Estadística

CIF 5141

José Miguel Zúñiga Núñez.

Universidad de Playa Ancha.

05 de Abril 2022

Tabla de contenidos:

La Estadística.

ILa estadística

La estadística es la ciencia que estudia la recolección, organización, análisis e interpretación de un conjunto de datos, y sus conceptos generales pueden aplicarse a distintas disciplinas, como la ingeniería, la agricultura, la economía (llamada econometría),o la psicología(donde se conoce como biometría). Cuando se aplica a las ciencias de la salud , se utiliza el término bioestadística.

La Estadística.

Introducción

En término globales, la estadística puede dividirse en descriptiva y analítica. La Estadística descriptiva, como su nombre lo indica, solo pretende describir o caracterizar un conjunto de datos. La estadística analítica, en cambio, plantea hipótesis respecto a una población, usando un subconjunto de datos disponibles de esta.

La Estadística.

La Estadística

Para llevar a cabo un estudio descriptivo, es necesario conocer los conceptos de población y muestra(aleatorio) y sus propiedades, los tipos de variables aleatorias posible de encontrar en la práctica y la forma como se describen: tablas de frecuencias, medidas de tendencia central, medidas de dispersión y medidas de posición. Todos estos conceptos los encontrarán en esta primera unidad del curso.

Población y muestra.

Población y muestra.

La Población(también llamada universo) se definen como el conjunto total de objetos o personas de interés en un estudio. Una característica relevante de la población es que todos sus elementos deben cumplir con un conjunto predefinido de características. El conjunto de características debe permitir entender sin lugar a duda cuál es la población en estudio. Por ejemplo, si un estudio plantea lo siguiente: Se quiere determinar el porcentaje de personas de la ciudad de Valdivia que usa detergente X, se está dando a entender que la población en estudio corresponde a todos los habitantes de la ciudad de Valdivia(los niños usarán el detergente x?).Quizás sería más adecuado plantear: Se quiere terminar el porcentaje de dueña de casa de la ciudad de Valdivia que usa el detergente x.

Población y muestra.

Población y muestra.

En este mismo problema, sí los investigadores contactaran a las dueñas de casa por teléfono para averiguar cuántas usan el detergente x, entonces sería necesario incorporar esta nueva característica (tener teléfono) a nuestra definición de la población.

Población y muestra.

Población y muestra

Lo habitual es que la población esté constituida por un gran número de personas u objetos, por lo que normalmente se hace inviable acceder a todos ellos. El proceso de recopilación de datos de toda la población se denomina censo. Aunque sería atractivo acceder a toda la población, existen varios problemas para llevar a cabo esta idea:

Población y muestra.

Población y muestra

- Un censo usualmente requiere invertir mucho tiempo y recursos, mientras que los estudios en salud se hacen cumpliendo ciertos plazos y con recursos limitados.
- Como las poblaciones son dinámicas, el objeto en estudio puede ser distinto en los primeros individuos estudiados que en los últimos, sobre todo si éstos son vistos mucho tiempo después que los primeros. Por ejemplo, si un investigador quiere determinar la prevalencia de estrés en la octava región, y recopila los datos entre enero y marzo del 2010, sus resultados se verán afectados por la ocurrencia del terremoto del 27 de febrero de este año. En ocasiones no es posible identificar con facilidad a los sujetos que componen la población. Por ejemplo, la población de personas que viven con VIH, o el número de aves acuáticas que existen en una reserva ecológica.

Población y muestra.

Población y muestra

Dados los inconvenientes que se presentan en estudiar una población, lo habitual es que las Investigaciones Científicas se basen en una muestra de la población de interés, es decir, en un subconjunto de los elementos de la población.Para que lo averiguado en la muestra sea cierto para la población en su conjunto, la muestra debe cumplir con los siguientes requisitos:

Población y muestra.

Población y muestra

- la muestra debe ser aleatoria, Esto es, los sujetos en la muestra deben ser escogidos al azar(mediante un sorteo), de modo que todas las personas u objetos de la población tengan una probabilidad mayor que cero de estar presente en las muestra.
- la muestra debe ser de un tamaño mínimo adecuado. Se entenderá por adecuado, si el número de individuos seleccionados al azar de la población(el tamaño de la muestra), permite obtener estimaciones con un margen de error acotado.

