Práctica 2: Tipología y ciclo de vida de los datos

Componentes del grupo

- Francisco Jesús Montes Mantero

Preguntas

1. Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder?

He elegido el dataset de Kaggle:

Medical Appointment No-Shows
https://www.kaggle.com/joniarroba/noshowappointments

Contiene información básica sobre más de cien mil citas médicas que se produjeron en una población de Brasil. Los datos nos dicen si la persona terminó asistiendo a la cita o si por el contrario faltó. La utilidad principal de desarrollar análisis a partir de estos datos es que podrían servir para contestar a la pregunta de si un paciente asistirá o no a su cita con una probabilidad razonable. Esta información tiene un gran valor ya que permitiría identificar grupos de citas que por sus características son más propensas a acabar en una falta con el consiguiente coste para el sistema sanitario. Disponer de información para identificar este tipo de citas resultaría de gran ayuda ya que permitiría construir una línea de trabajo basada en los datos y dirigida a minimizar su ocurrencia.

En concreto y para esta práctica intentaré identificar cuáles de los datos disponibles para una cita médica podrían ayudar a contestar la pregunta y en la medida de lo posible en cada caso, en qué medida.

2. Limpieza de los datos.

2.1. Selección de los datos de interés a analizar. ¿Cuáles son los campos más relevantes para responder al problema?

La primera parte del código carga el dataset, selecciona y establece los tipos de datos adecuados en R para poder tratarlos adecuadamente más adelante.

Hay que destacar el hecho de que he seleccionado aquellas citas no relacionadas con niños. Por tanto he seleccionado sólo aquellas para pacientes adultos mayores de edad

(más de 18 años). Vamos a considerar que estas personas son más responsables por el hecho de acudir o no a una cita y las que más nos interesa estudiar de acuerdo con el problema que nos ocupa. Además, algunos aspectos como el hecho de recibir un SMS o no a modo de recordatorio o si recibe ayudas del Estado tendrá más sentido con esta selección de pacientes.

Los datos más relevantes que he considerado para intentar resolver el problema han sido los siguientes:

Gender: Sexo del paciente. Variable dicotómica (M/F). Sin embargo esta variable se transforma a True (hombre) y False (mujer) para facilitar su análisis posterior.
Age: Edad del paciente (discretizada en años y a partir de 18).
Scheduled Day: Fecha y hora en que se planificó la cita.
Appointment Day: Fecha en la que la cita médica tuvo lugar.
Scholarship: Si el paciente cuenta con ayudas del estado o no. Variable dicotómica (True/False).
Handcap: Número de condiciones de minusvalía que tiene el paciente. Para esta práctica he considerado interesante transformar este valor a una variable dicotómica para indicar si el paciente tiene simplemente una minusvalía o no. (True/False).
SMS_received: Si el paciente recibió un SMS recordatorio de su cita o no. (True/False).

2.2. ¿Los datos contienen ceros o elementos vacíos? ¿Y valores extremos? ¿Cómo gestionarías cada uno de estos casos?

El dataset no presenta elementos vacíos. Existen varios campos binarios como SMS_received o Handcap por lo que los ceros para representar el valor de “falso” es normal. Las fechas de la cita “appointment day” tienen todas a cero la parte de hora, minutos y segundos. Esto es normal y es solo una falta de precisión que no se ha capturado en el dataset.

Sin embargo sí existen valores extremos para la edad, teniendo máximos de 115. Voy a considerar como sospechosas aquellas citas de pacientes de más de 100 años. Al comprobar que solo existen 7 registros en esta situación y sin tener posibilidad de enmendarlos con el valor correcto, he decidido descartarlos como parte de la limpieza previa al análisis.

También he detectado 4 registros erróneos que contienen fechas de cita anteriores a su fecha de planificación. Debido a su reducido número, y por las mismas razones que en el caso de la edad y no poder recuperarlos, he decidido descartarlos.

