Propuesta metodológica para la validación de bases de datos aplicada a las ciencias de la salud

Methodological proposal for validation of databases applied to health sciences

Tamara Otzen T.[1,2,3,4], *Carlos Manterola*[1,2,4], *Nayely García-Méndez N.*[1,2,4], *Kristina Varela*[3] *y Guissela Quiroz*[2,5]

[1] Centro de Estudios Morfológicos y Quirúrgicos (CEMyQ), Universidad de La Frontera, Chile.
[2] Programa de Doctorado en Ciencias Médicas, Universidad de La Frontera, Chile.
[3] Fundación OPA (Observo, Protejo y Aprendo), Temuco, Chile.
[4] Núcleo Milenio de Sociomedicina.
[5] Universidad Mayor, Temuco, Chile.

Financiamiento: Parcialmente financiado por el proyecto DI19-0030, Universidad de La Frontera.
Conflictos de interés: Ninguno.

Recibido: 18 de agosto de 2021 (tercera versión: 27 de julio de 2022) / Aceptado: 18 de agosto de 2022

Resumen

La estadística es uno de los pilares de la ciencia, especialmente para describir y analizar los datos, la que ha progresado exponencialmente en los últimos años. Se le debe entender como un apoyo fundamental para la toma de decisiones en las distintas disciplinas. Los análisis exploratorios son descritos como el primer paso en la estadística, buscando organizar, representar y describir los datos, pero muchas veces este proceso se vuelve complejo, siendo importante realizar una revisión exhaustiva de la matriz de datos. Es objetivo de este manuscrito describir una propuesta metodológica para la validación de bases de datos aplicada a las ciencias de la salud. Esta metodología consta de seis etapas: cuatro de ellas obligatorias y sucesivas, y las otras dos opcionales. En la literatura médica, estos procedimientos generalmente son pasados por alto; buscamos, en consecuencia, recalcar la importancia de este proceso como previo a los análisis exploratorios.

Palabras clave: base de datos; exactitud de los datos; precisión de los datos; estadística; bioestadística; métodos; metodología; ciencias de la salud; validación.

Abstract

Statistics is one of the pillars of science, especially to describe and analyze data, it has progressed exponentially in recent years. Being the fundamental support for decision-making in different disciplines. Exploratory analyzes are described as the first step in statistics, seeking to organize, represent and describe the data, but many times this process becomes complex, and it is important to carry out an exhaustive review of the data matrix. The aim of this manuscript was to describe a methodological proposal for databases validation applied to health sciences.

This methodology consists of 6 stages, 4 of them compulsory and successive, and the other two, optional. In the literature these procedures are generally overlooked, our purpose is thus to emphasize the importance of this process prior to exploratory analyzes.

Keywords: database "[Mesh]; data accuracy "[Mesh]; statistics "[Mesh]; biostatistics"[Mesh]; methodology "[Mesh]; methods"[Mesh]; health sciences"[Mesh]; "data accuracy"[Mesh].

Introducción

En la actualidad, la estadística es uno de los pilares de la ciencia y su aplicación para describir y analizar los datos; ha progresado exponencialmente en los últimos años, convirtiéndose en el apoyo fundamental para la toma de decisiones en distintas disciplinas, entre ellas el campo de la salud[1,2].

Estadística se define históricamente como una disciplina matemática que debe cumplir ciertas condiciones para poder aplicarse a otras disciplinas[3]. Sus leyes y fórmulas se basan en poblaciones, bien definidas, casi siempre infinitas, casi sin errores de medición[1,2,4]. Sin embargo, la extrapolación de la estadística a otras ciencias requiere de su aplicación a poblaciones finitas, con algunas variables cualitativas, y errores de medida, desarrollándose por consiguiente la estadística aplicada[1,2].

Correspondencia a:
Tamara Otzen T.
tamara.otzen@ufrontera.cl

ORCID ID:
Tamara Otzen T.
https://orcid.org/0000-0001-6014-124

Carlos Manterola
https://orcid.org/0000-0001-921

La bioestadística es la rama de la estadística aplicada que estudia la utilización de métodos estadísticos en problemas médicos y biológicos; su uso se divide en descripción y análisis[5,6]. La estadística descriptiva pretende sintetizar y resumir la información contenida en datos[6]. En los últimos años, se conjuga la estadística descriptiva con los análisis exploratorios de datos, pudiendo de esta forma realizar un análisis previo de ellos, considerando las características de las variables objeto de estudio[7,8]. El principal objetivo del abordaje exploratorio es maximizar lo que podemos aprender de nuestros datos y lo que nos pueden sugerir[9], organizando, representando y describiendo los datos, pudiendo de esta forma extraer la información que contienen[10]. Tanto como hemos podido investigar en la literatura técnica, no evidenciamos que haya consenso de si el análisis exploratorio de los datos es parte de la estadística descriptiva[10], o es un paso previo[6,9], pero si hay acuerdo sobre la importancia de estos análisis para la estadística aplicada[8,10].

