Tema 58

Educación Secundaria

Matemáticas

POBLACIÓN Y MUESTRA. CONDICIONES DE REPRESENTATIVIDAD DE UNA MUESTRA. TIPO DE MUESTREO. TAMAÑO DE UNA MUESTRA

1. INTRODUCCIÓN

1.1. Localización

Este es un tema del bloque de Estadística y Teoría de Probabilidades que incluye básicamente los temas del 57 al 68 del temario. El tema que nos ocupa pertenece a la parte del bloque dedicada a la estadística.

En el pueden distinguirse varias partes:

- Aspectos generales de la Estadística y sus aplicaciones.
- Teoría del muestreo.
- Teoría de la probabilidad.
- Inferencia Estadística.

1.2. Objetivos y contenidos

Los objetivos específicos del tema son :

1. Exponer los aspectos básicos de la teoría del muestreo aleatorio.
2. Introducir los parámetros característicos de poblaciones y muestras.
3. Enunciar e interpretar la ley de los grandes números y el teorema central del límite.
4. Describir la influencia del tamaño muestral en la estimación de parámetros de la población a través del análisis de muestras.

El tema comienza con una descripción del origen de la Estadística como ciencia y su relación con otras ciencias.

En la sección siguiente se aborda el concepto de población, muestra y tamaño muestral haciendo distinciones y marcando las características fundamentales. Se introduce igualmente la definición de censo y marco así como la de muestreo y muestra representativa.

En la cuarta sección se abordan los tipos básicos de muestreo dividiendo primariamente en aleatorios o no aleatorios dando condiciones de aleatoriedad. Se describen, estudian las ventajas y sesgos que pueden tener cada uno de los modelos sean o no aleatorios.

En la última sección se dan las definiciones de los parámetros fundamentales de una población y una muestra: media y desviación típica. A continuación se desarrolla brevemente la teoría inferencial suficiente para introducir dos resultados fundamentales que relacionan a la media poblacional y la media muestral: la ley de los grandes números y el teorema central del límite. Se dan aplicaciones para la inferencia de la media poblacional a partir de la muestral mediante intervalos de confianza y en este mismo sentido, se observa la acotación mínima para el tamaño muestral con una confianza determinada.

1.3. Indicaciones

La estructura global del tema es en resumen la siguiente:

- Concepto de estadística y fenómeno aleatorio.
- Población, muestra y tamaño muestral.
- Concepto de muestreo y consideraciones sobre la representatividad de la muestra.
- Tipos de muestreo.
- Parámetros fundamentales de poblaciones y muestras: media y desviación típica.
- Ley de los grandes números
- Teorema fundamental del límite.
- Aplicaciones para el cálculo de la media poblacional a través de la media muestral.

No hacemos ninguna demostración extensa, aunque se incluyen diversas demostraciones cortas de propiedades. Los requisitos para abordar el estudio del tema son fundamentalmente los conceptos básicos de estadística y probabilidad.

Para una exposición de dos horas de duración podría seguirse la siguiente estrategia:

1. Situar el tema dentro del temario e indicar sus relaciones con otros temas. Indicar los objetivos. Escribir en la pizarra el orden de la exposición sin demasiados detalles.
2. Explicar la doble vertiente histórica que da lugar a la Estadística y algunos de sus matemáticos más importantes.
3. Dar una definición de Estadística.
4. Caracterizar lo que es un fenómeno aleatorio y determinista.
5. Definir el concepto de población e individuo de un fenómeno aleatorio.
6. Definir censo y marco de un fenómeno aleatorio.
7. Definir el concepto de muestreo.
8. Introducir la definición de muestra y tamaño muestral.
9. Dar una explicación sobre lo que se considera muestra representativa.
10. Definir encuesta y el error de muestreo o sesgo.
11. Dar las características del muestreo aleatorio o probabilístico.
12. Dar los seis ejemplos básicos de muestreo aleatorio con sus ventajas y sesgos.
13. Dar las características del muestreo no aleatorio o no probabilístico.
14. Dar los seis ejemplos básicos de muestreo no aleatorio con sus ventajas y sesgos.
15. Definir la media, varianza y desviación típica poblacional.
16. Introducir el concepto de muestra aleatoria simple a partir de una variable aleatoria.
17. Dar brevemente la definición de estimador y la definición de estimador insesgado y consistente.
18. Dar como ejemplos de estimadores la media, varianza y cuaivarianza muestral.
19. Enunciar las propiedades de la media muestral.
20. Enunciar las propiedades de la varianza y cuasi varianza muestral.
21. Introducir y demostrar la ley de los grandes números. Explicar su importancia en relación a la media muestral y poblacional para muestras grandes.

22. Enunciar el teorema central del límite y aplicar para la estimación de la media poblacional mediante intervalos de confianza para tamaños muestrales elevados.
23. Explicar el concepto de valor esperado o esperanza matemática.

Para la realización de una prueba escrita podrá elaborarse un esquema similar al anterior. Sin embargo dado que en general no ser posible desarrollar los contenidos con igual amplitud son aconsejables las siguientes modificaciones:

- Escribir el título del tema destacado del resto al comienzo de la primera hoja del ejercicio.
- Dividir el ejercicio en secciones en una forma similar a la que empleamos en el texto.
- A continuación del título comenzar con la sección de introducción. En esa sección comenzar describiendo con brevedad la situación del tema dentro del temario.
- Dejar un espacio libre a continuación para completar la introducción al final del resto del ejercicio. En ese espacio comentaremos finalmente, también con brevedad, los contenidos que incluimos del tema.
- Desarrollar el tema diferenciando claramente definiciones, listas de propiedades, teoremas y comentarios generales.
- Si no se recuerda con exactitud una demostración (que se considere importante para el desarrollo del tema) deberían incluirse algunos comentarios sobre el enunciado, y uno o dos ejemplos ilustrativos.
- Como norma general podría ser útil a la hora de realizar el examen escrito empezar cada sección en una nueva hoja. De esta forma si hemos eliminado partes del desarrollo del tema será posible, si disponemos de tiempo, completar algunos aspectos antes de entregar el ejercicio.
- Otro aspecto muy importante en el ejercicio escrito es la presentación. Sería recomendable numerar los gráficos y las tablas que se realicen (para poder referirnos a ellos en el texto) y situarlos siempre al mismo lado de la hoja. Dentro de las posibilidades de cada cual, sería también recomendable cuidar la escritura (letra clara, líneas de escritura más o menos paralelas,etc.)

2. CONCEPTOS BÁSICOS SOBRE LA ESTADÍSTICA

2.1. Breve reseña histórica acerca de la Estadística

En nuestros días los métodos estadísticos ocupan un lugar prominente en las distintas ciencias tanto naturales como sociales y constituyen una de las herramientas más utilizadas y apreciadas por los investigadores. La estadística actual es el resultado de la confluencia de dos disciplinas que evolucionaron independientemente hasta unirse en un cuerpo común hacia el siglo XIX:

- El cálculo de probabilidades que nace en el siglo XVII como teoría matemática de los juegos de azar (dados, barajas, lotería, etc.).

- La "estadística" (o ciencia del estado, del latín *Status*) que estudia la descripción de datos y tiene raíces muy antiguas (los primeros censos conocidos se remontan a los chinos, realizados 2.000 años a.J.C.).

La integración de ambas líneas de investigación ha dado lugar a una ciencia experimental interdisciplinar basada en el empleo de modelos matemáticos propios. La estadística proporciona también una metodología para evaluar y juzgar las discrepancias de sus modelos respecto de la realidad.

Los primeros resultados de la estadística matemática se encuentran en los escritos de Bernoulli, Laplace y Poisson.

Los orígenes de la estadística inferencial están ligados a la teoría de errores y al método de ajuste por mínimos cuadrados (Gauss y Markov). La teoría moderna del muestreo, la estimación y el contraste de hipótesis se debe a la escuela anglo-americana (Fisher, Pearson y Neyman).

La estadística es, aparte de un importante objeto de estudio como parte de las matemáticas, una de las más importantes herramientas de las ciencias aplicadas y de la técnica. Los datos numéricos son la información más habitual que extraemos del mundo que nos rodea y la estadística se utiliza siempre que es necesario analizar datos como, por ejemplo, en las ciencias de la naturaleza como la física, la química, la biología, economía, sociología o medicina.

2.2. Definición de Estadística: fenómenos aleatorios

Definición de estadística: podemos definir la estadística como la ciencia de los datos, de la generación de datos interesantes, de su descripción de forma clara y útil y de su interpretación para obtener conclusiones válidas sobre fenómenos aleatorios.

La estadística estudia por tanto fenómenos aleatorios o de azar que se pueden definir como aquellos experimentos caracterizados por las dos propiedades siguientes:

- Presentan notables variaciones en los efectos, de modo que resulta imposible predecir el resultado de una experiencia particular.
- Todos los posibles resultados se conocen de antemano.
- Se verifica la “ley de la estabilidad de las frecuencias” que enuncia:

  “Si se repite el experimento indefinidamente, la frecuencia relativa con la que se presenta un suceso cualquiera de ese experimento tiende a estabilizarse a medida que aumenta el número de repeticiones del mismo”.

Los fenómenos que no son aleatorios se denominan fenómenos deterministas.

Ejemplos de fenómeno aleatorio son el sondeo de intención de voto en unas elecciones, mientras que ejemplo de un fenómeno determinista es el cálculo de la velocidad de un móvil en el vacío.

3. POBLACIÓN, INDIVIDUO Y MUESTRA ALEATORIA DE UN FENÓMENO ALEATORIO

3.1. Población e individuo de un fenómeno aleatorio

Dado un fenómeno aleatorio que se pretende investigar, debemos caracterizar inicialmente dónde y cuántos son los elementos que presentan tal fenómeno. Para ello definimos individuo y población.

Definición de individuo de un fenómeno aleatorio: Los fenómenos aleatorios se estudian sobre elementos concretos que llamaremos individuos. Así pues, entendemos por individuo a cada uno de los elementos donde se presenta el fenómeno aleatorio.

Ejemplo: cada una de las personas con posibilidad de votar en unas elecciones bajo el fenómeno aleatorio "Intención de voto".

Definición de población de un fenómeno aleatorio: llamamos población al conjunto, finito o infinito, de todos los individuos o sistemas bajo el influjo del fenómeno aleatorio. La definición de lo que constituye la población depende del experimentador y de la naturaleza del fenómeno en estudio.

Ejemplo: el conjunto formado por todas las personas que tienen posibilidad de votar en unas elecciones bajo el fenómeno aleatorio "Intención de voto".

Se distinguen los siguientes tipos de poblaciones:

- **Población objetivo o teórica:** está formada por todos los individuos bajo el influjo del fenómeno aleatorio.

- **Población disponible:** es la que resulta tras la depuración de los individuos de la población objetiva no accesibles a priori. Sin embargo, esta no es la población final que manejaremos en muchos casos puesto que pueden existir factores que no conocemos que hagan inaccesibles a otros individuos a priori accesibles.