3. Análisis de los datos.

3.1. Selección de los grupos de datos que se quieren analizar/comparar.

Los dos grupos básicos a comparar son el de las citas donde el paciente asistió y aquellas donde faltó. Considerando estos dos grupos compruebo si las proporciones de faltas y asistencias depende significativamente de las siguientes variables:

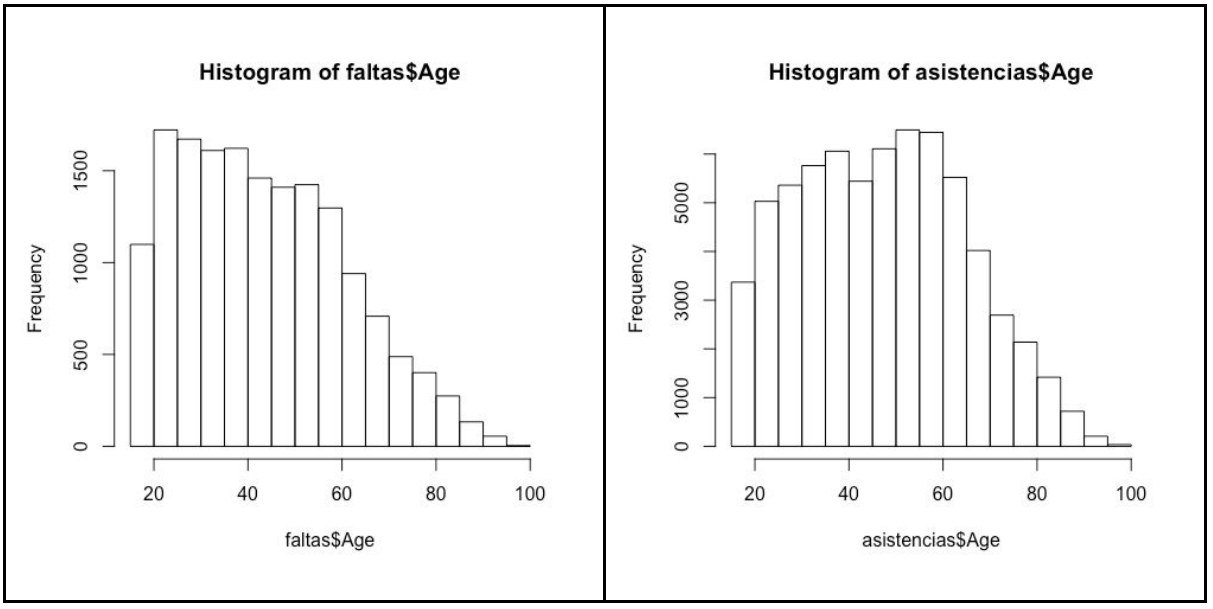
Age: ¿La edad del paciente puede marcar una pauta hacia una falta o una asistencia?
Gender: ¿Puede existir una relación entre el sexo del paciente y el hecho de que falte o no a su cita?
Scholarship: el hecho de el paciente reciba ciertas ayudas o subvenciones del Estado o no. ¿puede influir?
Handcap: Si un paciente tiene un grado de minusvalía, ¿puede jugar esto en su contra a la hora de asistir a su cita?
SMS_received: Recibir un SMS de recordatorio, ¿reduce realmente las posibilidades de una falta al intentar evitar un olvido?

3.2. Comprobación de la normalidad y homogeneidad de la varianza. Si es necesario (y posible), aplicar transformaciones que normalicen los datos.

En este apartado entraría el análisis de la variable edad (Age). Se trata de una variable numérica continua pero se encuentra discretizada y expresada en enteros (años).

Para analizar el efecto de la edad recurrimos a efectuar pruebas de normalidad y así saber si podemos utilizar pruebas paramétricas o no.

Histogramas de las edades de pacientes en citas que acabaron en falta y asistencia:



Las pruebas de normalidad dan un resultado muy pobre sobre la distribución de edades y rechazan la hipótesis de normalidad. Para todas se ha utilizado el grado de significación estándar de 0.05.

