Análisis e interpretación de datos estadísticos

*María del Mar Ballesteros Aguado**

Introducción

En este capítulo se resumen, en primer lugar, los conceptos básicos más utilizados dentro de la disciplina de Estadística, incluida, dentro de las técnicas de análisis habituales en un proceso de investigación; a continuación se define el concepto de parámetros estadísticos, se describen los más utilizados y se detallan sus respectivas fórmulas de cálculo, así como las propiedades, utilidad y principales características de cada uno; en tercer lugar se trata la inferencia estadística, que se compone de la recogida de información, junto con los diferentes tipos de muestreo, los procesos de análisis para hacer comparaciones, validaciones y medición de errores, con el objetivo de poder deducir conclusiones acerca el comportamiento de las variables a medir, así como dimensionar el grado de incertidumbre o error al que nos enfrentamos. A continuación se describen los principales tipos de fuentes de información, junto con los principales problemas y limitaciones asociados a ellas, también se justifica la necesidad de depurar (o estilizar) los datos; y, finalmente, se conciencia al lector sobre la importancia de la calidad de la información, así como de la correcta aplicación de las técnicas de análisis, ya que marcan la diferencia entre un estudio válido y un estudio inútil, enfatizando que la recogida y depuración de datos es una fase crucial de un estudio estadístico.

Conceptos de Estadística

La *Estadística* trata de una serie de procedimientos para el recuento, ordenación y clasificación de información y datos, que caracterizan una población, obtenidos a través de observación empírica, para poder realizar comparaciones y obtener conclusiones, que sirvan de apoyo en el proceso de toma de decisiones.

El principal objetivo de la Estadística consiste en poder *decir algo* con respecto al comportamiento de un gran conjunto (población) de personas, mediciones u otros entes, en

* Licenciada en Ciencias Económicas, Especialidad Economía Cuantitativa, Universidad Complutense de Madrid. Master en Economía y Dirección de Empresas, I.E.S.E. (Harvard Business School), Universidad de Navarra. Doctora en Economía Aplicada, Universidad Rey Juan Carlos de Madrid. maria.ballesteros@acague.cl

base a las observaciones hechas sobre solo una parte (muestra) de dicho gran conjunto. La capacidad para *decir algo* sobre poblaciones con base en muestras, está basada en supuestos con respecto a algún modelo de probabilidad que permitirán explicar las características del fenómeno bajo observación.

El recuento, ordenación y clasificación de datos de cada variable entran dentro del ámbito de la *Estadística Descriptiva*; la comparación, identificación de relaciones entre variables y el análisis de supuestos o hipótesis, está dentro del campo de la *Inferencia Estadística.*

Un *estudio estadístico* consta de cuatro fases:

– Recogida de datos.
– Organización y representación de datos.
– Análisis de datos.
– Obtención de conclusiones.

Los principales conceptos que se manejan en los estudios estadísticos son los siguientes:

Población: una población se define como el conjunto de todos los elementos posibles que se van a estudiar –a través de un estudio estadístico. *(Ejemplo: los postulantes a las Escuelas Matrices militares de Chile).*

*Individuo: u*n individuo o unidad estadística es cada uno de los elementos que componen la población. *(Ejemplo: cada uno de los postulantes).*

Muestra y Muestreo: una muestra es un conjunto representativo de la población en estudio, por tanto el número de individuos de una muestra siempre será menor que el de la población. El muestreo es la técnica de reunión de datos de una proporción reducida y representativa de la población que se desea estudiar. Más adelante se describen diversas formas de realizar un muestreo.

Variable estadística: es cada una de las características o cualidades que poseen los individuos de una población. *(Ejemplo: edad, género, colegio, comuna de residencia, etc., de cada postulante).*

Valor de una Variable: un valor es cada uno de los distintos resultados que se pueden obtener en un estudio estadístico. *(Ejemplo: las diferentes edades de los postulantes).*

Dato: un dato es cada uno de los valores que se ha obtenido al realizar un estudio estadístico. *(Ejemplo: si estudiamos 100 postulantes, obtendremos 100 datos de edad).*

Frecuencia: es el número de veces que aparece cada valor.

Inferencia estadística: es un conjunto de técnicas que se utilizan para sacar conclusiones generales del comportamiento de una población, a partir del estudio de una muestra, y para medir el grado de fiabilidad o confianza de los resultados obtenidos. Para ello, previamente, hay que diseñar y elegir una muestra de la población objeto de estudio.

Variable estadística

Dentro del concepto, ya definido, de variable estadística se enmarcan dos tipos de variables:

Variable cualitativa: una variable cualitativa es aquella que recoge una característica o cualidad que no se puede medir con un número. *(Ejemplo: la comuna de residencia de los postulantes).* Podemos distinguir dos tipos**:**

– *Variable cualitativa nominal:* aquella variable que presenta valores no numéricos y que no admite un criterio de orden. (*Ejemplo: el estado civil, con las siguientes modalidades: soltero, casado, separado, divorciado y viudo).*

– *Variable cualitativa ordinal o variable cuasicuantitativa:* aquella **v**ariable cualitativa que presenta modalidades no numéricas, pero a las que se puede asignar un orden y, en algunos casos asignar un número de orden. (*Ejemplo: ranking en un campeonato deportivo).*

Variable cuantitativa: una variable cuantitativa siempre se expresa mediante un número, y se pueden realizar operaciones aritméticas con ella. *(Ejemplo: la edad de los estudiantes de una promoción).* Podemos distinguir dos subtipos:

– *Variable discreta:* es una variable que solo puede tomar valores concretos dentro de un rango. (*Ejemplo: La edad de un postulante: 17, 18, 19, 20, 21, 22, 23 o 24).*

– *Variable continua:* es una variable que puede tomar un número infinito de valores dentro de un rango. (*Ejemplo: La altura de los postulantes: 1,70, 1,82, 1,75, etc.).*

Distribución de frecuencias

La distribución de frecuencias o tabla de frecuencias es una ordenación, en forma de tabla, de los datos estadísticos obtenidos para una variable, asignando a cada uno de los datos su frecuencia correspondiente. Existen varios tipos de frecuencias:

Frecuencia absoluta: la frecuencia absoluta es el número de veces que aparece un determinado valor en un estudio estadístico. Se representa por f_i. La suma de las frecuencias absolutas es igual al número total de datos, que se representa por *N*. Para indicar resumidamente estas sumas se utiliza la letra griega Σ (sigma mayúscula) que se lee suma o sumatoria.

$$\sum_{i=1}^{i=n} f_i = N$$

$$f_1 + f_2 + f_3 + \ldots + f_n = N$$

Frecuencia relativa: la frecuencia relativa es el cociente entre la frecuencia absoluta de un determinado valor y el número total de datos. Se puede expresar en tantos por ciento y se representa por $\mathbf{n_i}$. La suma de las frecuencias relativas es igual a 1.

$$n_i = \frac{f_i}{N}$$

Frecuencia acumulada: la frecuencia acumulada es la suma de las frecuencias absolutas de todos los valores inferiores o iguales al valor considerado. Se representa por F_i.

Frecuencia relativa acumulada: la frecuencia relativa acumulada es el cociente entre la frecuencia acumulada de un determinado valor y el número total de datos. Se puede expresar en porcentajes.

