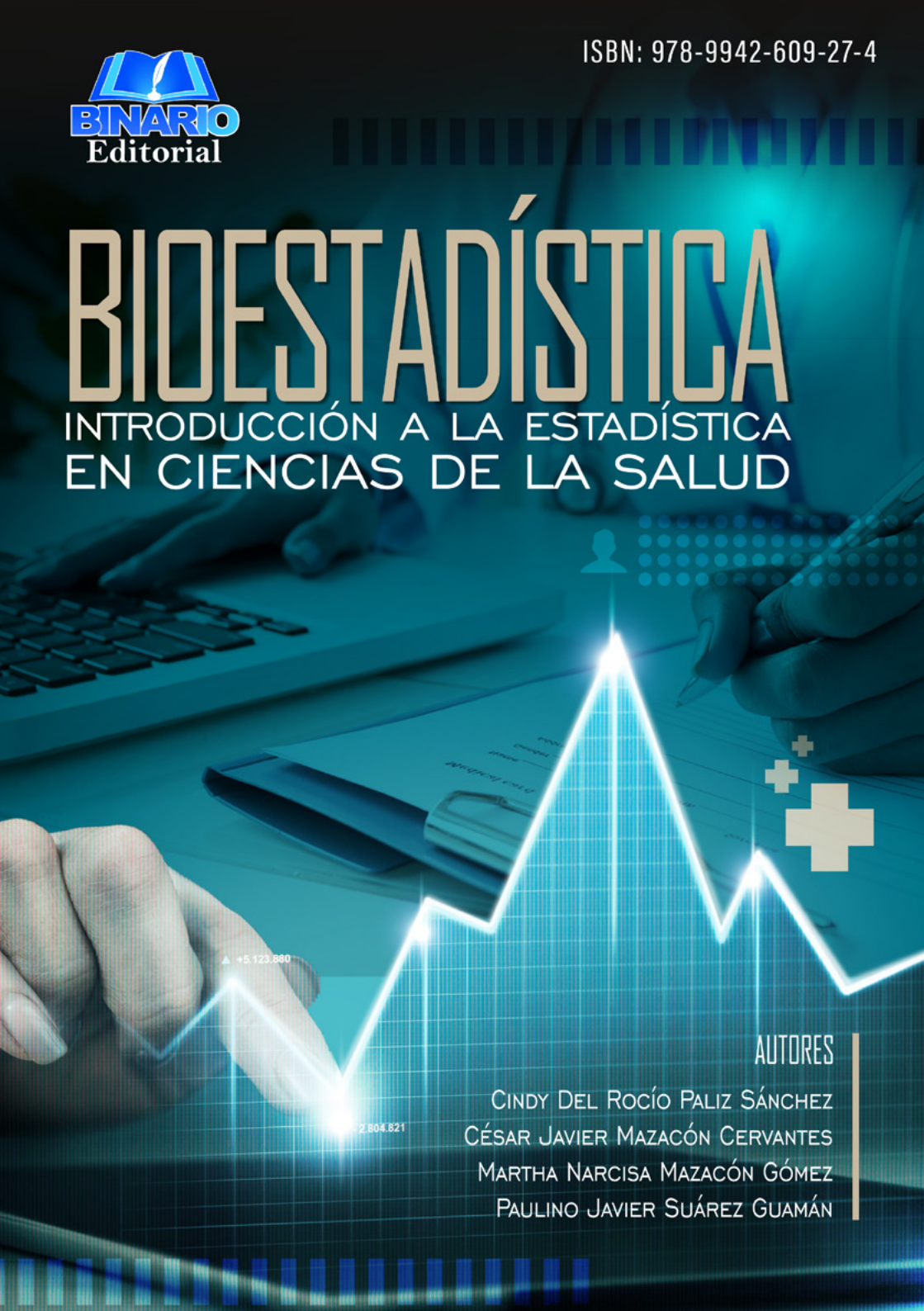
ISBN: 978-9942-609-27-4

BINARIO Editorial

BIOESTADÍSTICA

INTRODUCCIÓN A LA ESTADÍSTICA EN CIENCIAS DE LA SALUD



AUTORES

Cindy Del Rocío Paliz Sánchez

César Javier Mazacón Cervantes

Martha Narcisa Mazacón Gómez

Paulino Javier Suárez Guamán

Instituto de Investigaciones Transdisciplinarias Ecuador
BINARIO

BIOESTADÍSTICA

INTRODUCCIÓN A LA ESTADÍSTICA EN CIENCIAS DE LA SALUD

Autores

| Cindy Del Rocío Paliz Sánchez | César Javier Mazacón Cervantes

Martha Narcisa Mazacón Gómez | Paulino Javier Suárez Guamán

La revisión técnica de los documentos correspondió a especialistas expertos en el área.

