Trabajo Práctico I

Bioestadística

Validación de datos en R

Facultad de Ciencias Económicas y Estadística, Universidad Nacional de Rosario

Alfonsina Badin, Valentina Salvarezza, Camila Sebastiani

Abril de 2024

Introducción

En un ensayo clínico aleatorizado se comparan dos drogas para tratar la hemorragia posparto de mujeres embarazadas. El estudio consiste en reclutar mujeres que voluntariamente aceptan formar parte del ensayo, a estas se les asignará aleatoriamente uno de los dos tratamientos y se recolectó información sobre ellas.

El formulario de parto `DEL` registra las características basales de la mujer e información adicional sobre el nacimiento y administración del tratamiento experimental (ver en anexo). Este ingreso de información se hace en base a las respuestas de un formulario que, aunque no se desea, puede cometer errores.

El objetivo del presente informe es realizar un análisis exhaustivo de validación de datos, empleando técnicas que permiten evaluar la calidad de los datos ingresados y detectar errores sistemáticos en el llenado de los formularios.

Desarrollo de reglas de validación

A continuación se detallan las reglas globales que tienen que cumplirse en todas las preguntas y las reglas específicas de cada una con sus respectivas descripciones. Esto se hace con la finalidad de hacer un recorrido detallado del formulario y la validación a realizarse.

Al final de esta sección, se encontrará el listado de reglas empleadas en un código de R que será probado en la base provista para identificar:

- Participantes limpios.
- Participantes con más inconsistencias.
- Errores más frecuentes.

Reglas globales

Como regla principal y general para todas las respuestas del formulario, ninguna de ellas puede ser vacía, a no ser que alguna regla específica diga lo contrario.

Además, se entiende que en las respuestas categóricas con codificación numérica ningún participante puso ni letras ni números que no formaban parte del listado de opciones. Se asume, por ejemplo, que nadie colocó *"No"* en lugar de 0, y que nadie utilizó otro número que no sea 0, 1 o 9 ya que, al ser un campo cerrado sólo se admiten respuestas válidas.

Reglas específicas

Las reglas específicas se detallan para cada una de las preguntas del formulario que lo requieran, teniendo en cuenta cuáles son las condiciones que cada una requiere y qué implican sus respuestas.

1 - Edad materna

- No puede ser mayor o igual a 60 años, ya que se considera que pasados los 45 años hay muy poca probabilidad de embarazo. Una madre de más de 60 años sería extremadamente atípica y probablemente sea un error de entrada.

2 - Edad gestacional

- No puede ser menor a 20 semanas, en dicho caso se considera aborto.
- No puede ser mayor a 50 semanas, ya que la gestación ronda por lo general entre las 38 a 42 semanas. Un tiempo de gestación mayor o igual a 50 semanas sería prácticamente imposible y probablemente sea un error de entrada.

3 - Cantidad de embarazos previos

- Si es 0, la respuesta 4 debe ser vacía.
- Si es 0, la respuesta 5 debe ser vacía.

4 - ¿La mujer tuvo HPP en los embarazos anteriores?

- Sólo puede ser vacía si la respuesta 3 es 0.

5 - ¿La mujer tuvo parto por cesárea anteriormente?

- Sólo puede ser vacía si la respuesta 3 es 0.

6 - ¿Le indujeron el parto a la mujer cuando la ingresaron al hospital para este parto?

- Si es 0, las respuestas 6a1, 6a2, 6a3, 6a4 y 6as deben ser vacías.

6a1 - Oxitocina

- No puede ser vacía si la respuesta 6 es 1.

6a2 - Misoprostol

- No puede ser vacía si la respuesta 6 es 1.

6a3 - Sonda de Foley con globo

- No puede ser vacía si la respuesta 6 es 1.

6a4 - Otro

- No puede ser vacía si la respuesta 6 es 1.

6s - Si la respuesta es 'Otro', especifique.

