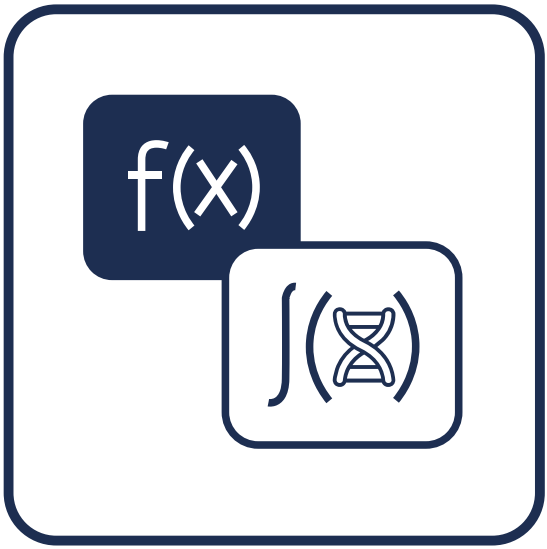
UCUENCA
CIENCIAS MÉDICAS

BIOESTADÍSTICA 1

Juan Chuchuca Pillajo
Paola Delgado Andrade
Nathalie Pinos Vélez
Verónica Pinos Vélez
José Roldán Fernández



LIBRO DE
CÁTEDRA

Bioestadística 1

Juan Chuchuca Pillajo
Paola Delgado Andrade
Nathalie Pinos Vélez
Verónica Pinos Vélez
José Roldán Fernández

Bioestadística 1

@Universidad de Cuenca

Autores: Juan Chuchuca Pillajo, Paola Delgado Andrade, Nathalie Pinos Vélez, Verónica Pinos Vélez y José Roldán Fernández.

Luis Rodrigo Mendieta Muñoz
Rector

Carmen Eulalia Calle Palomeque
Vicerrectora Académica

María de Lourdes Huiracocha Tutivén
Vicerrectora de Investigación e Innovación

Nancy Eulalia Auquilla Diaz
Decana de la Facultad de Ciencias Médicas

Jenny Carola Cardenas Carrera
Vicedecana de la Facultad de Ciencias Médicas

David Ricardo Achig Balarezo
Director de publicaciones de la Facultad de Ciencias Médicas

Centro Editorial UCuenca Press

Dirección (e): Ángeles Martínez Donoso. **Corrección:** Mihaela Ionela Badin.
Diseño: Geovanny Gavilanes Pando y John Macas Aguilar.
Preprensa: Juan Tigre Amón.

Ciudadela Universitaria
Av. Doce de Abril y calle Agustín Cueva
+593 (07) 413 4520
Casilla postal
editorial.ucuenca.edu.ec

Primera edición digital

ISBN: 978-9978-14-656-9

Derechos de autor: CUE-006341

El presente libro ha sido arbitrado por pares externos bajo el sistema doble ciego. Fieles al espíritu de la universidad pública, los libros de nuestra editorial son de acceso abierto y descarga libre para democratizar el conocimiento. Queda prohibida su venta.

La reproducción de este material para grupos o fines específicos, que no son personales, deben contar con la autorización de la Universidad de Cuenca.

Cuenca, Ecuador

Diciembre 2025-enero 2026

Índice

Prólogo **5**

Capítulo 1. Generalidades y método científico **7**

Capítulo 2. Diseños de investigación **13**

Capítulo 3. Variables **21**

Capítulo 4. Programas estadísticos **27**

Capítulo 5. Recolección de datos **37**

Capítulo 6. Gestión de datos: diseño de base de datos **47**

Capítulo 7. Gestión de los datos: depuración, codificación y cálculo de variables **53**

Capítulo 8. Niveles de análisis y tipos de pruebas **59**

Capítulo 9. Análisis descriptivo numérico: variables cualitativas **75**

Capítulo 10. Tasas y ajuste de tasas **87**

Capítulo 11. Análisis descriptivo gráfico: variables cualitativas **101**

Capítulo 12. Análisis descriptivo numérico: variables cuantitativas **109**

Capítulo 13: Análisis descriptivo gráfico: variables cuantitativas y epidemiológico 131

Capítulo 14: Distribución de los datos..... 145

Capítulo 15. Universo y muestra 161

Capítulo 16. Muestreo.....173

Biografía de los autores 183

Prólogo

La bioestadística es la disciplina científica que transforma los datos numéricos en información válida y de calidad. No se trata solo de un conjunto de fórmulas, sino de un lenguaje que facilita la interpretación de los datos biológicos y su conversión en conocimiento científico. En un mundo saturado de evidencia científica, la bioestadística es una herramienta indispensable para los profesionales en las ciencias de la vida.

Este libro surge de la necesidad de crear un recurso local, adaptado específicamente al ámbito de la salud, que apoye el proceso de comprensión y aplicación de la estadística en contextos biológicos y sanitarios. A lo largo de sus capítulos, se exploran los principios teóricos fundamentales, siempre con un enfoque práctico y contextualizado en situaciones reales de investigación. El lector encontrará explicaciones conceptuales claras, acompañadas de ejemplos, ejercicios y estudios de caso que fomentarán una visión crítica frente a los datos.

El texto ha sido diseñado como una herramienta valiosa tanto para estudiantes que comienzan en el área como para profesionales que desean reforzar sus capacidades analíticas. Más allá de enseñar cálculos, su propósito es promover un enfoque basado en la evidencia, el rigor metodológico y la interpretación responsable de los resultados.

Capítulo 1
Generalidades y método científico

Introducción

La estadística es la ciencia que se encarga de generar conclusiones a partir de datos numéricos. Cuando el estudio que se realiza es acerca de los hechos biológicos, éste se denomina bioestadística. Es decir, que es la ciencia que nos ayuda a generar información nueva a partir de la recolección ordenada y planificada de datos.

La bioestadística no se circunscribe al análisis de fenómenos biológicos aislados; es una herramienta esencial para abordar la complejidad propia de las ciencias de la vida o biociencias —conjunto de disciplinas que estudian a los organismos vivos, como plantas, animales, microorganismos y seres humanos—. Su aplicación es clave en la evaluación de intervenciones médicas, el análisis de factores de riesgo, la medición de resultados en salud, la vigilancia epidemiológica y la toma de decisiones sustentadas en evidencia científica.

El origen de esta ciencia surge de la imperiosa necesidad de llenar los diferentes vacíos existentes en la información. Por ejemplo, en medicina, existe una gran cantidad de enfermedades cuyo origen aún no se conoce, lo que hace que su tratamiento no sea óptimo. El desarrollo de investigaciones con el objetivo de descubrir esta causalidad ha generado una constante evolución, lo que ha llevado a que el tratamiento de las enfermedades cambie y evolucione para mejorar la calidad y la supervivencia de los enfermos.

Para que las conclusiones generadas por los estudios de investigación sean válidas, deben ser el resultado de un estudio de investigación ordenado, con una planificación impecable, que garantice la fiabilidad de los datos obtenidos. Por ello, es tan importante conocer los diferentes pasos a seguir para iniciar un estudio de investigación válido.

El objetivo de este capítulo es analizar las distintas etapas del método científico, desde el planteamiento del problema y la recolección y análisis de datos, hasta la formulación de conclusiones y recomendaciones fundamentadas.

Desarrollo

El método científico

Son las etapas que debe seguir un investigador para obtener conocimiento científico válido. Su finalidad es evitar errores causados por la subjetividad, generando observaciones fidedignas y exactas. Para ello, ordena y clasifica los datos recolectados para su correcto análisis, y finalmente ayuda al investigador a concluir objetivamente si la información encontrada es relevante y adecuada. Se basa en la reproductibilidad y la reutilización, lo que significa que cualquier investigador, ya sea de otro país o entorno, debería poder replicar un estudio, ya sea para verificarlo o rechazarlo.

Se caracteriza por ser un método teórico, sistemático, inductivo, deductivo y con una base empírica. Sus resultados no son definitivos y pueden ir cambiando de acuerdo con la generación de nuevo conocimiento con el tiempo, por lo que es un proceso circular. Además, busca evitar todos los factores externos que podrían generar malas interpretaciones y conclusiones erróneas. En conclusión, busca aislar de forma intencional el problema de estudio para colocarlo en las condiciones ideales y así evitar errores.

Existen varias técnicas que se pueden utilizar para la generación de nueva información:

- Técnica inductiva: cuando la información se genera a partir de la observación de casos individuales. Por ejemplo: "He visto que un pájaro vuela, he visto otro pájaro que vuela, por lo tanto, concluyo que los pájaros vuelan".
- Técnica deductiva: se basa en el razonamiento de acuerdo con premisas encontradas que son asumidas como ciertas. Volviendo al ejemplo anterior, podríamos decir que los pájaros son aves, los pájaros vuelan, luego las aves vuelan.
    - Hipotético-deductivo: es aplicado a las ciencias formales.
    - Analítico: divide al objeto de estudio para estudiarlo por partes.
    - Sintético: integra todos los componentes del objeto de estudio y lo analiza como un todo.
    - Histórico-comparativo.

  - Cuantitativo: usa la recolección de datos numéricos para probar diferentes teorías.
  - Cualitativo: usa la recolección no numérica de datos.
- El método científico tiene seis etapas:
  - Definición y planteamiento del problema: se determina cuál es el problema que nos preocupa y se genera una pregunta que debe ser resoluble.
  - Formulación de la hipótesis: explicación del porqué creemos que existe el problema observado. Es una verdad provisional que debe ser confirmada o rechazada.
  - Recogida y análisis de datos: como su nombre lo indica, es la recolección y análisis de los datos necesarios para verificar o rechazar nuestra hipótesis.
  - Confrontación de la hipótesis: aquí se confrontan los datos recolectados con nuestra pregunta de investigación.
  - Conclusiones y resultados: tras confrontar los datos, se decide si se acepta o rechaza la hipótesis generada.
  - Nuevas predicciones: son las preguntas que surgen mientras se desarrolla la investigación.

Fases de la investigación

Cuando queremos realizar un estudio de investigación, debemos seguir cinco pasos de forma lógica y ordenada:

- Planteamiento del problema
- Búsqueda y evaluación de la información
- Formulación de la hipótesis
- Verificación de la hipótesis
- Conclusiones y recomendaciones

El primer paso para cualquier estudio es identificar un problema que deba ser resuelto. Este problema puede surgir de la vida diaria, del trabajo, de nuestro entorno, etc. Tras identificar este problema, debemos intentar

determinar qué tan importante es, cuál es la posible causa que lo está generando y cómo podríamos comprobarlo. Para ello, nos planteamos la pregunta ¿qué?, y a partir de ahí respondemos las siguientes interrogantes: ¿por qué?, donde analizamos las razones que originan el problema, identificando las causas subyacentes que lo generan; ¿para qué?, donde reflexionamos sobre la relevancia de resolver este problema, estableciendo su importancia dentro de un contexto más amplio y sus implicaciones; y, ¿cómo?, donde definimos la estrategia y las acciones necesarias para abordar y resolver el problema de manera efectiva, detallando los métodos y enfoques que se utilizarán.

Esto nos permite comprender la raíz del problema, su relevancia y las posibles soluciones de manera estructurada y coherente.

Por ejemplo, un problema que enfrentan los estudiantes de primer año de medicina es que, en su gran mayoría, llegan tarde a las aulas el primer día de clases. Este problema podría estar generado por el desconocimiento de las instalaciones, la deficiente señalética existente, el nerviosismo, la ansiedad, la falta de una adecuada inducción preuniversitaria, etc. Para hacer este ejercicio podríamos preguntarnos:

¿Qué está haciendo que los estudiantes de primer año de medicina lleguen tarde a clases el primer día? Esto es importante porque en la primera hora de clases se da la introducción a la asignatura y se explica el sistema de calificación. Saber por qué llegan tarde y tratar de solventar estos errores nos ayudará para que todos los estudiantes tengan una correcta introducción a las asignaturas y sepan cómo se les va a calificar.

¿Cómo podría saber cuál de todas las posibles alternativas es la verdadera causante de este retraso? Probablemente, debería hacer una encuesta a todos los estudiantes de primer año, mediante la cual se les indague acerca de las razones de su impuntualidad.

Búsqueda y evaluación de la información

Tras tener clara nuestra pregunta central, el siguiente paso es empaparnos del tema, con la finalidad de buscar diferentes alternativas de respuesta y descartar que el estudio o la pregunta de investigación ya hayan sido respondidos por algún otro investigador. Para esto, debemos utilizar todo el material publicado hasta la fecha, tanto en textos como en repositorios bibliográficos físicos y digitales.

Se deben analizar los estudios individuales con detalle para determinar si son válidos y qué tan semejantes son a mi tema de estudio. Es decir, si existió una correcta planificación, si los sujetos de estudio fueron los adecuados, si las conclusiones son congruentes, cuán antiguo es el estudio, si están respondiendo a mi pregunta de investigación de manera adecuada, etc. Esto nos ayudará a descartar información de mala calidad científica, además de darnos una guía acerca de cómo podemos continuar con nuestro estudio.

Formulación de la hipótesis

La hipótesis no es más que la respuesta provisional a nuestra pregunta de investigación ¿qué? Esta debe ser clara y concisa, pues es la que nos guiará en los siguientes pasos de nuestra investigación, ya que todo lo que hagamos a partir de aquí dependerá de ella. Por esto, debe ser explicada de manera sencilla y entendible, sin dejar espacio a las subjetividades, que podrían generar confusión en los siguientes pasos de la investigación.

Verificación de la hipótesis

Aquí están todos los pasos a seguir para generar nuestras conclusiones. En esta parte del libro solo se dará un breve vistazo a lo que debemos hacer, para que el estudiante se haga una idea general del tema. Una investigación tiene dos partes: diseño y ejecución.

Diseño de la investigación

Aquí definiremos cómo vamos a recolectar, elaborar y analizar los datos, además de determinar cuáles son los recursos humanos y materiales de los cuales disponemos. Es decir, vamos a responder las siguientes preguntas: ¿Cómo vamos a realizar nuestro estudio? ¿Cuántas personas vamos a incluir en el mismo? ¿Cómo vamos a escoger a estas personas? ¿Cómo vamos a recolectar la información? ¿Existen conflictos éticos? ¿Qué tipo de estudio vamos a realizar? ¿Cuánto tiempo nos llevará el estudio? ¿Cuánto nos costará desarrollar el estudio?, etc.

En esta parte también es de suma importancia aclarar conceptos para evitar confusiones o subjetividades. Por ejemplo, si quiero hacer un estudio acerca de la obesidad en las parroquias urbanas de Cuenca, debo definir claramente los valores de IMC (índice de masa corporal) a partir de los cuales se van a interpretar como patológicos. Si, por ejemplo, yo determino que mi punto de corte es un IMC mayor de 30, ¿las personas que tengan 30, ¿a qué columna deberían ir? Para evitar estos errores, sería mejor poner que mi punto de corte es un IMC igual o mayor de 30.

Ejecución de la investigación

Esta es la etapa que sigue; tras diseñar el estudio, vamos a la recolección de la información, elaboración y análisis de los datos.

Conclusiones y recomendaciones

Después de haber pasado por todas las etapas anteriores, llega el turno de las conclusiones y recomendaciones. Aquí el investigador debe dar sus impresiones finales. Tras la ardua tarea realizada, se debe realizar un resumen de los datos más importantes y decidir si se acepta o rechaza la hipótesis generada. Es relevante entender que no solo los estudios de investigación que confirmen la hipótesis son válidos, puesto que cuando se rechaza una hipótesis también se genera conocimiento científico.

Conclusiones y recomendaciones

En resumen, el método científico es el conjunto de pasos que todo investigador debe seguir si desea realizar un estudio científico de calidad, con la menor cantidad de errores. La idea de este método es tener una guía sólida, ordenada y universal, para que todos los investigadores, sin excepción, puedan seguirla de una manera fácil y ágil, minimizando errores.

Por esta razón, se recomienda que todos los autores, previo a realizar un estudio de investigación, se empapen de este método y lo sigan paso a paso, sin saltarse niveles, para poder desarrollar el estudio de manera diligente, mientras minimizan errores.

Referencias

Celis, A., y Labrada, V. (2014). Bioestadística (3ra ed.). Editorial El Manual Moderno.

Fayad, V. (1970). *Estadística médica y de salud pública* (10ª ed.). Universidad de los Andes.

Ruggiero, M., Bustamante, Y., y Mark, C. (2015). *Guía teórico-práctica de Bioestadística I.* Universidad Central de Venezuela.

Capítulo 2
Diseños de investigación

Introducción

La investigación científica es el eje central del progreso en las ciencias de la salud. A través de ella, se generan conocimientos que permiten mejorar la atención médica, optimizar recursos, entender enfermedades y desarrollar políticas de salud pública. Para lograr estos objetivos, es fundamental que los estudios se diseñen de manera rigurosa, siguiendo principios metodológicos sólidos y empleando herramientas apropiadas para el análisis y la interpretación de datos.

En este capítulo se analizan los enfoques más comunes de la investigación científica, como son: el enfoque cuantitativo y el enfoque cualitativo. Se explorarán sus características, los procesos metodológicos que implican, los tipos de estudios que derivan de cada uno y sus aplicaciones en la práctica médica. Además, se incluirán ejemplos que permitan al lector aplicar estos conceptos en contextos reales de investigación.

Desarrollo

Investigación científica

La investigación científica es un proceso sistemático, organizado, metódico y objetivo mediante el cual se busca generar nuevos conocimientos o profundizar en los ya existentes. En el contexto de las ciencias de la salud, la investigación permite entender mejor los problemas que afectan a la población, evaluar la efectividad de intervenciones, identificar factores de riesgo y contribuir al diseño de políticas de salud.

Su importancia radica en que brinda una base sólida y verificable para la toma de decisiones clínicas, educativas, preventivas o administrativas. A través de la investigación, se promueve la mejora continua en los sistemas de salud, se sustenta la práctica basada en la evidencia y se fomenta el pensamiento crítico entre los profesionales.

Una investigación científica válida se caracteriza por seguir una metodología rigurosa, tener objetivos claros, seleccionar adecuadamente las herramientas de recolección y análisis de datos, y presentar resultados de forma transparente y reproducible.

Enfoques de investigación

La investigación científica puede abordarse desde diferentes perspectivas metodológicas. Las dos más conocidas son el enfoque cuantitativo, que se apoya en la medición numérica y el análisis estadístico, y el enfoque cualitativo, que busca comprender fenómenos complejos desde la perspectiva de los participantes.

El enfoque cuantitativo tiene como objetivo explicar, predecir y controlar fenómenos a través del análisis estadístico de variables observables y medibles. En cambio, el enfoque cualitativo se interesa en interpretar significados, experiencias y contextos sociales mediante el análisis de discursos, comportamientos y narrativas.

A continuación, se presentan las principales diferencias entre ambos enfoques, lo que permite al investigador seleccionar el más adecuado según el objetivo de su estudio.

Tabla 1

Comparación entre enfoques cualitativo y cuantitativo

Característica	Enfoque cuantitativo	Enfoque cualitativo
Finalidad	Medir fenómenos, probar hipótesis	Comprender fenómenos en profundidad
Tipo de datos	Números (variables cuantificables)	Palabras, imágenes, experiencia
Diseño	Estructurado, secuencial	Flexible, adaptable
Recolección de datos	Encuestas, pruebas, registros	Entrevistas, observación, grupos focales
Análisis	Estadístico, inferencial	Interpretativo, categorial
Muestra	Grande, representativa	Pequeña, intencionada
Resultados	Generalizables	Contextualizados

Ambos enfoques pueden combinarse en diseños de investigación mixtos, lo que permite una comprensión más integral de los fenómenos estudiados.

Procesos metodológicos en cada enfoque

El desarrollo de una investigación sigue etapas comunes que se adaptan según el enfoque seleccionado. En el enfoque cuantitativo, el proceso es más rígido y estructurado; en el cualitativo, se prioriza la flexibilidad y la construcción progresiva del conocimiento.

- Proceso cuantitativo: parte de una hipótesis formulada con precisión, plantea objetivos específicos y se orienta a verificar relaciones entre variables.
- Proceso cualitativo: se inicia con preguntas abiertas y exploratorias, busca significados y construye categorías a partir del análisis continuo de la información.

Tabla 2

Secuencia del proceso de investigación según el enfoque

Fase	**Enfoque cuantitativo**	**Enfoque cualitativo**
Idea	Claramente definida	General, exploratoria
Problema	Específico, medible	Abierto, centrado en el significado
Revisión bibliográfica	Exhaustiva	Inicial y luego progresiva
Diseño	Cerrado, secuencial	Flexible, adaptable
Recolección	Instrumentos estandarizados	Técnicas interactivas
Análisis	Estadístico	Categorial, inductivo
Interpretación	Valores numéricos	Significados emergentes
Validación	Control de sesgos	Triangulación

Diseños de investigación cuantitativa

Los diseños de investigación cuantitativa son estrategias metodológicas que permiten responder preguntas de investigación a través de la medición

numérica y el análisis estadístico de variables. Estos diseños se centran en establecer relaciones entre variables, probar hipótesis y, en muchos casos, determinar efectos causales. Su objetivo principal es obtener resultados objetivos, replicables y generalizables a una población más amplia.

Dependiendo del objetivo del estudio, los diseños cuantitativos pueden clasificarse en tres grandes grupos:

- Estudios descriptivos: buscan caracterizar una población, fenómeno o situación en términos de frecuencia, distribución o proporción. No establecen relaciones causa-efecto, sino que proporcionan una visión panorámica del fenómeno investigado.
- Estudios analíticos observacionales: permiten establecer asociaciones entre variables. Se subdividen según la temporalidad y el tipo de comparación en: transversales, casos y controles, y cohortes.
- Estudios experimentales: en los cuales el investigador manipula intencionalmente una o más variables independientes para observar sus efectos sobre una o más variables dependientes. El diseño más conocido en esta categoría es el ensayo clínico aleatorizado.

Estos diseños utilizan instrumentos estandarizados para la recolección de datos (encuestas, formularios, fichas clínicas), técnicas de muestreo probabilístico y pruebas estadísticas para el análisis, lo que permite validar las hipótesis planteadas y generar conocimiento aplicable en contextos clínicos y de salud pública.

A continuación, se describen los principales tipos de diseño cuantitativo con sus características y aplicaciones:

Tabla 3

Principales diseños de investigación cuantitativa

Categoria principal	Tipo de diseño	Descripción	Ejemplo en Salud
Observacional descriptivo	Descriptivo	Describe frecuencia o distribución sin analizar asociaciones	Prevalencia de HTA en adultos
Observacional analítico	Transversal-analítico	Asociación medida en un solo momento	Ansiedad y rendimiento académico
	Casos y controles	Comparación retrospectiva entre expuestos y no expuestos	Cáncer gástrico y alimentos ahumados
	Cohortes	Seguimiento temporal de grupos expuestos y no expuestos	Diabetes en trabajadores nocturnos
	Test diagnóstico	Evaluación de validez de un test frente a un estándar	Sensibilidad test COVID-19
Experimental	Experimental (ensayo clínico)	Manipulación deliberada con aleatorización	Antibiótico en neumonía

Diseños de investigación cualitativa

Los diseños de investigación cualitativa permiten explorar fenómenos complejos desde la perspectiva de los actores involucrados, profundizando en las experiencias, creencias, significados y contextos sociales. A diferencia de los diseños cuantitativos, no buscan medir ni generalizar resultados, sino comprender la riqueza y profundidad de los fenómenos humanos en su entorno natural.

Estos diseños son especialmente útiles cuando se investigan temas poco explorados, aspectos subjetivos como emociones, percepciones, decisiones, procesos sociales, dinámicas culturales, y cuando se requiere contextualizar los hallazgos en una situación específica. En el ámbito de la salud, la investigación cualitativa ha sido clave para comprender la adherencia al tratamiento, las barreras en el acceso a servicios, el significado de la enfermedad, la comunicación entre profesionales y pacientes, entre otros.

El investigador cualitativo actúa como un instrumento activo del proceso investigativo, y su relación con los participantes es cercana y dialogante. Los métodos más comunes de recolección de datos son las entrevistas en profundidad, los grupos focales, la observación participante y el análisis de documentos o narrativas. El análisis suele implicar la codificación de la información, la identificación de patrones y la construcción de categorías emergentes.

Los diseños cualitativos se clasifican en función del objetivo y el tipo de análisis que se pretende realizar. A continuación, se describen los principales tipos y sus aplicaciones en el campo de la salud.

Tabla 4

Principales diseños de investigación cualitativa

Tipo de diseño	Descripción	Ejemplo: campo de la salud
Etnográfico	Estudio de culturas o comunidades	Parto en comunidades rurales
Fenomenológico	Experiencias vividas	Pacientes con cáncer terminal
Teoría fundamentada	Teorización a partir de datos	Adherencia en VIH
Investigación-acción	Resolución de problemas con participación	Ambiente en salas de hospital
Narrativo	Análisis de relatos personales	Madres adolescentes

Conclusiones y recomendaciones

La elección del diseño de investigación es uno de los pilares fundamentales de todo estudio científico. A través del enfoque metodológico seleccionado —ya sea cuantitativo o cualitativo— se estructura el camino para responder adecuadamente a la pregunta de investigación, recolectar datos válidos y producir conocimiento confiable.

El enfoque cuantitativo permite observar patrones, medir variables y establecer relaciones estadísticas entre fenómenos. Es ideal para estudios que requieren evidencias sólidas, objetivas y generalizables, como aquellos

que evalúan intervenciones clínicas, determinan prevalencias o identifican factores de riesgo.

Por su parte, el enfoque cualitativo brinda profundidad, sensibilidad y contexto. Es indispensable cuando se busca comprender la experiencia humana, los procesos sociales y las dinámicas culturales que subyacen en los problemas de salud. No busca generalizar, sino interpretar, revelar significados y construir marcos teóricos desde la vivencia de los actores.

Ambos enfoques son válidos y complementarios. En la actualidad, muchas investigaciones en salud adoptan diseños mixtos, integrando herramientas y perspectivas de ambos paradigmas, lo que enriquece la comprensión de los fenómenos y fortalece la toma de decisiones basadas en múltiples fuentes de evidencia.

Se recomienda claridad en la pregunta de investigación antes de elegir un diseño, pues este debe guardar congruencia lógica con el enfoque elegido, el tipo de datos requeridos y la población a estudiar. Los estudiantes y profesionales de la salud deben desarrollar competencias básicas tanto en metodología cuantitativa como cualitativa, lo que les permitirá interpretar y generar evidencia científica con mayor criterio. Esto evitará elegir un enfoque por moda o facilidad aparente, sino por su adecuación al problema de investigación.

En problemas complejos de salud, combinar métodos puede ofrecer una visión más integral y contextualizada, complementando lo que cada enfoque por sí solo no puede abarcar completamente. Además, las universidades y centros de salud deben fortalecer las líneas de investigación, ofrecer formación continua y crear espacios de asesoría metodológica para garantizar la calidad científica de los proyectos.

Referencias

Creswell, J. W., y Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches (5th ed.)*. SAGE Publications.

Flick, U. (2015). I*ntroducción a la investigación cualitativa* (5ª ed.). Morata.

Hernández Sampieri, R., Fernández Collado, C., y Baptista Lucio, P. (2014). *Metodología de la investigación (6ª ed.)*. McGraw-Hill.

Martínez-González, M. A., Sánchez-Villegas, A., y Faulin Fajardo, J. (2016). *Bioestadística amigable.* Elsevier España.

Patton, M. Q. (2015). *Qualitative research & evaluation methods* (4th ed.). SAGE Publications.

Polit, D. F., y Beck, C. T. (2012). *Investigación en enfermería: principios y métodos* (8ª ed.). McGraw-Hill.

Ruggiero, M., Bustamante, Y., y Mark, C. (2015). *Guía teórico-práctica de bioestadística I.* Universidad Central de Venezuela.

Capítulo 3
Variables

Introducción

En el análisis estadístico, el concepto de variable resulta fundamental, pues toda investigación parte del análisis de las características que varían entre los individuos de una población. Una variable puede definirse como toda característica medible u observable que puede adoptar distintos valores entre sujetos, por ejemplo, edad, peso, nivel de glucosa en sangre, sexo o la presencia de una enfermedad.

La identificación de las variables del estudio constituye el primer paso del análisis estadístico, pues determina qué tipo de herramientas pueden aplicarse para describir, comparar o establecer relaciones entre los datos. Conocer y clasificar adecuadamente las variables durante el proceso de investigación es una competencia básica que permite manejar con mayor precisión la información.

Una buena gestión de las variables de estudio mejora la calidad del análisis y contribuye a una interpretación crítica y fundamentada de la información. El objetivo de este capítulo es conocer el concepto de variables y sus diferentes clasificaciones y usos.

Desarrollo

Concepto de variable

Una variable es una característica individual dentro de la población en estudio, que se puede observar, medir y diferenciar. En bioestadística, las variables son fundamentales para estudiar fenómenos, eventos en salud y enfermedades, por lo cual es indispensable identificarlas, recolectarlas, analizarlas y procesarlas.

Una variable no es sinónimo de pregunta. Una pregunta puede dar origen a una o varias variables, dependiendo de cómo se estructure la información. Por ejemplo, la pregunta: "¿Qué servicios básicos tiene usted en su vivienda?" puede generar varias variables independientes. En este caso, cada servicio constituye cuatro variables distintas:

1. Disponibilidad de luz eléctrica (sí / no)

2. Disponibilidad de agua potable (sí / no)

3. Disponibilidad de teléfono (sí / no)

4. Disponibilidad de alcantarillado (sí / no)

Cada una de estas variables es cualitativa nominal dicotómica, ya que solo admite dos categorías de respuesta mutuamente excluyentes.

O, al contrario, podemos tener el ejemplo del test DASS-21, en donde siete preguntas corresponden a una sola variable, que sería la depresión. Cada una de las preguntas tiene diferentes opciones de respuesta que van desde "nunca", "a veces", "a menudo" y "casi siempre". Cuando las personas contestan "nunca", o cuando contestan "casi siempre", se obtienen puntajes de 0 a 3, con escalas intermedias. Al sumar el puntaje de las siete preguntas, se obtiene la variable nivel de depresión, que va desde "no tiene depresión", "depresión leve", "depresión moderada", "depresión severa" y "extremadamente severa". Es decir, siete preguntas construyen una variable.

Las variables se clasifican según su naturaleza, la función que cumplen en el análisis, el nivel de medición y su estructura.

Según su naturaleza

- Variables cuantitativas: las denominadas numéricas son aquellas que representan cantidades y se expresan mediante números. Se dividen en:

    1. Variables cuantitativas discretas o discontinuas: son aquellas que cuentan o contabilizan caracteristicas que no se pueden fraccionar, por lo que no tienen decimales. Es decir, toman valores enteros y contables. Ejemplos: número de hijos, de hermanos, de días de hospitalización, de consultas médicas en un año.

    2. Variables cuantitativas continuas: son aquellas que pueden tomar cualquier valor dentro de un intervalo, es decir, tienen unidades de medida que podrían incluir decimales. Ejemplos: talla (1,57 m) , peso (34,5 kg) , glicemia (93 mg/dl).

- Variables cualitativas: son aquellas que describen cualidades, atributos o características de la población. Se clasifican en dos tipos principales:

    1. Variables cualitativas nominales o marginales: son aquellas en las que las categorías no tienen un orden específico; se subdividen en dicotómicas, cuando tienen solo dos categorías. Ejemplo: sexo (hombre, mujer); y politómicas o policotómicas, cuando tienen más de dos categorías, pero no tienen un orden o secuencia. Ejemplo: estado civil (soltero, casado, viudo, divorciado, unión libre).
    2. Variables cualitativas ordinales: son aquellas en las que las categorías tienen un orden jerárquico. Ejemplos: nivel de educación (primaria, secundaria, tercer nivel y cuarto nivel); grado de dolor (leve, moderado, severo).

Según el nivel

- Nivel nominal: corresponde al nivel de las variables nominales, tanto las dicotómicas como las politómicas, en las cuales no hay un orden específico.
- Nivel ordinal: como su nombre lo indica, corresponde a las variables ordinales, en donde las categorías se organizan según un orden o jerarquía.
- Nivel de intervalo: corresponde a variables cuantitativas en las cuales los datos se pueden organizar según un orden definido, y las distancias entre los valores son iguales o significativas. En este tipo de nivel, las variables no contienen un cero absoluto y pueden tener valores negativos, por lo que no se analizan bajo proporciones. Ejemplo: temperatura en grados Celsius (-10 grados).
- Nivel de razón: este tipo de nivel también corresponde a variables cuantitativas, pero en este caso contiene el cero absoluto (punto de partida), lo que permite realizar operaciones matemáticas como sumas, restas, multiplicaciones y divisiones; no permite tener valores negativos. Ejemplos: peso corporal, frecuencia respiratoria, edad.

Según la función en el análisis

- Variable independiente: es la variable predictora que se puede manipular; se utiliza para explicar o predecir, puede ser considerada como el factor asociado o de riesgo o la causa del efecto. Solo los estudios que buscan causalidad la tienen. Ejemplo: la actividad física como variable independiente de la presión arterial.

- Variable dependiente o principal: es la llamada variable resultado; representa el efecto o la respuesta ante la variable independiente. Todos los estudios tienen esta variable. Ejemplo: la presión arterial, que depende de la actividad física.
- Variable confusora: o de confusión, es una variable que se relaciona tanto con la variable dependiente como con la independiente, y que puede distorsionar o enmascarar la verdadera relación entre ambas. Ejemplo: en el análisis de presión arterial y actividad física, la edad puede ser una variable confusora, ya que influye tanto en el nivel de actividad física como en los valores de presión arterial.
- Variables universales: son variables base, que nos ayudan a dar un panorama general de la población de estudio, generalmente son los datos de filiación. Ejemplo: edad, sexo, raza, estado civil.

Según su estructura

- Variable simple: es aquella en la que se observa una sola característica o dimensión, y se la obtiene de manera directa. Ejemplo: peso y talla.
- Variable compuesta: es la que está formada por dos o más variables simples. Ejemplo: Índice de masa corporal (IMC: peso/talla^2).

Tabla 1

Ejemplos de variables de uso común

Variable	Tipo de variable	Nivel de medición	Ejemplo clínico
Fumador	Cualitativa nominal dicotómica	Nominal	Sí/No
Grupo sanguíneo	Cualitativa nominal policotómica	Nominal	Tipo O, A, B, AB
Estadío del cáncer	Cualitativa ordinal	Ordinal	Estadio I Estadio II Estadio III Estadio IV
Pulso	Cuantitativa discontinua	Razón	80 por minuto
Peso al nacer	Cuantitativa continua	Razón	3,2 kg
Número de consultas médicas	Cuantitativa discreta	Razón	5 visitas/ año
Temperatura ambiental	Cuantitativa continua	Intervalo	-12 °C

Conclusiones y recomendaciones

La correcta clasificación y conocimiento de las variables constituye una parte fundamental del proceso investigativo. Las variables permiten transformar fenómenos complejos en datos medibles y analizables, lo que posibilita describir poblaciones, establecer comparaciones y describir relaciones causales o de asociación.

