

Apuntes estadísticos sobre el procesamiento de datos en un trabajo de investigación cuantitativo de orden organizacional

Statistical notes on data processing in a quantitative research work of organizational order

Henry Hurtado B¹

Ricardo Gómez N.²

1 Licenciado en Matemáticas, Magíster en Ciencias de la Organización, investigador grupo llama, Escuela de Ciencias Administrativas, Contables, Económicas y de Negocios, Universidad Nacional Abierta y a Distancia, UNAD.

2 Matemático, Magíster en La enseñanza de las Ciencias Exactas y Naturales, profesor de la Escuela de Ciencias de la Educación, Universidad Nacional Abierta y a Distancia, UNAD.

Resumen

El diseño de la investigación parte de una pregunta adecuadamente formulada y delimitada que además corresponda a un problema real, no a un problema supuesto que el investigador ha creado para realizar un estudio. Seguidamente se formulan los objetivos, los cuales deben estar en concordancia con el título del estudio y la pregunta de investigación, los objetivos específicos deben derivarse del objetivo general y ser medibles. El diseño incluye la tipología del estudio, descriptiva, cuantitativa, cualitativa, mixta, correlacional, experimental, no experimental, cuasi experimental, histórica, documental, etc.

La recolección de los datos y su respectivo procesamiento debe de ser acorde con la tipología del estudio definida en la metodología, es decir, si se definió que el estudio es descriptivo, se esperaría encontrar tablas de frecuencias, de contingencia, medidas de centralización, de dispersión, el coeficiente de desviación. Si el estudio es correlacional, se espera encontrar en el procesamiento, una prueba de correlación, unas hipótesis nula y alterna, con sus respectivas pruebas de hipótesis. Si se ha definido que el estudio es experimental, se espera que se haya definido una variable de control y las condiciones para ejecutar el experimento, etc.

El estudio analizó 50 trabajos de grado en el campo de la administración y sus áreas afines, de tipología cuantitativa, se seleccionaron al azar y se observó la metodología planteada por los autores, se estudió el procesamiento estadístico de datos que usaron los autores.

Este documento de trabajo, como su nombre lo indica, ofrece apuntes estadísticos para el procesamiento de los datos, en estudios cuantitativos, que sean consistentes con la tipología de la investigación cuantitativa.

Palabras clave:

Investigación cuantitativa, pruebas estadísticas.

Abstract

The design of investigation begin with a key question, formulated and delimited correctly, is a real problem, not a invention problem developed by the investigator. Nextly the objectives are formulated, they must be according with the title of study and according with the key question, the secondary objectives are derivates from de main objective, and be measures. The design include the class of study, such as descriptive, quantitative, qualitative, mix, correlational, experimental, non-experimental, quasi experimental historical, documental or others.

The data recollections and its process must be according with the class of study, if study is descriptive, may be are frequency tables, contingency tables, central measures, dispersion measures and deviation coefficient. If the study is correlational may be are a correlation test, null and altern hypothesis with its mechanisms of test. If the study is experimental, may be are a control variable and the conditions to make the experiment, etc.

The study analyzed 50 grades papers in the administration fields and similar areas, with quantitative typology, the elements were selected with the random way and in the methodology was studied, as well as, its statically data processing used by the authors.

This paper as its title said, offers statistical notes to process data in quantitative studies according with the quantitative investigation.

Keywords:

Quantitative investigation, statistical test.

Introducción

Los estudios de tipo cuantitativo requieren precisión en las pruebas matemáticas o estadísticas que se vayan a aplicar para procesar los datos. No toda prueba es aplicable a todo tipo de estudio, ni todo tipo de estudio requiere un compendio de pruebas. Las pruebas a utilizar deben de aportar información para que el investigador responda la pregunta fundamental y cumpla con los objetivos propuestos.

A manera de ejemplo, un área de una curva salarial, se podría calcular con una técnica de integración (potencia, sustitución, partes) o con una fórmula de integración, no tendría sentido, calcular el área de la curva con una medida de centralización estadística; una correlación de dos variables, se calcula con una prueba estadística de correlación y no con un caso de factorización, así sucesivamente. Se quiere expresar que las técnicas matemáticas y/o estadísticas, no son generales, sino de acuerdo al tipo de datos y a la clase de estudio.