- **Población investigada:** es la parte realmente accesible de la disponible.

3.2. Censo y marco de un fenómeno aleatorio

Para el estudio del fenómeno aleatorio sobre una población, es necesario inicialmente contrastar el resultado de dicho fenómeno en cada individuo de alguna de las poblaciones definidas anteriormente. A este procedimiento lo denominaremos censo o marco.

- **Definición de censo:** es el estudio del fenómeno aleatorio en todos los elementos de la población objetivo.

- **Definición de marco:** es el estudio del fenómeno, cuando se realiza sobre el total de la población disponible o investigada, y no sobre la población objetivo.

En muchas ocasiones, no es posible realizar un censo o marco por muy diferentes motivos, y entre otros:

- **El estudio puede deteriorar el sistema.** Por ejemplo, la medida de la resistencia de una pieza industrial o la resistencia ante el impacto de un vehículo. Pensemos también en el estudio del efecto de una sustancia medicinal en cada individuo. No podemos correr riesgos ni destruir el sistema o individuo de la población.

- **Es inviable desde un punto de vista económico.** El estudio de todos los individuos de una población, en tamaños grandes, puede constituir un problema costoso en tiempo, dinero, etc., que no estamos dispuestos a admitir.

- **Es inviable desde un punto de vista práctico:** puede ocurrir que a población sea infinita o tan numerosa que excede las posibilidades del estudio.

3.3. Concepto de muestreo: muestra y tamaño muestral, muestra representativa, encuesta y sesgos.

En las situaciones donde no es posible realizar un estudio del influjo del fenómeno aleatorio mediante censos o marcos, se aplica la teoría de muestreo.

Definición de Muestreo: la teoría del muestreo es la encargada de determinar la forma de elegir y las condiciones que debe cumplir un subconjunto de la población bajo el influjo de un fenómeno aleatorio, que denominamos muestra, para que su estudio permita hacer inferencias correctas sobre dicha población.

En estas condiciones definimos muestra en los siguientes términos:

Definición de muestra y tamaño muestral en un fenómeno aleatorio: llamaremos muestra de una población al conjunto de individuos que elegimos en la población para el estudio del fenómeno aleatorio. Al número de individuos elegidos se denomina tamaño de la muestra.

Definición de muestra representativa: es aquella muestra que logra una versión simplificada de la población y reproduce, de algún modo, el mismo comportamiento y características ante el fenómeno aleatorio que esta pero a pequeña escala y tal que su estudio sea viable.

La recogida de los datos sobre los individuos de un censo o muestra se llama **encuesta**.

Es evidente que todo proceso de muestreo lleva asociado un error, debido a la obtención de conclusiones respecto de valoraciones sobre una muestra y no sobre el total de la misma. A esto se denomina error o sesgo del muestreo.

Por otra parte, llamaremos falacia a las inferencias o conclusiones erróneas que podamos realizar a partir de una muestra respecto del estudio de un fenómeno aleatorio sobre una determinada población. La teoría del muestreo trata de impedir y minimizar la aparición de falacias aunque pueden llegar a presentarse incluso con muestras representativas, si las condiciones de estudio no son las adecuadas o están sesgadas intencionadamente.

4. TIPOS DE MUESTREO

Existen diferentes criterios de clasificación de los diferentes tipos de muestreo, aunque en general pueden dividirse en dos grandes grupos:

- Muestreo aleatorio o probabilístico.
- Muestreo no aleatorio o probabilístico.

4.1. Muestreo aleatorio o probabilístico.

Definición de muestreo o muestra aleatorio: aquella muestra que se toma en condiciones de aleatoriedad. La condición de aleatoriedad en esta definición viene determinada por un proceso de selección de los individuos que debe verificar cuanto menos los siguientes requisitos:

- Cada individuo de la población tiene la misma probabilidad de ser elegido en la extracción de cada muestra.
- La población es la misma ante cada extracción de una muestra.

Las consecuencias de la aleatoriedad en cada muestra garantiza la representatividad de la misma respecto de la población estudiada ya que:

- Elimina influencia subjetiva del experimentador en el proceso de selección.
- Compensa la influencia de factores desconocidos relevantes que pueden estar variando en el transcurso del experimento.
- Permite el análisis estadístico de los resultados mediante la teoría de la probabilidad.

Por lo tanto, el muestreo aleatorio asegura la representatividad de la muestra extraída.

Aunque la aleatoriedad absoluta en cada muestra es muy complicada de obtener, en la práctica existen forma útiles para buscar muestras aleatorias, como por ejemplo:

1. Muestreo aleatorio mediante números aleatorios.

Una tabla de números aleatorios es un conjunto ordenado de dígitos generado de tal forma que cada dígito 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 tiene la misma probabilidad de aparecer en cada lugar del conjunto.

Las calculadoras y los ordenadores poseen programas sencillos para generar tablas de números aleatorios. De forma rudimentaria, podemos generar una tabla de números aleatorios poniendo en una bolsa diez etiquetas con los diez dígitos y sacándolas una a una con reposición.

Para crear una muestra aleatoria de tamaño *n* con una tabla de números aleatorios se procede de la forma siguiente:

- Se asigna una etiqueta numérica con la misma longitud a cada elemento de la población. Si hay 100 elementos empezaríamos con 00 y acabaríamos en 99. Si hay 1000 elementos tendríamos que empezar con el 000 y acabar con el 999.
- Se elige un lugar cualquiera en al tabla de números aleatorios y a partir de ese lugar se forman etiquetas numéricas del mismo tamaño que el usado en el paso anterior. Se eliminan aquellas etiquetas que no corresponden a ningún elemento de la población, y también aquellas que se repiten. Los elementos de la población que corresponden a las n primeras etiquetas formadas admisibles constituyen la muestra aleatoriamente extraída.

Entre las ventajas de este tipo de muestreo aleatorio destacamos:

- Sencillo y fácil de realizar.
- Cálculo rápido de medias y varianzas.
- Se basa en la teoría estadística, y por tanto existen paquetes informáticos para analizar los datos.

Entre sus inconvenientes:

- Requiere que se posea de antemano un listado completo de toda la población.
- Cuando se trabaja con muestras pequeñas es posible que no represente a la población adecuadamente y se genera sesgo.

2. Muestreo aleatorio sistemático

Si los elementos de la población están ordenados en una lista elaborada al azar con un número de individuos N y se desea extraer una muestra de tamaño *n* ($n \leq N$), se puede proceder en la forma siguiente: se divide la población en n subconjuntos de los que se extraen un elemento de cada uno de ellos. El modo de proceder para la selección de un elemento de cada conjunto es el siguiente: primero se determina el entero *k* más próximo al cociente N / *n*. Después se elige al azar un elemento n_1 de los *k* primeros de la lista y, por último, se van tomando sucesivamente los elementos $n_1 + k$, $n_1 + 2k$ hasta $n_1 + (n - 1)k$.

Entre las ventajas de este tipo de muestreo aleatorio destacamos:

- Sencillo y fácil comprensión.
- Se basa en la teoría estadística, y por tanto existen paquetes informáticos para analizar los datos.
- No siempre es necesario tener un listado de toda la población sino sólo fragmentado.
- Cuando la población está ordenada siguiendo una tendencia conocida, asegura una representación de cada subconjunto.

Entre sus inconvenientes:

- Si la constante de muestreo *k* guarda relación o depende del fenómeno o incluso el modo de orden de la muestra está relacionado, la muestra puede contener sesgo de selección.

3. Muestreo aleatorio estratificado

Divide a la población en estratos mediante características de homogeneidad de individuos de la población (sexo, profesión, edad, lugar de residencia, etc.). En ese sentido se trata de que todas las características de los individuos de la población estén representadas en la muestra. Dentro de cada estrato, se funciona de manera independiente y se aplica nuevas estratificaciones o se realiza un muestreo aleatorio de otra clase.

La distribución de la muestra en función de los diferentes estratos se denomina afijación, y puede ser de diferentes tipos:

- **Afijación Simple:** cada estrato tiene el mismo número de individuos.
- **Afijación Proporcional:** cada estrato tiene la proporción correcta respecto a la de aparición de la característica en la población.
- **Afijación Óptima:** cada estrato tiene la proporción y dispersión adecuada respecto a la de la aparición de la característica que representa en la población.

Entre las ventajas de este tipo de muestreo aleatorio destacamos:

- Tiende a asegurar que la muestra represente adecuadamente a la población.
- Se obtienen estimaciones más precisas.
- Permite la viabilidad económica.

Entre sus inconvenientes:

- La manera de elección de los estratos puede contener sesgo.
- La afijación puede introducir errores.

4. Muestreo aleatorio por conglomerados

Es un tipo de muestreo en el que inicialmente la población está definida mediante grupos o conglomerados que consideraremos homogéneos. En este sentido para tomar una muestra de tamaño n, consideraremos en tomar mediante muestreo aleatorio el número necesario de conglomerados para asegurarnos que la suma de los individuos de todos ellos sea el tamaño muestral buscado.

Por ejemplo, si necesitamos tomar una muestra mediante conglomerados respecto al estudio sobre la preparación académica de los enfermeras/os respecto a determinadas técnicas, suponiendo que hay n hospitales (conglomerados) y cada hospital tiene m enfermeros/as, trataremos de elegir mediante muestreo aleatorio un número de hospitales de tal modo que sumando las enfermeras/os de tales hospitales obtengamos el tamaño muestral pedido.

Entre las ventajas de este tipo de muestreo aleatorio destacamos:

- Es más fácil y cómodo seleccionar conglomerados que realizar un muestreo aleatorio sobre los propios individuos.
- Permite la viabilidad económica.

Entre sus inconvenientes:

- La manera de elección de los conglomerados puede presentar sesgos cuando dichos conglomerados no presentan condiciones de homogeneidad.

5. Muestreo aleatorio compuesto, polietápico

Es un muestreo aleatorio bastante complejo que trata de obtener la muestra mediante etapas y que puede presentar en cada etapa diferentes tipos de muestreo aleatorio como por ejemplo conglomerados, estratificados, etc.

Es el que se utiliza, por ejemplo, en las encuestas y estudios sobre poblaciones distribuidas en amplias zonas geográficas. Se divide el país en donde se encuentra la población en zonas denominadas unidades primarias de muestreo (UPM). Dentro de cada UPM se forman áreas más pequeñas del orden de 500 habitantes que se denominan distritos del censo (DC).

Para elaborar la encuesta se selecciona una muestra aleatoria de un cierto número de UPM, en cada una de estas unidades se extrae una muestra aleatoria de DC y, finalmente, en cada DC de la muestra aleatoria se extrae una muestra de individuos.

Esta forma de muestreo tiene ventajas obvias. Por un lado no es necesario manejar la lista de todos los individuos del país. Además los individuos elegidos para la encuesta se encuentran distribuidos en grupos concentrados geográficamente en unas pocas regiones, con lo que es más barato el trabajo de los encuestadores.