Se han utilizado para ello las siguientes pruebas en R:

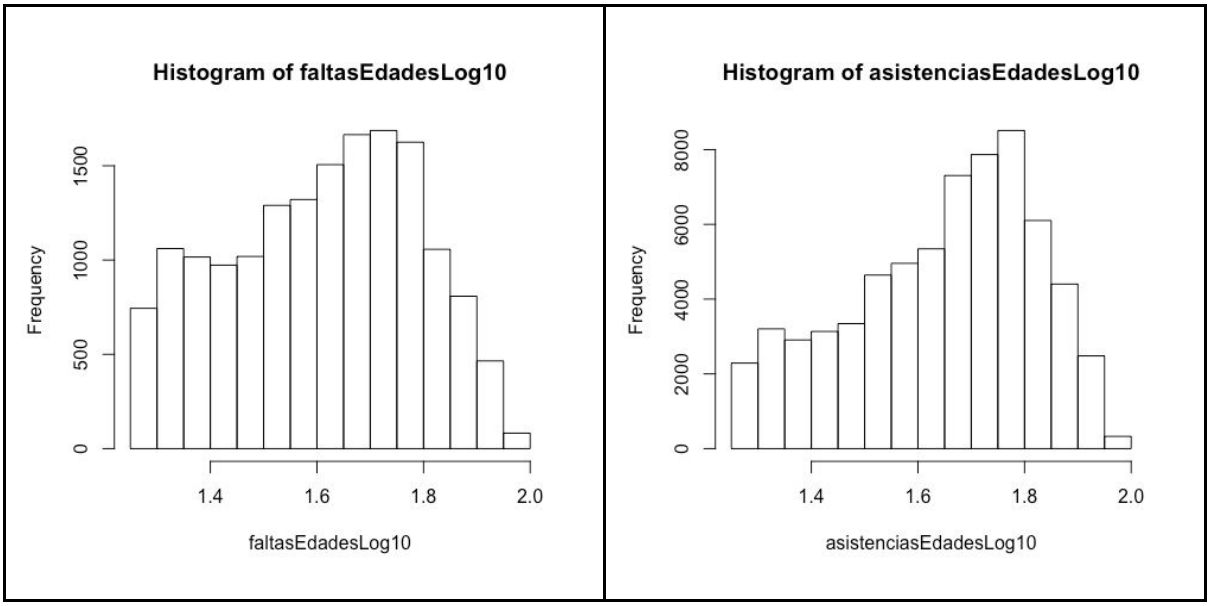
- Test de Anderson-Darling
- Test de Cramer-Von Mises
- Test de Lilliefors (Kolmogorov-Smirnov)
- Test de Pearson
- Test de Shapiro-Wilk: Este test no pudo usarse al estar limitado a 5000 valores de muestra.

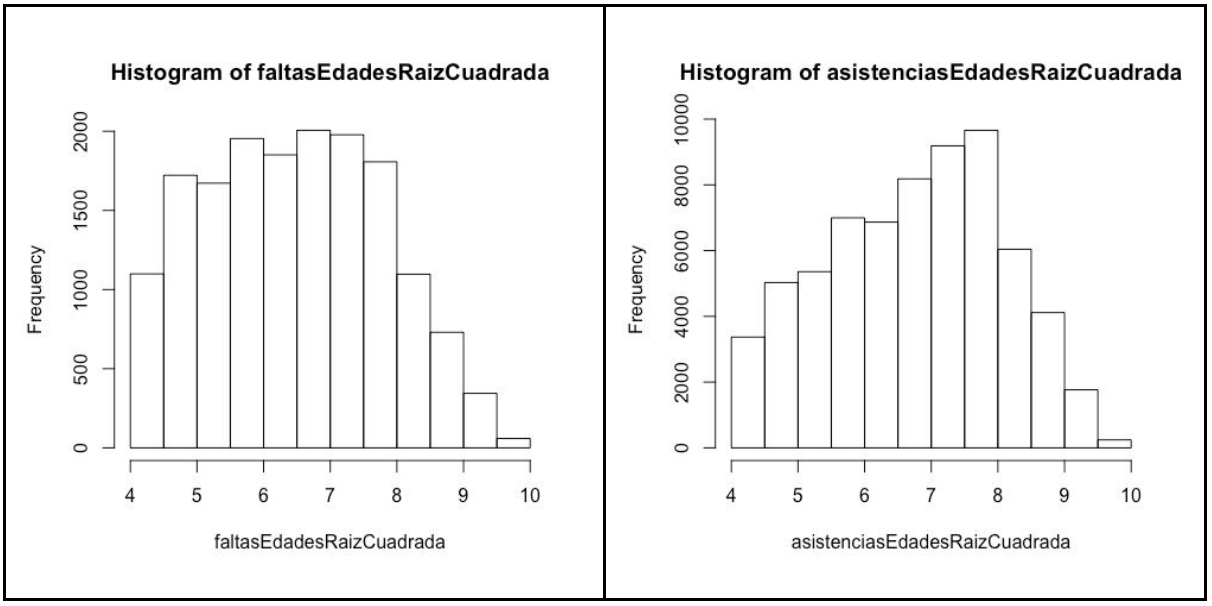
Todos rechazaron la hipótesis nula de estos tests que es precisamente lo que queríamos demostrar, que la curva se aproximaba a una distribución normal.