En términos coloquiales, usar cualquier técnica de forma general, en cualquier estudio es un sinsentido, como pretender medir la altura de una persona con un tensiómetro, o el peso con un decámetro.

Habiéndose planteado ya que las pruebas estadísticas deben ajustarse a la necesidad de lo que se busca medir, es importante, considerar que una prueba estadística tampoco aplica a cualquier clase de distribución de probabilidad de datos. Por un lado, determinar el tamaño de la población, indicaría si es necesario calcular un tamaño muestral o no, y si hay una muestra, tendría sentido conocer el error muestral y el error estándar de la muestra.

De otro lado, tampoco hay una "fórmula" para calcular el tamaño muestral, su cálculo depende si el estudio tiene una población finita o infinita, y si el estudio es, además, cuantitativo o cualitativo. Adicionalmente, si se calcula un tamaño muestral, hay que determinar ¿cómo se van a elegir los elementos de la población que conforman la muestra, de forma aleatoria, sistemática, por conglomerados, por ejemplo?

Al calcular una correlación, o probar una hipótesis, se observa la aplicación de determinadas pruebas, sin previamente identificar, si la distribución de datos es simétrica (normal) o asimétrica. La simetría de los datos conduce a pruebas paramétricas y su asimetría a pruebas no paramétricas. Entonces, ¿usar una determinada prueba, dígame Pearson para calcular una correlación sin saber si los datos son simétricos o asimétricos tiene sentido? Por supuesto que no.

Estos son los apuntes estadísticos que se pretende discutir en este documento de trabajo.

Marco teórico

En forma general, la estadística puede clasificarse en tres grupos: estadística descriptiva, estadística inferencial o estadística de probabilidades. La estadística descriptiva como su nombre lo indica muestra las características que acompañan a un objeto o acontecimiento en estudio, para ello, se cuenta con las medidas de centralización (rango, media, mediana y moda), debe considerarse que estas medidas tienen cierta diferencia tanto si los datos son agrupados o no agrupados, así mismo consta de métodos y procedimientos para organizar, resumir y presentar los datos transformados en información (Lind, 2015). Esta clase de estadística es fundamental en estudios descriptivos y puede usarse en los otros tipos de estudios, a manera informativa. Entre los métodos para presentar los datos en forma de información se cuenta con las tablas de frecuencias, de contingencia, los gráficos, entre ellos los polígonos y los histogramas.

La importancia de la estadística descriptiva está en su poder transformador de los datos en información útil y significativa que describa un objeto o suceso en estudio y su utilidad para la toma de decisiones que pueden ser variadas: correcciones, ajustes, inversiones, programas, etc.

La estadística de probabilidades es importante en estudios de prospectiva, análisis del presente para anticiparse al futuro, se fundamenta en una buena cantidad de datos previos, a los cuales se les identifica un patrón de compor-

tamiento que permita intuir el futuro, de una manera probabilística, es decir, no hay certeza, en este sentido, en la prospectiva se pueden afirmar muchos eventos posibles, basados en probabilidades, que finalmente resulten o no resulten verdaderos. El tiempo es el que juzga si lo que se predijo basado en probabilidades ocurrió o no ocurrió.

La estadística de probabilidades se fundamenta en variables discretas, números enteros no negativos, y en variables continuas, números racionales de la forma p/q con q mayor a 0.

El teorema de Bayes es un método de probabilidad útil para definir eventos bajo probabilidad condicionada, e.g., calcular la probabilidad de que un elemento vendido haya sido devuelto en determinada sucursal. Entre los métodos de probabilidad con distribuciones discretas está la distribución binomial, la de Poisson y la hipergeométrica. Las tres distribuciones pueden resolverse con la noción de factorial y específicamente con permutaciones o combinaciones. La distribución binomial también puede resolverse con la factorización de binomios con exponente n , requiere conocimiento algebraico, no obstante, es un conocimiento al alcance de cualquier investigador que con una explicación es suficiente para su entendimiento (de menos de media hora). Con estos modelos puede calcularse el valor probable del éxito de un nuevo producto, de expansión del mercado o la probabilidad de fidelización de un cliente.