ISBN: 978-9942-609-27-4

1era. Edición enero 2024
Edición con fines educativos no lucrativos
Hecho en Ecuador
Diseño y Tipografía: Greguis Reolón Ríos

BINARIO Editorial
SELLO 978 9942 8754
Instituto de Investigaciones Transdisciplinarias
Ecuador - BINARIO
Cel.: +593 96 766 4864

http.://www.binario.com.ec

EDITORIAL BINARIO

Mgs. María Gabriela Mancero Arias
Directora ejecutiva

Lcdo. Wilfrido Rosero Chávez
Gerente operaciones generales

Dra. Sherline Chirinos
Directora de publicaciones y revistas

Lcda. Greguis Reolón Ríos
Directora de marketing y RRSS

AUTORES

CINDY DEL ROCÍO PALIZ SÁNCHEZ
correo: cpalizs@utb.edu.ec
Economista, Magister en Docencia Universitaria
Universidad Técnica de Babahoyo

CESAR JAVIER MAZACÓN CERVANTES
jmazaconce@utb.edu.ec
Licenciado en Ciencias de la Educacion Especializacion Informática Educativa, Magister en Gerencia de Innovaciones Educativas
Universidad Técnica De Babahoyo

MARTHA NARCISA MAZACÓN GÓMEZ
mmazacon@utb.edu.ec
Ingeniera Comercial Magister en Administración de Empresas
Universidad Técnica de Babahoyo

PAULINO JAVIER SUÁREZ GUAMÁN
jmazaconce@utb.edu.ec
Ingeniero en Sistemas, Magister en Informática empresarial
Universidad Técnica de Babahoyo

PRÓLOGO

En las últimas décadas se ha producido una amplia y creciente utilización de los métodos estadísticos en todas las disciplinas de las ciencias de la salud, dando lugar, a la incorporación de la estadística en los planes de estudios de numerosas titulaciones, entre estas, la Medicina, Enfermería, Psicología, Biología, etcétera. De esta forma, la estadística habría pasado a desempeñar un papel básico en la formación de los profesionales de la salud.

Por otro lado, pocos son los artículos científicos que se publican sin la inclusión de estudios estadísticos. En realidad, no es posible pensar el desarrollo de la ciencia moderna sin el uso creciente de la estadística. De allí la relevancia y pertinencia académico de esta obra como recurso de apoyo a la formación y al aprendizaje de la estadística como parte fundamental del método científico.

La estadística proporciona al profesional de la salud (clínico, docente o investigador) un conjunto de herramientas de análisis que le permiten resumir y describir la información sobre determinadas características de interés de los individuos o elementos objeto de estudio, así como inferir o extraer conclusiones sobre una población a partir de los resultados obtenidos en una muestra.

En este libro se ofrece al lector un breve, pero académicamente riguroso, recorrido por los temas fundamentales de la ciencia estadística: medidas de tendencia central, diferentes técnicas de resumen y presentación de la información estadística, y los conceptos de mayor importancia de la estadística descriptiva, asimismo, se presentan los elementos básicos de probabilidad necesarios para la comprensión de los modelos teóricos de probabilidad y los conceptos, procedimientos y técnicas del análisis de

la estadística Inferencial: muestreo estadístico, estimación estadística y contraste de hipótesis.

La obra brida también información sobre los principales programas informáticos estadísticos disponibles en la Web (de acceso libre y privado) que posibilitan la realización de los cálculos más complejos.

Para facilitar la lectura y comprensión de la de la obra, los autores se han esforzado por no abusar de las fórmulas y el lenguaje matemático, haciéndose, más bien hincapié, en el desarrollo de las ideas, conceptos y procedimientos del análisis estadístico que permita una utilización adecuada, en el campo de la investigación o del ejercicio procesional, de estos recursos estadístico del conocimiento científico.

Cada uno de los capítulos de libro se acompaña de tablas, figuras y variados ejemplos, así como de referencias bibliográficas donde el lector interesado puede ampliar el alcance de sus conocimientos.

Al finalizar el estudio del contenido de este libro, los lectores interesados habrán adquirido las competencias siguientes:

- Entender la importancia de los conocimientos y métodos estadísticos en las diversas disciplinas de las ciencias de la salud.
- Desarrollar los conceptos básicos de la estadística descriptiva.
- Adquirir nociones básicas de probabilidad.
- Desarrollar los conceptos básicos de la estadística inferencial.
- Aplicar los métodos estadísticos como herramienta fundamental en investigación en ciencias de la salud.
- Analizar e interpretar los datos estadísticos referidos a estudios poblacionales.
- Si estos objetivos son alcanzados, los autores del libro estarían realmente satisfechos. ¡Que así sea...!