Entre las ventajas de este tipo de muestreo aleatorio destacamos:

- Tiende a asegurar que la muestra represente adecuadamente a la población.
- Se obtienen estimaciones más precisas.
- Permite la viabilidad económica.

Entre sus inconvenientes:

- La manera de elección de las UPM o en DC puede contener sesgo dependiendo del muestreo aleatorio utilizado.

4.2. Muestreo no aleatorio o no probabilístico

Son aquellos muestreos en los que se incumplen las principales características del muestreo aleatorio en lo que se refiere a la equiprobabilidad de cada individuo de la población en la aparición en la muestra o que la población se mantenga constante. Sin embargo, en muchas ocasiones son utilizados consciente o inconscientemente, puesto que son más sencillos de llevar a cabo y la utilización de muestreos aleatorios suele presentar costes muy elevados o inasumibles.

Hay que tener en cuenta que este tipo de muestreo genera muestras que no tienen porque ser representativas y por lo tanto, no se puede garantizar la fiabilidad acerca de las inferencias y generalizaciones respecto al estudio del fenómeno aleatorio sobre la población mediante estas muestras.

Pese a no haber condiciones de aleatoriedad, estos métodos tienen sus propios criterios para la selección de los individuos de la muestra que procuran de alguna manera la representatividad de la misma.

1. Muestreo no aleatorio de conveniencia

Es un muestreo no aleatorio en el que se procede a seleccionar los individuos que menos esfuerzo presentan por cuestiones económicas, físicas o temporales.

Por ejemplo, si queremos hacer una encuesta en una ciudad, lo más sencillo es ponerse en un lugar concreto y preguntar a la gente que pasa.

Entre las ventajas de este tipo de muestreo no aleatorio destacamos:

- Sencillo y fácil de realizar.
- Presenta costes mínimos en cuanto a tiempo, dinero y esfuerzo.

Entre sus inconvenientes:

- Entra la subjetividad del encuestador o investigador a la hora de elegir a cada individuo. Siguiendo con el ejemplo anterior, el encuestador tenderá a preguntar de forma preferente a personas con aspecto aseado y a ignorar a personas con aspecto menos agradable.
- La localización o condiciones preconcebidas para la selección de los individuos de la muestra pueden sesgar la misma. En el ejemplo anterior, el lugar donde el encuestador está situado puede introducir sesgos pues dependiendo de la localización puede encontrar una cierta respuesta ante el fenómeno más acusada que en otro lugar.

2. Muestreo no aleatorio voluntario o de voluntarismo

Es un muestreo no aleatorio en el que se procede a seleccionar los individuos que deciden por propia iniciativa y voluntariamente ser elementos de la muestra. Además, este tipo de muestreo puede admitir la aparición en la muestra de repeticiones de individuos.

Por ejemplo, los muestreos vía SMS que proponen los programas de televisión y radio son un claro muestreo por voluntarismo.

Entre las ventajas de este tipo de muestreo no aleatorio destacamos:
- Sencillo y fácil de realizar.
- Presenta costes mínimos en cuanto a tiempo, dinero y esfuerzo.

Entre sus inconvenientes:
- La localización o condiciones preconcebidas para la selección de los individuos de la muestra pueden sesgar la misma. En el ejemplo anterior, la cadena o periódico ya están sesgando al tipo de individuo que puede llegar a presentarse voluntario para ser objeto del estudio.
- La aparición de repetición de individuos en la muestra crean inevitablemente sesgos intencionados por el propio elemento de la muestra.

3. Muestreo no aleatorio por cuotas o accidental

Se toma a los individuos más "representativos" de la población como representantes de la población en la muestra. Por lo tanto, utiliza una técnica parecida al muestreo aleatorio por estratos en cuanto a que determina divisiones de la población mediante características pero se diferencia en la elección no aleatoria de los individuos dentro de cada estrato. En este sentido, es muy típico de este muestreo elegir los individuos con una serie de condiciones mediante "cuotas". La elección de los elementos dentro del estrato puede entrar la aleatoriedad o, por el contrario, el voluntarismo.

Por ejemplo, dentro del estrato "mujer" la cuota puede ser la selección de 30 mujeres con edades comprendidas entre los 18 y 25. La selección final de las 30 mujeres puede llevarse a cabo con o sin aleatoriedad.

Entre las ventajas de este tipo de muestreo aleatorio destacamos:
- Tiende a asegurar que la muestra represente adecuadamente a la población.
- Permite la viabilidad económica.

Entre sus inconvenientes:
- La manera de elección de los estratos puede contener sesgo.
- Las distinta cuota de elección en cada estrato puede sesgar a la muestra.
- Existe voluntarismo al tomar los elementos finales para la cuota.

Este método se utiliza mucho en las encuestas de opinión.

4. Muestreo no aleatorio opinático o intencional

Se trata de tomar muestras deliberadas que en ocasiones anteriores se tiene la certeza que han tenido comportamientos similares o parecidos al de la población en el fenómeno aleatorio en estudio. Estas muestras son consideradas como "representativas".

Por ejemplo, en las encuestas preelectorales, algunos estudios se centran en determinados pueblos o barrios que tradicionalmente, su intención de voto ha sido muy similar a la que tuvo la población total de la votación.

Entre las ventajas de este tipo de muestreo aleatorio destacamos:

- Son sencillas y de fácil aplicación.
- Son menos costosos que otro tipo de muestreos.

Entre sus inconvenientes:

- El comportamiento similar entre muestra y población en otras ocasiones no asegura que esto vuelva a ocurrir.
- La muestra puede sesgarse si tiene conocimiento de su "representatividad".
- La muestra puede variar sustancialmente con el tiempo.
- Permite sesgo intencionado a la hora de la elección de la muestra "representativa"

5. Muestreo no aleatorio "Bola de nieve"

La muestra se crea a sí misma a partir de la selección de varios individuos los cuales conducen a otros, y estos a otros, y así hasta conseguir una muestra suficiente.

Entre las ventajas de este tipo de muestreo aleatorio destacamos:

- Son sencillas y de fácil aplicación.
- Permite la viabilidad económica.

Entre sus inconvenientes:

- La elección del primer individuo o de cualquiera de los demás puede sesgar a toda la muestra a partir de él.
- Permite sesgo intencionado a la hora de la elección de los primeros individuos.

Este tipo se emplea muy frecuentemente cuando se hacen estudios con poblaciones "marginales", delincuentes, determinados tipos de enfermos, etc.

6. Muestreo no aleatorio discrecional

El encuestador selecciona intencionadamente los elementos de la muestra en base a los prejuicios o conocimientos del investigador acerca de la representatividad de esos individuos respecto del estudio sobre la población.

Entre las ventajas de este tipo de muestreo aleatorio destacamos:

- Son sencillas y de fácil aplicación.
- En la mayoría de los casos, permite la viabilidad económica.

Entre sus inconvenientes:

- Se tiene que conocer de antemano los estratos representativos de la población.
- Permite sesgo intencionado a la hora de la elección de los individuos.

Este tipo se emplea muy frecuentemente cuando se hacen estudios con poblaciones "marginales", delincuentes, determinados tipos de enfermos, etc.

5. ESTUDIO DE PARÁMETROS EN POBLACIONES Y MUESTRAS

5.1. Media, varianza y desviación típica poblacional

Sea una población cuyos individuos poseen una característica común descrita por una variable estadística cuantitativa X que toma valores diferentes $x_1, x_2, \ldots, x_r$ mediante una distribución de frecuencias relativas $f_1, f_2, \ldots, f_n$.

- Se denomina **media poblacional** al valor media aritmética de todos los valores de la población:

$$\mu_X = M(X) = \sum_i x_i \cdot f_i$$

- Se denomina **varianza poblacional** al valor media aritmética de la función $(X - \mu_X)^2$,

$$\sigma_X^2 = V(X) = M((X - \mu_X)^2) = \sum_i (x_i - \mu_X)^2 \cdot f_i$$

- Se denomina **desviación típica poblacional** a la raíz cuadrada de la varianza poblacional,

$$\sigma_X = \sqrt{V(X)} = \sqrt{\sum_i (x_i - \mu_X)^2 \cdot f_i}$$

Observaciones

1. Los valores o parámetros poblacionales media y desviación típica son esenciales para la descripción y valoración de cada población.

2. En muchas ocasiones el estudio estadístico no se puede efectuar mediante censo y hay que recurrir a muestras para el estudio de la población. En este sentido, uno de los objetivos consistirá en inferir el valor de la media y la desviación poblacionales.

Algunas de las propiedades de los parámetros media poblacional y desviación típica poblacional son:

1. **Si X es una variable aleatoria y λ un número real fijo entonces $\mu_{\lambda \cdot X} = \lambda \cdot \mu_X$.**

Demostración. Por la definición de media aritmética de una muestra, tendremos que:

$$\mu_{\lambda \cdot X} = M(\lambda \cdot X) = \sum_i f_i \cdot (\lambda \cdot x_i) = \lambda \cdot \sum_i f_i \cdot x_i = \lambda \cdot M(X)$$

2. **Dadas las variables X e Y se verifica que $\mu_{X+Y} = \mu_X + \mu_Y$.**

Demostración. Dadas las variables aleatorias X e Y,

$$\mu_{X+Y} = M(X+Y) = \sum_i \sum_j f_{i,j} \cdot (x_i + y_j) = \sum_i \sum_j (f_{i,j} \cdot x_i + f_{i,j} \cdot yx_j) =$$

$$= \sum_i \sum_j f_{i,j} \cdot x_i + \sum_i \sum_j f_{i,j} \cdot y_i = \sum_i f_i \cdot x_i + \sum_j f_j \cdot y_j =$$

$$= M(X) + M(Y) = \mu_X + \mu_Y$$

3. **Si X e Y son dos variables independientes, entonces $\mu_{X \cdot Y} = \mu_X \cdot \mu_Y$**

Demostración. Sean X e Y variables aleatorias independientes, entonces:

$$\mu_{X \cdot Y} M(X \cdot Y) = \sum_{i,j} f_{i,j} \cdot (x_i \cdot y_j)$$

Puesto que las variables X_1 y X_2 son independientes, entonces:

$$f_{i,j} = f_{i,\bullet} \cdot f_{\bullet,j}$$

y por tanto,

$$\mu_{X \cdot Y} = M(X \cdot Y) = \sum_{i,j} f_{i,j} \cdot (x_i \cdot y_j) = \sum_{i,j} f_{i,\bullet} \cdot f_{\bullet,j} \cdot (x_i \cdot y_j) =$$

$$= \left[\sum_i f_{i,\bullet} \cdot x_i \right] \cdot \left[\sum_j f_{\bullet,j} \cdot y_j \right] = M(X) \cdot M(Y) = \mu_X \cdot \mu_Y$$