Entre los métodos de distribuciones de probabilidad continua, el más usado e importante es la familia de distribuciones de probabilidad normal, que es simétrica, asintótica y con forma de campana. Simétrica significa que tanto la media como la mediana y la moda están en un mismo punto que divide la gráfica en dos partes iguales, con media cero y desviación estándar de uno (1), esto significa, que a la derecha del punto donde se hallan las medidas de centralización corresponde al 50% de los datos y a la izquierda el otro 50%, si se trabaja con un margen de error, por ejemplo del 5%, entonces el 47,5% de los datos están a la derecha e izquierda del punto medio, y el margen de error, se distribuye en partes iguales en ambas colas, 2,5% en la cola derecha y 2,5% en la cola izquierda que sumado cada uno al 47,5% abarca el 50% de los datos en cada lado.

La distribución de probabilidad normal con media cero y desviación estándar de uno, se denomina distribución de probabilidad normal estándar, se representa con la tabla Z y se usa para calcular la probabilidad de ocurrencia de un evento bajo la curva normal. La tabla de probabilidad normal, le asigna un valor constante Z, a todas las probabilidades comprendidas entre 0 y 0,499. Si se desconoce la desviación estándar se usa la tabla de distribución T de Student que comprende los grados de libertad, los intervalos de confianza, el nivel de significancia de una prueba de una cola o de dos colas.

La inferencia estadística explica el comportamiento de una población a partir de lo que se observa en la muestra, un método inductivo de investigación. Hacen parte de la inferencia estadística temas como la correlación, las pruebas de hipótesis, el análisis de varianza y los análisis de regresión múltiple.

El análisis de correlación es una prueba que mide la potencia de la asociación entre dos variables, el coeficiente de correlación recibe la denominación de r, y es un valor comprendido entre -1 y 1. Si r es mayor que cero, se trata de una correlación positiva o directa entre las variables, si r es menor que cero, se trata de una correlación negativa o inversa entre las variables. Si r es igual a cero, se afirma que no hay correlación entre las variables. El valor de r se puede aproximar a uno, a cero o a menos 1, el autor dirá entonces que la correlación entre las variables es positiva alta, positiva baja, negativa alta o negativa baja.

La prueba de hipótesis de una muestra o de dos muestras, consiste en determinar qué hipótesis, entre las hipótesis nula y alterna, del estudio se acepta; el procedimiento es definir el nivel de confianza, calcular el estadístico p con la tabla de distribución normal estándar, y luego calcular la relación entre el error muestral y la desviación típica de la muestra, a este valor se le denomina Z. si el valor de Z está dentro de los parámetros establecidos por el estadístico p, se halla en la zona de aceptación, por tanto, se aprueba la hipótesis nula, de lo contrario se acepta la hipótesis alterna. El error tipo I consiste en rechazar la hipótesis nula siendo la hipótesis nula verdadera y el error tipo II consiste en aceptar la hipótesis nula, siendo la hipótesis nula falsa.

El análisis de varianza conocido como ANOVA, compara dos varianzas poblacionales que deben de seguir la distribución normal, tener desviaciones estándar iguales y ser independientes; se usa la distribución F, la cual es continua, no negativa, con sesgo positivo y asintótica. La herramienta ANOVA permite resolver situaciones del tipo, se tienen dos formas de pago en un almacén, con un funcionario o con un dispositivo electrónico, con una muestra, un nivel de significancia, α , una media, \bar{x} , y una desviación típica, s , ¿cuál de los dos sistemas es preferido por los clientes?

Los análisis de regresión múltiple, permite predecir el comportamiento de una variable dependiente con relación a dos o más variables independientes, de ahí la denominación de múltiple, en el análisis hay que calcular el error estándar y el coeficiente de determinación múltiple para conformar la ecuación de regresión múltiple con la que se predice el valor de la variable dependiente. Obsérvese que es común el uso de términos como probabilidad o predicción, no se habla de certeza, sino de probable.

Concluyendo la teoría estadística, si se trata de estudios descriptivos, las pruebas de la estadística descriptiva pueden usarse en cualquier clase de estudios, pero si se trata de la estadística inferencial, hay que, en primer lugar, determinar de qué clase de distribución de probabilidad de datos se trata, simétrica o normal o en su defecto, asimétrica.