En un intento de obtener una distribución más normal y que pase las pruebas decidí utilizar las transformaciones más usuales aprovechando que no existen edades negativas:

- Logarítmica
- Raíz cuadrada

Histogramas de las transformadas logarítmica y de raíz cuadrada.





De nuevo, los resultados mostraron que la distribución no podía considerarse normal. Debido a ésto, nos quedan utilizar test no paramétricos como el de Wilcoxon de dos muestras.

El test dio como resultado que la diferencia de edades de paciente entre las citas que acaban en falta y asistencia es significativa estadísticamente.

3.3. Aplicación de pruebas estadísticas (tantas como sea posible) para comparar los grupos de datos.

Para la verificación de proporciones de las variables **Gender**, **Scholarship**, **Handcap** y **SMS_received** podemos recurrir a los tests siguientes:

- Test de dos proporciones

Planteamiento:

Aquí nos planteamos la hipótesis nula y su alternativa como sigue, es la misma para cada una de las variables analizadas:

H0 = La proporción de faltas para ambos valores posibles de la variable analizada (true/false) es la misma. Es decir, no hay incidencia de la variable analizada en las faltas.
H1 = La proporción de faltas para ambos valores posibles de la variable analizada (true/false) no es la misma. Hay incidencia de la variable analizada en las faltas.

Dado que tenemos una muestra grande podemos esperar al menos 5 ocurrencias de faltas en cada grupo con lo que podemos asumir una aproximación a la normal de la distribución de las diferencias entre ambas proporciones para este test.

Elegimos un nivel de significación estándar y utilizado en muchos dominios: 5% (0.05), 95% de confianza.

Resultado:

Tras comparar el grado de significación con el resultado de este test para todas las variables analizadas, tenemos que la hipótesis nula es rechazada con lo que las proporciones son distintas en cada caso y no se debe al azar de la muestra.

- Test de Chi-Cuadrado y test exacto de fisher

Planteamiento:

H0 = El hecho de asistir o faltar a una cita es independiente de la variable analizada.
H1 = El hecho de asistir o faltar a una cita está relacionado con la variable analizada.

Al que igual que para los tests anteriores el grado de significación elegido es del 5% (0.05)