La distribución de frecuencias[1] se suele agrupar en tablas denominadas distribución de frecuencias agrupadas o tabla con datos agrupados, y se emplea cuando las variables toman un gran número de valores o cuando la variable es continua. Para ello, se agrupan los valores en intervalos que tengan la misma amplitud denominados clases, y a cada clase se le asigna su frecuencia correspondiente**.**

- *Límites de la clase:* cada clase está delimitada por el límite inferior de la clase y el límite superior de la clase.
- *Amplitud de la clase:* la amplitud de la clase es la diferencia entre el límite superior e inferior de cada una.
- *Marca de clase:* la marca de clase es el punto medio de cada intervalo y es el valor que representa a todo el intervalo para el cálculo de algunos parámetros**.**

Diagramas de distribución de frecuencias[2]

Las distribuciones de frecuencias se representan a través de diagramas, para facilitar su estudio. De forma generalizada, los diagramas de frecuencias se agrupan en tres tipos:

Diagrama de barras: Un diagrama de barras se utiliza para de presentar datos cualitativos o datos cuantitativos de tipo discreto. Se representan sobre unos ejes de coordenadas, en el eje de abscisas se colocan los valores de la variable, y sobre el eje de ordenadas las frecuencias absolutas o relativas o acumuladas. Los datos se representan mediante barras de una altura proporcional a la frecuencia.

1 Ver ejemplos detallados en www.vitutor.com, http: //www.vitutor.com/estadistica/descriptiva/a_3.html
2 Ver ejemplos detallados en www.vitutor.com, http: //www.vitutor.com/estadistica/descriptiva/a_4.html

Polígonos de frecuencia: un polígono de frecuencias se forma uniendo los extremos de las barras mediante segmentos. También se puede realizar trazando los puntos que representan las frecuencias y uniéndolos mediante segmentos.

Diagrama de sectores: un diagrama de sectores –o gráfico de torta– generalmente se usa para representar variables cualitativas, aunque se puede utilizar para todo tipo de variables. Los datos se representan en un círculo, de modo que el ángulo de cada sector es proporcional a la frecuencia absoluta correspondiente**.**

$$\alpha = \frac{360^o}{N} \cdot f_i$$

Histogramas[3]

Un histograma es una representación gráfica de una variable en forma de barras. Los histogramas se utilizan cuando se manejan muchos datos, se pueden utilizar tanto para variables continuas como para variables discretas, previamente agrupadas en clases.

En el eje abscisas se construyen unos rectángulos cuya base tiene como ancho la amplitud del intervalo, y su altura toma el valor de la frecuencia absoluta de cada intervalo. La superficie de cada barra es proporcional a la frecuencia de los valores representados.

Polígono de frecuencias: para construir el polígono de frecuencias, en un histograma, se toma la marca de clase que coincide con el punto medio de cada rectángulo.

Histograma y polígono de frecuencias acumuladas: representando las frecuencias acumuladas de una tabla de datos agrupados, se puede obtener el histograma de frecuencias acumuladas y su correspondiente polígono.

Histogramas con intervalos de amplitud diferente: Para construir un histogramas con intervalo de amplitud diferente tenemos que calcular las diversas amplitudes de los intervalos y las alturas de cada uno delos rectángulos del histograma.

$$h_i = \frac{f_i}{a_i}$$

h_i es la altura del intervalo.
f_i es la frecuencia del intervalo.
a_i es la amplitud del intervalo.

Parámetros Estadísticos

Un parámetro estadístico es un número que se obtiene a partir de los datos de una distribución estadística. Los parámetros estadísticos sirven para sintetizar la información dada por una tabla o por una gráfica. Existen tres tipos de parámetros estadísticos:

[3] Ver ejemplos detallados en www.vitutor.com, http: //www.vitutor.com/estadistica/descriptiva/a_6.html

- Medidas de centralización: media aritmética, moda y mediana.
- Medidas de posición: los más utilizados son cuartiles, quintiles, deciles y percentiles.
- Medidas de dispersión: rango o recorrido, desviación media, varianza y desviación típica (o estándar).

Medidas de centralización

Se basan en el cálculo del centro de una muestra. Indican en torno a qué valor (centro) se distribuyen los datos de dicha muestra. La definición y forma de cálculo de las medidas de centralización se detallan a continuación.

Estas medidas sirven para poder visualizar como se distribuye una muestra alrededor de su promedio o de su valor central y, se utilizan para poder decidir si este valor es una buena representación de dicha muestra y, por tanto, utilizarlo de forma generalizada. Por ejemplo la renta per cápita o PIB per cápita, es una medida de centralización, pero debemos estudiar su distribución para decidir si realmente es una representación del poder adquisitivo generalizado de los habitantes de un país.

- *Media aritmética:* La media aritmética es el valor promedio de la distribución. Corresponde al valor obtenido al sumar todos los datos y dividir el resultado entre el número total de datos. También se conoce como *esperanza matemática*. $\overline{X}$ es el símbolo de la media aritmética. La media se puede hallar solo para variables cuantitativas.
  La fórmula general de cálculo es la siguiente:

$$\overline{X} = \frac{X_1 + X_2 + X_3 + \ldots X_n}{N} \qquad \overline{X} = \frac{\sum_{i=1}^{n} X_i}{N}$$

  En caso de datos agrupados la media aritmética se calcula considerando la frecuencia de cada valor, agrupados previamente en una tabla de frecuencias, su expresión se adjunta en el formulario resumen incluido en el Anexo de este artículo.

- *Moda*: La moda es el valor que más se repite en una distribución, o el que tiene mayor frecuencia absoluta. Se representa por *Mo*. Se puede hallar la moda para variables tanto cualitativas como cuantitativas. Solamente para datos cuantitativos: si dos valores consecutivos tienen la frecuencia máxima, la moda es el promedio de las dos puntuaciones adyacentes.

- *Mediana:* La mediana es la puntación que se sitúa en el centro de la distribución, separa la mitad superior de la distribución de la inferior, es decir divide la serie de datos en dos partes iguales (en número de datos, no en valor). Ocupa el lugar central de todos los datos cuando estos se han ordenado. La mediana se representa por *Me*, solamente se puede calcular en variables cuantitativas.

La mediana[4] es independiente del tamaño de los intervalos. Se calcula en tres pasos:

a) Ordenar los datos de menor a mayor.

b) Si la serie tiene un número impar de medidas la mediana es la puntuación que ocupa el lugar central de la misma.

c) Si la serie tiene un número par de puntuaciones la mediana es la media entre las dos puntuaciones centrales.

En el caso de datos agrupados, la mediana se encuentra en el intervalo donde la frecuencia acumulada llega hasta la mitad de la suma de las frecuencias absolutas.

En el caso de variables cuantitativas, cuando la media la moda y la mediana coinciden, nos encontramos ante una distribución normal y, en una primera aproximación, podríamos tomar cualquiera de estas medidas de centralización, o el valor central de la muestra como una representación válida de la muestra.

Medidas de posición

Las medidas de posición solamente se pueden utilizar para variables cuantitativas y dividen un conjunto de datos en grupos con el mismo número de individuos, o de observaciones. Esta división de una muestra en grupos o segmentos, con el mismo número de datos cada uno, nos permitirá confirmar si la distribución de los mismos es simétrica o si es asimétrica.

Si fuese asimétrica, la utilización de una medida de centralización global no sería un procedimiento adecuado, siendo más aconsejable estudiar por separado cada uno de los segmentos de dicha muestra. Por ejemplo, la distribución del PIB per cápita se analiza generalmente utilizando quintiles.

