Notas de Estadística y Probabilidad

Carlos Martínez-Rodríguez
Academia de Matemáticas
Casa Libertad
carlos.martinez@uacm.edu.mx

Agosto 2025

UACM
Universidad Autónoma de la Ciudad de México
Nada humano me es ajeno

UACM
Universidad Autónoma de la Ciudad de México
Nada humano me es ajeno

La Estadística es una ciencia formal que estudia la recolección, análisis e interpretación de datos de una muestra representativa, ya sea para ayudar en la toma de decisiones o para explicar condiciones regulares o irregulares de algún fenómeno o estudio aplicado, de ocurrencia en forma aleatoria o condicional. Sin embargo, la estadística es más que eso, es decir, es el vehículo que permite llevar a cabo el proceso relacionado con la investigación científica. Es transversal a una amplia variedad de disciplinas, desde la física hasta las ciencias sociales, desde las ciencias de la salud hasta el control de calidad. Se usa para la toma de decisiones en áreas de negocios o instituciones gubernamentales.

La Estadística es mucho más que sólo números apilados y gráficas bonitas. Es una ciencia con tanta antigüedad como la escritura, y es por sí misma auxiliar de todas las demás ciencias. Los mercados, la medicina, la ingeniería, los gobiernos, etc. La Estadística que conocemos hoy en día debe gran parte de su realización a los trabajos matemáticos de aquellos hombres que desarrollaron la teoría de las probabilidades, con la cual se adhirió a la Estadística a las ciencias formales.

Definición

La Estadística es la ciencia cuyo objetivo es reunir una información cuantitativa concerniente a individuos, grupos, series de hechos, etc. y deducir de ello gracias al análisis de estos datos unos significados precisos o unas previsiones para el futuro.

La estadística, en general, es la ciencia que trata de la recopilación, organización presentación, análisis e interpretación de datos numéricos con el fin de realizar una toma de decisión más efectiva. Otros autores la definen como la expresión cuantitativa del conocimiento dispuesta en forma adecuada para el escrutinio y análisis.

- la palabra estadística, en primer término se usa para referirse a la información estadística;
- también se utiliza para referirse al conjunto de técnicas y métodos que se utilizan para analizar la información estadística; y
- el término estadístico, en singular y en masculino, se refiere a una medida derivada de una muestra.

Los métodos estadísticos tradicionalmente se utilizan para propósitos descriptivos, para organizar y resumir datos numéricos. La estadística descriptiva, por ejemplo trata de la tabulación de datos, su presentación en forma gráfica o ilustrativa y el cálculo de medidas descriptivas.
Es difícil conocer los orígenes de la Estadística. Desde los comienzos de la civilización han existido formas sencillas de estadística, pues ya se utilizaban representaciones gráficas y otros símbolos en pieles, rocas, palos de madera y paredes de cuevas para contar el número de personas, animales o ciertas cosas.

Su origen inicia posiblemente en la isla de Cerdeña, donde existen monumentos prehistóricos pertenecientes a los Nuragas, las primeros habitantes de la isla; estos monumentos constan de bloques de basalto superpuestos sin mortero y en cuyas paredes de encontraban grabados toscos signos que han sido interpretados con mucha verosimilidad como muescas que servían para llevar la cuenta del ganado y la caza.
Los babilonios usaban pequeñas tablillas de arcilla para recopilar datos en tablas sobre la producción agrícola y los géneros vendidos o cambiados mediante trueque. Otros vestigios pueden ser hallados en el antiguo Egipto, cuyos faraones lograron recopilar, hacia el año 3050 antes de Cristo, datos relativos a la población y la riqueza del país. De acuerdo al historiador griego Heródoto, dicho registro de riqueza y población se hizo con el objetivo de preparar la construcción de las pirámides. En el mismo Egipto, Ramsés II hizo un censo de las tierras con el objeto de verificar un nuevo reparto.

En el antiguo Israel la Biblia da referencias, en el libro de los Números, de los datos estadísticos obtenidos en dos recuentos de la población hebrea. El rey David por otra parte, ordenó a Joab, general del ejército hacer un censo de Israel con la finalidad de conocer el número de la población.También los chinos efectuaron censos hace más de cuarenta siglos. Los griegos efectuaron censos periódicamente con fines tributarios, sociales (división de tierras) y militares (cálculo de recursos y hombres disponibles). La investigación histórica revela que se realizaron 69 censos para calcular los impuestos, determinar los derechos de voto y ponderar la potencia guerrera.

