Clase 1

Estadística descriptiva

1.1 Introducción

No hace falta ser un experto para estar en contacto con la estadística. Con frecuencia, los medios de comunicación presentan información de naturaleza estadística, como encuestas de intención de voto y de opinión, información de rating, evolución de las tasas de cambio de diferentes divisas y otros indicadores económicos, datos del clima, etc.

Estos ejemplos son solo son una parte del trabajo estadístico y detrás de ellos existe una amplia disciplina, una forma de razonar y un método de generación de conocimiento.

Los problemas que resuelve la estadística surgen de preguntas en diversas áreas que se fundamentan en el conocimiento inductivo. En un nivel más pragmático, la estadística tiene que ver con la forma más conveniente de obtener, resumir y analizar información. Algunos autores la definen como el arte de aprender de los datos.

La estadística se ha convertido en el lenguaje aceptado por la comunidad científica para la puesta a prueba, validación o rechazo de hipótesis de investigación; una descripción estadística es una forma de comunicación con un lenguaje especial.

La estadística se puede clasificar por su intención como ***descriptiva o inferencial***.

La primera tiene que ver con la mención de los hechos observados o la descripción de características de un conjunto de datos.

La segunda se refiere a la obtención de propiedades generales, basadas en una muestra de un conjunto de datos.

En esta materia estudiaremos principalmente la estadística descriptiva, estudiando algunas técnicas para hacer descripciones estadísticas y luego realizaremos un enfoque sobre la probabilidad, que es la base de la inferencia estadística. Finalmente trabajaremos algunos métodos informáticos centrados en las bases de datos, que nos permitan exploración y obtener datos, que al fin y al cabo son los "ingredientes" de la estadística descriptiva.

1.2 La estadística

Hay diversas definiciones sobre la estadística, pero todas ellas rondan acerca de la obtención y procesamiento de datos con el fin de obtener conclusiones y tomar decisiones.

Definición

La estadística es la rama de la matemática que estudia la variabilidad, colección, organización, análisis, interpretación, y presentación de conjuntos de datos, así como el proceso aleatorio que los genera siguiendo las leyes de la probabilidad; teniendo como finalidad principal el de obtener conclusiones y poder tomar decisiones.

1.2.1 Relación con el método científico

El uso de métodos estadísticos en la manufactura, el desarrollo de productos alimenticios, el software para computadoras, las fuentes de energía, los productos farmacéuticos y muchas otras áreas implican el **acopio de información o datos científicos**. Por supuesto que la obtención de datos no es algo nuevo, ya que se ha realizado por más de mil años. Los datos se han recabado, resumido, reportado y almacenado para su examen cuidadoso.

Este acopio de información y procesamiento de la información es lo que caracteriza a la estadística descriptiva.

Un paso más allá se encuentra la estadística inferencial, que ha recibido mucha atención en las últimas décadas dado que provee de un número enorme de "herramientas" de los métodos estadísticos que utilizan los profesionales.

Estos métodos estadísticos contribuyen al proceso de poder tomar decisiones con bases científicas frente a la incertidumbre y a la variación que caracterizan a la mayoría de los procesos.

Los métodos estadísticos se diseñan para contribuir al proceso de realizar juicios científicos frente a la incertidumbre y a la variación. Por ejemplo, dentro de un proceso de manufactura, la densidad de producto de un material específico no siempre será la misma. Los métodos estadísticos se utilizan para analizar datos de procesos como estos; con el objetivo de tener una mejor orientación respecto de que cambios se deben realizar en el proceso para mejorar su calidad.

En un caso de un estudio biomédico de un nuevo fármaco, que reduce la hipertensión, 85% de los pacientes experimentaron alivio; aunque se reconoce que el medicamento actual o el "viejo" alivia a 80% de los pacientes que sufren hipertensión crónica.

Sin embargo, el nuevo fármaco es más caro de elaborar y podría tener algunos efectos colaterales. ¿Se debería adoptar el nuevo medicamento? Éste es un problema con el que las empresas farmacéuticas, junto con la FDA (Federal Drug

Administration), se encuentran a menudo (a veces es mucho más complejo). De nuevo se debe tomar en cuenta las necesidades de variación y para estos estudios es evidente que la estadística tiene mucho que aportar.