Ejemplo: Supongamos que es de interés estimar el porcentaje de fumadores en cierta población, y que el porcentaje real de fumadores es de alrededor del 40 por ciento.Luego , si se quiere estimar con un margen de error de 5 puntos porcentuales, entonces el tamaño de la muestra debiera permitir obtener entre un 35 por ciento y un 45 por ciento de fumadores en la muestra.

Población y muestra.

Población y muestra

-La muestra debe ser representativa de la población de la que procede. Esto se cumple cuando las características de la población relevantes para la investigación, están presentes en la misma la misma proporción o promedio en la muestra. Por ejemplo si la población tiene 30 por ciento de hombres, esta proporción se mantiene la muestra estudiada. Si la edad promedio poblacional es de 50 años, en la muestra se observa aproximadamente lo mismo, etc.

Población y muestra.

Población y muestra

Sin embargo, es posible determinar si efectivamente cada una de las características poblacionales está presente en la misma proporción o promedio en la muestra. En consecuencia, se asume que si una muestra es aleatoria y de tamaño mínimo adecuado, entonces esta es representativa de la población de interés.

Población y muestra.

Población y muestra

La aleatoriedad y el tamaño de una muestra son características que podemos controlar,(el tamaño muestral se puede calcular y el investigador suele escoger, entre varios métodos de selección al azar, el que sea adecuado mejor a su estudio). La representatividad, en cambio, es una cualidad de la muestra.

Parámetros y estimadores

.

Parámetros y estimadores.

Llamaremos inferencia estadística a las deducciones que hacemos acerca de una población de interés, a partir de los resultados obtenidos mediante una muestra aleatoria de dicha población. Por ejemplo, si en una muestra aleatoria se calcula que el promedio de edad es de 20 años, entonces se inferirá que el promedio de edad de la población de la cual procede la muestra debiera ser de aproximadamente 20 años, con un margen de error dado por el tamaño de la muestra, O bien, si se calcula que 38 por ciento de los individuos en la muestra de fumador, entonces se deducirá que el porcentaje de fumadores poblacional debiera ser aproximadamente 38 por ciento..

Parámetros y estimadores

.

Parámetros y estimadores.

El promedio de edad y el porcentaje de fumadores poblacionales se denominan parámetros poblacionales (o simplemente parámetros). En general, un parámetro es cualquier función de los datos calculada en la población. El promedio de edad y el porcentaje de fumadores calculados en la muestra, y utilizados para aproximar el verdadero valor poblacional, se denominan estimadores muestrales o estadísticos. Por lo común , un estimador puede ser cualquier función calculada con los datos muestrales y , como es un valor que representa a la muestra completa, suele llamarse también medida resumen.

Parámetros y estimadores

.

Parámetros y estimadores.

La muestra debiera ser un buen reflejo de la población. De esta forma, el objetivo cuando se estiman parámetros poblacionales es que: El estimador o estadístico sea aproximadamente igual a los parámetros.
Esto solo es posible si la muestra escogida es representativa de la población de interés. Un promedio o una proporción poblacional no es el único parámetro de interés en un estudio. Como puede ser cualquier función de los datos, podría interesar la mediana, la varianza, la desviación estándar, algún percentil u otras funciones menos conocidas.

Parámetros y estimadores

.

Parámetros y estimadores.

Por ejemplo, si se asume que la edad fértil es entre los 15 años y 49 años, según el censo del 2002, el porcentaje de mujeres en edad fértil es de 52 por ciento.Si se quiere hacer un estudio sobre el número de hijos promedio por mujer, en base a una muestra de 400 mujeres de Puente Alto, y se observan 190 mujeres en edad fértil, entonces el 52 por ciento es el parámetro poblacional y el 47,5 por ciento (190 mujeres de un total de 400) es el estimador de ese parámetro.

Variables Aleatorias.

Variables aleatorias.

Una vez que seleccionamos un conjunto de individuos de la población para que formen parte de la muestra aleatoria, cada uno de estos individuos es caracterizado por un conjunto de variables de interés en el estudio.

Variables Aleatorias.

Variables aleatorias.

Se denomina unidad muestral, a cada elemento susceptible de ser seleccionado. Habitualmente, la unidad muestral corresponde a un individuo, aunque no siempre es así. Por ejemplo, en un estudio de contaminación intra domiciliaria la unidad muestral podría ser un hogar,(y no los sujetos que viven en ella). En un estudio en que se interesa analizar el cambio a través de los años en el número de aves acuáticas en una reserva ecológica, la unidad muestral será un número de aves en cada momento del tiempo.