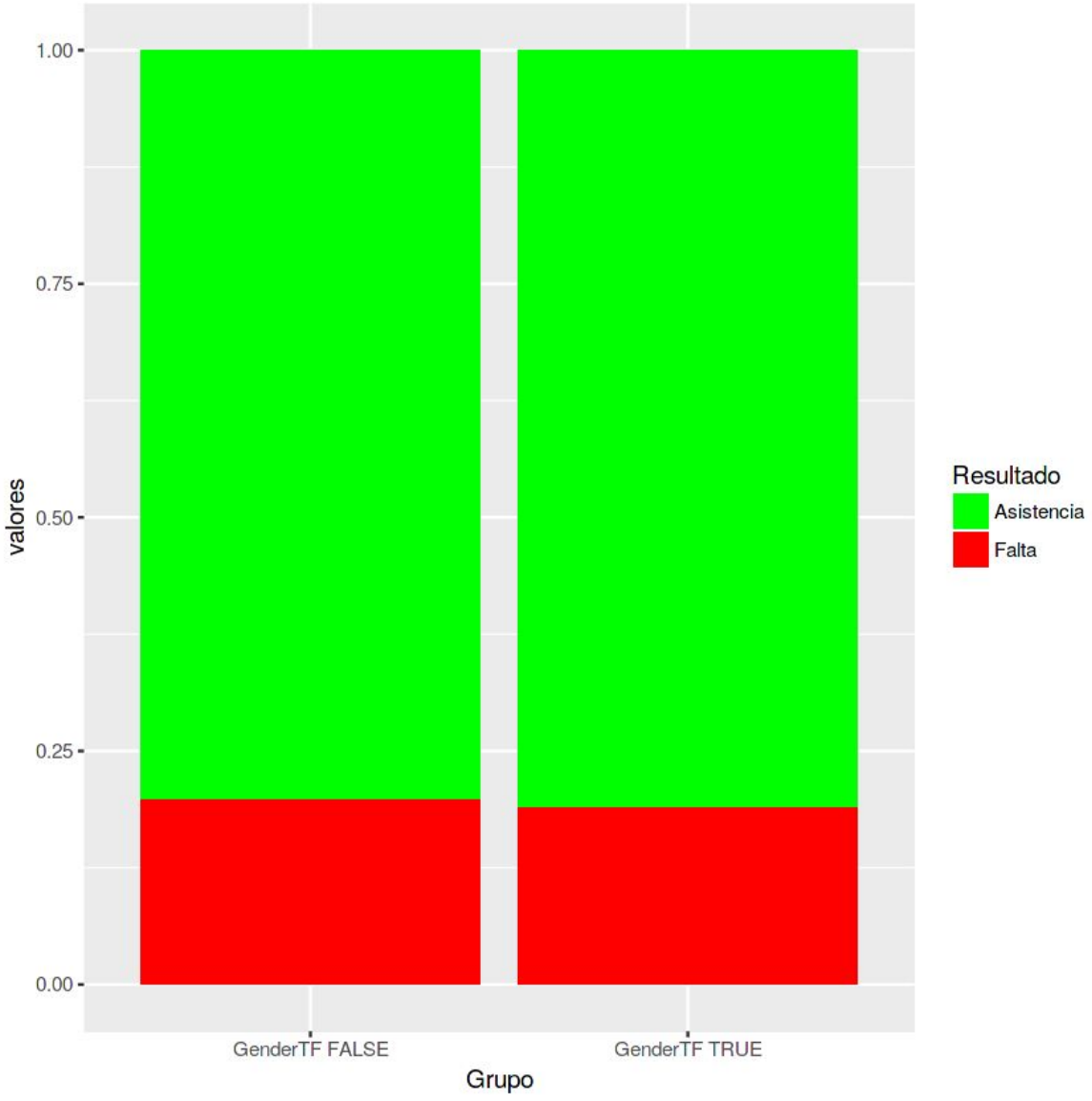
Resultados:

En todos los tests la hipótesis nula es rechazada con lo que apunta a una asociación entre cada una de las variables y el hecho de asistir o faltar a una cita.