Pero fueron los romanos, maestros de la organización política, quienes mejor supieron emplear los recursos de la estadística. Cada cinco años realizaban un censo de la población y sus funcionarios públicos tenían la obligación de anotar nacimientos, defunciones y matrimonios, sin olvidar los recuentos periódicos del ganado y de las riquezas contenidas en las tierras conquistadas. Para el nacimiento de Cristo sucedía uno de estos empadronamientos de la población bajo la autoridad del imperio. Durante los mil años siguientes a la caída del imperio Romano se realizaron muy pocas operaciones Estadísticas, con la notable excepción de las relaciones de tierras pertenecientes a la Iglesia, compiladas por Pipino el Breve en el 758 y por Carlomagno en el 762 DC. Durante el siglo IX se realizaron en Francia algunos censos parciales de siervos. En Inglaterra, Guillermo el Conquistador recopiló el Domesday Book o libro del Gran Catastro para el año 1086, un documento de la propiedad, extensión y valor de las tierras de Inglaterra.

Esa obra fue el primer compendio estadístico de Inglaterra. Aunque Carlomagno, en Francia; y Guillermo el Conquistador, en Inglaterra, trataron de revivir la técnica romana, los métodos estadísticos permanecieron casi olvidados durante la Edad Media.

Durante los siglos XV, XVI, y XVII, hombres como Leonardo de Vinci, Nicolás Copérnico, Galileo, Neper, William Harvey, Sir Francis Bacon y René Descartes, hicieron grandes operaciones al método científico, de tal forma que cuando se crearon los Estados Nacionales y surgió como fuerza el comercio internacional existía ya un método capaz de aplicarse a los datos económicos.

Para el año 1532 empezaron a registrarse en Inglaterra las defunciones debido al temor que Enrique VII tenía por la peste. Más o menos por la misma época, en Francia la ley exigió a los clérigos registrar los bautismos, fallecimientos y matrimonios. Durante un brote de peste que apareció a fines de la década de 1500, el gobierno inglés comenzó a publicar estadística semanales de los decesos. Esa costumbre continuó muchos años, y en 1632 estos Bills of Mortality (Cuentas de Mortalidad) contenían los nacimientos y fallecimientos por sexo.

En 1662, el capitán John Graunt usó documentos que abarcaban treinta años y efectuó predicciones sobre el número de personas que morirían de varias enfermedades y sobre las proporciones de nacimientos de varones y mujeres que cabría esperar. El trabajo de Graunt, condensado en su obra *Natural and Political Observations...Made upon the Bills of Mortality*, fue un esfuerzo innovador en el análisis estadístico.

Los eruditos del siglo XVII demostraron especial interés por la Estadística Demográfica como resultado de la especulación sobre si la población aumentaba, decrecía o permanecía estática. En los tiempos modernos tales métodos fueron resucitados por algunos reyes que necesitaban conocer las riquezas monetarias y el potencial humano de sus respectivos países. El primer empleo de los datos estadísticos para fines ajenos a la política tuvo lugar en 1691 y estuvo a cargo de Gaspar Neumann, un profesor alemán que vivía en Breslau. Este investigador se propuso destruir la antigua creencia popular de que en los años terminados en siete moría más gente que en los restantes, y para lograrlo hurgó pacientemente en los archivos parroquiales de la ciudad.

Después de revisar miles de partidas de defunción pudo demostrar que en tales años no fallecían más personas que en los demás. Los procedimientos de Neumann fueron conocidos por el astrónomo inglés Halley, descubridor del cometa que lleva su nombre, quien los aplicó al estudio de la vida humana. Sus cálculos sirvieron de base para las tablas de mortalidad que hoy utilizan todas las compañías de seguros.
Durante el siglo XVII y principios del XVIII, matemáticos como Bernoulli, Francis Maseres, Lagrange y Laplace desarrollaron la teoría de probabilidades. No obstante durante cierto tiempo, la teoría de las probabilidades limitó su aplicación a los juegos de azar y hasta el siglo XVIII no comenzó a aplicarse a los grandes problemas científicos.

Godofredo Achenwall, profesor de la Universidad de Gotinga, acuñó en 1760 la palabra estadística, que extrajo del término italiano statista (estadista). Creía, y con sobrada razón, que los datos de la nueva ciencia serían el aliado más eficaz del gobernante consciente. La raíz remota de la palabra se halla, por otra parte, en el término latino status, que significa estado o situación; Esta etimología aumenta el valor intrínseco de la palabra, por cuanto la estadística revela el sentido cuantitativo de las más variadas situaciones. Jacques Quételect es quien aplica las Estadísticas a las ciencias sociales. Este interpretó la teoría de la probabilidad para su uso en las ciencias sociales y resolver la aplicación del principio de promedios y de la variabilidad a los fenómenos sociales. Quételect fue el primero en realizar la aplicación práctica de todo el método Estadístico, entonces conocido, a las diversas ramas de la ciencia.

