

MAESTRÍA EN CIENCIAS DE LA INFORMACIÓN Y LAS TELECOMUNICACIONES

Herramientas matemáticas para el manejo de la información

Taller #2

Salazar Ortiz, Jaiver

jesalazaro@correo.udistrital.edu.co 20221495012

Forero Castro, Diego

ddforefoc@correo.udistrital.edu.co 20221495005

FECHA: 16 de enero de 2024

FECHA DE ENTREGA: 23 de abril de 2022

1. Problema

OBJETIVO El siguiente enunciado ha sido dado por el taller:

"Tarea:Çon respecto a los inmuebles del sector B, estimar el porcentaje (proporción) y error están dar de inmuebles residenciales que hace parte del estrato bajo entre de 60 y 80 (inclusive) metros cuadrados construidos

Sector A						
Inmueble	tipo	estrato	numero habitantes	metros cuadrados		
1	Residencial	Medio	3	67,11		
2	Residencial	Bajo	7	65,21		
3	Residencial	Bajo	1	75,96		
4	No Residencial	Alto	1	71,34		
5	Residencial	Alto	5	67,8		
6	Residencial	Bajo	5	67,82		
7	No Residencial	Alto	2	72,82		
8	Residencial	Alto	4	80,56		
9	Residencial	Bajo	2	74,38		
10	Residencial	Bajo	3	72,72		
91	Residencial	Medio	5	67,11		
92	Residencial	Bajo	3	65,21		
93	Residencial	Bajo	4	75,96		
94	No Residencial	Alto	3	71,34		
95	Residencial	Alto	3	67,8		
96	Residencial	Bajo	6	67,82		
97	Residencial	Bajo	3	73,72		
98	Residencial	Bajo	4	74,68		
99	Residencial	Bajo	3	69,19		
100	No Residencial	Alto	3	72,82		

Tabla 1.1: INFORMACIÓN DEL SECTOR A

Sector B						
Inmueble	tipo	estrato	numero habitantes	metros cuadrados		
1	Residencial	Bajo	3	88,46		
2	Residencial	Bajo	2	80,78		
3	Residencial	Bajo	7	93,28		
4	No Residencial	Alto	7	67,11		
5	Residencial	Alto	5	65,21		
6	Residencial	Bajo	7	75,96		
7	Residencial	Bajo	2	71,34		
8	Residencial	Medio	3	67,8		
9	Residencial	Bajo	4	67,82		
10	No Residencial	Alto	7	72,82		
61	Residencial	Bajo	7	72,82		
62	No Residencial	Alto	7	80,56		
63	Residencial	Alto	5	74,38		
64	Residencial	Bajo	7	72,72		
65	Residencial	Bajo	2	59,4		
66	Residencial	Medio	3	68,55		
67	Residencial	Bajo	4	64,49		
68	No Residencial	Alto	7	65,10		
69	Residencial	Alto	5	65,95		
70	Residencial	Bajo	6	66,80		
71	Residencial	Bajo	3	67,65		

Tabla 1.2: INFORMACIÓN DEL SECTOR B

SOLUCIÓN 1 (PROPORCIÓN Y ERROR ESTÁNDAR) se desea estimar el porcentaje de inmuebles residenciales del sector b que pertenecen al estrato bajo y que el área de residencia se encuentre entre 60 y 80 m^2 , a partir de esto se debe determinar el error estándar. Tomando como referencia la tabla 1.2 del sector b se filtraron los datos que cumplieran con las condiciones establecidas, lo cual se logro determinar que existen 38 unidades residenciales de estrato bajo con áreas entre 60 y $80\,m^2$ de construcción, a partir de esto se logra determinar que el 53,5 % de los 71 unidades residenciales cumplen las características descritas, por lo tanto la proporción sera.:

$$\widehat{p} = \frac{n_i}{N} * 100\% = \frac{38}{71} * 100\% = 53,5211\%$$
(1.1)

por lo tanto el \hat{p} es 53.5211 %, por lo tanto para calcular el error estándar de la proporción a partir de la siguiente ecuación

$$S_{\widehat{p}} = \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$$

$$S_{\widehat{p}} = \sqrt{\frac{0,535211(1-0,535211)}{71}}$$

$$S_{\widehat{p}} = 0,05919175887556613$$
(1.2)

