¿A qué nos referiremos con la exploración de datos?

Como hemos mencionado anteriormente, va a ser fundamental que identifiquemos patrones, tendencias, comportamientos entre dos, tres o más grupos de variables, para a partir de ello poder realizar una predicción o poder anticipar el comportamiento futuro de una variable aleatoria de nuestro interés. Permíteme ponerte un ejemplo. Piensa en el área de compensaciones de una organización, responsable de definir un nivel de sueldo y que sea un nivel de sueldo competitivo para los empleados que desea tener laborando dentro de ella. Para definir ese nivel de sueldo es necesario revisar cómo está el entorno y revisar cómo están sueldos para puestos similares considerando experiencia, trayectoria, nivel de sueldo, geografías y zonas geográficas similares a partir de las cuales pueda revisar qué es lo que está ocurriendo ahí afuera, que no está bajo su control, para a partir de ello definir

Un nivel de sueldo que sea competitivo y que sea atractivo de acuerdo al tipo de personas que desea atraer dentro de su organización. Un segundo caso de mucho interés es, ¿qué tan probable es que tú decidas comprar un producto B dado que previamente tu compraste un producto A? Es decir, ¿cuál es tu interés de compra por determinado producto dado que en el pasado has adquirido otros?

 Eso lo podrá identificar una organización al revisar patrones de comportamiento de compra en clientes similares a lo que tu realizas precisamente también como cliente. Es decir, ¿qué tipos de compras has realizado? ¿Qué tipo de productos has adquirido? ¿Por qué tipo de productos has hecho búsquedas o has hecho consultas? Para con base en ello hacer buscar hacerte una recomendación sobre qué producto puedes adquirir.

Eso es precisamente lo que realizan empresas como Amazon. Cuando tú haces una búsqueda de un producto, automáticamente Amazon te realizará en su sitio web recomendaciones sobre productos similares que otros clientes que han hecho búsquedas sobre este tipo de productos han eventualmente adquirido.

Es decir, te está haciendo una recomendación con base en la exploración y revisión de datos para a partir de ello definir una recomendación o una tendencia sobre posible interés de compra con base en esa información.

En la lectura anexa a esta lección encontrarás una guía que te podrá apoyar en forma muy importante para aprender y utilizar esos modelos estadísticos y herramientas computacionales como la que estaremos manejando para realizar esta exploración de datos. Es decir, agrupar variables, identificar patrones, tendencias o comportamientos entre ellas, valores mínimos, valores máximos, relaciones que sean importantes para a partir de ello predecir comportamientos futuros.

Es muy importante que seas sensible a que no vas a encontrar una respuesta única o una respuesta correcta. Tendrás que estar realizando distintos análisis con distintas alternativas y a partir de ello revisarlo bajo distintos enfoques para poder irlo mejorando en forma continua. Te invito a que trabajes de manera muy detallada, muy estructurada, que lo practiques, pues esa práctica te permitirá desarrollar una mayor y mejor capacidad y, en consecuencia, poder realizar mejores análisis y poder tomar mejores decisiones. Trabájalo de manera ardua, de manera muy dura y muy estructurada y estoy seguro que encontrarás aspectos muy interesantes de beneficio para ti y

Análisis de datos

Una vez concluidas las etapas de colección y procesamiento de datos se inicia con una de las más importantes fases de una investigación: el análisis de datos. El análisis de datos es el precedente para la actividad de interpretación. La interpretación se realiza en términos de los resultados de la investigación. Esta actividad consiste en establecer inferencias sobre las relaciones entre las variables estudiadas para extraer conclusiones y recomendaciones (Kerlinger, 1982). La interpretación se realiza en dos etapas:

1. Interpretación de las relaciones entre las variables y los datos que las sustentan con fundamento en algún nivel de significancia estadística.
2. Establecer un significado más amplio de la investigación, es decir, determinar el grado de generalización de los resultados de la investigación.

El propósito es, entonces, describir el comportamiento de una variable a partir de la información correspondiente, contenida en un banco de datos. Así, la información está formada por los datos que se encuentran en una columna del banco y la principal característica de esos datos es su variabilidad.

“Analizar significa establecer categorías, ordenar, manipular y resumir los datos,” (Kerlinger, 1982, p. 96). En esta etapa del proceso de investigación se procede a racionalizar los datos colectados a fin de explicar e interpretar las posibles relaciones que expresan las variables estudiadas.