Para identificar si una distribución es simétrica o asimétrica, la forma más sencilla es calcular la media, mediana y moda de la población y si son iguales, se trata de una distribución normal, graficar los datos y observar si se forma una gráfica en forma de campana lo que corresponde a una distribución normal. Otros métodos más avanzados, dependen del tamaño de la muestra, si la muestra es menor o igual a treinta elementos, puede usarse la prueba de Shapiro para calcular la asimetría o asimetría, de lo contrario, con muestras mayores a treinta elementos puede usarse la prueba de Kolmogorov – Smirnov.

Identificándose la simetría o asimetría de los datos, se procede a determinar si se usan pruebas paramétricas que corresponden a distribuciones de probabilidad normales o pruebas no paramétricas para distribuciones no paramétricas.

En síntesis, para distribuciones de probabilidad normales, úsese pruebas paramétricas como Pearson para las correlaciones, T-Student, Chi Cuadrado o Anova para probar hipótesis; para pruebas no paramétricas las distribuciones no paramétricas, Spearman para correlaciones o la U de Mann Whitney para pruebas de hipótesis, Kruskal Wallis, equivalente a la ANOVA, entre otras.

Resultados

Se revisaron 50 trabajos de grado de egresados de programas de Administración de Empresas, en la parte cuantitativa, todos muestran información descriptiva, usando tablas de frecuencias, con clases cualitativas, ocasionalmente se encontraron tablas de frecuencias para datos cuantitativos que utilizaran intervalos de clase, marca de clase o pruebas de estadística descriptiva para datos agrupados. El tipo de gráfico más común es el gráfico de barras y el gráfico circular. Fueron pocos los hallazgos de trabajos que usaran tablas de contingencia.

No se encontraron trabajos que desarrollaran aplicaciones de probabilidad para variables discretas o continuas, ni el uso de permutaciones o combinaciones.

Se encontró que se asume que las distribuciones de probabilidad de datos son continuas, no hay pruebas para determinar la simetría o asimetría de las poblaciones de datos, en tal sentido, cuando se trata de correlaciones, todos usaron la prueba de Pearson, sin saber si los datos eran normales o asimétricos y que lo más recomendable sería usar otro tipo de prueba no paramétrica como Spearman, en el caso de distribuciones asimétricas.

No se hallaron trabajos donde la tipología fuera correlacional y en la metodología se definieran hipótesis, no se encontró trabajos con pruebas

estadísticas para las hipótesis. Algunos trabajos tienen hipótesis, más de dos hipótesis y dícese probarse por los datos arrojados por las tablas de frecuencias, es decir, de forma intuitiva.

Los trabajos suelen realizarse en grupos de dos o más estudiantes, a veces, hasta cinco estudiantes, no hay preferencia por los trabajos individuales.

Se buscó el curriculum vitae de los directores de trabajo de grado, y se observó que para estudios cuantitativos tienen una formación matemática baja o media, algunos son ingenieros y ninguno era Estadístico, Matemático, Licenciado en Matemáticas o afines.

Los trabajos de grado en el área se orientan a implementar modelos, caracterizar poblaciones, realizar diagnósticos, temas que resultan en estudios descriptivos, algunos trabajos se titulan impacto lo que conduce a estudios de correlación entre variables e incluso al planteamiento de hipótesis. No se encontró temas de estudio que ameriten el uso de análisis de varianza o regresiones múltiples.

Los datos se suelen recolectar con la aplicación de encuestas, se observó que se les mide la confiabilidad y en algunos casos la validez del instrumento. El software de procesamiento más empleado es Excel, se usan pocos programas estadísticos como SPSS u otros semejantes.

Discusión y conclusiones

El uso de tablas de frecuencia es muy importante en la descripción de datos, es una forma útil de comprender la información y tomar decisiones por parte del investigador. No obstante, los trabajos pueden profundizar en las técnicas de la estadística descriptiva, por ejemplo, es importante usar las tablas de contingencia para cruzar dos o más variables o las medidas de centralización y de dispersión para luego calcular el coeficiente de dispersión, la relación entre la desviación estándar y la media, un valor entre 0 y 1 que muestra la

heterogeneidad u homogeneidad de los datos, esta clase de estadística es importante para observar el comportamiento de los datos, si es normal, tiene determinado sesgo o los datos son tan homogéneos o heterogéneos que su gráfico muestra la curtosis (alargamiento, achatamiento o normalidad de los datos). El uso de polígonos es importante para destacar el punto medio de los intervalos de clase. Es importante incentivar la creación de los intervalos de confianza con los datos en estudio y de acuerdo al nivel de confianza que determine el investigador.