Para calcular las medidas de posición siempre va a ser necesario que los datos estén ordenados de menor a mayor. Las medidas de posición más utilizadas son cuartiles, deciles y percentiles. La definición y forma de cálculo de las medidas de posición es la siguiente:

– *Cuartiles:* los cuartiles dividen la serie de datos en cuatro partes iguales**.** Los cuartiles son los tres valores de la variable que dividen a un conjunto de datos ordenados en cuatro partes iguales. Q_1, Q_2 y Q_3 determinan los valores correspondientes al 25%, al 50% y al 75% de los datos. Q_2 coincide siempre con la mediana.

  Para calcular los cuartiles hay que seguir dos pasos:

  a) Ordenar los datos de menor a mayor.

  b) Buscar el lugar que ocupa cada cuartil mediante la expresión: $\frac{k \cdot N}{4}, k = 1,2,3$

  Para calcular cuartiles en el caso de datos agrupados, En primer lugar se busca la clase donde se encuentra el valor $\frac{k \cdot N}{4}, k = 1,2,3$, en la tabla de frecuencias acumuladas y después se aplica la fórmula:

[4] Se pueden consultar las fórmulas y ejemplos detallados para el cálculo de la mediana en http: //www.vitutor.com/estadistica/descriptiva/a_9.html

$$Q_k = L_i + \frac{\frac{k \cdot N}{4} - F_{i-1}}{f_i} \cdot a_i \qquad k = 1,2,3$$

Donde:
L_i es el límite inferior de la clase donde se encuentra el cuartil
N es la suma de las frecuencias absolutas.
F_{i-1} es la frecuencia acumulada anterior a la clase del cuartil.
a_i es la amplitud de la clase.

– *Deciles:* los deciles dividen la serie de datos en diez partes iguales. Los deciles son los nueve valores que dividen la serie de datos en diez partes iguales. Los deciles muestran los valores correspondientes al 10%, al 20%... y al 90% de los datos. D_5 coincide con la mediana. D_5 coincide con Q_2. Para calcular los deciles también hay que ordenar los datos de menor a mayor y, después, buscar el lugar que ocupa cada decil, mediante la expresión: $\frac{k \cdot N}{10}, k = 1,2,\ldots 9$

  Para calcular deciles en el caso de datos agrupados, se sigue el mismo procedimiento, aplicando las fórmulas que se detallan en el formulario incluido en el Anexo 1.

– *Percentiles:* los percentiles dividen la serie de datos en cien partes iguales. Los percentiles son los 99 valores que dividen la serie de datos en 100 partes iguales. Los percentiles dan los valores correspondientes al 1%, al 2%... y al 99% de los datos. P_{50} coincide con la mediana. P_{50} coincide con D_5 y con Q_2. Para calcular percentiles también hay que ordenar los datos de menor a mayor y, después, buscar el lugar que ocupa cada percentil mediante la expresión: $\frac{k \cdot N}{100}, k = 1,2,\ldots 99$

  Para calcular percentiles en el caso de datos agrupados, el mismo procedimiento, aplicando las fórmulas que se detallan en el Anexo 1.

Medidas de dispersión

Las medidas de dispersión reflejan cuánto se alejan del centro los valores de la distribución y *todas están basadas en la desviación respecto a la media.* Solamente se pueden utilizar para variables cuantitativas, y nos permiten medir el grado de error o de incertidumbre que manejamos al utilizar las medidas de centralización como aproximación generalizada del valor de una muestra.

Por ejemplo utilizaríamos este tipo de medidas si sabemos que la edad promedio observada de entrada a la Escuela Militar es de 21 años y quisiéramos saber si efectivamente la mayor concentración de alumnos tiene alrededor de 21 años o la realidad es que solo el 20% tienen 21, mientras que el 40% tienen 18 y otro 40% tienen 24. En ambos casos el promedio sería 21 años, pero la dispersión sería diferente.

La desviación respecto a la media se define como la diferencia entre cada valor de la variable estadística y la media aritmética (promedio) y su fórmula es $D_i = (x - x)$. Las medidas de dispersión son: rango o recorrido, desviación media, varianza y desviación típica o estándar.

– *Rango o recorrido:* es la diferencia entre el mayor y el menor de los datos de una distribución estadística.

– *Desviación media:* es la media aritmética de los valores absolutos de las desviaciones respecto a la media. La desviación media se representa por $D_{\bar{X}}$. No es la más utilizada.

$$D_{\bar{X}} = \frac{\left|X_1 - \bar{X}\right| + \left|X_2 - \bar{X}\right| + \ldots + \left|X_n - \bar{X}\right|}{N} \qquad D_{\bar{X}} = \frac{\sum_{i=1}^{n} \left|X_i - \bar{X}\right|}{N}$$

– *Varianza:* es la media aritmética del cuadrado de las desviaciones respecto a la media de una distribución estadística. La varianza de una se representa por . Es muy utilizada.

$$\sigma^2 = \frac{\left(X_1 - \bar{X}\right)^2 + \left(X_2 - \bar{X}\right)^2 + \ldots + \left(X_n - \bar{X}\right)^2}{N} \qquad \sigma^2 = \frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2}{N}$$

Es importante mencionar algunas propiedades y características de la varianza:

a) La varianza nunca toma valor negativo, será siempre positiva o cero (por ser un cuadrado).

b) Si a todos los valores de la variable se les suma el mismo número, la varianza no cambia.

c) Si todos los valores de la variable se multiplican por el mismo número, la varianza queda multiplicada por el cuadrado de dicho número.

d) Si tenemos varias distribuciones con la misma media y conocemos sus respectivas varianzas se puede calcular la varianza total:

Si todas las muestras tienen el mismo tamaño: $\sigma^2 = \frac{\sigma_1^2 + \sigma_2^2 + \ldots + \sigma_n^2}{N}$

Si las muestras tienen distinto tamaño: $\sigma^2 = \frac{k_1 \cdot \sigma_1^2 + k_2 \cdot \sigma_2^2 + \ldots + k_n \cdot \sigma_n^2}{k_1 + k_2 + \ldots + k_n}$

e) La varianza, al igual que la media, es un parámetro muy sensible a los valores extremos.

f) Si no se puede calcular la media, tampoco se puede calcular la varianza, ya que depende de ella.

g) La varianza no viene expresada en las mismas unidades que los datos, ya que las desviaciones están elevadas al cuadrado.

También se puede calcular la varianza cuando los datos son agrupados y además existen algunas fórmulas simplificadas (ver formulario resumen al final del artículo).

– *Desviación típica*: la desviación típica es la raíz cuadrada de la varianza. Es decir, la raíz cuadrada de la media de los cuadrados de las puntuaciones de desviación. La desviación típica se representa por σ.

$$\sigma = \sqrt{\frac{(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \ldots + (X_n - \overline{X})^2}{N}} \qquad \sigma = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{N}}$$

Es importante mencionar algunas propiedades y características de la desviación típica:

a) La desviación típica toma siempre valor positivo o cero.
b) Si se suma el mismo número a todos los valores de la variable, la desviación típica no cambia.
c) Si se multiplican todos los valores de la variable por el mismo número, la desviación típica queda multiplicada por dicho número.
d) Si tenemos varias distribuciones con la misma media y conocemos sus respectivas desviaciones típicas se puede calcular la desviación típica total:

   Si todas las muestras tienen el mismo tamaño: $\sigma = \sqrt{\frac{\sigma_1^2 + \sigma_2^2 + \ldots + \sigma_n^2}{n}}$

   Si las muestras tienen distinto tamaño: $\sigma = \sqrt{\frac{k_1 \cdot \sigma_1^2 + k_2 \cdot \sigma_2^2 + \ldots + k_n \cdot \sigma_n^2}{k_1 + k_2 + \ldots + k_n}}$

e) La desviación típica, es un parámetros muy sensible a las puntuaciones extremas, igual que la media y la varianza, ya que depende de ambas.
f) Si no se puede calcular la media, tampoco se puede calcular la desviación típica.
g) Cuanto menor sea la desviación típica mayor será la concentración de datos alrededor de la media.