SOLUCIÓN 2 (ESTIMACIÓN DE PROPORCIÓN CON INTERVALO DE CONFIANZA) . Proponga un intervalo de confianza del 97.39 %, para establecer la pro porción de inmuebles de tipo residencial y que al mismo tiempo sea de estrato medio para el sector A. A partir de esta afirmación debemos recurir a las expreciones de intervalo de confianza.

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \tag{1.3}$$

esto se representa como

$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$
 (1.4)

Si la distribución es normal, se sigue que

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1) \tag{1.5}$$

En una distribución $Z \sim N(0,1)$ puede calcularse fácilmente un intervalo dentro del cual caiga un determinado porcentaje de las observaciones, esto es, es sencillo hallar z_1 y z_2 tales que $P[z_1 \leq Z \leq z_2] = 1 - \alpha$ donde $(1-\alpha)100\%$ es el porcentaje deseado. En esta distribución normal de medias se puede calcular el intervalo de confianza donde se encontrará la media poblacional si solo se conoce una media muestral \bar{x}), con una confianza determinada.

Habitualmente se manejan valores de confianza del 95 y del 99 por ciento. A este valor se le llamará $1-\alpha$ (debido a que α es el error que se cometerá, un término opuesto Para ello se necesita calcular el punto $Z_{\alpha/2}$ o valor critico.

Dicho punto es el número tal que:

$$P[z \ge z_{\alpha/2}] = \alpha/2 \tag{1.6}$$

Y en la versión estandarizada se cumple que: $z_{-\alpha/2} = -z_{\alpha/2}$ Así:

$$P\left[\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right] = 1 - \alpha \tag{1.7}$$

De lo cual se obtendrá el intervalo de confianza

$$\left(\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) \tag{1.8}$$

Obsérvese que el intervalo de confianza viene dado por la media muestral (\bar{x}) el producto del valor critico $Z_{\alpha/2}$ por el error estándar $\frac{\sigma}{\sqrt{n}}$.

Aproximaciones para el valor $z_{\alpha/2}$ para los niveles de confianza estándar son 1,96 para $1-\alpha=95\,\%$ y 2,576 para $1-\alpha=99\,\%$.

El intervalo de confianza para estimar una proporción p, conocida como una proporción muestral " p_n " de una muestra de tamaño n, a un nivel del $(1-\alpha)100\%$ de confianza es:

$$\left(p_n - z_{\alpha/2} \sqrt{\frac{p_n(1-p_n)}{n}}, \ p_n + z_{\alpha/2} \sqrt{\frac{p_n(1-p_n)}{n}}\right)$$
 (1.9)

Apartir de este corto resumen podemos determinar la estimación para el intervalo de confianza de 97.39 % para establecer la proporción de inmuebles de tipo residencial de estrato medio para el sector A. Para tal fin, a partir de la Tabla 1.1, tenemos inicialmente que 11 inmuebles residenciales y de estrato medio. por lo tanto la proporción es:

$$\widehat{p} = \frac{n_i}{N} * 100\% = \frac{11}{100} * 100\% = 11\%$$
(1.10)

Apartir de la ecuacion 1.9 se procede a calcular el intervalo de confianza asumiendo normalidad.

$$\left(\widehat{p}-z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}},\ \widehat{p}+z_{\alpha/2}\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\right) \tag{1.11}$$

Donde valor critico $Z_{\alpha/2}$ y esta determinado por

$$P[z \ge z_{\alpha/2}] = \frac{1 + (1 - \alpha)}{2}$$

$$P[z \ge z_{\alpha/2}] = \frac{1 + 0.9739}{2}$$

$$P[z \ge z_{\alpha/2}] = 0.98695$$
(1.12)