4. **$\sigma^2_X = M(X^2) - M(X)^2$**

Demostración. Utilizando las propiedades de la media y la propia definición de la varianza:

$$\sigma^2(X) = M((X - M(X))^2) = \sum_i f_i \cdot (x_i - M(X))^2 =$$

$$= \sum_i f_i \cdot x_I^2 + \sum_i f_i \cdot M(X)^2 - 2 \cdot M(X) \cdot \sum_i f_i \cdot x_i =$$

$$= M(X^2) + M(X)^2 \cdot \sum_i f_i - 2 \cdot M(X) \cdot M(X) =$$

$$= M(X^2) + M(X)^2 - 2M(X)^2 = M(X^2) - M(X)^2$$

5. Para todo número real λ se verifica que $\sigma^2_{\lambda \cdot X} = \lambda^2 \cdot \sigma^2_X$

Demostración. Basta ver que:

$$\sigma^2_{\lambda \cdot X} = M((\lambda \cdot X - M(\lambda \cdot X))^2) = M(\lambda^2 \cdot (X - M(X))^2) =$$

$$= \lambda^2 \cdot M((X - M(X))^2) = \lambda^2 \cdot \sigma^2_X$$

6. Si X e Y son variables independientes, entonces $\sigma s^2_{X+Y} = \sigma^2_X + \sigma^2_Y$.

Demostración. Sean X e Y variables aleatorias independientes. Utilizando las propiedades de la media, y teniendo en cuenta que X e Y son independientes, se obtiene:

$$\sigma^2_{X+Y} = M((X+Y)^2) - [M(X+Y)]^2 =$$

$$= M(X^2 + Y^2 + 2X \cdot Y) - M(X)^2 - M(Y)^2 - 2M(X) \cdot M(Y) =$$

$$= M(X^2) + M(Y^2) + 2M(X \cdot Y) - M(X)^2 - M(Y)^2 - 2M(X) \cdot M(Y) =$$

$$= M(X^2) + M(Y^2) + 2M(X) \cdot M(Y) - M(X)^2 - M(Y)^2 - 2M(X) \cdot M(Y) =$$

$$= M(X^2) - M(X)^2 + M(Y^2) - M(Y)^2 = \sigma^2_X + s^2_Y$$

5.2. Muestra aleatoria simple

Definición de muestra aleatoria simple: Sea un fenómeno aleatorio en el que se estudia una característica de una población sobre el que se mide una cierto parámetro asociado a la variable estadística X con determinada distribución de probabilidades. Llamaremos muestra aleatoria simple de la variable aleatoria X a la variable aleatoria n-dimensional:

$$(X_1, X_2, \ldots, X_n)$$

cuyas variables aleatorias unidimensionales X_k con k = 1, 2, . . . n, que la componen son independientes y con la misma distribución de probabilidad (idénticamente distribuidas) que X.

Por tanto, si X es discreta con función de masa p(x), la función de masa conjunta de la muestra aleatoria simple (X_1, X_2, . . . , X_n) será:

$$p(x_1, x_2, \ldots, x_n) = \prod_{k=1}^{n} p(x_k) = p(x_1) \cdot p(x_2) \cdot \ldots \cdot p(x_n)$$

Si X es continua con función de densidad f(x), la función de densidad conjunta de la muestra aleatoria simple (X_1, X_2, . . . , X_n) es:

$$f(x_1, x_2, \ldots, x_n) = \prod_{k=1}^{n} f(x_k) = f(x_1) \cdot f(x_2) \cdot \ldots \cdot f(x_n)$$

Ejemplos

1. Sea (X_1, X_2, . . . , X_n), la muestra aleatoria simple de una variable aleatoria X con distribución de Poisson P(ς) con ς un parámetro desconocido. La distribución de esta muestra aleatoria simple vendrá dada por la función de masa:

$$p(x_1, x_2, \ldots, x_n) = \prod_{k=1}^{n} p(x_k) = \prod_{k=1}^{n} \frac{\lambda^{x_k} \cdot e^{-\lambda}}{(x_k)!} = \frac{\lambda^{\sum_{k=1}^{n} x_k} \cdot e^{-n\lambda}}{\prod_{k=1}^{n} (x_k)!}$$

2. Sea (X_1, X_2, . . . , X_n), la muestra aleatoria simple de una variable aleatoria X con distribución normal N(Ω, α) con Ω y α parámetros desconocidos. La distribución de esta muestra aleatoria simple vendrá dada por la función de densidad:

$$f(x_1, x_2, \ldots, x_n) = \prod_{k=1}^{n} f(x_k) = \prod_{k=1}^{n} \frac{e^{-\frac{(x_k-\mu)^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}} = \frac{e^{-\frac{\sum_{k=1}^{n}(x_k-\mu)^2}{2\sigma^2}}}{\left(\sigma\sqrt{2\pi}\right)^n}$$

5.3. Definición de estimadores: media, varianza y cuaivarianza muestral

En la inferencia estadística se suele plantear el problema de calcular uno o varios parámetros desconocidos θ_k de los que depende la distribución conocida de la variable aleatoria X a partir de la variable aleatoria simple. Muy usualmente estos parámetros son la media y desviación poblacionales.

Para ello definiremos funciones n-dimensionales que estimen de un modo tan aproximado como sea necesario al parámetro desconocido. Por tanto, estimar un parámetro es asignarle un valor obtenido a partir de una medición de una función (estimador) del parámetro sobre una muestra.

Definición de estimador: sea (X_1, X_2, . . . , X_n), una muestra aleatoria simple de una variable aleatoria X con distribución conocida p(θ) y con θ parámetro desconocido. Se llama estimador de un parámetro desconocido θ, a toda función T(X_1, X_2, . . . , X_n) que permita aproximar a dicho parámetro. Se verifica que, como (X_1, X_2, . . . , X_n) es una muestra o variable aleatoria que tiene una función de distribución, el estimador también gozará de los mismos privilegios.

Definición de estimador centrado o insesgado: se dirá que el estimador es centrado o insesgado respecto del parámetro θ si E [T(X_1, X_2, . . . , X_n)] = θ

Esta propiedad indica que el estimador elegido es una buena función para aproximar al parámetro θ.

Definición de estimador consistente: se dirá que el estimador es consistente para el parámetro θ si se verifica que:

$$\lim_{n\to\infty} E[T(X_1, X_2, \ldots, X_n)] = \theta \qquad y \qquad \lim_{n\to\infty} V[T(X_1, X_2, \ldots, X_n)] = 0$$

La elección del estimador adecuado para la inferencia del parámetro desconocido es una cuestión bastante delicada y hay varios procedimientos para encontrarlos de manera centrada.

Algunos de los estimadores más utilizados son:

- **La media muestral:** sirve para estimar el parámetro media Ω si es que este es desconocido para la distribución de X.

$$T(X_1, X_2, \ldots, X_n) = \frac{1}{n} \cdot \sum_{k=1}^{n} X_k$$

Se suele representar mediante $\overline{X}$.

- **La varianza muestral:** sirve para estimar el parámetro varianza α^2 si es que este es desconocido para la distribución de X.

$$T(X_1, X_2, \ldots, X_n) = \frac{1}{n} \cdot \sum_{k=1}^{n} \left(X_k - \overline{X}\right)^2$$

Se suele representar mediante s^2.

- **La cuasivarianza muestral:** sirve para estimar el parámetro varianza α^2 si es que este es desconocido para la distribución de X al igual que la media Ω.

$$T(X_1, X_2, \ldots, X_n) = \frac{1}{n-1} \cdot \sum_{k=1}^{n} \left(X_k - \overline{X}\right)^2$$

Se suele representar mediante S^2.

5.4. Propiedades de la media muestral

1. **La media de la media muestral es la media poblacional.** Si queremos estimar la media Ω_X de una variable estadística X con distribución $p(\Omega_X)$, un candidato natural a estimador es la media muestral que es un estimador centrado ya que, por las propiedades del operador media:

$$M\left[\frac{1}{n}\cdot\sum_{k=1}^{n}X_k\right]=\frac{1}{n}\cdot M\left[\sum_{k=1}^{n}X_k\right]=\frac{1}{n}\cdot\sum_{k=1}^{n}M\left[X_i\right]=\frac{1}{n}\cdot n\cdot\mu_X=\mu_X$$

Por lo tanto, la media muestral es un estimador insesgado.

2. **La varianza de la media muestral es el cociente de la varianza poblacional entre el tamaño muestral.** El estimador media muestral tiene la buena propiedad de que su varianza coincide con $\alpha_X^2/\ n$, con α_X^2 la varianza de la distribución de la variable X y n el tamaño muestral.

$$s^2\left[\frac{1}{n}\cdot\sum_{k=1}^{n}X_k\right]=\frac{1}{n^2}\cdot s^2\left[\sum_{k=1}^{n}X_k\right]=\frac{1}{n^2}\cdot\sum_{k=1}^{n}s^2\left[X_i\right]=\frac{1}{n^2}\cdot n\cdot\sigma_X^2=\frac{\sigma_X^2}{n}$$

Concluimos entonces que la media muestral es un estimador consistente para la media poblacional, ya que al hacer tender el tamaño muestral a valores grandes, la varianza tiende a cero.

Esta propiedad hace de la media muestral sea un gran estimador para la media Ω ya que "*cuanto más grande sea el tamaño muestral de las muestras usadas, menos dispersa será la media muestral"*. Es decir, los datos de la medición de la media muestral estarán tanto más localizados cuanto más grande sea el tamaño de las muestras que empleamos. Por tanto, según lo dicho antes, el estimador media muestral es consistente.

5.5. Propiedades de la varianza y cuasi varianza muestral

1. La media de la cuasi varianza muestral es la varianza poblacional. La varianza muestral y cuasivarianza muestral proporcionan estimaciones de la varianza de la población. Como podemos observar mediante los siguientes cálculos:

Si calculamos la media de la varianza muestral,

$$M\left[s^2\right]=M\left[\frac{1}{n}\cdot\sum_{k=1}^{n}\left(X_k-\overline{X}\right)^2\right]=\frac{1}{n}\cdot M\left[\sum_{k=1}^{n}X_k^2+\sum_{k=1}^{n}\overline{X}^2-2\sum_{k=1}^{n}X_k\cdot\overline{X}\right]=$$

$$=\frac{1}{n}\cdot M\left[\sum_{k=1}^{n}X_k^2+n\cdot\overline{X}^2-2\cdot\overline{X}\cdot\sum_{k=1}^{n}X_k\right]=\frac{1}{n}\cdot M\left[\sum_{k=1}^{n}X_k^2+n\cdot\overline{X}^2-2\cdot\overline{X}\cdot n\cdot\overline{X}\right]=$$

$$= \frac{1}{n} \cdot M\left[\sum_{k=1}^{n} X_k^2 + n \cdot \overline{X} - 2n \cdot \overline{X}^2\right] = \frac{1}{n} \cdot M\left[\sum_{k=1}^{n} X_k^2 - n \cdot \overline{X}^2\right] = \frac{1}{n} \cdot \sum_{k=1}^{n} M(X_k^2) - M(\overline{X}^2) =$$

$$= \frac{1}{n} \cdot n \cdot M(X^2) - M(\overline{X}^2) = M(X^2) - M(\overline{X}^2)$$

