Contribución de la estadística en la investigación. Elementos básicos de inferencia estadística.

Contribution of statistics in research. Basic elements of statistical inference.

Celina Beltrán; Ivana Barbona
Facultad de Ciencias Agrarias, Universidad Nacional de Rosario, Argentina
beltranc36@yahoo.com.ar

Abstract

This article takes up the concepts and applications presented in Beltrán, C. and Barbona, I. (2016). It is aimed at those who are initiated in the investigations that involve data to demonstrate the veracity of a hypothesis. It includes from data base design and definition of the variables to the election of the suitable analysis technique. The objective is to show the contribution of statistical methodology in data analysis and decision making.

Keywords: statistics, inference, data analysis

Resumen

Este artículo retoma los conceptos y aplicaciones presentados en Beltrán, C. y Barbona, I. (2016). Está dirigido a quienes se inician en las investigaciones que involucran datos mediante los cuales se pretende demostrar la veracidad de una hipótesis. Contempla desde la construcción de la base de datos y definición de las variables, hasta la elección de la técnica adecuada. El objetivo del mismo es evidenciar el aporte de la metodología estadística en el análisis de datos y la toma de decisiones.

Palabras claves: estadística, inferencia, análisis de datos.

1. INTRODUCCION

¿Qué es la ESTADÍSTICA? De todas las definiciones que se le ha brindado a la ESTADÍSTICA la más completa establece que es una ciencia que tiene por objetivo realizar inferencias basadas en datos, para lo cual proporciona técnicas de recolección, resumen y presentación de éstos como así también permite cuantificar el error cometido en dicho proceso de inferencia.

¿Quién debe llevar a cabo el ANÁLISIS ESTADÍSTICO de los datos? La inferencia estadística proporciona una asistencia fundamental en una gran variedad de campos de investigación en los que se trabaja con observaciones. Sin embargo, el lector debe tener muy en claro que sería sospechoso que un estadístico pudiese sacar conclusiones sobre economía, medicina, lingüística o cualquier otra disciplina sin la ayuda del conocimiento del experto en dicha área. De la misma manera hay que

resaltar que, debido a la facilidad de utilización de los software estadísticos, los expertos en las distintas disciplinas se ven tentados a aplicar metodología estadística y sacar conclusiones sin contar con el conocimiento necesario para llevar a cabo dicha tarea. Estos dos aspectos revelan la esencia interdisciplinaria de la estadística.

2. LAS OBSERVACIONES

2.1.Población y Muestra. Muestreo aleatorio.

Un aspecto fundamental en estadística es llegar a conclusiones, mediante una técnica de inferencia estadística, sobre un grupo "grande" de datos (población) habiendo observado y analizado sólo una parte de dicho grupo (muestra). En este sentido la población sería el conjunto total de elementos definidos en un determinado tiempo y lugar, que son objeto de estudio. La muestra entonces se define como un subconjunto de la población.

A cada elemento de la población se lo denomina unidad experimental, y un grupo de ellas conformarán la muestra. Sobre estas unidades se van a estudiar una o más características denominadas variables. Los parámetros son medidas asociadas a estas variables calculadas sobre toda la población y las estadísticas son aquellas medidas calculadas sobre una muestra. Dado que un parámetro sólo se obtiene al observar la totalidad de la población, es un valor fijo mientras que una estadística es un valor que varía de una muestra a otra, por lo tanto las estadísticas o estimadores son variables aleatorias.

Puesto que se va a concluir sobre la población observando la muestra, se buscará entonces que ésta sea lo "más parecida posible" a la población para llegar a conclusiones acertadas o válidas. Es frecuente observar que se le exige a la muestra que sea "representativa" de la población de la cual fue extraída. Sin embargo, esa representatividad buscada, siempre es respecto a alguna característica de la población que se está estudiando y si es una característica que no se conoce (y por eso es objeto de estudio) poco se podrá decir acerca de la representatividad de la muestra, ya que si se supiera como debe ser la muestra para parecerse a la población ya no existiría incertidumbre y dejaría de ser necesario el muestreo de la misma. Entonces, ¿cómo se debe elegir a las unidades que pertenecerán a la muestra? ¿Qué significa una muestra aleatoria? ¿En qué consisten los distintos tipos de muestreos?

El procedimiento por el cual se obtiene la muestra se denomina método de muestreo. Cuando un método de muestreo asigna a cada unidad de la población una probabilidad de pertenecer a la muestra, se denomina muestreo aleatorio o probabilístico. Una muestra aleatoria permite realizar inferencias confiables sobre la población de la que se ha seleccionado debido a que se puede estimar la magnitud del error. Las distintas maneras de asignar una probabilidad a cada unidad de la población dan origen a los distintos métodos de muestreo.