Entretanto, en el período del 1800 al 1820 se desarrollaron dos conceptos matemáticos fundamentales para la teoría Estadística; la teoría de los errores de observación, aportada por Laplace y Gauss; y la teoría de los mínimos cuadrados desarrollada por Laplace, Gauss y Legendre. A finales del siglo XIX, Sir Francis Gaston ideó el método conocido por Correlación, que tenía por objeto medir la influencia relativa de los factores sobre las variables. De aquí partió el desarrollo del coeficiente de correlación creado por Karl Pearson y otros cultivadores de la ciencia biométrica como J. Pease Norton, R. H. Hooker y G. Udny Yule, que efectuaron amplios estudios sobre la medida de las relaciones. Los progresos más recientes en el campo de la Estadística se refieren al ulterior desarrollo del cálculo de probabilidades, particularmente en la rama denominada indeterminismo o relatividad, se ha demostrado que el determinismo fue reconocido en la Física como resultado de las investigaciones atómicas y que este principio se juzga aplicable tanto a las ciencias sociales como a las físicas.

La historia de la estadística está resumida en tres grandes etapas o fases.

- Primera Fase. **Los Censos:** Desde el momento en que se constituye una autoridad política, la idea de inventariar de una forma más o menos regular la población y las riquezas existentes en el territorio está ligada a la conciencia de soberanía y a los primeros esfuerzos administrativos.

- Segunda Fase. **De la Descripción de los Conjuntos a la Aritmética Política:** Las ideas mercantilistas extrañan una intensificación de este tipo de investigación. Colbert multiplica las encuestas sobre artículos manufacturados, el comercio y la población: los intendentes del Reino envían a París sus memorias. Vauban, más conocido por sus fortificaciones o su Dime Royale, que es la primera propuesta de un impuesto sobre los ingresos, se señala como el verdadero precursor de los sondeos. Más tarde, Bufón se preocupa de esos problemas antes de dedicarse a la historia natural. La escuela inglesa proporciona un nuevo progreso al superar la fase puramente descriptiva. Sus tres principales representantes son Graunt, Petty y Halley. El penúltimo es autor de la famosa Aritmética Política. Chaptal, ministro del interior francés, publica en 1801 el primer censo general de población, desarrolla los estudios industriales, de las producciones y los cambios, haciéndose sistemáticos durantes las dos terceras partes del siglo XIX.

- Tercera Fase. **Estadística y Cálculo de Probabilidades:** El cálculo de probabilidades se incorpora rápidamente como un instrumento de análisis extremadamente poderoso para el estudio de los fenómenos económicos y sociales y en general para el estudio de fenómenos cuyas causas son demasiados complejas para conocerlos totalmente y hacer posible su análisis.

La Estadística para su mejor estudio se ha dividido en dos grandes ramas: **la Estadística Descriptiva y la Estadística Inferencial**.

- **Descriptiva:** consiste sobre todo en la presentación de datos en forma de tablas y gráficas. Esta comprende cualquier actividad relacionada con los datos y está diseñada para resumir o describir los mismos sin factores pertinentes adicionales; esto es, sin intentar inferir nada que vaya más allá de los datos, como tales.
- **Inferencial:** se realiza a partir de muestras, observaciones hechas sólo acerca de una parte de un conjunto numeroso de elementos y esto implica que su análisis requiere de generalizaciones que van más allá de los datos. Como consecuencia, la característica más importante del reciente crecimiento de la estadística ha sido un cambio en el énfasis de los métodos que describen a métodos que sirven para hacer generalizaciones. La Estadística Inferencial investiga o analiza una población a partir de una muestra.

UACM Universidad Autónoma de la Ciudad de México Nada humano me es ajeno

UACM Universidad Autónoma de la Ciudad de México Nada humano me es ajeno

El conjunto de los métodos que se utilizan para medir las características de la información, para resumir los valores individuales, y para analizar los datos a fin de extraerles el máximo de información, es lo que se llama *métodos estadísticos*. Los métodos de análisis para la información cuantitativa se pueden dividir en los siguientes seis pasos:

1. Definición del problema.
2. Recopilación de la información existente.
3. Obtención de información original.
4. Clasificación.
5. Presentación.
6. Análisis.