Como tenemos que $s^2(X) = M(X^2) - M(X)^2$, entonces, para uno de los elementos de la resta ocurrirá:

$$M\left(X^2\right) = s^2\left(X\right) + M\left(X\right)^2 = \sigma_X^2 + \mu_X^2$$

$$M\left(\overline{X}^2\right) = s^2\left(\overline{X}\right) + M\left(\overline{X}\right)^2 = \frac{\sigma_X^2}{n} + \mu_X^2$$

Y por tanto, sustituyendo tendremos:

$$M\left[s^2\right] = M(X^2) - M(\overline{X}^2) = \sigma_X^2 + \mu_X^2 - \frac{\sigma_X^2}{n} - \mu_X^2 = \frac{n-1}{n} \cdot \sigma_X^2$$

Obsérvese que la media de la varianza muestral no coincide con la varianza de la población. Ello motiva la introducción de la cuasivarianza muestral, ya que como

$$S^2 = \frac{1}{n-1} \cdot \sum_{k=1}^{n} \left(x_k - \overline{X}\right)^2 = \frac{n}{n-1} \cdot \left[\frac{1}{n} \cdot \sum_{k=1}^{n} \left(x_k - \overline{X}\right)^2\right] = \frac{n}{n-1} \cdot s_X^2$$

Entonces, teniendo en cuenta el cálculo de la media de la varianza muestra, se obtiene

$$M\left[S^2\right] = M\left[\frac{n}{n-1} \cdot s^2\right] = \frac{n}{n-1} M(s^2) = \frac{n}{n-1} \cdot \frac{n-1}{n} \sigma_X^2 = \sigma_X^2$$

Es decir, en el lenguaje de la estadística inferencial, la cuasivarianza muestral es un estimador insesgado de la varianza de la población.

5.6. Relación entre la media muestral y poblacional: la ley de los grandes números

Entre los resultados más importantes en la estadística están la ley de los grandes números y el teorema central del límite. Ambos tratan sobre la media muestral.

Lo que se deduce de estos resultados es que la media muestral es una variable estadística tanto más localizada alrededor de su media cuanto más grande sea el número n de variables que la forman. En el límite de grandes valores de n solo hay probabilidad de encontrar la media muestral concentrada en su valor medio.

El teorema de Chebyshev, es el germen de la demostración de la ley de los grandes números que veremos a continuación. El enunciado del teorema de Chebyshev dice lo siguiente:

Teorema de Chebyshev. Sea una variable aleatoria X con función de probabilidad asociada P_x. Sean μ_X y σ_X la media y desviación típica poblacionales respectivamente. Se verifica que,

$$P_x(|X-\mu_x| \le k\cdot\sigma_x) \ge 1-\frac{1}{k^2} \qquad \forall k>0$$

Enunciamos ahora la ley de los grandes números. Este resultado es consecuencia fundamental de la desigualdad de Chebyshef y juega un importante papel en la teoría de probabilidades y estadística relacionando la media muestral y poblacional de una variable aleatoria X. En particular:

- Proporciona una interpretación de la media como un valor esperado.
- Proporciona en el caso binomial una formulación del concepto de probabilidad en términos de límites de frecuencias.

Ley de los grandes números: dada una variable aleatoria simple (X_1, X_2, . . . , X_n) de una variable estadística X con distribución conocida p(θ) y con θ parámetro desconocido. Se verifica que

$$\lim_{n\to\infty} P_x\left(\left|\overline{X}-\mu_x\right| \le \varepsilon\right)=1$$

Demostración

Dada la variable aleatoria simple (X_1, X_2, . . . , X_n) de una variable estadística X con distribución conocida y teniendo en cuenta que la v.a $\overline{X}$ tiene media y desviación:

$$E\left[\overline{X}\right]=\mu_x \qquad s\left[\overline{X}\right]=\frac{\sigma_x}{\sqrt{n}}$$

Podemos aplicar a la media muestral el teorema de Chebyshev de tal modo que

$$P_x\left(|\overline{X}-\mu_x| \le k\cdot\frac{\sigma_x}{\sqrt{n}}\right) \ge 1-\frac{1}{k^2} \qquad \forall k>0$$

Sea un valor cualquiera ε > 0, procedemos a calcular la tendencia de probabilidad de que la diferencia entre la media muestral y la poblacional sea menor o igual que el valor ε elegido.

$$\lim_{n\to\infty} P_x\left(\left|\overline{X}-\mu_x\right| \le \varepsilon\right)$$

En la desigualdad de Chevyshev, sea $k=\frac{\varepsilon\cdot\sqrt{n}}{\sigma_x}>0$, en ese caso tendremos que:

$$P_x\left(|\overline{X}-\mu_x| \le \frac{\varepsilon\cdot\sqrt{n}}{\sigma_x}\cdot\frac{\sigma_x}{\sqrt{n}}\right) \ge 1-\frac{1}{\left(\frac{\cdot\varepsilon\sqrt{n}}{\sigma_x}\right)^2}$$

Simplificando,

$$P_x\left(|\overline{X}-\mu_x| \le \varepsilon\right) \ge 1-\frac{\sigma_x^2}{n\cdot\varepsilon^2}$$

Tomando límite en n entonces:

$$\lim_{n\to\infty} P_x\left(\mid \overline{X} - \mu_x \mid \leq \varepsilon\right) \geq \lim_{n\to\infty}\left[1 - \frac{\sigma_x^2}{n\cdot\varepsilon^2}\right] = 1$$

Y concluimos

$$\lim_{n\to\infty} P_x\left(\left|\overline{X} - \mu_x\right| \leq \varepsilon\right) = 1$$

Observaciones

1. La ley de los grandes números explica que la media muestral es una variable aleatoria que, bajo cualquier distribución, tanto más cercana es a la media poblacional de la variable X de la que procede al aumentar el número de variables que la forman.

2. En el caso de que la v.a. X tenga distribución Bernuilli de parámetro p, sabemos que $\mu_x = p$ y $\sigma_x = p\cdot(1 - p)$. En ese caso, la ley de los grandes números

$$\lim_{n\to\infty} P_x\left(\left|\overline{X} - p\right| \leq \varepsilon\right) = 1$$

   Es decir, que cuantas más repeticiones o experimentos de Bernouilli independientes e idénticos realicemos, más aproximación existe entre la media muestral o proporción de éxitos y el valor poblacional p.

3. El valor medio o media aritmética de una variable aleatoria M(X) o E(X) recibe el nombre a menudo de valor esperado o esperanza matemática de la v.a. X. La razón de esta denominación está en la ley de los grandes números ya que cuando se repite un número grande de veces la medición de la v.a. X, la media aritmética en cada caso es una media muestral que, de acuerdo con la ley de los grandes números será con gran probabilidad muy cercana al valor de la media poblacional μ_x.

4. La ley de los grandes números da base para un posible modelo de definición de la probabilidad.

5.7. El teorema central del límite: estimación de la media poblacional a partir de intervalos de confianza para tamaños muestrales elevados

La estimación o inferencia mediante intervalos de confianza de la media poblacional a partir de la media muestral se basa inicialmente en el teorema central del límite que enunciamos a continuación:

Teorema central del límite: sea X una variable estadística. Dado un intervalo I < ♣ cualquiera de la recta. Para cada n natural sea $p_n(\overline{X} \in I)$, la probabilidad de que la media muestral esté en el intervalo I. En este caso,

$$\lim_{n\to\infty} p_n(\overline{X} \in I) \approx \int_I N(x, \mu, \frac{\sigma}{\sqrt{n}})dx$$

Observaciones

1. Lo que el teorema expresa es un resultado general que nos dice que para tamaños muestrales suficientemente grandes, la media muestral es una variable estadística con distribución normal de media, la media poblacional y de desviación, el cociente de la desviación poblacional entre la raíz cuadrada del tamaño muestral.

$$\overline{X} \approx N\left(\mu_X, \frac{\sigma_X}{\sqrt{n}}\right)$$

Veamos ahora la aplicación directa de este teorema para la inferencia de la media poblacional a partir de la media muestral utilizando intervalos de confianza.

- **Si dada una variable estadística X con distribución de desviación típica conocida σ_X,** deseamos formar un intervalo de confianza ($\overline{X} - \varepsilon$, $\overline{X} + \varepsilon$) para la media poblacional para tamaños muestrales suficientemente grandes entonces el teorema central del límite asegura que:

$$\lim_{n\to\infty} p_n(\mu_X - \varepsilon < \overline{X} < \mu_X + \varepsilon) \approx \int_I N(x, \mu, \frac{\sigma}{\sqrt{n}})dx$$

O lo que es lo mismo,

$$\lim_{n\to\infty} p_n(-\varepsilon < \overline{X} - \mu_X < +\varepsilon) \approx \int_I N(x, \mu, \frac{\sigma}{\sqrt{n}})dx$$

Si condicionamos a que la probabilidad de que la media muestral y la media poblacional difieran en ε con tamaño muestral lo suficientemente grande y sea de confianza 1 – α entonces podemos tipificar según el procedimiento para distribuciones normales de tal modo que:

$$\lim_{n\to\infty} p_n(-\varepsilon < \overline{X} - \mu_X < +\varepsilon) = 1-\alpha$$

$$\Leftrightarrow \quad \lim_{n\to\infty} p_n\left(\frac{-\varepsilon}{\sigma_X/\sqrt{n}} < \frac{\overline{X} - \mu_X}{\sigma_X/\sqrt{n}} < \frac{+\varepsilon}{\sigma_X/\sqrt{n}}\right) \approx 1-\alpha \quad \Leftrightarrow$$

$$\Leftrightarrow \quad \frac{\varepsilon}{\sigma_X/\sqrt{n}} = z_{\alpha/2}$$

En tal caso,

$$\varepsilon = z_{\alpha/2} \cdot \frac{\sigma_X}{\sqrt{n}}$$

Y el intervalo deseado será

$$\left(\mu_X - z_{\alpha/2} \cdot \frac{\sigma_X}{\sqrt{n}} \quad , \mu_X + z_{\alpha/2} \cdot \frac{\sigma_X}{\sqrt{n}}\right)$$

De este modo, para inferir acerca de la media poblacional μ_X podemos utilizar de nuevo la media muestral ya que con una confianza 1 – α, la media poblacional se encontrará, para muestras suficientemente grandes, en el intervalo

$$\left(\overline{X} - z_{\alpha/2} \cdot \frac{\sigma_X}{\sqrt{n}} \quad , \overline{X} + z_{\alpha/2} \cdot \frac{\sigma_X}{\sqrt{n}} \right)$$

- **Si dada una variable estadística X con distribución de desviación típica desconocida σ_X**, deseamos formar un intervalo de confianza ($\overline{X}$ – ε, $\overline{X}$ + ε) para la media poblacional con tamaños muestrales suficientemente grandes entonces podemos reemplazar la desviación típica (que no conocemos) por la cuasivarianza muestral (estimador insesgado respecto a la varianza) y utilizar el resultado siguiente

$$\frac{\overline{X} - \mu_X}{S_n / \sqrt{n}} \approx t_{n-1}$$

Es decir, la variable aleatoria,

$$\frac{\overline{X} - \mu_X}{S_n / \sqrt{n}}$$

se distribuye según una t de Student con n – 1 grados de libertad., a la que denotamos por t_{n-1}.