Donde 0.98695 es valor localizado en la normal N(0,1) a partir de esto se calcula $z_{\alpha/2}$ cuyo valor es 2.22

$$\begin{pmatrix} 0.11 - 2.22\sqrt{\frac{0.11(1 - 0.11)}{100}}, & 0.11 + 2.22\sqrt{\frac{0.11(1 - 0.11)}{100}} \end{pmatrix}$$

$$(0.11 - 2.22(0.03129), & 0.11 + 2.22(0.03129))$$

$$(0.11 - 0.06946, & 0.11 + 0.06946)$$

$$(0.04053847, & 0.17946152)$$

Donde se obtiene el intervalo de confianza de (0,04053847, 0,17946152)

SOLUCIÓN 3 (PRUEBA DE HIPÓTESIS PARA UNA PROPORCIÓN) ¿Es posible inferir que más del 75 % de los inmuebles del sector A tienen menos de 70 metros cuadrados? Para este proposito, se lleva a cabo una prueba chi-cuadrado (χ^2) de Pearson, para realizar la prueba de hipotesis para una proporcion. Para realizar la prueba, inicialmente se definen las hipotesis nula H_0 y alternativa H_1 , de esta forma:

$$H_0: p = 0.75$$

 $H_1: p > 0.75$ (1.14)

La aproximación z puede usarse para una prueba de hipótesis con una sola proporción $(p_0, \text{ observada})$, donde esta se compara con otra proporción conocida (\widehat{p}) , control o teórica. Sigue el mismo raciocinio ya visto donde z = diferencia de proporciones/error estándar, calculado aquí usando la proporción control:

$$Z = \frac{\widehat{p} - p_o}{\sqrt{\frac{p_0(1 - p_0)}{n}}} \tag{1.15}$$

Se determina los casos x exitosos, aquellos inmuebles residenciales que tienen menos de $70m^2$ en la muestra. se estableció que 46 unidades residenciales cumplen con esta condición sobre las 100 existentes . Por lo tanto la proporción es.

$$\widehat{p} = \frac{x}{n} * 100 \%$$

$$\widehat{p} = \frac{46}{100} * 100 \%$$

$$\widehat{p} = 46 \%$$
(1.16)

La prueba de hipotesis se lleva a cabo en RStudio, utilizando la funcion prop.test, como se muestra a continuacion:

```
> n<-100
> metros_A<-sector_a$ 'metros cuadrados '
> x<-length(metros_A[metros_A<70])
> prop.test(x, n, p = 0.75, alternative = 'greater')
$p.value
[1] 1
```

Se solicito en RStudio únicamente el p-valor para verificar o no la hipótesis nula. Como resultado, se obtuvo un p-valor = 1. Por lo tanto, debido al valor es igual a uno, la hipótesis nula no se puede rechazar, entonces, se acepta.

Como la hipótesis nula que se planteo es que la proporción de inmuebles con menos de 70 metros cuadrados en el sector A es menor o igual que 0.75, entonces, NO SE PUEDE DETERMINAR QUE EL 75% DE LAS UNIDADES RESIDENCIALES DEL SECTOR A TIENEN MENOS DE 70m.

 $SOLUCIÓN\ 4\ (COMPARACIÓN\ DE\ VARIANZA\ Y\ MEDIA\ DE\ DOS\ MUESTRAS)\ .$ Usando un a prueba de hipótesis con el 98.93 % de confianza e indicar si hay diferencias

significativas entre los metros cuadrados de los inmuebles de sector A y del sector B. (suponer normalidad) (No olvide estudiar la varianza de los metros cuadrados de cada sector para decidir que prueba usar.