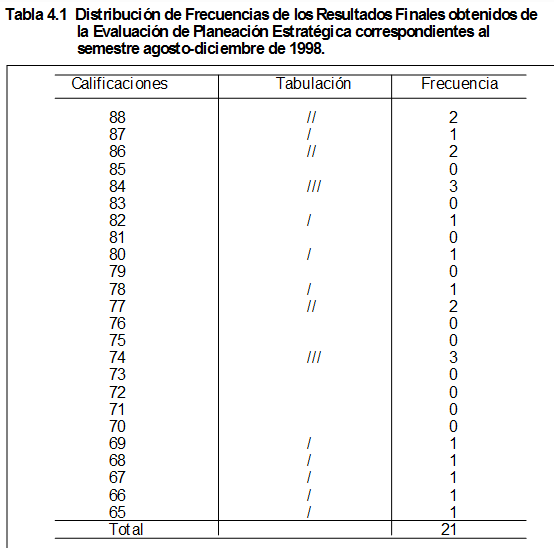
**Organización de datos**

Entre los sistemas para ordenar los datos se encuentran principalmente dos: **a) la distribución de frecuencias y b) la representación gráfica.**

**Distribución de frecuencias**

Comúnmente llamada tabla de frecuencias, se utiliza para hacer la presentación de datos provenientes de las observaciones realizadas en el estudio, estableciendo un orden mediante la división en clases y registro de la cantidad de observaciones correspondientes a cada clase. Lo anterior facilita la realización de un mejor análisis e interpretación de las características que describen y que no son evidentes en el conjunto de datos brutos o sin procesar. Una distribución de frecuencias constituye una tabla en el ámbito de investigación.

La distribución de frecuencias puede ser simple o agrupada. La distribución de frecuencias simple es una tabla que se construye con base en los siguientes datos: clase o variable (valores numéricos) en orden descendente o ascendente, tabulaciones o marcas de recuento y frecuencia.



La distribución de frecuencia agrupada se utiliza cuando el número de variables es muy grande o la variable es continua, para este caso se definen clases. En las clases se agrupan valores en intervalos de la misma amplitud y se calcula la frecuencia correspondiente.

**Interpretación de las relaciones entre las variables**

Correlación entre variables

La correlación estadística constituye una técnica estadística que nos indica si dos variables están relacionadas o no. Por ejemplo, considera que las variables son el ingreso familiar y el gasto familiar. Se sabe que los aumentos de ingresos y gastos disminuyen juntos. Por lo tanto, están relacionados en el sentido de que el cambio en cualquier variable estará acompañado por un cambio en la otra variable.

Si el cambio en una variable está acompañado de un cambio en la otra, entonces se dice que las variables están correlacionadas. Por lo tanto, podemos decir que el ingreso familiar y gastos familiares y el precio y la demanda están correlacionados.

La correlación puede decir algo acerca de la relación entre las variables. Se utiliza para entender si la relación es positiva o negativa y la fuerza de la relación.

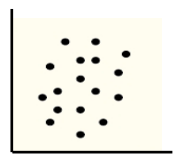
La correlación es una herramienta poderosa que brinda piezas vitales de información. La correlación estadística es medida por lo que se denomina coeficiente de correlación (r). Su valor numérico varía de 1,0 a -1,0. Nos indica la fuerza de la relación.

En general, r> 0 indica una relación positiva y r <0 indica una relación negativa, mientras que r = 0 indica que no hay relación (o que las variables son independientes y no están relacionadas). Aquí, r = 1,0 describe una correlación positiva perfecta y r = -1,0 describe una correlación negativa perfecta.

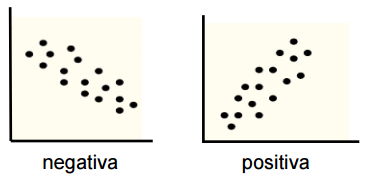
La correlación es solamente apropiada para examinar la relación entre datos cuantificables significativos (por ejemplo, la presión atmosférica o la temperatura) en vez de datos categóricos, tales como el sexo, el color favorito, etc.

El tipo de correlación se puede deducir según la forma de la nube de puntos representada en un gráfico de dispersión:

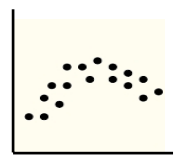
* Correlación nula: No existe ninguna relación entre las variables. Se dice que ambas son independientes.



* Correlación lineal: Existe una relación lineal negativa si al aumentar los valores de la variable independiente disminuyen los valores de la variable dependiente y relación lineal positiva si al aumentar los valores de la variable independiente aumentan los valores de la variable dependiente.