Comprender la naturaleza de las variables y el rol que cumplen dentro del análisis es clave, tanto para elegir las pruebas estadísticas apropiadas como para realizar un buen análisis estadístico. Además, este conocimiento permite interpretar los resultados obtenidos evitando cometer errores.

Una buena gestión de las variables del estudio fortalece la calidad de la información obtenida a través de su análisis, contribuyendo en la generación de conocimiento aplicable, útil y responsable.

Referencias

Celis de la Rosa, A. J., y Labrada Martagón, V. (2014). *Bioestadística* (3ª ed.). Editorial El Manual Moderno.

Fayad Cámel, V. (1970). *Estadística médica y de salud pública* (2ª ed.). Universidad de los Andes.

Gómez Barrantes, M. (2012). *Elementos de la estadística descriptiva* (4ª ed.). Universidad Estatal a Distancia de Costa Rica.

Hernández Sampieri, R., Fernández Collado, C., y Baptista Lucio, P. (2014). *Metodología de la investigación* (6ª ed.). McGraw-Hill Education.

Martínez González, M. A., et al. (2020). Bioestadística amigable (4ª ed.). Elsevier.

Ministerio de Sanidad y Consumo de España. (2008). *Manual de Bioestadística para Ciencias de la Salud*. *Ministerio de Sanidad*.

Organización Panamericana de la Salud. (2016). EPIDAT 4.2: *Programa para análisis epidemiológico de datos tabulados* [Internet]. OPS. Recuperado de https://www.sergas.es/Saude-publica/EPIDAT

Capítulo 4
Programas estadísticos

Introducción

Durante el proceso de investigación, la cantidad de datos recolectados hace necesario el uso de programas estadísticos que faciliten la tabulación, análisis, procesamiento y presentación de resultados. Estas herramientas pueden ser incorporadas desde etapas iniciales en el proceso investigativo, como la recolección y organización de datos. En este punto, el uso de herramientas como Google Forms o Kobo Toolbox permite optimizar la recolección de información y su posterior incorporación a bases de datos en software como Microsoft Excel, convirtiéndose en puentes hacia programas más avanzados, donde se realizará el análisis y procesamiento estadístico.

Por ello, el dominio básico de varios programas se ha convertido en una competencia indispensable en el proceso de investigación en salud. Saber aplicar los conceptos teóricos de la estadística en un software capaz de generar gráficos, tablas o correr pruebas no solo mejora la eficiencia en el análisis, sino que también eleva la calidad del trabajo, reduce errores y facilita la replicabilidad de los resultados.

Finalmente, el acceso a software libre, que no requiere suscripciones y que puede complementarse entre ellos, facilita los procesos de enseñanza-aprendizaje, así como su uso diario en las aulas y en los ordenadores personales. El objetivo de este capítulo es recopilar los diferentes programas que pueden ser usados en los diversos momentos de la investigación.

Desarrollo

Gestores de texto

Los gestores de texto son importantes dentro de los procesos de investigación, ya que cumplen un papel esencial como herramienta para la redacción de protocolos, informes finales, elaboración de instrumentos para

la recolección de datos, y para la visualización de reportes, tablas, gráficos o interpretaciones, que son muy comunes dentro del campo de la salud.

Entre los gestores de texto más utilizados en el ámbito académico se encuentran Documentos de Google y Microsoft Word, cada uno con características propias y diferentes al momento de trabajar en aspectos relacionados con la investigación y el análisis estadístico, con accesibilidad en dispositivos móviles y de escritorio.

Documentos de Google:

Gestor de texto en línea que forma parte de un conjunto de herramientas funcionales ofrecidas por Google Drive, al que se accede únicamente con una cuenta gratuita de Google. Tiene ventajas importantes, como permitir el trabajo colaborativo en línea y en tiempo real, ya que se puede usar en el sitio web. Interactúa directamente con Google Sheets y Google Forms, lo que facilita la inserción de gráficos, tablas y resultados de manera automática desde estas herramientas. Además, permite integrar bibliografías desde gestores como Mendeley y Zotero, aplicando normas y estilos como APA o Vancouver, entre otras. Al funcionar en línea, guarda automáticamente los cambios, lo que evita pérdidas de información. Si bien tiene funciones limitadas en cuanto a estilos y formatos avanzados, en la actualidad es una herramienta muy funcional y práctica para estudiantes universitarios, sobre todo para trabajos colaborativos.

Microsoft Word:

Herramienta de Office que permite gestionar textos en el ámbito académico, institucional y profesional. A diferencia de Documentos de Google, ofrece una mayor funcionalidad en lo que respecta a estilos y formatos de texto. Al ser parte del paquete Office, requiere instalaciones previas y licencias para acceder a distintos niveles de funcionalidades que permiten redactar, editar y estructurar documentos de manera avanzada.

También se integra con gestores de referencias bibliográficas como Zotero, Mendeley y EndNote, lo que facilita la aplicación de normas como APA, Vancouver, entre otras. A diferencia de Documentos de Google, Microsoft Word tiene limitaciones en cuanto al trabajo colaborativo en tiempo real, aunque permite revisar cambios y añadir comentarios de manera organizada. Sin embargo, una de sus principales ventajas es la integración directa con programas como Excel, SPSS y RStudio, lo que facilita insertar gráficos, tablas y resultados dentro del mismo documento.

Herramientas para la recolección de datos

En la actualidad, a diferencia de décadas anteriores, en las que la recolección de datos dependía exclusivamente de técnicas como encuestas aplicadas con instrumentos físicos (formularios impresos que los encuestadores utilizaban de manera directa con los encuestados), existen diversas herramientas digitales que permiten diseñar, aplicar y gestionar instrumentos de recolección de datos estadísticos de forma digital.

Entre las principales herramientas de acceso gratuito se encuentran Google Forms y Kobo Toolbox. Sin embargo, también existe una amplia variedad de plataformas que requieren licencias de pago para su uso, las cuales ofrecen funcionalidades específicas o más complejas.

El uso de los formularios digitales brinda algunas facilidades, como el acceso al instrumento a través de enlaces, códigos QR y la posibilidad de parametrizar las preguntas para evitar datos incompletos, además de permitir la concatenación de preguntas o secciones. De esta manera, se garantiza que las bases de datos obtenidas sean mucho más completas en comparación con los registros físicos, en los cuales la pérdida de datos era mucho más frecuente.

Resultan muy funcionales al momento de realizar pruebas piloto, ya que permiten identificar posibles errores y hacer correcciones de forma simultánea, mientras se desarrollan los procesos de pilotaje.

Algunas de las ventajas adicionales que ofrecen los formularios digitales son: la optimización en el uso de papel, la facilidad para generar automáticamente bases de datos, y la capacidad de elaborar reportes rápidos que tabulan las preguntas incluidas en el instrumento de recolección de datos. Esto permite obtener resultados preliminares que orientan y ajustan los procesos de recolección a futuro.

Los formularios digitales resultan claves en los procesos de capacitación, ya que permiten familiarizar al personal con el instrumento, y también ayudan a definir los tiempos estimados para la recolección de datos, así como el número de encuestados que deben ser considerados en el proceso de levantamiento.

Google Forms:

La herramienta de Google Forms es gratuita, lo único que requiere es una cuenta de Gmail que tenga acceso a Google Drive. A partir de Google Drive, la creación de formularios digitales en Google Forms es bastante intuitiva, fáciles de parametrizar, tienen opciones de preguntas abiertas, preguntas cortas, preguntas largas, preguntas de opción múltiple, preguntas de varias opciones de respuesta. También permite recolectar fechas, horas, fotografías o archivos en diferentes formatos como PDF o Word.

Las funcionalidades de Google Forms son bastante fáciles de usar, permiten diseñar formularios con múltiples pantallas, personalizando los encabezados, los colores, y diseño; también tiene la característica de poder concatenar secciones de preguntas, es decir, si una persona responde a una opción de una pregunta, como por ejemplo, si autoriza o si está de acuerdo en participar en la encuesta, en caso de que la respuesta sea "sí", automáticamente permite enviar a una sección de preguntas del cuestionario y si la respuesta es "no", puede ser parametrizado para enviar a un mensaje de agradecimiento.

Una vez terminado el diseño del formulario, se genera automáticamente un enlace que puede ser compartido por diferentes mecanismos como redes sociales, dispositivos móviles, correos electrónicos, etc. La desventaja que tiene es que, para poder hacer el proceso de recolección de datos, la persona tiene que estar conectada a una red que disponga de internet, sea a través de wifi, datos móviles o una red fija en computador, lo que dificulta el uso en zonas de difícil acceso, como áreas rurales.

Una vez que se realiza el proceso de recolección de datos, Google Forms genera una base de datos utilizando hojas de cálculo de Google, lo que facilita las descargas y el manejo de los datos *a posteriori*. Finalmente, tiene un módulo de reportes inmediatos con gráficos de barras o de sectores y tablas simples de frecuencia.

KoboToolBox:

Herramienta gratuita creada por la Harvard Humanitarian Initiative, que paulatinamente ha ido condicionando el acceso gratuito a una capacidad de almacenamiento limitada y a un número específico de formularios. Sin embargo, en la actualidad sigue siendo una herramienta importante para la creación, aplicación y gestión de formularios.

Al igual que Google Forms, es bastante intuitiva. Requiere la creación de una cuenta para poder desarrollar diferentes tipos de proyectos, cada uno con su propio formulario, el cual permite agregar preguntas abiertas, cerradas, de opción múltiple y de selección múltiple. También permite recolectar fechas, horas, audios, videos, fotografías y coordenadas geográficas con geolocalización.

A diferencia de Google Forms, KoboToolbox permite gestionar los datos en tiempo real, conforme van llegando a la plataforma web. Esto posibilita un proceso de depuración inmediata de cada formulario, pudiendo aceptar, rechazar o mantener las respuestas en modo de edición. La información ingresada se puede visualizar individualmente por cada formulario, y cuenta con la particularidad de generar un reporte rápido basado en gráficos, que pueden ser modificados entre gráficos de barras, de sectores o de anillos. Además, se complementa con tablas simples de frecuencia, identificando variables cualitativas, cuantitativas, discretas y continuas, y generando estadísticas descriptivas como media, mediana y moda, según el caso.

Permite descargar las bases de datos en formatos compatibles con Excel, SPSS, entre otros. A diferencia de Google Forms, su principal ventaja es que los usuarios pueden acceder al formulario mediante un enlace que requiere conexión a internet solo en el momento de apertura, pero pueden completar el cuestionario sin estar conectados, y sincronizar la información posteriormente cuando tengan acceso a una red.

Gestores de bases de datos

Una vez que se ha realizado el proceso de recolección de datos, ya sea mediante instrumentos físicos o digitales, la creación de bases de datos que permitan visualizar, depurar, filtrar y gestionar la información se vuelve, en muchos casos, una necesidad fundamental. Para ello, es indispensable el uso de programas que permitan organizar y trabajar adecuadamente con los datos recolectados.

En la actualidad, los más utilizados son las hojas de cálculo de Google y Excel de Microsoft Office, cada uno con sus propias bondades. Muchos de estos programas, además de permitir la gestión de datos, facilitan ciertos procesos de análisis estadístico, así como la generación de resultados representados en tablas o gráficos.

Sin embargo, es importante señalar que existen programas específicos diseñados para analizar y procesar la información bajo una lógica bioestadística más avanzada.

Hojas de cálculo de Google:

Al igual que Documentos de Google y Google Forms, las hojas de cálculo de Google son herramientas gratuitas, disponibles dentro del entorno de Google, que únicamente requieren contar con una cuenta de Gmail. Estas hojas se almacenan directamente en Google Drive y son fáciles de compartir entre diferentes usuarios, a quienes se les puede asignar roles como lectores, editores o revisores. Esa es una de las ventajas principales de las hojas de cálculo de Google, ya que permiten y facilitan el trabajo en equipo.

Dentro de sus funcionalidades, existen ciertas limitaciones para el análisis avanzado de información. Sin embargo, contienen funciones básicas como sumar, restar, dividir, multiplicar, promediar y filtrar, que son herramientas fundamentales dentro del proceso de investigación, especialmente para estudiantes universitarios.

Las hojas de cálculo de Google también permiten generar gráficos simples, como gráficos de barras y gráficos de sectores. De igual manera, permiten obtener estadísticas básicas como media, mediana, moda, conteos y porcentajes. Tienen la particularidad de manejar hasta 10 millones de celdas por cada hoja de cálculo, lo cual, en la práctica, no representa una limitación para proyectos de investigación o tesis de pregrado en el área de la salud.

Finalmente, es importante mencionar que las hojas de cálculo de Google se sincronizan de manera directa con los formularios de Google Forms, lo que facilita el proceso de creación de bases de datos, así como con Looker Studio, que permite construir tableros visuales, que una vez diseñados solo actualizan los datos a partir de las hojas de cálculo. Además, son altamente compatibles con Microsoft Excel, ya que permiten copiar y pegar información entre ambas plataformas sin mayores inconvenientes.

Microsoft Excel:

Programa que forma parte del paquete Office y que requiere instalación en un ordenador de escritorio o un computador portátil. A diferencia de las hojas de cálculo de Google, necesita ejecutarse como software instalado y no puede ser usado directamente en un sitio web.

Excel tiene la ventaja de ser una herramienta robusta que permite manejar hasta 1'048 576 filas por 16 384 columnas, lo que sin duda marca una diferencia significativa frente a lo permitido en las hojas de cálculo de Google. Contiene funciones similares a las que se encuentran en Google Sheets, sin embargo, tiene la capacidad adicional de trabajar con tablas dinámicas y un sinnúmero de funciones avanzadas, las cuales dependen de la licencia adquirida dentro del paquete de Office. También permite generar distintos tipos de archivos compatibles con programas como Epidat, Epi Info, SPSS o RStudio, lo cual representa una ventaja en investigaciones que requieren análisis bioestadístico más avanzado.

Una de sus principales desventajas, en comparación con las hojas de cálculo de Google, es la limitación para trabajar en línea en tiempo real con un equipo de trabajo, lo que impide la colaboración simultánea.

Procesadores de bases de datos

Con la finalidad de cerrar este capítulo, abordando todo el proceso del método estadístico, que corresponde a la recolección, el análisis, el procesamiento y la presentación de resultados, es importante mencionar que, una vez que se cuenta con las diferentes bases de datos, los procesadores de datos para la generación de resultados estadísticos se vuelven fundamentales para el cumplimiento de este proceso.

Como principales procesadores de datos en la actualidad se tienen programas como EpiDat, Epi Info, RStudio, SPSS, entre otros. Sin embargo, se debe tener en consideración que algunos de ellos son gratuitos, mientras que otros requieren suscripciones pagadas para su uso. Por esta razón, en el ámbito académico, se ha priorizado el uso de *software* libre, como es el caso de EpiData y Epi Info, por su facilidad de acceso, comprensión y manejo por parte de los estudiantes universitarios. Cabe señalar que ambos programas son compatibles con formatos XLS, CSV, XLSX, provenientes de Excel.

Algunos de estos programas, además del procesamiento de datos, también incluyen módulos para la elaboración o construcción de instrumentos de recolección de datos. Sin embargo, al no contar con la posibilidad de compartir los formularios mediante enlaces o códigos QR, esta funcionalidad ha ido quedando en desuso. Actualmente, su mayor fortaleza se encuentra en

el procesamiento de datos y en la generación de resultados estadísticos, que permiten presentar información de manera ordenada, visual y útil para la toma de decisiones en salud.

EpiDat:

Es un programa gratuito para el procesamiento de bases de datos que fue conceptualizado por la Junta de Galicia, en España, en colaboración con la Organización Panamericana de la Salud, y está enfocado específicamente al análisis estadístico en epidemiología y salud pública.

Tiene funcionalidades muy importantes en estadística descriptiva, permitiendo obtener los principales indicadores para variables cualitativas y cuantitativas, como porcentajes, media, mediana, moda, desviación estándar, varianza, cuartiles, deciles y rangos intercuartílicos. Además, permite realizar contrastes de normalidad y analizar variables cualitativas mediante tablas simples y tablas cruzadas (tablas de contingencia).

Cuenta con opciones en gráficos de barras simples, barras dobles, barras apiladas, histogramas con curva de normalidad, cajas y bigotes, líneas y dispersión.

De igual manera, cuenta con pruebas de inferencia estadística, que permiten realizar inferencias sobre una población o dos poblaciones, incluyendo pruebas como T de Student, prueba Z y McNemar. También permite realizar análisis de correlación, de tasas, y es especialmente fuerte en lo que corresponde al cálculo del tamaño muestral y a la selección de muestras, tanto para estudios descriptivos como para estudios con contrastes de hipótesis. Su versión 3.1 incluye la posibilidad de calcular pruebas diagnósticas y riesgos.

Cuando se generan los diferentes reportes, ya sea de cálculos, estadísticos, tablas o gráficos, Epidat contempla opciones que permiten exportar el informe a Excel o Word de manera muy sencilla e inmediata. Cuando se realiza el proceso de descarga, el programa tiene la particularidad de incluir en la carpeta un manual de ayuda, así como ejemplos de bases de datos. Esto permite al estudiante revisar, capítulo por capítulo, los conceptos teóricos de las diferentes funciones que ofrece Epidat, y al mismo tiempo replicar los ejercicios prácticos utilizando las bases que se facilitan junto con el software. Esta es una bondad que vale la pena resaltar desde el ámbito académico, ya que contribuye significativamente a la formación universitaria.

Epi info:

Herramienta diseñada por los Centros para el Control y la Prevención de Enfermedades (CDC) de los Estados Unidos, con un enfoque específico en vigilancia sanitaria, epidemiología y salud pública. A diferencia de Epidat, Epi Info incorpora diversas funciones que pueden ser utilizadas tanto en el ámbito técnico como académico.

Entre las más utilizadas en la actualidad se encuentra una calculadora estadística que permite obtener resultados derivados del cruce de dos variables, aplicando la lógica de las tablas tetracóricas. De igual manera, dispone de tableros visuales que permiten cargar bases de datos en formatos XLS o XLSX provenientes de Excel. Una vez cargada la base, los datos se visualizan directamente dentro de Epi Info, lo que facilita el análisis guiado y progresivo de la información.

Con la base ya incorporada en el tablero visual, se pueden realizar diferentes análisis descriptivos e inferenciales, obteniendo resultados de estadísticos similares a los que se obtienen en Epidat: como media, mediana, moda, desviación estándar, varianza, cuartiles, deciles y rangos intercuartílicos. Sin embargo, una de sus principales ventajas es que la visualización de los datos es mucho más amigable y presenta la información de forma más clara y dinámica.

También permite generar gráficos que son fáciles de personalizar, ya que ofrece opciones para modificar colores, etiquetas, títulos y otros elementos visuales, lo cual mejora notablemente la presentación de resultados. En cuanto a estadística inferencial básica, Epi Info es bastante robusto, permitiendo realizar pruebas como: Chi cuadrado, T de Student, ANOVA, U de Mann Whitney, Kruskal Wallis, Wilcoxon, así como análisis de correlación de Pearson y correlación de Spearman. Al igual que EpiDat, EpiInfo permite migrar los resultados del análisis estadístico muy fácilmente a diferentes gestores de texto o Excel.

Conclusiones y recomendaciones

Durante el proceso investigativo, la utilización de programas estadísticos constituye una necesidad, dada la importancia de generar información de alto valor. En este capítulo se han abordado las particularidades de varios

programas que tienen un papel protagónico en diferentes fases del proceso investigativo, desde gestores de texto, recolectores de datos, organizadores de bases de datos y procesadores estadísticos.

El dominio de estas herramientas en la investigación en ciencias de la salud es indispensable para promover eficiencia, precisión y calidad en el manejo de la información.

Referencias

Belliard, M., y Cristofoli, M. E. (2004). *Manual de estadística con Microsoft Excel* [Internet]. Buenos Aires: Omicron System. ISBN: 978-9871046249.

Celis de la Rosa, A. J., y Labrada Martagón, V. (2014). *Bioestadística* (3ª ed.). Editorial El Manual Moderno.

Centers for Disease Control and Prevention (CDC). (s.f.). *Epi Info™ User Guide* [Internet]. Atlanta (GA): CDC. Recuperado de https://www.cdc.gov/epiinfo/user-guide/index.html

Martínez González, M. A., et al. (2020). *Bioestadística amigable* (4ª ed.). Elsevier.

Organización Panamericana de la Salud. (2016). *EPIDAT 4.2: Programa para análisis epidemiológico de datos tabulados* [Internet]. OPS. Recuperado de https://www.sergas.es/Saude-publica/EPIDAT

Capítulo 5
Recolección de datos

Introducción

Todo investigador, una vez que tiene clara su hipótesis de investigación, debe iniciar con la preparación y elaboración de las herramientas que va a utilizar para la recolección de los datos. Estas herramientas serán desarrolladas por el investigador y su equipo, y dependen de varios factores, como el tipo de estudio, la fuente de información, las variables a ser recolectadas, el medio en el cual se van a recolectar los datos, la edad de la población, las características de la población, el tipo de datos a ser recolectados, etc.

La herramienta creada por el investigador o grupo de investigadores es el pilar fundamental del proyecto, puesto que, si esta recolecta los datos de manera clara, precisa, completa y objetiva, lo más probable es que estos sean de utilidad para sacar nuestras conclusiones. Sin embargo, si estos datos son incompletos o subjetivos, pueden hacer que las conclusiones sean difíciles de abordar o asumir.

Los formularios o encuestas a ser utilizados pueden ser de creación propia del autor o autores, o bien pueden ser formularios o cuestionarios prediseñados, ya validados en otros estudios. Independientemente de la herramienta que se decida utilizar, esta debe ser validada para la población de estudio antes del inicio de la recolección de los datos, con el fin de minimizar errores u omisiones.

Desarrollo

Fuentes de información

Las fuentes de información se refieren al origen de la información, es decir, de dónde proviene esta, si la información es nueva o si ya está disponible en la literatura científica. Según esta característica de origen, las fuentes se dividen en primarias y secundarias.

Fuentes primarias:

Llamamos fuentes primarias a toda la información de nueva generación que no se encuentra publicada en ninguna literatura y que el investigador o los investigadores fueron los encargados de crearla para su artículo científico. Entre estas tenemos la observación y el interrogatorio.

La observación, como su nombre lo indica, es la técnica de recolección de información mediante la cual los investigadores contemplan al individuo de estudio con el fin de recaudar datos. Es la que sirve para obtener los datos más objetivos, ya que hace que el investigador o grupo de investigadores estén pendientes de todo lo que ocurre con el sujeto de investigación minuto a minuto. Esta técnica es bastante demandante en tiempo y recursos materiales y humanos, por lo que normalmente se utiliza para grupos pequeños. En conclusión, su objetivo principal es recolectar la información del presente inmediato.

Por ejemplo, un investigador quiere saber el efecto adictivo del tabaco, para lo cual elige un grupo de 20 ratones blancos a los que expone a altas concentraciones de humo de tabaco durante un mes. Una vez concluido este tiempo, cesa la exposición e inicia la recolección de los cambios de comportamiento de estos roedores.

Desde el inicio, los investigadores anotaron en una libreta todas las acciones, cambios o comportamientos que observaron, tanto los que vieron al inicio, cuando existió la exposición, como los que fueron consecuencia del cese de la misma, para poder compararlos y ver si existían realmente cambios significativos.

El interrogatorio, a pesar de que su fuente también es primaria, no proporciona datos tan exactos como la observación. Esto se debe a que se basa en la recolección de información en un tiempo no real, es decir, recolectamos información del pasado o del futuro, lo que puede generar sesgos o errores de memoria (cuando una persona no recuerda los datos con exactitud y distorsiona la realidad o la cambia por completo). Esto implica que dependemos en gran medida de los sujetos de investigación, de sus ganas de colaborar y de su capacidad de colaboración. Para que esta fuente nos dé la mejor información posible, debemos ser simples y concretos en la creación de nuestra encuesta, evitando preguntas en negativo o positivo, de doble sentido, parcializadas o poco claras, además de intentar ser sintácticos y concretos.

El interrogatorio puede ser realizado de manera directa o indirecta, dependiendo del tipo de estudio, la población a estudiar y el número de individuos.

Forma directa:

Cuando el investigador se dirige al lugar de recolección para obtener los datos necesarios en persona. La ventaja de esta técnica es que ayuda a evitar errores de interpretación y subjetividades, puesto que, si alguna pregunta es poco clara, los sujetos de estudio pueden aclarar sus dudas. La desventaja es el costo operacional, pues limita la recolección de datos en poblaciones extremadamente amplias.

Forma indirecta:

El investigador no requiere movilizarse, ya que las encuestas se envían a los sujetos de investigación de manera online o física. La ventaja es que, al no tener grandes costos operacionales, se puede realizar a un grupo realmente amplio de individuos. La desventaja es que, si los sujetos de investigación tienen dudas, el investigador no podrá aclararlas, lo que podría afectar los datos recolectados.

Ejemplo: un investigador quiere saber el nivel de ansiedad de los alumnos de primer año de fisioterapia. Para ello, decide utilizar una encuesta ya validada para diagnóstico de ansiedad. El cuestionario podría ser aplicado por el profesor de bioestadística (forma directa), o se podría enviar un enlace al formulario a los correos electrónicos de los alumnos (forma indirecta).

Fuentes secundarias:

Estas corresponden a todo el material ya creado y publicado por otros autores. El conocimiento puede ser nuevo o antiguo, pero no es de autoría directa del investigador, quien se limita a revisar libros, revistas, textos, artículos, etc., con el fin de dar respuesta a su pregunta de investigación. Este tipo de fuente no crea nuevos conocimientos. Las fuentes secundarias más utilizadas antiguamente eran los libros de texto o publicaciones en papel que se encontraban en grandes bibliotecas, pero hoy en día, con el avance de la tecnología, es más frecuente el uso de los repositorios bibliográficos. Los repositorios son bibliotecas digitales enfocadas en dar información en forma de artículos de texto, resúmenes, fotos, videos, etc.

En medicina específicamente, algunos de los repositorios más utilizados son Scielo, la Biblioteca Virtual en Salud y Google Académico (en español), y PubMed y Scopus (en inglés).

Métodos de recolección

Los métodos de recolección de información se dividen según la periodicidad, es decir, cuántas veces se va a recoger la información. En esta clasificación, se dividen en encuestas, experimentos, censos y sistemas de registro.

Encuestas:

Buscan recolectar datos existentes de manera metódica y ordenada mediante el interrogatorio. Se deben incluir solo las preguntas que ayuden al investigador a recolectar información útil para el estudio. Estas pueden ser:

- Estructuradas: todas las preguntas ya están prediseñadas.
- Semiestructuradas: parte de las preguntas están prediseñadas y otras surgen durante el proceso de la entrevista.
- Abiertas: las preguntas surgen según el avance de la entrevista.

Ejemplo: Un investigador quiere realizar un estudio sobre el hábito tabáquico y la hipertensión arterial. Para ello, realiza una encuesta centrada en la carga tabáquica y el desarrollo de la enfermedad en los individuos. Esta encuesta puede aplicarse en un solo momento.

Experimentos:

Se utilizan para recolectar información nueva mediante la observación de fenómenos a medida que estos se van generando. Se emplean en estudios experimentales o ensayos clínicos. Por ejemplo, se pueden seleccionar dos grupos de animales, exponiendo uno al humo del tabaco y al otro no, y mediante la observación, se puede determinar cuántos desarrollan la enfermedad.

Censo:

Es un método de recolección que se realiza a toda la población, generalmente cada 10 años. Este método busca conocer la composición de la población de diferentes territorios, co mo la densidad poblacional y la división por grupos de edad. Un ejemplo sería un censo para saber cuántas personas se espera que estudien en la universidad el próximo año.

Sistemas de registro:

Este método se caracteriza por ser continuo y dinámico, permitiendo al investigador actualizar la información permanentemente. Los ejemplos más

utilizados en medicina son el registro de nacidos vivos y de defunciones, así como las historias clínicas.

Diseño de formularios

Los formularios son los instrumentos mediante los cuales el investigador recolecta la información. Estos sirven para registrar los datos de forma escrita, y pueden ser digitales o físicos (en forma impresa), dependiendo de su propósito y forma de recolección. Además, pueden ser originales o validados, dependiendo de su origen. Cuando no existen formularios que se adapten al estudio, el investigador debe crear uno desde cero; sin embargo, si en algún estudio similar se creó un formulario que se adecúa a la pregunta de investigación, éste podría ser utilizado, siempre y cuando esté validado en el idioma en el cual se van a recolectar los datos. Si no es así, se debe validar el formulario previo a su uso, teniendo en cuenta que dicho proceso de validación es largo, puesto que se requiere de la realización de otro estudio paralelo enfocado únicamente a este fin.

Cuando el formulario es original, el investigador debe tener en cuenta que lo más importante es tener claro el grupo poblacional al que va a aplicar el formulario, para que todas las preguntas estén enfocadas a una correcta comprensión e interpretación. El formulario debe ser fácil de llenar para que todos los sujetos puedan completarlo en su totalidad sin complicaciones, utilizando el tamaño, tipo de letra, color y material adecuados.

El objetivo de los formularios es obtener información específica de un problema de investigación. Por lo tanto, el autor debe ser consciente de la importancia de recolectar solo datos relevantes y dejar de lado información innecesaria que podría causar confusión o exceso de datos, lo que podría desembocar en problemas al momento del análisis.

Los formularios se pueden dividir en segmentos dependiendo de los datos que se quieren adquirir, pero como norma general, deben tener al menos las siguientes partes:

- Título
- Descripción
- Datos de filiación o administrativos
- Datos enfocados a resolver el problema de estudio

Título:

El título debe ser corto, conciso, claro y completo. Generalmente está asociado al tema del proyecto de investigación y es la carta de presentación del formulario, por lo que debe contener toda la información relevante para los sujetos de investigación. Ejemplo: Si se quiere hacer un estudio sobre la ansiedad en los estudiantes de primer año de medicina de la Universidad de Cuenca, el título podría ser:

- "Ansiedad en estudiantes de primer año de medicina de la Universidad de Cuenca"
- "Prevalencia de ansiedad en estudiantes de primer año de medicina de la Universidad de Cuenca"
- "Desarrollo de ansiedad en estudiantes de primer año de medicina de la Universidad de Cuenca"

Descripción:

La descripción debe ir de:bajo del título y debe ser una breve explicación de los objetivos del estudio, seguido de una descripción de cómo se va a realizar la contribución del sujeto de investigación y cómo debe colaborar. Finalmente, se informa a los sujetos de investigación cómo se usarán esos datos.

Ejemplo: "El siguiente estudio pretende determinar la prevalencia de ansiedad en los estudiantes de primer año de medicina. Para ello, le solicitamos que nos ayude llenando el siguiente formulario, que es de opción múltiple y solo tiene una opción de respuesta por pregunta. Todos los datos recolectados serán analizados para generar un artículo científico que será publicado en revistas nacionales y/o internacionales".

Datos de filiación:

Los datos de filiación, también conocidos como datos administrativos o de identificación, son datos generales que caracterizan a la población. Estos incluyen:

- Edad
- Sexo
- Raza
- Instrucción

- Procedencia
- Residencia
- Estado civil
- Profesión

Condición socioeconómica:

Ejemplo: Para el estudio sobre ansiedad en estudiantes de primer año de medicina, los datos de filiación podrían incluir: cédula, sexo, raza, procedencia, residencia, edad, condición socioeconómica y grupo al que pertenecen.

Datos enfocados a resolver el problema de estudio

Los datos enfocados a resolver el problema de estudio son específicos y ayudarán a responder la pregunta de investigación. Aquí se deben evitar preguntas innecesarias, banales o que recojan datos por recolectar. La importancia de esta parte del formulario es crucial para la validez del estudio, por lo que se deben formular preguntas claras, concisas y útiles. Ejemplo: Para el estudio sobre ansiedad en estudiantes de primer año de Medicina, las preguntas importantes podrían ser:

- ¿Tiene familiares directos con ansiedad?
- ¿Duerme bien en las noches?
- ¿Tiene dificultades para conciliar el sueño?
- ¿Le sudan las manos o alguna otra parte del cuerpo en exceso?
- ¿Tiene dificultad para controlar sus emociones?

Ejemplo adicional: el proceso investigativo que realiza cada médico en su consultorio cuando recibe a un nuevo paciente se enfoca en determinar su diagnóstico. Se comienzan preguntando los datos de filiación: edad, sexo, raza, antecedentes, procedencia, residencia, etc. Luego se hacen preguntas relacionadas con los síntomas y la temporalidad de cada uno de ellos, además de los exámenes realizados (sangre, radiografías, tomografías, etc.). Después de recolectar toda la información, se toma una decisión diagnóstica para iniciar el tratamiento.

Tipos de preguntas en formularios

Dependiendo de la información que se desea recolectar, se pueden utilizar diferentes tipos de preguntas:

- Tipo de párrafo o abiertas: se da la opción al sujeto de investigación para escribir su respuesta con sus propias palabras. Son difíciles de analizar debido a la diversidad de respuestas, pero proporcionan información detallada. Son útiles en estudios cualitativos.
- Opcion múltiple: son las más utilizadas en investigaciones cuantitativas, donde el investigador ofrece varias opciones de respuesta y el sujeto elige la que mejor se adapte a su situación. Son fáciles de analizar y compilar, pero se pierden detalles personales.
- Numéricas: se utilizan para recolectar datos que se expresan en cantidades o medidas, con o sin decimales. Son fáciles de analizar y compilar.
- lineales: se utilizan cuando el autor quiere dar un orden a la respuesta. Un ejemplo común es la Escala Visual Análoga (EVA) para medir el grado de dolor. La escala va de 0 (sin dolor) a 10 (mucho dolor).
- Tipo matriz: se utiliza cuando se quieren hacer varias preguntas con las mismas opciones de respuesta.
- Video: se emplean cuando se quiere captar la atención de los niños o personas analfabetas, presentando información a través de videos.
- Audio: se utilizan para personas que no saben leer o que tienen problemas visuales.
- Imágenes: se utilizan especialmente con niños para captar su atención.

