Tarea2

Luis Mantilla

2023-09-26

Contents

1 Diferencia de medias **1**

1 Diferencia de medias

En la página Link se descararga las bases de datos del Censo de Habitantes de Calle 2020 y 2021. Realizaremos los siguientes numerales en lenguaje R:

```
library(readr)
CHC_2020 <- read_delim("CHC_2020.csv", delim = ";", escape_double = FALSE, trim_ws = TRUE)
```

```
# Rows: 5043 Columns: 130
# -- Column specification --------------------------------------------------------
# Delimiter: ";"
# chr   (3): P1, P1S1, P2S1
# dbl (118): DIRECTORIO, TIP_FOR, P2, P5, CTL_1, P8R, P9, P10R, P11R, P12, P13...
# lgl   (9): P17S1, P20, P23, P26, P30, P32S4, P32S5, P33, P33_2
#
# i Use `spec()` to retrieve the full column specification for this data.
# i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
CHC_2021 <- read_csv("CHC_base_anonimizada09-09-2021.csv")
```

```
# Rows: 6250 Columns: 130
# -- Column specification --------------------------------------------------------
# Delimiter: ","
# chr   (3): P1, P1S1, P2S1
# dbl (125): DIRECTORIO, TIP_FOR, P2, P5, CTL_1, P8R, P9, P10R, P11, P12, P13,...
# lgl   (2): P23, P32S4
#
# i Use `spec()` to retrieve the full column specification for this data.
# i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Tomemos la columna de la edad de cada base de datos:

```
Edad_2020=CHC_2020$P8R
Edad_2021=CHC_2021$P8R
#Quitemos los datos faltante
Edad_2020=na.omit(Edad_2020)
Edad_2021=na.omit(Edad_2021)
```

1. Haga las pruebas de hipótesis correspondientes para establecer si las variables aleatorias de edad de un habitante de calle en el 2020 y en el 2021 presentan normalidad o no. Veamos si los datos presentan

normalidad usando una prueba de Anderson-Darlinng ya que tenemos más de 50 datos:

```
library("nortest")
ad.test(Edad_2020)
```

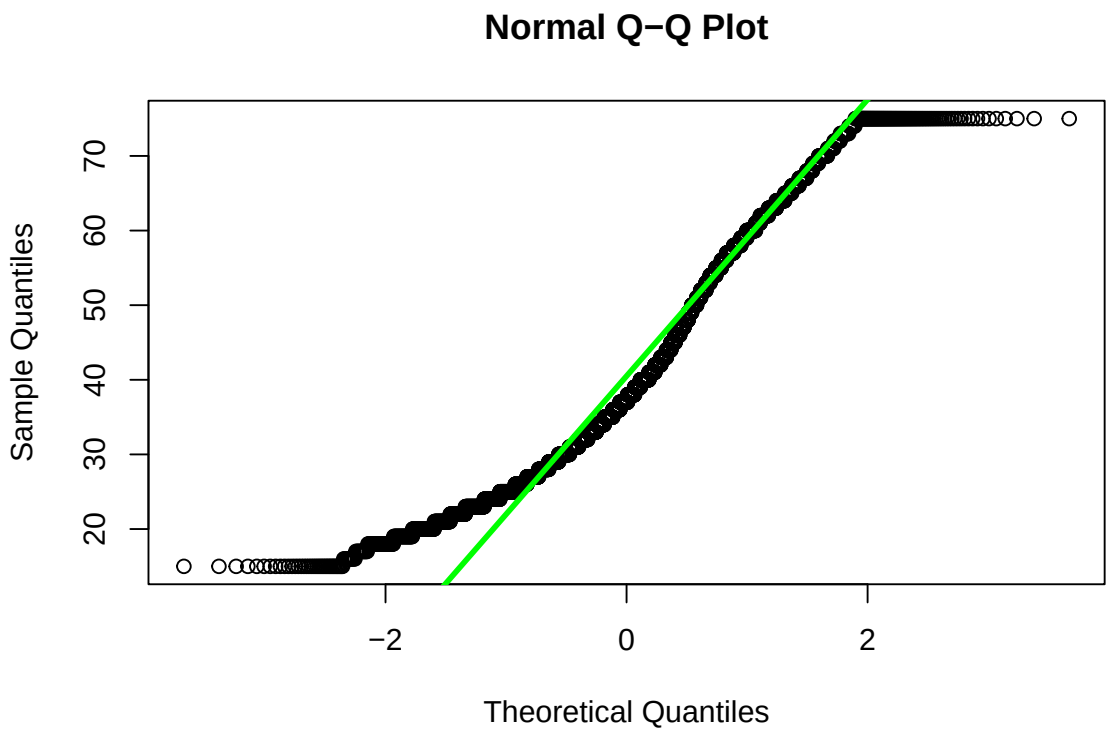
```
#
#   Anderson-Darling normality test
#
# data:  Edad_2020
# A = 69.827, p-value < 2.2e-16
```