Los análisis exploratorios de datos permiten detectar anomalías o errores en la distribución univariante de los datos[11,12], como parte de un proceso reflexivo que permite corregir a tiempo estos errores para garantizar que los análisis posteriores sean adecuados[13], y pudiendo, por consiguiente, extrapolar de manera correcta los resultados de los análisis a otras poblaciones[14,15].

Por otra parte, la tecnología aplicada a la investigación ha variado sustancialmente en los últimos treinta años, existiendo una diversidad de herramientas informáticas que simplifican la confección de base de datos y los análisis estadísticos, llevando a cometer errores asociados a la incorrecta aplicación de técnicas estadísticas y la interpretación inadecuada de los resultados basándose en la significación estadística en lugar de significación clínica[1].

Además, la experiencia nos muestra que los problemas con los análisis estadísticos no están solamente en la mala utilización de la estadística en sí misma, si no que en muchas oportunidades las bases de datos vienen con errores que no son percibidos por el investigador, pero que perjudican la inferencia estadística[16]. En relación a esto, varios manuales para análisis estadísticos con distintos *softwares* proponen la necesidad de preparar los datos, previo a cualquier análisis, pero no señalan ni proponen una metodología clara que lo permita, dejándolo a criterio de cada investigador[17,18].

Por otra parte, existe evidencia sobre la importancia del análisis de datos perdidos y de valores fuera de rango, así como, de técnicas específicas para identificarlos, entregando información respecto de las propiedades de las estimaciones obtenidas[19-22], y proponiendo estrategias para reemplazar estos valores[20-22].

En relación a lo anterior, en el mundo de la ingeniería, una gran cantidad de investigadores se han focalizado en lidiar con los problemas de integración y traducción de datos, pero solo una pequeña porción de ellos está orientada a resolver los problemas de la calidad de los datos[23,24]. Por ejemplo, se han hecho esfuerzos por detectar posibles anomalías secundarias al proceso de mecanografía[25,26]. Se estima que entre 3 y 5% de todos los datos se introducen erróneamente, pudiendo alcanzar a 27% en el caso de datos de doble entrada o datos duplicados[26,27], es decir, datos introducidos por dos investigadores independientes; por lo tanto, es importante que después de diseñar la base de datos, deban ser previamente depuradas, para evidenciar posibles anomalías secundarias al proceso de mecanografía[26,27]. Por otra parte, con el fin de resolver problemas que surgen de la fusión de datos provenientes de fuentes heterogéneas, se han propuesto esquemas para la integración de ellos, finalizando con la tarea de eliminación de duplicados[26]. Al parecer, estas aproximaciones no son incorporadas en las investigaciones que se realizan en las ciencias de la salud (CS).

Es fundamental la rigurosidad estadística, aplicada especialmente en las CS, ya que existe la creencia extendida de que no es necesario tener grandes conocimientos estadísticos para obtener buenos resultados[28,29]. La calidad de los análisis estadísticos requiere de un proceso reflexivo y directrices que sean explícitas en las publicaciones[29].

El objetivo de este manuscrito fue describir una propuesta metodológica para la validación de bases de datos aplicada a las CS.

Metodología

La metodología propuesta tiene por objetivo realizar la optimización del conjunto de datos obtenidos para su posterior análisis estadístico. Cabe recalcar que no consideramos esta metodología como parte de los análisis exploratorios de datos, ni de la estadística descriptiva, sino como un paso previo, en el que la sugerencia de utilización de técnicas estadísticas (tablas o gráficos), solamente tiene como finalidad, la de facilitar el proceso de validación de datos, no entregando información estadística sobre las variables involucradas, por lo que no se considera para esto el tipo de variable. Por otra parte, en la literatura técnica se puede encontrar el concepto de depuración de base de datos[23,24], centrado principalmente en la limpieza de los datos para su utilización con algún *software*, incorporándola como una parte de la validación de base de datos.

Además, nos parece fundamental clarificar los principales conceptos de este artículo, definiendo *análisis exploratorios* como "un conjunto de estrategias para evaluar los datos, sin ningún sesgo, prejuicio o parcialidad *a priori* sobre ellos, permitiendo escoger el método estadístico apropiado y obtener más información de los propios datos"[30]. Por otra parte, entendemos el concepto de *estadística descriptiva* como "el conjunto de técnicas

utilizadas para describir y resumir el conjunto de datos obtenidos por el investigador"[30]. Y por último definimos la *validación de base de datos* como "el conjunto de técnicas utilizadas para optimizar el conjunto de datos para los posteriores análisis estadísticos". Así también consideramos que este proceso debiera ser sucesivo, empezando con la validación de base de datos, siguiendo con los análisis exploratorios, para continuar con la estadística descriptiva.

Por otra parte, cabe señalar que el proceso de validación de base de datos no puede ser desarrollado por un *software* específico, ya que se requiere del involucramiento activo por parte de los autores del estudio, donde se realice un proceso razonado, teniendo en cuenta las hipótesis que se pretenden comprobar, pudiendo utilizar un *software* para facilitar y/o apoyar este proceso.

El lector debe prestar especial atención sobre el objetivo de cada etapa, y considerar las sugerencias para cumplir cada objetivo, teniendo la libertad de efectuar los procedimientos que más le acomoden, asociados a sus propias habilidades en el manejo de los datos. Donde se puede utilizar el *software* estadístico con el que el ejecutor se sienta con mayores habilidades en su manejo.