Prueba piloto

Antes de aplicar el formulario a la población de estudio, es recomendable realizar una prueba piloto con un número mínimo de personas (de 10 a 20). Esto permite detectar errores de forma (ortografía y gramática) y fondo (preguntas poco claras o que no recolectan lo que se desea). Durante esta etapa se pueden identificar preguntas faltantes o preguntas que están de más. Después de la prueba piloto, se deben corregir todos los errores antes de la recolección real de los datos.

Es importante señalar que la prueba piloto solo se realiza en formularios de nueva creación por parte del investigador, no en los que ya han sido validados previamente, ya que en el proceso de validación ya se realizó este paso.

Forma de recolección

Una vez que se ha finalizado el formulario, ya sea digital o físico, se debe decidir cómo se realizará la recolección de los datos:

- Recolección directa: el investigador aplica el formulario directamente a los sujetos de investigación. La ventaja de este método es que el investigador puede ayudar a resolver dudas en caso de que surjan. Sin embargo, hay un riesgo de que las respuestas se parcialicen si el sujeto de investigación responde lo que cree que el investigador quiere escuchar.
- Recolección indirecta: el formulario se envía de manera virtual o en papel, y el investigador no tiene contacto directo con los sujetos de investigación. La ventaja de este método es que las respuestas son imparciales, respetando el "ciego" del estudio. Sin embargo, si los sujetos de investigación tienen dudas, el investigador no podrá aclararlas, lo que podría afectar la calidad de las respuestas.

Ejemplo: Si el estudio se refiere a la ansiedad en los estudiantes de primer año de medicina, y una de las preguntas es "¿Cuántas veces ha llorado en la última semana?", si la encuesta es directa, el investigador puede ayudar a aclarar la pregunta si el sujeto ha perdido un familiar recientemente. Si la encuesta es indirecta, es posible que el sujeto responda afirmativamente, no por ansiedad, sino por el evento puntual ocurrido.

Conclusiones y recomendaciones

Las fuentes de información y la forma en que se recolecta son fundamentales para la validez de un estudio. Si los datos recolectados son adecuados y verídicos, las conclusiones serán confiables. Cualquier falencia en la recolección de datos invalidará completamente el estudio.

Se recomienda que, al generar la herramienta de recolección de datos, el investigador tenga claro el problema de investigación, las variables a recolectar y el tipo de cada una de estas variables. De esta forma, podrá crear un

formulario adecuado, evitando la recolección de datos irrelevantes que puedan causar confusión durante el análisis.

Referencias

Fayad, V. C. (1970). *Estadística médica y de salud pública* (10ª ed.). Universidad de los Andes.

Celis, A., y Labrada, V. (2014). *Bioestadística* (3ª ed.). Editorial El Manual Moderno.

Ruggiero, M., Bustamante, Y., y Mark, C. (2015). *Guía teórico-práctica de Bioestadística I*. Universidad Central de Venezuela.

Capítulo 6
Gestión de datos: diseño de base de datos

Introducción

En toda investigación cuantitativa, la recolección y organización de datos constituye una etapa esencial que garantiza la validez de los resultados. La calidad del análisis estadístico depende directamente de la calidad de los datos ingresados. Es por ello que el diseño de una base de datos adecuada es una tarea fundamental en el proceso investigativo, especialmente en el ámbito de las ciencias médicas y la salud pública.

En términos simples, una base de datos es una estructura organizada que permite almacenar información de forma sistemática y accesible. Esta estructura debe permitir la identificación clara de cada caso o unidad de análisis, la codificación precisa de las variables estudiadas y la facilidad para revisar, modificar y analizar la información cuando sea necesario.

En el contexto de la enseñanza de la Bioestadística en carreras de la salud, se utiliza el programa Microsoft Excel como una herramienta introductoria para el diseño de bases de datos, debido a su accesibilidad, facilidad de uso y compatibilidad con otros programas de análisis estadístico, como EpiDat 4.2. A través de esta herramienta es posible no solo ingresar datos, sino también realizar un control preliminar de calidad, aplicar filtros para detectar errores y estructurar la información de forma clara y ordenada.

Este capítulo tiene como propósito enseñar al estudiante a diseñar adecuadamente una base de datos utilizando Excel, reconociendo las etapas del proceso, identificando buenas prácticas para el ingreso de datos y conociendo herramientas para su validación, previa al análisis estadístico.

Desarrollo

Base de datos

Una base de datos es un conjunto estructurado de datos que permite almacenar y gestionar información de manera sistemática. En el ámbito de la estadística aplicada, una base de datos contiene los valores recolectados de las variables en estudio, organizados por casos u observaciones.

En la era predigital, la tabulación manual requería mucho tiempo y era propensa a errores. Actualmente, las hojas de cálculo como Excel permiten gestionar grandes volúmenes de datos de forma más eficiente y ordenada.

Etapas del diseño de una base de datos en Excel

El diseño de una base de datos debe realizarse con cuidado, ya que una estructura mal organizada puede dificultar el análisis estadístico, generar errores de interpretación y limitar la validez de los resultados. A continuación, se describen las principales etapas que deben seguirse para construir una base de datos funcional en Microsoft Excel:

Identificación de variables y tipo de medida:

Antes de crear la base de datos, es indispensable revisar el formulario de recolección de datos o el instrumento utilizado (por ejemplo, encuesta, ficha clínica o cuestionario). A partir de este instrumento, se deben listar todas las variables que serán recolectadas. Para cada variable es necesario identificar:

- Tipo de variable: si es cualitativa (por ejemplo, sexo, estado civil) o cuantitativa (por ejemplo, edad, peso, número de hijos).
- Tipo de medida: nominal, ordinal, de intervalo o de razón, según el nivel de información que aporta.

Esta etapa es clave para estructurar las columnas de la base de datos y para definir posteriormente los métodos estadísticos de análisis.

Creación de una columna de identificación:

La primera columna de la base de datos debe contener un identificador único para cada caso. Este identificador puede ser un número correlativo que representa el orden de ingreso de los datos o el número del formulario correspondiente. Esta columna permite localizar fácilmente errores, hacer verificaciones cruzadas y controlar el orden de los registros.

Creación de columnas para las variables:

Cada variable debe tener su propia columna en Excel. El nombre de la columna debe ser claro, breve y representativo de la variable que contiene. Por ejemplo, si se registra la edad del paciente, la columna puede llamarse "Edad"; si se registra el sexo, puede llamarse "Sexo". Evitar nombres largos o confusos. En lo posible, no usar espacios ni caracteres especiales.

Instrucciones para el ingreso de datos (codificación):

Es recomendable que cada columna contenga un comentario o nota que explique cómo se deben ingresar los datos. Por ejemplo, si la variable es "Sexo", la codificación puede ser: 0=Masculino y 1=Femenino. Esta codificación debe ser consistente en todos los registros. La claridad en la codificación evita errores de interpretación y facilita el análisis posterior.

Ingreso de datos:

El ingreso de datos debe realizarse con cuidado, fila por fila. Cada fila representa un caso individual (un paciente, un formulario, una observación), y cada columna representa una variable. Se debe verificar constantemente que los datos se registren en la columna correcta y con la codificación correspondiente.

Revisión y corrección de datos:

Una vez ingresados los datos, es fundamental realizar una revisión exhaustiva para identificar posibles errores. Excel permite aplicar filtros por columna para identificar rápidamente celdas vacías, valores fuera de rango, datos duplicados o mal codificados. Además, se puede convertir el conjunto de datos en una tabla de Excel para facilitar la navegación, filtrado y análisis preliminar.

Esta revisión puede complementarse con programas estadísticos como EpiDat 4.2, que permiten detectar errores mediante tablas de frecuencias, estadísticos descriptivos y gráficos como diagramas de caja.

Tabla 1

Etapas del diseño de una base de datos en Excel

Etapa	Descripción
Identificación de variables y su medida	Se debe definir qué variables serán incluidas, si son cualitativas o cuantitativas, y cómo serán medidas (por ejemplo, sexo como cualitativa categórica; edad como cuantitativa discreta).
Columna de identificación	Crear la primera columna que contenga la numeración o código de cada formulario o caso. Esta columna no se analiza estadísticamente, pero permite controlar y organizar los datos.
Columnas para las variables	Se crean columnas individuales para cada variable contenida en el formulario de recolección de datos.
Instrucciones de ingreso de datos	Es buena práctica incluir comentarios o notas explicativas sobre la codificación de cada variable (por ejemplo, 1=Masculino, 2=Femenino).
Ingreso de datos	Cada fila representa un caso, y en cada columna se ingresan los valores correspondientes de ese caso.
Revisión y corrección de datos	Usar herramientas como el formato tabla de Excel y los filtros automáticos para identificar valores vacíos, duplicados o fuera de rango.

Herramientas de validación de datos

Una vez construida e ingresada la base de datos, es indispensable aplicar un proceso de validación de los datos. Este paso busca garantizar que la información ingresada sea coherente, esté completa y no contenga errores de codificación, omisiones o valores extremos que puedan distorsionar los análisis estadísticos.

Validación de datos con Microsoft Excel:

- Conversión a tabla: insertar la base de datos como una tabla permite aplicar filtros automáticos por columna, facilitando la búsqueda de valores vacíos, duplicados o fuera de rango.
- Filtros por columna: permite revisar rápidamente los valores posibles en cada variable, ayudando a identificar categorías mal escritas o códigos erróneos.

- Formato condicional: destaca visualmente celdas con errores, valores atípicos o vacíos.
- Validación de datos: a través de esta opción, se pueden restringir los valores permitidos en una celda (por ejemplo, solo permitir "0" o "1" para la variable "Sexo"), reduciendo errores al momento del ingreso.

Validación de datos con EpiDat 4.2

Después de la revisión inicial en Excel, es recomendable realizar una validación estadística preliminar utilizando un software especializado como EpiDat 4.2, que permite identificar errores más complejos mediante análisis descriptivos. Las principales funciones útiles para la revisión son:

- Tablas de frecuencia: identifican categorías con pocos casos, valores inesperados o datos ausentes.
- Estadísticos descriptivos: proveen medidas como mínimo, máximo, media y desviación estándar. Si se detecta un valor de "150 años" en la variable "Edad", es probable que se trate de un error de digitación.
- Diagrama de cajas (*boxplot*): visualiza la distribución de los datos y destacan valores extremos (*outliers*) que pueden corresponder a errores.
- Estos procedimientos permiten una revisión cruzada entre lo visual (gráficos), lo cuantitativo (estadísticos) y lo categórico (frecuencias), favoreciendo la depuración completa de la base de datos.

Conclusiones y recomendaciones

El diseño de una base de datos es un paso esencial en todo proceso de investigación cuantitativa, ya que garantiza la correcta organización y almacenamiento de la información recolectada. Una base de datos bien estructurada no solo facilita el análisis estadístico, sino que también reduce los errores humanos durante el ingreso de datos y mejora la calidad de los resultados obtenidos.

Microsoft Excel, por su accesibilidad y versatilidad, constituye una herramienta valiosa en las primeras etapas del tratamiento de datos. Permite diseñar estructuras claras, aplicar codificaciones consistentes y realizar revisiones preliminares mediante filtros, formatos condicionales y validaciones.

Sin embargo, el diseño de una base de datos va más allá del simple registro de información; requiere una planificación cuidadosa que considere el tipo de variables, la codificación adecuada, el control de calidad del ingreso y la preparación para su posterior análisis en programas estadísticos como EpiDat. Por ello, el estudiante debe adquirir competencias no solo técnicas, sino también metodológicas, que le permitan comprender la lógica detrás de la construcción de una base de datos confiable.

Se recomienda utilizar una plantilla común para el diseño de bases de datos en investigaciones estudiantiles. Esta debe incluir columnas claras, instrucciones de codificación y una columna de identificación. Además, se debe asegurar que todos los participantes del proyecto (especialmente quienes ingresan los datos) reciban formación básica en codificación, digitación y uso de herramientas como Excel y EpiDat.

Finalmente, es fundamental no confiar únicamente en una revisión visual; se debe combinar el uso de Excel con herramientas de análisis preliminar como tablas de frecuencia, estadísticas descriptivas y gráficos para detectar errores. Todo dato sospechoso o inconsistente debe ser verificado contra su fuente original antes de proceder al análisis.

Se sugiere incluir un documento anexo al archivo de datos que explique la estructura de la base, las codificaciones utilizadas y cualquier decisión metodológica tomada en el diseño.

Referencias

Ministerio de Salud Pública del Ecuador. *Guía metodológica para la investigación en salud*. Quito: MSP; 2020.

Norman, G. R., y Streiner, D. L. (2015). *Bioestadística: Bases para el análisis de las ciencias de la salud* (4ª ed.). México: El Manual Moderno.

Organización Panamericana de la Salud (OPS), Xunta de Galicia. EpiDat 4.2: *Programa para análisis epidemiológico de datos*. Manual del usuario. Santiago de Compostela: Consellería de Sanidade; 2016.

Rodríguez, G., y Pérez, M. (2019). *Bioestadística para Ciencias de la Salud: Fundamentos y aplicaciones* (2ª ed.). Madrid: Editorial Médica Panamericana.

Roldán Fernández, J. V. (2025). *Presentación: Base de Datos*. Cátedra de Bioestadística 1. Universidad de Cuenca.

Capítulo 7
Gestión de los datos: depuración, codificación y cálculo de variables

Introducción

La gestión de datos es un proceso integral y sistemático que comprende todas las etapas relacionadas con el manejo de datos de una investigación, desde su planificación y recolección, almacenamiento, organización, depuración, hasta su análisis, conservación y eventual eliminación.

La gestión de datos no se limita únicamente a la construcción de una base de datos, sino que abarca todo el análisis, asegurando que cada etapa del proceso investigativo se realice bajo criterios técnicos y éticos apropiados procurando la calidad de los resultados.

Una vez elaborada la base de datos recolectada, se debe realizar una depuración, que consiste en la identificación y corrección de errores o inconsistencias que podrían alterar los resultados, asegurando que los datos que serán utilizados en el análisis sean fiables. Posteriormente, podría ser necesario realizar el cálculo de variables que permitan generar nueva información a partir de los datos existentes, como el cálculo del índice de masa corporal (IMC) a partir de datos como el peso y la talla.

Por último, se considera la categorización de variables, un proceso de reorganización de la información para facilitar su análisis, por ejemplo, a partir de la edad en años cumplidos (un dato cuantitativo) se pueden generar categorías.

Una gestión adecuada de los datos no solo mejora la calidad del análisis estadístico, sino que también garantiza una interpretación válida y útil de los resultados en el proceso investigativo. El objetivo de este capítulo es recapitular los pasos a seguir para tener una base de datos sólida y fiable.

Desarrollo

Depuración de bases de datos

Una vez que se cuenta con una base de datos, ya sea proveniente de formularios físicos o digitales, el ejercicio de realizar la depuración es un paso fundamental antes de iniciar el análisis y el procesamiento de los datos. La depuración de bases de datos consiste en revisar varios aspectos clave para garantizar la calidad de la información.

Uno de los puntos principales a verificar es que los campos obligatorios hayan sido correctamente llenados, es decir, que las variables principales del estudio contengan la información necesaria. Por ejemplo, si estamos realizando un estudio sobre depresión, ansiedad y estrés, todas las preguntas relacionadas con estos tres eventos de salud deben estar respondidas. En caso contrario, se debe analizar la posibilidad de eliminar ese registro de la base de datos.

Sin embargo, es importante recalcar que, si bien las variables sociodemográficas complementarias (como edad, sexo o nivel de instrucción) pueden no estar completas en todos los formularios, esto no invalida el registro cuando las variables principales del estudio sí fueron respondidas. En estos casos, se recomienda mantener el registro y aclarar en la presentación de resultados por qué ciertas variables tienen una cantidad menor de respuestas. Esto aporta transparencia al análisis y evita interpretaciones erróneas.

En el caso de las variables cuantitativas, cuando los formularios físicos o digitales permiten que el encuestado ingrese los datos de manera directa, es importante identificar valores extremos o inconsistentes que puedan afectar estadísticos como la media o el promedio. Para ello, es fundamental que, de forma previa, los investigadores definan criterios claros de exclusión frente a este tipo de errores, ya que esto permitirá tomar decisiones adecuadas en el momento de depurar y analizar los datos recolectados.

Si la base de datos proviene principalmente de registros físicos, es importante revisar que las categorías de las variables cualitativas estén homologadas a un solo formato. Por ejemplo, si se encuentran registros con términos como “hombre”, “masculino” o simplemente la letra “M”, todos deben ser unificados bajo una misma categoría, por ejemplo: “hombre”.

En las variables codificadas, es importante revisar que dicha codificación esté acorde con el manual de códigos previamente establecido por los investigadores. Cuando los cuestionarios contemplan preguntas como, por ejemplo,

fecha de nacimiento y edad, se puede realizar una validación cruzada, es decir, verificar que entre ambas respuestas exista coherencia. Otro ejemplo es cuando se pregunta si la persona tiene hijos y, posteriormente, se incluyen preguntas relacionadas con esos hijos. En este caso, no debería existir ninguna respuesta en las preguntas posteriores si el número de hijos reportado es cero.

Los formularios digitales generan las bases de datos, y los nombres de las columnas suelen aparecer con los textos completos de cada pregunta. Por esta razón, es importante renombrar cada columna utilizando nombres cortos, sin caracteres especiales como tildes, eñes o espacios, para evitar problemas posteriores con los diferentes programas que procesan las bases de datos.

Finalmente, se debe revisar con mucha precisión, en la medida de lo posible, si existen duplicados de registros.

Tratamiento de datos faltantes

En el proceso de gestión de datos es frecuente encontrar datos faltantes, es decir, valores no registrados para una o más variables en determinados sujetos. Por ejemplo, un participante pudo no haber respondido su edad, o no haberse registrado su peso al momento de la encuesta. Los datos pueden faltar por diversas razones, entre ellas:

- El participante no respondió una pregunta.
- El registro clínico estaba incompleto.
- Se produjo un error durante la digitación o el almacenamiento de la información.

La presencia de datos faltantes no invalida automáticamente un estudio, pero sí requiere una decisión metodológica adecuada. Antes de decidir cómo se procederá con el dato faltante, el investigador debe evaluar:

- La cantidad de datos faltantes.
- El tipo de variable afectada.
- La importancia de la variable dentro del estudio.
- El posible patrón de ausencia de los datos.

Una vez identificados los datos faltantes, el investigador debe decidir cómo manejarlos. Existen tres estrategias generales:

1. Eliminar los registros incompletos: consiste en excluir del análisis a los sujetos que presentan datos faltantes en una o más variables. Esta estrategia es apropiada cuando la cantidad de datos faltantes es pequeñao las variables faltantes no son centrales para el estudio. Es la opción más simple y más segura en estudios descriptivos, pero puede reducir el tamaño de la muestra.

2. Mantener los registros sin modificar los datos faltantes: en algunos análisis descriptivos, los registros con datos faltantes pueden mantenerse, aclarando que ciertas variables tendrán un número menor de observaciones. En estos casos, es obligatorio indicar el número de datos válidos (n) utilizado en cada análisis.

3. Imputación de datos: consiste en reemplazar un dato faltante por un valor estimado, con el fin de poder incluir ese registro en el análisis. Este procedimiento requiere criterios estadísticos claros y puede introducir sesgos si no se aplica correctamente. Ejemplo: si falta el peso de un paciente, podría reemplazarse por el promedio de peso del grupo.

Codificación y decodificación de variables

El proceso de codificación y decodificación de variables se aplica únicamente a las variables cualitativas, ya que son las que contienen diferentes categorías. Este proceso consiste en asignar un valor numérico a cada categoría con la finalidad de facilitar ciertos análisis, como la sumatoria de respuestas para obtener un puntaje total en instrumentos específicos.

Ejemplo: en el caso del DASS-21, en el que 7 preguntas están destinadas a medir el nivel de depresión, las opciones de respuesta suelen ser: nunca, a veces, a menudo y casi siempre. Estas pueden ser codificadas con los valores 0, 1, 2 y 3, respectivamente. Esta codificación permite que, al finalizar la aplicación del instrumento, se pueda realizar una sumatoria de las respuestas para obtener un puntaje total que refleje el nivel de depresión del encuestado.

Una recomendación importante es que, independientemente de donde provenga la base de datos, se mantenga en una hoja de cálculo la base original, generando una segunda hoja de cálculo para la base decodificada y una tercera con la base codificada. Esto facilita el proceso de análisis y procesamiento de los datos en los diferentes programas estadísticos.

Categorización de una variable cuantitativa

Para las variables cuantitativas, existe la posibilidad de categorizarlas. Un ejemplo común es la variable edad, la cual, según recomendaciones de la Organización Mundial de la Salud (OMS), puede organizarse en rangos como:

- De 0 a 4 años
- De 5 a 9 años
- De 10 a 14 años
- De 15 a 19 años

Y así sucesivamente. Este proceso se realiza en función de la naturaleza de la variable y los objetivos del estudio. Otra posibilidad de categorización ocurre con el DASS-21: una vez calculado el puntaje total de depresión, se clasifica en diferentes categorías como:

- Depresión leve
- Moderada
- Severa
- Extremadamente severa

Es decir, el proceso de categorización consiste en tomar una columna con valores numéricos continuos y generar una nueva variable categórica que agrupe esos valores en rangos previamente definidos. Los valores de los rangos pueden ser renombrados según sea el caso y se pueden mostrar en las tablas o gráficos.

Cálculo de una variable

Dentro de la base de datos, también se pueden crear nuevas variables a partir de las existentes. Un ejemplo claro es tomar la columna de peso y la columna de talla, y a partir de ellas calcular una nueva columna con la variable índice de masa corporal (IMC). El IMC se obtiene aplicando la fórmula:

$$\text{IMC}=\frac{peso\ (kg)}{talla\ (m)^2}$$

Conclusiones y recomendaciones

La gestión de datos constituye un pilar fundamental en el proceso de investigación, puesto que asegura la calidad y consistencia de los datos a analizar. Una depuración rigurosa, la codificación de datos cualitativos, la categorización de variables cuantitativas y el cálculo de nuevas variables permiten transformar datos puros en el sustrato para el análisis estadístico.

Es fundamental que los investigadores se familiaricen con el proceso de gestión de datos y la correcta aplicación de herramientas como Microsoft Excel y EpiDat para garantizar la precisión de los resultados. Además, siempre que sea posible, deben realizarse verificaciones cruzadas de los datos mediante análisis descriptivos, filtros y validaciones, ya sea en Excel o en otros programas especializados.

Referencias

Berenson, M. L., Levine, D. M., y Krehbiel, T. C. (2006). *Fundamentos de estadística* (6ª ed.). México: Pearson Educación.

Celis de la Rosa, A. J., y Labrada Martagón, V. (2014). *Bioestadística* (3ª ed.). Editorial El Manual Moderno.

Martínez González, M. A., et al. (2020). *Bioestadística amigable (4ª ed.). Madrid: Elsevier.*

Organización Panamericana de la Salud. (2016). EPIDAT 4.2: *Programa para análisis epidemiológico de datos tabulados* [Manual del usuario]. OPS.

Szklo, M., y Nieto, F. J. (2007). *Epidemiología intermedia: conceptos y aplicaciones* (2ª ed.). Buenos Aires: Lugar Editorial.

Wayne, D. W. (2009). *Bioestadística: base para el análisis de las ciencias de la salud* (7ª ed.). México: Limusa.

Capítulo 8
Niveles de análisis y tipos de pruebas

Introducción

En el ámbito de la investigación científica, especialmente en las ciencias de la salud, el análisis de los datos representa una etapa crítica que permite transformar los datos recolectados en conocimiento significativo. Una vez que se ha definido el problema de investigación, establecido el diseño del estudio y recolectada la información pertinente, el siguiente paso fundamental es analizar esos datos con rigurosidad para responder a las preguntas planteadas.

El análisis de datos estadísticos implica una serie de procedimientos técnicos que permiten resumir, organizar, explorar e interpretar los datos obtenidos de una muestra o población. Estos procedimientos se estructuran en dos grandes niveles:

- Análisis descriptivo, que permite visualizar de forma clara y resumida las características principales de los datos mediante medidas numéricas y representaciones gráficas.
- Análisis inferencial, que permite extender o generalizar los hallazgos obtenidos en una muestra hacia la población de interés, evaluando hipótesis, identificando asociaciones o prediciendo desenlaces con base en modelos estadísticos.

La correcta elección de los métodos de análisis depende de diversos factores, como el tipo de variable, la escala de medición, la distribución de los datos y el tamaño de la muestra. Esta elección también determina si se deben aplicar pruebas estadísticas paramétricas, que requieren supuestos más estrictos (como normalidad y homogeneidad de varianzas), o pruebas no paramétricas, que son útiles cuando esos supuestos no se cumplen.

Comprender estos niveles de análisis y las herramientas estadísticas asociadas no solo permite realizar una adecuada interpretación de los resultados, sino

que también aporta validez científica a las conclusiones que se derivan de una investigación. En el contexto médico y epidemiológico, donde las decisiones clínicas y de salud pública deben basarse en evidencia sólida, el análisis correcto de los datos se convierte en un componente esencial de la práctica basada en evidencias.

Este capítulo tiene como propósito brindar al estudiante de medicina los fundamentos teóricos y prácticos necesarios para entender y aplicar correctamente los distintos tipos de análisis estadístico, así como para seleccionar la prueba estadística más adecuada según la naturaleza de los datos disponibles.

Desarrollo

Análisis de datos

La estadística es la ciencia que proporciona métodos objetivos para recoger, organizar, analizar e interpretar datos con el fin de obtener conclusiones válidas y tomar decisiones razonadas. En el campo de la salud, el análisis estadístico permite describir fenómenos, estimar parámetros poblacionales, identificar asociaciones entre variables y evaluar la eficacia de intervenciones sanitarias o tratamientos médicos.

El análisis de datos es una etapa crítica en el proceso de investigación. A través de él, se transforma la información cruda en evidencia significativa que guía la toma de decisiones clínicas, epidemiológicas o de gestión en salud. Su aplicación rigurosa aporta objetividad y sustento empírico a los resultados obtenidos.

Proceso de análisis estadístico

- El análisis de datos sigue un proceso sistemático que puede describirse en las siguientes fases:
- Revisión y limpieza de datos: validar la consistencia, detectar errores y preparar la base de datos.
- Análisis descriptivo: resumir y explorar los datos mediante estadísticas y gráficos.
- Evaluación de supuestos: verificar si se cumplen los criterios para aplicar determinadas pruebas (por ejemplo, distribución normal).

- Análisis inferencial: aplicar pruebas estadísticas para contrastar hipótesis o establecer relaciones.
- Interpretación de resultados: traducir los resultados numéricos en conclusiones que respondan a los objetivos del estudio.
- Presentación de los hallazgos: elaboración de tablas, gráficos e informes para la comunicación de resultados.

Programas utilizados en el análisis de datos

El análisis estadístico puede realizarse manualmente para ejercicios simples, pero en la práctica profesional y académica se emplean programas estadísticos especializados. Estos permiten mayor precisión, rapidez y capacidad para manejar grandes volúmenes de datos. A continuación, se presentan algunos de los programas más utilizados y recomendados:

Tabla 1

Programas de acceso libre

Propiedad	Explicación
EPIDAT 4.2	Desarrollado por la Junta de Galicia, útil para análisis epidemiológicos y clínicos.
Epi Info 7	Desarrollado por la CDC, útil para encuestas, gestión de datos y análisis básico.
PSPP	Alternativa libre a SPSS, interfaz similar, útil para análisis estadísticos básicos.
R / RStudio	Potente, flexible y extensible mediante paquetes; ideal para análisis avanzados.
JASP	Interfaz amigable basada en R, permite análisis frecuentistas y bayesianos.
Jamovi	Basado también en R, muy intuitivo, útil para enseñanza y análisis profesional.
SOFA Statistics	Enfocado en la presentación de resultados; adecuado para análisis básicos.
BioStat	Software libre para estadística biomédica básica, incluye tests comunes.

Tabla 2

Programas de licencia comercial

Propiedad	Explicación
SPSS	Muy usado en ciencias sociales y de la salud, con amplia documentación.
Stata	Robusto para análisis econométrico y epidemiológico, con lenguaje de comandos.
Minitab	Intuitivo, usado en control de calidad y también útil para análisis biomédico.
SAS	Potente para análisis complejos y grandes bases de datos; uso común en biomedicina.
GraphPad Prism	Muy utilizado en investigación biomédica y farmacología por su enfoque gráfico.
Statistica	Versátil, con capacidades gráficas avanzadas y módulos especializados.

La elección del programa dependerá de diversos factores: nivel de experiencia del investigador, tipo de análisis requerido, presupuesto disponible y preferencias institucionales. Para fines académicos y de docencia, programas como Epi Info, EPIDAT, PSPP, JASP y Jamovi ofrecen una excelente alternativa libre y funcional.

Niveles de análisis

En estadística, los niveles de análisis se refieren a los diferentes enfoques mediante los cuales se examinan los datos con el fin de obtener información relevante. Estos niveles permiten organizar el análisis en función de los objetivos del estudio, el tipo de datos disponibles y las preguntas de investigación. Los dos grandes niveles de análisis son:

Estadística descriptiva

La estadística descriptiva tiene como objetivo resumir, organizar y presentar los datos de forma clara, permitiendo una primera aproximación a la información contenida en una base de datos. No pretende generalizar los hallazgos a toda la población, sino describir lo que se observa en la muestra.

La descripción se puede hacer utilizando dos tipos de procedimientos: mediante el cálculo de índices estadísticos que son números que resumen de modo sencillo la información; o bien utilizando representaciones gráficas.

Tabla 3

Principales índices estadísticos

Variables	**Índices**
Cualitativas	Frecuencia absoluta: conteo de los casos.
	Frecuencia relativa: conteo de los casos de una categoría con relación a otra categoría o el total (razón, proporción, tasas)
Cuantitativas	Medidas de tendencia central: valor alrededor del cual se agrupan los datos (media, mediana, moda).
	Medidas de dispersión: cuantifican la variabilidad de la distribución de los datos (varianza, desviación estándar, rango)
	Medidas de posición: ubicación en la distribución de los datos (cuartiles y percentiles).
	Medidas de forma: describen aspectos de la distribución (asimetría y curtosis).

Tabla 4

Principales gráficos estadísticos

Variables	Índices
Cualitativa nominal	Gráficos de barras: representa frecuencias por categorías con barras separadas.
	Gráfico de sectores: muestra proporciones relativas como partes de un círculo.
	Pictograma: usa símbolos o dibujos repetidos para representar frecuencias.
	Gráfico de anillos: variante del gráfico de pastel en forma de anillo.
Cualitativa ordinal	Gráfico de barras ordenado: similar al de barras, pero ordenado según jerarquía o frecuencia.
	Gráfico de escalera: categorías ordenadas en forma progresiva.
Cuantitativa discreta	Gráfico de barras: frecuencia de valores numéricos discretos (por ejemplo, número de hijos).
	Polígono de frecuencias: línea que une los puntos medios de cada clase; útil para comparar distribuciones.
Cuantitativa continua	Histograma: barras contiguas que representan intervalos de datos continuos.
	Diagrama de cajas: resume distribución, mediana, cuartiles y valores atípicos.
	Ojivas: gráfico acumulativo que muestra frecuencias acumuladas.
	Polígono de frecuencias: representa frecuencias mediante líneas; útil para comparar varios grupos.
	Gráfico de líneas: muestra evolución de valores en el tiempo (series temporales).
Dos variables (comparativo)	Gráfico de barras agrupadas/apiladas: comparación de categorías entre grupos.
	Diagrama de cajas: compara la distribución de una variable cuantitativa en distintos grupos cualitativos.
	Diagrama de dispersión: representa la relación entre dos variables cuantitativas.
	Gráfico de líneas comparativas: muestra cambios simultáneos en dos o más series.

Figura 1

Diagrama de barras que muestra la frecuencia de diferentes diagnósticos en pacientes hospitalizados

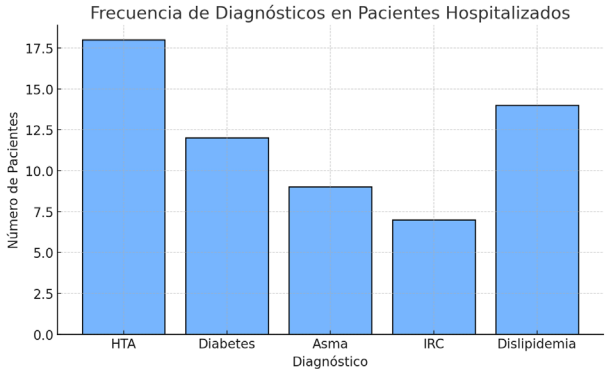
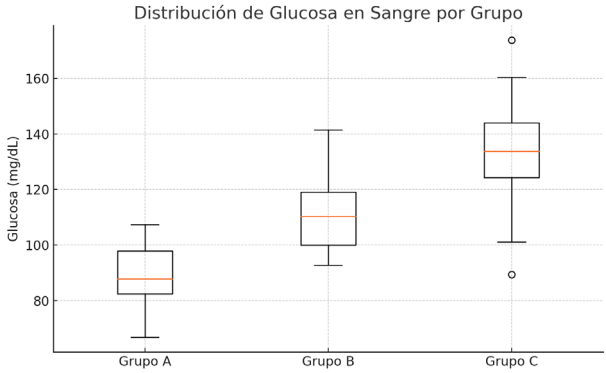


Figura 2

Diagrama de cajas que muestra la distribución de los niveles de glucosa en sangre en tres grupos distintos de pacientes



Estadística inferencial

La estadística inferencial permite extraer conclusiones sobre una población a partir de una muestra representativa, evaluando hipótesis, estimando parámetros y determinando relaciones entre variables. Se basa en la probabilidad y el uso de pruebas estadísticas que permiten tomar decisiones con un cierto nivel de confianza.

Funciones principales:

- Estimar parámetros poblacionales: como la media o proporción de una población a partir de la muestra.
- Contrastar hipótesis estadísticas: se plantea una hipótesis nula (H_0) y una alternativa (H_1), y se decide cuál es más plausible con base en los datos.
- Determinar relaciones: de dependencia o independencia entre variables (ej. Chi-cuadrado); comparación de grupos (ej. T de Student, ANOVA); correlación y regresión.

Tipos de análisis inferencial

- Bivariado: evalúa la relación entre dos variables. Ejemplo: ¿el sexo se asocia con la presencia de hipertensión?
- Multivariado: analiza la relación entre tres o más variables simultáneamente. Ejemplo: ¿la edad, el índice de masa corporal y la actividad física influyen en el nivel de glucosa?