El valor del "85%" se basa en cierto número de pacientes seleccionados para el estudio. Tal vez si se repitiera el estudio con nuevos pacientes ¡el número observado de "éxitos" sería de 75%. Se trata de una variación natural de un estudio a otro que se debe tomar en cuenta en el proceso de toma de decisiones.

En los problemas analizados anteriormente los métodos estadísticos empleados tienen que ver con la variabilidad y en cada caso la variabilidad que se estudia se encuentra en datos científicos.

Es bien conocida la importancia del pensamiento estadístico para los administradores y el uso de la inferencia estadística para el personal científico. Los investigadores obtienen mucho de los datos científicos. Los datos proveen conocimiento acerca del fenómeno científico y este conocimiento permite determinar las modificaciones que se requieren realizar para mantener el proceso en el nivel de calidad deseado.

1.2.2 La utilización de la estadística en informática

En esta sección se trata de dar una visión general de cómo la estadística aporta a la informática. No confundir con el caso inverso, pues es evidente que la informática, desde su propia definición como ciencia que se enfoca en el tratamiento automático de la información, ha favorecido el procesamiento de datos que requiere la estadística, y se ha vuelto fundamental para esa tarea.

La estadística es importante porque en muchas ocasiones un profesional de la informática debe trabajar con situaciones inciertas, pero que se desean predecir con el mayor nivel de precisión posible. Otras tantes veces la estadística ayuda al procesamiento de grandes cantidades de datos con el fin de obtener resultados o bien aprender de dichos datos.

Tiene su aplicación en diferentes áreas de especialización de la informática, tales como son la ***minería de datos***, el ***machine learning*** y la ***inteligencia artificial***. Estas tres áreas están a su vez relacionadas entre sí.

La **minería de datos** es el proceso de hallar anomalías, patrones y correlaciones en grandes conjuntos de datos para predecir resultados. Empleando una amplia variedad de técnicas, puede utilizar esta información para incrementar sus ingresos, recortar costos, mejorar sus relaciones con clientes, reducir riesgos y más.

Es evidente, desde la misma definición, que la estadística tiene mucho que aportar a la minería de datos.

El proceso de hurgar en los datos para descubrir conexiones ocultas y predecir tendencias futuras tiene una larga historia. Conocido algunas veces como

"descubrimiento de conocimientos en bases de datos", el término "**minería de datos**" recién se comenzó a utilizar en la década de 1990.

Su base comprende tres disciplinas científicas entrelazadas: **la estadística** (el estudio numérico de relaciones de datos), **inteligencia artificial** (inteligencia similar a la humana exhibida por software y/o máquinas) y **machine learning** (algoritmos que pueden aprender de datos para hacer predicciones). Lo que era antiguo es nuevo otra vez, ya que la minería de datos continúa evolucionando para igualar el ritmo del potencial sin límites del big data y poder de cómputo asequible.

En la última década, los métodos estadísticos junto con los avances en el poder y la velocidad de procesamiento nos han permitido llegar más allá de las prácticas manuales, tediosas y que toman mucho tiempo al análisis de datos rápido, fácil y automatizado. Cuanto más complejos son los conjuntos de datos recopilados, mayor es el potencial que hay para descubrir "*insights*"[1] relevantes.

Los comerciantes detallistas, bancos, fabricantes, proveedores de telecomunicaciones y aseguradoras, entre otros, utilizan la minería de datos para descubrir relaciones entre todas las cosas, desde precios, promociones y demografía hasta la forma en que la economía, el riesgo, la competencia y los medios sociales afectan sus modelos de negocios, ingresos, operaciones y relaciones con clientes.

Para trabajar con grandes volúmenes de información, la minería utiliza técnicas modernas de aprendizaje automático (machine learning), tanto para identificar patrones y tendencias de datos sin intervención humana como para predecir datos a priori desconocidos en función del comportamiento de los grupos. Los sistemas con machine learning aprenden por sí solos mediante la repetición, sin necesidad de que el programador introduzca nuevos datos o códigos. Es una tecnología muy utilizada por ejemplo en el uso de chats bot.