La propuesta metodológica para validación de bases de datos aplicada a las CS está comprendida de seis etapas. Cuatro de ellas obligatorias y sucesivas, y dos opcionales (Figura 1). A continuación, se describirán en detalle las etapas, acompañadas de ejemplos basados en datos ficticios, utilizando el *software* Microsoft® Excel para Mac, versión 16.19.

Etapas de la propuesta (Figura 1)

Etapa I

Análisis comprensivo de los datos

Descripción

Es la primera etapa por desarrollar, la que se fundamenta en la observación racional de la base de datos para comprender las temáticas que se pretenden abordar y tener información sobre las posibles características de los sujetos en estudio.

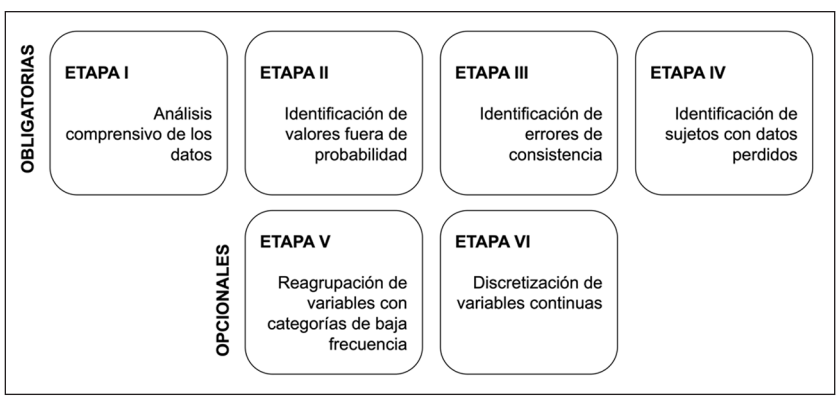

Figura 1. Esquema de la propuesta para la validación de bases de datos.

Objetivo

Identificar las temáticas asociadas a la base de datos, considerando los objetivos, las hipótesis y sujetos en estudio.

Pasos sugeridos

1. Conocer la población en estudio, objetivos e hipótesis del estudio base.
2. Realizar un análisis visual de los datos, identificando la coherencia de lo reportado en el estudio original y lo que está presente en la base de datos.
3. Revisar el libro de códigos[31], para obtener la descripción *in extenso* de las variables en estudio, identificando la coherencia entre los datos de la base de datos y el libro de códigos.

Ejemplo

La población del estudio está compuesta por un grupo de sujetos de un centro de salud público del sur de Chile. El objetivo del estudio fue identificar las características sociodemográficas asociadas a la realización del examen de próstata. La hipótesis fue: existen variables sociodemográficas asociadas a la realización del examen de próstata. En la base de datos de ejemplo (Tabla 1), al realizar un análisis visual se observa que corresponde a una población de 20 sujetos donde se consultó por: edad, la que podemos ver que varía en rangos de 2 a 72 años; sexo, siendo las posibles respuestas hombre y mujer; se realizó o no el examen de próstata, siendo las posibles respuestas si y no; religión, donde los sujetos se dividen en cuatro religiones: ateo, católico, bautista, luterana; peso, está considerado en kilos desde 44,4 a 110,4.

Conclusión del ejemplo

Los datos son coherentes con la población, objetivo e hipótesis del estudio. Identificando además la coherencia con el libro de códigos (Tabla 2).

Situaciones frecuentes a considerar

1. Es importante considerar que las variables deben ser precisas en recopilar la información de los participantes, de forma segmentada, para su posterior análisis. Siendo frecuente que las personas que no están vinculadas a los análisis estadísticos intenten lo contrario, es decir, incorporar la mayor cantidad de información en una sola celda. Ejemplo: Se crea una variable para registrar el grado de miopía ocular, incorporando el valor D o I asociado a la lateralidad,

acompañado del valor específico de miopía en números (ejemplo: D -1,5). Lo correcto en esta situación sería crear una variable que indique el grado de miopía del ojo derecho y otra variable que indique el grado de miopía del ojo izquierdo.
2. En la práctica, se suelen encontrar bases de datos con celdas resaltadas con colores para destacar cierta información. Ejemplo: Al analizar la base de datos se encuentran algunas casillas "pintadas" en la variable talla en cm de niños, indicando aquellos sujetos que están bajo el percentil 10. En este caso se debería agregar otra variable indicando si los sujetos se encuentran bajo el 10% de talla de los estándares poblacionales, con valores de respuesta si/no.

Etapa II

Identificación de valores fuera de probabilidad

Descripción

La identificación de los valores fuera de probabilidad consiste en identificar cuáles deberían ser los valores de las variables.

Objetivo

Identificar en cada una de las variables los valores que se encuentran fuera de las opciones esperadas.

