Capítulo 1. Introducción al R

2022-08-16

Contents

Introducción

Una de las interrogantes iniciales conlleva a preguntar ¿En qué consiste el análisis de datos ecológicos y ambientales (ANDEA)? En este sentido, el ANDEA, suele aplicarse en áreas como la estadística, como una herramienta asociada a la recolección, análisis, representación y la interpretación de datos y, por lo tanto, es fundamental para la mayoría de las actividades científicas en esta área del conocimiento. Para el profesional de las ciencias biológicas y ambientales, se convierte en una herramienta fundamental, debido a que ofrece elementos descriptivos e inferenciales de importancia en el diseño de experimentos y en la prueba de hipótesis, para cumplir con los requerimientos del método científico.

La jerarquía más elemental del análisis de datos se divide en poblaciones y muestras estadísticas, así como de variables. Las poblaciones estadísticas son referidas a todas las observaciones posibles, de la cual puede extraerse una muestra o subconjunto sobre la cual se puede llegar a conclusiones. Estas poblaciones pueden representar unidades ecológicas naturales, como la población de serpientes de cascabel de los cerros en Santa Marta, aunque pueden dividirse en subunidades más artificiales si se limita a la población de hembras de estas serpientes.

Los parámetros son atributos o características sobre los cuales se pueden sacar conclusiones de una población. La generalidad es que no sea posible hacer inferencias sobre la población, por lo que los parámetros son estimados a partir de estadísticos que caracterizan partes de una población. Estos parámetros están relacionados con las muestras (ej. La media o varianza muestral), siempre y cuando estas representen adecuadamente a una población (tener un tamaño mínimo y representativo de la población). Por esta razón, es frecuente realizar diseños experimentales que asuman que una muestra haya sido obtenida de la población de forma aleatoria o por el azar (esto maximiza la posibilidad de que la muestra represente a la población), pero también hay casos en los que el interés se orienta a que la muestra sea sistemática, en la que su configuración sea definida por el investigador.

El objetivo del ANDEA se centrará en poder descifrar patrones o señales que brinden información ofrecida por una o múltiples variables, observaciones y en algunos casos su integración con factores que representan a los datos. Adicionalmente, en la actualidad existe una gran disponibilidad de técnicas y de programas estadísticos robustos que facilitan su aplicación e interpretación.

Table 1: Términos comunes en estadística, referidos a una investigación sobre el contenido de amoniaco en las excretas de serpientes de cascabel

Termino	Definición	Ejemplo
Medición	Dato único cuya información refleja una característica de interés (ej. Longitud de un pez, número de individuos en un cuadrante, etc.)	Contenido de amoniaco en la excreta de una serpiente hembra.
Observación	Medida única o unidad experimental (ej. Un cuadrante, un sitio, un transepto, etc.)	mg/L de amoniaco en una serpiente hembra.
Población	El total de observaciones posibles que pueden ser medidas de la unidad en la cual se busca sacar conclusiones.	Contenido de amoniaco presente en las excretas todas las serpientes hembras.
Muestra	Un subgrupo representativo de la población a evaluar.	mg/L de amoniaco en 20 serpientes hembras.
Variable	Conjunto de mediciones del mismo tipo que componen a la muestra. La características medidas difieren (varían) de una observación a otra.	Contenido de amoniaco de cada serpiente hembra
Factor	Conjunto de observaciones tomadas en una población y que son clasificadas por algún atributo particular (gradientes de profundidad o altura, épocas climáticas, etc.).	Amoniaco en excreta de serpientes en un gradiente de altura

Otra de las interrogantes se describe así: ¿Por qué es necesario el análisis de datos en las ciencias naturales o ambientales? Y para dar respuesta, se debe ponderar el nivel de complejidad de la naturaleza, en donde sus patrones fluctúan considerablemente en espacio y en tiempo. Esto se puede corroborar al evaluar el comportamiento de los organismos y su respuesta a múltiples variables bióticas y abióticas, que pueden actuar de forma simultánea.

