Análisis de las solicitudes de uniones de hecho por la población de castilla y león

Autor:

Gonzalo Rodríguez Castro

Grado:

Ingeniería en Sistemas de la Información

Fecha:

10/12/2024

Índice

1. Introducción

- 1.1 Contexto
- 1.2 Objetivo
- 1.3 alcance

2. Metodología

- 2.1 Poblaciones y propiedad estudiada
- 2.2 Muestras y origen de los datos
- 2.3 Normalidad de los datos
- 2.4 Hipótesis
- 2.5 Herramientas

3. Resultados

- 3.1 Normalidad de los datos
- 3.2 Análisis de homogeneidad de varianzas
- 3.3 Intervalos de confianza
- 3.4 Contrastes de hipótesis

4. Análisis de resultados
5. Conclusiones

Introducción

Contexto

En la comunidad de Castilla y León se solicitan anualmente miles de solicitudes de uniones de hecho, debido a la gran cantidad de solicitudes en el ultimo año, se han dividido el numero de solicitudes en dos grandes grupos, el primero se engloba con todas las solicitudes mensuales desde el año 2010 hasta el año 2016 y el segundo grupo contiene todas las solicitudes menusales que comprenden desde el año 2017 hasta el año 2023

Objetivo

El objetivo de la investigación es: "Realizar un análisis estadístico inferencial de las solicitudes de hecho en la comunidad de Castilla y León comparando las diferencias que existen respecto a número de solicitudes de los ultimos años con las solicitudes de años pasados".

Alcance

La investigacion se limita al número de solicitudes mensuales de uniones de hecho localizadas en Castilla y León en los años que comprenden entre 2010 y 2023.

Metodología

Poblaciones y propiedad estudiada

Las poblaciones sobre las que se realizará la inferencia son dos y se desconoce su tamaño:

- Uniones de hecho entre los años 2010 y 2016
- Uniones de hecho entre los años 2017 y 2023 La propiedad de las uniones que se han utilizado en el estudio son:
- Tipo de union (ya sea una union homosexual o heterosexual)

Muestras y origen de los datos

Se han seleccionado dos muestras de cada poblacion para realizar el estudio:

- Muestras grandes: Se ha seleccionado una muestra de 63 meses de la primera franja y una muestra de 63 meses de la segunda franja obteniendo todas las solicitudes realizadas en cada uno de los meses, en el que podemos diferenciar dos tipos de uniones de hecho, las uniones homosexuales y las uniones heterosexuales
- Muestras pequeñas: Para comparar resultados utilizando tambien muestras pequeñas, se han seleccionado aleatoriamente una muestra de 20 meses de la primera franja y 20 meses de la segunda franja

Se ha trabajado con un archivo "uniones_de_hecho.csv" que incluye todas las solicitudes presentadas en Castilla y León en los ultimos años. El archivo ha sido presentado por el ayuntamiento de la comunidad de Castilla y León: (https://datosabiertos.jcyl.es/web/jcyl/set/es/demografia/registro-parejas-hecho/1285090203163.)

Normalidad de los datos

Se realizaron evaluaciones de normalidad para las muestras grandes y pequeñas mediante pruebas gráficas y estadísticas. Los resultados no proporcionan evidencia suficiente para asumir normalidad en ninguno de los casos.

- Muestras Grandes
    - Gráficos QQ (Quantile-Quantile): Para ambas franjas temporales (primera y segunda), los diagramas QQ muestran desviaciones significativas respecto a la diagonal de normalidad. Esto indica que los datos no siguen una distribución normal de manera evidente.
    - Prueba de Jarque-Bera: Los resultados de la prueba rechazaron la hipótesis nula de normalidad en ambas franjas (primera y segunda) a un nivel de significación convencional (p<0.05).
- Muestras Pequeñas
    - Gráficos QQ: En los diagramas QQ de las muestras pequeñas (20 valores seleccionados aleatoriamente por franja), se observa un patrón no alineado con la diagonal, reforzando la falta de normalidad.
    - Prueba de Shapiro-Wilk: La prueba de Shapiro-Wilk confirmó que las muestras pequeñas tampoco cumplen con los criterios de normalidad, rechazando la hipótesis nula de distribución normal (p<0.05).
- Aproximación de los datos: Como los datos no tenían una distribución normal para aproximarlo a una distribución normal hubo que hacer el logaritmo de todos los datos seleccionados para poder aproximarlo, tras la transformacion de los datos todos los tests de normalidad pasaron correctamente