El centro de la metodología estadística comienza a desplazarse hacia técnicas de computación intensiva aplicadas a grandes masas de datos, y se empieza a considerar el método estadístico como un proceso iterativo de búsqueda del modelo ideal.
Al momento de recopilar los datos que serán procesados es posible cometer errores así como durante los cómputos de los mismos. No obstante, hay otros errores que no tienen nada que ver con la digitación y que no son tan fácilmente identificables. Algunos de éstos errores son:

- **Sesgo:** Es imposible ser completamente objetivo o no tener ideas preconcebidas antes de comenzar a estudiar un problema, y existen muchas maneras en que una perspectiva o estado mental pueda influir en la recopilación y en el análisis de la información. En estos casos se dice que hay un sesgo cuando el individuo da mayor peso a los datos que apoyan su opinión que a aquellos que la contradicen. Un caso extremo de sesgo sería la situación donde primero se toma una decisión y después se utiliza el análisis estadístico para justificar la decisión ya tomada.

- **Datos No Comparables:** el establecer comparaciones es una de las partes más importantes del análisis estadístico, pero es extremadamente importante que tales comparaciones se hagan entre datos que sean comparables.
- **Proyección descuidada de tendencias:** la proyección simplista de tendencias pasadas hacia el futuro es uno de los errores que más ha desacreditado el uso del análisis estadístico.
- **Muestreo Incorrecto:** en la mayoría de los estudios sucede que el volumen de información disponible es tan inmenso que se hace necesario estudiar muestras, para derivar conclusiones acerca de la población a que pertenece la muestra. Si la muestra se selecciona correctamente, tendrá básicamente las mismas propiedades que la población de la cual fue extraída; pero si el muestreo se realiza incorrectamente, entonces puede suceder que los resultados no signifiquen nada

La estadística es un conjunto de procedimientos para reunir, clasificar, codificar, procesar, analizar y resumir información numérica adquirida sistemáticamente. Permite hacer inferencias a partir de una muestra para extrapolarlas a una población. Aunque normalmente se asocia a muchos cálculos y operaciones aritméticas, y aunque las matemáticas están involucradas, en su mayor parte sus fundamentos y uso apropiado pueden dominarse sin hacer referencia a habilidades matemáticas avanzadas. De esta manera la estadística se relaciona con el método científico complementándolo como herramienta de análisis y, aunque la investigación científica no requiere necesariamente de la estadística, ésta valida muchos de los resultados cuantitativos derivados de la investigación.

La obtención del conocimiento debe hacerse de manera sistemática por lo que deben planearse todos los pasos que llevan desde el planteamiento de un problema, pasando por la elaboración de hipótesis y la manera en que van a ser probadas; la selección de sujetos (muestreo), los escenarios, los instrumentos que se utilizarán para obtener los datos, definir el procedimiento que se seguirá para esto último, los controles que se deben hacer para asegurar que las intervenciones son las causas más probables de los cambios esperados (diseño); hasta la elección del plan de análisis idóneo para el tipo de datos que se están obteniendo, es aquí donde la estadística entra en el estudio, aunque pueden existir otras herramientas de análisis si se está haciendo una investigación de corte cualitativo.

Una buena planeación permitirá que los resultados puedan ser reproducidos, mediante la comprobación empírica, por cualquier investigador interesado en refutar o comprobar las conclusiones que se hagan del estudio. De esta manera también se logrará la predicción de los fenómenos que se están estudiando, ayudando a conocer y prevenir los problemas sociales e individuales que forman parte del objeto de estudio de la psicología. El tratamiento de los datos de la investigación científica tiene varias etapas:

- En la etapa de recolección de datos del método científico, se define a la población de interés y se selecciona una muestra o conjunto de personas representativas de la misma, se realizan experimentos o se emplean instrumentos ya existentes o de nueva creación, para medir los atributos de interés necesarios para responder a las preguntas de investigación.Durante lo que es llamado trabajo de campo se obtienen los datos en crudo, es decir las respuestas directas de los sujetos uno por uno, se codifican (se les asignan valores a las respuestas), se capturan y se verifican para ser utilizados en las siguientes etapas.
- En la etapa de recuento, se organizan y ordenan los datos obtenidos de la muestra. Esta será descrita en la siguiente etapa utilizando la estadística descriptiva, todas las investigaciones utilizan estadística descriptiva, para conocer de manera organizada y resumida las características de la muestra.

- En la etapa de análisis se utilizan las pruebas estadísticas (estadística inferencial) y en la interpretación se acepta o rechaza la hipótesis nula.
- En investigación, el fenómeno en estudio puede ser cualitativo que implicaría comprenderlo y explicarlo, o cuantitativo para compararlo y hacer inferencias. Se puede decir que si se hace análisis se usan métodos cuantitativos y si se hace descripción se usan métodos cualitativos. Medición Para poder emplear el método estadístico en un estudio es necesario medir las variables.