Apartir de una prueba de hipotesis del 98.93% de confianza se prentende analisar si existen diferencias significativas apartir de la comparacion del area de las muestras del sectorA con respecto a las Sector B. Se debe suponer que son variables independientes con distribuciones normales y posteriormente relaizar una hipotesis sobre la similitud de las varianzas. Lo cual se debes definir la hipotesis nula H_0 y la alternativa H_1

$$H_0 = \sigma_A^2 = \sigma_B^2$$

$$H_1 = \sigma_A^2 \neq \sigma_B^2$$

Donde σ_A es la varianza poblacional de los metros cuadrados de los inmuebles en el sector A y σ_B es la varianza poblacional de los metros cuadrados de los inmuebles en el sector B.

```
> metros_A <— sector_a$ 'metros cuadrados '
> metros_B <— sector_b$ 'metros cuadrados '
> var.test(metros_A, metros_B)$p.value
[1] 0.08866177
```

SOLUCIÓN 5 (Modelo de Poisson para habitantes por inmueble) .

Se calcula la media del número de habitantes del sector B debido a que está es la estimación por máxima verosimilitud(MLE) de λ , también es necesario calcular el error estándar mediante la ecuación:

$$\sigma = \frac{s}{\sqrt{n}} \tag{1.17}$$

O alternativamente usando la biblioteca plotrix y la función std.error(). Por tanto al realizar esto en R:

```
> numeroHabitantes = sectorB$'numero habitantes'
> lamb = mean(numeroHabitantes)
> lamb
[1] 4.830986
> std.error(numeroHabitantes)
[1] 0.2045616
> sd(numeroHabitantes)/sqrt(length(numeroHabitantes))
[1] 0.2045616
```

Se puede concluir que el valor de habitantes por inmueble es de 4.8 con un error estándar de 0.2.

Para el calculo de probabilidad de que un inmueble tenga entre 3 y 5 personas usamos la distribución de probabilidad de Poisson, primero para la probabilidad de que habiten 5 habitantes o menos en el inmueble, luego la probabilidad para 3 habitantes o menos, como se quiere el intervalo de 3 a 5 simplemente restamos la probabilidad de 3 a 5, al realizar esto en R:

```
> probabilidad5 = ppois(5, lambda = lamb)
> probabilidad5
[1] 0.6455884
> probabilidad3 = ppois(3, lambda = lamb)
> probabilidad3
[1] 0.2895569
> probabilidad3_5 = probabilidad5 - probabilidad3
> probabilidad3_5
[1] 0.3560315
```

Por tanto se puede concluir que la probabilidad de que un inmueble tenga de 3 a 5 personas es de 0.3560315.

SOLUCIÓN 6 (COMPARACIÓN DE PROPORCIONES)

En este caso, se busca evidencia estadística para decidir si la proporción de inmuebles residenciales en el sector B, es 3 veces la proporción de inmuebles no residenciales. Entonces, se plantean las siguientes hipótesis nula H_0 y alternativa H_1 :

•
$$H_0 = p_{residencial} = 3p_{no-residencial}$$

■
$$H_1 = p_{residencial} \neq 3p_{no-residencial}$$

Donde $p_{residencial}$ es la proporción de inmuebles residenciales en el sector B y $p_{no-residencial}$ es la proporción de inmuebles no residenciales en el mismo sector. En este caso, se utiliza la estadística de prueba, considerando en este caso que:

$$\widehat{p} = \widehat{p}_{residencial} = \frac{57}{71} = 0,8028$$

$$p_0 = 3\widehat{p}_{residencial} = \frac{3*14}{71} = 05915$$
(1.18)

Notese que el total de inmuebles residenciales del sector B es 57, el total de inmuebles no residenciales es 14, y el total de inmuebles 71. Ahora, la prueba de hipótesis se lleva a cabo en RStudio, utilizando la función prop.test, como se muestra a continuación:

```
>residencial_B <- DATOS_SECTOR_B$ 'tipo '
>n_resi_B <- length (residencial_B
  [residencial_B=='Residencial'])
>n_noresi_B <- length (residencial_B
  [residencial_B=='NouResidencial'])

> prop.test(n_resi_B, n_B,
  p = 3*n_noresi_B/n_B, alternative = 'two.sided' $p.value
  [1] 0.0004637764
```

Se solicito en RStudio únicamente el p-valor para verificar o no la hipótesis nula. Como resultado, se obtuvo un p-valor = 0.0004638. Por lo tanto, debido al valor tan peque no, la hipótesis nula se puede rechazar con un nivel de confianza superior al 99.9 %. Por lo tanto, se obtuvo evidencia estadística para decidir que la proporción de inmuebles residenciales del Sector B NO es tres veces la proporción de inmuebles no residenciales.