2.2.Datos observacionales y experimentales

Los datos observacionales provienen de estudios observacionales en los cuales se observa el efecto "crudo" de la naturaleza en las unidades de análisis sin que se explore una acción diseñada por el investigador.

En los estudios experimentales, el investigador manipula un fenómeno en estudio y evalúa el efecto que dicha intervención causa sobre éste.

Los datos observacionales presentan la desventaja de que la observación rigurosa suele estar afectada por muchas variables o factores que no permiten determinar el verdadero efecto de un determinado factor en estudio. A esas variables se las denomina variables de confusión. Una manera de solucionar esta participación de las variables de confusión es controlar, esto es, medir todas aquellas variables o factores que de alguna manera se piensan asociados con el evento en estudio.

No obstante, sólo se pueden controlar las variables que se conocen como relacionadas, mientras que otras que se ignoran van a permanecer ocultas y quizás modificando las relaciones entre las variables que estamos estudiando. ¿Cómo se soluciona? La solución es la ejecución de diseños de experimentos con asignación aleatoria de las unidades experimentales, ya que implementando un adecuado proceso de aleatorización, éste se encarga de que todas las variables de confusión se distribuyan de manera similar en todos los grupos en estudio. La aleatorización no elimina el efecto de la variable de confusión, sino que hace que su efecto esté presente de la misma manera en cada grupo, pretendiendo que estos grupos a comparar y evaluar sólo difieran en la característica que se estudia y sean homogéneos respecto a las restantes variables.

Asimismo, los experimentos diseñados también presentan limitaciones. Una limitación son los costos, cuanto más complejo es el diseño del experimento, mayor es el costo y mayor será el requerimiento de personal y recursos técnicos para ejecutarlo. Asimismo, cuando existen eventos raros y difíciles de hallar en la población, se debe contar con un tamaño de muestra suficientemente grande para lograr estimaciones consistentes. Por otro lado, quizás más frecuente en biología y medicina, también hay limitaciones éticas en aquellos casos en los cuales se somete a un individuo a un determinado factor y se le impide recibir un tratamiento que ya se ha demostrado que es beneficioso.

2.3.Relevamientos puntuales y longitudinales

Los relevamientos de los datos pueden clasificarse en puntuales y longitudinales. Los primeros se caracterizan porque los datos se refieren a un mismo momento del tiempo. Se denominan también "de corte transversal" precisamente porque los datos se obtienen mediante un corte en el eje del tiempo. En este tipo de datos se supone que los valores de una variable entre los distintos elementos que conforman la muestra son independientes.

En los relevamientos longitudinales los datos corresponden a observaciones de una o varias variables a intervalos de tiempo. La característica principal en este tipo de datos, y que debe considerarse al momento de elegir la técnica estadística para su análisis, es que no es adecuado asumir o suponer que los valores de una variable, medida sobre una misma unidad experimental, en distintos momentos de tiempo son independientes.

2.4.Variables. Clasificación

Las variables, como se menciona previamente, son aquellas características que se miden o registran en cada una de las unidades en estudio. Las mismas pueden clasificarse de la siguiente manera:

- ✓ Cualitativas: Son aquellas que clasifican a las unidades en categorías. Éstas pueden ser nominales u ordinales.

En las cualitativas nominales las categorías no presentan un determinado orden, son ejemplos de este tipo de variables: sexo, lengua materna, tipo de escuela, categoría morfosintáctica, etc.

En las cualitativas ordinales las categorías presentan un orden natural, por ejemplo Severidad de una enfermedad (Leve/Moderada/Severa), nivel de instrucción alcanzado (primario incompleto/primario completo/secundario incompleto/…./universitario completo), etc.

En algunos casos a las observaciones cualitativas se les puede asignar un número, por ejemplo: sexo masculino=1 y sexo femenino=2 sin que eso signifique un orden entre las categorías, a diferencia de la codificación de 1 a 3 en el ejemplo de la variable severidad de una enfermedad (Leve=1, Moderada=2, Severa=3).

- ✓ Cuantitativas: Son aquellas cuyos valores provienen de mediciones o conteos. Ellas se clasifican en continuas y discretas.

Variables cuantitativas discretas: Son aquellas tales que la unidad de medida no es divisible, es decir la unidad de medida se debe definir sólo en términos de números enteros. De esta manera, una variable discreta sólo puede tomar un conjunto finito o infinito numerable de valores.

Son variables cuantitativas discretas: número de hijos por familia, número de verbos por texto, etc.

Variables cuantitativas continuas: Son aquellas que pueden asumir cualquier valor dentro de intervalo de números reales.

Las medidas de peso, altura, temperatura, etc. representan variables continuas. Dentro de las investigaciones algunas variables continuas podrían ser porcentajes o proporciones. Algunas técnicas estadísticas dan como resultado de su aplicación algún índice calculado sobre las variables originales (por ejemplo Análisis de Componentes Principales) y estos nuevos índices son variables continuas que resumen información original y pueden utilizarse en posteriores estudios.