ÍNDICE DE CONTENIDO

ÍNDICE DE FIGURAS xviii

ÍNDICE DE TABLAS xix

INTRODUCCIÓN 21

CAPÍTULO I. DEFINICIONES Y APLICACIONES 25

Los fenómenos y su descripción 25

Concepto de estadística 26

¿Cómo surge la estadística? 26

Estadística. Definición 28

La estadística como ciencia 29

Investigaciones estadísticas 30

Planificación 30

Recogida de datos 31

Organización de los datos 32

Presentación de los datos 32

Análisis de los datos presentados 32

Interpretaciones y conclusiones 32

Rol de la estadística en la investigación biomédica 32

Aplicaciones 33

La bioestadística y su papel en la investigación en salud 33

Estadística descriptiva e inferencial 34

Estadística descriptiva 35

Estadística inferencial 35

Conceptos básicos 37

Población 37

Parámetros y estadígrafos 37
Variable 37
Muestra 38
Tipo de muestreo 39
Distribución 41
CAPÍTULO II. INTRODUCCIÓN AL ESTUDIO DE LA ESTADÍSTICA DESCRIPTIVA 45
Métodos numéricos. Medidas de tendencia central 45
Medidas de tendencia central en datos no agrupado 46
Media aritmética ($\ddot{x}$) 46
Mediana (Me) 47
Moda (Mo) 49
Media geométrica 50
Media armónica 52
Media ponderada 53
Media cuadrática 53
Medidas de tendencia central y de dispersión en datos agrupados 54
Media en datos agrupados 54
Mediana en datos agrupados 55
Moda en datos agrupados 55
Varianza en datos agrupados 55
Medidas de variabilidad 56
Rango 56
Desviación media 57
Varianza 59
Desviación estándar 60

Posición de un dato respecto de la media 61
Medidas de posición relativa 62
Percentiles 62
Cuartiles 64
Deciles 64
Estudio de las distribuciones de frecuencias 65
Distribución de frecuencia 65
Tablas de distribuciones de frecuencias 65
Distribuciones de frecuencias relativas 67
Distribuciones de frecuencias acumulativas 69
Distribuciones de frecuencias relativas acumulativas 69
Distribuciones agrupadas 69
Intervalo de clase 70
Forma de distribución 71
CAPÍTULO III. DATOS: TIPOS Y CARACTERÍSTICAS 75
Variable 75
Variables cualitativas 75
Variables cuantitativas 76
Escalas de medición 77
Escala nominal 77
Escala ordinal 77
Escala de intervalo 78
Escala de razones 78
Características de los conjuntos de datos 79
Tipos de datos 79
Datos categóricos o cualitativos 80

Datos numéricos 82
Representaciones gráficas 83
Gráficas o diagramas de barras 83
Histograma y polígonos de frecuencia 85
Polígono de frecuencia 86
Ojivas o polígonos de frecuencias acumuladas 87
Diagrama de barras 87
Gráficas de tallo y hojas 88
Tabulación de datos binarios o cruzados 90
Tabla de contingencia 91
Tablas de asociación: exposición–enfermedad 93
Tablas de frecuencias relativas 97
CAPÍTULO IV: ELEMENTOS BÁSICOS DE PROBABILIDAD 101
Introducción 101
Conceptos básicos de la teoría de la probabilidad 101
Experimento 101
Fenómenos determinísticos y aleatorios 102
Fenómeno o experimento aleatorio 102
Suceso aleatorio 104
Definición de probabilidad 105
Definición clásica 106
Definición frecuentista 108
Definición subjetiva de probabilidad 110
Teoría de conjuntos, probabilidad y operaciones con sucesos 110
Probabilidad de la unión de dos sucesos 111
Probabilidad de intersección de los sucesos: 112

Probabilidad de la diferencia de dos sucesos 113
Complemento de un conjunto (Suceso contrario) 113
Probabilidad condicional e independencia 115
Probabilidad condicionada 115
Eventos estadísticamente independientes 116
Tablas de contingencias y cálculo de probabilidad 117
Teorema de la probabilidad total 119
Enunciado del Teorema de Bayes 120
Axiomas y propiedades de la probabilidad 120
Aplicación de la teoría de probabilidad a la ciencia médica 121
Valor predictivo de pruebas diagnósticas: sensibilidad y especificidad 121
Prevalencia e incidencia 126
Valoración del riesgo 127
Riesgo relativo 127
Odds ratio o razón de productos cruzados 129
CAPÍTULO V: DISTRIBUCIONES TEÓRICASDE PROBABILIDADES .. 133
Introducción 133
Concepto de variable aleatoria 133
Distribuciones de probabilidad discretas 136
Distribución binomial 138
Distribución de Poisson 139
Distribuciones de probabilidad continuas 142
Distribución normal 143
La distribución normal y su función de densidad 145
La Distribución Normal (μ, σ2) 145