Hipótesis

La investigación pretende comprobar si se cumplen las siguientes hipotesis:

1. Hipótesis: Existe una diferencia significativa entre la media de las solicitudes de la primera franja y la media de la segunda franja.
2. Hipótesis: Existe una diferencia significativa entre la mediana de las solicitudes de la primera franja y la mediana de las solicitudes de la segunda franja.
3. Hipótesis: Existe una diferencia significativa entre la proporción de solicitudes de parejas homosexuales en la primera franja y la proporción de solicitudes de parejas homosexuales en la segunda franja.
4. Hipótesis: Existe una diferencia significativa entre la proporción de solicitudes de parejas heterosexuales en la primera franja y la proporción de solicitudes de parejas heterosexuales en la segunda franja.

Herramientas

Se han procesado los datos utilizando la aplicación RStudio para linux, versión 2024.09.0+375 y los paquetes: "BSDA" para utilizar la función "z.test" para calcular intervalos de confianza y contrastes de hipótesis sobre muestras grandes, "car" para dibujar diagramas QQ para la comprobación de normalidad de los datos y realizar el test de Levene sobre la homogeneidad de varianzas, y el paquete "tseries" para realizar el test de normalidad Jarque-Bera. El trabajo de informe estadístico ha sido realizado en el lenguaje de marcado ligero: Markdown, utilizando como entorno de desarrollo Visual Studio Code en su versión 1.94.2. Como control de versiones se ha utilizado GitHub: https://github.com/gonzalorg8799/Estadistica-Inferencial.

Resultados

Normalidad de los datos

Se ha realizado una comprobación visual para cada una de las muestras, obteniendo los histogramas de las figuras 1 y 2 que se aproximan a la distribución normal.