El big data

El trabajo con estos grandes volúmenes de datos (big data), su procesamiento y obtención de información de ellos, se ve influenciado por la estadística.

Cada una de las acciones que realizamos diariamente mediante dispositivos electrónicos, redes sociales y espacios de internet, generan información. Esta información puede revelar comportamientos, preferencias y sentimientos de los usuarios. Un ejemplo evidente es la controversia acerca de la recopilación de datos realizada por la empresa Facebook, que ha sido renombrada a Meta, y la duda de con qué finalidad y por quienes son utilizados nuestros datos. Esta controversia, lejos de disiparse, se incrementa por el actual interés de Meta por

[1] Hace referencia a los datos concretos que nos aportan una información valiosa para la optimización general, un ejemplo de insight puede ser detectar un patrón de comportamiento en el consumidor, habitualmente estos puntos concretos solo pueden ser vistos por expertos en Big Data.

ser la primera empresa en desarrollar un Metaverso y la necesidad de contar con una cuenta de Facebook para conectarse a dicho espacio virtual.

Imagen del presidente de *Meta*, Mark Zuckerberg, publicitando su versión del Metaverso.

En la era digital, la información es sinónimo de poder. Por ello, desentrañar el significado de este tipo de datos resulta vital para las empresas actuales. Dicha acción puede lograrse mediante el Big Data o análisis masivo de datos, que es considerada como una disciplina del futuro.

Se entiende por Big Data a enormes conjuntos de datos que solo pueden entenderse mediante su procesamiento a través de aplicaciones informáticas. Estas aplicaciones buscan patrones y coincidencias dentro de los datos, clasificándolos y organizándolos de forma en que sean de utilidad y puedan comprenderse sencillamente.

La estadística y el big data comparten un mismo objetivo: la claridad. Mediante diferentes acciones, buscan organizar los datos de tal forma que estos resulten claros y comprensibles. A su vez, tienen un segundo objetivo definido: la toma de decisiones. Con la información analizada pueden diseñarse caminos a recorrer y acciones a realizar, por lo que tanto la Estadística como el Big Data pueden definirse como necesarios para la toma de decisiones.

La inteligencia artificial

Las aplicaciones de la inteligencia artificial (I.A.) en estadística persiguen integrar distintos contrastes, estimaciones, transformaciones y modelos para conseguir una aproximación coherente y total en Análisis de Datos, estableciendo estrategias que dirijan el proceso de modelización, de elección de técnicas y transformaciones a aplicar, y de ayuda a la interpretación de los resultados.

La Estadística aporta a la Inteligencia Artificial no sólo herramientas para la resolución de problemas específicos, sino también el tratamiento probabilístico de los problemas de incertidumbre, así como al estudio de los procesos de aprendizaje y desarrollo conceptual.

Hasta hace relativamente no muchos años, la comunidad de investigadores en IA no se mostraba atraída por el razonamiento probabilístico. Las razones de esta desafección inicial parecían claras: las probabilidades son números y la contribución de la IA se ha de producir, de acuerdo con su fundamento más exigente, aportando herramientas y técnicas para el razonamiento no numérico. Sin embargo, actualmente se llegó al acuerdo de que la única descripción satisfactoria de la incertidumbre se tiene incorporando la probabilidad; la probabilidad es inevitable.

Por otro lado, la principal aportación de la Estadística al tratamiento del problema del aprendizaje en IA, la ha realizado a través de las técnicas de clasificación y ordenamiento de datos

En conclusión, la relación entre las estadísticas, el big data, el aprendizaje automático y la inteligencia artificial es mutuamente beneficiosa y está siendo rápidamente reconocida, los científicos y estadísticos descubren áreas útiles donde estas disciplinas se superponen. A diferencia de lo que ocurrió desde el nacimiento de la Inteligencia Artificial, hasta hace relativamente pocos años, el modelo probabilístico es hoy considerado por un número creciente de investigadores en Inteligencia Artificial, presumiblemente ya mayoritario, como el mejor modelo numérico para el tratamiento de la incertidumbre.