Uno de los aspectos a tener en cuenta en la pregunta anterior, es la correlación que pueda existir entre las variables colectadas, consiste en descifrar la información que subyace a la interacción de estos parámetros, algunos pueden descartarse dado a que generan ruido (variables colineales o distorsionadas), aportan la misma información que otras (redundantes), por lo que existen procedimientos que permitirán descartarlas y dejar solo aquellas que muestren estructuras importantes en la correlación de los datos. Esta dinámica hace parte importante del "principio de la parsimonia" orientado a cómo lograr explicar patrones o estructuras de los datos, con el menor número posible de variables, resumiendo o simplificando a la mayor cantidad de información expresada en las variables.

De acuerdo con lo anterior, se suele utilizar técnicas univariadas o multivariadas, que buscan identificar estructuras de los datos, basado en un conjunto reducido de dimensiones, por lo general se usan dos ejes construidos de forma matemática o computacional, en donde además se pueden visualizar a las variables y/o las observaciones, orientado principalmente a la exploración de hipótesis construidas previamente. Si el objeto del análisis es hacia la prueba de hipótesis, se aplican otras técnicas, que permitan evaluar diferentes variables, conservando el nivel de significancia (alfa).

1.1 Etapas del ANDEA.

El análisis de datos en el que interactúan múltiples variables suele realizarse en dos etapas generales, partiendo en de variables y de observaciones (Tablas 1 y 2), para intentar identificar patrones generales y que puedan ser explicados desde el contexto en estudio, para poder validar hipótesis propuestas previamente. De igual

forma se busca descifrar patrones ocultos a la complejidad de bases de datos que pueden presentar pocas hasta cientos de variables, para lo cual el ANDEA suele presentar una potencia importante en su entorno gráfico.

En este sentido, el ANDEA suele dividirse en dos grandes grupos de técnicas: (1) las **descriptivas** que explorar la estructura de los datos, definidos por variables, observaciones y en algunos casos incluyen factores. Estas técnicas buscan identificar variables que presenten algún tipo de correlación o combinación preferiblemente lineal para que facilite a la generación de modelos con una buena base matemática (ver capítulos de ordenación y de clasificación) y la inferencia estadística que corresponde al segundo grupo de técnicas que se describen posteriormente.

Es importante aclarar que las técnicas descriptivas, no necesariamente corresponden al **análisis exploratorio**, el cuál debería hacer parte inicial y de rutina para cualquier análisis de datos y está constituido principalmente por figuras univariadas y/o multivariadas, que permiten visualizar patrones generales en el comportamiento de nuestros datos (ver capítulo de figuras exploratorias), este componente suele ser subvalorado en muchos análisis, sin tener en cuenta que en algunos casos, basta con un buen análisis exploratorio para identificar los patrones de nuestros datos.

(2) la **inferencia multivariada**, es desarrollada por técnicas orientadas a la prueba de hipótesis y/o a la generación de modelos que pueden ser construidos a través de ecuaciones matemáticas y que exigen el cumplimiento de diferentes requisitos o supuestos numéricos (ver capítulo de pruebas de hipótesis). Estas técnicas permiten a su vez, elegir a las variables de mayor importancia en la explicación de un experimento realizado, independiente del número de variables analizadas de manera simultánea, debido a que son reguladas por un valor **alfa**, que corresponde al nivel de significancia, establecido previamente por el investigador (normalmente una probabilidad menor de 0,05). Estas técnicas aportan control sobre la tasa de error de los diseños estadísticos.

La diferencia fundamental entre la inferencia y la descriptiva, es que las últimas no requieren del cumplimiento de supuestos (homogeneidad de la matriz de varianza-covarianza, normalidad multivariada o la independencia, entre otros), o que las variables presenten una distribución esperada. Hace más de una década, la generalidad consistía en la aplicación de las técnicas descriptivas y exploratorias, pero con el desarrollo de la teoría estadística y de herramientas computacionales, en la actualidad es mucho más frecuente la aplicación de inferencia estadística, de pruebas que cumplen los supuestos (paramétricas), como de aquellas que no exigen de supuestos (no paramétricas o permutacionales).