Una variable cuantitativa puede ser transformada en una categórica ordinal definiendo puntos de corte. Por ejemplo la variable EDAD puede trabajarse como una variable ordinal con 3 categorías: 0-15 (de 0 a 15 años) 16-45 (de 16 a 45 años) y 46-+ (46 y más años). Es importante remarcar que si se registra la variable como cuantitativa, luego puede ser transformada a una variable categórica, pero si se registra en forma categórica no es posible volver al dato exacto de la edad sin recurrir a la unidad experimental nuevamente. Esto sugiere medir las variables de la manera más desagregada posible para tener mayor información.

2.5. Organización de la información

La materia prima de la cual se nutre todo análisis estadístico son los datos. El dato se define como una representación simbólica de un atributo o variable de una entidad (o unidad experimental). Al conjunto de datos pertenecientes a un mismo contexto, es decir, que provienen de un estudio observacional o experimental se lo denomina Base de Datos. Finalmente, la información se obtiene mediante la organización, procesamiento y análisis dicha base.

Los datos se organizan en una matriz que se define como un conjunto de números ordenados en filas y columnas. Una matriz de dimensión nxp se refiere a una matriz de n filas y p columnas. Considérese un experimento en el que se obtiene la información correspondiente a p variables medidas sobre n unidades experimentales. Se usará la notación xij para indicar el valor de la variable j en el individuo i. Por lo tanto, las n mediciones de las p variables pueden disponerse de la siguiente manera:

	variable 1	variable 2	⋯	variable j	⋯	variable p
unidad 1	x_{11}	x_{12}	⋯	x_{1j}	⋯	x_{1p}
unidad 2	x_{21}	x_{22}	⋯	x_{2j}	⋯	x_{2p}
⋮	⋮	⋮		⋮		⋮
unidad i	x_{i1}	x_{i2}	⋯	x_{ij}	⋯	x_{ip}
⋮	⋮	⋮		⋮		⋮
unidad n	x_{n1}	x_{n2}	⋯	x_{nj}	⋯	x_{np}

es decir, en una matriz rectangular, **X**, de dimensión nxp (n filas y p columnas)

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2p} \\ x_{31} & x_{32} & \cdots & x_{3j} & \cdots & x_{3p} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nj} & \cdots & x_{np} \end{bmatrix}$$

Esta forma de organizar la información es la requerida por la mayoría de los programas que se ocupan del análisis estadístico de los datos. Cabe destacar que la forma de disponer matricialmente los datos es la que se corresponde con una planilla de cálculo.

En el momento de diseñar y cargar los datos en la base es importante tener en cuenta algunas recomendaciones:

- ✓ Una variable cuantitativa puede ser transformada en una categórica o cualitativa definiendo puntos de corte.
- ✓ Como se indicó anteriormente, recordar que si se registra la variable como cuantitativa, luego puede ser transformada a una variable categórica, en cambio si se registra en forma categórica *no es posible volver al dato exacto* sin recurrir a la unidad experimental nuevamente.
- ✓ Respetar el formato de fechas, es decir, definir en qué formato se van a registrar las variables cuyos valores sean fechas y respetar ese formato para todas las observaciones, por ejemplo dd/mm/aa, dd/mm/aaaa, etc
- ✓ Para datos longitudinales registrar todos los valores en cada fila por más que sea redundante en alguna columna. Por ejemplo:
- ✓ Hacer un listado de las variables con su nombre, clasificación, categorías detalladas (en el caso de cualitativas) o rango de valores posibles (en el caso de cuantitativas).
- ✓ Considerar una columna identificando la unidad experimental en la matriz de datos. Este es un detalle importante y útil ya que hay situaciones en las que se debe "volver" a la unidad experimental de la cual provino el registro para chequearlo y revisarlo y si no se registró o identificó a la unidad experimental esta tarea es casi imposible de llevar a cabo en muestras numerosas.

2.6.Análisis Descriptivo de los Datos.

Luego de organizar la información en la base de datos, resulta conveniente realizar un análisis descriptivo de las variables. Este tipo de análisis consiste en un examen previo de los datos que permite explorar el comportamiento de las variables, tanto en forma individual como conjunta (sin generalizar conclusiones a nivel población), y así poder inspeccionar, entre otras cosas:

- ✓ Presencia de outliers o valores extremos, es decir observaciones que toman valores extremadamente altos o bajos dentro del campo de variación.

- ✓ Inconsistencias en las variables, como valores fuera del campo de variación o imposibles de observar según la naturaleza de la variable.
- ✓ Datos faltantes.
- ✓ Relaciones, patrones, tendencias o correlaciones entre las variables.
- ✓ La estructura de covariancias entre variables.
- ✓ El cumplimiento de supuestos como normalidad, linealidad, homocedasticidad, etc.