Tabla 5

Pruebas bivariantes según tipo de variable y relación entre muestras

Variable dependiente	Muestra	Supuestos	Prueba
Cualitativa	Independiente	Frec. esperadas ≥ 5	Chi-cuadrado
Cualitativa	Independiente	Frec. esperadas < 5	Fisher
Cuantitativa	Independiente	Normalidad + varianzas iguales	U de Mann Withney

Cuantitativa	Independiente	Normalidad sin varianzas iguales	t de Welch
Cuantitativa	Independiente	No normal	Prueba de Wilconxon
Cuantitativa	Apareadas	Normalidad	Prueba de Kruskal Walls
Cuantitativa	Apareadas	No normal	Prueba de Friedman

Importancia del diseño y la distribución

El uso correcto de técnicas inferenciales requiere:

- Muestreo adecuado.
- Conocimiento de la distribución de los datos (normal o no).
- Tipo de variable (cualitativa, ordinal, cuantitativa).

Tabla 6

Comparación entre ambos niveles de análisis

Variables	**Estadística descriptiva**	**Estadística inferencial**
Objetivo	Describir los datos observados	Generalizar a la población de referencia
Aplicación	Resumen de variables	Pruebas de hipótesis, estimaciones
Datos requeridos	Muestra	Muestra (representativa de la población)
Técnicas	Frecuencias, medidas y gráficos	Pruebas estadísticas, intervalos de confianza
Ejemplo	Promedio de edad en la muestra	Comparar promedios entre grupos

Pruebas paramétricas y no paramétricas

En el análisis estadístico inferencial, la elección de la prueba adecuada depende de varios factores, entre los que destacan el tipo de variable, la escala de medición, la distribución de los datos y el tamaño de la muestra. En función de estos criterios, las pruebas estadísticas se dividen en paramétricas y no paramétricas.

Las pruebas paramétricas son aquellas que se basan en supuestos específicos sobre la distribución de los datos, especialmente que siguen una distribución normal. Se aplican principalmente a variables cuantitativas continuas y ofrecen mayor potencia estadística, es decir, mayor capacidad para detectar diferencias reales cuando estas existen.

Supuestos principales:

- La variable es cuantitativa continua (intervalo o razón).
- Los datos siguen una distribución normal.
- Homogeneidad de varianzas entre grupos (prueba de Levene).
- Muestra aleatoria e independiente.
- Tamaño muestral moderado o grande (n > 30) (las pruebas paramétricas pueden utilizarse incluso con tamaños muestrales pequeños, como la T de Student, siempre que los supuestos de normalidad aproximada y homogeneidad de varianzas sean razonables).

Tabla 7

Ejemplos de pruebas paramétricas

Objetivo	Prueba paramétrica	Ejemplo clínico
Comparar medias de dos grupos independientes	T de Student	Comparar niveles de glucosa entre pacientes diabéticos y no diabéticos.
Comparar medias de más de dos grupos independientes	ANOVA (una vía)	Comparar presión arterial entre tres grupos de tratamiento.

Evaluar relación lineal	Correlación de Pearson	Relación entre edad y nivel de colesterol.
Analizar predicción de variable	Regresión lineal	Predecir peso en base a talla y edad.

Las pruebas no paramétricas se utilizan cuando no se cumplen los supuestos de normalidad, o cuando se trabaja con variables ordinales o categóricas, o muestras pequeñas. Aunque tienen menor poder estadístico, son más robustas ante violaciones de supuestos.

Situaciones donde se aplican:

- Datos de escala ordinal (como niveles de dolor: leve, moderado, severo).
- Datos cuantitativos con distribución no normal.
- Se pueden aplicar con un tamaño muestral pequeño ($n < 30$).
- Algunas pruebas se aplican cuando las varianzas no son homogéneas.
- Presencia de valores atípicos (*outliers*).

Tabla 8

Ejemplos de pruebas no paramétricas

Tipo de comparación o análisis	**Prueba**	**Ejemplo clínico**
Comparar dos grupos independientes	U de Mann-Whitney	Comparar duración de hospitalización entre hombres y mujeres.
Comparar dos grupos apareados	Wilcoxon	Comparar presión arterial antes y después de un tratamiento.
Más de dos grupos independientes	Kruskal-Wallis	Comparar dolor en tres grupos con distintos analgésicos.

Más de dos grupos apareados	Friedman	Comparar nivel de glucosa antes, durante y despúes de un programa nutricional.
Tablas RxC (frecuencias esperadas ≥ 5)	Chi-cuadrado (χ^2)	Evaluar la asociación entre sexo (hombre/mujer) y presencia de hipertensión (sí/no).

Tabla 9

Comparación entre prueba

Criterio	**Prueba paramétrica**	**Prueba no paramétrica**
Tipo de variable	Cuantitativa continua	Cuantitativa discreta, cualitativa ordinal, cualitativa
Distribución de datos	Normal (campana de Gauss)	No normal o desconocida
Medidas utilizadas	Media y desviación estándar	Mediana y cuartiles
Tamaño muestral	Usualmente requiere muestras grandes (> 30).	Se pueden utilizar para muestras pequeñas (< 30).
Homogeneidad de varianzas	Requiere varianzas similares usualmente.	No requiere homogeneidad en algunas pruebas.
Potencia estadística	Alta (mayor capacidad de detectar diferencias)	Menor que la paramétrica
Ejemplos	T de Student, ANOVA, Pearson	U de Mann-Whitney, Wilcoxon, Spearman, Chi Cuadrado

Decisión según la distribución de los datos

Un paso clave antes de aplicar pruebas paramétricas es verificar la normalidad de los datos mediante:

- Gráficos: histograma, Q-Q plot.
- Pruebas estadísticas: Kolmogorov-Smirnov, Shapiro-Wilk.
- Si los datos son normales, se utilizan pruebas paramétricas.
- Si no son normales, se opta por pruebas no paramétricas.

Conclusiones y recomendaciones

El análisis de los datos constituye un componente esencial en todo proceso de investigación, ya que permite transformar datos crudos en información relevante, comprensible y útil para la toma de decisiones clínicas, epidemiológicas o de política sanitaria. A lo largo de este capítulo se han abordado los fundamentos que sustentan los dos grandes niveles de análisis estadístico: el descriptivo, orientado a resumir y organizar la información observada, y el inferencial, cuyo propósito es extraer conclusiones válidas y generalizables hacia una población de referencia.

La estadística descriptiva proporciona herramientas numéricas y gráficas para sintetizar los datos, lo cual es fundamental como paso previo a cualquier análisis más complejo. Por su parte, la estadística inferencial permite probar hipótesis, estimar parámetros y evaluar relaciones entre variables, contribuyendo a validar científicamente los hallazgos observados.

Un aspecto crítico abordado ha sido la elección entre pruebas paramétricas y no paramétricas. Esta decisión no solo depende del tipo de variable o del número de grupos a comparar, sino también de la distribución de los datos, la escala de medición y el tamaño muestral. La comprensión adecuada de estos criterios es indispensable para evitar errores metodológicos y garantizar la validez estadística de los resultados.

En el contexto médico, un análisis estadístico bien realizado permite identificar factores de riesgo, evaluar la efectividad de tratamientos, predecir desenlaces clínicos y guiar la toma de decisiones basadas en evidencia. Así, el conocimiento y la aplicación de estos métodos se convierten en una competencia esencial para los profesionales de salud.

Se recomienda adquirir dominio progresivo de los fundamentos estadísticos, comenzando por el análisis descriptivo y avanzando hacia la inferencia, comprendiendo no solo los cálculos, sino también su interpretación clínica. También es crucial verificar siempre los supuestos estadísticos antes de

aplicar una prueba, especialmente los relacionados con la distribución de los datos y el tipo de variable. La aplicación incorrecta puede llevar a conclusiones erróneas. Además, se debe utilizar herramientas informáticas adecuadas para el análisis de datos, priorizando aquellas con soporte técnico y documentación científica.

Se sugiere familiarizarse con al menos un software de análisis libre (como R, Jamovi o Epidat) y uno de licencia institucional (como SPSS o STATA).

Se debe complementar los resultados cuantitativos con su interpretación clínica, sin caer en el error de sobrevalorar la significancia estadística por encima de la relevancia médica o del contexto del estudio. Fomentar el pensamiento crítico y la ética en el análisis de datos, evitando manipulación de resultados, selección sesgada de pruebas o presentación intencionada de gráficos para apoyar hipótesis sin sustento.

Es útil practicar regularmente con ejercicios reales o simulados, para fortalecer las habilidades analíticas, la toma de decisiones estadísticas y la comprensión integral del proceso de investigación.

Por último, se recomienda consultar con un experto en estadística o un bioestadístico cuando se enfrenten análisis complejos, estudios multivariantes o diseños poco convencionales, especialmente en investigación clínica y estudios de intervención.

Referencias

Centers for Disease Control and Prevention. (2025). *Epi Info™ 7.* [Internet] Disponible en: https://www.cdc.gov/epiinfo/index.html

JASP Team. (2024). *JASP* (Versión 0.18.1) [Software]. Disponible en: https://jasp-stats.org

Pagano, M., y Gauvreau, K. (2006). *Principios de bioestadística*. México: Thomson.

Pérez, C. (2010). *Bioestadística aplicada a las ciencias de la salud* (2ª ed.). Madrid: Díaz de Santos.

Roldán, J. V. (2024). *Niveles de análisis y pruebas estadísticas.* Presentación de clase, Bioestadística 1.

The jamovi project. (2024). *Jamovi* (Versión 2.4) [Software]. Disponible en: https://www.jamovi.org

Xunta de Galicia, OPS/OMS. (2024). *Epidat 4.2: programa para análisis epidemiológico de datos tabulados.* [Internet] Disponible en: https://www.sergas.gal/Saude-publica/Epidat

Wayne, D. W. (2009). *Bioestadística: base para el análisis de las ciencias de la salud* (7ª ed.). México: Limusa.

Capítulo 9
Análisis descriptivo numérico: variables cualitativas

Introducción

Una vez que los investigadores han depurado y codificado la base de datos, en la cual se incluyen todas las variables, tanto cualitativas como cuantitativas, el siguiente paso es el análisis de los datos utilizando estadísticos específicos diseñados para tal fin. Este análisis debe realizarse desde un enfoque general hacia uno más específico, de modo que el lector pueda comprender el estudio de manera adecuada. Antes de comenzar con el análisis, es importante tener claro qué se desea comunicar a los lectores y qué datos son esenciales para el entendimiento del estudio.

En este capítulo, nos enfocaremos únicamente en el análisis de variables cualitativas, es decir, aquellas que no son numéricas, y cómo presentar los resultados de estas variables mediante tablas e índices estadísticos.

Desarrollo

Conceptos generales

Para comenzar con el análisis de las variables cualitativas, es necesario recordar algunos conceptos básicos que, aunque generales, son fundamentales para comprender este capítulo.

- Variables: una variable es cualquier propiedad o característica que presenta un objeto o persona y que puede tomar diferentes valores en distintos momentos, dependiendo de las condiciones dadas.
- Tipos de variables: se pueden clasificar en cualitativas y cuantitativas (Figura 1).

Variables cualitativas:

- Nominales**:** son aquellas que solo indican diferencias entre categorías. Por ejemplo, la variable "color del objeto", que puede tomar las categorías azul, verde o rojo. Las variables nominales muestran una relación de equivalencia, es decir, todos los elementos u objetos que pertenecen a una categoría comparten una característica común.
- Ordinales**:** estas variables, además de indicar una característica, también tienen un orden o jerarquía. Un ejemplo es el "nivel socioeconómico" de los encuestados, clasificado como alto, medio o bajo.

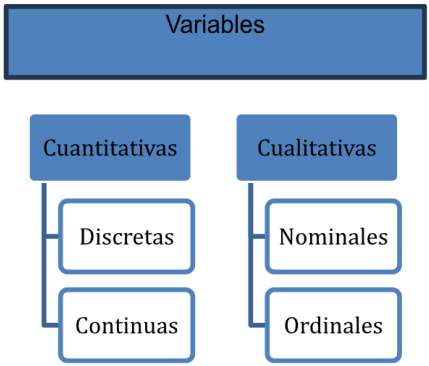
Variables cuantitativas:

Incluyen las características de las variables nominales y ordinales, pero además poseen una unidad numérica de medición definida. Se dividen en:

- Discretas: toman valores enteros, como 1, 2, 3, etc. Un ejemplo de variable discreta es el "número de hijos".
- Continuas: pueden tomar cualquier valor dentro de un rango determinado, como 1.1, 1.2, 1.3, etc. Un ejemplo de variable continua es la "altura en metros".

Figura 1

Clasificación de las variables



Las variables también pueden clasificarse como de intervalo o de razón. Las de intervalo emplean números cardinales, tiene la misma distancia entre dos puntos y emplea un cero arbitrario como punto de partida (tienen números positivos y negativos). Las variables de razón poseen las características de las variables de intervalo, pero con un punto cero verdadero, donde el cero significa ninguno o ausencia de la característica.

Tabla 1

Características de los niveles de medición

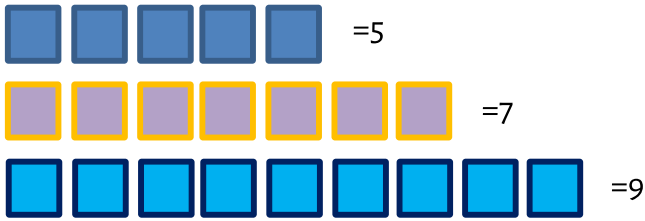
Nivel de medición	Cualidad	Ejemplo
Nominal	Clasificación en categorías	Género, estado civil, profesión
Ordinal	Clasificación y ordenamiento	Nivel educativo, escalas de aptitud
De intervalo	Clasificación, ordenamiento y distancia entre puntuaciones.	Temperatura, escalas de aptitud.
De razón	Clasificación, ordenamiento y distancia entre puntuaciones considerando un punto real cero.	Peso, talla, pulso.

Una vez que hemos identificado claramente las variables cualitativas y las hemos diferenciado de las cuantitativas, el siguiente paso es determinar la forma más adecuada de presentar los datos obtenidos de estas variables. La forma más básica de presentación es mediante tablas, las cuales pueden ser de frecuencias absolutas, relativas o de contingencia.

Tablas de frecuencias

Distribución de frecuencias absolutas (f): representa el número de veces que se repite un valor en un conjunto de datos. La distribución de frecuencias es un listado de todas las puntuaciones observadas de una variable, junto con la frecuencia (f) de cada puntuación o categoría. El conteo de frecuencias corresponde al número de veces que aparece cada elemento en el conjunto de datos.

Por ejemplo, supongamos que tenemos las siguientes puntuaciones:



En el ejemplo anterior vemos que son variables nominales, donde podemos discriminar por color. Es decir, la variable puede tomar tres tonos de azul y de acuerdo con esto clasificamos cada elemento.

La frecuencia absoluta (f) es el número de observaciones que pertenecen a una clase; en otras palabras, es el número de veces que se ha encontrado que los elementos analizados pertenecen a determinada categoría. En la Tabla 2, observamos cómo se han distribuido de acuerdo al color los elementos analizados.

Tabla 2

Tabla de distribución de frecuencias absolutas por color

Unidad	Frecuencia absoluta (f)
	5
	7
	9

Distribución de frecuencias relativas (fr): es la relación entre las frecuencias absolutas y el número total de datos. Para calcularla, podemos utilizar la ecuación 1:

$$f_r = \frac{n_i}{n_T}$$

Donde:

n_i es el número total de elementos que pertenecen a la categoría *i*, mientras que n_T es el número total de elementos de toda la distribución.

La sumatoria de las frecuencias relativas debe darnos como resultado la unidad (1). Podemos, expresar la frecuencia relativa como porcentaje, multiplicando la frecuencia relativa por cien, en este caso la sumatoria debería resultar en el 100 %.

En la Tabla 3, podemos observar una distribución de frecuencias relativas.

Tabla 3

Tabla de distribución de frecuencias relativas por color

Unidad	f	fr	fr %
	5	(5/21) = 0,24	(5×100/21) = 24
	7	(7/21) = 0,33	(7×100/21) = 33
	9	(9/21) = 0,43	(9×100/21) = 43
n_T	21	(0,24 + 0,33 + 0,43) = 1	(24 % + 33 % + 43 %) = 100

Distribución de frecuencias acumuladas (f_a): indica el número de datos que caen por debajo del límite real superior de cada intervalo. Podemos emplear para su cálculo la ecuación 2:

$$f_a = f1 + f2 + f3 .....$$

Se puede obtener la frecuencia acumulada absoluta sumando las frecuencias absolutas, mientras que la frecuencia relativa acumulada se obtiene sumando las frecuencias relativas. En el primer caso, la última casilla de la distribución de frecuencias acumuladas debe coincidir con el número total de elementos. En el caso de la frecuencia relativa acumulada, la última casilla debe ser igual a 1 o 100 %, ya que representa el total de la muestra.

Tabla 4

Tabla de distribución de frecuencias absolutas por color

Unidad	f	fa	fa %
	5	5	24
	7	(5+7) = 12	24 + 33 = 57
	9	(5+7+9) = 21	57 + 43 = 100

Una vez elegido el tipo de tabla más adecuado para la presentación de nuestros datos, debemos tener en cuenta las siguientes partes que componen una tabla:

- Título: debe responder a las preguntas: ¿Qué? ¿Cómo? ¿Dónde? ¿Cuándo?
- Encabezado: nombre o denominación de las columnas, en este espacio se encuentran las medidas de frecuencia utilizadas.
- Columna matriz: en esta se encuentran las escalas o categorías de la variable.
- Cuerpo: aquí se colocan los datos numéricos.
- Total: en esta sección se incluyen las sumatorias.
- Fuente: indica de dónde se obtuvieron los datos y quién elaboró la tabla.
- Interpretación: líneas finales que esclarecen o explican lo observado en la tabla.

Tablas de contingencia

La tabla de contingencia, también conocida como tabla de tabulación cruzada o tabla de doble entrada, es una herramienta estadística utilizada para registrar las frecuencias de las variables dependientes e independientes. Su principal función es analizar la relación entre dos o más variables categóricas.

La tabla de contingencia más común es la de 2x2, que representa dos variables categóricas, cada una con dos categorías. Un ejemplo de esta se muestra en la Tabla 5.

Tabla 5

Distribución por cáncer de pulmón y hábito tabáquico

	Cáncer	**Sano**	**Total**
Fumadores	20	30	50
No fumadores	15	35	50
Total	35	65	100

También es posible construir una tabla de contingencia en forma de proporciones, relacionando el número de elementos de cada casilla con el total general.

Tabla 6

Tabla de contingencia

	Cáncer	**Sano**	**Total**
Fumadores	20/100	30/100	0,5
No fumadores	15/100	35/100	0,5
Total	0,35	0,65	1

Al igual que las tablas de frecuencia, las tablas de contingencia también se estructuran en varias partes clave:

- Título: debe responder a las preguntas ¿qué? ¿cómo? ¿dónde? ¿cuándo?
- Encabezado: corresponde al nombre o denominación de las columnas. En este espacio se registran las enfermedades objeto de estudio (variable dependiente).

- Columna matriz: aquí se encuentran los factores de riesgo, tratamientos recibidos o pruebas diagnósticas realizadas (variable independiente).
- Cuerpo: se colocan los datos numéricos correspondientes.
- Total: se incluyen las sumatorias de cada columna y fila.
- Fuente: indica de dónde se obtuvieron los datos y quién elaboró la tabla.
- Interpretación: son unas líneas finales que esclarecen lo observado en la tabla.

Para finalizar esta primera parte, es importante tener en cuenta los errores más comunes en la creación de tablas, tales como disposiciones incorrectas, títulos incompletos o inadecuados, la presentación de porcentajes sin frecuencias absolutas acompañantes, o la inclusión de una cantidad excesiva de información que pueda confundir al lector, entre otros.

Una vez realizada la presentación en forma de tabla, se pueden calcular los índices estadísticos para las diferentes variables. En relación con las variables cualitativas, los índices más comunes son las razones, proporciones, porcentajes y tasas.

Razones

Una razón es una comparación entre dos cantidades mediante división. Un ejemplo de razón es a/b, lo que nos indica cuántas veces b está contenido en a. Solo puede calcularse en variables cualitativas nominales dicotómicas. Por ejemplo, si queremos calcular la razón de personas sanas/enfermas en la Tabla 5, tendríamos que hacer el siguiente cálculo:

$$Razón = \frac{Número\ de\ personas\ sanas}{Número\ de\ personas\ enfermas}$$

Razón: 65/35 = 1,85

La conclusión sería que hay 1,85 personas sanas por cada persona enferma.

Proporciones

Una proporción es un término matemático relacionado con fracciones. Las fracciones son una forma de expresar qué parte del todo constituye una observación. Por ejemplo, si queremos saber cuál es la proporción de enfermos en la Tabla 5, debemos aplicar la siguiente fórmula:

$$Proporción = \frac{Número\ de\ personas\ enfermas}{Total\ de\ personas\ sanas\ mas\ enfermas}$$

Proporción: 35/100 = 0,35

La conclusión sería que existen 0,35 personas enfermas en la distribución.

En pocas palabras, la proporción representa la cantidad total o el número de observaciones, expresada en forma decimal. En la siguiente ecuación 3 se presenta la fórmula general de las proporciones:

$$Proporción\ de\ una\ categoría = \frac{número\ total\ de\ una\ categoría}{Número\ de\ observaciones\ total}$$

Porcentajes

Los porcentajes son proporciones multiplicadas por 100. Por ejemplo, si queremos saber cuál es el porcentaje de enfermos en la Tabla 5, aplicamos la siguiente fórmula:

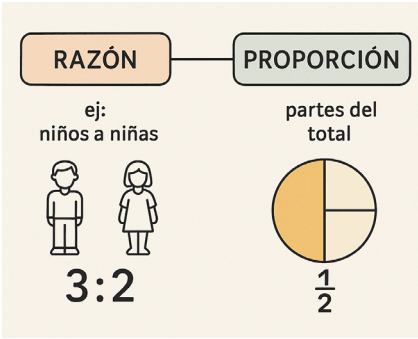
$$Porcentaje = \frac{Número\ de\ personas\ enfermas}{Total\ de\ personas\ sanas\ mas\ enfermas} \times 100$$

Proporción: (35/100)×100 = 35 %

La conclusión sería que hay 0,35 del total de la distribución.

Figura 2

Razones y proporciones



Conclusiones y recomendaciones

Lo primero que debemos hacer cuando iniciamos un análisis estadístico es reconocer qué variables valen la pena ser representadas mediante tablas y/o índices estadísticos, ya que no es necesario traducir al papel todos los datos recolectados en un estudio de investigación. Por esta razón, los investigadores deben saber discernir las variables útiles para el análisis de aquellas que no aportarán información significativa.

Una vez seleccionadas las variables de interés, debemos identificar el tipo de variable con la que estamos tratando, ya que el análisis variará según su tipo. Como se observó, las variables cualitativas son de fácil análisis y representación, pero esto no implica que deban ser reportadas en su totalidad. Como se mencionó anteriormente, solo se deben reportar los datos más relevantes.

Se recomienda presentar tablas con los datos generales o introductorios e índices con los datos más significativos o pertinentes del estudio.

Referencias

Barón, F. (1995). Bioestadística: Métodos y Aplicaciones. Universidad de Málaga. Recuperado de https://www.bioestadistica.uma.es/baron/bioestadistica.pdf

Govinden, L. P. (2000). *Introducción a la estadística* (2da ed.). McGraw-Hill.

Navidi, W. (2006). Estadística para ingenieros y científicos (1ra ed.). McGraw-Hill.

Ritchey, F. J. (2001). *Estadística para las ciencias sociales:* El potencial de la imaginación estadística. McGraw-Hill.

Capítulo 10
Tasas y ajuste de tasas

Introducción

En el ámbito de la medicina y la salud pública, es fundamental contar con herramientas que permitan describir cuantitativamente el comportamiento de enfermedades, muertes, nacimientos y otros fenómenos relevantes. Entre estas herramientas se encuentran las tasas, que permiten relacionar el número de eventos de interés con la población en riesgo, facilitando su interpretación en términos proporcionales y su comparación entre distintos grupos o períodos de tiempo.

Las tasas son ampliamente utilizadas en la práctica médica y epidemiológica. Nos permiten responder preguntas como: ¿Cuál es la magnitud de una epidemia? ¿En qué grupo etario hay mayor riesgo de muerte? ¿Ha mejorado la cobertura prenatal en los últimos años?, entre muchas otras. Así, las tasas constituyen medidas esenciales para la toma de decisiones clínicas, la planificación sanitaria, la asignación de recursos y la evaluación de intervenciones.

No obstante, la simple comparación de tasas entre dos o más poblaciones puede resultar engañosa si estas difieren significativamente en su estructura demográfica, especialmente en variables como edad y sexo. Por ejemplo, comparar la tasa de mortalidad general entre dos países, sin considerar que uno tiene una población más envejecida, puede conducir a conclusiones erróneas. Para evitar este tipo de sesgos y realizar comparaciones justas, se recurre al ajuste de tasas.

El ajuste de tasas es una técnica estadística que busca estandarizar la comparación de medidas entre poblaciones, eliminando el efecto de factores confusores. Existen principalmente dos métodos de ajuste: el método directo y el método indirecto, cada uno con sus requisitos y aplicaciones específicas. La elección del método dependerá del tipo de información disponible y del tamaño de la población en estudio.

En este capítulo se abordarán los conceptos fundamentales de tasas, los principales tipos utilizados en salud, la importancia del ajuste de tasas y los

procedimientos para aplicar correctamente ambos métodos de ajuste. El contenido se complementará con ejemplos aplicados, un ejercicio práctico y una sección de conclusiones y recomendaciones útiles para el contexto médico y epidemiológico.

Desarrollo

Tasas

Una tasa es una medida que relaciona el número de casos de un evento (enfermedad, nacimiento, muerte, etc.) con una población en riesgo, multiplicada por una constante. Se usa para facilitar la comparación entre grupos o períodos de tiempo distintos.

$$Tasa = \left(\frac{Número\ de\ eventos\ en\ un\ tiempo\ dado}{Población\ en\ riesgo\ en\ el\ mismo\ tiempo}\right) \times k$$

Donde k suele ser 100, 1 000, 10 000 o 100 000, dependiendo de la rareza del evento.

Principales tasas en salud

Las tasas son herramientas estadísticas fundamentales para medir fenómenos de salud en una población. Permiten cuantificar cuán frecuente es un evento dentro de un grupo definido de personas en un período específico, facilitando su análisis y comparación entre grupos, regiones o momentos distintos.

Las tasas pueden aplicarse tanto para describir el estado de salud de una población como para evaluar la efectividad de intervenciones sanitarias o políticas públicas. Entre las más relevantes en medicina se encuentran las siguientes:

Prevalencia

Definición: la proporción de individuos que presentan una enfermedad o condición de salud específica en un momento determinado, o durante un periodo concreto, dentro de una población definida.

Fórmula:

$$Tasa\ de\ prevalencia\ =\ \left(\frac{Número\ total\ de\ casos\ (nuevos+existentes)}{Población\ total\ en\ el\ mismo\ tiempo}\right)\times k$$

La prevalencia es una medida de frecuencia que se expresa como una proporción, ya que no incorpora el tiempo en el denominador. Ejemplo: en un centro de salud se identificaron 200 personas con hipertensión arterial en una población de 2 000 adultos.

$$Prevalencia\ =\ \left(\frac{200}{2.000}\right)\times 100\ =\ 10\ \%$$

Tasa de incidencia

Definición: mide la ocurrencia, o velocidad de ocurrencia, de nuevos casos de un evento de salud en una población susceptible durante un periodo de observación determinado.

Fórmula:

$$Tasa\ de\ incidencia\ =\ \left(\frac{Número\ de\ nuevos\ casos}{Población\ en\ riesgo\ en\ el\ inicio\ del\ periodo}\right)\times k$$

Ejemplo:

Durante un año se identificaron 50 nuevos casos de dengue en una población de 10 000 habitantes.

$$Incidencia\ =\ \left(\frac{50}{10\ 000}\right)\times 10\ 000\ =\ 50\ por\ cada\ 10\ 000\ habitantes$$

Tasa de mortalidad general

Definición: indica el número total de muertes ocurridas en una población, por cualquier causa, durante un período determinado.

Fórmula:

$$Tasa\ de\ mortalidad\ general\ =\ \left(\frac{Número\ total\ de\ defunciones}{Población\ total\ en\ el\ mismo\ tiempo}\right)\times k$$

Ejemplo:

En una comunidad de 100 000 personas hubo 850 fallecimientos durante el año.

$$Mortalidad\ General = \left(\frac{850}{100\,000}\right)\times 1\,000 = 8,5\ muertes\ por\ cada\ 1000\ habitantes$$

Tasa de mortalidad infantil

Definición: mide el número de muertes de niños menores de 1 año por cada 1 000 nacidos vivos en un año.

Fórmula:

$$Tasa\ de\ mortalidad\ infantil = \left(\frac{Número\ de\ muertes\ en\ menores\ de\ 1\ año}{Nacidos\ vivos\ en\ el\ mismo\ periodo}\right)\times k$$

Ejemplo:

Se reportaron 45 muertes de menores de 1 año entre 5 000 nacidos vivos.

$$Mortalidad\ Infantil = \left(\frac{45}{5\,000}\right)\times 1\,000 = 9\ muertes\ infantiles\ por\ cada\ 1000\ nacidos\ vivos$$

Tasa de mortalidad materna

Definición: indica el número de muertes de mujeres por causas relacionadas con el embarazo, parto o puerperio, por cada 100 000 nacidos vivos.

Fórmula:

$$Tasa\ de\ mortalidad\ materna = \left(\frac{Número\ de\ muertes\ maternas}{Nacidos\ vivos\ en\ el\ mismo\ periodo}\right)\times k$$

Ejemplo:

En una región hubo 4 muertes maternas con un total de 20 000 nacidos vivos.

$$Mortalidad\ Materna = \left(\frac{4}{20\,000}\right)\times 100\,000 = 20\ muertes\ maternas\ por\ cada\ 100.000\ nacidos$$

Tasa de natalidad

Definición: mide el número de nacimientos ocurridos durante un año por cada 1.000 personas en la población total.

Fórmula:

$$Tasa\ de\ natalidad\ =\ \left(\frac{Número\ de\ nacimientos}{Población\ total}\right)\times k$$

Ejemplo:

En un cantón se registraron 2 000 nacimientos y la población total fue de 120 000 personas.

$$Natalidad\ =\ \left(\frac{2\ 000}{120\ 000}\right)\times 1\ 000\ =\ 16,7\ nacimientos\ por\ cada\ 1000\ habitantes$$

Tasa de letalidad

Definición: refleja la gravedad de una enfermedad al indicar la proporción de personas que mueren a causa de una enfermedad, entre quienes la padecen.

Fórmula:

$$Tasa\ de\ letalidad\ =\ \left(\frac{Muertes\ por\ una\ enfermedad}{Total\ de\ casos\ diagnosticados\ de\ una\ enfermedad}\right)\times k$$

Ejemplo:

De 300 personas diagnosticadas con meningitis, 60 fallecieron.

$$Letalidad\ =\ \left(\frac{60}{300}\right)\times 100\ =\ 20\ \%$$

A continuación, una tabla resumen de las principales tasas en el campo de la medicina:

Tabla 1

Tabla resumen de las principales medidas de frecuencia

Tasa	¿Qué mide?	Ejemplo
Prevalencia	Casos existentes (nuevos + antiguos) de una enfermedad en un tiempo determinado.	200 hipertensos entre 2 000 → 10 %
Incidencia	Casos nuevos de una enfermedad en un tiempo determinado	50 nuevos casos de dengue entre 10 000 → 50 por 10 000 habitantes
Mortalidad general	Muertes por todas las causas	850 muertes en 100 000 → 8,5 por 1 000 habitantes
Mortalidad infantil	Muertes en menores de 1 año por cada 1 000 nacidos vivos	45 muertes en <1 año entre 5 000 nacidos vivos → 9 por 1 000 nacidos vivos
Mortalidad materna	Muertes maternas por cada 100.000 nacidos vivos	4 muertes maternas entre 20 000 nacidos → 20 por 100 000 nacidos vivos
Natalidad	Nacidos vivos por cada 1 000 habitantes	2 000 nacimientos en 120 000 personas → 16,7 por 1 000 habitantes
Letalidad	Muertes por una enfermedad entre quienes la padecen	60 muertes en 300 casos de meningitis → 20 %

Ajuste de tasas

Las tasas son medidas útiles para describir la frecuencia de eventos de salud en una población. Sin embargo, cuando se desea comparar tasas entre diferentes poblaciones, es fundamental tener en cuenta que dichas poblaciones pueden diferir en características estructurales importantes, como la distribución por edad, sexo, ocupación, nivel socioeconómico, entre otras.

Una comparación directa de las tasas crudas podría llevar a conclusiones erróneas si estas variables actúan como factores confusores. Por ejemplo, una población más envejecida naturalmente tendrá una mayor tasa de mortalidad, no necesariamente porque las condiciones sanitarias sean peores, sino porque el riesgo de morir aumenta con la edad.

El ajuste de tasas es un procedimiento estadístico que permite eliminar o minimizar el efecto de variables confusoras para lograr comparaciones justas entre tasas de diferentes poblaciones. Este procedimiento implica modificar las tasas observadas, como si las poblaciones tuvieran una estructura homogénea respecto a la variable confusora (por ejemplo, edad).

Importancia del ajuste de tasas

- Permite comparaciones válidas entre dos o más poblaciones con estructuras demográficas diferentes.
- Controla el efecto de variables confusoras, mejorando la interpretación de los datos.
- Es fundamental en estudios de vigilancia epidemiológica, evaluaciones de intervenciones sanitarias y diseño de políticas públicas.
- Evita errores de interpretación que podrían llevar a decisiones inadecuadas en salud pública.

Factores confusores comunes

- Edad: uno de los factores más relevantes, especialmente en estudios de mortalidad, incidencia y enfermedades crónicas.
- Sexo: algunas condiciones de salud son más frecuentes en hombres o en mujeres.
- Condición socioeconómica, ocupación, etnia, área geográfica, entre otros, también pueden influir dependiendo del contexto.

Requerimientos para el ajuste

- Tasas específicas por subgrupo: por ejemplo, tasas de mortalidad por grupo etario.
- Población estándar: sirve como referencia para estandarizar las tasas. Puede ser:
- Interna: obtenida de la propia muestra o combinación de las poblaciones en estudio.
- Externa: definida por organismos internacionales (por ejemplo, OMS, CELADE), permite comparaciones más amplias.

Situaciones en que se recomienda ajustar tasas:

- Cuando las poblaciones tienen distribuciones diferentes por edad o sexo.
- Cuando se comparan tasas entre países, regiones o períodos de tiempo con estructuras poblacionales distintas.
- Cuando se estudian enfermedades asociadas a variables demográficas (como cáncer, enfermedades cardiovasculares o diabetes).
- Cuando se desea analizar el impacto de una intervención sanitaria, independientemente de la composición demográfica.