Son diferentes los retos a los que se somete el investigador, al realizar el análisis e interpretación de bases de datos con diferentes variables. Entre las situaciones que pueden presentarse se destacan (1) Enfrentarse a numerosas variables y de diferente naturaleza, que han sido tomadas en diferentes localidades y periodos de muestreo, para lo cual se debe establecer una estrategia parsimoniosa que resuma y maximice la información ofrecida por los datos, (2) desarrollar modelos mentales que previo al análisis, permitan establecer la mejor ruta en el ANDEA, dada la infinidad de técnicas que tiene a disposición, (3) en cuanto a la inferencia estadística, el tener la posibilidad de decidir si es necesario realizarla a nivel multivariado o si basta con pruebas univariadas para desarrollar el análisis requerido, (4) tener la capacidad de interactuar con diversas disciplinas, para abordar diseños multivariados que integren variables de diversas tipologías.

Estas y otras situaciones son las que hacen de esta disciplina un campo amplio y de mucha importancia para responder a problemáticas del entorno, que exigen el tratamiento de múltiples variables, en diferentes escalas espaciales y temporales, sin dejar a un lado la realización previa de un buen diseño experimental, que permita dar respuesta a hipótesis, controlando los tipos de error que se pueden presentar. Resumiendo lo anterior, el presente documento intentará proponer un esquema general para el ANDEA, que inicie con la exploración de los datos, a través de estadísticos básicos de tendencia central y de dispersión, así como de relaciones entre variables y muestras, para luego ser complementados con pruebas descriptivas e inferenciales (en caso de ser posible), para dar respuesta a preguntas e hipótesis sobre datos univariados y multivariados de naturaleza ecológica y ambiental.

1.2 Requisitos.

El requisito inicial, es el de contar con múltiples variables (más de tres), para dar respuestas a preguntas de investigación, de muestras que han sido tomadas de manera apropiada y que representan a sus poblaciones estadísticas. Existen situaciones en que la aplicación de pruebas individuales, con cada una de las variables que caracterizan a la muestra en estudio limita el poder revelar la estructura completa de los datos, dada las interrelaciones que se pueden presentar entre variables.

Se requieren bases de conocimiento sobre aspectos de la estadística univariada, en especial de la bioestadística. Importante tener conocimiento de la distribución normal, pruebas para comparar dos o más muestras, regresiones y análisis de varianza (paramétricos y no paramétricos), así como de los supuestos que deben cumplirse y las estandarizaciones o transformaciones que se pueden realizar a las variables. Las pruebas anteriores presentan un análogo multivariado, que será mucho más fácil de entender con las bases mencionadas.

Otro requerimiento importante es el conocimiento que se debe tener del algebra lineal, especialmente del contexto matricial y vectorial que representa el insumo numérico estos análisis estadísticos (ver capítulo algebra lineal). Vale la pena mencionar que el componente multivariado del ANDEA se soporta completamente sobre matrices y vectores, tanto en las operaciones realizadas como en los insumos numéricos y gráficos. Si bien este documento no se enfoca al contexto matemático de cada prueba, si es necesario que el lector se familiarice con el origen y el significado de cada insumo matricial y de su importancia en cuanto al enfoque ambiental.

Otro aspecto relevante es la experiencia previa que se tenga con el análisis y reporte de resultados estadísticos en pruebas univariadas, como el registro de valores de significancia "p", del estadístico o de los grados de libertad, elementos que son universales en el análisis de hipótesis univariadas y multivariadas. Mucho más importante es la capacidad de análisis e interpretación de resultados numéricos y gráficos desde una perspectiva ecológica o ambiental, sin demeritar la importancia algebraica de las pruebas, aunque esto último no sea una prioridad para este manual.

1.3 Objetivos.

Debido a su aplicación multidisciplinaria, el ANDEA presenta numerosos objetivos específicos, que, para el caso de este texto, pueden ser asignados a tres grupos más generales, que se describen a continuación.

Un primer objetivo se orienta a brindar información, especialmente práctica, sobre los detalles y el significado biológico o ambiental de diferentes técnicas multivariadas descriptivas e inferenciales, en cuanto a su contexto general, su aplicación, restricciones y su articulación con otras pruebas complementarias, priorizando en el componente práctico.

Como segundo objetivo, se busca promover la capacidad intuitiva para seleccionar las técnicas más apropiadas en el análisis de datos, dependiendo de la situación que se presente. Propendiendo porque cada técnica sea antecedida por un buen análisis exploratorio, a partir de gráficas que constituyen una de las fortalezas de la plataforma R.