Pasos sugeridos

1. Determinar cuáles son los valores posibles en cada una de las variables.
2. Buscar una estrategia que facilite la identificación de valores que se encuentran fuera de las opciones esperadas.
3. Revisar el motivo de los errores identificados, por ejemplo, si fueron producto de errores de digitación. En caso de que no sea un error de digitación, y no se pueda acceder a la información original se deberán eliminar o reemplazar según se acuerde con el equipo de investigadores.

Estadística sugerida

Tablas de frecuencias con variables cualitativas y medidas de resumen (valor máximo y valor mínimo) con variables cuantitativas**.**

Ejemplo 1

En la base de datos de ejemplo (Tabla 3) se observan sujetos menores de 18 años, lo que se encuentra fuera de los rangos de valores esperados para la variable edad, encontrándose 2 sujetos; el sujeto número 4 de 4 años, y el sujeto número 12, de 2 años, debiendo ser eliminados sus valores de edad.

Tabla 1. Base de datos de ejemplo

Sujeto	Edad	Sexo	Prost Exam	Religión	Peso
1	26	11	Sí	Ateo	60,5
2	57	Mujer	No	Ateo	80,3
3	41	Hombre	Sí	Católico	78,4
4	4	Hombre	No	Bautista	58,1
5	25	Hombre	Sí	Ateo	98,5
6	39	Hombre	Sí	Católico	57,3
7	47	Hombre	No	Ateo	70,4
8	54	Hombre	No	Católico	69,2
9	38	Hombre	No	Luterano	104,3
10	30	Hombre	No	Católico	55,1
11	19	Hombre	Sí	Bautista	85,1
12	2	Mujer	Sí	Católico	73,5
13	25	Hombre	Sí	Católico	86,7
14	34	Hombre	Sí	Bautista	99,3
15	37	Mujer	No	Bautista	62,1
16	48	Mujer	No	Católico	61,2
17	72	Mujer	No	Luterano	110,4
18	65	Mujer	No	Católico	44,4
19	39	Mujer	No	Ateo	64,5
20	45	22	No	Ateo	79,9

Tabla 2. Libro de códigos

Variable	Descripción	Categorías o valores
Edad	Edad del sujeto en años	Números enteros mayores a 17
Sexo	Identidad sexual	Mujer, Hombre
Religión	Creencia religiosa	Ateo, Católico, Bautista, Luterano
Posteexam	¿Se ha realizado el examen de próstata?	Sí, No
Peso	Peso del sujeto en kilos	Números con un decimal separado por coma.

Ejemplo 2

En la base de datos de ejemplo (Tabla 3) se observa los valores esperados para la variable sexo, encontrando el sujeto 1 con valor 11 y el sujeto 2 con valor 22, considerando que no se relaciona a ningún valor esperado es que se deciden eliminar los valores específicos.

Ejemplo 3

En la base de datos de ejemplo (Tabla 3) se observa los valores esperados para la variable examen de próstata,

Tabla 3. Base de datos con valores fuera de probabilidad

Sujeto	Edad	Sexo	Prost Exam	Religión	Peso
1	26	**11**	Sí	Ateo	60,5
2	57	Mujer	No	Ateo	80,3
3	41	Hombre	Sí	Católico	78,4
4	**4**	Hombre	No	Bautista	58,1
5	25	Hombre	Sí	Ateo	98,5
6	30	Hombre	Sí	Católico	**57.3**
7	47	Hombre	No	Ateo	70,4
8	54	Hombre	No	Católico	69,2
9	38	Hombre	**No**	Luterano	104,3
10	30	Hombre	No	Católico	55,1
11	19	Hombre	Sí	Bautista	**85.1**
12	**2**	Mujer	Sí	Católico	73,5
13	25	Hombre	**1**	Católico	**86.7**
14	34	Hombre	**Sí**	Bautista	99,3
15	37	Mujer	**0**	Bautista	62,1
16	48	Mujer	No	Católico	61,2
17	72	Mujer	No	Luterano	110,4
18	65	Mujer	No	Católico	44,4
19	30	Mujer	No	Ateo	64,5
20	45	**22**	No	Ateo	79,9

encontrando los sujetos 9, 13, 14 y 15 con valores fuera de lo esperado (no; 1; sí; 0); en el caso de estos valores se decide hacer el reemplazo ya que se aproximan a los valores originales esperados (No; Sí; Sí; No).

Ejemplo 4

En la base de datos de ejemplo (Tabla 3) se observa los valores esperados para la variable peso, encontrando los sujetos 6, 11 y 13 con valores fuera de lo esperado (57,3; 85,1; 86,7), en el caso de estos valores se decide hacer el reemplazo del punto por la coma.

Situaciones frecuentes que considerar

1. Cuando se consideran variables asociadas a la temporalidad, por ejemplo, fecha u hora, es fundamental homogenizar el formato de estas, dejándolo claramente establecido en el libro de códigos. Ejemplo: en el caso de la variable fecha de nacimiento 15 de junio de 2018, en algunos sujetos se escribe 15-06-2018 (formato DD-MM-AAAA) y en otros 2018/06/15 (formato AAAA/MM/DD).
2. A veces se encuentran valores posibles, pero poco frecuentes. Por lo que antes de considerar cada valor como posible o no, es importante contrastar la observación cuestionada con el resto de la información proveniente del mismo sujeto. Ejemplo: en la variable peso del recién nacido (RN), donde se consideraron solo RN de 38 o más semanas de gestación intrauterina, el valor para un RN fue de 2.000 g. Si el RN de 2.000 g nació de una madre con problemas durante el embarazo, indicaría que hay una alta chance de que sea un valor válido.