El uso de la teoría de probabilidad, la teoría de conjuntos y el uso del factorial son instrumentos fundamentales para los estudios de prospectiva, no es comprensible el planteamiento de escenarios de prospectiva sin el uso de la probabilidad, por ejemplo, el Teorema de Bayes es fundamental cuando se trata de probabilidad condicional,

Un error muy frecuente es aplicar pruebas paramétricas en distribuciones de datos sin profundizar la simetría o asimetría de los datos, lo que conduce a un error considerable cuando se trata de procesar datos, por tanto, es importante que los trabajos de grado muestren las pruebas de normalidad de los datos, desde las más sencillas, como un gráfico de barras o más complejas como las pruebas de Shapiro o Kolmogorov que están disponibles en programas de estadística o incluso están disponibles en las funciones de Excel.

Los estudios que se orientan a analizar el impacto de determinadas variables son una excelente oportunidad para realizar estudios de correlación y usar las pruebas de Pearson para los datos simétricos o de Spearman para los datos asimétricos. Los estudios de impacto no se resuelven con datos descriptivos, estos datos solo dan indicios de la correlación, pero no son concluyentes.

Es importante que en los estudios de impacto o de correlación, se calcule el valor r para determinar la clase y grado de correlación, usar los gráficos de dispersión de puntos para observar si hay correlación o no entre las variables. Es importante que se planteen hipótesis y probar las hipótesis estadísticamente con las pruebas paramétricas T-Student o Chi Cuadrado o

con las pruebas no paramétricas como U de Mann Whitney según sea el caso. Los jóvenes investigadores deben entender cómo se grafica en la curva normal los estadísticos Z, T o F para crear las áreas de aceptación o de rechazo que permiten probar las hipótesis. Una característica en algunos estudios es definir más de dos hipótesis, sin determinar cuál es la nula y cuál es la alterna, y sin que los enunciados de las hipótesis sean excluyentes entre sí, esta situación imposibilita la prueba estadística de hipótesis.

Igualmente, es importante el uso de sistemas de datos multivariados y superar los análisis bivariados, los datos en la realidad de los negocios, suelen ser multivariados.

Es necesario avanzar un poco más en la frontera del procesamiento estadístico de datos y explorar novedosas técnicas como el uso de ecuaciones estructuradas en el análisis de datos, en investigaciones empíricas, propias de las ciencias sociales, en las cuales están las investigaciones de la Organización y la gerencia.

Es indispensable y necesario incentivar la creatividad en los investigadores, la investigación no puede observarse como un algoritmo, necesita de la invención y la chispa de la creatividad para que tenga sentido y no sea un resultado mecanicista sin mayores aportes.

Por último, se sugiere que los investigadores no familiarizados con la Estadística y la Matemática, reciban seminarios de estadística inferencial permanentemente y de esta manera logren maximizar los resultados y la utilidad de la estadística en sus investigaciones.

Referencias

Gil, J. (2006). *Estadística e informática en la investigación descriptiva e inferencial*. Editorial Tirant.

Gómez, M. et al. (2003). Sinopsis de pruebas estadísticas no paramétricas. Cuándo usarlas. *Revista Mexicana de Pediatría*, 20 (2).

Gutiérrez, E. y Vladirimovna, O. (2014). *Probabilidad y estadística*. Grupo editorial Patria.

Flórez, E. et al. (2017). *Cómo elegir la prueba estadística adecuada*. *Estadística inferencial*.

Lind, D. et al (2020). *Estadística aplicada a los negocios y la economía*. Mc Graw Hill.

Nolberto, V. y Ponce, M. (2008). *Estadística inferencial aplicada*. Unidad de Post Grado de la Facultad de Educación de la Universidad Nacional Mayor de San Marcos.

Romero, D. et al (2018). Los modelos de ecuaciones estructuradas como herramienta de análisis estadístico en la investigación social. *Laurus, revista de educación*, 17(1).