Métodos de ajuste de tasas

El ajuste de tasas puede realizarse utilizando principalmente dos métodos: directo e indirecto. Ambos buscan eliminar el efecto de una variable confusora, como la edad, para comparar tasas como si las poblaciones tuvieran la misma estructura. La elección del método depende de la disponibilidad y calidad de los datos.

El método directo calcula cuál sería la tasa que tendría cada población si compartieran la misma estructura demográfica que una población estándar. Es el más preciso, pero requiere información detallada sobre la tasa específica en cada subgrupo y la distribución de la población estándar. Se utiliza cuando se conoce la población de cada subgrupo (edad, sexo, etc.) y sus respectivas tasas específicas.

Ofrece una alta precisión en los resultados, lo que permite realizar comparaciones válidas entre múltiples poblaciones de manera confiable. Sin embargo, este método requiere contar con información detallada sobre las tasas por subgrupo, lo cual puede ser un desafío en algunos contextos. Además, no es aplicable cuando se desconoce la estructura demográfica de la población en estudio.

Tabla 2

Pasos del método directo

Paso	Descripción	Ejemplo
1	Calcular las tasas específicas por subgrupo (e.g., por grupo de edad) para cada población.	Tasa por edad en Cali y São Paulo
2	Seleccionar una población estándar. Puede ser la suma de las poblaciones o una externa.	CELADE o suma de ambas ciudades
3	Multiplicar cada tasa específica por el número de personas en ese subgrupo de la población estándar y dividir para 1 000.	Casos ajustados esperados por grupo de edad
4	Sumar todos los casos ajustados para obtener el total ajustado para cada población.	Suma de casos esperados por ciudad
5	Dividir el total ajustado por el total de la población estándar y multiplicar por 1 000 para obtener la tasa ajustada.	Tasa ajustada = (casos ajustados / población estándar) x 1 000

Supongamos que queremos comparar la tasa de mortalidad por cáncer de cuello uterino entre Cali y São Paulo, ajustando por edad. Usamos como población estándar la suma de ambas ciudades.

Tabla 3

Ejemplo de tabla de aplicación (método directo)

Grupos de edad	Tasa específica en Cali	Población estándar	Casos esperados en Cali
15-24	2,0	10 000	(2 x 10 000)/1 000=20
25-34	3,5	8 000	(3,5 x 8 000)/1 000=28
35-44	5,0	5 000	(5 x 5 000)/1 000=25
45-54	6,0	4 000	(6 x 4 000)/1 000=24
55-64	7,5	3 000	(7,5 x 3 000)/1 000=22,5
Total	4,8	30 000	119,5

Teniendo una tasa bruta en Cali de 4,8 por cada 1 000 habitantes, una vez ajustada queda:

$$Tasa\ ajustada\ en\ Cali = \left(\frac{119{,}5}{30\,000}\right) \times 1\,000 = 3,98\ por\ cada\ 1\,000\ \text{habitantes.}$$

El método indirecto calcula cuántos casos se esperaría observar en la población de estudio si tuviera las mismas tasas específicas que la población estándar. Se utiliza cuando no se dispone de tasas específicas de la población de estudio, pero sí se tiene su estructura demográfica.

Además, es principalmente ventajoso en poblaciones pequeñas o cuando se dispone de datos incompletos. Requiere de menos información que el método directo, lo que lo hace más accesible en situaciones con recursos limitados.

Pero, es un método menos preciso, ya que utiliza tasas de una población externa.

Tabla 4

Pasos del método indirecto

Paso	Descripción	Ejemplo
1	Seleccionar una población estándar con tasas específicas por subgrupo.	São Paulo (estándar)
2	Conocer la distribución por subgrupos de la población de estudio (e.g., Cali).	Población por edad en Cali
3	Multiplicar las tasas específicas estándar por la población del subgrupo de la población de estudio y dividirlo para 1 000.	Casos esperados en Cali
4	Calcular el total de casos esperados.	Suma de todos los casos esperados
5	Calcular el factor de corrección (FC): observados/ esperados.	Observados/ Esperados
6	Multiplicar el FC por la tasa bruta de la población estándar para obtener la tasa ajustada.	Tasa ajustada = FC × tasa bruta estándar

Supongamos que no se conocen las tasas específicas por edad de Cali, pero sí se conoce la distribución de su población por grupos de edad. En cambio, sí se disponen de las tasas específicas por edad de São Paulo, que usaremos como estándar.

Tabla 5

Ejemplo de tabla de aplicación (método indirecto)

Grupos de edad	Tasa específica estándar (por 1 000)	Población en Cali	Casos esperados en Cali
15-24	2,0	10 000	(2 x 10 000)/1 000 = 20
25-34	3,5	8 000	(3,5 x 8 000)/1 000 = 28
35-44	5,0	5 000	(5 x 5 000)/1 000 = 25
45-54	6,0	4 000	(6 x 4 000)/1 000 = 24
55-64	7,5	3 000	(7,5 x 3 000)/1 000 = 22,5
Total	4,8	30 000	119,5

Ahora supongamos que Cali reporta 100 muertes observadas por cáncer de cuello uterino.

$$Factor\ de\ corrección = \left(\frac{100\ casos\ observados}{119{,}5\ casos\ esperados}\right) = 0,837$$

Y si la tasa bruta de São Paulo (población estándar) es 4,8 por 1 000 habitantes, entonces:

$$Tasa\ ajustada\ en\ Cali = 0,837 \times 4,8 = 4,01\ por\ cada\ 1\ 000\ \text{habitantes.}$$

Conclusiones y recomendaciones

Las tasas son indicadores fundamentales en epidemiología y salud pública, ya que permiten cuantificar la frecuencia y distribución de fenómenos de salud en una población definida, durante un tiempo determinado.

Estas herramientas son indispensables para el diagnóstico de situación, la priorización de problemas y la evaluación de políticas sanitarias.

Existen diversos tipos de tasas, como la tasa de prevalencia, incidencia, mortalidad general, mortalidad infantil, mortalidad materna, natalidad y letalidad. Cada una responde a un propósito específico y aporta información crítica para la toma de decisiones clínicas y poblacionales.

La comparación de tasas crudas entre diferentes poblaciones puede conducir a conclusiones erróneas si no se considera la estructura demográfica de cada población. Variables como la edad y el sexo influyen notablemente en la ocurrencia de eventos de salud y deben ser tenidas en cuenta como posibles factores de confusión.

El ajuste de tasas permite eliminar el efecto de variables confusoras, estandarizando la comparación entre poblaciones. Es una técnica estadística esencial para asegurar que las diferencias observadas en las tasas se deban al fenómeno de estudio, y no a distorsiones por estructura poblacional.

El método directo de ajuste es el más preciso, ya que utiliza tasas específicas de la población de estudio aplicadas a una población estándar. El método indirecto, en cambio, es útil cuando no se conocen tasas específicas, pero sí la distribución de la población por subgrupos; es especialmente aplicable a poblaciones pequeñas o con información incompleta.

El conocimiento y correcta aplicación de estas técnicas fortalece la validez interna y externa de los análisis epidemiológicos, contribuyendo a una mejor comprensión de los patrones de salud y enfermedad, así como al diseño de intervenciones más eficaces y equitativas.

Es necesario capacitar a estudiantes y profesionales de la salud en el uso correcto de las tasas y su interpretación, incluyendo la necesidad de ajustarlas cuando se identifiquen posibles variables confusoras como la edad, el sexo u otras características sociodemográficas. Además, se debería incluir el cálculo de tasas ajustadas como parte rutinaria del análisis epidemiológico y estadístico en investigaciones clínicas, evaluaciones de programas y elaboración de reportes de situación de salud, utilizando poblaciones estándar internacionalmente reconocidas (como las propuestas por la OMS o CELADE) cuando se pretenda comparar tasas entre países o regiones, o cuando se busque una referencia común para evaluaciones longitudinales.

Es importante saber elegir el método de ajuste más adecuado según la disponibilidad de datos; si se dispone de tasas específicas por subgrupo, se debe

usar el método directo, y si se desconoce la distribución por subgrupos o los datos son escasos, se debe recurrir al método indirecto.

Se recomienda aprovechar herramientas digitales como Excel y EPIDAT para realizar cálculos de ajuste de tasas de forma sistemática, reproducible y didáctica en entornos académicos y profesionales. Además, se debe fomentar el análisis crítico de los resultados: una tasa ajustada debe interpretarse en su contexto, considerando también otros indicadores complementarios (como proporciones, razones, tasas brutas y razones estandarizadas) para una comprensión integral del fenómeno de salud.

Referencias

Bonita, R., Beaglehole, R., y Kjellström, T. (2006). *Epidemiología Básica* (2.ª ed.). Ginebra: Organización Mundial de la Salud.

Celade. (1991). *América Latina: proyecciones de población años calendario 1950-2000*. Boletín Demográfico, (48), 31.

Gordis, L. (2014). *Epidemiología* (5.ª ed.). Madrid: Elsevier.

Ministerio de Salud Pública del Ecuador. (2023). *Indicadores Básicos de Salud 2022*. Quito: MSP.

Pagano, M., & Gauvreau, K. (2000). *Principles of Biostatistics* (2nd ed.). Belmont: Duxbury Press.

Rothman, K. J., Greenland, S., & Lash, T. L. (2008). *Modern Epidemiology* (3rd ed.). Philadelphia: Lippincott Williams & Wilkins.

Capítulo 11
Análisis descriptivo gráfico: variables cualitativas

Introducción

La representación gráfica de datos cualitativos obtenidos tras el análisis estadístico es fundamental, ya que a través de los gráficos se logra representar de forma clara y sintetizada la distribución de los datos, facilitando al lector su comprensión.

Cuando hablamos del análisis descriptivo gráfico de variables cualitativas, los gráficos que se abordarán son: barras simples, dobles, apiladas y gráficos de sectores. Estos gráficos deben ser elaborados siguiendo ciertas directrices que permitan al lector identificar detalles como las variables estudiadas, la población y el tiempo en el que se ha realizado el análisis. Es importante mencionar que los softwares estadísticos ofrecen una gran variedad de opciones y facilitan la elaboración de estos gráficos.

El objetivo de este capítulo es describir las características de cada tipo de gráfico, los criterios para su correcta elaboración e interpretación, además de señalar los errores comunes que deben evitarse.

Desarrollo

Cuando se busca representar los resultados de variables cualitativas, los gráficos son herramientas muy útiles para complementar la información que se presenta en tablas simples de frecuencia o en tablas cruzadas. Sin embargo, es importante señalar que su uso debe ser complementario, evitando la repetición exacta de los mismos datos ya tabulados.

Los gráficos para variables cualitativas permiten visualizar con mayor claridad la distribución de frecuencias entre categorías, así como identificar diferencias entre distintos grupos poblacionales. Entre los gráficos más utilizados se

encuentran el gráfico de sectores, el gráfico de barras simples, el gráfico de barras dobles y las barras apiladas.

Gráfico de sectores

El gráfico de sectores, también conocido como gráfico de pastel o gráfico circular, es una representación visual que divide el total en porciones o sectores. Cada sector corresponde a las frecuencias relativas (porcentajes) de las distintas categorías de la variable analizada. Este tipo de gráfico permite visualizar de forma clara cómo se distribuye un total entre las diferentes características, facilitando la comprensión de la composición general de una variable. Es especialmente útil cuando se desea resaltar la categoría que predomina dentro del conjunto de datos.

Figura 1

Distribución porcentual de enfermedades en una población simulada

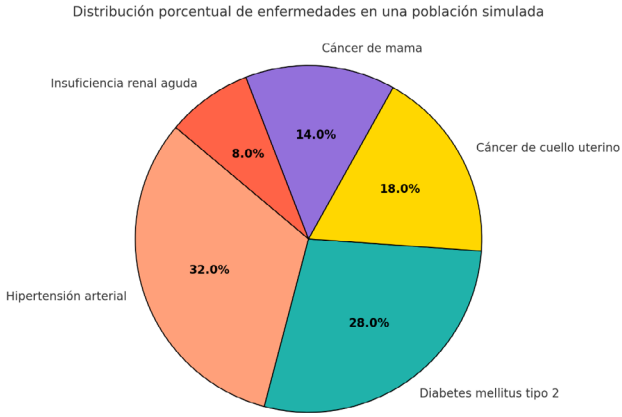


Gráfico de barras simples

Los gráficos de barras simples constituyen el tipo más básico y ampliamente utilizado para representar una sola variable cualitativa, ya sea de tipo |

ordinal o nominal (por ejemplo: estado civil, nivel de escolaridad, servicios básicos, comorbilidades).

En este tipo de gráfico, las variables son representadas mediante barras rectangulares separadas, que pueden estar dispuestas en sentido vertical u horizontal. La longitud de cada barra representa la frecuencia absoluta o relativa (porcentaje) de cada categoría.

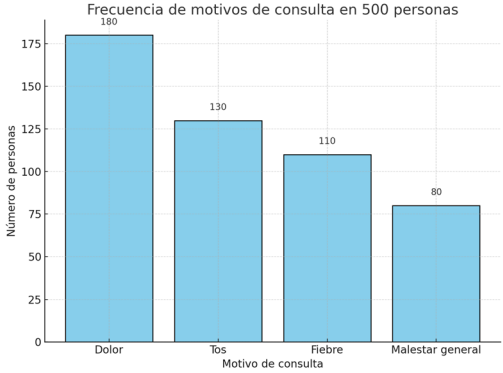
Las categorías se deben mostrar en el eje horizontal (X), mientras que la frecuencia absoluta o relativa se coloca en el eje vertical (Y) en el caso de barras verticales. Si se utilizan barras horizontales, esta disposición se invierte.

Para la correcta elaboración del gráfico de barras simples, se recomienda lo siguiente:

- En el caso de variables nominales: ordenar las categorías de mayor a menor, según su frecuencia, para facilitar la lectura e interpretación del gráfico. Esto permite identificar rápidamente la categoría con mayor frecuencia y proporciona una visión clara de la distribución de los datos. Este tipo de gráfico es ideal para comparar las frecuencias de diferentes categorías de una variable y es especialmente útil cuando se desea mostrar la variabilidad entre categorías de una sola variable cualitativa.

Figura 2

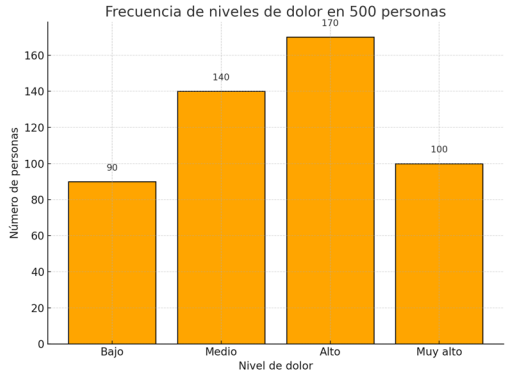
Simulación de variable nominal



- En el caso de variables ordinales, se debe respetar el orden lógico, independientemente de su frecuencia.

Figura 3

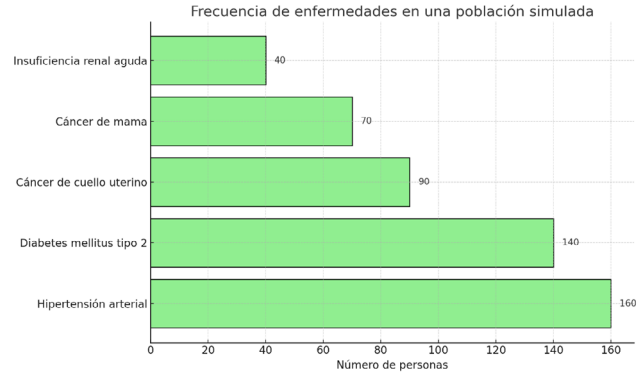
Simulación de variable ordinal vertical



La elección entre barras verticales u horizontales depende principalmente de la cantidad de categorías y la longitud de sus etiquetas. Deben preferirse barras orientadas en sentido horizontal para aquellas variables con muchas categorías y etiquetas extensas, para evitar que los nombres se inclinen o superpongan, lo cual dificultará la lectura.

Figura 4

Simulación de variable ordinal horizontal



Barras dobles

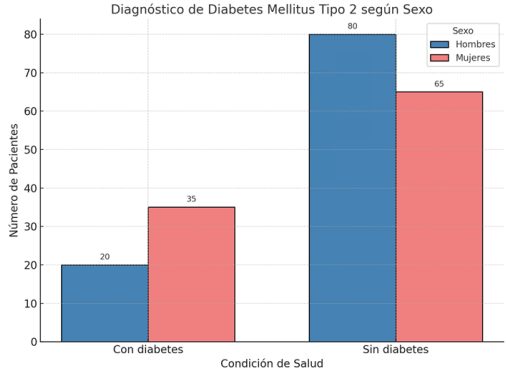
Se trata de un gráfico de barras útil para comparar dos grupos diferentes respecto a una misma variable cualitativa. Cada categoría debe ser representada mediante dos barras verticales paralelas, una para cada grupo. Este gráfico se utiliza con frecuencia cuando el objetivo es determinar si existen diferencias entre dos grupos (por ejemplo: hombres y mujeres) de una misma variable (por ejemplo: diagnóstico de diabetes mellitus tipo 2).

En el eje horizontal (X) se debe colocar la variable principal (nominal u ordinal) además de sus respectivos grupos de comparación, normalmente una variable dicotómica (por ejemplo: sexo, residencia urbana/rural).

Al elaborar este gráfico se deben usar colores diferentes para identificar claramente a cada grupo e incluir una leyenda que aclare que barra representa a cada grupo.

Figura 5

Simulación de barras dobles



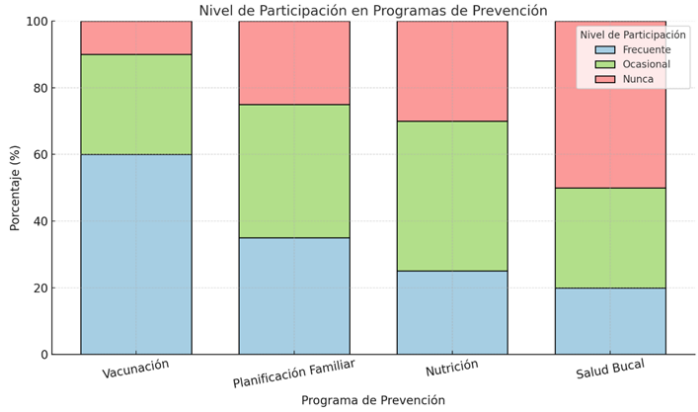
Barras apiladas

Se trata de un gráfico que permite representar la distribución conjunta de dos variables cualitativas. Se utiliza cuando se requiere visualizar tanto el total como la composición interna de cada categoría de una variable cualitativa (por ejemplo: programa de prevención), comparando a su vez dos o más subgrupos (por ejemplo: nivel de participación) dentro de cada barra.

En lugar de comparar los grupos a través de barras dobles, se representa una única barra por cada categoría, la cual está dividida o apilada en segmentos de diferentes colores, uno para cada subgrupo. Cada segmento representa el porcentaje acumulado de la variable con la que se realiza la comparación. De tal manera que cada barra alcanza el 100 % y los segmentos dentro de ella indican qué proporción corresponde a cada subgrupo.

Figura 6

Simulación de barras apiladas



Conclusiones y recomendaciones

Los gráficos constituyen una herramienta fundamental a la hora de comunicar los resultados de variables cualitativas. La elección del tipo de gráfico dependerá del número de variables que se deseen representar, del

tipo de variables y del objetivo del análisis, como, por ejemplo, describir una distribución simple o comparar diferentes grupos.

Es recomendable que, al momento de elaborar los gráficos, se utilicen colores diferentes para cada variable, así como leyendas claras que permitan diferenciar fácilmente las categorías y grupos analizados. Esta práctica ayuda a mejorar la legibilidad y comprensión de la información presentada.

Dominar la elaboración de estos gráficos permite no solo mejorar la comprensión de los resultados obtenidos durante la investigación, sino también facilitar la comunicación de los datos más relevantes, haciendo que los resultados sean accesibles y fáciles de interpretar para un público amplio.

Referencias

Celis de la Rosa, A. J., y Labrada Martagón, V. (2014). *Bioestadística* (3.ª ed.). Ciudad de publicación desconocida: Editorial El Manual Moderno.

Fayad Cámel, V. (1970). *Estadística médica y de salud pública* (2.ª ed.). Mérida (VE): Universidad de los Andes.

Gómez Barrantes, M. (2012). *Elementos de la estadística descriptiva* (4.ª ed.). Costa Rica: Universidad Estatal a Distancia de Costa Rica.

Martínez González, M. A., et al. (2020). *Bioestadística amigable* (4.ª ed.). Madrid: Elsevier.

Ministerio de Sanidad y Consumo de España. (2008). *Manual de Bioestadística para Ciencias de la Salud*. Madrid: Ministerio de Sanidad.

Pagano, M., y Gauvreau, K. (2012). *Principios de Bioestadística* (2.ª ed.). México: Cengage Learning.

Capítulo 12
Análisis descriptivo numérico: variables cuantitativas

Introducción

El análisis descriptivo numérico permite resumir, organizar y comprender de manera clara la información recopilada. Las variables cuantitativas, como la edad, el peso, la talla, los niveles de colesterol o glucosa, requieren el uso de pruebas que permitan describir su comportamiento en la población de estudio.

Es importante recalcar que este tipo de análisis no busca hacer generalizaciones ni establecer relaciones causales, sino ofrecer una primera aproximación a los datos, permitiendo observar cómo se distribuyen y qué tan homogéneos o heterogéneos son.

Comprender e interpretar adecuadamente los datos obtenidos a partir del análisis descriptivo numérico permite al investigador obtener una idea clara del comportamiento de los datos, facilitando la identificación de características relevantes dentro del conjunto analizado. Esta etapa del análisis estadístico sienta una base sólida para futuras decisiones metodológicas, como la elección de pruebas estadísticas apropiadas o el cálculo de la muestra.

El objetivo de este capítulo es conocer las principales medidas del análisis descriptivo numérico, como son las medidas de tendencia central (media, mediana y moda), medidas de posición (cuartiles, deciles, percentiles), medidas de dispersión (rango, varianza y desviación estándar), y medidas de forma (asimetría y curtosis), ya que cada una de estas proporciona información complementaria sobre el comportamiento de los datos.

Desarrollo

Cabe señalar que, una vez realizadas las encuestas, ya sea mediante formularios físicos o digitales, es fundamental identificar las variables

cuantitativas, es decir, aquellas en las que se obtiene un dato numérico puro o directo y no una categoría predefinida. Un ejemplo claro de ello es la pregunta "¿cuál es su edad en años cumplidos?", donde las respuestas de los encuestados serán datos como 17, 18, 19 o 20 años.

Es importante recalcar que las variables cuantitativas deben ser formuladas como tales dentro de los cuestionarios. No se recomienda preguntar por la edad utilizando rangos categóricos como: de 10 a 15, de 16 a 19, de 20 a 25, etc., ya que esto impide un análisis más detallado y limita la posibilidad de categorizarlas posteriormente según los objetivos del estudio.

Antes de ingresar al análisis descriptivo de una variable cuantitativa, es importante señalar que, cuando se obtiene el dato puro, se puede realizar posteriormente un proceso de categorización, de acuerdo con la operacionalización de las variables planteadas en el estudio. Una vez realizada esta categorización, la variable puede ser tratada como una variable cualitativa ordinal. Esto permite aplicar distintos procedimientos, como tablas de frecuencias absolutas y relativas, tablas cruzadas, gráficos de sectores, gráficos de barras simples, dobles o apiladas. Además, se pueden calcular medidas como razones, proporciones y otros indicadores útiles en la descripción y comparación de los datos descritos en el capítulo de variables cualitativas.

Una vez realizada esta aclaración, es importante señalar que las principales medidas utilizadas para analizar descriptivamente una variable cuantitativa se agrupan en diferentes tipos. Primero, están las medidas de tendencia central, como la media, la mediana y la moda; las medidas de posición, como los cuartiles, deciles y percentiles; las medidas de dispersión, como el rango, la varianza, la desviación estándar y el intervalo intercuartilar; y finalmente, las medidas de forma, como la asimetría y la curtosis.

Medidas de tendencia central

Son las medidas que buscan resumir en un valor único el centro de la distribución, estas son la media, la mediana y la moda. En el presente capítulo usaremos algunos ejemplos para la mejor comprensión de las diferentes medidas. Para las medidas de tendencia central partiremos de un conjunto de 10 individuos y sus edades:

Media o promedio

Se obtiene sumando todos los datos de cada uno de los individuos y dividiendo para el total de las personas.

$$\bar{x} = \frac{\sum x_i}{n}$$

x_i son los valores individuales

n es el numero total de datos

Paso 1: sumar los valores.

$$16 + 17 + 17 + 17 + 17 + 18 + 18 + 19 + 19 + 19 = 177$$

Paso 2: contar cuántos valores hay.

$$n = 10$$

Paso 3: aplicar la fórmula.

Conclusión: la media de edad es de 17,7 años.

Mediana

Es la medida que se ubica en el 50 % de los datos, es decir, en el punto medio de los individuos estudiados, por esta razón a más de ser una medida de tendencia central es una medida de posición. Para su cálculo, es indispensable que los datos se encuentren ordenados. Cuando el número total de datos es impar, existe un valor único que representa exactamente ese punto medio. Sin embargo, cuando el número de datos es par, como en el ejemplo utilizado que contiene 10 observaciones, la mediana se obtiene a partir del promedio de los dos valores centrales del conjunto ordenado.

Paso 1: verificar que los datos estén ordenados.

$$16 + 17 + 17 + 17 + 17 + 18 + 18 + 19 + 19 + 19$$

Paso 2: como hay 10 datos (n par), se toma el promedio del 5° y 6° dato.

- 5° dato: 17
- 6° dato: 18

Conclusión: la mediana de edad es de 17,5 años.

Moda

Es el valor que más se repite dentro de los individuos estudiados.

$$16 + 17 + 17 + 17 + 17 + 18 + 18 + 19 + 19 + 19 = 177$$

Conclusión: la moda de edad es de 17 años.

Una vez que hemos identificado cómo se calcula la media, la mediana y la moda, es importante señalar que, entre las tres medidas, la media es la más sensible a la presencia de valores extremos. A continuación, mostramos el porqué. En el ejemplo que venimos desarrollando, vamos a suponer que ocurrió un error durante la recolección de datos, ya sea por una falla en las mallas de validación de los formularios digitales o en el registro de los formularios físicos. En lugar de registrar correctamente la última edad como 19 años, se ingresó erróneamente 91 años. Al realizar nuevamente el cálculo de los estadísticos, se puede observar que la mediana y la moda se mantienen, mientras que la media se ve notablemente alterada.

$$\bar{x} = \frac{\sum x_i}{n}$$

$$\bar{x} = \frac{16 + 17 + 17 + 17 + 17 + 18 + 18 + 19 + 19 + 91}{10}$$

$$\bar{x} = \frac{259}{10} = 25{,}9$$

$$\text{Mediana} = \frac{17+18}{2} = 17{,}5$$

$$16 + 17 + 17 + 17 + 17 + 18 + 18 + 19 + 19 + 19 = 177$$

Este ejemplo, a pesar de contar con apenas 10 datos y haber modificado únicamente un valor, evidencia cómo la media se incrementa de 17,7 a 25,9. Este comportamiento explica por qué, ante la presencia de valores extremos, se debe tener precaución al utilizar la media o promedio como medida representativa. Este principio abre la discusión sobre la normalidad en el comportamiento de los datos, tema que será abordado con mayor profundidad en el capítulo correspondiente a la distribución de los datos.

Medidas de posición

Cuando hablamos de medidas de posición, nos referimos a la ubicación de los datos dentro de la distribución, estas son los cuartiles, percentiles y deciles. Para comprender mejor este concepto podemos tomar el concepto de la mediana que habíamos manifestado (valor que se ubica en el 50 % de los datos o de los individuos y los divide en dos partes iguales). Es decir, la mediana se convierte en el percentil 50, el decil 5 o el cuartil 2, como detallaremos a continuación.

Cuartiles

Si consideramos al total de los individuos como un 100 %, cuando hablamos de cuartiles significa dividir a ese 100 % en 4 partes iguales. Es decir, el cuartil 1 (Q1) ubicará el primer 25 % de individuos, el cuartil 2 (Q2) el 50 %, el cuartil 3 (Q3) el 75 %, y el cuartil 4 (Q4) finalmente el 100 % de los individuos. Se calcula de la siguiente manera:

Paso 1: verificar que los datos estén ordenados.

$$16 + 17 + 17 + 17 + 17 + 18 + 18 + 19 + 19 + 19$$

Paso 2: calcular Q2 (mediana).

Paso 3: calcular Q1 (primer cuartil).

Q1 es la mediana de la mitad inferior de los datos (los 5 primeros):

$$16 + 17 + \textcircled{17} + 17 + 17$$

valor central (3º dato) = 17

Paso 4: calcular Q3 (tercer cuartil).

Q3 es la mediana de la mitad superior de los datos (los últimos 5):

$$18 + 18 + \textcircled{19} + 19 + 19$$

valor central (3º dato) = 17

Paso 5: visualizar el Q4.

$$16 + 17 + 17 + 17 + 17 + 18 + 18 + 19 + 19 + \textcircled{19}$$

Esto para distribuciones pequeñas, pero para distribuciones grandes se suele usar la siguiente fórmula, la cual da como resultado la ubicación del cuartil buscado:

$$Q = \left(\frac{Q}{4}\right) * (n + 1) =$$

Si usamos esta fórmula con el ejemplo anterior nos da los siguientes resultados:

Q1:

$$Q = \left(\frac{1}{4}\right) * (10 + 1) = 2,75$$

Este resultado nos dice que el valor del Q1 se encuentra en la posición 2,75, es decir entre el 17 y el 17, por lo que se interpola ((0,75*(17-17)) +17 = 17.

$$16 + 17 + 17 + 17 + 17 + 18 + 18 + 19 + 19 + 19$$

Conclusión: El 25 % de observaciones se encuentran por debajo de los 17 años.

Q2:

$$Q = \left(\frac{2}{4}\right) * (10 + 1) = 5,5$$

Este resultado nos dice que el valor del Q2 se encuentra en la posición 5,5, es decir entre el 17 y el 18, por lo que se interpola (17+18)/2 = 17,5.

$$16 + 17 + 17 + 17 + 17 + 17 + 18 + 18 + 19 + 19 + 19$$

Conclusión: el 50 % de observaciones se encuentran por debajo de los 17,5 años.

Q3:

$$Q = \left(\frac{3}{4}\right) * (10 + 1) = 8,25$$

Este resultado nos dice que el valor del Q3 se encuentra en la posición 8,25, es decir entre el 19 y el 19, por lo que se interpola ((0,25*(19-19)) +19 = 19.

$$16 + 17 + 17 + 17 + 17 + 18 + 18 + 19 + 19 + 19$$

Conclusión: el 75 % de observaciones se encuentran por debajo de los 19 años.

Deciles

Siguiendo el mismo principio que los cuartiles, los deciles dividen al 100 % de los individuos en diez partes iguales, es decir, cada decil representa un 10 % del total. Por lo tanto, el decil 1 (D1) corresponde al valor que se ubica en el 10 % de los datos, el decil 3 (D3) al 30 %, el decil 7 (D7) al 70 %, el decil 9 (D9) al 90 %, y así sucesivamente con cada uno de los deciles. Su fórmula es la siguiente:

$$D = \left(\frac{D}{10}\right) * (n + 1) =$$

Por ejemplo, si de la anterior distribución queremos calcular el decil 2 (D2), la fórmula sería:

$$D = \left(\frac{2}{10}\right) * (10 + 1) = 2,2$$

Este resultado nos dice que el valor del D2 se encuentra en la posición 2,2, es decir entre el 17 y el 17, por lo que se interpola ((0,2*(17-17)) +17 = 17.

$$16 + 17 + 17 + 17 + 17 + 18 + 18 + 19 + 19 + 19$$

Conclusión: el 20 % de observaciones se encuentran por debajo de los 17 años.

Percentiles

De igual manera, los percentiles dividen a la totalidad de los individuos en 100 partes iguales, lo que significa que cada percentil representa el 1 % del total. Por lo tanto, el percentil 15 (P15) corresponde al valor que se ubica en el 15 % de los datos, el percentil 25 (P25) al 25 %, el percentil 70 (P70) al 70 %, el percentil 93 (P93) al 93 %, y así sucesivamente con cada uno de los percentiles. Los percentiles más usados son el P25 que corresponde al Q1, el P50 que corresponde al Q2 y a la mediana y el P75 que corresponde al Q3, por lo que se puede utilizar la misma fórmula para su cálculo. Para los demás percentiles se debe usar la siguiente fórmula:

$$P = \left(\frac{P}{100}\right) * (n + 1) =$$

Si del anterior ejemplo queremos calcular el percentil 25, la fórmula sería la siguiente:

$$P25:$$

$$Q = \left(\frac{25}{100}\right) * (10 + 1) = 2,75$$

Este resultado nos dice que el valor del P25 se encuentra en la posición 2,75, es decir entre el 17 y el 17, por lo que se interpola ((0,75*(17-17)) +17 = 17.

$$16 + 17 + 17 + 17 + 17 + 18 + 18 + 19 + 19 + 19$$

Conclusión: el 25 % de observaciones se encuentran por debajo de los 17 años.

Medidas de dispersión

Son las que nos ayudan a conocer la variabilidad de los datos, que tan juntos o separados se encuentran. Estas son la desviación estándar, varianza, rango e intervalo intercuartilar. Para comprender las medidas de dispersión, tomaremos como ejemplo las notas sobre 100 puntos de 300 estudiantes de la carrera de Medicina, teniendo como nota mínima 0 y como nota máxima 100, con un promedio de 52 puntos.

Desviación estándar

Teniendo en cuenta el ejemplo antes descrito, supongamos que tenemos un estudiante A con una nota de 70 y un estudiante B con una nota de 21. Para comprender el concepto de desviación estándar, es fundamental partir del concepto de desviaciones individuales.