   También se puede calcular la desviación típica cuando los datos son agrupados y además existen algunas fórmulas simplificadas equivalentes (ver formulario en el Anexo 1).

Inferencia Estadística

La inferencia estadística estudia cómo sacar conclusiones generales para toda la población, a partir del estudio de una muestra (tomada de esa población), así como valora el grado de fiabilidad o significación de los resultados obtenidos.

Para ello, inicialmente hay que determinar la elección de una muestra de una población al azar, que represente adecuadamente dicha población y que tenga tamaño suficiente; después tomar los datos de dicha muestra y tabularlos; estimar sus parámetros; y, finalmente, analizar su fiabilidad para poder obtener conclusiones y poder reducir la incertidumbre en la toma de decisiones.

Experimentos aleatorios y deterministas

Los diferentes hechos que pueden ser observados en la naturaleza, o los experimentos que se pueden realizar, se clasifican en dos categorías: deterministas y aleatorios.

Se denomina experimento o fenómeno determinista a aquél que siempre se produce de la misma forma cuando se dan las mismas condiciones. Esto ocurre, por ejemplo, el tiempo que tarda un misil en recorrer un espacio dado con movimiento uniforme, a velocidad constante.

Un fenómeno *aleatorio*, por el contrario, es aquel que incluye la posibilidad de que en idénticas condiciones puedan producirse resultados diferentes, que no pueden ser previstos con anticipación. Por ejemplo el tiempo de demora en establecer entre dos conexiones o transmisiones por radio. Sin embargo, si se hace una larga serie de experiencias, se puede observar cierta regularidad que puede ayudar a estudiar estos fenómenos. Esto se llama ley del azar o de estabilidad de las frecuencias[5].

A través del estudio de estos fenómenos, se han desarrollado una serie de técnicas estadísticas que permiten obtener conclusiones sobre el comportamiento de estos fenómenos. La inferencia estadística recoge estas técnicas para el análisis de una población, sin embargo, generalmente no es posible abarcar una población en su totalidad y se hace necesario reducir el estudio a una muestra[6].

Muestreo: Poblaciones, censos y muestras

Como ya se ha mencionado, la inferencia estadística estudia cómo sacar conclusiones generales para toda la población, a partir del estudio de una muestra, y valora el grado de fiabilidad o confianza de los resultados obtenidos. Para ello, inicialmente, es imprescindible determinar la elección de una muestra, al azar, que represente adecuadamente dicha población.

Una población (o universo) es el conjunto total de objetos que se deben estudiar para analizar un problema dado, o para responder a una pregunta de investigación planteada. Los objetos pueden ser personas, animales, productos fabricados, fenómenos naturales, etc. Cada uno de ellos recibe el nombre de elemento (o individuo) de la población. Por lo general, en los estudios estadísticos, el investigador analiza algún aspecto parcial de los individuos que componen la población: por ejemplo, la edad, profesión, nivel de estudios, el sueldo mensual, el número de personas de su familia, la opinión sobre el partido que gobierna, etc. Estos aspectos parciales reciben el nombre de caracteres de los individuos de una población y son variables, es decir, en distintos individuos pueden tomar valores diferentes.

5 Al repetir un mismo experimento "A" n veces, la frecuencia relativa, o el cociente n_A/n entre las veces que aparece A (n_A) y el número total de repeticiones, tiende a estabilizarse alrededor de un número (probabilidad de dicho resultado).

6 De acuerdo con el diccionario de la RAE, inferir significa *sacar una consecuencia o deducir algo de otra cosa.* Al conjunto de procedimientos estadísticos en los que interviene la aplicación de modelos de probabilidad y, mediante los cuales, se realiza alguna afirmación sobre poblaciones, con base en la información obtenida de muestras, se le llama Inferencia Estadística o Estadística Inferencial.

Si la población es finita, el mejor procedimiento será el estudio de todos y cada uno de los individuos. Un estudio estadístico realizado sobre la totalidad de una población se denomina censo. Estudios de este tipo son realizados periódicamente por el Gobierno y otras instituciones.

Sin embargo, la mayoría de los problemas en estudio, implican, poblaciones infinitas, o poblaciones finitas muy grandes que son difíciles, costosas o imposibles de inspeccionar en su totalidad. Esto obliga a seleccionar –de forma adecuada– un subconjunto de *n* elementos de la población, que constituyen una muestra de tamaño *n*, para examinar la característica que interesa y después poder generalizar estos resultados a la población.

Para que estas conclusiones ofrezcan las debidas garantías es preciso comprobar que las muestras están adecuadamente diseñadas y medidas, es decir, cumplen el requisito básico de que *la muestra es representativa de la población en estudio.* Estas comprobaciones, y la posterior generalización, de los resultados obtenidos, se realizan por medio de los procedimientos estadísticos recogidos dentro de la inferencia estadística.

Tipos de muestreo

Existen varios tipos de muestreo aleatorio, que dependen de la forma en que se elige una muestra representativa de la población objeto de estudio: simple, sistemático y estratificado.

Para que sea válido, el muestreo siempre debe ser aleatorio, de forma que la muestra y los resultados del análisis no estén sesgados y no pierdan la representatividad de la población y por tanto la validez.

– *Muestreo aleatorio simple.* Para obtener una muestra, se numeran los elementos de la población y se seleccionan al azar los *n* elementos que contiene la muestra.
– *Muestreo aleatorio sistemático*[7]. Se elige un individuo al azar y a partir de él, a intervalos constantes, se eligen los demás hasta completar la muestra.
– *Muestreo aleatorio estratificado.* Se divide la población en estratos y se escoge, aleatoriamente, un número de individuos de cada estrato proporcional al número de componentes de cada estrato.

Para garantizar la representatividad de la muestra es imprescindible determinar adecuadamente el tamaño muestral, es decir, el número mínimo necesario de observaciones para que los resultados de investigación sean válidos. Para calcular el tamaño de la muestra suele utilizarse la siguiente fórmula[8]:

[7] Ejemplo: Si tenemos una población formada por 100 elementos y queremos extraer una muestra de 25 elementos, en primer lugar debemos establecer el intervalo de selección que será igual a 100/25 = 4. A continuación elegimos el elemento de arranque, tomando aleatoriamente un número entre el 1 y el 4, y a partir de él obtenemos los restantes elementos de la muestra. 2, 6, 10, 14,..., 98.