Etapa III

Identificación de errores de consistencia

Descripción

Esta etapa se emplea para registrar y analizar la relación entre dos o más variables cualitativas. La búsqueda de errores de consistencia consiste en el supuesto de que las respuestas de las variables son dependientes unas de las otras. Por consiguiente, al revisar ambas variables de forma cruzada debería haber casilleros internos sin frecuencia, identificando de esta forma los errores de consistencia.

Objetivo

Identificar errores de consistencia asociados a las variables que se someterán a comprobación de hipótesis.

Pasos sugeridos

1. Identificar cuáles son las variables que tienen respuestas asociadas entre si.
2. Buscar una estrategia que facilite el cruce de las variables asociadas para la identificación de errores de consistencia.
3. Revisar si los errores identificados fueron producto de errores de digitación o no. En caso de que no sea un error de digitación se deberán eliminar los valores de ambas variables.

Estadística sugerida

Tablas de contingencia entre variables cualitativas. Reagrupar variables cuantitativas e incorporarlas en tablas de contingencia.

Ejemplo

En la base de datos de ejemplo se muestra el análisis de consistencia de datos mediante una tabla de contingencia (Tabla 4), que relaciona la variable examen de próstata con sexo[32], donde se demuestra que del total de los sujetos que se realizaron examen de próstata, 6 son hombres y una mujer. De acuerdo con lo anterior es necesario eliminar los

valores del sujeto número 12, asociados al sexo y que se realizó el examen de próstata, ya que existe un problema de coherencia entre ambos (Tabla 5).

Etapa IV

Identificación de sujetos con datos perdidos

Descripción

Por lo general, los investigadores abordan los datos faltantes al incluir en el análisis solo los casos con la totalidad de las respuestas válidamente emitidas, lo que dependiendo de los análisis estadísticos que se pretenden realizar posteriormente puede ser válido[33]. Sin embargo, si la estadística no lo requiere, pudiera ser que los resultados de los análisis de comprobación de hipótesis pudieran estar sesgados, por lo que se vuelve fundamental poder llegar a un acuerdo con el grupo de investigadores para definir qué se realizará con los sujetos que no tengan todos sus datos válidamente emitidos.

Objetivo

Identificar la influencia estadística en las hipótesis de los sujetos con alto porcentaje de errores u omisiones en sus respuestas.

Pasos sugeridos

1. Definir el porcentaje de datos perdidos por sujeto que se considerará como perjudicial en los análisis estadísticos posteriores.
2. Definir qué se realizará con el o los sujetos que cumplan este porcentaje o lo superen.
3. Buscar una estrategia que permita realizar un conteo de datos perdidos por sujeto y si se considera necesario asociar esto a un porcentaje.
4. Realizar la acción acordada por el equipo en relación con el o los sujetos identificados.

Estadística sugerida

Se recomienda que si la cantidad de datos perdidos supera el 10% de las observaciones totales se hagan análisis específicos para identificar la causa de estos errores u omisiones, revisando los consejos metodológicos y técnicas estadísticas avanzadas específicas para esta materia[34].

Ejemplo

En la base de datos de ejemplo (Tabla 6), se observa el porcentaje de datos perdidos por sujeto. De acuerdo con los criterios definidos por los investigadores se ha establecido eliminar a los sujetos que superen el 50% de datos perdidos. Por lo tanto, corresponde eliminar al sujeto número 12, que presenta un 60% de datos perdidos del total de los datos.

Etapa V

Reagrupación de variables con categorías de baja frecuencia

Descripción

Este proceso consiste en la reagrupación de los valores de variables categóricas para facilitar los análisis posteriores. Esta etapa guarda estrecha relación con el propósito del estudio, ante una eventual eliminación y/o reagrupación de algunas categorías. Es fundamental consi-

Tabla 4. Análisis de consistencia de datos

		Examen de próstata	
		Sí	**No**
Sexo	**Mujer**	1	6
	Hombre	6	5

Tabla 5. Base de datos de ejemplo con error de consistencia

Sujeto	Edad	Sexo	Prost Exam	Religión	Peso
1	26		Sí	Ateo	60,5
2	57	Mujer	No	Ateo	80,3
3	41	Hombre	Sí	Católico	78,4
4		Hombre	No	Bautista	58,1
5	25	Hombre	Sí	Ateo	98,5
6	30	Hombre	Sí	Católico	57,3
7	47	Hombre	No	Ateo	70,4
8	54	Hombre	No	Católico	69,2
9	38	Hombre	No	Luterano	104,3
10	30	Hombre	No	Católico	55,1
11	19	Hombre	Sí	Bautista	85,1
12		**Mujer**	**Sí**	Católico	73,5
13	25	Hombre	Sí	Católico	86,7
14	34	Hombre	Sí	Bautista	99,3
15	37	Mujer	No	Bautista	62,1
16	48	Mujer	No	Católico	61,2
17	72	Mujer	No	Luterano	110,4
18	65	Mujer	No	Católico	44,4
19	30	Mujer	No	Ateo	64,5
20	45		No	Ateo	79,9