1. Representación de los resultados a partir de tablas y gráficas.

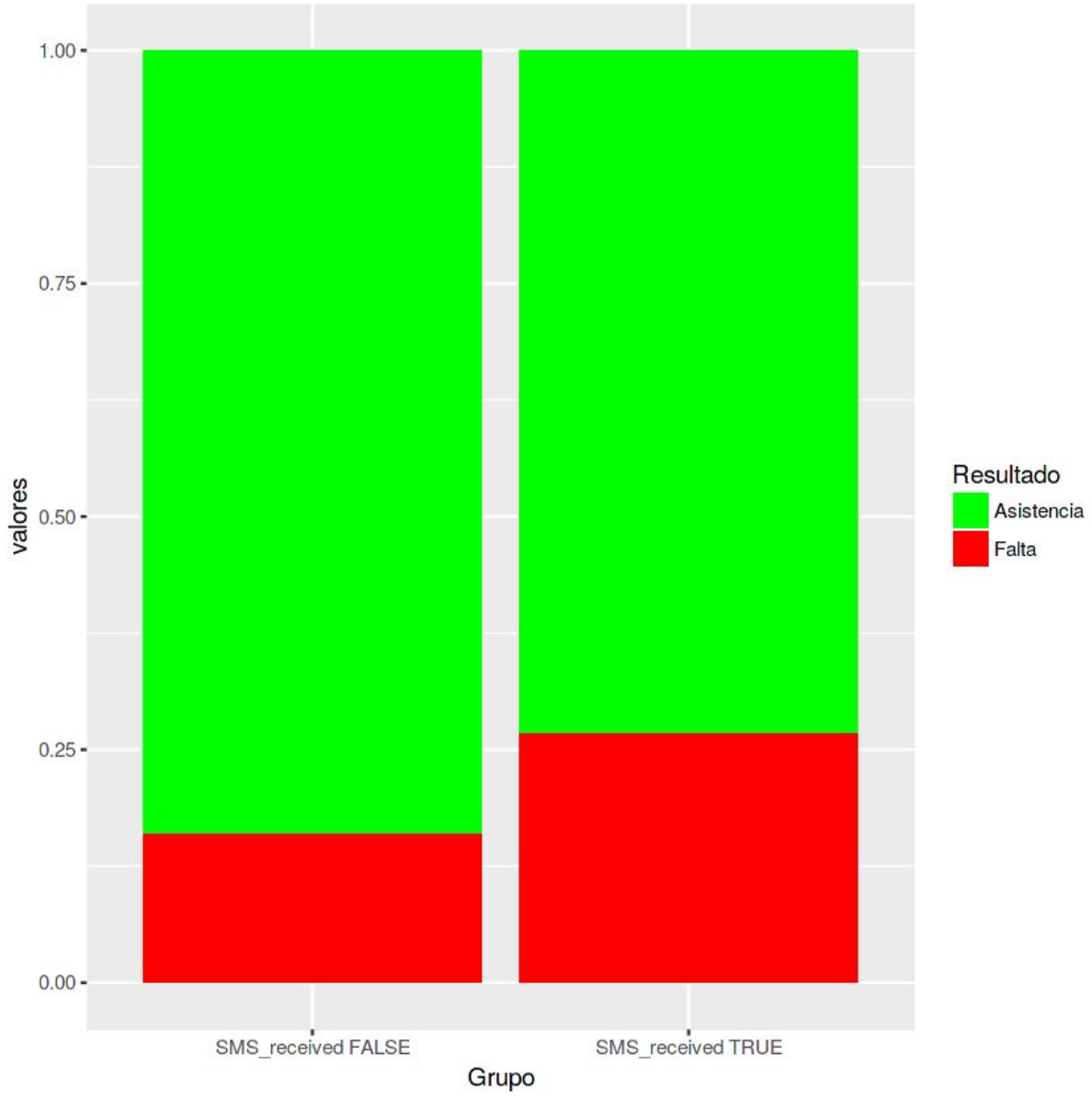
Proporción de faltas y asistencias para la variable **Gender**

Grupo GenderTF False: Aquellas citas médicas donde el paciente es mujer.
Grupo GenderTF True: Aquellas citas médicas donde el paciente es hombre.