* Correlación no lineal: Existe una relación entre las variables pero no es lineal.



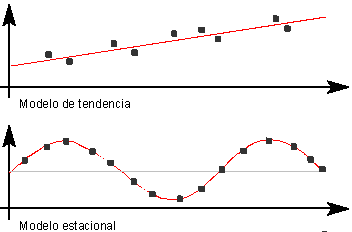
**Características de series de tiempo**

Por serie de tiempo nos referimos a datos estadísticos que se recopilan, observan o registran en intervalos de tiempo regulares (diario, semanal, semestral, anual, entre otros). El término serie de tiempo se aplica por ejemplo a datos registrados en forma periódica que muestran, por ejemplo, las ventas anuales totales de almacenes, el valor trimestral total de contratos de construcción otorgados, el valor trimestral del PIB.

**Componentes de la serie de tiempo**

Existen cuatro tipos básicos de variación, los cuales sobrepuestos o actuando en concierto, contribuyen a los cambios observados en un período de tiempo y dan a la serie su aspecto errático. Estas cuatro componentes son: Tendencia secular, variación estacional, variación cíclica y variación irregular.

1. **Tendencia secular:** La tendencia secular o tendencia a largo plazo de una serie es por lo común el resultado de factores a largo plazo. En términos intuitivos, la tendencia de una serie de tiempo caracteriza el patrón gradual y consistente de las variaciones de la propia serie, que se consideran consecuencias de fuerzas persistentes que afectan el crecimiento o la reducción de la misma, tales como: cambios en la población, en las características demográficas de la misma, cambios en los ingresos, en la salud, en el nivel de educación y tecnología. Las tendencias a largo plazo se ajustan a diversos esquemas. Algunas se mueven continuamente hacía arriba, otras declinan, y otras más permanecen igual en un cierto período o intervalo de tiempo.
2. **Variación estacional:** El componente de la serie de tiempo que representa la variabilidad en los datos debida a influencias de las estaciones, se llama componente estacional. Esta variación corresponde a los movimientos de la serie que recurren año tras año en los mismos meses (o en los mismos trimestres) del año poco más o menos con la misma intensidad. Por ejemplo: Un fabricante de albercas inflables espera poca actividad de ventas durante los meses de otoño e invierno y tiene ventas máximas en los de primavera y verano, mientras que los fabricantes de equipo para la nieve y ropa de abrigo esperan un comportamiento anual opuesto al del fabricante de albercas.
3. **Variación cíclica:** Con frecuencia las series de tiempo presentan secuencias alternas de puntos abajo y arriba de la línea de tendencia que duran más de un año, esta variación se mantiene después de que se han eliminado las variaciones o tendencias estacional e irregular. Un ejemplo de este tipo de variación son los ciclos comerciales cuyos períodos recurrentes dependen de la prosperidad, recesión, depresión y recuperación, las cuales no dependen de factores como el clima o las costumbres sociales.
4. **Variación Irregular:** Esta se debe a factores a corto plazo, imprevisibles y no recurrentes que afectan a la serie de tiempo. Como este componente explica la variabilidad aleatoria de la serie, es impredecible, es decir, no se puede esperar predecir su impacto sobre la serie de tiempo. Existen dos tipos de variación irregular: a) Las variaciones que son provocadas por acontecimientos especiales, fácilmente identificables, como las elecciones, inundaciones, huelgas, terremotos. b) Variaciones aleatorias o por casualidad, cuyas causas no se pueden señalar en forma exacta, pero que tienden a equilibrarse a la larga.



**Referencias**

Baray, H. L. (2006). Introducción a la Metodología de la Investigación. Juan Carlos Martínez Coll.

Explorable.com. (02 de Mayo de 2009). La Correlación Estadística. Obtenido de https://explorable.com/es/la-correlacion-estadistica

Instituto Nacional de Estadística. (s.f.). Tipos de gráficos. Obtenido de http://www.ine.es/explica/docs/pasos\_tipos\_graficos.pdf

Kerlinger, F. (1983). Investigación del Comportamiento. Técnicas y Metodología. México: Interamericana.

Ramírez, M. M. (s.f.). Departamento de Estadística y Centro de Estadística Aplicada. Obtenido de ITAM: http://allman.rhon.itam.mx/~mendoza/EyP.pdf

Universidad de Sonora. (s.f.). Series de tiempo. Obtenido de Departamento de Matemáticas: http://www.estadistica.mat.uson.mx/Material/seriesdetiempo.pdf