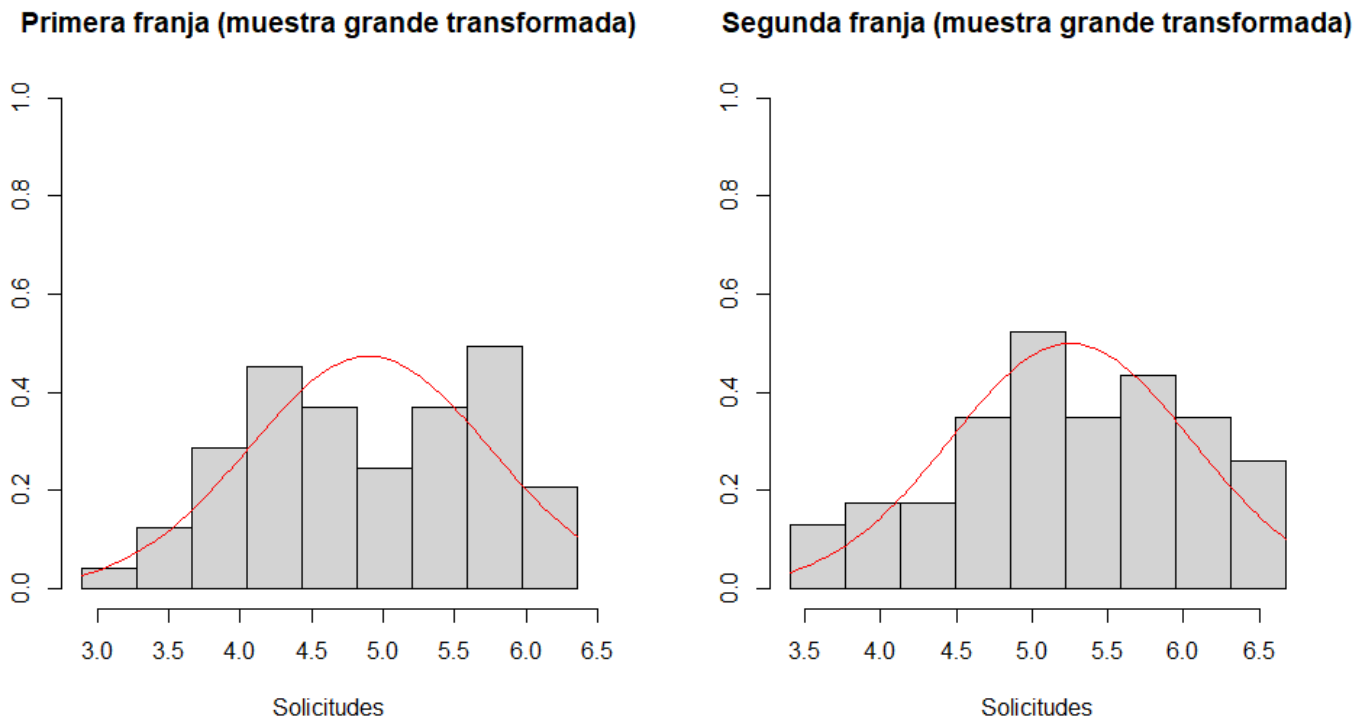

Figura 1

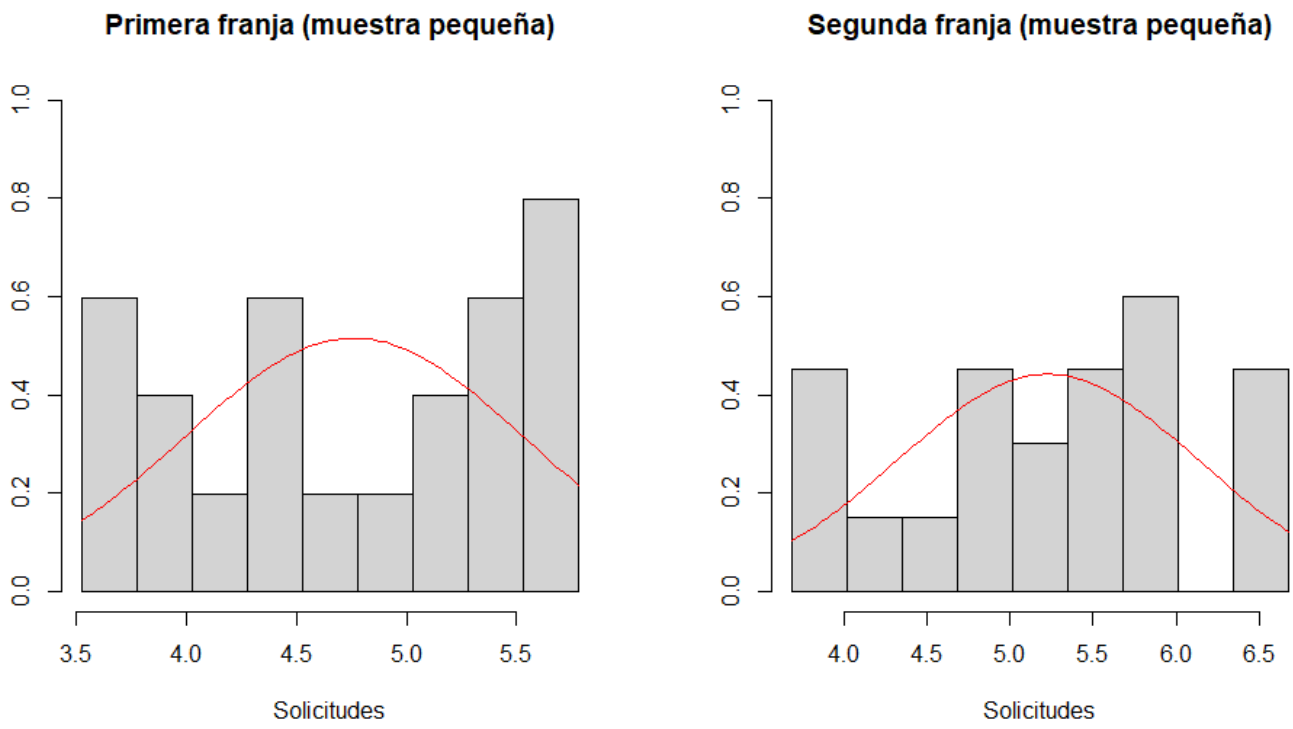

Figura 2

Tambien se han obtenido los diagramas QQ con región de aceptación del 95% que se representan en las figuras 3 y 4