El análisis exploratorio también se utiliza para generar nuevas hipótesis de interés a probar posteriormente mediante un análisis inferencial.

En primer lugar conviene realizar un análisis univariado, es decir, examinar cada una de las variables por separado para luego continuar con un análisis de las variables en forma conjunta. La inspección se puede llevar a cabo en forma gráfica, en tablas y/o calculando medidas descriptivas de posición y de dispersión.

Algunos gráficos y tablas que pueden utilizarse son:

- ✓ Variables Cualitativas
  - Tablas de Frecuencia. Tablas cruzadas.
  - Gráficos: Torta, Barras simples, Barras compuestas, barras subdivididas, etc
- ✓ Variables Cuantitativas
  - Tablas de distribución de frecuencias.
  - Gráficos: Histograma, Polígono acumulativo, Gráfico de bastones, Gráfico escalonado, Gráfico de Dispersión, Diagrama de caja, Gráfico de serie de tiempo, etc

Para el caso de las variables cuantitativas, además de los gráficos y tablas, pueden calcularse medidas descriptivas. Estas medidas consisten en valores que sirven para describir un conjunto de datos. Si son calculadas en base a las observaciones de una muestra se denominan estimadores, mientras que si su cálculo se realiza en base a observaciones de una población se denominan parámetros.

Existen dos tipos de medidas descriptivas, las denominadas de posición y las de dispersión. Las medidas de posición proveen información acerca de la posición de la distribución en el eje de valores de la variable. En cuanto a las medidas de dispersión, éstas se refieren a la forma en la cual los valores se encuentran dispersos o distribuidos alrededor de las medidas de posición. Es decir, miden la distribución de las observaciones alrededor de las medidas de posición.

Algunas de las medidas más utilizadas son:

- ***Promedio o media, mediana, modo, cuartiles (Medidas de posición)***
- ***Variancia, desvío estándar, rango, rango intercuartil, coeficiente de variación (Medidas de dispersión)***

Si bien pueden calcularse todas las medidas descriptivas para cada variable esto no implica que la totalidad de ellas van a brindar información útil, algunas hasta pueden resultar engañosas a la hora

de su interpretación. A continuación se presentan algunas consideraciones a tener en cuenta a la hora de aplicarlas:

- ✓ Tener en cuenta la forma de la distribución de la variable, es decir, si es simétrica o asimétrica. Para distribuciones simétricas trabajar con la media aritmética, mientras que para distribuciones asimétricas la mejor medida de posición es la mediana. Esto se debe a que la media es una medida de posición sensible a valores extremos mientras que la mediana por la forma en la que se realiza su cálculo es robusta respecto a la presencia de valores extremos o atípicos.
- ✓ Siempre que se describa una distribución mediante una medida de posición es conveniente acompañarla con una medida de dispersión. Si la medida de posición a utilizar es la media, la medida de posición correspondiente es el desvío estándar dado que éste mide da dispersión de las observaciones alrededor de la media aritmética. En el caso de utilizar la mediana como medida de posición la medida de dispersión correspondiente es el rango intercuartil, ya que éste da una idea de cuánto se dispersan las observaciones alrededor de la mediana.
- ✓ Cuando se quiere comparar dispersión en distribuciones de variables con distintas escalas de medida, o con la misma escala pero con promedios muy diferentes, la mejor medida es el coeficiente de variación. El coeficiente de variación es una medida de variación relativa (a diferencia del resto que son medidas de dispersión absoluta), ya que consiste en un porcentaje que nos dice cuánto del total de la media representa el desvío estándar. Al ser un porcentaje, es independiente de la escala de medida y permite la comparación entre distribuciones. Mientras que el resto de las medidas de dispersión absoluta se ven afectadas por la escala de medida de la variable.

Una vez realizado el análisis exploratorio de los datos, se puede utilizar la información captada por éste para tener una idea de qué métodos serán los más adecuados para aplicar en el análisis inferencial. También, como se ya se enunció anteriormente, del análisis exploratorio pueden surgir nuevas hipótesis de interés que serán evaluadas en el posterior análisis inferencial.

3. ELEMENTOS BÁSICOS DE ESTADÍSTICA INFERENCIAL

El propósito de un estudio estadístico generalmente es extraer conclusiones sobre alguna variable de interés de una determinada población que está en estudio. Dado que la población en general no se puede observar en forma exhaustiva (la población puede ser muy grande o infinita o bien la observación de la población puede implicar la destrucción de sus elementos por lo cual tampoco es posible observarla en forma completa), es necesario utilizar una muestra mediante la cual se obtendrán las conclusiones referidas a la población.