Tabla 6. Base de datos de ejemplo con análisis de datos perdidos

Sujeto	Edad	Sexo	Prost Exam	Religión	Peso	Frecuencia datos perdidos	% datos perdidos
1	26		Sí	Ateo	60,5	1	20
2	57	Mujer	No	Ateo	80,3	0	0
3	41	Hombre	Sí	Católico	78,4	0	0
4		Hombre	No	Bautista	58,1	1	20
5	25	Hombre	Sí	Ateo	98,5	0	0
6	30	Hombre	Sí	Católico	57,3	0	0
7	47	Hombre	No	Ateo	70,4	0	0
8	54	Hombre	No	Católico	69,2	0	0
9	38	Hombre	No	Luterano	104,3	0	0
10	30	Hombre	No	Católico	55,1	0	0
11	19	Hombre	Sí	Bautista	85,1	0	0
12				**Católico**	**73,5**	**3**	**60**
13	25	Hombre	Sí	Católico	86,7	0	0
14	34	Hombre	Sí	Bautista	99,3	0	0
15	37	Mujer	No	Bautista	62,1	0	0
16	48	Mujer	No	Católico	61,2	0	0
17	72	Mujer	No	Luterano	110,4	0	0
18	65	Mujer	No	Católico	44,4	0	0
19	30	Mujer	No	Ateo	64,5	0	0
20	45		No	Ateo	79,9	1	20

derar en este apartado la prueba estadística que se utilizará para la comprobación de hipótesis, ya que dependiendo de la prueba se requiere cierta de frecuencia de sujetos en los grupos de estudio para mantener la robustez de los análisis. Consideramos esta etapa como opcional dentro del proceso de validación de la base de datos.

Objetivo

Identificar aquellas categorías de las variables cualitativas que presentan baja frecuencia y reagruparlas para facilitar las pruebas estadísticas posteriores.

Pasos sugeridos

1. Buscar una estrategia que me permita identificar los porcentajes de representación de las categorías de todas las variables cualitativas.
2. Acordar con el equipo de investigadores cómo se podrían reagrupar los valores de las categorías que presentan baja frecuencia.
3. Crear una nueva variable con los valores reagrupados.

Estadística sugerida

Tablas de frecuencia.

Ejemplo

En la base de datos de ejemplo se analiza la variable religión por ser la única variable categórica con más de tres valores y para esto se realiza una tabla de frecuencia para analizar las opciones de reagrupación (Tabla 7). Posteriormente, se deciden reagrupar las categorías de baja frecuencia Bautista y Luterana en "Otras". Creando una nueva variable con las categorías reagrupadas (Tabla 8).

Tabla 7. Tabla de frecuencia variable religión

Religión		Frecuencia	Porcentaje
Categorías	Ateo	6	30
	Bautista	4	20
	Católico	8	40
	Luterano	2	10

Etapa VI

Discretización de variables continuas

Descripción

En ocasiones, durante la investigación se hace necesario categorizar una variable continua para efectos de los análisis estadísticos posteriores. Siendo ideal efectuar este procedimiento de forma racional considerando los objetivos y finalidad de la investigación. Existen situaciones donde al realizar la comprobación de las hipótesis no se encuentran los resultados esperados, por lo que se recomienda considerar la o las variables, tanto en su forma continua como categórica, en las hipótesis propuestas *a priori*. Consideramos esta etapa como opcional dentro del proceso de validación de la base de datos.

Objetivo

Generar una variable categórica basada en una variable cuantitativa.

Pasos sugeridos

1. Definir con el equipo de investigadores si existe una variable cuantitativa que pudiera discretizarse para enriquecer los análisis estadísticos posteriores.
2. Determinar con el equipo de investigadores cuál será el criterio de discretización considerando la evidencia teórica o los estadísticos de tendencia central o dispersión a utilizar.
3. Crear una nueva variable discretizada.

Ejemplo

Se decidió discretizar la variable edad de acuerdo con su mediana, la que corresponde a 37,5 años; creando una nueva variable nominal con dos categorías: "bajo la mediana" y "sobre la mediana" (Tabla 9).

Tabla 8. Base de datos de ejemplo con recodificación de la variable religión

Sujeto	Religión	Religión reagrupada
1	Ateo	Ateo
2	Ateo	Ateo
3	Católico	Católico
4	**Bautista**	**Otro**
5	Ateo	Ateo
6	Católico	Católico
7	Ateo	Ateo
8	Católico	Católico
9	**Luterano**	**Otro**
10	Católico	Católico
11	**Bautista**	**Otro**
12	Católico	Católico
13	Católico	Católico
14	**Bautista**	**Otro**
15	**Bautista**	**Otro**
16	Católico	Católico
17	**Luterano**	**Otro**
18	Católico	Católico
19	Ateo	Ateo
20	Ateo	Ateo