Distribución normal estandarizada N (0, 1) 147
Uso de tabla normal estándar 147
CAPÍTULO VI. MUESTREO Y ESTIMACIÓN ESTADÍSTICA 155
Introducción 155
Principales tipos de muestreo probabilístico 158
Muestreo aleatorio simple 158
Muestreo sistemático 160
Muestreo estratificado 162
Muestreo por conglomerados 164
Muestreo polietápico 168
Determinación del tamaño muestral 168
Tamaño muestral para la estimación de una media 170
El teorema central del límite 172
Estimación en el muestreo aleatorio simple 172
Algunos parámetros y sus estimadores puntuales 173
Estimación puntual de una media poblacional 174
Estimación mediante intervalos de confianza 175
Tablas de probabilidad para la distribución t de Student 178
Intervalo de confianza para un porcentaje poblacional 179
CAPÍTULO VII. CONTRASTE DE HIPÓTESIS 183
Estadística inferencial 183
Hipótesis estadística 183
Tipos de hipótesis 183
Contrastes de hipótesis 184
Contraste de hipótesis sobre la media de una población 185
Errores en un contraste de hipótesis 189

Hipótesis nula e hipótesis alternativa....190
Contraste y nivel de significación....191
Nivel de significación a posteriori o p-valor....192
Contrastes bilaterales y unilaterales....194
Potencia de un contraste....195
Contraste de hipótesis sobre una proporción....196
Comparación de dos medias poblacionales....198
Prueba t de comparación de medias para muestras independientes y varianzas iguales....198
Programas de cómputo y análisis estadístico....200
Procesadores de texto....201
Administradores de bases de datos....202
Hojas de cálculo....202
Programas para presentaciones....203
Programas estadísticos....203
SPSS (Statistical Package for Social Science)....205
Programa R....205
Open Epi....207
REFERENCIAS....209

ÍNDICE DE FIGURAS

Figura 1. Concepto de estadística 28
Figura 2. Tipo de variables 38
Figura 3. Conceptos estadísticos básicos 39
Figura 4. Gráfico de la distribución gaussiana (normal) 71
Figura 5. Gráfica de barras de las frecuencias relativas 84
Figura 6. Diagrama de sectores para la variable sexo..... 84
Figura 7. Histograma de frecuencia 85
Figura 8. Polígono de frecuencia 86
Figura 9. Polígono de frecuencias relativas acumuladas 87
Figura 10. Diagrama de barras horizontal..... 88
Figura 11. Gráfico de tallo y hoja 89
Figura 12. Diagrama de Venn..... 115
Figura 13. Función de densidad de una distribución normal con media μ y desviación típica σ..... 146
Figura 14. Distribución muestral del estadístico de contraste 187
Figura 15. Región critica de contraste y de aceptación de H_0..... 189
Figura 16. Situación del estadístico de contraste..... 189
Figura 17. Diferencia en la región crítica de contraste según el nivel de significación..... 192
Figura 18. Región que determina el valor de p 193
Figura 19. Valor p del contraste y nivel de significación α..... 194
Figura 20. Región crítica de contraste. Contraste de una proporción. 197
Figura 21. Región crítica de contraste. Comparación de medias. Varianzas iguales..... 200

ÍNDICE DE TABLAS

Tabla 1. Número de parto 54

Tabla 2. Cálculo de la desviación media y la varianza de la distribución A 58

Tabla 3. Cálculo de la desviación media y la varianza de la distribución B 58

Tabla 4. Distribución de frecuencias de las presiones sanguíneas de 144 adolescentes moderadamente obesos 63

Tabla 5. Distribución de mediciones de dolor percibido 68

Tabla 6. Tabla de contingencia para las variables tipo de ICC y estado de salud 92

Tabla 7. Resultados posible del lanzamiento de 2 dados 107

Tabla 8. Resultado del experiemnto de la vida util 109

Tabla 9. Tabla de contingencia que muestra las frecuencias de población 118

Tabla 10. Resultados de la enfermedad 125

Tabla 11. Evaluación de los factores de Riesgo (Diseño Prospectivo) 128

Tabla 12. Evaluación de los factores de Riesgo (Diseño Retrospectivo) 129

Tabla 13. Función de masa de probabilidad y función de distribución del número de supervivientes a los seis meses de cuatro pacientes con enfermedad catastrófica sometidos a tratamiento 137

Tabla 14. Distribución de probabilidad del número de muertes por cáncer de vesícula en periodos de 1 y 2 años en una población de 140.000 hombres 142

Tabla 15. Distribución del número de ancianos institucionalizados por residencia 166
Tabla 16. Tipos de error en un contraste de hipótesis 190

INTRODUCCIÓN

Aunque la estadística ha estado presente en la mayoría de las titulaciones universitarias de Ciencias de la Salud desde hace década, es desde su inclusión como materia básica en los planes de estudio de las titulaciones universitarias de grado incluidas en la rama de conocimiento de Ciencias de la Salud cuando ha venido a consolidar su presencia en todas las titulaciones de esta rama. De esta forma, la estadística ha pasado a desempeñar un papel básico en la formación de estos profesionales.