Variables Aleatorias.

Variables aleatorias.

LLamaremos variable a cualquier característica que tome dos o más valores en una población. Por ejemplo la edad, sexo, hábito tabáquico, presencia o ausencia de una patología, valores de colesterol total, triglicéridos en un examen de lípidos, etcétera. Nosotros estudiaremos variables aleatorias, para las cuales no es posible anticipar su resultado, aún cuando se intente controlar los factores que puedan afectarlas. Visto de otra manera, si al mantener constantes las condiciones experimentales no es posible predecir el valor de una variable, entonces se está frente a una variable aleatoria.

Variables aleatorias.

Nótese que si la característica toma solo un valor, entonces es una constante y no es de interés estadístico.Por ejemplo, en el estudio de las dueñas de casa que usan Detergente X, la ciudad de residencia es constante, por lo que no es útil para discriminar entre las mujeres que usan el detergente de las que no lo hacen.

Determinar cuáles variables aleatorias deben ser medidas a cada unidad muestral es de vital importancia para el estudio. Por ejemplo, si interesa investigar factores de riesgo de infarto al miocardio, no puede dejar de medirse la edad, el hábito tabáquico o el consumo de alcohol, ya que todos son factores que se asocian con el fenómeno en estudio.

Variabilidad Muestral.

Cuando tomamos una muestra la aleatoria de una población, lo que hacemos es observar una de muchas posibles muestras aleatorias de la población de interés.

Por ejemplo, si la población está compuesta de 50 individuos y decidimos tomar una muestra de tamaño 5 de ellos, entonces el número de muestras posibles de obtener es 2118760.Aunque el número de muestras posibles puede ser muy grande, en la práctica nosotros solo tenemos acceso a una de ellas.

En consecuencia, sí es de interés calcular el promedio muestral, lo que obtenemos es uno de muchos promedios muestrales posibles de conseguir.

Variabilidad Muestral.

Variabilidad Muestral.

Claramente, si tomamos distintas muestras, el estimador será siempre diferente. Esto es conocido como variabilidad muestral

Tipos de muestreo.

Tipos de muestreo

La selección de una muestra de la población de interés es de vital importancia para obtener resultados válidos. Si la muestra no es representativa de la población de la que procede, todos los cálculos que se hagan serán válidos solo para la muestra , sin posibilidad de extrapolar estos resultados a los individuos que no fueron incluidos en ella.

Tipos de muestreo.

Tipos de muestreo

En general, estaremos interesados en muestras aleatorias, las cuales implican una selección al azar de los individuos que componen la muestra, en alguna etapa del proceso de muestreo. Este tipo de muestreo se denomina muestreo probabilístico.

Tipos de muestreo.

Tipos de muestreo

Los principales tipos de muestreo aleatorio son el muestreo aleatorio simple, el muestreo estratificado y el muestreo sistemático. Además, actualmente adquieren mayor importancia tipos de muestreo más complejos, como el muestreo por conglomerados.

Tipos de muestreo.

Muestreo aleatorio simple.

Es un método selección en que todos los elementos de la población tienen la misma probabilidad de ser elegidos en la muestra. En este tipo de muestreo se asume que la población en estudio es homogénea respecto a la variable que afecta el fenómeno estudiado. Para aplicar este método es necesario tener un registro de todos los objetos poblacionales, por ejemplo, un listado de los rut, del número de ficha clínica, etcétera.

Tipos de muestreo.

Muestreo aleatorio simple.

La selección de los individuos muestrales podría hacerse con métodos tan simples como una bolsa con papeles numeradas o con una tómbola (si la población fuera muy pequeña, hasta el uso de tablas de números aleatorios o la generación de números aleatorios mediante un computador.

Tipos de muestreo.

Muestreo estratificado

Cuando la población es heterogénea respecto a una o más variables que afecten el fenómeno estudiado, seleccionar los datos mediante muestreo aleatorio simple podría resultar en una muestra una representativa de la población. En este caso, las conclusiones derivadas del análisis de los datos serían inválidas. Por ejemplo, si las variables de interés tienen un comportamiento distinto según el no ivel socioeconómico (NSE), un muestreo aleatorio simple podría resultar en una proporción de individuos en cada NSE distinta a la observada en la población y por lo tanto los estimadores serían incorrectos.