Tabla 9. Discretización de variable edad

Sujeto	Edad	Reagrupación de edad
1	26	Bajo la mediana
2	57	Sobre la mediana
3	41	Sobre la mediana
5	25	Bajo la mediana
6	30	Bajo la mediana
7	47	Sobre la mediana
8	54	Sobre la mediana
9	38	Sobre la mediana
10	30	Bajo la mediana
11	19	Bajo la mediana
13	25	Bajo la mediana
14	34	Bajo la mediana
15	37	Bajo la mediana
16	48	Sobre la mediana
17	72	Sobre la mediana
18	65	Sobre la mediana
19	30	Bajo la mediana
20	45	Sobre la mediana

Discusión

Este manuscrito pretende dar a conocer un detallado proceso metodológico, que debe ser realizado antes de los análisis exploratorios, llamado *validación de la base de datos*. Este proceso permite detectar y corregir a tiempo las anomalías de la base de datos, pudiendo garantizar que los análisis posteriores se realizaron con un base de datos sin incongruencias y errores que pudieran afectar la rigurosidad del análisis que se desarrollará.

La literatura técnica menciona que los análisis exploratorios son importantes; sin embargo, no hay consenso sobre cuál es el proceso que se debe llevar a cabo para dar cuenta de un análisis exploratorio riguroso. Así, tampoco existe consenso de cómo deben ser reportados estos análisis en los artículos que se publican. Los que al

parecer deberían ser cada vez más exigidos en las revistas que pretendan mejorar la calidad metodológica de sus publicaciones, considerando la importancia de estos y cómo pueden afectar los resultados y conclusiones de los estudios.

Una de las variantes más relevantes dentro de los análisis estadísticos es poder detectar los errores, los que comúnmente no son percibidos e inevitablemente perjudican las inferencias estadísticas. No solo los errores son importantes, también hay que tener en vista a los datos faltantes (omisiones), los que son inevitables en la investigación clínica, teniendo el potencial de socavar la validez de los resultados de la investigación, siendo a menudo pasados por alto en la literatura médica.

Es fundamental mantener una actitud de escepticismo ante los datos, independiente de la confianza o experiencia de quien haya realizado el ingreso de estos. Ya que la consecuencia de no realizar una buena validación de los datos no solo va a generar inferencias erróneas, sino que va a requerir de muchas horas para la re-revisión y nueva ejecución de análisis, escritura e interpretación de resultados y conclusiones.

No debemos olvidar que detrás de los errores o dificultades que encontremos en nuestros datos, puede haber múltiples causales, siendo fundamental preocuparnos de definir nuestras variables *a priori*, logrando de esta forma homogeneizar la información que se registra. Como por ejemplo en el caso de cuando se pregunta por el día en que un recién nacido recibió cierta vacuna. Puede haber sujetos que cuenten el primer día como "día 0" y otros como "día 1". O también, en el contexto de recién nacidos, cuando se pregunta por el número de hijos, se debe dejar explícito si se considera o no el niño recién nacido. O si preguntan por si la madre es fumadora y por el número de cigarrillos fumados. Se debe validar que, todos aquellas que reportan ser no fumadoras, tengan vacía la celda de número de cigarrillos. En las bases se suele encontrar no fumadoras con número de cigarrillos "0" o "vacíos". Similarmente ocurre con "¿Estuvo hospitalizada /UCI?" y "días de hospitalización/UCI". El proceso de la validación del instrumento de recolección de información es fundamental para asegurarnos que no se generen errores, así como también la correcta confección y socialización del libro de códigos.

Otro problema fundamental es sobre la decisión de qué se realizará con los valores extremos que se encuentren; posiblemente esto sea debate para otro artículo, donde están implicados múltiples factores para tomar una buena decisión al respecto[35], y en este artículo no se pretende orientar hacia una decisión específica sobre qué hacer con estos datos, siendo enfáticos en que estas decisiones deben ser tomadas por el equipo de investigadores, contando con a lo menos un experto en bioestadística.

Entendemos que, independientemente de nuestra propuesta de validación de base de datos, pueden surgir otros pasos en el proceso, especialmente considerando la disciplina que lo ejecute. La finalidad de este artículo es más que nada orientativa para los equipos de investigadores, especialmente aquellos en proceso de formación.

Esperamos que a futuro la descripción del proceso de validación de datos sea obligatoria en el proceso investigativo, fomentando de esta forma procesos reflexivos, con directrices que permitan establecer la sistematización en la validación de datos, de manera eficiente y ética dentro de cada disciplina.

Agradecimientos: Se agradece a ANID - MILENIO - NCS2021_013.