SOLUCIÓN 7 (Proporción entre sector A y sector B) .

Es necesario realizar una prueba de proporciones para dos muestras. Primero es necesario hallar determinados valores como el número de coincidencias de 'No Residencial', tanto para el sector A como el B, además de ello el número total n.

```
NoRA=sectorA$tipo[which(sectorA$tipo == "No_Residencial")]
length(NoRA)
[1] 20
NoRB=sectorB$tipo[which(sectorB$tipo == "No_Residencial")]
length(NoRB)
[1] 14
tipoTotalA=sectorA$tipo
length(tipoTotalA)
[1]100
tipoTotalB= ectorB$tipo
length(tipoTotalB)
[1]71
x_1x_2 = (length(NoRA)+length(NoRB))
n1_n_2 = (length(tipoTotalA)+length(tipoTotalB))
P = x_1x_2/n1_n_2
P
[1] 0.1988304
```

En resumen se obtienen los siguientes datos: $x_a = 20$, $x_b = 14$, $n_a = 100$, $n_b = 200$. Luego es necesario considerar la hipótesis nula:

$$H_0: p_1 = p_2$$
 (1.19)

Las hipótesis alternativas:

$$H_1: p_1 \neq p_2 \quad H_1: p_1 > p_2 \quad H_1: p_1 < p_2$$
 (1.20)

Calcular la desviación estándar de la proporción:

$$\sigma_p = \sqrt{p(1-p)[(1/n_1) + (1/n_2)]}$$
 (1.21)

Y el valor estadístico de la prueba z:

$$z = \frac{(x_1/n_1) - (x_2/n_2)}{\sigma_p} \tag{1.22}$$

Debido a que no se menciona el nivel de significación se va a asumir una significación del $0.05(5\,\%)$

$$\alpha = 0.05 \tag{1.23}$$

Por simplicidad se realizan los cálculos en R solo para observar si la hipótesis de que son equivalentes se cumple, esto se realiza mediante una función la cual viene por defecto con un nivel de confianza del 95 %:

res
$$\leftarrow$$
 prop.test(x = c(20, 14), n = c(100, 71)) res

Lo cual genera:

2-sample test for equality of proportions with continuity correction

data: c(20, 14) out of c(100, 71)X-squared = 6.4917e-31, df = 1, p-value = 1 alternative hypothesis: two.sided 95 percent confidence interval: -0.1212900 0.1269238 sample estimates:

prop 1 **prop** 2 0.2000000 0.1971831

como el p-valor generado por R es de 1, y por tanto $p-valor>\alpha$ se puede afirmar que la proporción entre el sector A y el sector B es muy similar.

2. SEGUNDA PARTE

SOLUCIÓN 1 (PROPORCIÓN ESTIMADA Y VARIABILIDAD)

El intervalo de confianza es calculado como la proporción \pm z por tanto dado que se tienen los intervalos de confianza 0.92 y 0.98 se puede decir que:

$$0.95 \pm 0.03$$
 (2.1)

Donde 0.95 es el valor de la proporción, y la variabilidad ± 0.03 .