Por ello, el investigador puede segmentar la población en estratos, los que corresponden a subconjuntos heterogéneos entre sí, pero que agrupan unidades homogéneas.

Tipos de muestreo.

Muestreo estratificado

El muestreo estratificado es un método de selección en que se obtiene una muestra aleatoria simple de cada estrato por separado y se calculan los estimadores de parámetros(medias, proporciones, etc) para cada estrato.

Tipos de muestreo.

Muestreo estratificado

Finalmente , se calcula un promedio ponderado de los estimadores de los estratos para obtener la medida resumen de interés.
Algunos problemas de investigación en lo que podría ser útil usar muestreo estratificado son los siguientes: - Interesa determinar el gasto promedio en la alimentación de los hogares de cierta ciudad. Como el nivel de gasto es una característica que depende fuertemente del nivel socioeconómico familiar(NSE), conviene hacer estratos de la ciudad agrupando los hogares con niveles socioeconómicos semejantes. Así, la ciudad se podría dividir en zonas de nivel socioeconómico bajo, medio y alto, formando 3 estratos. Al interior de cada estrato se toma una muestra aleatoria simple de hogares y se cuantifica el gasto en alimentación de cada hogar.

Tipos de muestreo.

Muestreo estratificado

- En un muestreo para estimar la cosecha total de café en un país centroamericano, se sabe que la región ecológica donde se ubican los árboles influye mucho en su productividad. Después, sería conveniente estratificar las regiones según altura sobre el nivel del mar, nivel de vientos y temperatura antes de seleccionar los predios y determinar la productividad.

Tipos de muestreo.

Muestreo estratificado

Respecto al número de individuos por seleccionar de cada estrato, existen dos criterios principales:
Asignación proporcional. El número de individuos por seleccionar de cada estrato es proporcional al tamaño poblacional del estrato. Por ejemplo, si el 25 por ciento de los habitantes de cierta ciudad son de nivel socioeconómico bajo, 65 por ciento de nivel medio y 10 por ciento de nivel alto, y se quiere una muestra estratificada de n 120 casos, entonces, usando asignación proporcional, se debieran muestrear 30, 78 y 12 casos de cada NSE, respectivamente.

Tipos de muestreo.

Muestreo estratificado

Asignación óptima. El número de individuos por seleccionar de cada estrato es proporcional a la variabilidad de la característica en estudio al interior del estrato. Por ejemplo, si el gasto en alimentación presenta el doble de variación en el nivel socioeconómico alto que en los niveles medio y bajo, entonces se podría mostrar el doble de los casos del nivel socioeconómico alto que de los otros dos niveles.

Tipos de muestreo.

Muestreo sistemático.

Este método de selección aleatoria es aplicable cuando los elementos de la población están ordenados físicamente y no existe un registro escrito o computacional, que permita hacer una selección por muestreo simple. Por ejemplo, las fichas clínicas de un hospital, ordenadas según fecha de hospitalización en un estante, sería una situación adecuada para usar este tipo de muestreo. Si la población tiene N elementos y se quiere una muestra aleatoria sistemática de n elementos , el procedimiento es el siguiente:

Tipos de muestreo.

Muestreo sistemático.

- Calcular el tamaño del salto sistemático k=N/n. Si k tiene decimal, se utiliza la parte entera del número. - Elegir un número entero al azar, r, entre 1 y k. - Seleccionar de la población ordenada los elementos en la posición. r, r+k, r+2k, ..., r+(n-1)k

Al final del proceso, se obtiene una muestra de n elementos seleccionados sistemáticamente.

Tipos de muestreo.

Muestreo sistemático.

Ejemplo: si tenemos 478 fichas clínicas y necesitamos seleccionar 55 para una encuesta de calidad de atención, tenemos: N=478 (tamaño de la población) n=55 (tamaño de la muestra) k=8.69. Usaremos saltos de 8 unidades.

Tipos de muestreo.

Muestreo por conglomerados.

Cuando es de alto costo realizar un muestreo aleatorio simple o cuando éste último es inaplicable debido a que los individuos que componen la población no están identificados, un muestreo por conglomerados puede ser un método de selección adecuado. Los conglomerados son divisiones de la población en que los elementos al interior de cada uno son heterogéneos, pero existe homogeneidad entre estas agrupaciones. Es decir, se quiere que haya diversidad al interior de cada conglomerado, pero que no importe cuáles conglomerados están presentes en la muestra, ya que entre ellos no hay mucha diferencia.