El tercer objetivo consiste en la habilidad que se pueda adquirir en el desarrollo de cada técnica, por lo que cada una finaliza con un pequeño cuestionario, orientado al fortalecimiento conceptual e intuitivo, aunque para algunos casos, será necesario continuar entrenando para controlar cada una de las situaciones que se puedan presentar, con estos análisis.

Estos objetivos pueden ser alcanzados, si se cuenta con el planteamiento de preguntas e hipótesis de forma adecuada, para brindar facilidad al desarrollo e interpretación de las técnicas aplicadas. Esta habilidad se adquiere con el aprendizaje constante, especialmente si se trabaja con herramientas de cierto nivel de complejidad como la plataforma R, que se aplicará en capítulos posteriores. En resumen, ¡con la práctica se hacen buenos maestros para el análisis de datos!

Table 2: Valores hipotéticos para datos multivariados

Sitios	Zonas	Regiones	Textura	Usos	T	K	Ca	Mg	Arcillas	Cyod	Dgbi	Lapae
1	Zona 1	Alta	G	1	27.2	19.54	98.2	126.64	10.75	10	0	3
2	Zona 1	Alta	G	1	27.5	21.94	63.8	113.49	21.35	12	8	4
3	Zona 1	Alta	G	1	26	23.89	63.3	118.42	28.61	11	9	7
4	Zona 1	Baja	SG	1	29	18.06	77.4	123.36	23.12	*df	3	9
5	Zona 1	Baja	SG	2	27.5	13.98	115.6	93.75	26.48	7	7	12
6	Zona 1	Baja	SG	2	28.5	13.69	98.1	100.33	32.67	8	9	12
7	Zona 1	Media	G	2	29	13.94	111.5	105.26	12.55	6	0	15
8	Zona 1	Media	G	2	*df	13.96	105.2	97.04	20.99	11	0	14
9	Zona 2	Media	G	1	28	4.33	15.1	15.62	19.67	11	0	13
10	Zona 2	Norte	F	1	25	5.76	17.8	16.45	21.71	13	5	1
11	Zona 2	Norte	F	1	24	1.50	7.3	2.63	30.15	15	0	7
12	Zona 2	Norte	F	1	25	1.24	5.8	1.97	19.67	17	0	0

1.4 Tipos de datos.

La forma común para presentar los datos es a través de matrices, en donde las filas agrupan, muestras y/u observaciones y las columnas relacionan a las variables medidas para cada observación. Las observaciones pueden representar a individuos, lugares, periodos, etc. (Tabla 2).

Las variables no necesariamente deben ser del mismo tipo y pueden ser agrupadas en cuatro grandes grupos: (1) Las nominales, que suelen ser categóricas (ej. colores, sexos, etc.), (2) Las ordinales, definen un orden, aunque no mantengan una magnitud proporcional (ej. Razas, niveles de contaminación, etc.), (3) Intervalos, presentan una magnitud proporcional a lo largo de una escala y la ubicación de su origen (cero) es arbitraria (ej. Escalas de temperatura en grados Celsius o Fahrenheit), (4) Razón o Proporción, corresponde a magnitudes relativas de las variables (ej. La edad, el peso, el tamaño, etc.).

En la Tabla 1, se presentan diferentes tipos de muestras y variables que pueden haber sido tomadas en estudios ambientales. (1) Sitios, corresponden a las observaciones, para este caso son los lugares visitados, es normal que puedan ser numerados o codificados. (2) Zonas y Regiones corresponden las muestras o factores que, a diferencia de los factores, cada una de sus categorías o niveles permiten presentar varios valores o replicas por cada nivel. (3) Texturas, corresponde a una variable nominal o categórica, que representa a texturas del suelo presente en los sitios evaluados. (4) Usos, representa a una variable de tipo ordinal, en la que se definen dos categorías de usos del suelo. (3) K, Ca, Mg y Arcillas, son variables continuas (con decimales), de tipo proporción, que corresponden a parámetros físicos y químicos del suelo. Cyod, Dgbi y Lapae, son variables discretas (conteos de individuos), de tipo proporción que representan la abundancia de tres especies vegetales, cuyos nombres han sido codificados.

Estas variables pueden ser evaluadas inicialmente a partir de estadística descriptiva univariada, como medidas de tendencia central (media, mediana, ...), de dispersión (desviaciones, varianzas, ...) o de posición (cuartiles, percentiles, ...). Posteriormente puede aplicarse técnicas multivariadas, para visualizar patrones con la integración simultánea de todas las variables.