Si condicionamos a que la probabilidad de que la media muestral y la poblacional difieran en ε con tamaño muestral lo suficientemente grande con un nivel de confianza 1 – α entonces podemos seguir un procedimiento similar al anteriormente expuesto de tal modo que:

$$\lim_{n\to\infty} p_n(-\varepsilon < \overline{X} - \mu_X < +\varepsilon) = 1 - \alpha$$

$$\Leftrightarrow \quad \lim_{n\to\infty} p_n\left(\frac{-\varepsilon}{S_n/\sqrt{n}} < \frac{\overline{X} - \mu_X}{S_n/\sqrt{n}} < \frac{+\varepsilon}{S_n/\sqrt{n}} \right) \approx 1 - \alpha \quad \Leftrightarrow$$

$$\Leftrightarrow \quad \lim_{n\to\infty} p_n\left(\frac{-\varepsilon}{S_n/\sqrt{n}} < t_{n-1} < \frac{+\varepsilon}{S_n/\sqrt{n}} \right) \approx 1 - \alpha \quad \Leftrightarrow$$

$$\Leftrightarrow \quad \frac{\varepsilon}{S_n/\sqrt{n}} = t_{n-1,\alpha/2}$$

En tal caso,

$$\varepsilon = t_{n-1,\alpha/2} \cdot \frac{S_n}{\sqrt{n}}$$

Y el intervalo deseado será

$$\left(\mu_X - t_{n-1,\,\alpha/2} \cdot \frac{S_n}{\sqrt{n}} \;\; , \mu_X + t_{n-1,\,\alpha/2} \cdot \frac{S_n}{\sqrt{n}} \right)$$

Concluimos que para inferir acerca de la media poblacional μ_X podemos utilizar de nuevo la media muestral ya que con probabilidad $1 - \alpha$ la media poblacional se encontrará, para muestras suficientemente grandes, en el intervalo

$$\left(\overline{X} - t_{n-1,\,\alpha/2} \cdot \frac{S_n}{\sqrt{n}} \;\; , \overline{X} + t_{n-1,\,\alpha/2} \cdot \frac{S_n}{\sqrt{n}} \right)$$

Observaciones

1. Este procedimiento nos determina una cota inferior para el tamaño muestral ya que dada la desviación conocida σ_X, un valor $\varepsilon > 0$ cualquiera y una confianza $1 - \alpha$ establecida de que la media poblacional y la muestral se diferencian como máximo en ε, el menor natural n que cumple es tal que,

$$\varepsilon \geq z_{\alpha/2} \cdot \frac{\sigma_X}{\sqrt{n}} \;\Leftrightarrow\; \sqrt{n} \geq z_{\alpha/2} \frac{\sigma_x}{\varepsilon} \;\Leftrightarrow\; n \geq \left(z_{\alpha/2} \frac{\sigma_x}{\varepsilon} \right)^2$$

2. Por un procedimiento análogo, si la desviación es desconocida, dado un valor $\varepsilon > 0$ cualquiera y una confianza $1 - \alpha$ establecida de que la media poblacional y la muestral se diferencian como máximo en ε, el menor natural n que cumple es tal que,

$$\varepsilon \geq t_{n-1,\,\alpha/2} \cdot \frac{S_n}{\sqrt{n}} \;\Leftrightarrow\; \sqrt{n} \geq t_{n-1\,\alpha/2} \cdot \frac{S_n}{\varepsilon} \;\Leftrightarrow\; n \geq \left(t_{n-1,\,\alpha/2} \cdot \frac{S_n}{\varepsilon} \right)^2$$

5.8. El concepto de valor esperado o esperanza matemática

El valor medio o media poblacional $\mu_X = M(X)$ de una variable estadística recibe a menudo el nombre de valor esperado o esperanza matemática de la variable X.

La razón de estas denominaciones radica en los resultados de los teoremas de la ley de los grandes números y central del límite. La cuestión es que cuando uno repite la medición de una variable cualquiera X un número grande de veces, el valor medio de los resultados es una media muestral de tamaño grande y por tanto de acuerdo con los teoremas anteriores ese valor medido será con gran probabilidad muy cercano al valor medio o media poblacional $\mu = M(X)$ de X.

Ejemplo. Muchos automóviles actuales poseen un indicador de consumo instantáneo X y otro de consumo medio M(X). La diferencia entre ambos es que X proporciona el consumo que se va produciendo al cabo de cada cierto tiempo (unos cuantos segundos), en cambio M(X) proporciona la media de todas las medidas de X desde la última puesta a cero del contador de M(X). Se observa que la indicación de X varía constantemente, dependiendo del tipo de lugar por el que se conduzca, el estado del tráfico, etc. Sin embargo el de M(X) es muy estable al

cabo de un cierto tiempo. Lo que ocurre es que M(X) es una media muestral y cuanto más tiempo transcurre más grande va siendo el tamaño de la muestra que emplea, con lo que se entra en un régimen en que se aprecia la ley de los grandes números.

Ejemplo. Supongamos un juego en que el premio es una cantidad a si se gana y tal que la probabilidad de ganar es p. Sea X la variable premio obtenido. Su valor medio es

$$\mu = M(X) = p \cdot a + (1 - p) \cdot 0 = p \cdot a.$$

Si jugamos un número grande de veces n y llamamos $X_1, \ldots, X_n$ las correspondientes variables premios obtenidos en las n jugadas, el premio total obtenido será el valor de la suma $X_1 + \ldots + X_n$ que es $n \cdot M(X)$. De acuerdo con los teoremas anteriores, ese número, lo que esperamos ganar, será con gran probabilidad muy cercano a $n \cdot \mu$.

Es decir con gran probabilidad el premio medio que esperamos obtener por jugada será muy cercano a $\mu = p \cdot a$ y este ha de ser el precio por participar en el juego si tal juego es justo.

BIBLIOGRAFÍA

[1] E. W. Weisstein, CRC Concise *Encyclopedia of Mathematics*, Chapman Hall, 1999.

[2] D. Peña Sánchez de Rivera: *Estadística, modelos y métodos*. Alianza Universal, 2000

[3] *Matemáticas Aplicadas a las ciencias sociales II*, Edelvives, 2016.

[4] *Matemáticas I Bachillerato LOGSE*, McGraw-Hill, 1995 .

[5] *Estadística*. D. Freedman, R. Pisani, R. Purves y A. Adhikari, Antoni Bosch Editor, 1993.

[6] *Statistics, Principles and Methods*, R. A. Johnson and G. K. Bhattacharyya, Wiley and Sons, Inc., 2014.

[7] *Estadística aplicada: Conceptos básicos*. Alfonso García Pérez. Editorial UNED. 2008.

[8] Carrasco JL. *El método estadístico en la investigación médica*. Madrid. Editorial Ciencia. 1983.

[9] Hulley SB, Cummings SR. *Diseño de la investigación clínica*. Ed Doyma. Barcelona 1993.

[10] Kelsey IL, Thompson WD, Evans A. *Methods in observational epidemiology*. New York. Oxford University Press 1986.

ESQUEMA-RESUMEN

2. CONSIDERACIONES GENERALES ACERCA DE LA ESTADÍSTICA

2.1.Breve reseña histórica acerca de la Estadística

En nuestros días los métodos estadísticos ocupan un lugar prominente en las distintas ciencias tanto naturales como sociales y constituyen una de las herramientas más utilizadas y apreciadas por los investigadores. La estadística actual es el resultado de la confluencia de dos disciplinas que evolucionaron independientemente hasta unirse en un cuerpo común hacia el siglo XIX:

- El cálculo de probabilidades que nace en el siglo XVII como teoría matemática de los juegos de azar (dados, barajas, lotería, etc.).
- La "estadística" (o ciencia del estado, del latín *Status*) que estudia la descripción de datos y tiene raíces muy antiguas (los primeros censos conocidos se remontan a los chinos, realizados 2.000 años a.J.C.).

La integración de ambas líneas de investigación ha dado lugar a una ciencia experimental interdisciplinar basada en el empleo de modelos matemáticos propios. La estadística proporciona también una metodología para evaluar y juzgar las discrepancias de sus modelos respecto de la realidad.

Algunos de los más insignes matemáticos que han trabajado en esta rama son los Bernoulli, Laplace, Poisson, Gauss, Markov, Fisher, Pearson y Neyman.

2.2.Definición de Estadística. Fenómenos aleatorios

Definición de estadística: podemos definir la estadística como la ciencia de los datos, de la generación de datos interesantes, de su descripción de forma clara y útil, y de su interpretación para obtener conclusiones válidas sobre fenómenos aleatorios.

La estadística estudia por tanto fenómenos aleatorios o de azar que se pueden definir como aquellos experimentos caracterizados por las dos propiedades siguientes:

- Presentan notables variaciones en los efectos, de modo que resulta imposible predecir el resultado de una experiencia particular.
- Todos los posibles resultados se conocen de antemano.
- Se verifica la “ley de la estabilidad de las frecuencias” que enuncia:

  “Si se repite el experimento indefinidamente, la frecuencia relativa con la que se presenta un suceso cualquiera de ese experimento tiende a estabilizarse a medida que aumenta el número de repeticiones del mismo”.

Los fenómenos que no son aleatorios se denominan fenómenos deterministas.

Ejemplos de fenómeno aleatorio son el sondeo de intención de voto en unas elecciones, mientras que ejemplo de un fenómeno determinista es el cálculo de la velocidad de un móvil en el vacío.

3. POBLACIÓN, INDIVIDUO Y MUESTRA ALEATORIA DE UN FENÓMENO ALEATORIO

3.1. Población e individuo de un fenómeno aleatorio

Definición de individuo de un fenómeno aleatorio: los fenómenos aleatorios se estudian sobre elementos concretos que llamaremos individuos. Así pues, entendemos por individuo a cada uno de los elementos donde se presenta el fenómeno aleatorio.

Definición de población de un fenómeno aleatorio: llamamos población al conjunto, finito o infinito, de todos los individuos o sistemas bajo el influjo del fenómeno aleatorio. La definición de lo que constituye la población depende del experimentador y de la naturaleza del fenómeno en estudio.

Se distinguen los siguientes tipos de poblaciones:

- **Población objetivo o teórica:** está formada por todos los individuos bajo el influjo del fenómeno aleatorio.
- **Población disponible:** es la que resulta tras la depuración de los individuos de la población objetiva no accesibles a priori. Sin embargo, esta no es la población final que manejaremos en muchos casos puesto que pueden existir factores que no conocemos que hagan inaccesibles a otros individuos a priori accesibles.
- **Población investigada:** es la parte realmente accesible de la disponible.