- No puede ser vacío si 6a4 es 1.

```
reglas_parto <- tribble(
  ~name, ~description, ~rule,
  "R01", "{del01} no puede ser vacía", "is.na(del01)",
  "R02", "Si {del01} no es vacío, no puede ser mayor o igual a 60", "del01 >= 60",
  "R03", "{del02} no puede ser vacía", "is.na(del02)",
  "R04", "{del02} no puede ser menor a 20", "del02 <= 20",
  "R05", "{del02} no puede ser mayor a 50", "del02 >= 50",
  "R06", "{del03} no puede ser vacía", "is.na(del03)",
  "R07", "si {del03} es 0, {del04} debe ser vacía", "del03 ==  0 & !is.na(del04)",
  "R08", "si {del03} es 0, {del05} debe ser vacía", "del03 ==  0 & !is.na(del05)",
  "R09", "si {del03} no es 0, {del04} no puede ser vacía", "del03 !=  0 & is.na(del04)",
  "R10", "si {del03} no es 0, {del05} no puede ser vacía", "del03 !=  0 & is.na(del05)",
  "R11", "{del06} no puede ser vacía", "is.na(del06)",
  "R12", "Si {del06} = 0, todas las 6a son vacías", "del06 == 0 & !is.na(del06a1) &
      !is.na(del06a2) & !is.na(del06a3) & !is.na(del06a4) & !is.na(del06as)",
  "R13", "Si {del06a4} = 1, {del06as} no puede ser vacío", "del06a4 == 1 &
      is.na(del06as)",
  "R14", "Si {del06} = 1, {del06a1} no puede ser vacía", "del06 == 1 & is.na(del06a1)",
  "R15", "Si {del06} = 1, {del06a2} no puede ser vacía", "del06 == 1 & is.na(del06a2)",
  "R16", "Si {del06} = 1, {del06a3} no puede ser vacía", "del06 == 1 & is.na(del06a3)",
  "R17", "Si {del06} = 1, {del06a4} no puede ser vacía", "del06 == 1 & is.na(del06a4)",
  "R18", "{del07} no puede ser vacía", "is.na(del07)"
)
```

Resultados

Luego de realizar el proceso de validación de datos de acuerdo a las reglas establecidas, se obtuvieron los siguientes resultados que revelan la calidad de la información recopilada en la primera sección del formulario de parto de las 50 mujeres. Entre los aspectos de mayor relevancia se encuentran la identificación de los errores más frecuentes y la clasificación de los pacientes en función de la presencia o ausencia de inconsistencias en sus datos.

Errores más frecuentes

Con la finalidad de identificar cuáles son los errores más frecuentes, se construye la Figura 1 en donde se puede observar que ninguna de las reglas evaluadas supera el 25% de inconsistencias, es decir la mayoría de las respuestas parecen cumplir las reglas establecidas por el formulario. Es de destacar que aquella con más inconsistencias es la regla 07 y que la regla 13 es la que más observaciones no disponibles acumula.

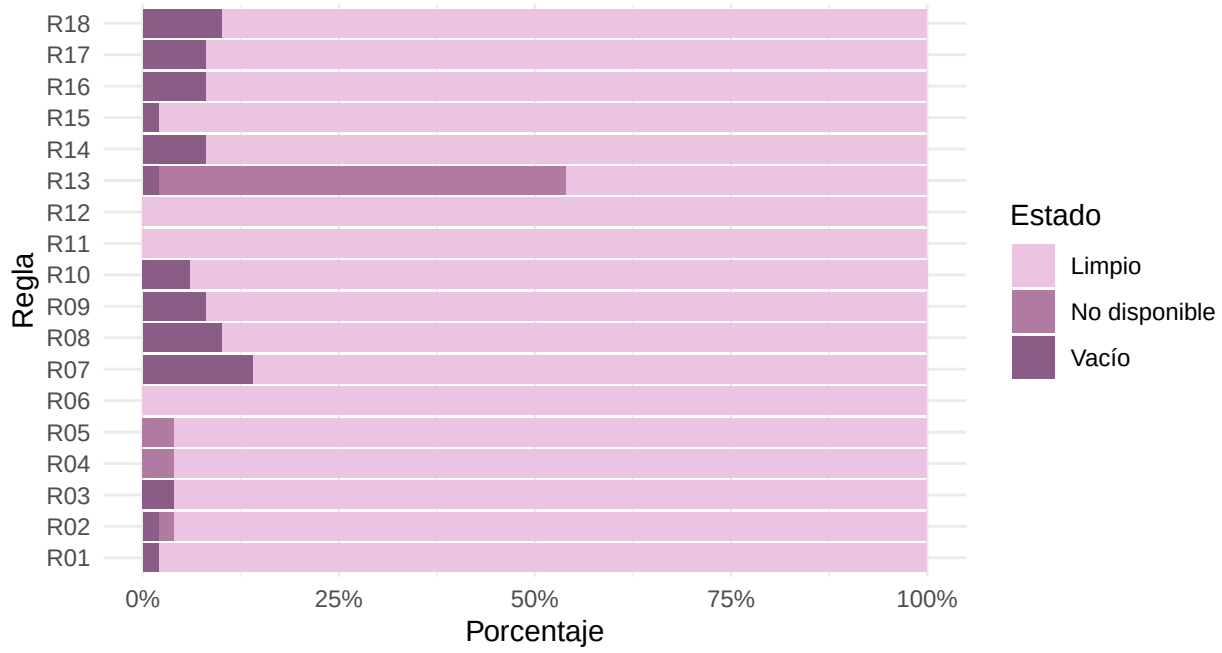

Figura 1: Porcentaje de cada estado en las reglas analizadas

Esta información es valiosa ya que permite identificar qué preguntas pueden no ser comprendidas correctamente, es decir falla en la redacción, para corregirlo en futuros formularios.