[8] Mario Suárez, *Interaprendizaje de Estadística Básica,* Ecuador: Ed. Gráficas Planeta, 2011.

$$n = \frac{N\sigma^2 Z^2}{(N-1)e^2 + \sigma^2 Z^2}$$

Donde:
n = el tamaño de la muestra.
N = tamaño de la población.
σ = Desviación estándar de la población que, generalmente cuando no se conoce, se toma valor constante de 0,5.
Z = Valor obtenido mediante niveles de confianza. Valor constante que, si no se conoce, se toma por defecto 95% de confianza equivale a 1,96 (como más usual). Otros valores quedan a criterio del investigador.
e = Límite aceptable de error muestral que, cuando no se conoce, se establece entre el 1% (0,01) y 9% (0,09), a criterio del encuestador[10].

Distribución muestral

Un muestreo puede hacerse con o sin reposición, y la población de partida puede ser infinita o finita. Generalmente, en ciencias sociales, se asumen poblaciones de partida infinitas o muestreos con reposición.

Cuando se consideran todas las posibles muestras de tamaño n en una población, es posible calcular, para cada muestra, sus parámetros estadísticos (media, desviación típica,...) que variarán de una a otra. De esta forma se puede obtener una distribución del parámetro que se llama distribución muestral.

Generalmente no conocemos la media ni la dispersión de la población total por lo que, para tomar decisiones y a través del estudio de esa muestra y sus parámetros, inferiremos (extrapolaremos) que los valores muéstrales representan, o son similares, a los poblacionales, estableciendo y considerando, al tomar decisiones, el grado de incertidumbre resultante de los análisis realizados.

El nivel de confianza (ρ) se designa mediante $1 - \alpha$.[10] El nivel de significación se designa mediante α. El valor crítico (k) como $Z\alpha/2$. $P(Z > z\alpha/2) = \alpha/2$ $P[-Z\alpha/2 < z < Z\alpha/2] = 1-\alpha$.

9 Ejemplo: En una fábrica que consta de 600 trabajadores queremos tomar una muestra de 20. Sabemos que hay 200 trabajadores en la sección A, 150 en la B, 150 en la C y 100 en la D. ¿Cuántos se toman de cada sección?

$\frac{20}{600} = \frac{X_1}{200}$ $X_1 = 6,6 \approx 7$ trabajadores de A

$\frac{20}{600} = \frac{X_2}{150}$ $X_2 = 5$ 5 trabajadores de B

$\frac{20}{600} = \frac{X_3}{150}$ $X_3 = 5$ 5 trabajadores de C

$\frac{20}{600} = \frac{X_4}{100}$ $X_4 = 3,3 \approx 3$ trabajadores de D

10 Se pueden consultar las fórmulas y ejemplos detallados para el cálculo de intervalos de confianza en http: // www.vitutor.com/estadistica/ /inferencia/intervalos_1.html

Estimación de Parámetros[11] y niveles de confianza

Cuando se el comportamiento de una población, se toma una muestra, y se pueden calcular los parámetros de esa muestra. En la realidad estamos estudiando una población, para poder sacar conclusiones, a través de una muestra, por lo que se busca poder aproximar los parámetros de dicha población, basándose en los resultados obtenidos de la muestra.

- *Estimación de parámetros* es el procedimiento utilizado para conocer las características de un parámetro poblacional, a partir del conocimiento de la muestra. Con una muestra aleatoria, de tamaño *n*, se puede aproximar el valor de un parámetro de la población, pero dado que no es exacto, necesitamos precisar un intervalo de confianza y un error de estimación admisible.
- *Intervalo de confianza* es un intervalo (valores mínimo y máximo) en el que sabemos que está un parámetro, con un nivel de confianza específico. Siendo el nivel de confianza la probabilidad de que el parámetro a estimar se encuentre dentro del intervalo de confianza. Error de estimación admisible es el grado de error máximo que se acepta en la estimación de un parámetro y está relacionado con el radio (o la amplitud) del intervalo de confianza.

  El investigador debe establecer los niveles de confianza exigidos a cada parámetro (media, desviación o varianza), así como determinar y cuáles son los intervalos, los valores mínimos o los valores máximos admisibles para cada uno de dichos parámetros de la población.

  Elegir el análisis de intervalos, de valores mínimos o de valores máximos, dependerá del problema planteado y, todos ellos, se hacen a través del planteamiento y test de hipótesis que se explica más adelante en el epígrafe *"3.5 Hipótesis Estadísticas*".

Relaciones entre variables: Correlación, Regresión y Análisis Factorial

Dentro del análisis del comportamiento de las poblaciones, otro objetivo puede ser la identificar la existencia de tendencias comunes o variaciones simultáneas en dicho comportamiento, no solamente el estudio de cada una por separado.

Estas evoluciones comunes se identifican estudiando la evolución de los valores que van tomando los datos, la técnica estadística utilizada para ver si dos variables están relacionadas o no se denomina *correlación estadística.* Solamente es aplicable para variables cuantitativas.

Por ejemplo, si analizamos el ingreso familiar y el gasto familiar, se observa sabe que ingresos y gastos aumentan o 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.

Una medición matemática de esta relación es el *Coeficiente de Correlación de Pearson*[12], que mide el grado de relación lineal entre dos variables aleatorias cuantitativas (varía entre 0 y 1

[11] Los parámetros referidos son los definidos previamente en el epígrafe *Parámetros Estadísticos* de este artículo.

[12] La validez o fiabilidad de la correlación entre variables se analiza con la *Prueba χ^2 de Pearson:* es una prueba no paramétrica que mide la discrepancia entre dos distribuciones. {\displaystyle \chi ^{2}=\sum _{i}{\frac {(\mathrm {observada} _{i}-\mathrm {teorica} _{i})^{2}}{\mathrm {teorica} _{i}}}} cuanto más se acera a cero el valor de chi-cuadrado, más ajustadas están ambas distribuciones. Ver Análisis de Hipótesis, Epígrafe de este artículo {\displaystyle \chi ^{2}}

y es independiente de la escala de las variables). De forma menos formal, podemos definir este coeficiente como un índice que puede utilizarse para medir el grado de relación de dos variables.

Otra medición estadística de la relación lineal entre variables es la *Regresión Lineal* o *Ajuste Lineal*[13]. Esta consiste en un modelo matemático usado para aproximar la relación de dependencia entre una variable dependiente (y) y otras variables independientes (x_i). Este modelo se basa en el análisis de las *Varianzas* de las variables y en las *Covarianzas,* y también mide la relación entre ellas. Pero, a diferencia del anterior, los coeficientes resultantes permiten calcular (o estimar) el valor de una variable si se conoce el valor de las otras. Relaciona directamente la magnitud de las variables, no solamente el grado de relación y, al igual que el Coeficiente de Correlación, solamente es aplicable para variables cuantitativas.

Otra técnica estadística utilizada para analizar relaciones entre variables es el *Análisis Factorial,* se trata de simplificar el análisis, explicando las correlaciones entre las variables observadas por medios de un número menor de variables, no observadas directamente, que se llaman factores.[14] Este análisis se utilizar para variables cualitativas.

El análisis factorial exploratorio, AFE, se usa para tratar de descubrir la estructura interna de un número relativamente grande de variables. La hipótesis a priori del investigador es que pueden existir una serie de factores asociados a grupos de variables. Es el tipo de análisis factorial más común.

El análisis factorial confirmatorio, AFC, trata de determinar si el número de factores obtenidos corresponden con los que cabría esperar en base a una teoría previa acerca de los datos. La hipótesis a priori es que existen factores preestablecidos y que cada uno de ellos está asociado con un determinado subconjunto de las variables. Esto entregaría un nivel de confianza para poder aceptar o rechazar dicha hipótesis.