- Medir: es asignar valores a las propiedades de los objetos bajo ciertas reglas, esas reglas son los niveles de medición.
- Cuantificar: es asignar valores a algo tomando un patrón de referencia. Por ejemplo, cuantificar es ver cuántos hombres y cuántas mujeres hay.
- Variable: es una característica o propiedad que asume diferentes valores dentro de una población de interés y cuya variación es susceptible de medirse. Las variables pueden clasificarse de acuerdo al tipo de valores que puede tomar como:

- Discretas o categóricas: en las que los valores se relacionan a nombres, etiquetas o categorías, no existe un significado numérico directo.
- Continuas: los valores tienen un correlato numérico directo, son continuos y susceptibles de fraccionarse y de poder utilizarse en operaciones aritméticas De acuerdo a la cantidad de valores.
- Dicotómica: sólo tienen dos valores posibles, la característica está ausente o presente.
- Policotómica: pueden tomar tres valores o más, pueden tomarse matices diferentes, en grados, jerarquías o magnitudes continuas.

- Niveles de Medición
  - Nominal Las propiedades de la medición nominal son:
    - Exhaustiva: implica a todas las opciones.
    - A los sujetos se les asignan categorías, por lo que son mutuamente excluyentes. Es decir, la variable está presente o no; tiene o no una característica
  - Ordinal: Las propiedades de la medición ordinal son:
    - El nivel ordinal posee transitividad, por lo que se tiene la capacidad de identificar que esto es mejor o mayor que aquello, en ese sentido se pueden establecer jerarquías
    - Las distancias entre un valor y otro no son iguales.

- Intervalo
  - El nivel de medición por intervalo requiere distancias iguales entre cada valor. Por lo general utiliza datos cuantitativos. Por ejemplo: temperatura, atributos psicológicos (CI, nivel de autoestima, pruebas de conocimientos, etc.)
  - Las unidades de calificación son equivalentes en todos los puntos de la escala. Una escala de intervalos implica: clasificación, magnitud y unidades de tamaños iguales.
  - Se pueden hacer operaciones aritméticas
  - Cuando se le pide al sujeto que califique una situación del 0 al 10 puede tomarse como un nivel de medición por intervalo, siempre y cuando se incluya el 0.

- Razón
  - La escala empieza a partir del 0 absoluto, por lo tanto incluye sólo los números por su valor en sí, por lo que no pueden existir los números con signo negativo. Por ejemplo: Peso corporal en kg., edad en años, estatura en cm.
  - Convencionalmente los datos que son de nivel absoluto o de razón son manejados como los datos intervalares.

Datos estadísticos inadecuados Los datos estadísticos son usados como la materia prima para un estudio estadístico. Cuando los datos son inadecuados, la conclusión extraída del estudio de los datos se vuelve obviamente inválida. Por ejemplo, supongamos que deseamos encontrar el ingreso familiar típico del año pasado en la ciudad Y de 50,000 familias y tenemos una muestra consistente del ingreso de solamente tres familias: 1 millón, 2 millones y no ingreso. Si sumamos el ingreso de las tres familias y dividimos el total por 3, obtenemos un promedio de 1 millón. Entonces, extraemos una conclusión basada en la muestra de que el ingreso familiar promedio durante el año pasado en la ciudad fue de 1 millón. Es obvio que la conclusión es falsa, puesto que las cifras son extremas y el tamaño de la muestra es demasiado pequeño; por lo tanto la muestra no es representativa.

Hay muchas otras clases de datos inadecuados. Por ejemplo, algunos datos son respuestas inexactas de una encuesta, porque las preguntas usadas en la misma son vagas o engañosas, algunos datos son toscas estimaciones porque no hay disponibles datos exactos o es demasiado costosa su obtención, y algunos datos son irrelevantes en un problema dado, porque el estudio estadístico no está bien planeado.

Sesgo significa que un usuario dé los datos perjudicialmente de más énfasis a los hechos, los cuales son empleados para mantener su predeterminada posición u opinión. Hay dos clases de sesgos: conscientes e inconscientes. Ambos son comunes en el análisis estadístico. Hay numerosos ejemplos de sesgos conscientes.

- Un anunciante frecuentemente usa la estadística para probar que su producto es muy superior al producto de su competidor. '.item Un político prefiere usar la estadística para sostener su punto de vista.
- Gerentes y líderes de trabajadores pueden simultáneamente situar sus respectivas cifras estadísticas sobre la misma tabla de trato para mostrar que sus rechazos o peticiones son justificadas.

Es casi imposible que un sesgo inconsciente esté completamente ausente en un trabajo estadístico. En lo que respecta al ser humano, es difícil obtener una actitud completamente objetiva al abordar un problema.