```
ad.test(Edad_2021)
```

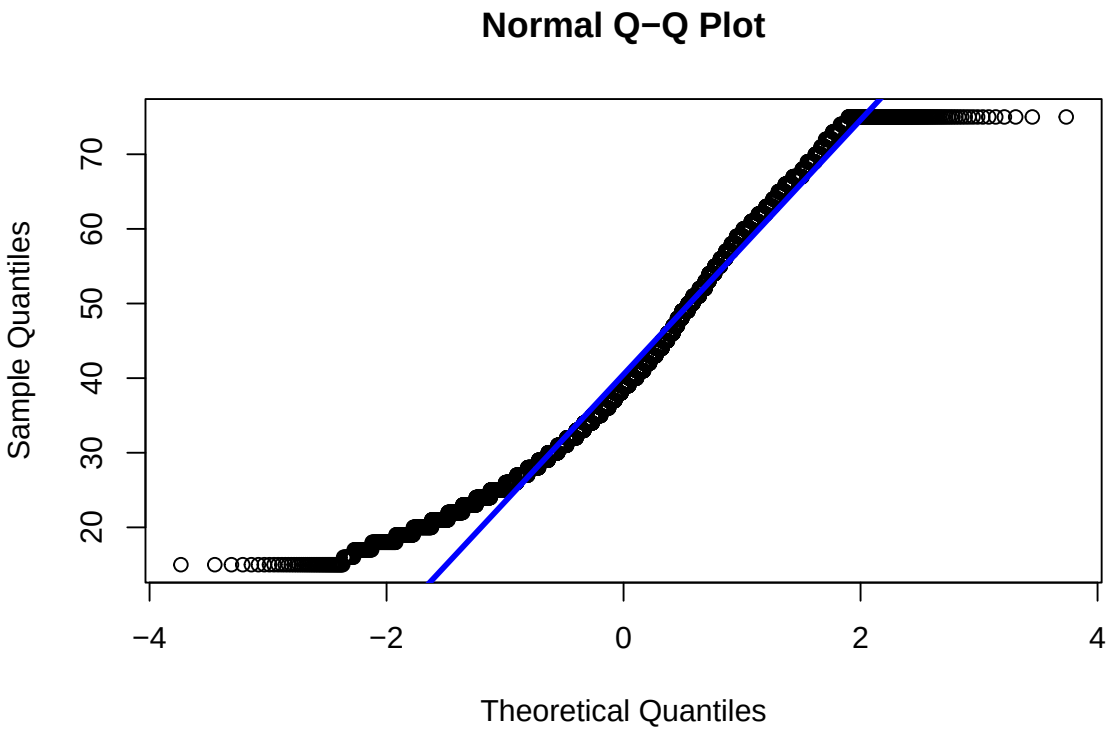
```
#
#   Anderson-Darling normality test
#
# data:  Edad_2021
# A = 63.114, p-value < 2.2e-16
```

Como el valor p para ambas pruebas es menor al 0.05 entonces decimos que con una confianza del 95% la edad del 2020 y 2021 no presentan normalidad. Además veamos de una manera descriptiva, si los datos se aproximan a la normalidad:

- Veamos Las edades del 2020:



- Veamos Las edades del 2021:



2. Asumiendo los supuestos de misma población, independencia entre, independencia dentro y normalidad, encuentre e interprete el intervalo de confianza de la diferencia de medias de edad entre un habitante de calle en el 2020 y un habitante de calle en el 2021. No olvide verificar si se tiene o no igualdad de varianzas a nivel poblacional.

   Sol: Como asumimos normalidad, entonces veamos si mantienen igualdad de varianza: Usando una distribución F , calculemos y las hipotesis

$$\begin{array}{ll} H_0: & \sigma_X^2/\sigma_Y^2 = 1 \\ Ha: & \sigma_X^2/\sigma_Y^2 \neq 1 \end{array}.$$

   Calculamos el F_c (F calculado):