Esta obra puede ser útil como libro de aprendizaje y/o como libro de consulta, pues el material que se presenta en este documento parte de la reflexión y la experiencia de sus autores durante varios años y del contacto con alumnado en formación para ser futuros profesionales de las Ciencias de la Salud o profesionales en ejercicio con deseos de aumentar sus conocimientos en estadística.

El texto está organizado en siete capítulos. El capítulo I expone las definiciones y aplicaciones de la estadística, la investigación en estadística, el rol de la estadística en la investigación biomédica, se introduce algunos elementos de la estadística descriptiva e inferencial y conceptos básicos.

El capítulo II se refiere a la estadística descriptiva, es decir, a los métodos numéricos y medidas de tendencia central en datos agrupado y no agrupados, medidas de variabilidad, medidas de posición relativa y distribución de frecuencia. El capítulo III trata sobre los datos, tipos y características, iniciando con las variables, escala de medición, las características de los conjuntos de los datos, representaciones gráficas y tabulaciones de datos binarios o cruzados.

En el capítulo IV se aborda los elementos básicos de probabilidad, comenzando con los conceptos básicos de la teoría de la probabilidad, definición

de probabilidad, teoría de conjunto, probabilidad y operaciones de sucesos, probabilidad condicional e independencia, tablas de contingencias y cálculo de probabilidad, teorema de la probabilidad total, aplicación de la teoría de probabilidad a la ciencia médica y valoración del riesgo.

El capítulo V estudia las distribuciones teóricas de probabilidades, donde se introduce los conceptos de variables aleatorias, distribución de probabilidad discretas y continuas, distribución normal. En el capítulo VI se introduce el estudio de muestreo y estimación estadística donde se desarrolla los principales tipos de muestreo probabilístico, determinación del tamaño muestral, el teorema central del límite, estimación en el muestreo aleatorio simple, estimación mediante intervalos de confianza intervalo de confianza para un porcentaje poblacional. Finalmente, en el capítulo VII se enfoca en el contraste de hipótesis, para ello se describe la hipótesis estadística, nivel de significación a posteriori o p-valor, contraste de hipótesis sobre una proporción, comparación de dos medias poblacionales y por último los programas de cómputo y análisis estadístico.

Aunque la bibliografía en este campo es extensa, consideramos oportuno redactar un libro restringido a los contenidos específicos que se incluyen en un curso introductorio de estadística y que el lector cuente con los elementos y la base necesarios para comprender estos temas.

CAPÍTULO I.

DEFINICIONES Y APLICACIONES

CAPÍTULO I. DEFINICIONES Y APLICACIONES

Los fenómenos y su descripción

Las ciencias de la salud se incluyen dentro de las denominadas ciencias fácticas, puesto que en ellas el objeto de estudio es un conjunto de hechos o fenómenos implícitos en el concepto de salud. Al igual que en las demás ciencias que se incluyen bajo esa denominación, son de particular interés los hechos o fenómenos que varían al cambiar las circunstancias bajo las cuales se producen (1).

En tal sentido, los hechos de interés son definidos como variables, por lo cual para el trabajo en el campo científico se hace necesario identificarlas y diferenciarlas, a fin de poder analizarlas, evaluar las condiciones en que se producen y así intentar prever, prevenir o modificar su ocurrencia.

En ciencias de la salud, lo anterior significa la posibilidad de realizar acciones preventivas, diagnósticas o terapéuticas. Estas consideraciones se aplican en cualquiera de las actividades que se consideren dentro de las que realiza un profesional de la salud: asistenciales, de investigación o docentes.

Por otro lado, los objetivos más importantes relacionados con la estadística y que contribuyen al campo de la salud se tiene los siguientes:

- Permite comprender los fundamentos racionales en que se basan las decisiones en materia de diagnóstico, pronóstico y terapéutica
- Interpreta las pruebas de laboratorio y las observaciones y mediciones clínicas con un conocimiento de las variaciones fisiológicas y de las correspondientes al observador y a los instrumentos
- Proporciona el conocimiento y comprensión de la información acerca de la etiología y el pronóstico de las enfermedades, a fin de asesorar a los pacientes sobre la manera de evitar las enfermedades o limitar sus efectos

- Otorga un discernimiento de los problemas sanitarios para que eficientemente se apliquen los recursos disponibles para resolverlos

Las estadísticas de salud tienen uso individual y estadístico. El uso individual se refiere al uso de los registros médicos de cada persona que accede a los servicios de salud donde quedan registrados ciertas características del individuo y la historia de su enfermedad, muerte, tratamientos u otros servicios recibidos. Los registros médicos deben poseer los atributos de confidencialidad y custodia lo cual se regula por leyes y reglamentaciones con amparo legal. El uso estadístico se refiere al manejo de datos agregados donde se suman los datos relativos a cada individuo en modelos que compilan la información individual o de caso a caso con las periodicidades establecidas para los diferentes niveles del sistema nacional de salud (2).