Pacientes con y sin inconsistencias

Es de gran interés estudiar la distribución de las inconsistencias encontradas en cada paciente, para obtener una visión clara de la magnitud y la variabilidad de estas irregularidades en los registros de los formularios.

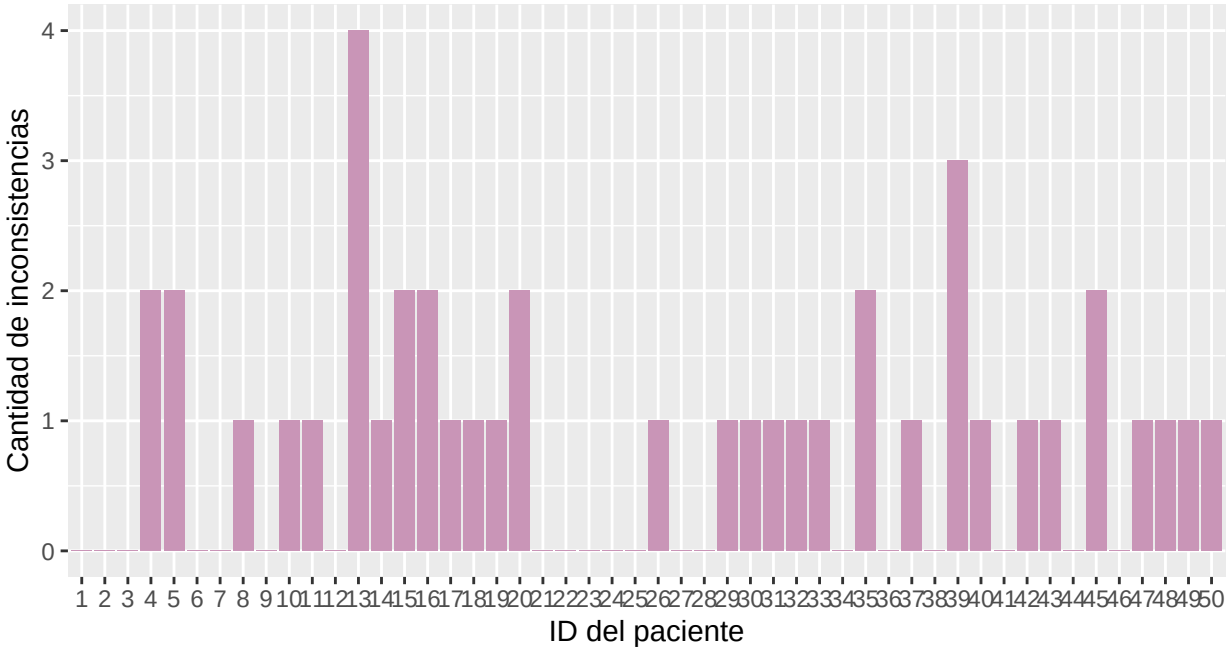

Figura 2: Cantidad de inconsistencias por paciente

En la Figura 2 se destaca el paciente con el ID número 13, quien presenta la mayor cantidad de inconsistencias detectadas, con un total de 4. Le sigue el paciente con ID 39, que muestra 3 inconsistencias. Por otro lado, es relevante mencionar que, varios pacientes han sido notablemente consistentes en sus registros, entr ellos: ID 1, 2, 3, 6, 7, entre otros.

Este hallazgo determina la integridad de los datos de estos pacientes en el contexto de las reglas establecidas.

Conclusión

La calidad de la información obtenida a través del formulario de parto se ve reflejada en las reglas de validación de datos establecidas para cada pregunta.

Identificar los errores más comunes ofrece una visión clara de los puntos críticos en la entrada de datos, lo que permite minimizar futuras inconsistencias. Asimismo, al clasificar a los pacientes según la presencia o ausencia de tales discrepancias, se obtiene una comprensión detallada de la calidad general de los registros.

Es importante señalar que la mayoría de las reglas muestran pocos casos de inconsistencia, lo cual indica un alto nivel de conformidad con los criterios de ingreso de datos del formulario. Sin embargo, existen áreas específicas donde se concentran las discrepancias, lo que subraya la necesidad de prestar especial atención a esas secciones para mejorar la precisión y consistencia de los datos recopilados.

Además, se destaca que el 60% de los pacientes presentan al menos una inconsistencia, lo que resalta la importancia de implementar medidas de validación para identificar y corregir estas discrepancias.

Es relevante mencionar que en algunos pacientes no se evidencian errores en sus registros. Este hallazgo subraya la efectividad de las reglas de validación para garantizar la coherencia en la información, así como la confiabilidad de los datos de estos pacientes para su análisis y uso futuro.

En conclusión, la aplicación de reglas de validación de datos no solo mejora la calidad y confiabilidad de la base de datos, sino que también proporciona información para la optimización continua de los procesos de captura y gestión de datos.