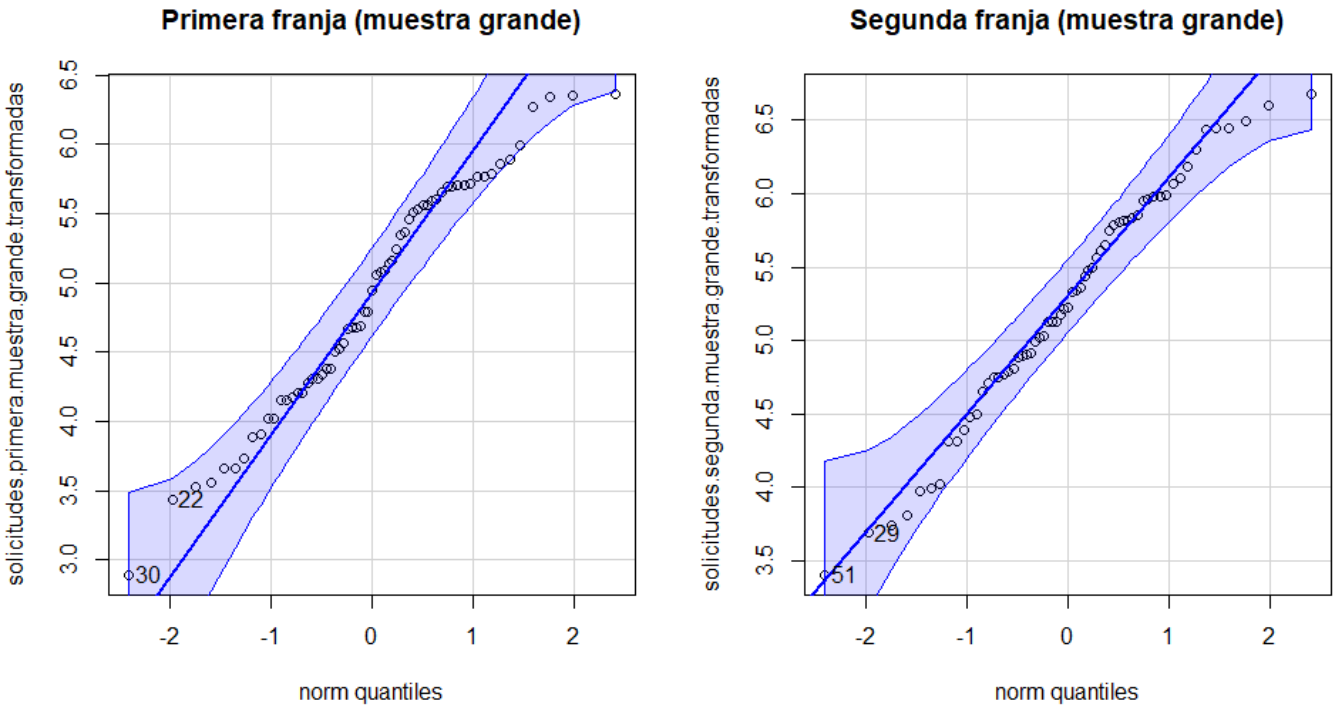

Figura 3

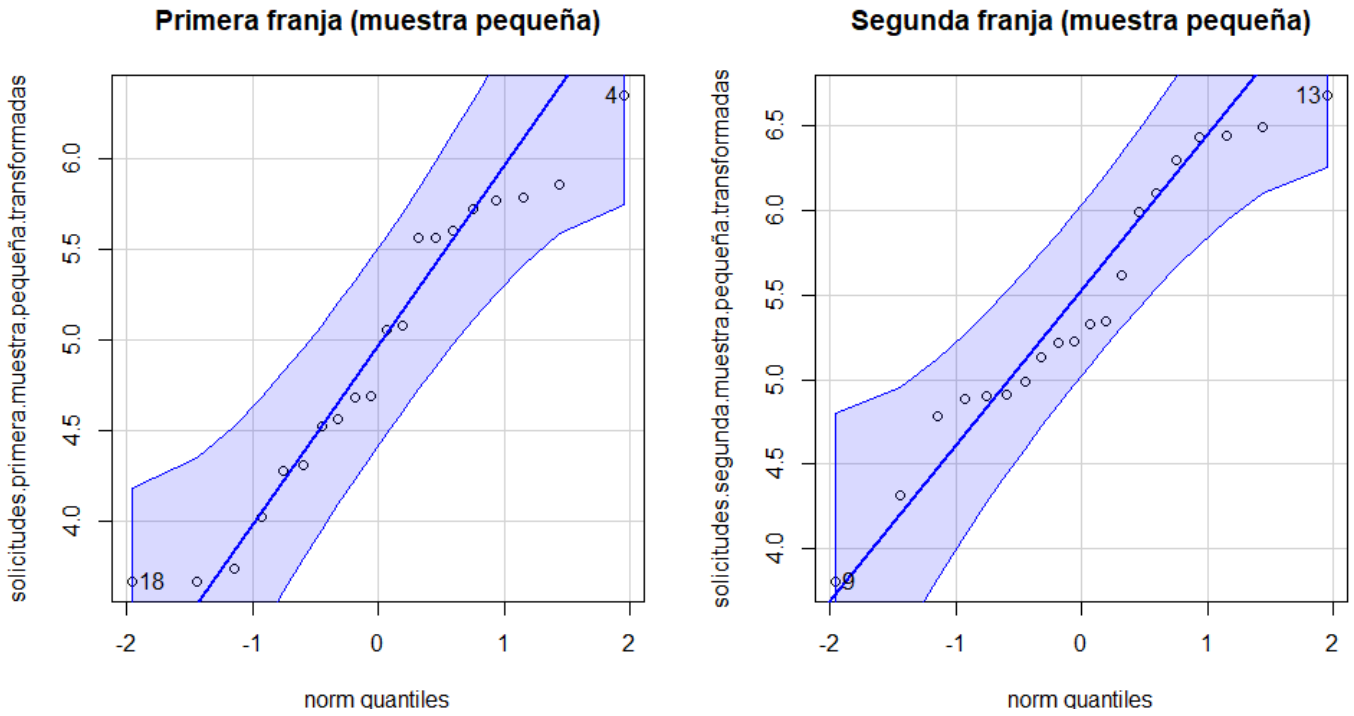

Figura 4

En la tabla 1 se muestran los resultados de los test de normalidad realizados despues de las transformaciones de los datos para cada una de las muestras

Test de Normalidad	Primera franja Muestra Grande	Segunda franja Muestra Grande	Primera franja Muestra Pequeña	Segunda franja Muestra Pequeña
Shapiro-Wilk	p-valor= 0.08548	p-valor= 0.3255	p-valor= 0.2182	p-valor= 0.337

Tanto en el caso de las muestras pequeñas como el de las muestras grandes, se ha superado el test de Shapiro-wilk, para un nivel de significación de 0,05 pues los p-valores superan dicho valor.

Análisis de homogeneidad de varianzas

Para evaluar si las varianzas de las poblaciones son iguales o diferentes, se llevó a cabo un test de Levene de homogeneidad de varianzas. En el caso de las muestras grandes, se obtuvo un p-valor = 0.3616, mientras que para las muestras pequeñas el p-valor fue = 0.5515. Dado que ambos valores son mayores al nivel de significación habitual de 0.05, se puede asumir que las varianzas de las franjas Primera y Segunda son iguales en ambos casos.

Intervalos de confianza