```
Fc= var(Edad_2020)/var(Edad_2021)
Fc
```

```
# [1] 1.040514
```

   Y el intervalo de confianza del 95% es:

```
c(qf(0.05/2, length(Edad_2020) - 1, length(Edad_2021) - 1),
  qf(1-0.05/2, length(Edad_2020) - 1, length(Edad_2021) - 1))
```

```
# [1] 0.9442442 1.0588320
```

   Como Fc se encuentra dentro del intervalo de confianza, entonces asumimos con una confianza del 95% que las varianzas son iguales. Entonces desvariamos calcular el intervalo de confianza para la diferencia de medias, para esto usamos la distribución t:

$$H_0 : \mu_{2020} - \mu_{2021} = 0$$

$$H_a : \mu_{2020} - \mu_{2021} \neq 0$$

```r
t.test(Edad_2020, Edad_2021, var.equal = T)
```

```
#
#   Two Sample t-test
#
# data:  Edad_2020 and Edad_2021
# t = -1.6326, df = 9516, p-value = 0.1026
# alternative hypothesis: true difference in means is not equal to 0
# 95 percent confidence interval:
#  -1.1447408  0.1043926
# sample estimates:
# mean of x mean of y
#  40.77030  41.29047
```

Como el cero se encuentra en el intervalo de confianza del 95%, entonces decimos que se asume que las medias son iguales con una confianza del 95%.

3. Asumiendo los supuestos de misma población, independencia entre e independencia dentro, encuentre e interprete el intervalo de confianza de la diferencia de medias de edad entre un habitante de calle en el 2020 y un habitante de calle en el 2021.

   Sol: Por el ejercicio 1, no podemos asumir normalidad en los datos, entonces debemos hacer una prueba de wilcoxon

```r
wilcox.test(Edad_2020, Edad_2021, conf.level = 0.95, conf.int = T)
```

```
#
#   Wilcoxon rank sum test with continuity correction
#
# data:  Edad_2020 and Edad_2021
# W = 10855479, p-value = 0.02503
# alternative hypothesis: true location shift is not equal to 0
# 95 percent confidence interval:
#  -1.000032e+00 -7.679469e-05
# sample estimates:
# difference in location
#             -0.9999971
```

Dado que el intervalo es de confianza es $(-1, -0.00007)$ entonces decimos que las edades del 2021 son mayores que las edades del 2020 a lo sumo 1 año, con una confianza del 95%.

4. Pruebe que la edad de primer consumo de cigarrillo en un habitante de calle de Bogotá es igual a la edad de primer consumo de Marihuana, con una confianza del 90%, para el año 2021. Asuma misma población, independencia entre, independencia dentro y normalidad. No olvide verificar si se tiene o no igualdad de varianzas a nivel poblacional.

   Sol: Tomemos los datos