Es muy frecuente que un análisis estadístico contemple supuestos. Un investigador debe ser muy cuidadoso en este hecho, para evitar que éstos sean falsos. Los supuestos falsos pueden ser originados por:

- Quien usa los datos
- Quien está tratando de confundir (con intencionalidad)
- Ignorancia
- Descuido.

- Variable: Característica o fenómeno que puede tomar distintos valores.
- Dato: Mediciones o cualidades que han sido recopiladas como resultado de observaciones.
- Población: área o conjunto del cual son extraídos los datos; es el conjunto de elementos o individuos que poseen una característica común y medible acerca de la cual se desea información. También llamada *universo*.
- Muestra: Subconjunto de la población, seleccionado de acuerdo con una regla o plan de muestreo.
- Censo: Recopilación de todos los datos de interés para la investigación de la población.
- Estadística: Función o fórmula que depende de los datos de la muestra (es variable).
- Parámetro: Característica medible de la población.

UACM Universidad Autónoma de la Ciudad de México Nada humano me es ajeno

UACM Universidad Autónoma de la Ciudad de México Nada humano me es ajeno

Ejemplo

La universidad está interesada en determinar el ingreso de las familias de sus alumnos.

- ***Variable:*** *Ingreso per cápita de las familias.*
- ***Dato:*** *Ingreso per cápita de la familia de un alumno específico.*
- ***Población:*** *Las familias de todos los alumnos de la universidad.*
- ***Estadística:*** *Ingreso per cápita promedio de las familias seleccionadas en la muestra.*
- ***Parámetro:*** *Ingreso per cápita promedio de la población.*

Una muestra es representativa en la medida en que es imagen de la población. En general, el tamaño de una muestra depende principalmente de:

- Nivel de precisión deseado.
- Recursos disponibles.
- Tiempo involucrado en la investigación.

Además, el plan de muestreo debe considerar:

- La población.
- Los parámetros a medir.

Existe una gran cantidad de tipos de muestreo. En la práctica, los más utilizados son los siguientes:

- **Muestreo aleatorio simple** Es un método de selección de n unidades extraídas de N, de tal manera que cada muestra posible tiene la misma probabilidad de ser escogida. En la práctica, se enumeran las unidades de 1 a N, y a continuación se seleccionan n números aleatorios entre 1 y N, ya sea de tablas o de una urna con fichas numeradas.

Ejemplo

Considere la producción de TV de una compañía en un determinado turno ($N = 35$ televisores). Para efectos de control de calidad de la pantalla, se desea extraer una muestra aleatoria simple de tamaño $n = 5$. Si los 35 TV producidos son numerados del 1 al 35, una posible muestra podría ser: $3, 5, 18, 23, 30$.
Â¿Cuántas muestras posibles hay? (respuesta combinatoria: $\binom{35}{5}$).

- **Muestreo estratificado aleatorio** Se usa cuando la población está agrupada en pocos estratos (cada uno con muchas entidades). Consiste en sacar una muestra aleatoria simple de cada uno de los estratos, generalmente de tamaño proporcional al estrato.
- **Muestreo sistemático** Se utiliza cuando las unidades de la población están ordenadas. Para seleccionar una muestra de n unidades, se divide la población en n subpoblaciones de tamaño $K = N/n$. Se toma al azar una unidad de las primeras K y, a partir de ahí, cada K-ésima unidad. Si n_0 es la primera unidad seleccionada, la muestra es:

$$\{\, n_0,\ n_0 + K,\ n_0 + 2K,\ \ldots,\ n_0 + (n-1)K \,\}.$$

- **Muestreo por conglomerado** Se emplea cuando la población está dividida en grupos o conglomerados pequeños. Consiste en obtener una muestra aleatoria simple de conglomerados y luego censar cada uno de éstos.

UACM Universidad Autónoma de la Ciudad de México Nada humano me es ajeno

UACM Universidad Autónoma de la Ciudad de México Nada humano me es ajeno

- **Muestreo en dos etapas (bietápico)** La muestra se toma en dos pasos:
  - Seleccionar una muestra de unidades primarias.
  - Seleccionar una muestra de elementos a partir de cada unidad primaria escogida.

Obsevación

En la realidad, a veces no es posible aplicar libremente un solo tipo de muestreo; incluso podemos estar obligados a mezclarlos.

Las variables se pueden clasificar en dos grandes grupos:

- **Variables categóricas** Son aquellas que pueden ser representadas a través de símbolos, letras o palabras. Los valores que toman se denominan *categorías*, y los elementos que pertenecen a estas categorías se consideran idénticos respecto a la característica que se está midiendo.

Ejemplo

Variable: Profesión. Valores posibles:Programador, Técnico en Control de Alimentos, Técnico en Prevención de Riesgos, Técnico en Control del Medio Ambiente, Químico Analítico, Técnico Mecánico, Etc.