3.2. Censo y marco de un fenómeno aleatorio

- **Definición de Censo:** es el estudio del fenómeno aleatorio en todos los elementos de la población objetivo.
- **Definición de marco:** es el estudio del fenómeno, cuando se realiza sobre el total de la población disponible o investigada, y no sobre la población objetivo.

En muchas ocasiones, no es posible realizar un censo o marco por muy diferentes motivos, y entre otros:

- **El estudio puede deteriorar el sistema.** Por ejemplo, la medida de la resistencia de una pieza industrial o la resistencia ante el impacto de un vehículo. Pensemos también en el estudio del efecto de una sustancia medicinal en cada individuo. No podemos correr riesgos ni destruir el sistema o individuo de la población.

- **Es inviable desde un punto de vista económico.** El estudio de todos los individuos de una población, en tamaños grandes, puede constituir un problema costoso en tiempo, dinero, etc. Que no estamos dispuestos a admitir.
- **Es inviable desde un punto de vista práctico: p**uede ocurrir que a población sea infinita o tan numerosa que excede las posibilidades del estudio.

3.3. Concepto de muestreo: muestra y tamaño muestral, muestra representativa, encuesta y sesgos

Definición de Muestreo: la teoría del muestreo es la encargada de determinar la forma de elegir y las condiciones que debe cumplir un subconjunto de la población bajo el influjo de un fenómeno aleatorio, que denominamos muestra, para que su estudio permita hacer inferencias correctas sobre dicha población.

Definición de muestra y tamaño muestral en un fenómeno aleatorio: llamaremos muestra de una población al conjunto de individuos que elegimos en la población para el estudio del fenómeno aleatorio. Al número de individuos elegidos se denomina tamaño de la muestra.

Definición de muestra representativa: es aquella muestra que logra una versión simplificada de la población, y reproduce, de algún modo, el mismo comportamiento y características ante el fenómeno aleatorio que esta pero a pequeña escala y tal que su estudio sea viable.

La recogida de los datos sobre los individuos de un censo o muestra se llama **encuesta**.

Es evidente que todo proceso de muestreo lleva asociado un error, debido a la obtención de conclusiones respecto de valoraciones sobre una muestra y no sobre el total de la misma. A esto se denomina error o sesgo del muestreo.

Por otra parte, llamaremos falacia a las inferencias o conclusiones erróneas que podamos realizar a partir de una muestra respecto del estudio de un fenómeno aleatorio sobre una determinada población. La teoría del muestreo trata de impedir y minimizar la aparición de falacias aunque pueden llegar a presentarse incluso con muestras representativas, si las condiciones de estudio no son las adecuadas o están sesgadas intencionadamente.

4. TIPOS DE MUESTREO

Existen diferentes criterios de clasificación de los diferentes tipos de muestreo, aunque en general pueden dividirse en dos grandes grupos:

- Muestreo aleatorio o probabilístico.
- Muestreo no aleatorio o probabilístico.

4.1. Muestreo aleatorio o probabilístico

Definición de muestreo o muestra aleatorio: es aquella muestra que se toma en condiciones de aleatoriedad. La condición de aleatoriedad en esta definición viene determinada por un proceso de selección de los individuos que debe verificar cuanto menos los siguientes requisitos:

- Cada individuo de la población tiene la misma probabilidad de ser elegido en la extracción de cada muestra.
- La población es la misma ante cada extracción de una muestra.

Las consecuencias de la aleatoriedad en cada muestra garantiza la representatividad de la misma respecto de la población estudiada ya que,

- Elimina influencia subjetiva del experimentador en el proceso de selección.
- Compensa la influencia de factores desconocidos relevantes que pueden estar variando en el transcurso del experimento.
- Permite el análisis estadístico de los resultados mediante la teoría de la probabilidad.

Por lo tanto, el muestreo aleatorio **asegura la representatividad** de la muestra extraída.

Aunque la aleatoriedad absoluta en cada muestra es muy complicada de obtener, en la práctica existen forma útiles para buscar muestras aleatorias, como por ejemplo:

Aunque la aleatoriedad absoluta en cada muestra es muy complicada de obtener, en la práctica existen forma útiles para buscar dicha aleatoriedad, como por ejemplo:

1. **Empleo de números aleatorios.** Una tabla de números aleatorios es un conjunto ordenado de dígitos generado de tal forma que cada dígito 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 tiene la misma probabilidad de aparecer en cada lugar del conjunto. Las calculadoras y los ordenadores poseen programas sencillos para generar tablas de números aleatorios.
2. **Muestreo sistemático.** Si los elementos de la población están ordenados en una lista elaborada al azar con un número de individuos N y se desea extraer una muestra de tamaño n ($n \leq N$), se puede proceder en la forma siguiente: Se divide la población en n subconjuntos de los que se extraen un elemento de cada uno de ellos. El modo de proceder para la selección de un elemento de cada conjunto es el siguiente: primero se determina el entero k más próximo al cociente N / n. Después se elige al azar un elemento n_1 de los k primeros de la lista y, por último, se van tomando sucesivamente los elementos $n_1 + k$, $n_1 + 2k$ hasta $n_1 + (n - 1)k$.
3. **Muestreo aleatorio estratificado.** Divide a la población en estratos mediante características de homogeneidad de individuos de la población (sexo, profesión, edad, lugar de residencia, etc.). Dentro de cada estrato, se funciona de manera independiente y se aplica nuevas estratificaciones o se realiza un muestreo aleatorio de otra clase. La distribución de la muestra en función de los diferentes estratos se denomina afijación, y puede ser de diferentes tipos:

- **Afijación Simple:** cada estrato tiene el mismo número de individuos.
- **Afijación Proporcional:** cada estrato tiene la proporción correcta respecto a la de aparición de la característica en la población.
- **Afijación Óptima:** cada estrato tiene la proporción y dispersión adecuada respecto a la de la aparición de la característica que representa en la población.

4. **Muestreo aleatorio por conglomerados.** Es un tipo de muestreo en el que inicialmente la población está definida mediante grupos o conglomerados que consideraremos homogéneos. En este sentido para tomar una muestra de tamaño n, consideraremos en tomar mediante muestreo aleatorio el número necesario de conglomerados para asegurarnos que la suma de los individuos de todos ellos sea el tamaño muestral buscado.
5. **Muestreo aleatorio compuesto, polietápico.** Es un muestreo aleatorio bastante complejo que trata de obtener la muestra mediante etapas y que puede presentar en cada etapa diferentes tipos de muestreo aleatorio como por ejemplo conglomerados, estratificados, etc. Es el que se utiliza por ejemplo en las encuestas y estudios sobre poblaciones distribuidas en amplias zonas geográficas. Se divide el país en donde se encuentra la población en zonas denominadas unidades primarias de muestreo (UPM). Dentro de cada UPM se forman áreas más pequeñas del orden de 500 habitantes que se denominan distritos del censo(DC).

4.2. Muestreo no aleatorio o no probabilístico

Son aquellos muestreos en los que se incumplen las principales características del muestreo aleatorio. Sin embargo, en muchas ocasiones son utilizados consciente o inconscientemente, puesto que son más sencillos de llevar a cabo y la utilización de muestreos aleatorios suele presentar costes muy elevados o inasumibles. Hay que tener en cuenta que este tipo de muestreo genera muestras que no tienen porque ser representativas y por lo tanto, no se puede garantizar la fiabilidad acerca de las inferencias y generalizaciones respecto al estudio del fenómeno aleatorio sobre la población mediante estas muestras.

1. **Muestreo no aleatorio de conveniencia.** Es un muestreo no aleatorio en el que se procede a seleccionar los individuos que menos esfuerzo presentan por cuestiones económicas, físicas o temporales.
2. **Muestreo no aleatorio voluntario o de voluntarismo.** Es un muestreo no aleatorio en el que se procede a seleccionar los individuos que deciden por propia iniciativa y voluntariamente ser elementos de la muestra. Además, este tipo de muestreo puede admitir la aparición en la muestra de repeticiones de individuos.
3. **Muestreo no aleatorio por cuotas o accidental.** Se toma a los individuos más "representativos" de la población como representantes de la población en la muestra. Por lo tanto, utiliza una técnica parecida al muestreo aleatorio por estratos en cuanto a que determina divisiones de la población mediante características pero se diferencia en la elección no aleatoria de los individuos dentro de cada estrato. En este sentido, es muy típico de este muestreo elegir los individuos con una serie de condiciones mediante "cuotas". La elección de los elementos dentro del estrato puede entrar la aleatoriedad o, por el contrario, el voluntarismo.

4. **Muestreo no aleatorio opinático o intencional.** Se trata de tomar muestras deliberadas que en ocasiones anteriores se tiene la certeza que han tenido comportamientos similares o parecidos al de la población en el fenómeno aleatorio en estudio. Estas muestras son consideradas como "representativas".
5. **Muestreo no aleatorio "Bola de nieve".** La muestra se crea a sí misma a partir de la selección de varios individuos los cuales conducen a otros, y estos a otros, y así hasta conseguir una muestra suficiente. Este tipo se emplea muy frecuentemente cuando se hacen estudios con poblaciones "marginales", delincuentes, determinados tipos de enfermos, etc.
6. **Muestreo no aleatorio discrecional.** El encuestador selecciona intencionadamente los elementos de la muestra en base a los prejuicios o conocimientos del investigador acerca de la representatividad de esos individuos respecto del estudio sobre la población.

5. ESTUDIO DE PARÁMETROS EN POBLACIONES Y MUESTRAS

5.1. Media, varianza y desviación típica poblacional.

Sea una población cuyos individuos poseen una característica común descrita por una variable estadística cuantitativa X que toma valores diferentes $x_1, x_2, \ldots, x_r$ mediante una distribución de frecuencias relativas $f_1, f_2, \ldots, f_n$.

- Se denomina **media poblacional** al valor media aritmética de todos los valores de la población: $\mu_X = M(X) = \sum_i x_i \cdot f_i$

- Se denomina **varianza poblacional** al valor media aritmética de la función

  $(X - \mu_X)^2$, $\sigma_X^2 = V(X) = M((X - \mu_X)^2) = \sum_i (x_i - \mu_X)^2 \cdot f_i$

- Se denomina **desviación típica poblacional** a la raíz cuadrada de la varianza poblacional, $\sigma_X = \sqrt{V(X)} = \sqrt{\sum_i (x_i - \mu_X)^2 \cdot f_i}$

Observaciones

1. Los valores o parámetros poblacionales media y desviación típica son esenciales para la descripción y valoración de cada población.
2. En muchas ocasiones el estudio estadístico no se puede efectuar mediante censo y hay que recurrir a muestras para el estudio de la población. En este sentido, uno de los objetivos consistirá en inferir el valor de la media y la desviación poblacionales.