En la tabla 2 se muestran los resultados del cálculo de los intervalos de confianza utilizando las muestras grandes, con una confianza del 95%, es decir, con una significación de 0.05. Se han calculado intervalos para la media poblacional de las solicitudes de las parejas de cada franja, y para las proporciones de diferentes tipos de parejas:

- Parejas Homosexuales

- Parejas Heterosexuales

Medidas	**Primera Franja**	**Segunda Franja**
Tamaño muestra	63	63
Media (muestra)	185.1746	254.127
Mediana (muestra)	141	186
IC (95%) media	[109.42, 165.9]	[114095, 281838]
Prop. Parejas Heterosexuales (muestra)	2.63	2.61
IC (95%) Prop. Parejas Heterosexuales	[2.32, 2.99]	[2.17, 2.77]
Prop. Parejas Homosexuales (muestra)	1.24	44.83
IC (95%) Prop. Parejas Homosexuales	[1.16, 1.35]	[1.25, 1.47]

En la tabla 3 se muestran los mismos cálculos, pero al utilizar las muestras pequeñas

Medidas	**Primera Franja**	**Segunda Franja**
Tamaño muestra	20	20
Media (muestra)	25.11	38.90
Mediana (muestra)	16.59	40.73
IC (95%) media	[3388, 199526]	[117489, 630957]
Prop. Parejas Heterosexuales (muestra)	5.88	5.48
IC (95%) Prop. Parejas Heterosexuales	[3.86, 7.70]	[3.63, 7.30]
Prop. Parejas Homosexuales (muestra)	1.69	2.21
IC (95%) Prop. Parejas Homosexuales	[1.29, 2.58]	[1.58, 3.35]

En las figuras 5 a 7 se muestran los intervalos de confianza para la media, la proporcion de parejas homosexuales y la proporcion de parejas heterosexuales