Las variables categóricas se dividen en dos tipos:

- **Ordinales:** Las categorías tienen un orden implícito; admiten grados de calidad (hay relación total entre categorías).

Ejemplo

Variable: Nivel de estudio de Enseñanza Básica. Valores: Primero Básico, Segundo Básico, Tercero Básico, . . . , Octavo Básico. A pesar de que admite grados de calidad, no es posible cuantificar la diferencia entre niveles adyacentes.

- Categóricas:
  - **Nominales:** No existe una relación de orden entre categorías.
- **Variables numéricas** Son aquellas que pueden tomar valores numéricos exclusivamente (mediciones). Se dividen en dos tipos: *discretas* y *continuas*.
  - **Discretas:** Toman valores en un conjunto finito o infinito numerable.
    Ejemplo (Variable: Número de sillas por sala). Valores: $0, 1, 2, 3, \ldots, n$.
  - **Continuas:** Toman valores en un subconjunto de los números reales, típicamente un intervalo.
    Ejemplo (Variable: Temperatura de Valparaíso en verano). Valores entre $5°$ y $30°$.

Obsevación

Para variables continuas, el ser humano ha debido definir unidades para medirlas (por ejemplo: el metro, la hora).

Los métodos básicos de la estadística inferencial son la **estimación** y el **contraste de hipótesis**; juegan un papel fundamental en la investigación.

- Calcular parámetros de la distribución de medias o proporciones muestrales de tamaño n, extraídas de una población de media y varianza conocidas.
- Estimar la media o la proporción de una población a partir de la media o proporción muestral.
- Utilizar distintos tamaños muestrales para controlar la confianza y el error admitido.

- Contrastar resultados obtenidos a partir de muestras.
- Visualizar gráficamente, mediante las respectivas curvas normales, las estimaciones realizadas.

En la mayoría de las investigaciones resulta imposible estudiar a todos los individuos de la población (por costo o inaccesibilidad). Mediante la inferencia estadística se obtienen conclusiones para una población no observada en su totalidad, a partir de estimaciones o resúmenes numéricos efectuados sobre la base informativa extraída de una muestra.

En definitiva: a partir de una población se extrae una muestra con alguno de los métodos existentes; con sus datos se generan *estadísticos* para realizar estimaciones o contrastes poblacionales.
Existen dos formas de estimar parámetros:

- **Estimación puntual:** con base en los datos muestrales, se propone un único valor para el parámetro.
- **Estimación por intervalo de confianza:** se determina un intervalo dentro del cual se encuentra el valor del parámetro con cierta probabilidad.

Si el objetivo del tratamiento inferencial es generalizar sobre poblaciones no observadas a partir de una parte de la población, la muestra debe ser **representativa y aleatoria**. Además, el tamaño muestral depende de múltiples factores: recursos (dinero y tiempo), importancia del tema, confiabilidad esperada, características del fenómeno, etc. A partir de la muestra se estiman parámetros como la media, varianza, desviación estándar o la forma de la distribución.

Recordemos los conceptos elementales

- Población: Conjunto de elementos sobre los que se observa un carácter común. Se representa con la letra N (tamaño poblacional).
- Muestra: Conjunto de unidades extraídas de la población. Cuanto más significativa, mejor será la muestra. Se representa con n (tamaño muestral).
- Unidad de muestreo: Está formada por uno o más elementos de la población. El total de unidades de muestreo constituye la población; son disjuntas entre sí.
- Parámetro: Resumen numérico de una variable de la población. Parámetros habituales: media poblacional μ, total poblacional T (p.ej. $T = N\mu$), proporción p.

- Estimador: Un estimador $\hat{\theta}$ de un parámetro θ es un *estadístico* usado para conocer el parámetro desconocido.
- Estadístico: Función de los valores muestrales; es una variable aleatoria cuya distribución se denomina *distribución muestral del estadístico*.
- Estimación: A partir de la muestra se extrapola el resultado a la población. Puede ser *puntual* o por *intervalo de confianza*.
- Prueba de Hipótesis: Determina si, con datos muestrales, es aceptable que una característica o parámetro poblacional tome cierto valor o pertenezca a un conjunto de valores.
- Intervalos de confianza: Proporción de veces que acertaríamos al afirmar que el parámetro θ está dentro del intervalo al seleccionar muchas muestras.

El objetivo de la inferencia es generalizar resultados de la muestra a la población. Interesa estudiar la distribución de ciertas funciones de la muestra (estadísticos muestrales).