Todas estas áreas giran en torno a una pregunta clave *¿cómo aprendemos de los datos?*

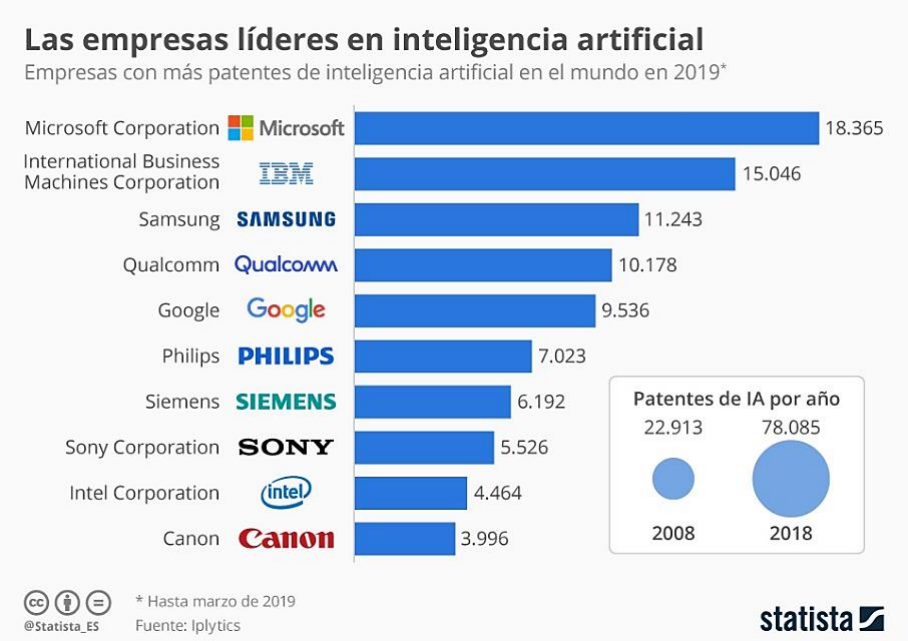
1.2.3 Estadística descriptiva

Como dijimos previamente, la estadística comprende la recolección de datos, el análisis de la información y la interpretación de los resultados.

*La **estadística descriptiva** busca resumir, entender y comunicar los datos de forma que se pueda generar una comprensión no disponible antes del análisis.*

Cuando apliquemos un promedio, mediana, moda, cuartiles, desviación media y estándar, cuando expongamos los datos en tablas de frecuencias, cuando los presentemos en gráficos estadísticos y hagamos usos de otros indicadores tales como los percentiles, coeficiente de variación, entre otros; estaremos usando descriptores de un conjunto de datos y aplicando estadística descriptiva.

Por ejemplo, a continuación, podemos ver un gráfico descriptivo de las empresas con más patentes de inteligencia artificial al año 2019.



La estadística descriptiva se vale de un conjunto de números o medidas estadísticas. Tales números ofrecen un sentido de la ubicación del centro de los datos, de la variabilidad en los datos y de la naturaleza general de la distribución de observaciones en la muestra.

El software estadístico moderno, tal como el conocido MS Excel o el Infostat, permite el cálculo de medias, medianas, desviaciones estándar y otros estadísticos de una sola cifra, así como el desarrollo de gráficas que presenten una "huella digital" de la naturaleza de la muestra.

Entre los principales valores que se utilizan en estadística descriptiva se destacan las medidas de tendencia central y de dispersión o variabilidad, así como los gráficos estadísticos. Los cuales iremos tratando a lo largo de las próximas clases.

1.3 Población objetivo y muestra

*En estadística se llama **población** a la colección total de elementos con algunas características comunes y sobre la que se desea obtener alguna información o realizar algún análisis.*

El tamaño de la población es el número de elementos de esta colección y, generalmente, se denota con la letra N.