Hipótesis estadísticas

Para poder extraer conclusiones de un estudio, previamente se plantean hipótesis de estudio o preguntas de investigación, que generalmente están relacionadas con el comportamiento de una o varias variables, ya sea de forma independiente o conjunta.

En el caso del análisis estadístico, las hipótesis previas se refieren al valor que toma un parámetro desconocido de una población y, se comprueba la validez del valor obtenido de dicho parámetro para poder extrapolarlo a la población estudiada.

Las hipótesis pueden realizarse sobre un intervalo de valores, un porcentaje o un valor absoluto concreto, que toma cualquier parámetro estadístico: media, varianza, coeficientes de regresión, diferencia de muestras, etc. Solamente cambia el tipo de test aplicado que se refleja en las tablas de probabilidad utilizadas, en función de la distribución de la variable.

[13] La validez de los coeficientes de regresión se analiza a través de la *Prueba t-Student,* o *Test-T* estos coeficientes. Ver Análisis de Hipótesis, Epígrafe 3.5 de este artículo.

[14] Por ejemplo, se concluye que dentro de la población estudiantil, aquellos que obtienen nota alta en una prueba de habilidad verbal también se desempeñan bien en pruebas que requieren habilidades verbales. Los investigadores explican esto mediante el uso de análisis factorial, aislando el factor llamado inteligencia cristalizada o inteligencia verbal, que representa el grado en el cual alguien es capaz de resolver problemas usando habilidades verbales.

Cuando se analiza si dos muestras o poblaciones son similares, se contrasta la hipótesis "*Diferencia de Medias es igual a 0*".

Las de uso más generalizado son la distribución normal Z y la prueba t-Student o Test-T para analizar medias, coeficientes de regresión lineal y calcular intervalos de confianza y la Prueba χ^2 de Pearson y el Test-F (de Fisher) para analizar dispersión y varianzas.

El procedimiento se llama *test estadístico*, y permite extraer conclusiones que permitan aceptar o rechazar una hipótesis previamente emitida sobre el valor de un parámetro desconocido de una población. Se utilizan dos hipótesis: la hipótesis a comprobar es H_0, y se llama hipótesis nula; la hipótesis contraria se designa por H_1 y se llama hipótesis alternativa.

El procedimiento se denomina "Contraste de Hipótesis" y consiste en aceptar o rechazar la hipótesis nula. Existen dos tipos de contraste, bilateral y unilateral, que se diferencian en la forma de plantear las hipótesis y las regiones de aceptación. Los pasos a seguir son los siguientes:

a) Enunciar la hipótesis nula H_0 y la alternativa H_1.
   - Si estamos haciendo un contraste bilateral, estamos planteando aceptar o rechazar un solo valor de un parámetro: $H_0 = k$ o $H_1 \neq k$
   - Si estamos haciendo un contraste unilateral, estamos planteando aceptar o rechazar un rango, es decir, que el valor del parámetro es mayor o menor a uno predeterminado[15]:
   - $H_0 \geq k$ o $H_1 < k$ o $H_0 \leq k$ o $H_1 > k$
   - A partir de un nivel de confianza $1 - \alpha$ o el de significación α: Determinar el valor $z_{\alpha/2}$ (bilaterales), o bien z_α (unilaterales), o el valor t y calcular la zona de aceptación del parámetro μ o p.

b) Calcular: x o p', a partir de la muestra.

c) Si el valor del parámetro muestral está dentro de la zona de la aceptación, se acepta la hipótesis H_0 con un nivel de significación α. Si no, se rechaza H_0 y se validaría la hipótesis alternativa.

Como ejemplo: el *Contraste Bilateral* se utiliza si la hipótesis nula es de tipo H_0: $\mu = k$ (o bien H_0: $p = k$) y la hipótesis alternativa, por tanto, es del tipo H_1: $\mu \neq k$(o bien H_1: $p \neq k$).

El nivel de significación α se concentra en dos partes (o colas) simétricas respecto de la media.

Figura
Curva de distribución

La *región de aceptación* en este caso es el correspondiente intervalo de probabilidad para μ o p.
El contraste unilateral tiene nivel de significación en una cola.

Fuente: Elaboración propia.

15 Para calcular los valores críticos de contraste, aunque no son los únicos, se usan frecuentemente el *Test-T* (Student) y el *Test-F* (Fisher) en contrastes bilaterales, así como la *Prueba* χ^2 *de Pearson* en contrastes unilaterales.

Los errores de estimación cometidos se clasifican en: Errores de tipo I y tipo II: a) Error tipo I se comete cuando la hipótesis nula es verdadera y, como consecuencia del contraste, se rechaza el parámetro obtenido tomando una decisión incorrecta; b) Error tipo II. Se comete cuando la hipótesis nula es falsa y, como consecuencia del contraste se acepta el valor de esta hipótesis, incurriendo en una decisión incorrecta.

La probabilidad de cometer Error de tipo I es el nivel de significación α. La probabilidad de cometer Error de tipo II depende del verdadero valor del parámetro. Se hace tanto menor cuanto mayor sea *n.*[16]

Cuadro
Error de estimación

H_0	Verdadera	Falsa
Aceptar	Decisión correcta Probabilidad = 1 – α	Decisión incorrecta: ERROR DE TIPO II *(Cuando se acepta el valor falso)*
Rechazar	ERROR DE TIPO I *(Cuando se rechaza el valor verdadero)* Probabilidad = α	Decisión correcta

Fuente: Elaboración propia.

Naturaleza y fuentes de información

La información puede tener diferente naturaleza, en función de cómo se registra.

Series Temporales: Se denomina así a todos los conjuntos de observaciones sobre los valores que toma una variable en diferentes momentos del tiempo, se recopilada en intervalos regulares (días, meses, años, etc.). Puede ser tanto cuantitativa (sueldo) como cualitativa (masculino o femenino). Para poder trabajar fácilmente con series de tiempo, estas series deben ser estacionarias, es decir, el valor de su media y varianza no deben variar sistemáticamente a través del tiempo.

Series de Corte Transversal: Son conjuntos de datos, de una o más variables recogidos en el mismo momento del tiempo (censos, encuestas regionales), etc. La información de corte transversal tiene problemas de heterogeneidad, por lo que debe tenerse en cuenta el efecto del tamaño o la escala.

Información combinada. Es una mezcla de ambas, se toman datos no aleatorios, de corte transversal a lo largo de diferentes momentos del tiempo. Permiten analizar la situación de una población en un momento del tiempo y, además su evolución temporal. (Ejemplo de ello es la encuesta de caracterización socioeconómica en Chile, CASEM, que recoge los datos de las mismas familias a lo largo del tiempo).

[16] Las fórmulas de cálculo de intervalos característicos, valores críticos de contraste y contrastes de hipótesis se pueden consultar en el formulario resumen final.

Fuentes de Información

Los datos pueden ser recogidos directamente por el investigador, procedentes del emisor de dicha información o de fuentes oficiales que registran y ordenan dicha información. Estas fuentes pueden ser institucionales, como el Military Balance, Ejército, ONU, etc., o encuestas realizadas por el propio investigador. Además la información puede ser experimental (cuando los datos están condicionados u obtenidos exclusivamente en el proceso de investigación), o no experimental (cuando los datos no están sujetos al control del investigador, es decir, están dados).