SOLUCIÓN 2(Parámetro y error estándar)

Se calcula la media de los datos debido a que está es la estimación por máxima verosimilitud(MLE) de λ , también se calcula el error estándar de la media, realizando esto en R:

```
> data <- c(3,3,4,3,2,3,5,3,3,2,2,3,4,5)
> sd(data)
[1] 0.9749613
> std_mean <- function(data) sd(data)/sqrt(length(data))
> std_mean(data)
[1] 0.2605694
```

Por tanto $\lambda = 0.974$ y el error estándar es 0,260

SOLUCIÓN 3(sesgo, estimador, estimador insesgado) .

El sesgo esta definido como:

$$B(n) = H(n) - n \tag{2.2}$$

Dado que se solicita calcular un determinador insesgado basado en B y H, por tanto se plantea un nuevo sesgo usando el sesgo anterior como determinador por tanto:

$$B_u(n) = B(n) - n \tag{2.3}$$

Reemplazando B(n) y solucionando:

$$B_{u}(n) = H(n) - n - n$$

$$B_{u}(n) = H(n) - 2n$$
(2.4)

como:

$$n = -B(n) + H(n) \tag{2.5}$$

Se tiene que:

$$B_{u} = H(n) - 2(-B(n) + H(n))$$

$$B_{u} = H(n) + 2B(n) - 2H(n)$$

$$B_{u} = -H(n) + 2B(n)$$
(2.6)

Como el estimador tiene que ser positivo y el sesgo debe ser 0 para que sea insesgado:

$$-(0) = -(-H(n) + 2B(n))$$

$$0 = H(n) - 2B(n)$$
(2.7)

 $SOLUCIÓN\ 4 ({\rm ESTADÍSTICOS\ DE\ ORDEN}) \quad \hbox{En estadística , el k-ésimo orden estadístico de una muestra estadística es igual a su k-ésimo valor más pequeño. [1] Junto con las estadísticas de clasificación, las estadísticas de orden se encuentran entre las herramientas más fundamentales en las estadísticas e inferencias no paramétricas .$

Los casos especiales importantes de las estadísticas de orden son el valor mínimo y máximo de una muestra y (con algunas calificaciones que se analizan a continuación) la mediana de la muestra y otros cuantiles de la muestra .

Cuando se usa la teoría de la probabilidad para analizar las estadísticas de orden de muestras aleatorias de una distribución continua , la función de distribución acumulativa se usa para reducir el análisis al caso de las estadísticas de orden de la distribución uniforme .

Aparte de esta definición se observan o se registran una muestra de tamaño 11 si los valores de la muestra son:

- **(**2,3,3,2,3,4,3,2,3,4,5)
- **(**2,2,2,3,3,3,3,4,4,5)

los estadísticos de orden se denotarían

■ $x_1 = 2$, $x_2 = 2$, $x_3 = 2$, $x_4 = 3$, $x_5 = 3$, $x_6 = 3$, $x_7 = 3$, $x_8 = 3$, $x_9 = 4$, $x_{10} = 4$, $x_{11} = 5$

$$X_{(1)} = min(x_1, \dots, x_n)$$

$$\vdots \qquad \vdots \qquad \vdots$$

$$X_{(n)} = max(x_1, \dots, x_n)$$
(2.8)

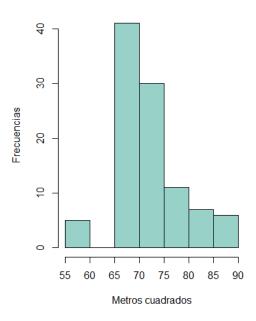
$$X_{(1)} \le X_{(2)} \le X_{(3)} \dots \le X_{(n-1)}$$
 (2.9)

El estadístico de primer orden (o estadístico de menor orden) es siempre el mínimo de la muestra, es decir, $X_{(1)} = min(x_1, ..., x_n)$ donde, siguiendo una convención común, usamos letras mayúsculas para referirnos a variables aleatorias y letras minúsculas (como arriba) para referirnos a sus valores reales observados.