Con frecuencia, el tamaño de la población es demasiado grande como para intentar examinar a todos sus elementos, en muchos casos se tiene interés por conocer uno o varios parámetros de una población y no es posible por razones económicas, limitación de tiempo y aspectos prácticos tomar de manera exhaustiva a todos los elementos de la población. En ese caso se tiene que tomar una **muestra** que contenga las características de interés para el investigador.

*Una **muestra** es, por lo tanto, una parte de los elementos de la población que refleje el comportamiento general de la población con respecto a lo que se desea estudiar.*

Por supuesto que esta muestra no reflejará de manera exacta el comportamiento de la población, la única manera de hacer esto es considerar la población completa; sin embargo, tomando una muestra los suficientemente grande y representativa de la población podremos obtener información suficientemente fiable.

Ejemplo 1

Para conocer algunas características de la calidad del agua o de un producto terminado en una empresa, no se necesita tomar toda el agua de un lago o todos los productos terminados en un día de producción, basta con acudir a las técnicas de muestreo, tomando por ejemplo una muestra de agua de partes distantes del lago, o un producto cada hora del día; y calcular los parámetros que nos interesan de la población, los cuales serán estimaciones obtenidas a partir de las muestras.

Debe quedar en claro que la población es todo aquello que estemos considerando; la población puede ser un país, una provincia, una ciudad, un aula de un colegio. En el caso del aula del colegio es factible considerar toda la población. Pero en general las muestras se reúnen a partir de poblaciones demasiado grandes para ser consideradas en su totalidad. Una muestra es un subconjunto de la población cuya característica general es que sus datos son factibles de obtener, con mayor o menor dificultad.

Ejemplo 2

Un fabricante de discos SSD para PC podría desear eliminar defectos. Un proceso de muestreo podría implicar recolectar información de 50 discos tomados aleatoriamente durante el proceso, en un periodo específico.

En este caso la población sería representada por todos los discos SSD producidos por la empresa en el periodo específico.

Si se lograra mejorar el proceso de producción de los discos para computadora y se reuniera una segunda muestra de discos, cualquier conclusión que se obtuviera respecto de la efectividad del cambio en el proceso podría extenderse a toda la población de discos que se produzcan en el "proceso mejorado".

1.3.1 Procedimientos de muestreo

Aunque el muestreo parece ser un concepto simple, la complejidad de las preguntas que se deben contestar acerca de la población, o las poblaciones, en ocasiones requiere que el proceso de muestreo sea muy complejo.

Muestreo aleatorio simple

La importancia del muestreo adecuado gira en torno al grado de confianza de la información que obtendremos con respecto a la población total.

El muestreo aleatorio simple significa que cierta muestra dada de un tamaño muestral específico tiene la misma probabilidad de ser seleccionada que cualquiera otra muestra del mismo tamaño.

El término tamaño muestral simplemente indica el número de elementos en la muestra y se lo simboliza con la letra *n*. El tamaño de la muestra se puede calcular según el nivel de confianza y el margen de error que deseemos obtener en los valores calculados.

La ventaja del muestreo aleatorio simple radica en que ayuda a eliminar el problema de tener una muestra que refleje una población diferente (quizá más restringida) de aquella sobre la cual se necesitan realizar las inferencias.

Ejemplo

Se elige una muestra para contestar diferentes preguntas respecto de las preferencias políticas en cierto país. La muestra implica la elección de, digamos, 1000 familias y una encuesta a aplicar.

Ahora bien, suponga que no se utiliza el muestreo aleatorio, sino que todas o casi todas las 1000 familias se eligen de una misma zona urbana.

Se considera que las preferencias políticas en las áreas rurales difieren de las de las áreas urbanas, incluso difieren las de otras zonas urbanas con respecto a esa que se tuvo en cuenta para la muestra. En otras palabras, la muestra obtenida en realidad confinó a la población y, por lo tanto, las inferencias también se tendrán que restringir a la "población confinada", y en este caso el confinamiento podría resultar indeseable.

Se dice que la muestra con un tamaño de 1000 familias aquí descrita es una muestra ***sesgada***[2].

[2] Se dice que la muestra está sesgada cuando hay diferencia entre los datos de la muestra y los datos de toda la población. Generalmente se produce a causa de una mala elección de la muestra.