Información Primaria es aquella que el investigador recoge directamente de la fuente y que está sin elaborar, es decir el investigador deberá procesar. Ejemplo de ello es la información recogida directamente a través de encuestas o de experimentos realizados por el investigador.

Información Secundaria es aquella que el investigador recoge indirectamente, generalmente de fuentes oficiales, que ha sido obtenida previamente de fuentes primarias por un tercero, y que ya está depurada, ordenada, resumida, etc. Esto no impide que el investigador pueda analizarla y/o reprocesarla para su investigación. (Ejemplos: Informes estadísticos del Military Balance).

Precisión y calidad de la información

La calidad de la información disponible no siempre es buena, por ello, el investigador debe tener siempre en mente que el resultado de la investigación solamente será tan bueno como lo sea la calidad de los datos, la depuración de los mismos, así como la correcta selección y aplicación de la(s) técnica(s) de análisis.

Podemos observar algunas fuentes de error que son bastante frecuentes:

- Errores de observación => cuando Información es no experimental.
- Errores de medición, debido a aproximación o redondeo.
- Diferencia de frecuencia en las serie temporales, debido a que los intervalos de medición pueden ser diferentes.
- Sesgo de selectividad (muestral) => cuando en las encuestas se omiten algunas respuestas.
- Los métodos de muestreo pueden variar => esto hace que las muestras no sean comparables.
- Alguna información está disponible a nivel altamente agregado (PIB, empleo, inflación) por lo que su análisis no puede hacerse de forma muy detallada.
- Mucha de la información es confidencial, por tanto de difícil acceso => condiciona la amplitud y profundidad del análisis.

Conclusiones

Es fundamental destacar que el investigador debe definir claramente, al inicio, el objeto y objetivo de la investigación para poder determinar la población a estudiar, identificar la información disponible y, en caso de que no exista y haya que tomarla directamente, ser capaz de dimensionar las muestras necesarias y elegir las técnicas de muestreo más adecuadas a utilizar.

Para evitar errores de observación y de medición en investigaciones no experimentales, es importante elegir adecuadamente las muestras y asegurar que la información recogida no está sesgada por juicios de valor ni por opiniones previas de los investigadores o de los encuestadores. También hay que considerar y definir las dimensiones de los datos porque no siempre es aconsejable utilizar redondeos.

Cuando se trata de información secundaria, no siempre se pueden evitar o corregir todos los errores de información, pero es importante identificarlos y documentarlos porque condicionan absolutamente la calidad de los resultados y por tanto las decisiones que se adopten a raíz de las conclusiones obtenidas.

En caso que se detecten distorsiones en algunos datos, por causas fortuitas o no habituales (como por ejemplo un incendio, un corte de energía o un atentado), se hace necesario eliminar dichas distorsiones. Estas se deberán corregir, eliminándolas cuando se conoce su magnitud y se pueda medir, o, en caso contrario, será necesario eliminar los datos distorsionados.

Si la información estudiada está asociada a series temporales que tienen diferente frecuencia, para compararlas se hará necesario unificar las frecuencias. Siempre se deberá hacer mediante agregación, nunca por desagregación, para no desvirtuar la información. (Ejemplo: si tenemos unos datos semanales y otros mensuales, hay que sumar los semanales para transformarlos en mensuales).

En aquellos casos en que el investigador observe la existencia de respuestas omitidas, o aprecie que los métodos de muestreo pudieran ser diferentes, la información no se puede procesar de la misma forma. Habrá que dividirla o separarla en grupos homogéneos, de esta forma se evita que las conclusiones estén desvirtuadas. Cuando no se puede separar en grupos homogéneos, dicha información deberá ser desestimada porque no debe formar parte del análisis.

Si el problema estuviera en que la información disponible es muy agregada, no es correcto dividirla a simple criterio del investigador, ya que estaría produciendo una distorsión o sesgo en la investigación y en sus resultados. Los datos deben tener siempre la misma frecuencia, por ejemplo, si tenemos datos anuales y trimestrales, no se debe dividir los datos anuales para transformarlos en trimestrales, sino que hay que sumar los trimestrales para que todos sean anuales. Aunque el analista sea consciente que la agregación va en detrimento de la profundidad y nivel de detalle del análisis, de forma similar a lo que ocurre con las diferencias en las frecuencias, deberá trabajar siempre con los datos más agregados. En caso contrario las conclusiones no serán correctas porque la información, de partida, estará desvirtuada.

Por último, es importante tener en cuenta que lo más habitual es disponer de información parcial, escasa y heterogénea que va contra la profundidad del análisis. El investigador

deberá, por tanto, documentar y acotar estas deficiencias al definir su análisis y también reflejarlo en sus conclusiones.

Bibliografía

Álvarez Contreras, Sixto Jesús, *Estadística aplicada, teoría y problemas* (Madrid, Ed. Clagsa, 2000).

Casas Sánchez, José Miguel; García Pérez, Carmelo; Rivera Galicia, Luis Felipe y Zamora Sanz, Ana Isabel, *Ejercicios de inferencia estadística y muestreo para economía y administración de empresas* (Madrid, Ed. Pirámide, 2006).

Gonzáles Rodríguez, Benito; Hernández Abreu, Domingo; Jiménez, Mateo; Marrero Rodríguez, María Isabel y Sanabria García, Alejandro, *Estadística descriptiva: problemas resueltos* (Tenerife, Universidad de La Laguna, 2013).

Gorgas García, Javier; Cardiel López, Nicolás y Zamorano Calvo, Jaime, *Estadística Básica para Estudiantes de Ciencias* (Madrid, Facultad de Ciencias Físicas, Universidad Complutense, 2011).

Mateo Rivas, María José, *Estadística en Investigación Social. Ejercicios Resueltos* (Madrid, Ed. Paraninfo, 1985).

Montero Lorenzo, José María, *Estadística para Relaciones Laborales* (Madrid, Ed. AC, 2003).

Ruíz-Maya Pérez, Luis y Martín-Pliego López, Francisco José, *Fundamentos de Inferencia Estadística* (Madrid, Paraninfo, 3ª Ed., 2005).

Salinas, Javier, *Problemas propuestos y Resueltos* [Granada, Universidad de Granada, www.ugr.es/~jsalinas/weproble/indice.htm 10.10.2017].

Suárez, Mario (2011), *Interaprendizaje de Estadística Básica*, Ecuador: Ed. Gráficas Planeta, 2011.

Verdoy, Pablo Juan; Beltrán, Modesto Joaquín y Peris, María José, *Problemas resueltos de estadística aplicada a las Ciencias Sociales* [Valencia, Universitat Jaume I, www.sapientia.ujoi.es. 16.10.2017].