Las inferencias estadísticas se refieren a los métodos mediante los cuales buscamos seleccionar una muestra aleatoria de una población, pretendiendo sacar conclusiones, sobre ésta, con las observaciones muestrales. Estas conclusiones pueden referirse a obtener aproximaciones de valores poblacionales o bien determinar si es factible que el parámetro desconocido pueda asumirse igual a un valor teórico predeterminado. En el primer caso se refiere a la estimación de parámetros y el segundo caso a pruebas de hipótesis estadísticas.

En este apartado se desarrollan los lineamientos generales de una prueba de hipótesis estadística sobre ejemplos presentados previamente.

3.1. Hipótesis

En toda prueba o contraste estadístico se pone en competencia dos hipótesis implícitas sobre la población en estudio. Estas hipótesis se denominan Hipótesis Nula e Hipótesis Alternativa. La hipótesis nula, denotada por H0, es la creencia convencional o status quo sobre una población, mientras que la hipótesis alternativa, denotada por H1, es una alternativa a la hipótesis nula, generalmente es el cambio en la población que el investigador está buscando.

Ejemplo 1:

Este ejemplo proviene de una investigación realizada por la Dra. Zulema Solana, Profesora Titular de la Facultad de Humanidades y Artes de la Universidad nacional de Rosario. Sólo se utiliza en este trabajo una parte de la información relevada.

Las estructuras del tipo

"El hombre se hizo alcanzar el diario"

es una estructura causativa con sujeto indeterminado que es de adquisición tardía. Se consideran tardías las estructuras que los niños adquieren después de los siete años. En este caso, según algunos autores no la adquieren hasta los 11 o 12 años.

Por lo tanto, si preguntamos ¿quién alcanza el diario al hombre?, los niños de menos de 10 años no podrán dar la respuesta esperada.

Se tomaron dos muestras aleatorias de 45 niños cada una. Una muestra compuesta por niños entre 8 y 9 años y una segunda muestra conformada por niños de edades entre 11 y 12 años, las muestras se dispusieron en dos aulas donde se escribió en el pizarrón la oración enunciada. A cada niño se le entregó impresa la pregunta

¿Quién alcanza el diario al hombre?

y se le indicó que respondiera la misma en forma escrita, registrando en cada caso si comprendió o no la estructura causativa escrita en el pizarrón. Para cada niño se registra Sí/No según el resultado favorable o no de su respuesta. La unidad experimental es el alumno y las variables relevadas en la base que se adjunta en imagen de la Figura 1 son:

- ALUMNO: Identificación del alumno
- Edad (Cuantitativa)
- Sexo (Cualitativa)
- Respuesta (Cualitativa)

	A	B	C	D
1	ALUMNO	EDAD	SEXO	RESPUESTA
2	1	12	F	CORRECTA
3	2	11	M	CORRECTA
4	3	8	F	INCORRECTA
5	4	8	M	INCORRECTA
6	5	9	M	CORRECTA

Figura 3.1: Esquema de planilla de datos del ejemplo 1.

Las hipótesis estadísticas en este caso son:

H0) La probabilidad de comprensión en niños de 8 y 9 años es similar que en los de 11 y 12 años.

H1) La probabilidad de comprensión en niños de 8 y 9 años es menor que en los de 11 y 12 años.

3.2. Tipos de errores

Para decidir cuál de estas teorías parece más razonable, recolectamos los datos que son los que nos brindan la información, los analizamos y nos preguntamos: ¿estos datos son más probables de ser obtenidos cuando H0 es cierta o si H1 es la correcta?

Basados en la evidencia que muestran los datos se decide si la hipótesis nula se rechaza, en cuyo caso se dice que se hallaron diferencias significativas, o si no alcanza la evidencia para descartarla. Por lo tanto, al tomar esta decisión puede ser que cometamos dos tipos de errores:

- Error de tipo I, (E I): consiste en rechazar H_0 siendo esta correcta. La probabilidad de cometer (E I) se denomina Nivel de significación del test y se lo denota con α, P(E I) = α. Este valor es la probabilidad de que la técnica de inferencia estadística indique que la hipótesis nula es falsa cuando en realidad es cierta, esto es, encontrar significación estadística cuando en realidad no existe (Falso positivo).

- Error de tipo II, (E II): consiste en no rechazar la H_0 siendo ella falsa. A la probabilidad de cometer un (E II) se lo denota con β, P(E II) = β. Este valor es la probabilidad de que la técnica de inferencia estadística indique que la hipótesis nula es cierta cuando en realidad es falsa, esto es, no encontrar significación estadística cuando en realidad existe (Falso negativo).

En el ejemplo planteado, un error de tipo I significaría decir que la probabilidad de comprensión de la estructura en estudio es mayor en los niños mayores de 11 años cuando en realidad en los dos grupos es la misma. Un error de tipo II significaría decir que la probabilidad de comprensión es similar en las dos edades cuando en realidad es superior en los niños mayores de 11 años.