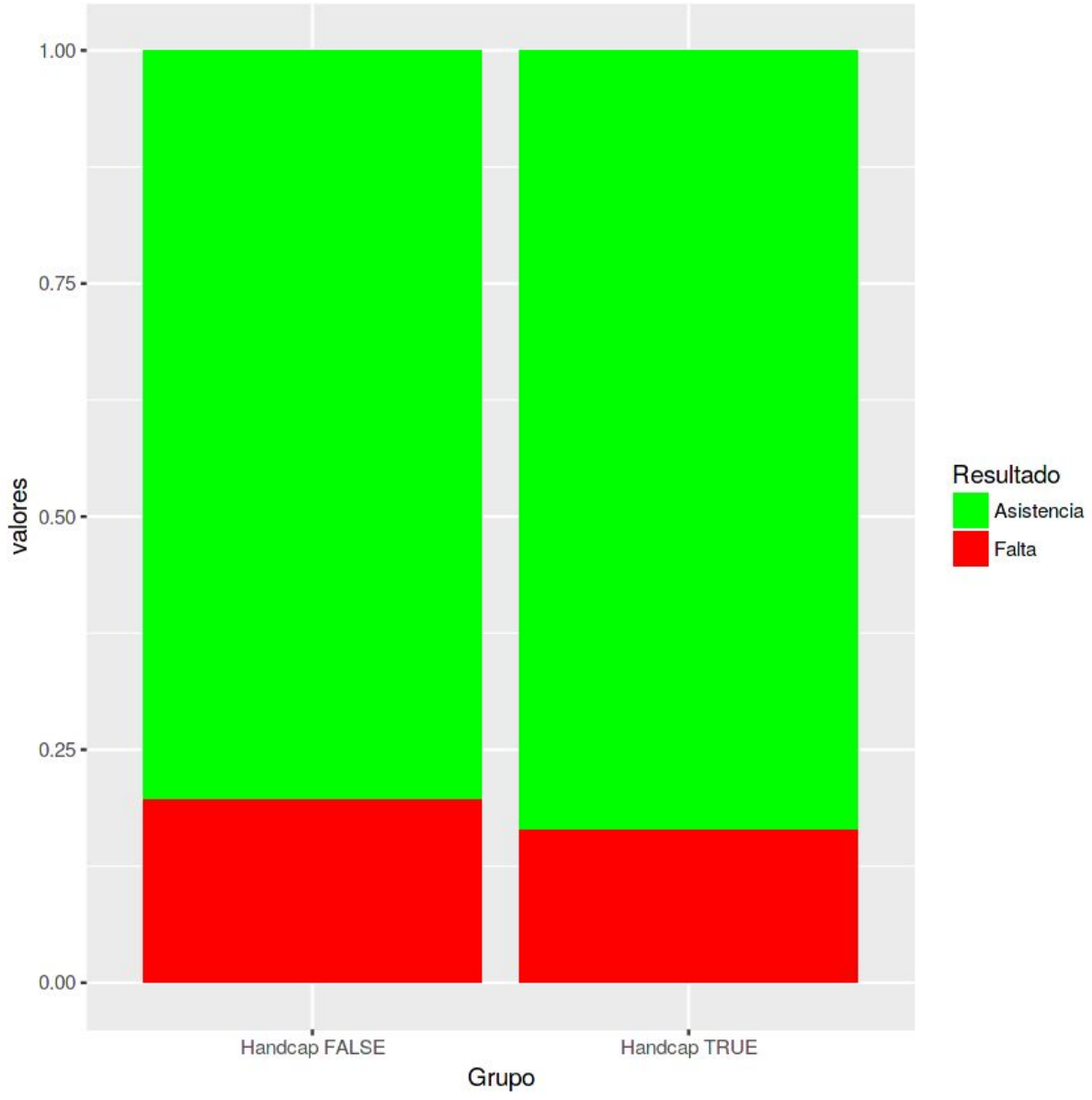
Proporción de faltas y asistencias para la variable **SMS_received**

Grupo SMS_received FALSE: Aquellas citas médicas donde el paciente no ha recibido ningún SMS de recordatorio.
Grupo SMS_received TRUE: Aquellas citas médicas donde el paciente ha recibido un SMS de recordatorio.