Algunas de las propiedades de los parámetros media poblacional y desviación típica poblacional son:

1. Si X es una variable aleatoria y λ un número real fijo entonces $\mu_{\lambda \cdot X} = \lambda \cdot \mu_X$.
2. Dadas las variables X e Y se verifica que $\mu_{X+Y} = \mu_X + \mu_Y$.
3. Si X e Y son dos variables independientes, entonces $\mu_{X \cdot Y} = \mu_X \cdot \mu_Y$
4. $\sigma^2_X = M(X^2) - M(X)^2$
5. Para todo número real λ se verifica que $\sigma^2_{\lambda \cdot X} = \lambda^2 \cdot \sigma^2_X$
6. Si X e Y son variables independientes, entonces $\sigma s^2_{X+Y} = \sigma^2_X + \sigma^2_Y$.

5.2. Muestra aleatoria simple.

Definición de muestra aleatoria simple: sea un fenómeno aleatorio en el que se estudia una característica de una población sobre el que se mide una cierto parámetro asociado a la variable estadística X con determinada distribución de probabilidades. Llamaremos muestra aleatoria simple de la variable aleatoria X a la variable aleatoria n-dimensional ($X_1, X_2, \ldots, X_n$) cuyas variables aleatorias unidimensionales X_k con k = 1, 2, . . . n, que la componen son independientes y con la misma distribución de probabilidad (idénticamente distribuidas) que X.

5.3. Definición de estimadores: media, varianza y cuaivarianza muestral

En la inferencia estadística se suele plantear el problema de calcular uno o varios parámetros desconocidos θ_k de los que depende la distribución conocida de la variable aleatoria X a partir de la variable aleatoria simple. Muy usualmente estos parámetros son la media y desviación poblacionales.

Definición de estimador: sea ($X_1, X_2, \ldots, X_n$), una muestra aleatoria simple de una variable aleatoria X con distribución conocida p(θ) y con θ parámetro desconocido. Se llama estimador de un parámetro desconocido θ, a toda función T($X_1, X_2, \ldots, X_n$) que permita aproximar a dicho parámetro.
Definición de estimador centrado o insesgado: se dirá que el estimador es centrado o insesgado respecto del parámetro θ si E [T($X_1, X_2, \ldots, X_n$)] = θ
Definición de estimador consistente: se dirá que el estimador es consistente para el parámetro θ si se verifica que $\lim_{n\to\infty} E[T(X_1, \ldots, X_n)] = \theta \quad y \quad \lim_{n\to\infty} V[T(X_1, \ldots, X_n)] = 0$.
Algunos de los estimadores más utilizados son:

- **La media muestral:** sirve para estimar el parámetro media Ω si es que este es desconocido para la distribución de X. $T(X_1, X_2, \ldots, X_n) = \frac{1}{n} \cdot \sum_{k=1}^{n} X_k$. Se suele representar mediante $\overline{X}$.
- **La varianza muestral:** sirve para estimar el parámetro varianza α^2 si es que este es desconocido para la distribución de X. $T(X_1, X_2, \ldots, X_n) = \frac{1}{n} \cdot \sum_{k=1}^{n} \left(X_k - \overline{X}\right)^2$. Se representa mediante s^2.

- **La cuasivarianza muestral:** sirve para estimar el parámetro varianza α^2 si es que este es desconocido para la distribución de X al igual que la media Ω.. $T(X_1, X_2, \ldots, X_n) = \frac{1}{n-1} \cdot \sum_{k=1}^{n} \left(X_k - \overline{X}\right)^2$. Se suele representar mediante S^2.

5.4. Propiedades de la media muestral

1. La media de la media muestral es la media poblacional.
2. La varianza de la media muestral es el cociente de la varianza poblacional entre el tamaño muestral.

   Esta propiedad hace de la media muestral sea un gran estimador para la media Ω ya que "*cuanto más grande sea el tamaño muestral de las muestras usadas, menos dispersa será la media muestral"*. Es decir, los datos de la medición de la media muestral estarán tanto más localizados cuanto más grande sea el tamaño de las muestras que empleamos. Por tanto, según lo dicho antes, el estimador media muestral es consistente.

5.5. Propiedades de la varianza y cuasi varianza muestral

1. La media de la varianza muestral no es proporcional a la varianza poblacional.
2. La media de la cuasi varianza muestral es la varianza poblacional.

5.6. Relación entre la media muestral y poblacional: la ley de los grandes números

Enunciamos ahora la ley de los grandes números. Este resultado es consecuencia fundamental de la desigualdad de Chebyshef y juega un importante papel en la teoría de probabilidades y estadística relacionando la media muestral y poblacional de una variable aleatoria X. En particular:

- Proporciona una interpretación de la media como un valor esperado.
- Proporciona en el caso binomial una formulación del concepto de probabilidad en términos de límites de frecuencias.

Ley de los grandes números: dada una variable aleatoria simple (X_1, X_2, . . . , X_n) de una variable estadística X con distribución conocida p(θ) y con θ parámetro desconocido. Se verifica que

$$\lim_{n\to\infty} P_x\left(\left|\overline{X} - \mu_x\right| \leq \varepsilon\right) = 1$$

Observaciones

1. La ley de los grandes números explica que la media muestral es una variable aleatoria que, bajo cualquier distribución, tanto más cercana es a la media poblacional de la variable X de la que procede al aumentar el número de variables que la forman.

2. En el caso de que la v.a. X tenga distribución Bernuilli de parámetro p, sabemos que $\mu_x = p$ y $\sigma_x = p\cdot(1 - p)$. En ese caso, la ley de los grandes números

$$\lim_{n\to\infty} P_x\left(\left|\overline{X} - p\right| \le \varepsilon\right) = 1$$

Es decir, que cuantas más repeticiones o experimentos de Bernouilli independientes e idénticos realicemos, más aproximación existe entre la media muestral o proporción de éxitos y el valor poblacional *p*.

3. El valor medio o media aritmética de una variable aleatoria M(X) o E(X) recibe el nombre a menudo de valor esperado o esperanza matemática de la v.a. X. La razón de esta denominación está en la ley de los grandes números ya que cuando se repite un número grande de veces la medición de la v.a. X, la media aritmética en cada caso es una media muestral que, de acuerdo con la ley de los grandes números será con gran probabilidad muy cercana al valor de la media poblacional μ_x.

4. La ley de los grandes números da base para un posible modelo de definición de la probabilidad.

5.7. El teorema central del límite: estimación de la media poblacional a partir de intervalos de confianza para tamaños muestrales elevados

Teorema central del límite: sea X una variable estadística. Dado un intervalo I < ♣ cualquiera de la recta. Para cada n natural sea $p_n(\overline{X} \in I)$, la probabilidad de que la media muestral esté en el intervalo I. En este caso,

$$\lim_{n\to\infty} p_n(\overline{X} \in I) \approx \int_I N(x, \mu, \frac{\sigma}{\sqrt{n}})dx$$

Observaciones

1. Lo que el teorema expresa es un resultado general que nos dice que para tamaños muestrales suficientemente grandes, la media muestral es una variable estadística con distribución normal de media, la media poblacional y de desviación, el cociente de la desviación poblacional entre la raíz cuadrada del tamaño muestral.

$$\overline{X} \approx N\left(\mu_X, \frac{\sigma_X}{\sqrt{n}}\right)$$

Veamos ahora la aplicación directa de este teorema para la inferencia de la media poblacional a partir de la media muestral utilizando intervalos de confianza.

- **Si dada una variable estadística X con distribución de desviación típica conocida σ_X,** deseamos formar un intervalo de confianza ($\overline{X} - \varepsilon$, $\overline{X} + \varepsilon$) para la media poblacional para tamaños muestrales suficientemente grandes entonces mediante el teorema central del límite, el intervalo deseado será:

$$\left(\overline{X} - z_{\alpha/2} \cdot \frac{\sigma_X}{\sqrt{n}} \ , \overline{X} + z_{\alpha/2} \cdot \frac{\sigma_X}{\sqrt{n}} \right)$$

- **Si dada una variable estadística X con distribución de desviación típica desconocida σ_X,** deseamos formar un intervalo de confianza ($\overline{X} - \varepsilon$, $\overline{X} + \varepsilon$) para la media poblacional con tamaños muestrales suficientemente grandes entonces podemos reemplazar la desviación típica (que no conocemos) por la cuasivarianza muestral y el intervalo deseado será

$$\left(\overline{X} - t_{n-1,\alpha/2} \cdot \frac{S_n}{\sqrt{n}} \ , \overline{X} + t_{n-1,\alpha/2} \cdot \frac{S_n}{\sqrt{n}} \right)$$

Observaciones

1. Este procedimiento nos determina una cota inferior para el tamaño muestral ya que dada la desviación conocida σ_X, un valor $\varepsilon > 0$ cualquiera y una confianza $1 - \alpha$ establecida de que la media poblacional y la muestral se diferencian como máximo en ε, el menor natural n que cumple es tal que, $n \geq \left(z_{\alpha/2} \frac{\sigma_x}{\varepsilon} \right)^2$
2. Por un procedimiento análogo, si la desviación es desconocida, dado un valor $\varepsilon > 0$ cualquiera y una confianza $1 - \alpha$ establecida de que la media poblacional y la muestral se diferencian como máximo en ε, el menor natural n que cumple es tal que, $n \geq \left(t_{n-1,\alpha/2} \cdot \frac{S_n}{\varepsilon} \right)^2$

5.8. El concepto de valor esperado o esperanza matemática

El valor medio o media poblacional $\mu_X = M(X)$ de una variable estadística recibe a menudo el nombre de valor esperado o esperanza matemática de la variable X.

La razón de estas denominaciones radica en los resultados de los teoremas de la ley de los grandes números y central del límite. La cuestión es que cuando uno repite la medición de una variable cualquiera X un número grande de veces, el valor medio de los resultados es una media muestral de tamaño grande y por tanto de acuerdo con los teoremas anteriores ese valor medido será con gran probabilidad muy cercano al valor medio o media poblacional $\mu = M(X)$ de X.

CUESTIONES

1. **Probar que la media muestral de la variable X es el único número respecto del cual la suma de las desviaciones de los datos es cero.**

   Solución

   Sea una variable aleatoria simple de la variable X, (X_1, X_2, . . ., X_n) y sea μ un número que verifique que la suma de las desviaciones de los datos con respecto a ese valor es cero:

$$\sum_i f_i \cdot (X_i - \mu) = 0$$

   Es claro que siempre se verifica:

$$\sum_i f_i \cdot (X_i - \mu) = \sum_i f_i \cdot X_i - \sum_i f_i \cdot \mu = \overline{X} - \mu \cdot 1 = 0 \quad \Rightarrow \quad \overline{X} = \mu$$

2. **Probar que $s^2(X + \lambda) = s^2(X)$.**

   Solución

   Sea una variable aleatoria simple de la variable X, (X_1, X_2, . . ., X_n). Basta operar en la forma siguiente:

$$s(X + \lambda) = \frac{1}{n} \cdot \sum_i f_i \cdot \left((X_i - \lambda) - (\overline{X} - \lambda)\right)^2 = \frac{1}{n} \cdot \sum_i f_i \cdot (X_i - \overline{X})^2 = s(X)$$