De manera similar, para una muestra de tamaño n , el estadístico de orden n (o el estadístico de orden más grande) es el máximo , es decir, $X_{(n)} = \max(x_1, \dots, x_n)$ El rango de muestra es la diferencia entre el máximo y el mínimo. Es una función de las estadísticas del pedido: $rango(x_1, \dots, x_n) = x_{(n)} - x_{(1)}$

3. ANEXOS

Metros cuadrados

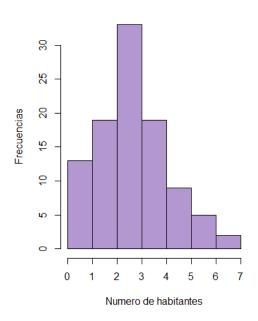


(a) numero de metros cuadrados sector a

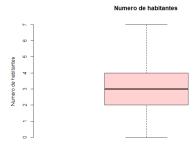
(b) numero de metros cuadrados sector a

Figura 3.1

Numero de habitantes



(a) numero de habitantes sector a



(b) Vista aérea

Figura 3.2: numero de habitantes sector a

Relacion de numero de habitantes por metro cuadrado

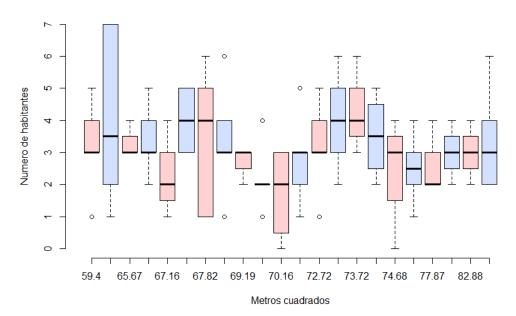
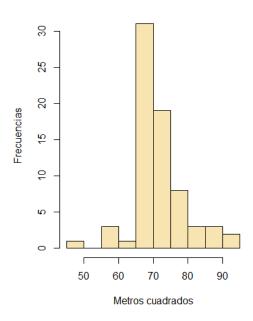


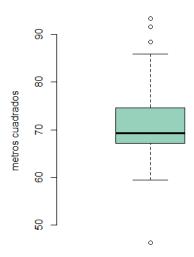
Figura 3.3

Metros cuadrados



(a) numero de metros cuadrados sector b

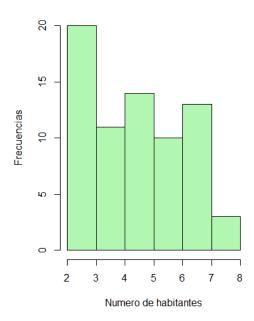
Metros cuadrados



(b) numero de metros cuadrados sector b

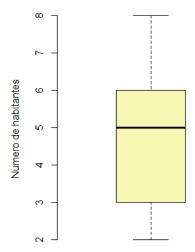
Figura 3.4

Numero de habitantes



(a) numero de habitantes sector b

Numero de habitantes



(b) numero de habitantes sector b

Figura 3.5

acion de numero de habitantes por metro cui

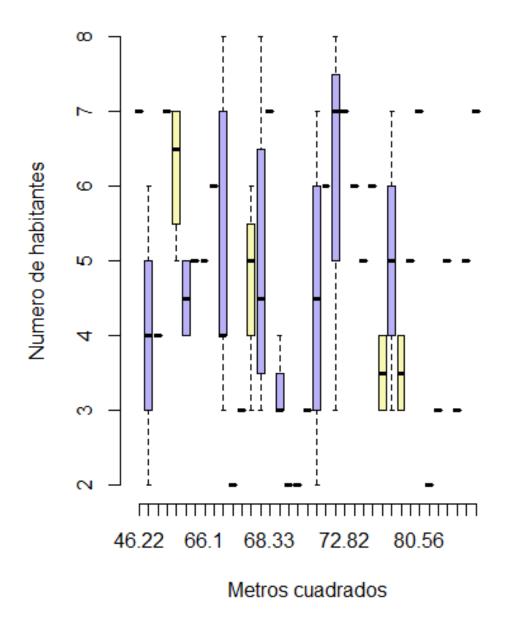


Figura 3.6