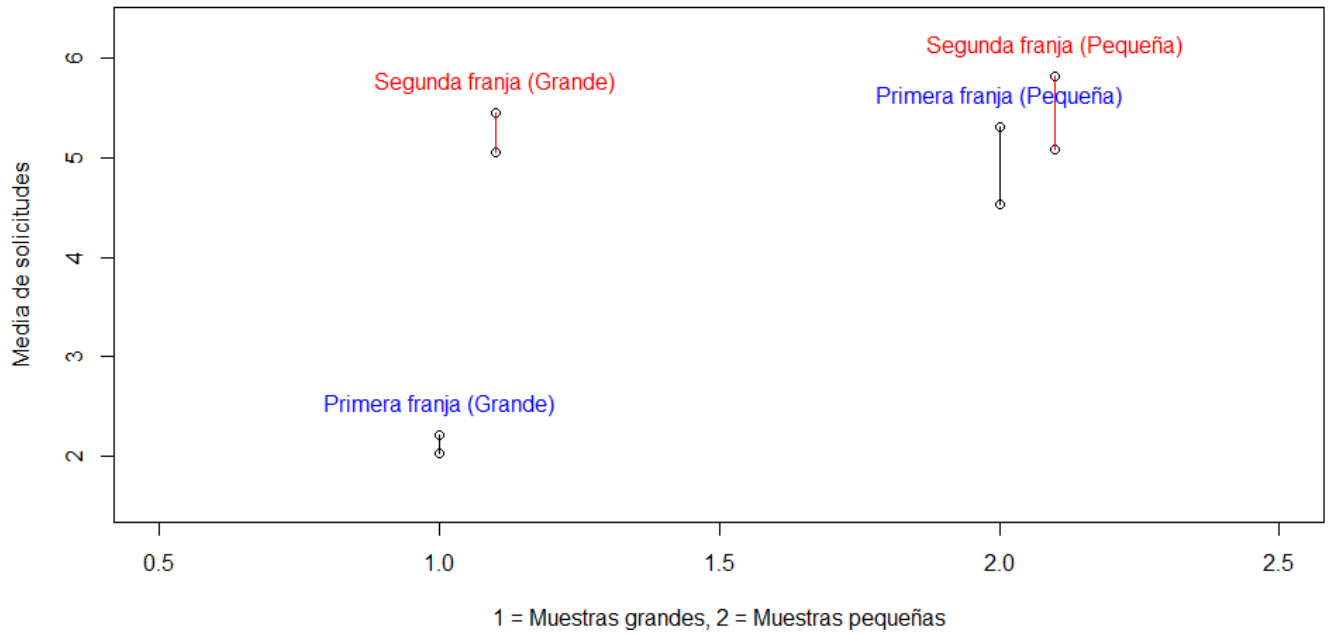

Figura 5

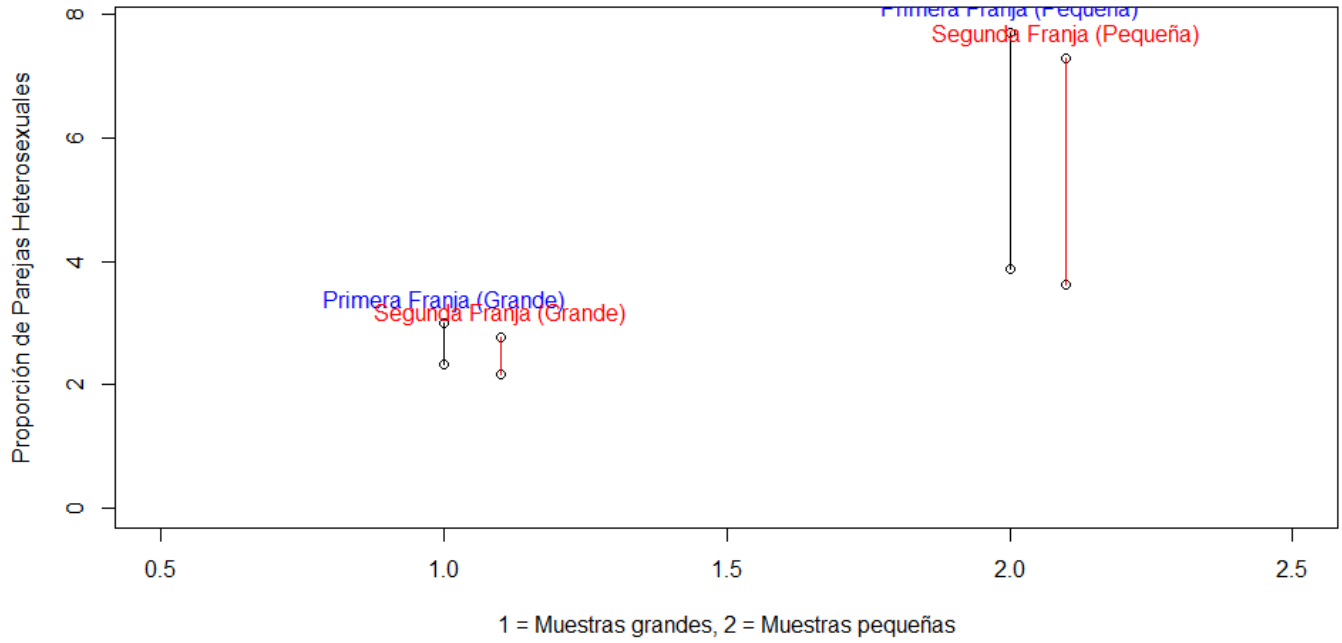

Figura 6

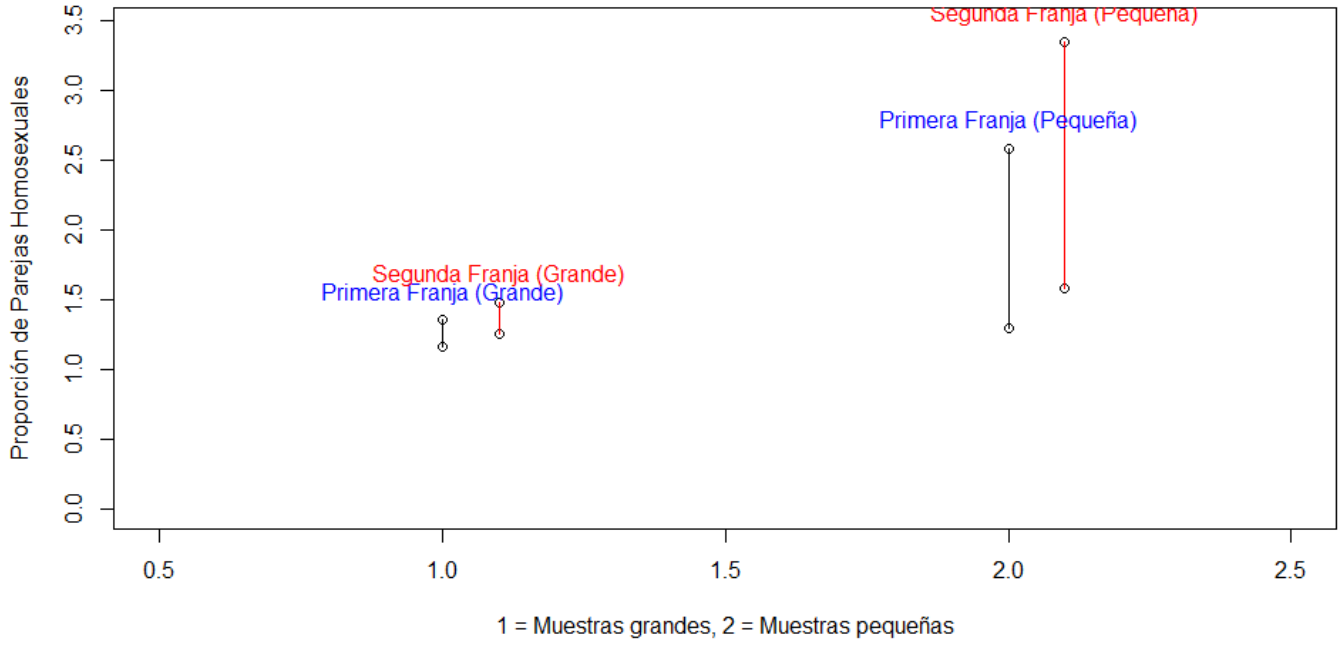

Figura 7

Contrastes de hipótesis

En la tabla 4 se muestran los resultados que se han obtenido, utilizando las muestras grandes, en los contrastes de hipótesis que se han planteado sobre la diferencia entre las medias y medianas poblacionales de las solicitudes de uniones de ambas franjas, y entre las proporciones de solicitudes de uniones homosexuales y heteresexuales.