Estas desviaciones individuales hacen referencia a cuánto se aleja cada valor con respecto a la media del grupo. Es decir, si calculamos la media de las notas y luego restamos a cada nota ese valor promedio, obtendremos la desviación individual.

$$d_i = x_i - \overline{x}$$

d_1=desviación individual del dato i

x_i = valor individual del dato

$\overline{x}$ = media (promedio) del conjunto de datos

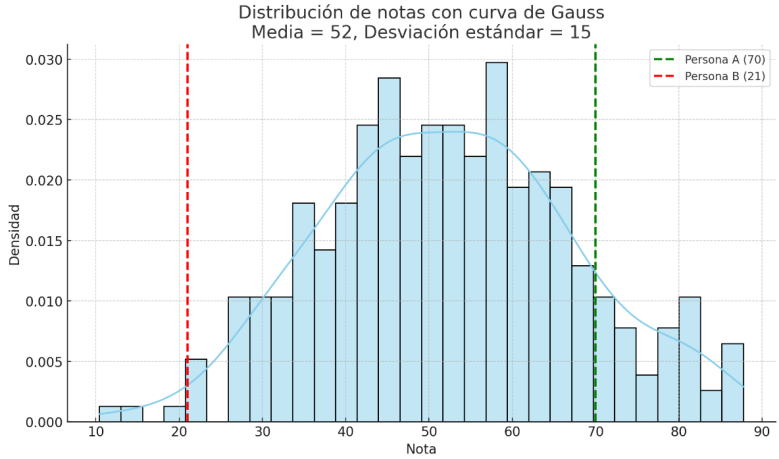
Para el ejemplo que estamos desarrollando las desviaciones de cada individuo serían:

$$d_1 = 70 - 52 = 18 \qquad d_2 = 21 - 52 = -31$$

Como se puede ver, las desviaciones pueden ser positivas o negativas dependiendo de si el valor está por encima o por debajo de la media.

Figura 1

Distribución de notas con curva de Gauss



Si consideramos que este ejercicio podría aplicarse a los 300 individuos del estudio, se entiende el por qué se requiere calcular la desviación estándar. Esta medida permite resumir el comportamiento de todos los datos, indicando en qué medida se dispersan o se concentran alrededor de la media. En lugar de analizar cada una de las 300 desviaciones individuales, la desviación estándar brinda un resumen numérico del grado de variabilidad que existe dentro del conjunto de datos.

Teniendo en cuenta la fórmula de las desviaciones individuales, se puede comprender de mejor manera la construcción de la fórmula de la desviación

estándar. Como las desviaciones individuales pueden ser negativas o positivas, al momento de sumarlas podría ocurrir que se anulen entre sí. Por esta razón, en la fórmula de la desviación estándar se elevan al cuadrado, con el fin de eliminar los signos negativos y considerar únicamente la magnitud de la desviación. Posteriormente, esta sumatoria se divide para el total de los datos. En el caso específico de trabajar con muestras, se divide para n - 1, como corrección del sesgo muestral. Finalmente, se extrae la raíz cuadrada del resultado, con lo cual se retorna a la unidad original de la variable analizada.

Desviación estándar para una muestra

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

s : Desviación estándar muestral
x_i : Cada uno de los valores individuales
$\bar{x}$: Media de la muestra
n : Numero totalde datos en la muestra

Desviación estándar cuando se trabaja con toda la población (universo)

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{n}}$$

σ : Desviación estándar poblacional
x_i : Cada uno de los valores individuales
μ : Media de lapoblación
n : Numero totalde datos en la población

Tabla 1

Ejemplo de cómo aplicar la fórmula en una pequeña muestra de 10 individuos (del ejemplo anterior)

Valor	xi - 69,5	(xi- 69,5)2
70	(70-69,5)=0,5	0,25
70	(70-69,5)=0,5	0,25
60	(60- 69,5)= -9,5	90,25
70	(70-69,5)=0,5	0,25
70	(70-69,5)=0,5	0,25
60	(60-69,5)=-9,5	90,25
70	(70-69,5)=0,5	0,25
80	(80-69,5)=10,5	110,25
70	(70-69,5)=0,5	0,25
75	(75-69,5)=5,5	30,25
695		322,5

Media= 69,5

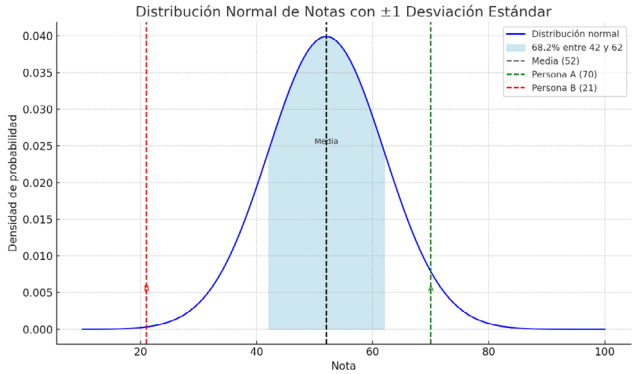
$$s = \sqrt[2]{322,5/(10 - 1)} = 5,98$$

Conclusión: la desviación estándar es de 5,98.

Si consideramos el ejemplo de los 300 individuos estudiados, en donde la media de las notas es 52 y la desviación estándar calculada es ± 10 puntos, y tomando en cuenta el concepto de normalidad que se abordará en capítulos posteriores, podemos señalar que el 68,2 % de la población se encuentra dentro del rango de una desviación estándar respecto a la media, es decir, entre 42 y 62 puntos. Por esta razón, como se observa en el gráfico, ni la persona A (con nota de 70) ni la persona B (con nota de 21) se encuentran dentro de ese 68,2 % de los individuos centrales.

Figura 2

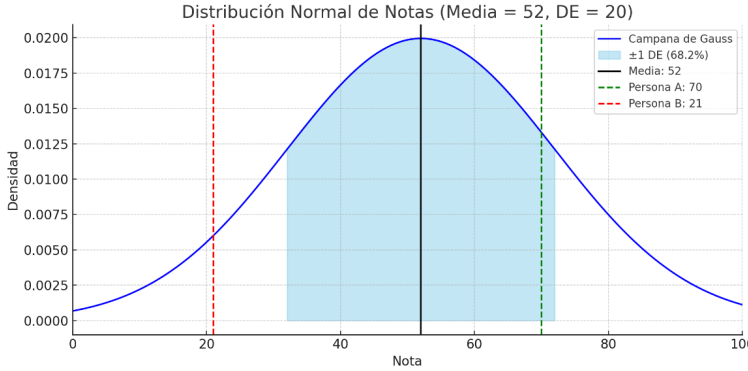
Distribución normal de notas con ±1 desviación estándar



Ahora vamos a suponer que, en el mismo ejercicio, el cálculo de la desviación estándar gracias al comportamiento de los datos resulta ser de ±20. Como se puede observar en el siguiente gráfico, la persona A ya se ubica dentro del 68,2 % de la población. Esto evidencia de manera clara que la desviación estándar es una medida de dispersión que describe qué tan alejados o concentrados están los datos con respecto a la media. En este caso, al tener una desviación estándar de 20, la distribución es mucho más amplia o dispersa, que con una desviación estándar de 15.

Figura 3

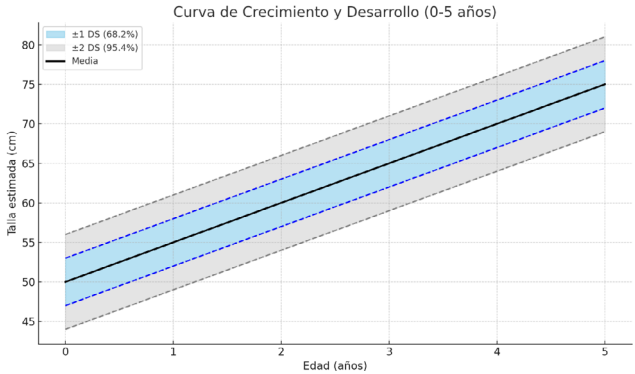
Distribución normal de notas



En la aplicación de las ciencias de la salud, el principio de la desviación estándar se evidencia de forma clara en las curvas de crecimiento y desarrollo, frecuentemente utilizadas en el seguimiento de niños y niñas sanas. Como se observa en la siguiente curva, cuando un individuo se encuentra por fuera del rango de ±2 desviaciones estándar (95,4 % de la población con relación a la media), se considera que podría presentar posibles alteraciones o retrasos en el crecimiento y desarrollo, especialmente en lo que respecta a la talla.

Figura 4

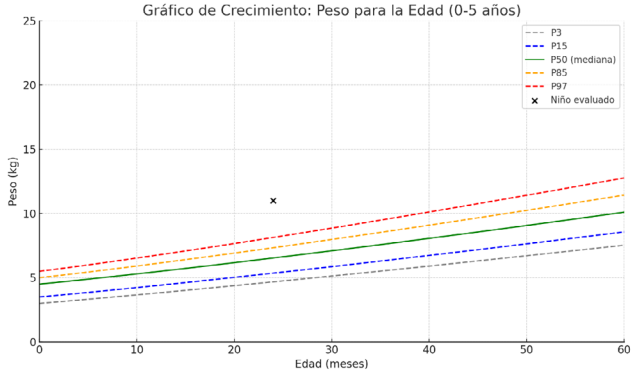
Curva de crecimiento y desarrollo (0-5 años)



Es importante tener en cuenta que los gráficos de crecimiento y desarrollo pueden estar bajo el concepto de cuartiles, deciles y percentiles antes descritos y que se observan en el siguiente gráfico.

Figura 5

Gráfico de crecimiento



Cabe indicar que ninguno de los gráficos antes descritos de crecimiento y desarrollo es real y no pueden ser usados en la práctica en salud, ya que han sido creados para la comprensión de los temas.

Varianza

Una vez comprendido el concepto de la desviación estándar, resulta más sencillo entender qué es la varianza. La varianza, en realidad, es el paso previo en el cálculo de la desviación estándar. Es decir, consiste en realizar la sumatoria de las desviaciones individuales elevadas al cuadrado y dividir ese resultado para el total de datos (n, o n-1 si se trata de una muestra). El valor que se obtiene corresponde a la varianza. En pocas palabras es la desviación estándar al cuadrado.

La principal diferencia entre ambas medidas radica en las unidades. La varianza mantiene las unidades al cuadrado, mientras que la desviación estándar, al aplicarle la raíz cuadrada, devuelve el resultado en las unidades originales de la variable. Esta es una de las razones por las cuales, en la mayoría de los estudios, se prefiere presentar la desviación estándar como medida de dispersión, ya que es más fácil de interpretar.

Rango

El rango no es más que la diferencia que existe entre el valor mínimo y el valor máximo, registrados entre los individuos. Esta medida proporciona una idea general de la dispersión de los datos. Por ejemplo, no es lo mismo tener un rango de notas entre 0 y 100 que entre 25 y 100, ya que en el primer caso la dispersión es mayor y refleja una variabilidad más amplia entre los individuos evaluados.

Rango intercualtílico (RIQ)

Corresponde a la resta entre los valores del Q3 y el Q1, y representa el comportamiento del 50 % de los valores centrales de la distribución. Es importante tener en cuenta que esta medida se relaciona directamente con la Mediana, ya que se enfoca en la parte central de los datos, dejando de lado los valores extremos que podrían distorsionar otras medidas de dispersión. En el ejemplo de los cuartiles el Q1 era 17 y el Q3 19, es decir en rango seria = 19-17 = 2.

Intervalo intercuartilar

Son los valores que existen entre el Q1 y el Q3 y representan el 50 % de las observaciones centrales. En el ejemplo de los cuartiles el Q1 era 17 y el Q3 19, es decir que el 50 % de las observaciones centrales se encontraban entre los 17 a los 19 años.

Medidas de forma

Las medidas de forma son las que nos muestran cómo se comporta la distribución alrededor de los valores centrales (media), entre estas tenemos la asimetría y la curtosis. Para abordar las medidas de forma utilizaremos el ejemplo de las notas con diferentes distribuciones de los datos para comprender cada uno de los conceptos.

Curtosis

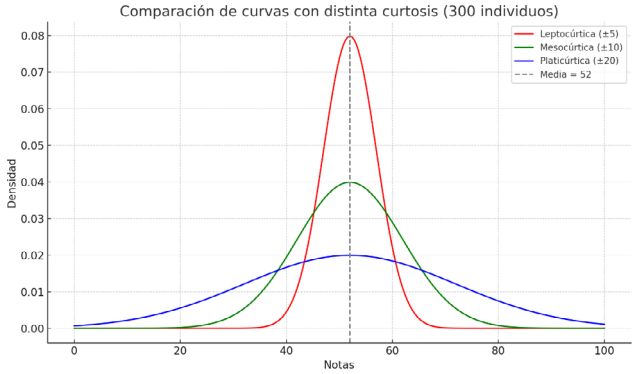
La curtosis describe cómo se distribuyen los datos alrededor de la media, específicamente el grado de concentración de los valores alrededor de la media, y la importancia de los valores extremos (colas de la distribución).

Tal como se observa en la figura, si la curva se presenta aplanada (curva azul), recibe el nombre de platicúrtica, en este caso se observa una menor concentración de datos en el centro, colas más ligeras y los datos están más repartidos. Cuando la curva muestra un apuntalamiento marcado hacia arriba, se denomina leptocúrtica (curva roja), observándose una mayor acumulación de datos cerca de la media y colas más pesadas (más valores extremos).

En cambio, si la curva mantiene una forma intermedia y centrada, se identifica como mesocúrtica (curva verde), indicando que tiene una concentración central y colas similares a la distribución normal de datos.

Figura 6

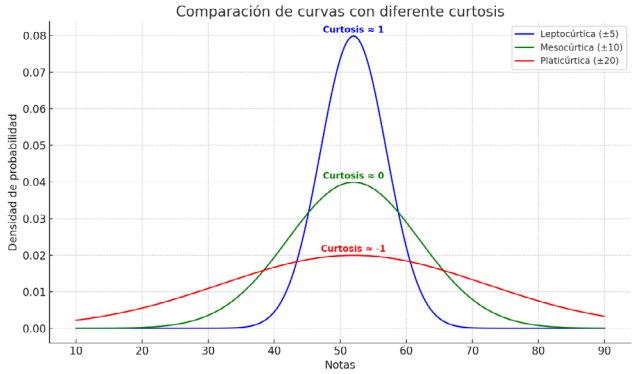
Comparación de curvas con curtosis distinta



Cabe señalar que algunos de los programas de procesamiento de bases de datos calculan el coeficiente de curtosis, el cual permite interpretar la forma de la curva. Cuando este coeficiente es mayor a cero, es decir, positivo, indica una curva leptocúrtica, con un mayor apuntalamiento, especialmente si el valor es alto. En cambio, cuando el coeficiente es menor a cero, es decir, negativo, la curva es platicúrtica, más aplanada, y esta característica se acentúa mientras más se aleja del 0. Y si el valor es de 0, la curva sería mesocúrtica

Figura 7

Comparación de curvas con curtosis diferente



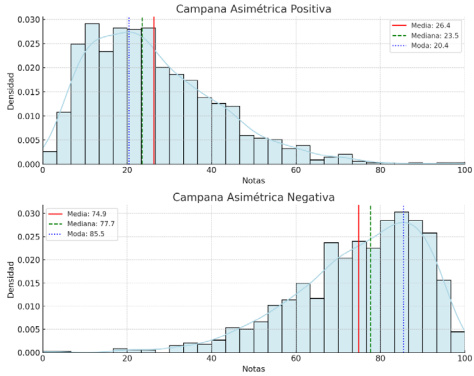
Simetría

Según las medidas de tendencia central, es decir, la media, la mediana y la moda, las curvas pueden presentar un comportamiento simétrico cuando los tres estadísticos coinciden en el mismo valor. Es importante recordar que la media es el estadístico más sensible a los valores extremos, y tiende a desplazarse hacia ellos. Por esta razón, cuando la moda permanece en el lado izquierdo y la media se desvía hacia la derecha, la curva adopta un comportamiento de asimetría positiva, lo que se traduce gráficamente en una cola más pronunciada hacia el lado derecho.

Por el contrario, si la moda se mantiene en el lado derecho y la media se desplaza hacia valores extremos en el lado izquierdo, estamos frente a una curva con asimetría negativa, caracterizada por una cola alargada hacia la izquierda.

Figura 8

Campana asimétrica positiva



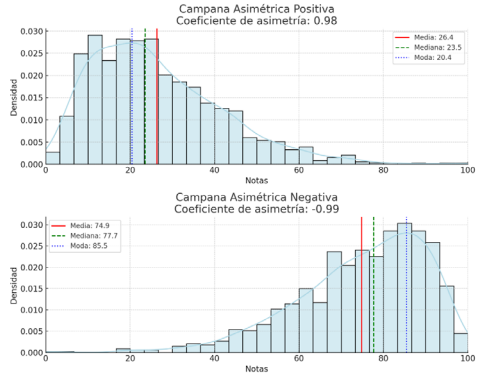
Algunos de los programas de procesamiento de bases de datos calculan el coeficiente de asimetría, el cual permite interpretar la simetría o sesgo de una distribución de datos. Cuando este coeficiente es igual o cercano a cero, indica una distribución simétrica, en la que los valores están distribuidos de manera equilibrada a ambos lados de la media.

Cuando el valor del coeficiente mayor a cero, es decir, es positivo, la distribución se considera asimétrica positiva, lo que significa que presenta una cola alargada hacia la derecha. En cambio, cuando el coeficiente es menor a cero, es decir, es negativo, la distribución se interpreta como asimétrica negativa, con una cola extendida hacia la izquierda.

Además, mientras más se aproxima el coeficiente a +1 o a –1, más marcada será la asimetría y más alargada la cola. Es decir, valores cercanos a +1 indican una asimetría positiva más acentuada, y valores cercanos a –1, una asimetría negativa más pronunciada.

Figura 9

Campanas asimétricas: positiva y negativa



Cabe indicar que los conceptos de curtosis y asimetría se relacionan con el concepto de normalidad que veremos en el capítulo de distribución de los datos.

Conclusiones y recomendaciones

El análisis descriptivo numérico comprende una serie de herramientas y medidas que permiten obtener una visión general del comportamiento de las variables cuantitativas, sirviendo como base para el uso posterior de pruebas estadísticas más complejas. Sin embargo, es importante señalar que este análisis solo ofrece un vistazo a los datos y no establece relaciones causales ni de asociación entre las variables.

Al calcular medidas de tendencia central, es fundamental tener en cuenta que la media es muy sensible a los valores extremos, lo que puede distorsionar la interpretación de los resultados. En este sentido, la mediana puede ser un indicador más confiable, especialmente en distribuciones asimétricas. Las medidas de posición permiten identificar la localización de los datos dentro de la distribución, mientras que las medidas de forma facilitan la interpretación de la distribución de los datos, ayudando a comprender si los datos están sesgados o presentan una forma particular.

Como recomendación, se sugiere diseñar los cuestionarios de recolección de datos con la posibilidad de obtener datos cuantitativos puros, ya que esto

permite realizar un análisis más completo y flexible, con la opción de categorizar las variables posteriormente. Además, se resalta la importancia de analizar las medidas de forma antes de elegir el tipo de prueba estadística, ya que conocer la distribución de los datos es clave para decidir si utilizar pruebas paramétricas o no paramétricas.

Referencias

Celis de la Rosa, A. J., y Labrada Martagón, V. (2014). *Bioestadística* (3.ª ed.). Ciudad de publicación desconocida: Editorial El Manual Moderno.

Fayad Cámel, V. (1970). *Estadística médica y de salud pública* (2.ª ed.). Mérida (VE): Universidad de los Andes.

Gómez Barrantes, M. (2012). *Elementos de la estadística descriptiva* (4.ª ed.). Costa Rica: Universidad Estatal a Distancia de Costa Rica.

Hernández Sampieri, R., Fernández Collado, C., y Baptista Lucio, P. (2014). *Metodología de la investigación* (6.ª ed.). México: McGraw-Hill Education.

Martínez González, M. A., Sánchez-Villegas, A., y Toledo Atucha, E. (2020). *Bioestadística amigable* (4.ª ed.). Madrid: Elsevier.

Ministerio de Sanidad y Consumo de España. (2008). *Manual de Bioestadística para Ciencias de la Salud*. Madrid: Ministerio de Sanidad.

Capítulo 13
Análisis descriptivo gráfico: variables cuantitativas y epidemiológico

Introducción

Este capítulo se centra en el abordaje de los principales tipos de gráficos utilizados para resumir y analizar variables cuantitativas.

Se describirán a detalle gráficos como diagrama de dispersión, útil para representar pequeñas cantidades de observaciones individuales; histograma, indicado para mostrar la distribución de frecuencias de variables continuas; polígono de frecuencias, a través del cual se puede visualizar la forma de la distribución y compararla entre grupos; y el gráfico de cajas y bigotes, una herramienta eficaz para describir la dispersión y simetría de los datos.

En el ámbito epidemiológico, se mencionará el uso de gráficos de series de tiempo, importantes al momento de evaluar la evolución de una variable a lo largo del tiempo, y la pirámide poblacional, utilizada para describir la distribución de una población según edad y sexo.

El objetivo de este capítulo es explicar la utilidad de cada tipo de gráfico, normas para su elaboración e interpretación, con ello se pretende desarrollar competencias en el manejo gráfico de datos cuantitativos y epidemiológicos.

Desarrollo

Principales gráficos para análisis de variables cuantitativas

Antes de abordar el presente capítulo, es importante recordar que una variable cuantitativa, al ser categorizada, pasa a comportarse como una variable cualitativa ordinal. Esto permite representarla utilizando gráficos

propios de variables cualitativas, como el gráfico de sectores, las barras simples, barras dobles o barras apiladas, según lo que se desee comparar o destacar.

Es importante mencionar que los gráficos utilizados en variables cuantitativas dependerán del objeto del estudio. Cuando se quiere describir una sola variable cuantitativa, por ejemplo, su distribución en una población, se pueden aplicar gráficos como el histograma, pero si lo que se busca es realizar cruces de variables, existen otras opciones: por ejemplo, el gráfico de dispersión es útil para representar el cruce entre dos variables cuantitativas, mientras que el polígono de frecuencias y el gráfico de cajas y bigotes permiten comparar la distribución de una variable cuantitativa entre varias poblaciones.

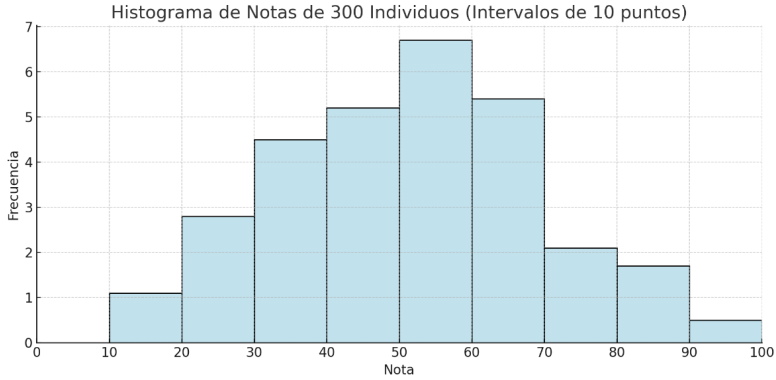
Histograma

Es un gráfico que, a primera vista, puede parecerse al gráfico de barras utilizado para variables cualitativas. Sin embargo, existe una diferencia clave: en el histograma, la variable cuantitativa se representa en el eje horizontal (x), pero de forma categorizada en intervalos continuos, no como categorías aisladas. A diferencia del gráfico de barras dobles, en el histograma las barras se encuentran pegadas entre sí, lo que refleja la continuidad de los datos y la agrupación en intervalos.

La altura de cada barra representa la frecuencia con la que los valores del intervalo correspondiente aparecen en la población. Como vimos en el capítulo anterior, el histograma no solo permite visualizar cómo se distribuyen los datos, sino también identificar su forma, evaluar su asimetría y observar si presenta una curtosis más apuntalada o aplanada.

Figura 1

Ejemplo gráfico de un histograma,



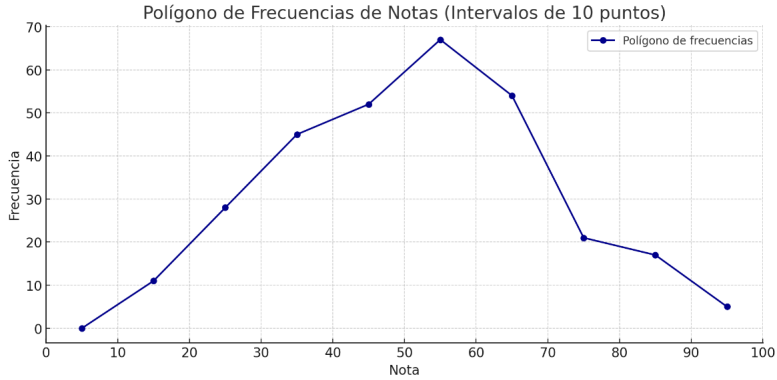
Nota. Muestra la distribución de una variable cuantitativa en intervalos continuos.

Polígono de frecuencias

Se construye a partir del gráfico del histograma ubicando los puntos centrales en la parte superior de cada barra y uniéndolos con una línea. De esta manera, se deja ver por debajo una tendencia general de los datos.

Figura 2

Ejemplo gráfico de un polígono de frecuencias

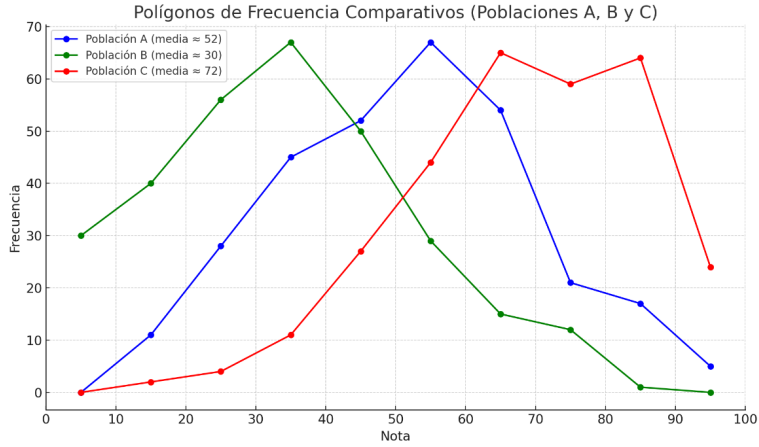


Nota. Construido a partir de un histograma, mostrando la tendencia general de los datos mediante una línea que conecta los puntos centrales de cada barra.

El uso principal que se da a los polígonos de frecuencia es superponerlos en un mismo gráfico cuando se trabaja con dos o más poblaciones, con el objetivo de comparar visualmente sus tendencias generales y observar similitudes o diferencias en la distribución de los datos.

Figura 3

Ejemplo gráfico de polígonos de frecuencias superpuestos



Nota. Se usa para comparar las tendencias generales de dos poblaciones.

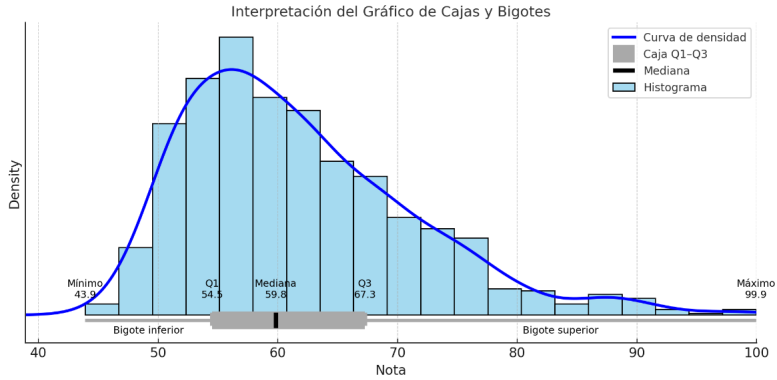
Gráfico de cajas y bigotes

En él se puede observar con facilidad la ubicación del cuartil 1 (Q1), la mediana (Q2) y el cuartil 3 (Q3), elementos que se relacionan directamente con la asimetría que ya habíamos abordado en el capítulo anterior.

Desde el valor mínimo hasta el primer cuartil se extiende el bigote inferior, mientras que desde el cuartil 3 hasta el valor máximo se traza el bigote superior. Entre Q1 y Q3 se encuentra la caja, que representa el intervalo intercuartilar e incluye al 50 % central de la población, lo que permite visualizar la concentración y dispersión de los datos de forma clara y sintética.

Figura 4

Ejemplo gráfico de un gráfico de cajas y bigotes

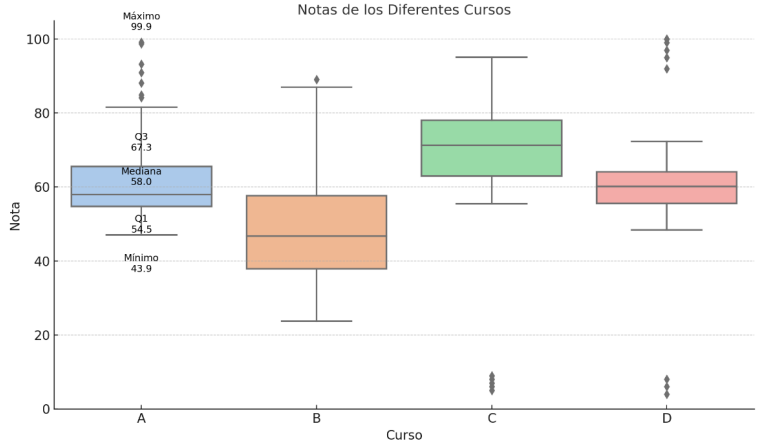


Nota. Muestra la ubicación de los cuartiles Q1, Q2 (mediana) y Q3, así como los bigotes inferior y superior, que representan la dispersión de los datos.

Como podemos ver en la figura anterior, los bigotes, tanto superior como inferior, pueden extenderse hasta abarcar la totalidad de los valores observados, incluyendo aquellos valores extremos. Sin embargo, como vamos a observar en el siguiente ejemplo, cuando se realizan comparaciones en diferentes poblaciones, los bigotes pueden excluir a los valores extremos y a cambio usar percentiles. Cuando dentro de las cajas se observa una x, este valor representa la media.

Figura 5

Ejemplo gráfico de un gráfico de cajas y bigotes con exclusión de valores extremos



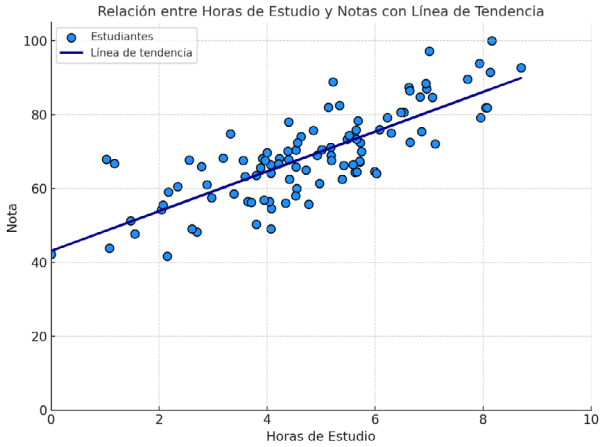
Nota. Utilizan percentiles en lugar de los bigotes para representar la dispersión, y con la media indicada por una "x" dentro de la caja

Dispersión

Es un gráfico que permite visualizar la relación entre dos variables cuantitativas. En el eje X se ubica la variable independiente, mientras que en el eje Y se representa la variable dependiente. Al graficar los valores correspondientes de cada individuo, se obtiene un conjunto de puntos dispersos que reflejan cómo se comportan ambas variables en conjunto, tal como se observa a continuación.

Figura 6

Ejemplo gráfico de un diagrama de dispersión,



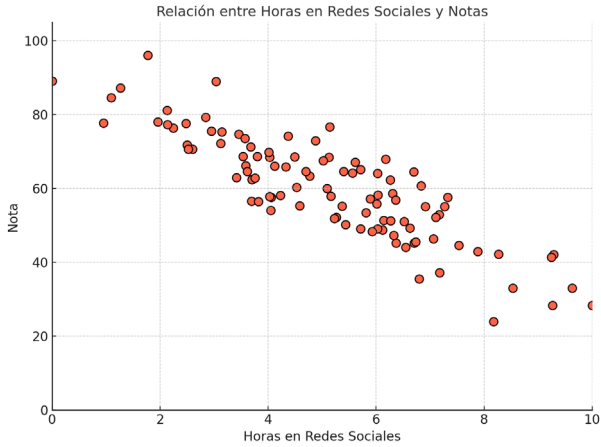
Nota. Muestra la relación entre dos variables cuantitativas, con la variable independiente en el eje X y la variable dependiente en el eje Y.

Como se puede observar, el gráfico de dispersión permite identificar una tendencia general en los datos. En este caso, se aprecia que, a mayor cantidad de horas de estudio, mayor es la nota obtenida, lo que evidencia una tendencia positiva entre ambas variables. Cuando a este gráfico se le incorpora una línea que resume dicha relación, pasa a denominarse gráfico de tendencia, el cual es clave en los análisis de regresión lineal.

También puede observarse una tendencia negativa, como en el caso del gráfico que muestra el cruce entre la variable independiente, horas dedicadas a las redes sociales y las notas. En este ejemplo, a mayor tiempo en redes sociales, se evidencian calificaciones más bajas.

Figura 7

Ejemplo gráfico de un diagrama de dispersión(con tendencia positiva entre las horas de estudio y las notas obtenidas) y un gráfico de tendencia



Gráficos adicionales para análisis epidemiológico

Es importante señalar que todos los gráficos descritos anteriormente tienen un amplio uso en el análisis epidemiológico de distintos eventos de salud. No obstante, existen otros tipos de gráficos que abordaremos a continuación, los cuales resultan especialmente útiles para análisis epidemiológicos.

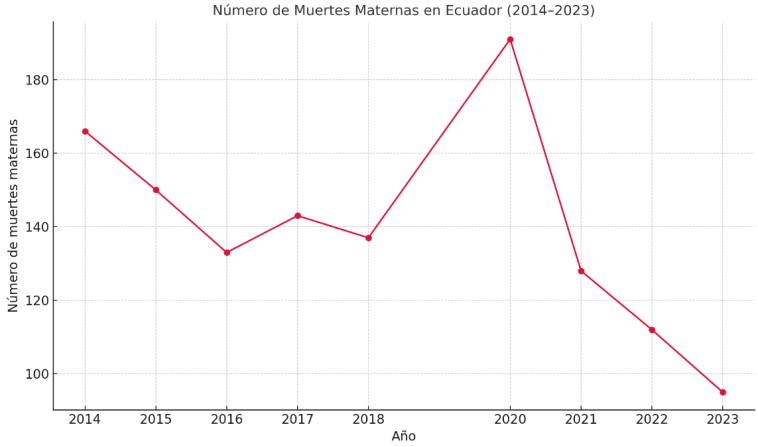
Gráfico de series de tiempo (gráfico de línea temporal)

Este gráfico presenta en el eje X el tiempo, que puede estar representado en años, meses o días, y en el eje Y el número de casos o las tasas de un evento de salud específico.