Sea $x_1, \ldots, x_n$ una muestra aleatoria simple (m.a.s.) de la v.a. X con función de distribución F_0. Un *estadístico* T es cualquier función de la muestra que no contiene cantidades desconocidas.

Los estadísticos más usuales (para un carácter cuantitativo) y sus distribuciones asociadas:

$$\text{Media muestral:} \quad \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad \bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right). \tag{1}$$

$$\text{Cuasivarianza:} \quad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2, \qquad \frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}. \tag{2}$$

$$\text{Total muestral:} \quad t = \sum_{i=1}^{n} X_i, \qquad t \sim \mathcal{N}(n\mu, n\sigma^2). \tag{3}$$

UACM Universidad Autónoma de la Ciudad de México Nada humano me es ajeno

UACM Universidad Autónoma de la Ciudad de México Nada humano me es ajeno

Un estimador de un parámetro poblacional es una función de los datos muestrales. Por ejemplo, para estimar la talla media de un grupo, se extrae una muestra y se usa la media muestral como estimación puntual.
La media de la muestra estima a la media poblacional; la cuasivarianza muestral estima la varianza poblacional; el total muestral estima el total poblacional.
Sea $X_1, \ldots, X_n$ una m.a.s. Decimos que $\hat{\theta}$ es estimador de θ si el estadístico empleado para conocer θ es $\hat{\theta}$.

- **Insesgadez:** $\mathbb{E}[\hat{\theta}] = \theta$. Si no coincide, es sesgado.
- **Eficiencia:** Dados $\hat{\theta}_1$ y $\hat{\theta}_2$ para θ, $\hat{\theta}_1$ es más eficiente si $\text{Var}(\hat{\theta}_1) < \text{Var}(\hat{\theta}_2)$.
- **Suficiencia:** Usa toda la información de la muestra relativa a θ.
- **Consistencia:** $\hat{\theta}_n \xrightarrow{p} \theta$, es decir, $\lim_{n\to\infty} \mathbb{P}\big(|\hat{\theta}_n - \theta| < \varepsilon\big) = 1$ para todo $\varepsilon > 0$.
- **Método de los momentos:** iguala momentos poblacionales a muestrales (suelen ser consistentes).
- **Mínimos cuadrados:** minimiza una función de pérdida (p.ej. suma de cuadrados de residuos).
- **Máxima verosimilitud:** elige el valor del parámetro que hace más verosímil la muestra (suele dar estimadores consistentes y eficientes).

La probabilidad de que $\bar{X}$ sea *exactamente* igual a μ es cero ($\mathbb{P}[\bar{X} = \mu] = 0$) en variables continuas; por ello, en la práctica se prefiere el uso de *intervalos de confianza* y *contrastes de hipótesis*.

Un intervalo de confianza está determinado por dos valores dentro de los cuales afirmamos que está el verdadero parámetro con cierta probabilidad (nivel de confianza $1 - \alpha$). Es una expresión del tipo $[L, U]$ o $L \leq \theta \leq U$.

Conceptos clave:

- **Variabilidad del parámetro:** si no se conoce σ, puede aproximarse con datos previos o un estudio piloto.
- **Error de la estimación:** amplitud del intervalo (precisión). A mayor precisión, intervalo más estrecho y mayor tamaño muestral. Denotemos $E = U - L$.
- **Nivel de confianza** $(1 - \alpha)$: probabilidad de que el intervalo contenga al parámetro (típicamente 95 % o 99 %).
- **Nivel de significación** α: $\alpha = 1 - (1 - \alpha)$. Para 95 %, $\alpha = 0{,}05$.
- **Valor crítico:** para normal estándar Z, $z_{\alpha/2}$ satisface $\mathbb{P}(|Z| \leq z_{\alpha/2}) = 1 - \alpha$ (p.ej. si $\alpha = 0{,}05$, $z_{0{,}025} \approx 1{,}96$).

Con estas definiciones, si tras extraer una muestra se afirma que "3 es una estimación de la media con un margen de error de 0,6 y un nivel de confianza del 99 %", el intervalo es [2,7, 3,3].
A tamaño muestral fijo, error y nivel de confianza están relacionados: mayor confianza requiere intervalos más amplios (o mayor n). El tamaño muestral puede despejarse de la fórmula del intervalo deseado según el error máximo permitido.

Unilaterales y bilaterales.

$$\text{Unilateral:} \quad \mathbb{P}(Z \leq z_\alpha) = 1 - \alpha \quad \text{o} \quad \mathbb{P}(Z \geq z_{1-\alpha}) = 1 - \alpha.$$

$$\text{Bilateral:} \quad \mathbb{P}\big(- z_{\alpha/2} \leq Z \leq z_{\alpha/2}\big) = 1 - \alpha.$$