FORMULARIO RESUMEN

Cuadro 1
Estadísticos descriptivos

MEDIA. Fórmula general $\bar{X} = \frac{X_1 + X_2 + X_3 + \ldots X_n}{N}$ $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{N}$	*MEDIA. Con datos agrupados* $\bar{X} = \frac{X_1 f_1 + X_2 f_2 + X_3 f_3 + \ldots X_n f_n}{N}$ $\bar{X} = \frac{\sum_{i=1}^{n} X_i f_i}{N}$
Cuartiles $Q_k = L_i + \frac{\frac{k \cdot N}{4} - F_{i-1}}{f_i} \cdot a_i$ $k = 1,2,3$	L_i *es el límite inferior de la clase donde se encuentra el cuartil* N *es la suma de las frecuencias absolutas* F_{i-1} *es la frecuencia acumulada anterior a la clase del cuartil* a_i *es la amplitud de la clase*
Deciles $D_k = L_i + \frac{\frac{k \cdot N}{10} - F_{i-1}}{f_i} \cdot a_i$ $k = 1,2,\ldots 9$	L_i *es el límite inferior de la clase donde se encuentra el decil* N *es la suma de las frecuencias absolutas* F_{i-1} *es la frecuencia acumulada anterior a la clase del decil* a_i *es la amplitud de la clase*
Percentiles $P_k = L_i + \frac{\frac{k \cdot N}{100} - F_{i-1}}{f_i} \cdot a_i$ $k = 1,2,\ldots 99$	L_i *es el límite inferior de la clase donde se encuentra el percentil* N *es la suma de las frecuencias absolutas* F_{i-1} *es la frecuencia acumulada anterior a la clase del percentil* a_i *es la amplitud de la clase*
Varianza. Fórmula general $\sigma^2 = \frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \ldots + (X_n - \bar{X})^2}{N}$ $\sigma^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{N}$	*Desviación Media* $D_{\bar{X}} = \frac{\lvert X_1 - \bar{X} \rvert + \lvert X_2 - \bar{X} \rvert + \ldots + \lvert X_n - \bar{X} \rvert}{N}$ $D_{\bar{X}} = \frac{\sum_{i=1}^{n} \lvert X_i - \bar{X} \rvert}{N}$
Varianza con datos agrupados $\sigma^2 = \frac{(X_1 - \bar{X})^2 f_1 + (X_2 - \bar{X})^2 f_2 + \ldots + (X_n - \bar{X})^2 f_n}{N}$ $\sigma^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2 f_i}{N}$	*Varianza. Fórmulas simplificadas* $\sigma^2 = \frac{X_1^2 + X_2^2 + \ldots + X_n^2}{N} - \bar{X}_2$ $\sigma^2 = \sum_{i=1}^{n} \frac{X_i^2}{N} - \bar{X}^2$
Desviación Típica. Fórmula general $\sigma = \sqrt{\frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \ldots + (X_n - \bar{X})^2}{N}}$ $\sigma = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{N}}$	*Desviación Típica. Fórmula simplificada* $\sigma = \sqrt{\frac{X_1^2 + X_2^2 + \ldots + X_n^2}{N} - \bar{X}_2}$ $\sigma = \sqrt{\sum_{i=1}^{n} \frac{X_i^2}{N} - \bar{X}^2}$
Desviación Típica con datos agrupados $\sigma = \sqrt{\frac{(X_1 - \bar{X})^2 f_1 + (X_2 - \bar{X})^2 f_2 + \ldots + (X_n - \bar{X})^2 f_n}{N}}$ $\sigma = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \bar{X})^2 f_i}{N}}$	*Desviación Típica. Fórmula simplificada* $\sigma = \sqrt{\frac{X_1^2 f + X_2^2 f_2 + \ldots + X_n^2 f_n}{N} - \bar{X}_2}$ $\sigma = \sqrt{\sum_{i=1}^{n} \frac{X_i^2 f_i}{N} - \bar{X}^2}$

Fuente: Elaboración propia.

Cuadro 2
Intervalos de confianza

Para la Media con varianza poblacional conocida	$\overline{X}-\frac{\sigma}{\sqrt{n}}z_{\frac{\alpha}{2}},\overline{X}+\frac{\sigma}{\sqrt{n}}z_{\frac{\alpha}{2}}$	*Para la Media con varianza poblacional desconocida (se utiliza la muestral)*	$\left(\overline{X}-t_{n=1,\frac{\alpha}{2}}\cdot\frac{S_{n-1}}{\sqrt{n}},\overline{X}+t_{n=1,\frac{\alpha}{2}}\cdot\frac{S_{n-1}}{\sqrt{n}}\right)$
Para la Media si no se conoce la varianza	$\left(-z_{\frac{\alpha}{2}},z_{\frac{\alpha}{2}}\right)$	*Para la Media con varianza poblacional y muestral desconocida*	$\left(-t_{n-1,\frac{\alpha}{2}},\ t_{n-1,\frac{\alpha}{2}}\right)$
Para la Varianza con varianza poblacional desconocida (se utiliza la muestral)	$\left(\frac{(n-1)S_{n-1}^2}{X^2_{n-1,\frac{\alpha}{2}}},\frac{(n-1)S_{n-1}^2}{X^2_{n-1,1\frac{\alpha}{2}}}\right)$	*Para la Varianza con varianza poblacional y muestral desconocida*	$\left(X^2_{n-1,\frac{\alpha}{2}},X^2_{n-1,\frac{\alpha}{2}}\right)$
Para una Proporción	$\left(\hat{p}-z_{\frac{\alpha}{2}}\cdot\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\hat{p}+z_{\frac{\alpha}{2}}\cdot\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$	*Para una Proporción si no se conoce la media muestral*	$\left(-z_{\frac{\alpha}{2}},z_{\frac{\alpha}{2}}\right)$

Fuente: Elaboración propia.

Cuadro 3
Contraste de Hipótesis unilateral

Contraste de la Media $H_0:\mu=\mu_0$ $H_1:\mu\neq\mu_0$ RECHAZO H_0 si $\left\lvert\frac{\overline{x}-\mu_0}{s/\sqrt{n}}\right\rvert>t_{n-1;\frac{\alpha}{2}}$	$H_0:\mu\leq\mu_0$ $H_1:\mu>\mu_0$ RECHAZO H_0 si $\frac{\overline{x}-\mu_0}{s/\sqrt{n}}>t_{n-1;\frac{\alpha}{2}}$	$H_0:\mu\geq\mu_0$ $H_1:\mu<\mu_0$ RECHAZO H_0 si $\frac{\overline{x}-\mu_0}{s/\sqrt{n}}<-t_{n-1;\frac{\alpha}{2}}$
Contraste de una Proporción $H_0:P=P_0$ $H_1:P\neq P_0$ RECHAZO H_0 si $\left\lvert\frac{p-P_0}{\sqrt{\frac{p_0q_0}{n}}}\right\rvert>z_{\frac{\alpha}{2}}$	$H_0:P\leq P_0$ $H_1:P>P_0$ RECHAZO H_0 si $\frac{p-P_0}{\sqrt{\frac{p_0q_0}{n}}}>z_a$	$H_0:P\geq P_0$ $H_1:P<P_0$ RECHAZO H_0 si $\frac{p-P_0}{\sqrt{\frac{p_0q_0}{n}}}<-z_a$
Contraste de la Varianza $H_0:\sigma^2=\sigma_0^2$ $H_1:\sigma^2\neq\sigma_0^2$ RECHAZO H_0 si $\frac{(n-1)s^2}{\sigma_0^2}\notin\left(\chi^2_{\frac{\alpha}{2}};\chi^2_{1-\frac{\alpha}{2}}\right)$	$H_0:\sigma^2\leq\sigma_0^2$ $H_1:\sigma^2>\sigma_0^2$ RECHAZO H_0 si $\frac{(n-1)s^2}{\sigma_0^2}>\chi^2_{\alpha}$	$H_0:\sigma^2\geq\sigma_0^2$ $H_1:\sigma^2<\sigma_0^2$ RECHAZO H_0 si $\frac{(n-1)s^2}{\sigma_0^2}<\chi^2_{1-\alpha}$

Fuente: Elaboración propia.