(prom,n)

varianza = sum((xi-prom_n)^2 )/(n-1)

# Desviacion estandar
sd = sqrt(varianza)

# Una vez que tenemos varianza y promedio podemos calcular la
probabilidad pi
```

```
pi = pnorm(lu, prom, sd) - pnorm (ll, prom, sd)


# Calculamos nuestro estadistico Chi-cuadrado

chi2_est = sum((fi-n*pi)^2/(n*pi))


# Como no se nos da el nivel de significancia alpha, calculamos el
p-valor

p-valor = 1 - pchisq(chi2_est, k-1-r)
```