Se puede presentar que falten datos para algunas variables, como se puede apreciar en la Tabla 2 (*df), lo cual puede ocurrir por varias razones, ya sea porque no fueron tomados o a que su valor es atípico por un posible problema en el instrumento de medida, lo cual indujo a su eliminación. Estos vacíos de información o de datos faltantes se pueden corregir con algunos métodos de imputación, como el realizado en técnicas de simulación con Monte Carlo en donde los valores faltantes son reemplazados por $m > 1$, en donde m es el número de simulaciones (3 a 10), esto debe ser contrastado con intervalos de confianza, para incorporar cierto nivel de incertidumbre a los datos requeridos, sin perder de vista que estos datos no son reales (estimados).

1.5 Generalidades del análisis de datos.

En el ANDEA es común que nos enfrentemos a diferentes situaciones, en la que se encuentren distribuidos los datos, para llevarnos a pensar que deben ser analizados en un contexto multivariado. La práctica y la intuición son las herramientas fundamentales para poderlos categorizar los siguientes tipos de casos.

1. Una muestra con diferentes variables (más de tres), en cada observación o muestra a evaluar.
2. Una muestra con dos grupos de diferentes variables medidas en cada observación.
3. Dos o más muestras (niveles de un factor) con distintas variables para cada observación.
4. Diferentes muestras en diferentes periodos (dos factores) con distintas variables en cada observación.

Para el caso (1) se pueden presentar diferentes tipos de análisis a realizar:

- Probar la relación de las medias de cada variable en la muestra evaluada (ver Análisis con Figuras Exploratorias).
- Encontrar algunas dimensiones que sean combinación lineal de las variables, que definan alguna estructura de los datos y permitan explorar modelos lineales (ver Componentes Principales).
- Determinar algunas dimensiones que caractericen a las variables y a sus intercorrelaciones (ver Análisis Factoriales).
- Clasificar a las observaciones, de acuerdo con su nivel de similitud determinada por las variables (ver Análisis de Clúster).

Para el caso (2) se pueden presentar diferentes tipos de análisis a realizar:

- Determinar el nivel de correlación en dos grupos de datos multivariados, evaluados sobre la misma muestra (ver análisis MANTEL).
- Relacionar los dos conjuntos de variables, en donde uno depende del otro (ver correspondencias canónicas y análisis de redundancias).
- Seleccionar el subconjunto de variables de un grupo (explicativas) que presente la máxima relación con el otro grupo de variables (ver Bioenv y Análisis de Redundancias).

Para el caso (3) se describen los siguientes análisis a realizar:

- Ordenar a las observaciones y las variables en pocas dimensiones, de acuerdo con su similitud y comparar gráficamente a las muestras (ver PCA, nMDS y AC).
- Comparar dos muestras de acuerdo con los promedios de sus variables (ver Prueba T2 de Hotelling, MANOVAS paramétricas y permutacionales).
- Comparar a más de dos muestras de acuerdo con los promedios de sus variables (ver MANOVAS paramétricas y permutacionales).
- Comparar a más de dos muestras de acuerdo con los promedios de sus variables, mediante métodos no paramétricos o que no requieran supuestos del Manova (ver MANOVAS permutacionales).
- Encontrar la combinación lineal de las variables que clasifique mejor a las muestras y verificar la membresía de cada observación a su muestra correspondiente (ver LDA y CDA).

- Determinar un método que permita validar la clasificación realizada de las observaciones o individuos a sus muestras, a partir de una escala de similitud o de distancia (ver Análisis de Clúster).

Para el caso (4) se describen los siguientes análisis a realizar:

- Ordenar a las observaciones y las variables en pocas dimensiones, de acuerdo con su nivel de similitud y comparar gráficamente a las muestras y/o a periodos (ver técnicas de Ordenación).
- Comparar a más de dos muestras y a periodos de acuerdo con los promedios de sus variables (ver MANOVAS paramétricas y permutacionales).
- Comparar a más de dos muestras y periodos de acuerdo con los promedios de sus variables, mediante métodos no paramétricos o que no requieran supuestos del Manova (ver MANOVAS permutacionales).