tabla 4 Contraste de hipótesis usando muestras grandes

Hipótesis nula (H0)	Hipótesis alternariva (HA)	P-valor
Medias iguales	Media de la primera franja mayor que la media de la segunda	0.00000000000000022
Medianas iguales	Mediana de la primera franja menor que la mediana de la segunda franja	0.00000000000000022
Porporciones de heterosexuales iguales	Proporciones de heterosexuales diferentes	0.4204
Proporciones de homosexuales iguales	Proporciones de homosexuales diferentes	0.01357
Proporciones de homosexuales iguales	Proporciones de homosexuales de la primera franja menor que la proporcion de homosexuales en la segunda	0.006786

tabla 5 Contraste de hipótesis usando muestras pequeñas

Hipótesis nula (H0)	Hipótesis alternariva (HA)	P-valor
Medias iguales	Media de la primera franja menor que la media de la segunda	0.02403
Medianas iguales	Mediana de la primera franja menor que la mediana de la segunda franja	0.02825
Porporciones de heterosexuales iguales	Proporciones de heterosexuales diferentes	1
Proporciones de homosexuales iguales	Proporciones de homosexuales diferentes	0.3851

Análisis de resultados

A partir de los resultados obtenidos y las pruebas realizadas, se han identificado las siguientes diferencias significativas entre los grupos de datos estudiados. A continuación, se interpretan los resultados resaltando los contrastes entre las franjas temporales analizadas (2010-2016 y 2017-2023) y haciendo referencia a las tablas y figuras presentadas en el apartado de resultados.

Medias y medianas de las solicitudes

Como se observa en la tabla 2, la media de solicitudes en la franja 2010-2016 fue de 185.17, mientras que en la franja 2017-2023 esta aumentó a 254.13, lo cual refleja un incremento notable. Los intervalos de confianza al 95% (figura 5) muestran que no hay solapamiento entre ambas franjas, confirmando la diferencia significativa.

En cuanto a las medianas, la tabla 2 y los intervalos de confianza (figura 6) también reflejan una diferencia significativa, con valores de 141 para la primera franja y 186 para la segunda.

Al realizar un contraste de hipótesis paramétrico, como se detalla en la tabla 4, se obtuvo un p-valor de 0.00000000000000022, mucho menor al nivel de significación de 0.05, lo que confirma que la media poblacional de la segunda franja es significativamente mayor. Resultados similares se observan para las medianas con un p-valor igual.

Proporción de uniones homosexuales y heterosexuales

Respecto a las proporciones de uniones homosexuales, los datos muestran un incremento significativo entre franjas. La primera franja tiene una proporción de 1.24 con un IC al 95% de [1.16, 1.35], mientras que en la segunda franja la proporción aumenta a 44.83 con un IC de [1.25, 1.47]. Este cambio es evidente tanto en las muestras grandes (tabla 2) como en las pequeñas (tabla 3), pero a la hora de hacer el contraste de hipótesis en la muestra grande los datos muestran un p-valor obtenido de 0.01 confirmando que en la segunda franja han aumentado las uniones homosexuales, mientras que en las muestras pequeñas el p-valor obtenido es de: 0.38 indicando que no hay diferencia significativa entre las dos franajas, aunque con menor robustez.

En el caso de las uniones heterosexuales, no se detectaron diferencias significativas entre las franjas. El p-valor obtenido (0.4204 en muestras grandes y 1 en muestras pequeñas) es superior al nivel de significación, como

se indica en la tabla 4 y 5. Además, los intervalos de confianza (figura 7) muestran un alto grado de solapamiento.

Muestras grandes y pequeñas

Para las muestras pequeñas, las diferencias en las medias y medianas también son significativas (tabla 5). Los p-valores obtenidos para la media (0.02403) y la mediana (0.02825) confirman estas diferencias.

Sin embargo, en el caso de las proporciones de uniones homosexuales y heterosexuales, los resultados con muestras pequeñas no permiten rechazar la hipótesis nula de igualdad, como se observa en los p-valores mayores a 0.05 (tabla 5).

Conclusiones

Los resultados reflejan un incremento significativo en la media y la mediana de las solicitudes de uniones de hecho en la franja 2017-2023 en comparación con la franja 2010-2016. Por otro lado, se aprecia un cambio significativo en la proporción de uniones homosexuales entre las franjas temporales, mientras que no se detectan diferencias significativas en las uniones heterosexuales. Estas conclusiones son consistentes basandonos en los datos tanto para muestras grandes como pequeñas, aunque con menor solidez en el caso de las muestras pequeñas.