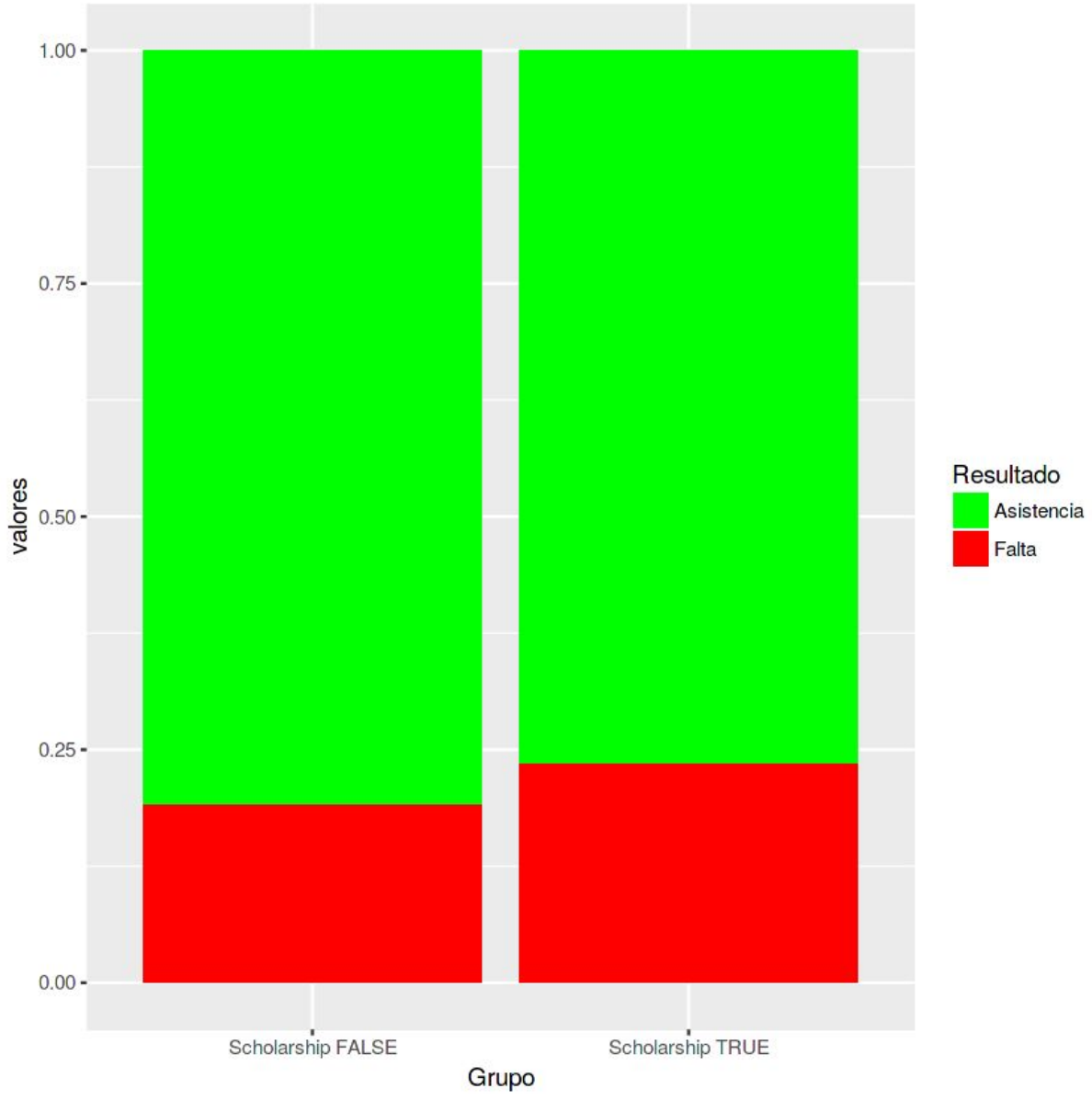
Proporción de faltas y asistencias para la variable **Handcap:**

Grupo HandCap FALSE: Aquellas citas médicas donde el paciente no presenta ninguna minusvalía.
Grupo HandCap TRUE: Aquellas citas médicas donde el paciente presenta alguna minusvalía.



Proporción de faltas y asistencias para la variable **Scholarship:**

Grupo Scholarship FALSE: Aquellas citas médicas donde el paciente no recibe ayudas del Estado.
Grupo Scholarship TRUE: Aquellas citas médicas donde el paciente recibe ayudas del estado.



2. Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?

Algunos resultados eran esperados y otros demuestran hechos contrarios a los esperados.

Para la columna **Age** de la muestra se obtiene que la media de edad del paciente de una cita que acaba en falta es de 43.92 años mientras que la media de edad de una cita que acaba en asistencia es de 47.61 años. La diferencia es significativa estadísticamente con lo que las faltas se produce en pacientes relativamente más jóvenes.

Tras una primera visualización de las variables **Gender**, **Scholarship**, **Handcap** y **SMS_received** pueden verse cierto grado de influencia en la posible falta a una cita médica.

Las tres pruebas efectuadas sobre las proporciones de las variables estudiadas rechazan las hipótesis nulas con lo que podemos estar seguros de que estadísticamente presenta un buen grado de significación y que además existe dependencia entre ellas y el hecho de que una persona asista o falte a una cita.

Sin embargo y de acuerdo con la gráficas:

- **Gender**: El sexo de un paciente no influye en gran medida de por sí en la asistencia o falta de una cita. Existe sin embargo una ligera tendencia de acabar en falta en caso de que el paciente sea mujer.
- **Scholarship**: Existe una mayor proporción de citas acabadas en falta cuando el paciente recibe ayudas del Estado.
- **Handcap**: Contrariamente a lo esperado, la proporción de citas acabadas en falta están del lado de pacientes que no tienen minusvalía.
- **SMS_received**: Sin duda aquí se muestra la diferencia en proporción más significativa siendo sorprendente el hecho de que un recordatorio por SMS no reduce la posibilidad de que una cita resulte en falta sino que parece que esté ligado con el efecto contrario. Por tanto los datos arrojan dudas sobre la eficacia de este sistema.