Concepto de estadística

¿Cómo surge la estadística?

Desde los comienzos de las distintas civilizaciones han existido formas sencillas de estadística, pues ya se utilizaban representaciones gráficas y otros símbolos en pieles, rocas, palos de madera, huesos, para contar el número de personas, animales o ciertas cosas. Desde que surgen los primeros estados (Babilonios (3000 a.C.), egipcios (2200 a. C.), se han recogido datos sobre sus habitantes con el objetivo principal de recaudar impuestos y tributos, y reclutar a jóvenes para el ejército.

En el antiguo Israel la Biblia da referencias, en el libro de los Números, de los datos estadísticos obtenidos en dos recuentos de la población hebrea. El rey David por otra parte, ordenó a Joab, general del ejército hacer un censo de Israel con la finalidad de conocer el número de la población. También los chinos efectuaron censos hace más de cua-

renta siglos. Los griegos efectuaron censos periódicamente con fines tributarios, sociales (división de tierras) y militares (cálculo de recursos y hombres disponibles).

Durante los siglos XVII y XVIII los estados europeos comienzan a realizar censos de población y a recopilar de manera sistemática datos demográficos, sociales y económicos. Por tanto, hasta el siglo XIX, la Estadística es una ciencia descriptiva que utiliza medias y gráficos para sintetizar datos sociales y económicos. La necesidad de estimar cantidades desconocidas a partir de muestras va transformando paulatinamente la disciplina en una ciencia normativa para extraer conclusiones de los datos, prever la evolución de las variables y guiar la toma de decisiones en ambiente de incertidumbre; esta transformación es posible por la incorporación del concepto de probabilidad. En el siglo XIX, Gauss introduce la distribución normal como modelo de los errores de medida y Quetelet, padre de la sociología cuantitativa, utiliza una distribución para describir y estimar las características sociales medias de los miembros de una comunidad.

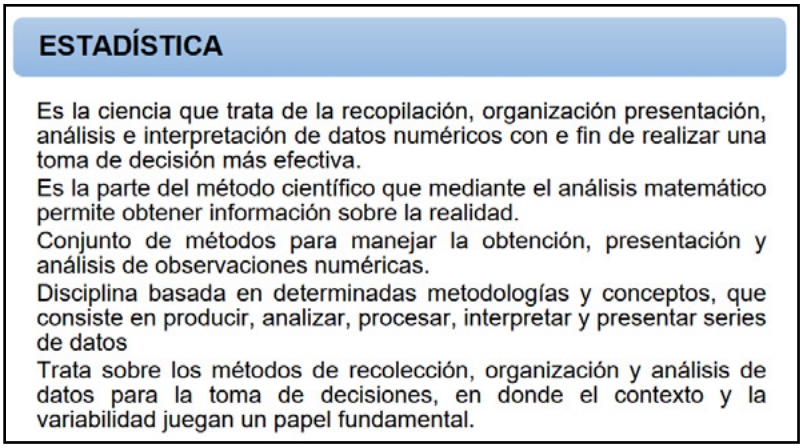
La Revolución Industrial dio un gran impulso a la necesidad de contar con información y datos permanentes, así que las estadísticas aplicables a fin de controlar la calidad de la producción, sumada después a la idea de experimentar y obtener productos nuevos, mejores y más baratos, tendrían reservado un lugar destacado en las fábricas y los comercios de la época.

La expansión de sus aplicaciones a todos los campos científicos ha dado lugar a disciplinas específicas como la Econometría, la Biometría o la Psicometría. En la actualidad, la estadística es probablemente una de las disciplinas científicas más utilizadas y estudiadas en todos los campos del conocimiento humano.

Estadística. Definición

El término estadística tiene dos acepciones fundamentales. Por un lado, la estadística como ciencia o método científico y, por otro lado, la estadística o estadísticas como conjunto o colecciones de datos. Este segundo concepto es muy usado hoy en día para referirse a resultados ya elaborados en un estudio en el que se empleó la estadística como método. Dado que la estadística es una disciplina muy amplia, existen diferentes definiciones de la misma según el enfoque en el que se plantee (figura 1).