Otra medida de probabilidad vinculada a la prueba inferencial es el nivel de potencia del test. La potencia de una prueba de hipótesis estadística es la probabilidad de rechazar la hipótesis nula cuando realmente es falsa, esto es, la capacidad de detectar diferencias significativas reales. Esta probabilidad es $\Pi=1-\beta$.

Es verdad que el valor de α es el que se establece para trabajar con un nivel aceptable de significación pero la potencia del test mide la capacidad de detectar diferencias cuando existen verdaderamente, es decir, determina el éxito en la búsqueda de diferencias reales o existentes. Esta situación sugiere entonces establecer no solo el nivel de α sino también el valor de β. Sin embargo esto no es posible ya que ambas probabilidades de error están inversamente relacionadas, cuando una probabilidad de error aumenta la otra disminuye y viceversa. Entonces, ¿cómo puede el estadístico aumentar la potencia del test para un determinado nivel de significación estipulado α?

La potencia del test depende de:

a) La magnitud de la diferencia o el efecto que se busca detectar. Está claro que si el efecto (por ejemplo la diferencia entre las medias de dos grupos) es grande será más probable hallar diferencias significativas.

b) El tamaño de la muestra. Al aumentar el tamaño de la muestra se incrementará la potencia del test. De esta manera, al trabajar con más observaciones se observa que efectos cada vez más chicos se tornan significativos. Este es un aspecto muy importante ya que el investigador debe tener claro que trabajar con muestras pequeñas conduce a la insensibilidad del test pero trabajar con muestras extremadamente grandes conduce a una extrema sensibilidad del test, esto es, detectar cualquier diferencia ínfima como significativa. En este punto se debe tener clara la diferencia entre diferencias importantes y diferencias significativas.

3.3. Regla de decisión. El valor de probabilidad asociada.

El procedimiento consiste en definir una estadística para el test, esto es, una función de los valores observados en la muestra que no involucra parámetros desconocidos cuya distribución, asumiendo que la hipótesis nula es verdadera, se conoce. La distribución de la estadística del test bajo la hipótesis nula permite calcular la probabilidad de hallar una diferencia como la observada en esta muestra o aún más extrema, asumiendo que es cierto lo postulado en H0. Esta probabilidad se la llama p-value o probabilidad asociada e indica cuán probables son los datos de provenir de una población en la cual la hipótesis nula se cumple. Este valor es comparado con el nivel de significación que generalmente es 0.05 (5%) o 0.01 (1%). Si este valor es inferior al nivel de significación con el que se ha decidido trabajar, la hipótesis nula es rechazada. En caso contrario, no habría evidencia para asumir que H0 es falsa, por lo tanto la hipótesis sustentada por los datos sería H0.

Para comprender este concepto se desarrolla el ejemplo mencionado considerando dos situaciones diferentes que se podrían dar en los datos:

Situación 1:

Tabla 3.1: Situación en la que los datos sustentan H1.

	8-9	**11-12**
Comprende	12	36
No comprende	33	9
TOTAL	**45**	**45**

Situación 2:

Tabla 3.2: Situación en la que los datos sustentan H0.

	8-9	**11-12**
Comprende	14	15
No comprende	31	30
TOTAL	**45**	**45**

En la primera situación (Tabla 3.1), los datos indican que el porcentaje de niños que responden correctamente es del 26% para los menores y un 80% para los mayores de 11 años. Mientras que en el segundo caso (Tabla 3.2) la muestra evidencia similares porcentajes de respuestas correctas en ambos grupos etarios, 31% y 33% respectivamente.

En una prueba de hipótesis se decide si la hipótesis nula será rechazada o no, mientras que la hipótesis alternativa es la que será admitida o aceptada en el caso que la nula se rechace. Ambas hipótesis se plantean previas a la realización del test.

Las hipótesis correspondientes serían:

H0: La probabilidad de comprensión en niños de 8 y 9 años es similar que en los de 11 y 12 años.

H1: La probabilidad de comprensión en niños de 8 y 9 años es menor que en los de 11 y 12 años.

En términos de la probabilidad de comprensión P8-9 y P11-12 para los niños de 8 - 9 años y 11 - 12 respectivamente, las hipótesis se expresan:

$$H_0) P_{8-9} = P_{11-12}$$
$$H_1) P_{8-9} < P_{11-12}$$

La primera muestra pareciera que sustenta la hipótesis alternativa H1, ya que hay una diferencia del 53% en los que responden correctamente, mientras que en la segunda situación pareciera que los datos muestrales sustentan la hipótesis nula, H0. Por lo tanto surge el siguiente interrogante:

¿A partir de qué diferencia de porcentajes podemos concluir confiados que se debe rechazar H0?

Se dice que las observaciones muestrales son estadísticamente significativas si ellas son poco probables de ser observadas bajo el supuesto de que H0 es verdadera. Si los datos son estadísticamente significativos, entonces nuestra decisión será rechazar H0.