Se suele acompañar con una línea que une los distintos puntos correspondientes a cada periodo, lo que permite visualizar de forma clara si existe una tendencia ascendente, descendente o estacionaria. Con este gráfico, se busca analizar cómo evoluciona un evento de salud a lo largo del tiempo.

Figura 8

Ejemplo gráfico de una serie de tiempo

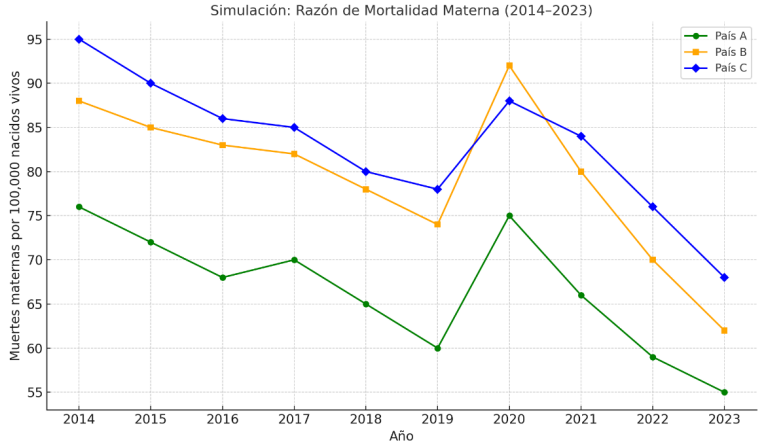


Nota. Muestra la evolución de un evento de salud específico a lo largo del tiempo, con el tiempo en el eje X y el número de casos o tasas en el eje Y.

Cuando se usan las tasas es factible usar el presente gráfico para comparar las tendencias en diferentes contextos, como vemos a continuación.

Figura 9

Ejemplo gráfico de una serie de tiempo



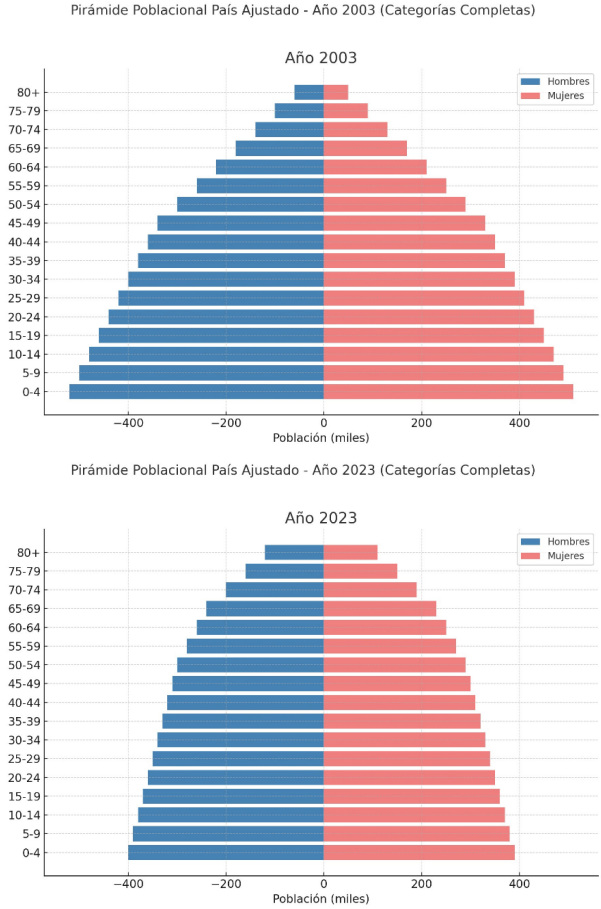
Nota. Compara las tendencias de un evento de salud en diferentes contextos, utilizando tasas para visualizar la evolución a lo largo del tiempo.

Pirámide poblacional

Se presenta como dos histogramas unidos por sus bases del eje de las X, permitiendo visualizar el comportamiento de la variable edad, con respecto al sexo (masculino y femenino) de una población determinada, (la edad se puede categorizar según las recomendaciones de la Organización Mundial de la Salud). Este tipo de gráfico facilita la identificación de cambios en los distintos grupos etarios a lo largo del tiempo al comparar diferentes pirámides a la vez.

Figura 10

Ejemplo gráfico de una pirámide poblacional



Nota. Muestra la distribución de la población por edad y sexo (masculino y femenino), permitiendo visualizar cambios en los grupos etarios a lo largo del tiempo

Conclusiones y recomendaciones

La aplicación de gráficos en el análisis cuantitativo permite visualizar de manera fácil y rápida el comportamiento y la distribución de variables cuantitativas. A través de esta herramienta, el lector puede obtener una representación que le permita identificar patrones, tendencias, simetría, presencia de valores extremos y realizar comparaciones entre grupos o series temporales.

Es importante recordar que cada tipo de gráfico cumple un objetivo específico. Por ejemplo, los histogramas permiten observar la forma de distribución de los datos a través de intervalos de variables continuas, mientras que el gráfico de cajas y bigotes muestra la distribución de los datos mediante cuartiles, indicando claramente la presencia de valores atípicos.

En el análisis gráfico epidemiológico, el gráfico de series temporales permite analizar el comportamiento de una variable a lo largo del tiempo, siendo útil para identificar brotes, estacionalidades o el impacto de estrategias aplicadas. De igual manera, la pirámide poblacional proporciona información sobre la estructura de una población de acuerdo al sexo y la edad, lo cual es fundamental para estudios demográficos.

Se recomienda tener en cuenta las variables a ser analizadas y el objetivo del análisis antes de elegir el gráfico adecuado para presentar los datos. Además, es esencial conocer el procedimiento para su construcción utilizando diferentes softwares estadísticos, lo que facilitará una correcta interpretación y presentación de los resultados.

Referencias

Berenson, M. L., Levine, D. M., y Krehbiel, T. C. (2006). *Fundamentos de estadística* (6.ª ed.). México: Pearson Educación.

Celis de la Rosa, A. J., y Labrada Martagón, V. (2014). *Bioestadística* (3.ª ed.). Ciudad de publicación desconocida: Editorial El Manual Moderno.

Fayad Cámel, V. (1970). *Estadística médica y de salud pública* (2.ª ed.). Mérida (VE): Universidad de los Andes.

Gómez Barrantes, M. (2012). *Elementos de la estadística descriptiva* (4.ª ed.). Costa Rica: Universidad Estatal a Distancia de Costa Rica.

Martínez González, M. A., Sánchez-Villegas, A., y Toledo Atucha, E. (2020). *Bioestadística amigable* (4.ª ed.). Madrid: Elsevier.

Ministerio de Sanidad y Consumo de España. (2008). *Manual de Bioestadística para Ciencias de la Salud*. *Madrid*: Ministerio de Sanidad.

Capítulo 14
Distribución de los datos

Introducción

El análisis estadístico de datos en el ámbito de la salud requiere una comprensión adecuada de cómo se distribuyen las variables que se estudian. La distribución de una variable describe la forma en la que se presentan sus valores en una población o muestra. Conocer dicha distribución permite seleccionar las herramientas estadísticas más apropiadas para el análisis, ya sea descriptivo o inferencial.

Una de las distribuciones más importantes en Bioestadística es la distribución normal, también conocida como distribución de Gauss, debido a su relevancia en fenómenos naturales y sociales. Muchas variables biológicas, como la presión arterial, la talla, el peso o los niveles de colesterol, tienden a seguir este tipo de distribución en poblaciones suficientemente grandes y homogéneas.

La distribución normal posee características matemáticas particulares que permiten aplicar con validez procedimientos estadísticos como el cálculo de intervalos de confianza, pruebas de hipótesis y regresiones. Sin embargo, no todas las variables se ajustan a este patrón. Por ello, verificar la normalidad de los datos se convierte en un paso previo obligatorio antes de aplicar métodos paramétricos.

Además de la distribución normal, existen otras distribuciones de probabilidad que modelan adecuadamente distintos tipos de datos y situaciones, como la distribución binomial, la distribución de Poisson, la t de Student, entre otras. Reconocer cuál es la más adecuada para un determinado análisis es parte de las competencias fundamentales en Bioestadística.

En este capítulo se abordarán los conceptos clave relacionados con la distribución de los datos, con énfasis en la distribución normal, el uso de la puntuación Z, la identificación de otras distribuciones relevantes y los métodos para contrastar la normalidad. Este conocimiento no solo es esencial para aplicar correctamente las pruebas estadísticas, sino también para interpretar con rigor los resultados de una investigación biomédica.

Desarrollo

Distribución normal

La distribución normal, también llamada distribución gaussiana, es una de las funciones de probabilidad más importantes y más utilizadas en estadística. Representa el comportamiento de muchas variables cuantitativas continuas que se observan en la naturaleza y en el ámbito de la salud. Se caracteriza por una forma de campana simétrica, en la que la mayoría de los valores se agrupan alrededor de la media, y las frecuencias disminuyen gradualmente hacia los extremos.

Esta distribución es teórica, pero muchas variables empíricas tienden a aproximarse a ella, especialmente cuando se cumple el supuesto de homogeneidad y el tamaño muestral es suficientemente grande. Por esta razón, la distribución normal es un modelo estadístico de referencia, y numerosos métodos de inferencia estadística (como la prueba t, ANOVA, intervalos de confianza, entre otros) están basados en la suposición de normalidad.

Función de densidad de probabilidad

Matemáticamente, la función de densidad de una variable aleatoria continua *X*, con distribución normal de media μ y desviación estándar σ, se expresa como:

$$\int(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{z-\mu}{\sigma}\right)^2}$$

Esta función cumple dos propiedades clave:

- El área bajo la curva es igual a 1 (representa el 100 % de las probabilidades).
- Es simétrica respecto a la media.

Parámetros de la distribución normal

- Media (μ): valor central de la distribución. Representa el promedio de los datos.
- Desviación estándar (σ): mide la dispersión de los datos respecto a la media. Cuanto mayor es σ, más "aplanada" es la curva; cuanto menor es σ, más "concentrada" es.

Propiedades esenciales

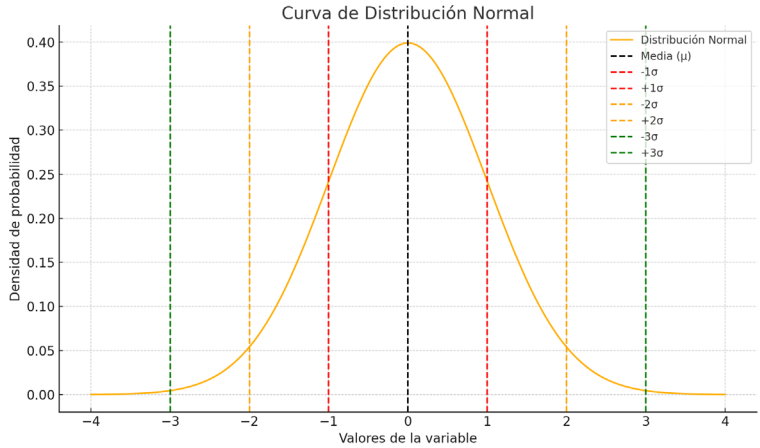
Tabla 1

Propiedades de la distribución normal

Propiedad	Explicación
Simetría	La curva es simétrica respecto a la media. La mitad de los valores se encuentra a cada lado.
Media=mediana=moda	En la distribución normal, estos tres estadísticos coinciden.
Campana mesocúrtica	Tiene una forma de campana moderada (ni muy picuda ni muy plana).
Área total=1	Toda el área bajo la curva representa el 100 % de las observaciones posibles.
Regla empírica	Aproximadamente el 68,3 % de los datos está entre ±1σ, el 95,4 % entre ±2σ, y el 99,7 % entre ±3σ de la media.
Extremos asintóticos	La curva nunca toca el eje horizontal, se extiende indefinidamente hacia ambos extremos.

Figura 1

Curva de distribución normal



Aplicaciones en medicina

La talla de los recién nacidos, el colesterol sérico, o la presión arterial suelen seguir distribuciones aproximadamente normales en poblaciones saludables.

Evaluar la normalidad de una variable es importante para aplicar los test paramétricos como el t de Student (para comparación de medias), o para calcular intervalos de confianza válidos.

Ejemplo práctico contextualizado: en una población de estudiantes de Medicina, la presión arterial sistólica tiene una media de 120 mmHg y una desviación estándar de 10 mmHg. Esta variable sigue una distribución normal. ¿Cuál es la probabilidad de que un estudiante tenga una presión sistólica entre 110 y 130 mmHg?

Aplicando la regla empírica ($\pm 1\sigma$), se concluye que:

$$P(110 \leq x \geq 130) \approx 68,3\%$$

Es decir, alrededor del 68,3 % de los estudiantes tendrán una presión arterial en ese rango, lo que es útil para establecer criterios de referencia, detección de outliers (valores extremos), o para estratificar riesgos en estudios clínicos.

Distribución normal estándar y puntuación Z

La distribución normal estándar es una forma particular de la distribución normal en la que la media es igual a 0 y la desviación estándar es igual a 1. Se representa como N (0,1) y se utiliza como modelo de referencia para comparar datos provenientes de cualquier distribución normal.

La ventaja de transformar cualquier distribución normal a su forma estándar es que permite comparar resultados provenientes de diferentes contextos, e interpretar probabilidades sin necesidad de calcular integrales complejas, ya que existen tablas de la distribución Z ampliamente disponibles o funciones en programas estadísticos como EPIDAT, Excel o R.

La puntuación Z (también llamada "valor Z" o "estandarización") indica a cuántas desviaciones estándar se encuentra un valor con respecto a la media de su distribución.

$$z = \frac{x - \mu}{\sigma}$$

Donde:

z = puntuación estándar

x = valor observado

μ = media de la distribución

σ = desviación estándar

Esta fórmula transforma cualquier valor de una distribución normal con media μ y desviación estándar σ, a la escala estándar con media 0 y desviación 1.

Interpretación de la puntuación Z

Si z = 0: el valor está exactamente en la media.

Si z > 0: el valor está por encima de la media.

Si z < 0: el valor está por debajo de la media.

Si z > 2: el valor se considera lejano a la media (posible outlier, dependiendo del contexto).

Ejemplo en contexto clínico: una paciente tiene un nivel de hemoglobina de 10,5 g/dL. En su población, la hemoglobina tiene una media de 13,5 g/dL y una desviación estándar de 1,5 g/dL. ¿Qué tan alejado está su valor de la media?

$$z = \frac{10{,}5 - 13{,}5}{1{,}5} = \frac{-3}{1{,}5} = -2$$

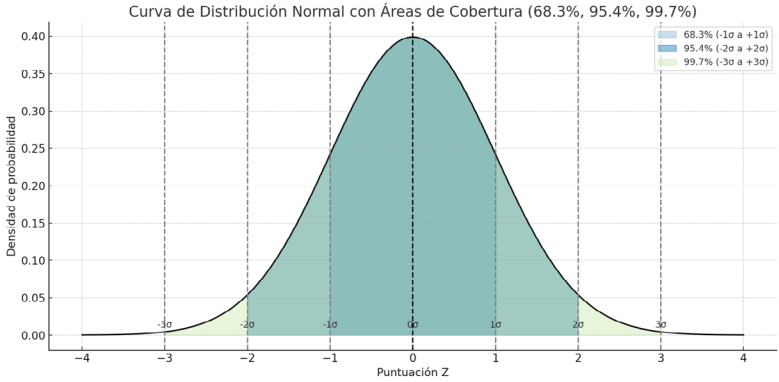
Interpretación: la paciente tiene un valor 2 de desviaciones estándar por debajo de la media. Esto indica un nivel significativamente bajo, compatible con anemia, y por tanto estadísticamente infrecuente (menos del 5 % de la población tendría un valor tan bajo si se distribuye normalmente).

Uso de la distribución normal estándar para calcular probabilidades

Una vez que el valor ha sido transformado a su puntuación Z, es posible calcular probabilidades asociadas a ese valor utilizando tablas o software. Las probabilidades corresponden al área bajo la curva normal estándar.

Figura 2

Valores críticos de la distribución normal estandarizada



Estas áreas representan la proporción de la población que se encuentra dentro de ese rango de valores, lo cual es útil para evaluar si un resultado es común o raro dentro de una población.

Ejemplo aplicado con pasos: en una panadería, la producción diaria de panes tiene una media de 158 unidades y una desviación estándar de 4 panes. ¿Cuál es la probabilidad de producir menos de 170 panes al día?

Paso 1: calcular el valor Z.

$$z = \frac{170 - 158}{4} = \frac{12}{4} = 3$$

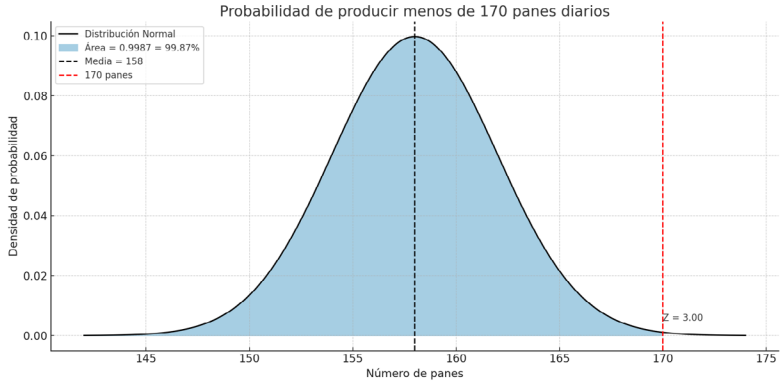
Paso 2: buscar el valor de Z en una tabla o software.

El área bajo la curva para Z = + 3 es de 0,4987

Como la pregunta era la probabilidad de producir menos de 170, a los 0,4987 que representa el +3 desviaciones estándar, se debe sumar 0,5 de los valores que están por debajo de la media, es decir: 0,4987 + 0,5 = 0,9987.

Figura 3

Probabilidad de producir menos de 170 panes al día



Conclusión: hay un 99,87 % de probabilidad de que en un día se produzcan menos de 170 panes.

Otros tipos de distribución

Aunque la distribución normal es ampliamente utilizada en Bioestadística, no todas las variables o situaciones del ámbito médico se ajustan a este patrón. Existen múltiples distribuciones de probabilidad que permiten modelar distintos tipos de fenómenos, según el tipo de variable (cualitativa o cuantitativa), su escala de medición y el objetivo del análisis.

Estas distribuciones se clasifican en dos grandes grupos:

- Distribuciones de variables discretas, que se utilizan cuando los valores posibles de la variable son finitos o contables (por ejemplo, número de hijos, número de consultas, número de eventos adversos).
- Distribuciones de variables continuas, que se aplican cuando los valores pueden tomar cualquier número dentro de un intervalo (por ejemplo, peso, talla, niveles de glucosa).

Comprender y seleccionar la distribución adecuada es crucial para aplicar pruebas estadísticas correctamente, interpretar sus resultados y estimar probabilidades o parámetros poblacionales con validez.

A continuación, se presenta una tabla resumen con algunas de las distribuciones más comunes en el análisis estadístico en salud.

Tabla 2

Otros tipos de distribución de los datos

Propiedad	Aplicacion principal	Tipo de variable
Binomial	Número de éxitos en n ensayos	Cualitativa dicotómica
Poisson	Conteo de eventos en un intervalo	Cuantitativa discreta
Chi-cuadrado (χ^2)	Comparación de frecuencias	Cualitativa
t de Student	Comparación de medias	Cuantitativa continua
F de Snedecor (F)	Comparación de varianzas	Cuantitativa continua

Ejemplos clínicos ilustrativos

- La distribución binomial se usa, por ejemplo, para modelar la probabilidad de que un paciente responda o no a un tratamiento en una serie de pruebas clínicas.
- La distribución de Poisson puede aplicarse para estimar la probabilidad de que ocurran ciertos efectos adversos poco frecuentes en una unidad hospitalaria en un día.
- La distribución Chi-cuadrado se emplea para evaluar si existe relación entre el tipo de dieta y la presencia de hipertensión en una muestra de pacientes.
- La t de Student es común en ensayos clínicos para comparar la media de una variable (como niveles de hemoglobina) entre dos grupos con tamaños pequeños.
- La distribución F se utiliza en el análisis de varianza (ANOVA), por ejemplo, para comparar los niveles promedio de glucosa entre tres tipos de tratamientos antidiabéticos.

- La exponencial se adapta al análisis del tiempo hasta la ocurrencia de eventos, útil en análisis de supervivencia o tiempos hasta complicaciones postquirúrgicas.

Contraste de normalidad

Muchos métodos estadísticos paramétricos, como la prueba t de Student, análisis de varianza (ANOVA), regresión lineal o cálculo de intervalos de confianza, requieren que los datos sigan una distribución normal. Sin embargo, en la práctica, no podemos asumir esta condición sin verificarla. Por ello, uno de los pasos previos más importantes en el análisis estadístico es contrastar si la distribución de los datos observados es normal.

Este proceso se conoce como prueba de normalidad o contraste de normalidad, y puede realizarse utilizando herramientas gráficas y/o pruebas estadísticas específicas.

Pruebas estadísticas de normalidad

Tabla 3

Pruebas estadísticas para contrastar la normalidad de los datos

Propiedad	Aplicación principal	Tipo de variable
Shapiro-Wilk	Muestras pequeñas (n<30)	Alta potencia. Recomendada en análisis clínicos individuales.
Kolmogórov-Smirnov	Muestras grandes (n≥30)	Compara la distribución empírica con la normal teórica.
Anderson-Darling	Todo tamaño muestral	Da más peso a las colas de la distribución.
D'Agostino y Jarque-Bera	Evaluación de simetría y curtosis	Útil para detectar desviaciones específicas respecto a la normalidad.
Chi-cuadrado de bondad de ajuste	Datos agrupados	Requiere agrupación previa de los datos en clases.
Shapiro-Francia (EpiDat)	Implementada en EPIDAT	Variante del Shapiro-Wilk, recomendada en enseñanza y práctica médica.

Criterio de decisión:

- Se formula una hipótesis nula. H0: los datos siguen una distribución normal. H1: los datos no siguen una distribución normal.
- Si el p-valor obtenido es menor que el nivel de significancia (por ejemplo, 0,05), se rechaza H_0 y se concluye que la distribución no es normal.

Herramientas gráficas para evaluar la normalidad

Además de las pruebas estadísticas, los gráficos ofrecen una herramienta visual poderosa para evaluar el comportamiento de los datos. Son particularmente útiles en el análisis exploratorio.

Tabla 4

Gráficos usados en la evaluación de la normalidad de los datos

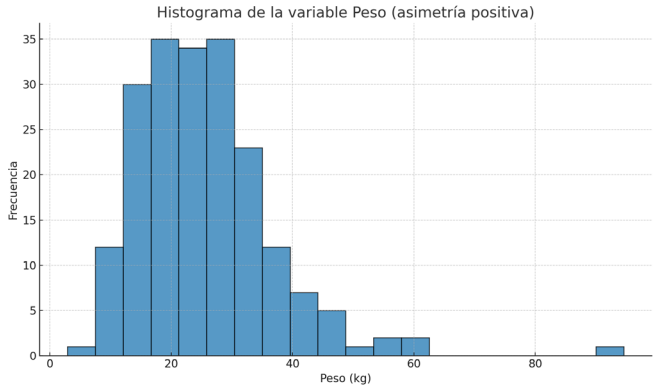
Propiedad	**Aplicación principal**
Histograma	Forma de campana, simétrica respecto a la media.
Diagrama de caja y bigotes	Mediana centrada, cuartiles simétricos, bigotes proporcionados.
Gráfico Q-Q (cuantiles)	Los puntos siguen una línea recta (valor observado ≈ valor teórico).
Gráfico P-P	Similar al Q-Q pero compara distribuciones acumuladas.

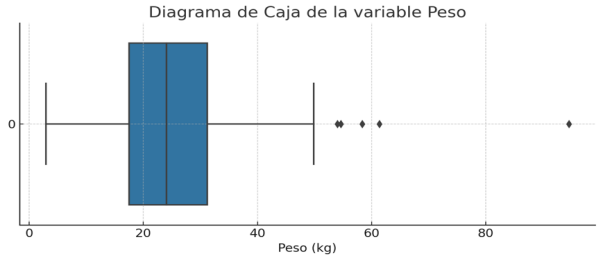
Ejemplo aplicado (EPIDAT 4.2):

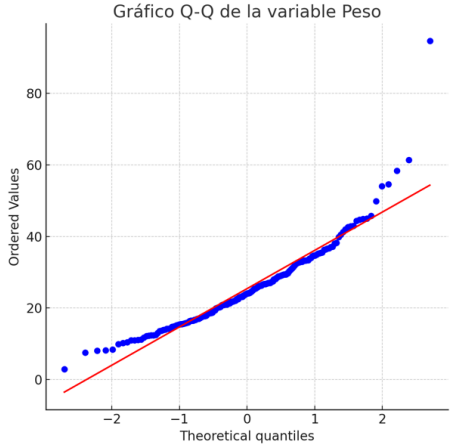
- Variable: peso de pacientes.
- Hipótesis:
    - H0: los datos siguen una distribución normal.
    - H1: los datos no siguen una distribución normal.
- Prueba utilizada: Shapiro-Francia:
    - Resultado: p-valor = <0.001 (valor aproximado)
- Conclusión: se rechaza H_0. El peso no se distribuye normalmente.

Figura 4

Gráficos de apoyo en EPIDAT







- El histograma muestra asimetría hacia la derecha (cola larga).
- El diagrama de cajas presenta un bigote superior alargado.
- El gráfico Q-Q no forma una línea recta.

Estos resultados combinados sugieren que los datos tienen una distribución asimétrica, por lo que no es apropiado aplicar métodos paramétricos sin transformarlos o sin optar por pruebas no paramétricas.

Teorema del Límite Central

El Teorema del Límite Central es uno de los principios fundamentales de la estadística inferencial. Este teorema establece que, cuando se extraen muestras aleatorias de tamaño suficientemente grande a partir de una población con media μ y desviación estándar σ, la distribución de las medias muestrales tiende a aproximarse a una distribución normal, independientemente de la forma que tenga la distribución original de la población.

Esto significa que, aun cuando los datos poblacionales no sigan una distribución normal, la distribución de las medias obtenidas a partir de múltiples muestras aleatorias será aproximadamente normal conforme aumenta el tamaño de la muestra. Este principio es clave para la aplicación de pruebas estadísticas paramétricas y para la construcción de intervalos de confianza.

La media de la distribución de las medias muestrales es igual a la media poblacional:

$$\mu_{\bar{x}} = \mu$$

Mientras que la desviación estándar de la distribución de las medias muestrales, conocida como error estándar de la media, se calcula como:

$$SD(\bar{X}) = \sigma / \sqrt{n}$$

El error estándar de la media refleja la variabilidad de las medias muestrales alrededor de la media poblacional y muestra que, a medida que aumenta el tamaño de la muestra, esta variabilidad disminuye. Es decir, muestras más grandes producen estimaciones más precisas de la media poblacional.

El Teorema del Límite Central permite comprender por qué muchas pruebas estadísticas asumen normalidad y constituye la base teórica para el uso de la distribución normal y de la distribución t de Student en la inferencia estadística, especialmente en estudios de salud y ciencias de la vida.

Condiciones para aplicar el TLC

- Muestras aleatorias independientes.
- Tamaño muestral suficientemente grande. Generalmente se acepta que:
  - »
  - » Si $n \geq 30$, es seguro aplicar el TLC.
  - » Si la distribución poblacional es muy asimétrica, podría requerirse un n mayor.
- La población puede tener cualquier forma de distribución: asimétrica, sesgada, bimodal, etc.

Importancia en la práctica médica

Este teorema justifica el uso de pruebas paramétricas, aun cuando los datos no son perfectamente normales, siempre que el tamaño muestral sea adecuado. Es particularmente útil en investigaciones clínicas y epidemiológicas, donde:

- Se realizan estimaciones sobre la media poblacional (por ejemplo, media de glucosa en sangre).
- Se calcula un intervalo de confianza para una proporción (por ejemplo, proporción de pacientes que responden al tratamiento).
- Se realizan pruebas de hipótesis para determinar diferencias significativas entre grupos.

Conclusiones y recomendaciones

La distribución de los datos es el punto de partida fundamental para aplicar correctamente cualquier análisis estadístico. Identificar si los datos siguen una distribución normal o no permite seleccionar las pruebas estadísticas más apropiadas y garantizar la validez de los resultados.

La distribución normal es una herramienta central en la estadística biomédica. Su forma simétrica y sus propiedades probabilísticas facilitan el análisis y la interpretación de fenómenos clínicos y epidemiológicos, como niveles de presión arterial, talla o colesterol.

La puntuación Z permite transformar cualquier valor a una escala estandarizada, facilitando la comparación entre sujetos y el cálculo de probabilidades. Esta herramienta es indispensable para la toma de decisiones clínicas y para la comparación de resultados entre estudios.

Existen muchas otras distribuciones que son útiles según el tipo de variable y el diseño del estudio. Las distribuciones binomial, de Poisson, Chi-cuadrado, t de Student y F de Snedecor son fundamentales en la estadística aplicada a la salud.

El contraste de normalidad es esencial para validar los supuestos de las pruebas paramétricas. Combinar herramientas gráficas y pruebas estadísticas proporciona una evaluación robusta de la normalidad de los datos.

El Teorema del Límite Central respalda el uso de la distribución normal en muestras grandes, incluso cuando los datos individuales no se distribuyen normalmente. Este teorema amplía el campo de aplicación de las pruebas paramétricas y fortalece la inferencia estadística.

Por lo tanto, se recomienda que, antes de aplicar cualquier prueba estadística, siempre se realice una exploración descriptiva de los datos, incluyendo gráficos (como histogramas y diagramas de caja) y estadísticos (como la media, mediana y desviación estándar).

No se recomienda asumir la normalidad sin verificarla mediante pruebas como Shapiro-Wilk o Kolmogorov-Smirnov, especialmente cuando se utilizarán métodos paramétricos sensibles a este supuesto. Estas pruebas se pueden calcular utilizando programas estadísticos como EPIDAT, SPSS o R, lo que facilita el análisis de la distribución y evita errores manuales. En caso de que los datos no sean normales, se debería considerar el uso de pruebas no paramétricas o técnicas de transformación de datos.

Se debe aprovechar el Teorema del Límite Central en el diseño de los estudios, planificando muestras grandes para ampliar las posibilidades analíticas y reducir los efectos de la distribución original de la variable.

Finalmente, se sugiere promover en los estudiantes y profesionales de la salud una comprensión conceptual y aplicada de las distribuciones de probabilidad, como parte del razonamiento crítico en investigación y en la toma de decisiones clínicas.

Referencias

Kirkwood, B. R., & Sterne, J. A. C. (2003). *Essential Medical Statistics* (2nd ed.). Malden: Blackwell Science.

Ministerio de Sanidad y Consumo de España. (2008). *Manual de Bioestadística para Ciencias de la Salud*. Madrid: Ministerio de Sanidad.

Organización Panamericana de la Salud. (2016). *EPIDAT 4.2: Programa para análisis epidemiológico de datos tabulados* [Internet]. OPS. Disponible en: https://www.sergas.es/Saude-publica/EPIDAT

Pagano, M., y Gauvreau, K. (2006). *Principios de Bioestadística (2.ª ed.)*. México: Cengage Learning.

Waines, D. W. (2009). *Bioestadística: base para el análisis de las ciencias de la salud* (7.ª ed.). México: Limusa.

Capítulo 15
Universo y muestra

Introducción

Todo estudio de investigación está dirigido a una población diana, la cual se debe conocer a detalle para una correcta elección de los sujetos de investigación. Todo lo que rodea la elección de la población del estudio requiere un profundo análisis estadístico para no caer en error de selección o de tamaño. Por esta razón, en este capítulo hablaremos a detalle que es un universo y sus tipos y, cómo a partir de este, debemos escoger a los individuos que van a formar nuestras investigaciones.

Desarrollo

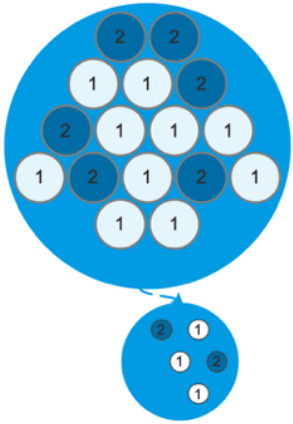
Universo y muestra

Para determinar la muestra es importante conocer algunos conceptos. Los elementos o individuos son los objetos o personas meta de la investigación. La población o universo es la colección completa de objetos o individuos meta de la investigación. El universo o población puede ser finito, cuando está conformado por un número limitado de elementos o individuos, e infinito si las unidades por las que está conformado son ilimitadas. Sin embargo, conocer el universo y las proporciones de las características que nos interesan en una población infinita o al menos muy grande no sería posible.

Por ello, es importante delimitar el universo a estudiar, definiendo las características que queremos estudiar en éste. La delimitación deberá realizarse en función del problema, objetivos, hipótesis, variables y tipo de estudio. Por ejemplo, estudiar el índice de masa corporal de todos los niños del mundo entre 5 y 6 años no sería posible, pues sería inviable tener una lista de todos los niños que cumplen con la condición; por esto, se deberá delimitar en espacio y en tiempo, entre otras características de interés para poder realizar el estudio.

Figura 1

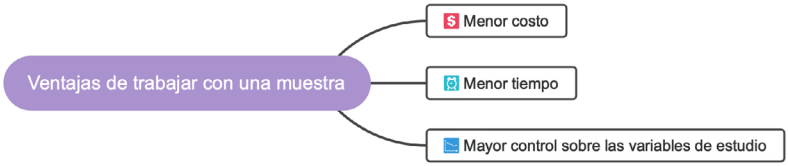
Universo y muestra



La muestra es un subconjunto de una población. Cualquier subconjunto de la población es una muestra. Por tanto, se pueden extraer infinito número de muestras de un mismo universo. La muestra tiene que ser independiente ya que el conocer el valor de algunos elementos de la muestra no debe ayudar a predecir los valores de otros. Trabajar con una muestra en lugar del universo tiene una serie de ventajas, como disminución de costos relacionados con la logística, menor tiempo y un mayor control sobre las variables.

Figura 2

Ventajas de trabajar con una muestra



Distribución de probabilidades y cálculo del tamaño de la muestra

Se debe considerar que, si hacemos un estudio con toda la población o universo, los resultados serán parámetros, mientras que al trabajar sobre una muestra los resultados serán estimaciones o estadísticos, los cuales son estimaciones de los parámetros de la población. En otras palabras, estimador de un parámetro o estadístico, , es cualquier valor que se exprese en función de una muestra aleatoria y su objetivo es aproximar el valor del parámetro; este no es un valor concreto sino una variable aleatoria pues depende de los valores de la muestra.