Figura 1. Concepto de estadística



En el campo de la estadística se diferencian dos partes:

- Estadística descriptiva o deductiva: es la que se limita a la descripción de un conjunto de datos sin llegar a generalizar con respecto a un grupo mayor
- Estadística inferencial o inductiva: es la que se dedica al análisis y trata de llegar a conclusiones o generalizaciones acerca de un grupo mayor, basado en un grupo menor o muestra

La estadística como ciencia

La ciencia es una de las empresas más humanas y productivas que haya desarrollado el hombre. Si lo que caracteriza al ser humano es su excepcional inteligencia, la cual le ha dotado de lenguaje y le ha permitido servirse de él para crear una singular organización social, de insólita eficacia, para dominar la naturaleza, entonces la ciencia es el logro humano más perfecto y contundente, el cual señala la cúspide de los frutos de su intelecto, único en el Sistema Solar y tal vez en el universo mismo.

La ciencia basada en un proceso analítico y crítico produce el conocimiento que ha permitido una mejor comprensión de la realidad circundante. Asimismo, ha facultado al hombre para penetrar en los secretos más profundos del mundo, incluido el ser del hombre mismo. La ciencia ha facilitado el desarrollo de teorías que exponen la realidad, con base en un examen de la relación entre los intentos de explicación teórica, evidencia empírica y congruencia lógica, tanto interna a la explicación como en lo relativo a otras teorías con las que tienen vínculos conceptuales (3).

Esto ha implicado que el científico pruebe sus teorías confrontándolas con la evidencia existente que, con el objeto de evaluar la teoría de que se trata, se acumula con procedimientos rigurosos. Asimismo, el científico está a la caza de inconsistencias internas en la lógica de las explicaciones, así como de las contradicciones entre las diversas teorías vinculadas.

Ahora bien, la estadística, como ya se ha indicado, es la ciencia que se encarga de recoger, organizar e interpretar los datos. Es la ciencia de los datos. Es una rama de las matemáticas que se dedica a entender los fenómenos que tienen un cierto grado de azar.

En la ciencia se enfrenta el problema de que los fenómenos son multicausales y existe una diversidad de aspectos de los que sólo se tiene un grado de control relativo. Frente a esta problemática, resulta útil

emplear un método que permita lidiar con datos con una cierta dosis de incertidumbre. En realidad, la estadística es un instrumento muy valioso para organizar la información científica y para tomar decisiones acerca de ella, sería imposible concebir la investigación científica moderna sin dicha estadística.

El objetivo fundamental de la estadística es obtener conclusiones de la investigación empírica usando modelos matemáticos. A partir de los datos reales se construye un modelo que se confronta con estos datos por medio de la estadística. Esta proporciona los métodos de evaluación de las discrepancias entre ambos. Por eso es necesaria para toda ciencia que requiere análisis de datos y diseño de experimentos.

Investigaciones estadísticas

Las actividades que involucra una investigación estadística pueden clasificarse con arreglo a diversos criterios. Uno de ellos se relaciona con el orden cronológico en que deben ser realizadas. En este sentido, puede hablarse de tres grandes etapas o fases:

Planificación

El éxito de una investigación estadística depende en gran parte de la planeación que de ella se haga y se basa fundamentalmente en los siguientes aspectos:

- Problema: un problema que debe ser refinado y planteado en términos estadísticos. Enseguida, se desarrolla una estrategia para determinar la variable a investigar, qué datos proporcionarán información sobre dicha variable, de qué forma obtenerlos o generarlos, y diseñar un plan para su análisis
- Objetivos de la investigación: Los objetivos refieren al propósito, al por qué de la propuesta de investigación. El propósito está relacio-

nado con ciertas hipótesis, ciertas necesidades de información en el marco de una teoría.

- Universo, unidad a investigar y unidad de observación: El universo refiere a la población que se desea investigar. La unidad a investigar es cada individuo del universo. La unidad de observación puede o no coincidir con la unidad a investigar. La unidad de observación es a la que se dirige el investigador para indagar por la unidad a investigar.
- Procedimientos de recolección: Los procedimientos de recolección más comunes son:

– Censo
– Muestreo
– Experimentación

El análisis puede hacerse a un nivel exploratorio para generar una hipótesis, o puede hacerse con la finalidad de hacer inferencias sobre la población de la cual se obtuvieron los datos.

Recogida de datos

Es un término general que puede significar, por ejemplo, la recogida de datos de un experimentador midiendo con un instrumento o de un entrevistador preguntando a la gente sobre sus opiniones.

Para fines estadísticos, los datos se clasifican como internos y externos. Los datos obtenidos de los propios archivos son datos internos. Sin embargo, en muchas ocasiones es necesario establecer comparaciones con datos de la misma índole, pero referidos a una escala de mayor magnitud o simplemente es necesario obtener la información de una fuente diferente a los propios archivos. Estos datos exógenos se denominan datos externos.