En este ejemplo la estadística del test es:

$$z = \frac{p_1 - p_2}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

donde p1 es la proporción de niños que comprenden en el primer grupo, p2 es la proporción de niños que comprenden en el segundo grupo, n1 y n2 son los tamaños de las dos muestras respectivamente. La distribución de esta estadística, bajo el supuesto que la hipótesis nula es cierta, es Normal estándar.

¿Cuál es nuestra decisión en cada caso?

- En el primer caso presentado, la estadística del test es Z=-6 y tiene una probabilidad asociada (p-value) igual a 0.0000. Esto significa que la probabilidad de obtener una diferencia de proporciones como la observada en el caso 1, o más extrema, cuando en realidad no existe diferencias entre las edades (H_0 verdadera) es muy baja (0.000<0.01 ó 0.05) y por lo tanto SE RECHAZA H_0, evidenciando que la probabilidad de comprensión es superior en los niños mayores.
- En el segundo caso, el valor de Z es -0.2256 con una probabilidad asociada de 0.4107. Esto significa que la probabilidad de obtener una diferencia de proporciones como la observada en el caso 2, o más extrema, cuando en realidad no existe diferencias entre las proporciones de ambos grupos de edades (H_0 verdadera) es superior al nivel de significación (0.4107> 0.05) y por lo tanto NO SE RECHAZA H_0, lo que indica que no hay evidencia muestral para pensar que la comprensión de este tipo de estructura difiere entre los 8 y 11 años de edad.

En general, cualquier prueba de hipótesis puede plantearse en los siguientes pasos:

- ✓ Enunciar H_0 y H_1.

La hipótesis de investigación debe traducirse en la hipótesis estadística, es decir, la hipótesis en términos estadísticos. Estas hipótesis estadísticas se formulan generalmente en término de parámetros poblacionales o de distribuciones poblacionales en los casos de pruebas de bondad de ajuste. Es importante destacar que una hipótesis de investigación puede ser traducida en distintas hipótesis estadísticas según las características de las variables en estudio y la técnica a utilizar.

- ✓ Especificar el nivel de significación del test.

Esto es probabilidad de error de Tipo I, la probabilidad de rechazar H0 cuando en realidad es verdadera.

- ✓ Definir la estadística del test.

Esto es una función de los valores de la muestra que no depende de parámetros desconocidos y cuya distribución, asumiendo que H0 se cumple, es conocida.

- ✓ Calcular el valor de probabilidad asociada al valor hallado en la estadística del test.

Una vez que se ha calculado el valor de la estadística del test, se debe decidir si ese valor obtenido sustenta la hipótesis nula o la hipótesis alternativa. Se nos plantea la siguiente pregunta ¿cuán

probable es obtener este valor de la estadística del test hallado o más extremo asumiendo que H0 es verdadera? Se calcula la probabilidad asociada al valor de la estadística del test.

- ✓ Tomar una decisión y realizar conclusiones.

Si es factible (probable) obtener un valor de la estadística del test como el obtenido con nuestros datos, bajo el supuesto que H0 es cierta, entonces no tenemos evidencia para rechazarla. En caso contrario los datos muestran cierta evidencia para pensar que H0 debe ser descartada. El valor de esta probabilidad es comparado con el nivel de significación del test, si la probabilidad asociada es menor que el nivel de significación se rechaza la hipótesis nula y en caso contrario ésta es aceptada. Una vez tomada la decisión se debe realizar la conclusión en términos del problema que se está estudiando, generalmente el experto en el área es quien formula la misma.

4. SELECCIÓN DE LA TÉCNICA ESTADÍSTICA.

Otro aspecto, referido al análisis estadístico en una investigación, es la elección de la o las técnicas estadísticas a aplicar. Una consideración a tener en cuenta al momento de buscar o elegir la técnica estadística que responde al objetivo de análisis es la existencia de variables dependientes en nuestros datos. Una variable dependiente es aquella variable respuesta que se pretende explicar por medio de las variables independientes o explicativas. Por ejemplo, si queremos determinar cómo las horas de estudio semanales de un alumno explican el rendimiento académico tendremos que la variable independiente o explicativa es la cantidad de horas semanales que el alumno estudia y la variable dependiente o respuesta es el rendimiento académico que podría estar medido como el número de materias aprobadas por año o el promedio académico.

Entonces una primera clasificación de las técnicas podría ser en técnicas para análisis de dependencia o interrelación. En las primeras existe una variable o un conjunto de variables dependientes que son explicadas por una o varias variables independientes. En las segundas, todas las variables se encuentran a un mismo nivel en este aspecto, y el objetivo es estudiar las interrelaciones entre ellas.