Sin embargo, es usual confundirlos. Supongamos que queremos determinar la intención de voto hacia dos candidatos A y B a través de una encuesta. Para ello, realizamos un muestreo para no trabajar con toda la población. Al usar una muestra y no el todo, hablamos de estadísticos y aceptamos un error determinado, por lo que lo correcto es presentar el resultado con sus márgenes de error. Si los candidatos tienen intenciones de voto muy similares, se podrá incurrir en un empate técnico, esto se da cuando los márgenes de error se superponen. Sin embargo, si tenemos ya los resultados reales de la votación obtenidos de las urnas, los resultados serán parámetros. Las diferencias son reales, si los resultados obtenidos por cada candidato son similares, se podrá decir que el candidato ganó con una diferencia mínima (real) y ya no hablar de empates técnicos, esto ya no aplica.

La inferencia estadística nos ayuda a hacer afirmaciones válidas acerca de la población o universo usando una muestra. La inferencia estadística se divide en estimación y prueba de hipótesis. Para realizar una buena estimación se debe tener una muestra adecuada considerando su tamaño y forma de colectarla.

La distribución de probabilidad de X relaciona el conjunto de los posibles valores que puede tomar X, con la probabilidad asociada a cada uno de estos valores. Estos se representan en una tabla o una función asociada de probabilidad o densidad. Una distribución de probabilidad o densidad si es continua depende de unos parámetros desconocidos θ , $f = (x; \theta_1, ..., \theta_k)$ Si tomamos varias muestras aleatorias simples de X, $(X_1, ..., X_n)$, tendremos una función para cada muestra. Podemos definir a la función de densidad conjunta considerando que se realiza un muestreo repetido (Barón, 1995):

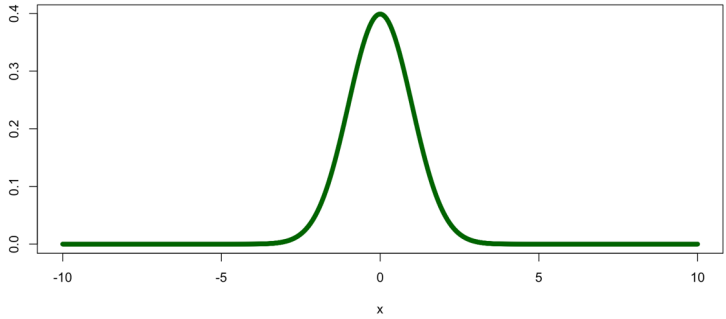
$$f_c(x_1, ..., x_n;\ \theta_1, ..., \theta_k) = (x_1; \theta_1, ..., \theta_k)(x_2; \theta_1, ..., \theta_k)(x_3; \theta_1, ..., \theta_k) ....(x_n; \theta_1, ..., \theta_k).$$

Se busca que el valor del estimador del parámetro de valor puntual obtenido de la muestra (), sea consistente (acerca su valor al del parámetro conforme el tamaño de la muestra crezca), que carezca de sesgo, eficiente (dispersión pequeña) y eficiente (que aproveche toda la información de la muestra). Los parámetros con los que más frecuentemente se quiere realizar inferencia son la media de la población (μ), desviación estándar (σ) y la proporción de aciertos (*p*) y sus correspondie ntes estimadores puntuales. Los estadísticos de los parámetros indicados anteriormente serían para la media $\bar{X}$, para la desviación estándar S y la proporción de aciertos $\hat{p}$.

En la Figura 3 podemos observar la distribución normal estándar o gaussiana de x con μ = 0 y σ = 1, esta es simétrica (coinciden media, mediana y moda) y puede ser representada como N(0, 1) y como se explicó anteriormente, podemos obtener de esta las puntuaciones Z o número de desviaciones estándar sobre la media.

Figura 3

Distribución de probabilidad normal estándar con μ = 0 y σ = 1, N (0, 1)



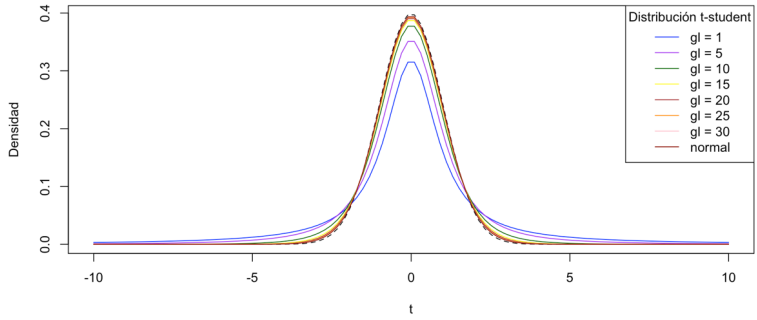
La ley de los números grandes nos dice que tenemos menor error a mayor tamaño de la muestra, mientras que muestras pequeñas tendrán errores grandes por lo que la curva de probabilidad será más aplanada, platocúrtica. Estas curvas son aproximaciones a la curva normal, se empieza a evidenciar en tamaños menores de 120 y es más crítico en tamaños menores de 30 elementos. Por ello, para muestras muy pequeñas no se aconseja usar la curva de desviación normal para obtener las probabilidades, sino una curva de distribución muestral aproximada a la normal. La distribución más usada

para muestras pequeñas es la distribución llamada t de Student, la cual es platocúrtica y simétrica.

De esta distribución se puede obtener puntuaciones estandarizadas, que son la distancia representada en el eje x que representa el número de errores estándar y está representada por la letra t minúscula. La puntuación crítica o estandarizada se obtiene de tablas o gráficas al nivel de significancia determinado y considerando sus grados de libertad. Los grados de libertad, gl, expresan el número de oportunidades en el muestreo para compensar las limitaciones, distorsiones y debilidades potenciales en los procedimientos estadísticos o el número de eventos muestrales independientes.

Figura 4

Comparación entre distribuciones t-Student para diferentes grados de libertad



Además del tamaño de la muestra es importante calcular el error estándar, que es la desviación estándar de una distribución normal, la cual es también una estimación, sus fórmulas se presentan en la Tabla 1. Recordemos que la varianza es el cuadrado de la desviación estándar.

Tabla 1

Fórmulas para el cálculo de los estadísticos para medias y proporciones

Entrada	Fórmulas	Nomenclatura considerando la variable X
1	$\underline{X} = \frac{X_1+X_2+\ldots+X_n}{n}$	p = éxito o aciertos. q = fracasos.
2	$\hat{p} = \frac{p}{p+q}$	$\hat{p}$ = Proporción de éxito o aciertos. $\hat{q}$ = proporción de fracasos.
3	$\hat{q} = 1 - \hat{p}$	σ_x = desviación estándar de la población $\sigma_{\underline{X}}$ = error estándar de una distribución
4	$\sigma_{\underline{X}} = \frac{\sigma_x}{\sqrt{n}}$	muestral de medias, σ_x conocida $S_{\underline{X}}$= error estándar de medias, σ_x
5	$S_{\underline{X}} = \frac{S_x}{\sqrt{n-1}}$	desconocida S_x = desviación estándar de la muestra
6	$\hat{S} = S_x\sqrt{n/(n-1)}$	$\hat{S}$ = estimador insesgado de la varianza $\hat{S}^2$ = cuasi varianza ponderada para la media
7	$\hat{S}^2 = \frac{(n_1-1)\hat{S}_1^2+(n_2-1)\hat{S}_2^2}{n_1+n_2-2}$	de dos poblaciones $\hat{S}_n^2$= varianza de cada muestra
8	$\sigma_p = \frac{\hat{p}\hat{q}}{\sqrt{n}}$	σ_p= error estándar de una distribución muestral de p. n = tamaño de la muestra

Intervalo de confianza

La estimación de un parámetro o estadístico es un valor puntual que es aleatorio debido a que depende de los elementos en la muestra tomada; por tanto, se podrá tener un valor puntual estimado del parámetro diferente por cada muestra. Para conocer cuan variable es el estimado, se deberá calcular el error o desviación estándar del estadístico, ver tabla 1. Para determinar la precisión de la estimación se debe calcular el intervalo de confianza, que es el

rango en el cuál definimos que el valor del parámetro se encontrará, considerando un nivel de significancia o error esperado α.

Cuando una muestra ha sido elegida mediante un muestreo aleatorio simple, no tiene sentido decir que el valor del parámetro se encuentra en el intervalo con probabilidad o nivel de confianza, 1- , pues sólo puede ocurrir que una vez fijada la muestra el parámetro esté o que no esté dentro del intervalo. Por lo que lo correcto es indicar que, si hubiésemos tomados muestras del mismo tamaño en una gran cantidad de ocasiones, hubiésemos acertado por lo menos en un nivel de confianza 100(1-) % de las ocasiones al decir que el parámetro estaba en el intervalo que cada muestra suministra (Barón, 1995). Para el cálculo del intervalo de confianza es importante el cálculo del término del error, el cual se calcula multiplicando el error estándar y la puntuación crítica al nivel de significancia determinado y considerando sus grados de libertad. De forma general podemos definir el intervalo de confianza (IC) de un parámetro un nivel de significancia , como:

$$IC(Límite\ inferior \leq \theta \geq Límite\ superior) = 1 - \alpha$$

En la Tabla 2 podemos encontrar las fórmulas para el cálculo de los intervalos de confianza.

Tabla 2

Cálculo de los intervalos de confianza (IC)

Entrada	Fórmulas	Nomenclatura considerando la variable X
1	$IC\ de\ \mu_x = \underline{X} \pm z_{(1-\alpha/2)}\sigma_{\underline{X}}$	$IC\ de\ \mu_x$ = IC de la media poblacional, σ_x conocida
2	$IC\ de\ \mu_x = \underline{X} \pm t_{(n-1,1-\alpha/2)}S_{\underline{X}}$	$IC\ de\ \mu_x$ = IC de la media poblacional, σ_x desconocida
3	$IC\ de\ \underline{X}_1 - \underline{X}_2 \mp t_{(n_1+n_2-2,\ 1-\frac{\alpha}{2})}\hat{S}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$ $IC\ de\ \underline{X}_1 - \underline{X}_2 \mp t_{(f,\ 1-\frac{\alpha}{2})}\sqrt{\frac{\hat{S}_1^2}{n_1}+\frac{\hat{S}_2^2}{n_2}}$	$IC\ de\ p$ = IC para una proporción $IC\ de\ p_1 - p_2$ = IC para dos proporciones $\hat{S}$ = desviación estándar desde la cuasi varianza para la media de dos poblaciones de varianza idéntica
4	$f = \frac{(\frac{\hat{S}_1^2}{n_1}+\frac{\hat{S}_2^2}{n_2})^2}{\frac{1}{n_1+1}(\frac{\hat{S}_1^2}{n_1})^2+\frac{1}{n_2+1}(\frac{\hat{S}_2^2}{n_2})^2} - 2$	α = nivel de significancia o error esperado $1-\alpha$ nivel de confianza gl = grados de libertad
5	$IC\ de\ p = \hat{p} \mp z_{1-\frac{\alpha}{2}}\sigma_p$	$z_{(1-\alpha/2)}$ = puntuación t crítica, tabla Z
6	$p_1 - p_2 \in \left(\hat{p}_1 - \hat{p}_2\right) \mp z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1}+\frac{\hat{p}_2\hat{q}_2}{n_2}}$	$t_{(n-1,1-\alpha/2)}$ = puntuación t crítica, tabla t student $t_{(n_1+n_2-2,\ 1-\frac{\alpha}{2})}$ = puntuación t crítica, tabla t Student

Cálculo del tamaño de la muestra

El tamaño de la muestra determina el error estándar. Como podemos ver en la taba 1, el tamaño de la muestra es el denominador de las ecuaciones

por lo que, a mayor tamaño de la muestra, menor error. Aunque lo ideal es elegir un tamaño de muestra acorde al grado de precisión que deseamos en los resultados, muchas veces la elección del tamaño de la muestra tiene otros criterios como lo económico, o lo práctico. En la tabla 3, tenemos algunas opciones para el cálculo de la muestra dependiendo del caso.

Tabla 3

Cálculo del tamaño de la muestra (n)

Caso	Fórmula	Nomenclatura
Población infinita para variables cuantitativas	$n \geq \frac{z_{\frac{1-\alpha}{2}}^2}{e^2} \hat{S}^2$	$\hat{S}^2$ = estimación puntual a priori de la varianza de la muestra $z_{\frac{1-\alpha}{2}}$ = tabla distribución z en función de α e = error propuesto s^2 = varianza t_α = tabla distribución t en función de . $\hat{p}$ = fracción de éxito $\hat{q}$ = fracción de fracaso N= tamaño de la población
Población infinita para variables cuantitativas	$n = \frac{s^2 t_\alpha^2}{e^2}$	
Población finita y conocida para variables cuantitativas	$n = \frac{N s^2 t_\alpha^2}{(N-1)e^2 + s^2 t_\alpha^2}$	
Población finita y conocida para variables cuantitativas	$n \geq \hat{p}\hat{q}\frac{z_{1-\frac{\alpha}{2}}^2}{e^2}$	
Población infinita para una proporción	$n = \frac{\hat{p}\hat{q}t_\alpha^2}{e^2}$	
Población finita y conocida para una proporción	$n = \frac{N\hat{p}\hat{q}t_\alpha^2}{(N-1)e^2 + pqt_\alpha^2}$	

Ejercicios resueltos

Se han medido las estaturas de los estudiantes varones de medicina de la ciudad de Cuenca, encontrándose una media de 165 cm y un error estándar de 5 cm obtenidos para una muestra de 61 estudiantes. Calcular el intervalo de confianza considerando un nivel de significancia del 10 %.

Para resolver vamos a usar la fórmula 2 de la Tabla 1.

$$IC\ de\ \mu_x = \underline{X} \pm t_{(n-1,1-\alpha/2)} S_{\underline{X}} = 165\ cm \pm t_{(61-1,1-0.1/2)} 5\ cm$$

Necesitamos obtener $t_{(60,\ 0.95)}$ de la tabla de t-student, dándonos un valor de 2.

$$IC\ de\ \mu_x = 165\ cm \pm 2\ *\ 5\ cm = 165\ cm \pm 10\ cm$$

Vuelva a calcular el IC considerando una distribución normal y considerando una desviación estándar de 5 cm.

Para resolver vamos a usar la fórmula 2 de la Tabla 1.

$$IC\ de\ \mu_x = \underline{X} \pm z_{(1-\alpha/2)} \sigma_{\underline{X}}$$

Necesitamos obtener $t_{(0.95)}$ de la tabla de z, dándonos un valor de 1,96

$$IC\ de\ \mu_x = 165\ cm \pm 1,96\ *\ 5\ cm = 165\ cm \pm 9,8\ cm$$

Si ajustamos la desviación estándar con la fórmula 6 de la tabla 1.

$$\hat{S} = S_x \sqrt{n/(n-1)} = 4,96$$

$$IC\ de\ \mu_x = 165\ cm \pm 1,96\ *\ 4,96\ cm = 165\ cm \pm 7,24\ cm$$

Calcular el intervalo de confianza de una proporción considerando una proporción de aciertos de 0,3 considerando n = 100, para un nivel de confianza del 95 %.

Se usará la fórmula 5 de la tabla 2

$$IC\ de\ p = \hat{p} \mp z_{1-\frac{\alpha}{2}} \sigma_p$$

$$\hat{p} = \frac{30}{100} = 0,3\ \hat{q} = 0,7$$

Para calcular la fórmula 8 de la Tabla 1 para el error estándar de una distribución muestral de una proporción.

$$\sigma_p = \sqrt{\frac{\hat{p}\hat{q}}{n}} = 0,046$$

Entonces el intervalo será, considerando que para este caso Z tienen un valor de 1,96.

$$p = 0.3 \pm 1.96 * 0,046 = 0.3 \pm 0,09$$

Se van a realizar las elecciones para elegir el alcalde de Cuenca, para este cargo se han postulado 2 candidatos. Se quiere estimar el resultado para esta elección mediante encuestas. Para ello, se deberá determinar el tamaño de la muestra considerando un nivel de significancia del 1 % y un error del 1 %.

Para calcular la fórmula 5 de la Tabla 3 para una población infinita para una proporción. Como no se conoce "p", se puede considerar que van al 50 %.

$$n = \frac{\hat{p}\hat{q}t_{\alpha}^{2}}{e^{2}} = 0,5 * 0,5 \frac{1,96^{2}}{0,01^{2}} = 9604$$

Referencias

Barón, F. (1995). *Bioestadística: Métodos y Aplicaciones.* Universidad de Málaga. https://www.bioestadistica.uma.es/baron/bioestadistica.pdf

Gutiérrez Pulido, H., y Vara Salazar, R. de la. (2008). *Análisis y diseño de experimentos.* McGraw-Hill.

Montgomery, D. (2016). *Diseño y Análisis de experimentos* (2da ed.). LIMUSA.

Navidi, W. (2006). *Estadística para ingenieros y científicos* (1ra ed.). McGraw-Hill. https://docs.google.com/viewer?a=v&pid=sites&srcid=ZGVmYXVsdGRvbWFpbnxvbWFyZGlhemhlcnxneDoyOGY2NDFlYmFkOTU5ZmI2

Pineda, E., De Alvarado, E., & De Canales, F. (1994). *Metodología de la Investigación* (2da ed.). Organización Panamericana de la Salud.

Ritchey, F. J. (2001). *Estadística para las ciencias sociales: El potencial de la imaginación estadística*. McGraw-Hill.

Capítulo 16
Muestreo

Introducción

Una vez que se ha determinado el tamaño muestral, el siguiente paso es definir cuáles son los individuos más idóneos para formar parte del estudio de investigación. La selección de estos individuos es crucial para obtener resultados válidos y confiables, y evitar sesgos o errores que puedan distorsionar los resultados. Existen diversos métodos diseñados estadísticamente con el fin de ayudar a escoger este grupo de personas, y la elección del método dependerá de los datos disponibles y de las posibilidades logísticas y financieras de los investigadores.

El objetivo de este capítulo es describir los diferentes tipos de muestreo, tanto los no probabilísticos o por conveniencia, como los probabilísticos, así como sus ventajas y desventajas, y explicar cómo se realiza cada uno de ellos.

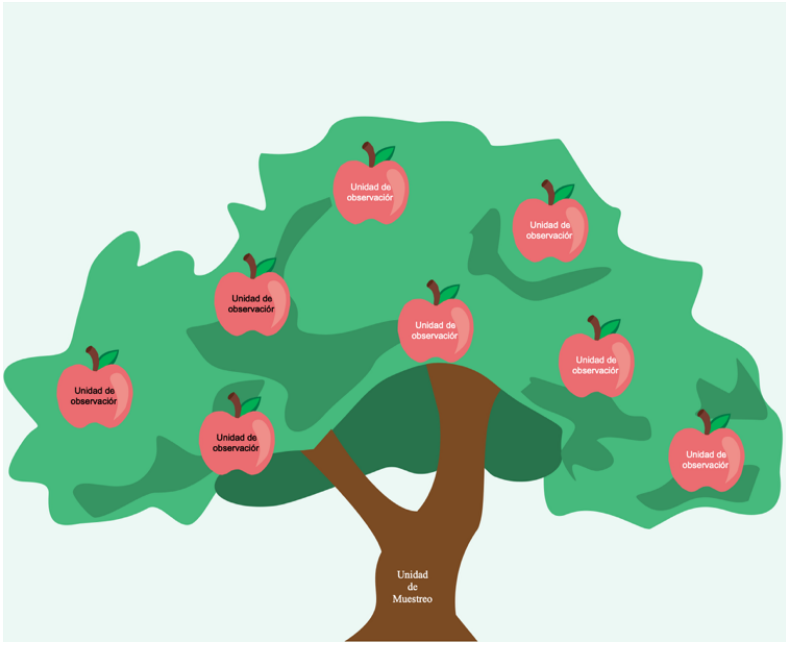
Desarrollo

Al tomar una muestra, es esencial identificar las unidades de observación, que son los elementos de la población o universo en los que se estudiarán las variables de interés. Además, se deben identificar las unidades de muestreo, que son los elementos que se seleccionan para conformar la muestra. Es importante señalar que las unidades de muestreo pueden o no coincidir con las unidades de observación.

Por ejemplo, si se desea conocer la talla promedio de los adolescentes, las unidades de observación serían los adolescentes (las personas a las que se les tomará la medida de la talla). Sin embargo, las unidades de muestreo podrían ser los colegios de una región, ya que se puede tomar una muestra de colegios y luego obtener los datos de los adolescentes dentro de esos colegios. En este caso, los colegios son las unidades de muestreo, y los adolescentes son las unidades de observación.

Figura 1

Unidades de observación y unidades de muestreo



Las muestras se consideran representativas cuando todas las características principales que nos interesan de la población están incluidas en la muestra, manteniendo las mismas proporciones que en el universo. Esto garantiza que la muestra refleje adecuadamente la diversidad de la población y permita extrapolar los resultados con mayor precisión.

Por ejemplo, supongamos que se desea estudiar el índice de masa corporal (IMC) en una escuela entre niños de 5 a 6 años, y se sabe que hay 1 000 estudiantes que cumplen con este requisito. De estos 1 000 estudiantes, el 60 % son mujeres y el 40 % son hombres. Para tomar una muestra representativa de 100 estudiantes, se debe asegurarse que las proporciones de género en la muestra sean las mismas que en la población total. Es decir, la muestra deberá incluir 60 niñas y 40 niños.

Tipos de muestras

No probabilística

El muestreo por conveniencia, también conocido como muestra no aleatoria, es un tipo de muestreo en el que no todos los individuos tienen las mismas posibilidades de ser seleccionados para el estudio. Este método implica utilizar los elementos que están disponibles en el momento del estudio, lo que puede incluir una selección sesgada de participantes.

Este tipo de muestreo se emplea generalmente en situaciones específicas en las que no hay mucha información sobre el tema o cuando se trata de enfermedades raras, que dificultan la inclusión de sujetos en el estudio. En estos casos, un experto en la materia es quien define, de acuerdo a sus criterios y experiencia, qué personas formarán parte de la muestra.

Por ejemplo, si se desea probar una nueva droga para tratar una enfermedad rara, con escasa prevalencia, el investigador podría decidir escoger a las primeras 30 personas que lleguen a su hospital de referencia con ese diagnóstico específico, y estas personas serían incluidas en el estudio. Aunque este método es práctico y rápido, puede no ser representativo de la población general y puede introducir sesgos en los resultados debido a la selección no aleatoria de los participantes.

Figura 2

Muestra por conveniencia

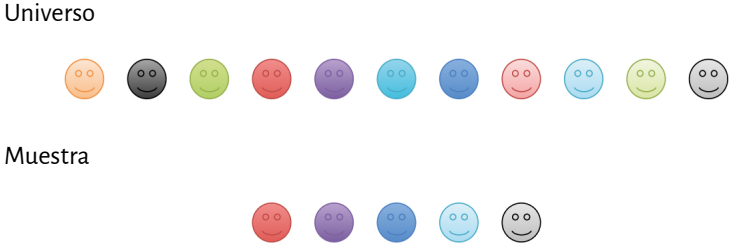
Probabilístico

En este tipo de muestreo todas las personas tienen una probabilidad conocida e igual de pertenecer a la muestra, por lo que formar parte de ella o no dependerá del azar. Existen diferentes tipos de muestreo probabilístico, entre los más importantes están: el azar simple, el sistemático, el estratificado, el por conglomerados y el procedimiento combinado.

La muestra aleatoria simple o azar simple es aquella en la que cada persona u objeto en la población tiene la misma oportunidad de ser seleccionada/o. Para tomar una muestra probabilística se deberá elegir al azar, por ejemplo, sortear entre todos los elementos o individuos de la población. Para poder usar este muestreo se debe tener un listado total de la población. Sus desventajas son que pueden determinar muestras dispersas y/o heterogéneas. En este tipo de muestreo se puede usar el método de la lotería o de la tabla de números, aunque actualmente la mayoría de investigadores, usan *softwares* estadísticos como el EPIDAT 4.2 para la elección de la muestra. Por ejemplo, se quiere realizar un estudio de investigación acerca de ansiedad en estudiantes de medicina, para esto los investigadores solicitan el listado total de los alumnos inscritos y de estos se eligen 50 al azar.

Figura 3

Ejemplo gráfico de selección aleatoria de 5 individuos en una población

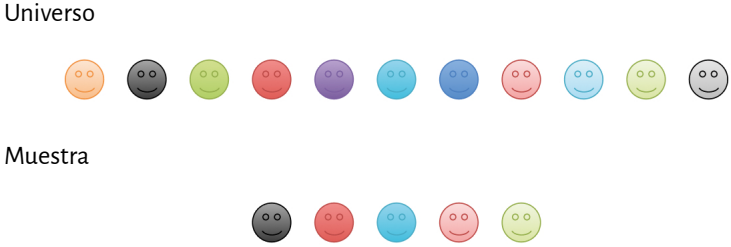


La muestra sistemática, también depende del azar, en ésta se crea un intervalo fijo que determina las personas que van a formar parte de la muestra, siendo el punto de partida elegido por el azar, aquí también necesitamos un

listado completo de la población, siendo sus desventajas la elección de muestras dispersas y/o heterogéneas. En este tipo de muestreo se puede usar el método de la lotería o de la tabla de números, para elegir el primer individuo y los demás se irán seleccionando según el intervalo requerido, de acuerdo al tamaño muestral. Actualmente la mayoría de investigadores, usan *softwares* estadísticos como el EPIDAT 4.2 para la elección de la muestra, con este tipo de muestreo. Por ejemplo, un investigador quiere realizar un estudio acerca de la ansiedad en medicina y para ello requiere una muestra de 50 individuos de un universo de 500 personas, por lo que elige al azar el primer individuo y a partir de este se elige cada 10 personas a un sujeto de investigación.

Figura 4

Ejemplo gráfico de selección por muestreo sistemático en una población de 10 personas, con un intervalo de 2



La muestra aleatoria estratificada se usa cuando queremos una muestra representativa aleatoria, pero conservando cierta o ciertas características de la población que son de interés; en ese caso, podemos identificar los estratos y sus proporciones en la población y luego sortear para completar la muestra. Para esto, necesitamos el listado total de la población y de las características de interés, para poder formar los estratos, esto nos dará la ventaja de tener muestras con mejor distribución y representación. Por ejemplo, un investigador quiere realizar un estudio acerca de la ansiedad en medicina, pero ha visto que existen más hombres que mujeres en la población, 70 % frente al 30 %, por lo que decide solicitar el listado de todos los alumnos inscritos y los estratifica en hombres y mujeres, tras lo que elige al azar a 70 hombres y 30 mujeres, para que conformen su muestra.

Figura 5

Ejemplo gráfico de selección por muestreo estratificado, eligiendo tres individuos de cada estrato en una población

Universo

Muestra

Las muestras por conglomerados, son aquellas en las cuales el investigador decide usar grupos naturales ya formados como sujetos de estudio, en este caso no se requiere un listado completo de la población puesto que no se escogen individuos, sino grupos completos ya formados, esto evita la dispersión y/o heterogeneidad de las muestras. Al no separar los individuos, se mantienen las características del grupo elegido en su totalidad. Por ejemplo, se quiere hacer un estudio de investigación de la ansiedad en los estudiantes de medicina, para esto se eligen a los alumnos de primero, tercero y cuarto año de carrera.

Figura 6

Ejemplo gráfico eligiendo un conglomerado de la población como muestra

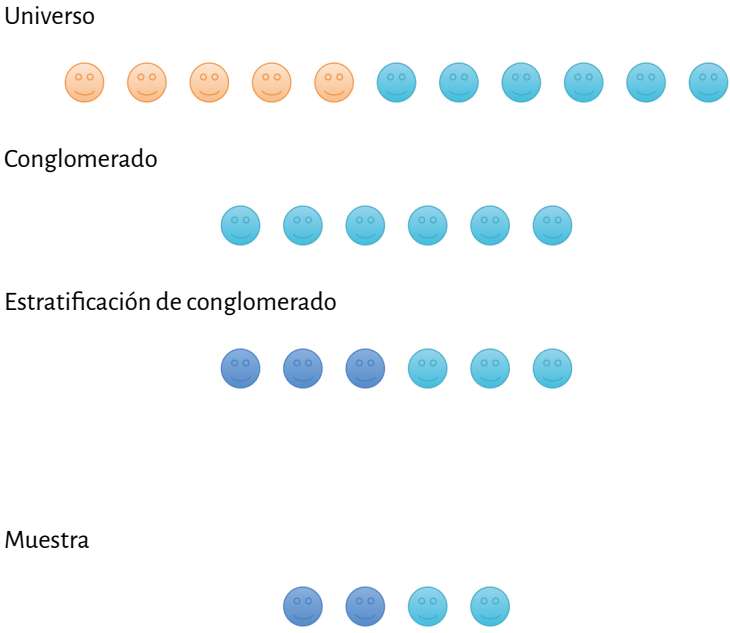
Universo

Muestra

Para finalizar, el procedimiento combinado es una fusión entre las muestras por conglomerados y las estratificadas, en este muestreo no se requiere de un listado completo de la población, pero sí de los conglomerados de estudio, una vez elegidos los conglomerados al azar, se realiza una estratificación dentro de ellos y por azar simple se elige un número representativo de cada estrato. Por ejemplo, se quiere hacer un estudio de investigación acerca de la ansiedad en medicina, para esto se eligen los alumnos de primero, tercero y cuarto año de carrera. A estos grupos se los estratifica en hombres y mujeres, tras lo que elige al azar a 10 hombres y 20 mujeres de cada grupo para que conformen la muestra.

Figura 7

Ejemplo gráfico de selección por muestreo por conglomerados y estratificación



Nota. Se elige un conglomerado, estratificándolo en A y B, y seleccionando dos personas al azar de cada estrato.

Conclusiones y recomendaciones

Como hemos visto, existen diferentes formas de seleccionar a los sujetos de investigación, y la elección del método dependerá de los datos disponibles. En el caso de enfermedades raras o estudios con pocos sujetos de investigación, se opta por muestreos no probabilísticos, dado que la población es difícil de acceder o está limitada. Para enfermedades más frecuentes, en las que se dispone de un listado completo de la población, se pueden emplear muestreos probabilísticos como el azar simple, sistemático o estratificado. En

situaciones donde no se dispone de la totalidad de los datos de la población, se pueden utilizar conglomerados naturales para la selección de la muestra.

Se recomienda, siempre que sea posible, realizar muestreos probabilísticos, ya que estos proporcionan un mayor nivel de confianza y reducen los sesgos de selección. Esto permite obtener resultados más robustos que pueden ser generalizados a la población completa.

Referencias

Barón, F. (1995).*Bioestadística: Métodos y Aplicaciones.* Universidad de Málaga. https://www.bioestadistica.uma.es/baron/bioestadistica.pdf

Camel Fayad, V. (1970).*Estadística médica y de salud pública* (10ª ed.). Universidad de los Andes.

Celis A, y Labrada V. (2014).*Bioestadística*(3ra ed.). Editorial El Manual Moderno.

Gutiérrez Pulido, H., y Vara Salazar, R. de la. (2008).*Análisis y diseño de experimentos.* McGraw-Hill.

Montgomery, D. (2016). *Diseño y Análisis de experimentos* (2da ed.). LIMUSA.

Navidi, W. (2006). *Estadística para ingenieros y científicos* (1ra ed.). McGraw-Hill.

Pineda, E., De Alvarado, E., y De Canales, F. (1994). *Metodología de la Investigación* (2da ed.). Organización Panamericana de la Salud.

Ritchey, F. J. (2001). *Estadística para las ciencias sociales:* El potencial de la imaginación estadística. McGraw-Hill.

Ruggiero, M., Bustamante, Y., y Mark, C. (2015). *Guía Teórico-Práctica de Bioestadística I.* U. Central Venezolana.

Biografía de los autores

Juan Chuchuca Pillajo, Mgt.

Médico titulado por la Universidad de Cuenca. Magíster en Salud Pública por la Universidad San Francisco de Quito.

Paola Delgado Andrade, Mgt.

Médica titulada por la Universidad de Cuenca. Magíster en Nutrición y Dietética por la Universidad de Cuenca. Magíster en Metodología de la Investigación en Ciencias de la Salud por la Universidad Internacional La Rioja.

Nathalie Pinos Vélez, MD, Ph. D.

Doctora en Medicina por la Universidad de Cuenca. Cirujana General y Cirujana de Tórax por la Universidad de Cuenca y el Ministerio de Sanidad de España, respectivamente. Además, posee un Máster en Investigación por la Universidad de Sevilla y un Doctorado en Investigación Clínica por la misma institución.

Ing. Verónica Pinos Vélez, Ph. D.

Ingeniera Química por la Universidad de Cuenca. Magíster en Gestión Tecnológica por la Universidad de Cuenca. Máster en Bioinformática y Bioestadística por la UOC y la Universidad de Barcelona. Doctora en Ingeniería Química, Ambiental y de Procesos por la Universitat Rovira i Virgili.

José Roldán Fernández, Mgt.

Médico titulado por la Universidad de Cuenca. Magíster en Investigación de la Salud por la Universidad de Cuenca.

Este libro fue publicado digitalmente en diciembre de 2025-enero de 2026
y pertenece al sello editorial UCuenca Press.

Cuenca - Ecuador

Bioestadística 1 es una obra imprescindible para estudiantes y profesionales de las ciencias médicas, porque proporciona una comprensión profunda y accesible de las herramientas estadísticas aplicadas a la investigación biomédica. El libro desglosa de manera clara los principios fundamentales de la estadística en salud, avanzando gradualmente hacia técnicas complejas de análisis. Su punto de vista práctico y bien estructurado permite comprender y aplicar eficazmente estos métodos, facilitando su integración en los contextos clínicos y epidemiológicos más desafiantes.

Con un enfoque especial en el análisis descriptivo gráfico y numérico, el texto aborda el manejo de variables cualitativas y cuantitativas, así como el análisis de conceptos clave como tasas, ajuste de tasas y la distribución de los datos. A través de ejemplos prácticos y gráficos intuitivos, esta obra permite a los profesionales de la salud interpretar y tomar decisiones informadas con base en datos, mejorando la calidad de la investigación y, en última instancia, la atención al paciente. Bioestadística 1 enseña a manejar números y a comprender su impacto en la vida real, desde una perspectiva más humana y rigurosa en el análisis y la toma de decisiones científicas.

UCUENCA

ISBN: 978-9978-14-656-9