Tipos de muestreo.

Muestreo por conglomerados.

Esto es opuesto a lo que ocurre con los estratos, ya que aquí interesa que los individuos al interior de cada uno sean homogéneos entre sí y allá heterogeneidad entre los estratos.
El muestreo por conglomerados es un sistema de selección al azar que consta de 2 fases principales: primero se eligen conglomerados al azar y luego se seleccionan elementos al interior de esto mediante un muestreo aleatorio simple.

Tipos de muestreo.

Muestreo por conglomerados.

Ejemplo: Si se quiere una muestra de 600 viviendas de una ciudad, podría ser de alto costo hacer un muestreo aleatorio simple, ya que con seguridad se tendría que recorrer toda la ciudad. Si se toma una muestra por conglomerados, se podrían seleccionar al azar 20 zonas de la ciudad(entendiendo por zona un conjunto de varias manzanas), luego seleccionar 10 manzanas de cada zona y por último 3 viviendas de cada manzana, teniéndose una muestra total de 600 viviendas.

Tipos de muestreo.

Selección con y sin reposición.

Se asume en los tipos de muestreo anteriores que esto se realicen sin reposición. Es decir, Sin devolver el elemento seleccionado a la población, después de ser observado. En este caso, la probabilidad de observar nuevamente el mismo elemento es cero, y la probabilidad de observar cualquier otro elemento se ve afectado por las observaciones anteriores. Un muestreo es con reposición cuando cada elemento seleccionado es devuelto a la población del cual procede después de ser observado. En este caso, la población siempre contiene los mismos elementos, por lo que todos conservan su probabilidad inicial de ser observados.

Tipos de muestreo.

Selección con y sin reposición.

Nótese que, aunque en el muestreo sin reposición se altera la probabilidad de seleccionar un elemento, cuando ya han sido seleccionados otros previamente, si la población es lo suficientemente grande, este probabilidad se puede considerar constante.

Tipos de variables.

Tipos de variables.

Cada variable registrada se puede clasificar en uno de los siguientes tipos :
Variable nominal
Es aquella en que podemos clasificar sus valores en clases o categorías, sin establecer un ordenamiento sugerido por la magnitud de sus valores. Esto significa que los valores con que se identifica cada nivel de la variable son arbitrarios. Por ejemplo, la variable sexo es nominal, ya que podemos identificar sus niveles mediante masculino y femenino; o bien hombre o mujer. Otras variables nominales son: estado civil causa de muerte, ciudad de residencia, tipo de parto, etcétera.

Tipos de variables.

Tipos de variables.

Variable Ordinal.
Es un tipo de variable en la que sus valores o clases se pueden ordenar. Incluye variables con categorías(como gravedad de una enfermedad definida como leve, moderada o severa) y scores(como el test de Apgar del recién nacido). En las Ciencias de la salud, se generan muchas variables ordinales que intentan cuantificar características difíciles o imposibles de medir directamente, como la gravedad cardíaca media usando scores como APACHE O TISS; el desarrollo puberal medido en escala de Tanner, el estado nutricional que se puede definir como bajo peso, normal, sobrepeso y obeso, etc.

Tipos de variables.

Tipos de variables.

Como se observa, las variables ordinales no tienen unidad de medida. Tampoco tiene sentido cuantificar la diferencia o la razón entre 2 valores ordinales. Por ejemplo, si una persona tiene un puntaje de gravedad igual a 30 y otra tiene un puntaje de gravedad igual a 60(asumiendo que mayor puntaje significa mayor gravedad), no podemos decir que la segunda tenga el doble de gravedad que la primera; solo podemos decir que la segunda está más grave.

Tipos de variables.

Tipos de variables.

Variable intervalar.
Es una variable cuantificable de manera objetiva, por lo que posee un orden natural en sus valores y es posible medir las diferencias entre 2 valores. Generalmente tiene unidad de medida.
Una variable intervalar se denomina discreta cuando no puede tomar decimales, como en las variables de conteo(número de hijos, número de caries, días de hospitalización, etc). Se denomina continua cuando toma cualquier valor en un intervalo(como el peso, la talla, el índice de masa corporal, los triglicéridos, etc).