Otro aspecto importante a tener en cuenta es el número de variables que se analizarán en forma simultánea. Esto lleva a la elección de una técnica multivariada frente a una univariada. Algunos autores definen al Análisis Multivariado simplemente como la aplicación de métodos que permiten trabajar con un número “grande” de medidas realizadas sobre cada unidad experimental en una o varias muestras simultáneamente. Otros hacen referencia a la cantidad de variables dependientes que se consideran a la vez.

La elección de la técnica tiene que ver con el objetivo del análisis y el tipo de variables con las que se va a trabajar, es decir, según la forma en que se registren las variables se podrá utilizar una u otra técnica.

5. CONCLUSIONES

El incremento en el uso de la estadística en las investigaciones ha sido evidente en las últimas décadas. Si bien este surgimiento de las técnicas estadísticas en las investigaciones se vincula directamente con el crecimiento de la informática, es claro que en gran medida también se debe a la necesidad de los investigadores de adquirir nuevos procedimientos de análisis de datos que respondieran a sus objetivos. Esto ha sido evidente en diversas áreas disciplinares. No obstante, el crecimiento informático ha contribuido en la utilización de técnicas estadísticas debido a la aparición de software estadísticos de fácil acceso y formato amigable para el usuario.

Referencias

Aliaga, Marta, y Gunderson B. 2002 “Interactive Statistics”, Prentice Hall.

Barbona, Ivana, 2015 Comparación de Métodos de Clasificación aplicados a textos Científicos y No Científicos. Revista INFOSUR. Número 7.

Beltrán, Celina 2008 Información lingüística y técnicas estadísticas en el análisis automático de textos, Tesis de Doctorado bajo la dirección de Gabriel Bès, Facultad de Humanidades y Artes, Universidad de Rosario.

Beltrán, C. 2009 Elementos de Inferencia Estadística aplicada a la Lingüística, (Capítulo 8). Título del libro: “La interlengua de los aprendientes del español como L2. Aportes de la Lingüística Informática. Centro de estudios de adquisición del lenguaje. Facultad de Humanidades y Artes. Grupo Infosur. Ediciones Juglaría.

Beltrán, C. 2011 Aplicación del análisis de regresión logística multinomial en la clasificación de textos académicos: Biometría, Filosofía y Lingüística informática Revista INFOSUR. Número 5.

Beltrán, C. 2012 Introducción al Análisis Estadístico en la Investigación. Ediciones Juglaría.

Beltrán, C. 2012 Aplicación de redes neuronales artificiales en la clasificación de textos académicos según disciplina: Biometría, Filosofía y Lingüística informática. Revista de Epistemología y Ciencias Humanas. Nro. 4

Beltrán, C. 2012 Árboles de clasificación y su comparación con análisis de regresión logística aplicado a la clasificación de textos académicos. Revista INFOSUR. Número 6.

Beltrán, C. 2015 Técnicas estadísticas de clasificación. Una aplicación en la clasificación de textos según el género: Textos Científicos y Textos No Científicos. Revista INFOSUR. Número 7.

Cuadras, C.M. 2008 NUEVOS MÉTODOS DE ANÁLISIS MULTIVARIANTE. CMC Editions. Barcelona, España.

Giubileo, M. G.; Bisaro, V.; Trevizan, A; Dalla Marta, N.; Cosolito, P.; Beltrán, C. 2005. Elementos de estadística descriptiva e inferencial. Ediciones Juglaría.

Giubileo, M. G.; Bisaro, V.; Trevizan, A; Cosolito, P.; Beltrán, C. 2006. Introducción al diseño y análisis de experimentos. Ediciones Juglaría.

Hastie, T., Tibshirani, R., Friedman, J. 2009. The Elements of Statistical Learning. Data Mining, Inference, and Predicion. Springer Series in Statistics.

Hosmer, D., Lemeshow, S., Sturdivant, R. 2013. Applied Logistic Regression. John Wiley & Sons.

Khattre R. y Naik D. 2000 Multivariate Data Reduction and Discriminatio with SAS Software. SAS Institute Inc. Cary, NC. USA

Manning, Christopher D. y Schütze, Hinrich. 1999. Foundations of Statistical Natural Language Processing. Cambridge Mass, The MIT Press.

Montgomery, Douglas y Peck, Elizabeth. 1992. Introduction to Linear Regression Analysis. John Wiley & Sons, Inc.

Solana, Zulema. 1999 Un estudio cognitivo del proceso de adquisición del lenguaje. Centro de Estudios de Adquisición del Lenguaje. UNR. Ediciones Juglaría.

Trevizan, A.; Beltrán, C.; Cosolito, P. 2009 Variables que condicionan la deserción y retención durante el trayecto universitario de alumnos de la carrera de Ingeniería Agronómica de la Universidad Nacional de Rosario". Revista de Epistemología y Ciencias Humanas. Nro. 1 Año 2009.

Witten, I., Frank, E. 2005. Data Mining. Practical Machine Learning Tools and Techniques. Elsevier.