Referencias bibliográficas

1. Álvarez Cácerez R. Estadística Aplicada a las Ciencias de la Salud. España: Díaz de Santos; 2007.
2. Altman D G, Goodman S N. Transfer of technology from statistical journals to the biomedical literature. JAMA 1994; 272: 129-32. https://doi.org/10.1001/jama.1994.03520020055015.
3. Hogg R V, McKean J, Craig A T. Introduction to mathematical statistics. Upper Saddle River, N.J: Pearson Education; 2005.
4. DeGroot MH, Schervish MJ. Probability and statistics. Pearson Education. 2012.
5. Pagano M, Gauvreau K. Principles of Biostatistics. Chapman and Hall/CRC. 2018;2:584.
6. Martin L. Bioestadística para las Ciencias de la Salud. 2ª. Norma; 1994.
7. Holcomb Z. Fundamentals of Descriptive Statistics. Routledge; 2016.
8. Tukey J. Exploratory Data Analysis. Pearson, editor. Reading, Massachusetts. Estados Unidos; 1977. 712 p.
9. San Luis C. Análisis de datos en investigación. Primeros pasos. 1ª. Barcelona, España: Universidad Miguel Hernández; 2018. 513 p.
10. Vilar D, Xu J, D'Haro L F, Ney H. Error analysis of statistical machine translation output. In: Proceedings of LREC. 2006; 697-702. Disponible en: http://lorien.die.upm.es/~lfdharo/Papers/ErrorAnalysis_2006.pdf.
11. Jebb A T, Parrigon S, Woo S E. Exploratory data analysis as a foundation of inductive research. HRMR 2017; 27: 265-76. https://doi.org/10.1016/j.hrmr.2016.08.003.
12. Martín Q, Cabero M, Paz Y. Tratamiento estadístico de datos con SPSS. Paraninfo Cengage Learning. Madrid, España; 2007. 593 p.
13. Wang RY, Storey VC, Firth CP. A framework for analysis of data quality research. IEEE Trans Knowl Data Eng. 1995; 7: 623-40. https://doi.org/10.1109/69.404034.
14. Silverman B. Density Estimation for Statistics and Data Analysis. Routledge2. 2018;
15. Dasu T, Vesonder GT, Wright JR. Data quality through knowledge engineering. In: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge discovery and data mining - KDD '03. 2003. pages: 705-10. https://doi.org/10.1145/956750.956844
16. Benítez MÁ, Arias Á. Curso de Introducción a la Administración de Bases de Datos. IT Campus Academy.; 2015.
17. Halter CP. The PSPP Guide. Second Edition. An Introduction to Statistical Analysis. Creative Minds Press Group; 2017. 186 p.

18. GraphPad Software Inc. Prism7 GraphPad Statistics Guide. GraphPad. 2012;
19. Little R J, Rubin D B. Stactistical Analysis with missing data. John Wiley Sons. 2019;793.
20. Durán P. Los datos perdidos en estudios de investigación ¿son realmente datos perdidos? Arch Argent Pediatr 2005; 103 :566-8. https://www.sap.org.ar/docs/publicaciones/archivosarg/2005/566.pdf
21. Cuesta M, Fonseca-Pedrero E, Vallejo G, Muniz J. Datos perdidos y propiedades psicométricas en los tests de personalidad. An Psicol 2013; 29: 285-92. https://doi.org/10.6018/analesps.29.1.137901.
22. Graham J W. Missing data analysis: Making it work in the real world. Annu Rev Psychol. 2009; 60: 549-76. https://doi.org/10.1146/annurev.psych.58.110405.085530.
23. Hasperué W. Extracción de conocimiento en grandes bases de datos utilizando estrategias adaptativas. 2014.
24. Pym A. Exploring Translation Theories. Routledge. 2017;
25. Brüß C, Rocha EG, Kudchadker G, Singhal M, Ellouze R, Becher MMSD, et al. Process Mining and Natural Language. 2019.
26. Rahm E, Do H. Data cleaning: Problems and current approaches. IEEE Data Eng Bull. 2000; 23: 3-13. Disponible en: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.98.8661&rep=rep1&type=pdf.
27. Wienand D, Paulheim H. Detecting incorrect numerical data in DBpedia. Springer, Cham. 2014; 504-18. https://doi.org/10.1007/978-3-319-07443-6_34
28. Lindgren B W. Statistical Theory. Routledge. 2017;
29. Van Den Broeck J, Cunningham S A, Eeckels R, Herbst K. Data cleaning: detecting, diagnosing, and editing data abnormalities. PLoS Medicine 2005; 2: e267. https://doi.org/10.1371/journal.pmed.0020267.
30. Pérez-Tejada H E. Estadística para las ciencias sociales, del comportamiento y de la salud. 3a Edición. Cengaje Learning. Mexico; 2008. 664 p/318 p.
31. Babbie E. Fundamentos de la Investigación Social. Mexico: International Thomson Editores; 2000. 474 p.
32. Netter FH. Atlas De Anatomía Humana - 6ª Edición. Vasa. 2015.
33. Sterne J A C, White I R, Carlin J B, Spratt M, Royston P, Kenward M G, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. Br Med J 2009; 338:b2393. https://doi.org/10.1136/bmj.b2393.
34. Dagnino J. Datos faltantes (missing values). Rev Chil Anest 2014; 43: 332-4. https://revistachilenadeanestesia.cl/datos-faltantes-missing-values/
35. Gaspar J, Catumbela E, Marques B, Freitas A. A systematic review of outliers detection techniques in medical data - Preliminary study. HEALTHINF 2011 - Proceedings of the International Conference on Health Informatics. 575-82. https://doi.org/10.5220/0003168705750582