```r
CHC_2021=data.frame(CHC_2021)

datos=data.frame(cigarrillo=CHC_2021$P30S1A1, marihuana=CHC_2021$P30S3A1)

datos=na.omit(datos)
```

Veamos si la varianza son iguales:

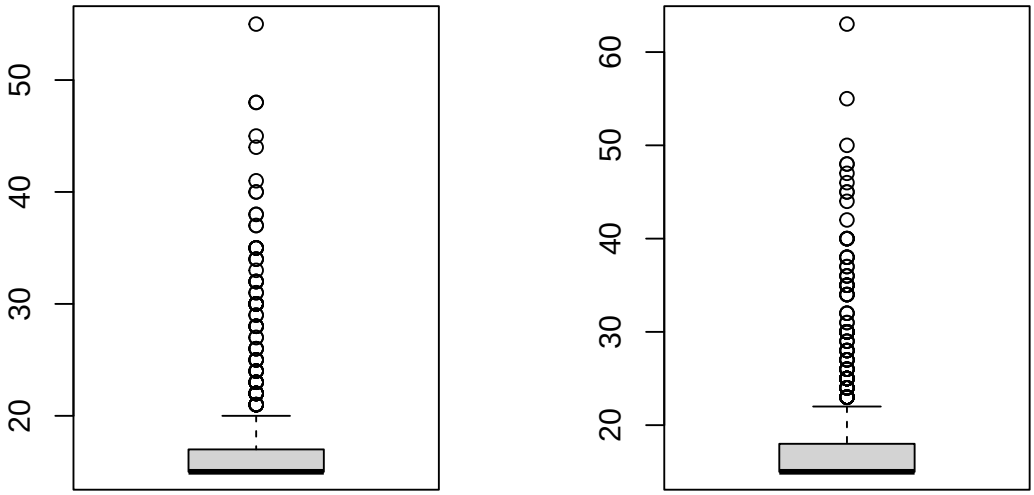
```r
Fc= var(datos$cigarrillo)/var(datos$marihuana)
Fc
```

```
# [1] 0.6337432
```

```r
c(qf(0.1/2, length(datos$cigarrillo) - 1, length(datos$marihuana) - 1),
  qf(1-0.1/2, length(datos$cigarrillo) - 1, length(datos$marihuana) - 1))
```

```
# [1] 0.9225664 1.0839329
```

Con una confianza del 90% rechazamos la hipotiposis nula, es decir rechazamos que las varianzas sean iguales. Por otro lado, de una manera descriptiva veamos si las cajas,en los diagramas de caja se parecen



Vemos que las cajas tienen diferente tamaño. Entonces asumiendo que las varianzas no son iguales, calculamos el intervalo de confianza para la diferencia de medias con varianzas distintas:

```r
t.test(datos$cigarrillo, datos$marihuana, var.equal = F, conf.level = 0.90)
```

```
#
#   Welch Two Sample t-test
#
# data:  datos$cigarrillo and datos$marihuana
# t = -4.1888, df = 3174.5, p-value = 2.881e-05
# alternative hypothesis: true difference in means is not equal to 0
# 90 percent confidence interval:
#  -0.8767609 -0.3822319
# sample estimates:
# mean of x mean of y
#  16.61751  17.24700
```

Con una confianza del 90% rechazamos que las medias sean iguales. Es decir rechazamos que las edades de inicio de conumo de cigarrillo y marihuana sean iguales, más aun por el intervalo de confianza $(-0.87, -0.38)$ decimos que las edades iniciales de consumo de marihuana es mayor entre 0.38 y 0.87 años que las edades del 2020.

5. Desarrolle este ejercicio MANUALMENTE: En un grupo de 12 personas se mide el cambio de ritmo cardiaco antes de levantarse y antes de acostarse. Haga las correspondientes pruebas de hipótesis utilizando la siguiente tabla:

Antes de levantarse	-2	4	8	25	-5	16	3	1	12	17	20	9	15
Antes de acostarse	-3	-2	7	20	-3	17	7	-1	13	15	22	8	15

Pruebe la hipótesis de que el ritmo cardíaco de las personas es igual antes de levantarse al ritmo cardíaco de antes de acostarse con una confianza del 95%. Concluya. Para responder este numeral vea el vídeo wilcoxon.wmv subido en Google Classroom.

Sol: Tomemos los datos:

```
AL=c(-2,4,8,25,-5,16,3,1,12,17,20,9)
AC=c(-3,-2,7,20,-3,17,7,-1,13,15,22,8)
```

Ahora se hace la diferencia de uno por uno

```
#  [1]  1  6  1  5 -2 -1 -4  2 -1  2 -2  1
```

Ahora el absoluto de cada termino:

```
#  [1] 1 6 1 5 2 1 4 2 1 2 2 1
```

Ahora agrupando todo en una matriz, tenemos en la primera columna la diferencia, luego el valor absoluto de la diferencia, en la tercer columna el R^+ y en la ultima $S(x_i - \theta_0)$:

```
#       [,1] [,2] [,3] [,4]
#  [1,]    1    1  3.0    1
#  [2,]    6    6 12.0    1
#  [3,]    1    1  3.0    1
#  [4,]    5    5 11.0    1
#  [5,]   -2    2  7.5    0
#  [6,]   -1    1  3.0    0
#  [7,]   -4    4 10.0    0
#  [8,]    2    2  7.5    1
#  [9,]   -1    1  3.0    0
# [10,]    2    2  7.5    1
# [11,]   -2    2  7.5    0
# [12,]    1    1  3.0    1
```

Ahora calculamos el estadistico de prueba W_s :

```
S=0
for (i in 1:12){
  S=S+ R[i,3]*R[i,4]
}
S
```

```
# [1] 47
```

Entonces $W_c = 47$, con base en tabla de wilcoxon

	alpha values						
n	0.001	0.005	0.01	0.025	0.05	0.10	0.20
5	--	--	--	--	--	0	2
6	--	--	--	--	0	2	3
7	--	--	--	0	2	3	5
8	--	--	0	2	3	5	8
9	--	0	1	3	5	8	10
10	--	1	3	5	8	10	14
11	0	3	5	8	10	13	17
12	1	5	7	10	13	17	21
13	2	7	9	13	17	21	26
14	4	9	12	17	21	25	31
15	6	12	15	20	25	30	36
16	8	15	19	25	29	35	42
17	11	19	23	29	34	41	48
18	14	23	27	34	40	47	55
19	18	27	32	39	46	53	62
20	21	32	37	45	52	60	69
21	25	37	42	51	58	67	77
22	30	42	48	57	65	75	86
23	35	48	54	64	73	83	94
24	40	54	61	72	81	91	104
25	45	60	68	79	89	100	113
26	51	67	75	87	98	110	124
27	57	74	83	96	107	119	134

	alpha values						
n	0.001	0.005	0.01	0.025	0.05	0.10	0.20
28	64	82	91	105	116	130	145
29	71	90	100	114	126	140	157
30	78	98	109	124	137	151	169
31	86	107	118	134	147	163	181
32	94	116	128	144	159	175	194
33	102	126	138	155	170	187	207
34	111	136	148	167	182	200	221
35	120	146	159	178	195	213	235
36	130	157	171	191	208	227	250
37	140	168	182	203	221	241	265
38	150	180	194	216	235	256	281
39	161	192	207	230	249	271	297
40	172	204	220	244	264	286	313
41	183	217	233	258	279	302	330
42	195	230	247	273	294	319	348
43	207	244	261	288	310	336	365
44	220	258	276	303	327	353	384
45	233	272	291	319	343	371	402
46	246	287	307	336	361	389	422
47	260	302	322	353	378	407	441
48	274	318	339	370	396	426	462
49	289	334	355	388	415	446	482
50	304	350	373	406	434	466	503

y con $\alpha = 0.05$, a dos colas, entonces la región critica está en los valores menores que 10, y dado que nuestro valor estimado es 47, no rechazamos la hipótesis nula, es decir tenemos igualdad de medianas o el ritmo cardíaco de las personas es igual al del levantarse y acostarse.