Organización de los datos

Al recolectarse la información, esta debe ser organizada, esto es, que los datos antes de ser totalizados y utilizados para un análisis deben ser sometidos a un proceso de crítica, es decir a un examen crítico y severo con el fin de detectar si los datos son: exactos, completos, precisos y representativos

Presentación de los datos

Es este el cuarto paso de una investigación estadística. Los datos pueden presentarse para los lectores potenciales, mediante enunciados textuales, cuadros, tablas, o gráficos. Estas alternativas de presentación ayudan al lector a comprender de una manera ágil, amena, resumida y comprensible la información resultante de la investigación.

Análisis de los datos presentados

Después de que los datos son recolectados, organizados y presentados de forma comprensible a través de los cuadros, gráficos y enunciados; la información debe ser analizada, proceso éste, que involucra una serie de operaciones matemáticas.

Interpretaciones y conclusiones

La interpretación de la información es un campo que compete a personas especializadas en el tema que es materia de investigación.

Rol de la estadística en la investigación biomédica

La estadística aplicada a las ciencias biológicas dentro de las cuales se encuentran todas las ciencias de la salud, se denomina Bioestadística.

El término bioestadística ha sido ampliamente definido por diferentes autores como la rama de la estadística aplicada que estudia la utilización de métodos estadísticos en problemas médicos y biológicos (4).

La bioestadística se ocupa entonces de la recolección, clasificación, análisis y presentación de los datos, a partir del uso de métodos estadísticos en el campo de las ciencias biológicas y de la salud cuya finalidad es la toma de decisiones en esta área.

Aplicaciones

La bioestadística tiene múltiples aplicaciones, dentro de las cuales se encuentran:

- Favorece la toma de decisiones en el campo de la salud y áreas biológicas
- Facilita la predicción de eventos, a través del empleo de métodos estadísticos
- Es empleada en el desarrollo de pruebas en nuevos fármacos
- Permite el entendimiento de enfermedades como modos de propagación
- Permite caracterizar individuos que presentan un determinado evento en salud.
- Sirve en la evaluación de programas sanitarios y políticas públicas
- Se emplea en el área de la demografía para establecer relaciones entre muertes y nacimientos o su relación con diferentes características etéreas, sexo e incluso condición socioeconómica.

La bioestadística y su papel en la investigación en salud

El análisis y las técnicas estadísticas son un componente esencial en toda investigación biomédica, y la utilización de las técnicas estadísticas ha evolucionado considerablemente en los últimos años en las áreas de la investigación de ciencias de la salud. No hay duda de que tanto la actividad investigadora como los profesionales de la salud necesitan métodos estadísticos para el análisis de sus observaciones debido al crecimiento

incesantemente de los mismos. El empleo de técnicas estadísticas más específicas en investigación ha ido en aumento en las últimas décadas, motivado por la inclusión de la bioestadística en el currículo de los profesionales de la salud y por la inclusión de perfiles expertos en metodología en los equipos de investigación. Los análisis estadísticos empleados en un estudio dependen en gran medida del tipo de estudio, del objetivo que se pretende abordar y del tamaño de la muestra, así como del grado de conocimiento por parte de los investigadores de las técnicas estadísticas y del software para su implementación.

Es por ello que la estadística juega un papel fundamental en la investigación en ciencias de la salud, y a través de un equipo multidisciplinar que engloba a profesionales del ámbito sanitario, académico y perfiles expertos en metodología estadística se obtienen investigaciones de mayor calidad. El estadístico se encarga de gestionar y monitorear el proceso de recolección, análisis, difusión y uso de la información en salud, así como, aplicar las tecnologías de la información y comunicación, como instrumentos de soporte en el proceso automatizado de los sistemas de que permitan la generación de la información ágil, consistente y oportuna para la toma de decisiones en la planeación, operación, monitoreo y evaluación de los servicios de salud.

También es el responsable de obtener información confiable que permita contar con los indicadores trazadores para observar las tendencias de coberturas alcanzadas a nivel institucional. Las estadísticas de salud son todos aquellos datos numéricos debidamente capturados, validados, elaborados analizados e interpretados que se requieren para las acciones de salud.

Estadística descriptiva e inferencial

Como ya se indicó, la estadística se ocupa del procesamiento numérico de datos. Esta disciplina incluye dos grandes capítulos en función del objetivo

final de su aplicación. En uno de esos capítulos, las técnicas estadísticas se utilizan para resumir los datos obtenidos en un conjunto de situaciones que tienen algo en común. Por ejemplo, para resumir el resultado obtenido en un grupo de individuos con una determinada enfermedad y que fueron sometidos a un tratamiento específico, o ante la presencia de casos de una determinada condición en los habitantes de una región geográfica específica.

Estadística descriptiva

La aplicación del tratamiento estadístico tiene dos fases fundamentales:

- Organización y análisis inicial de los datos recogidos.
- Extracción de conclusiones válidas y toma de decisiones razonables a partir de ellos.

Los objetivos de la estadística descriptiva son los que se abordan en la primera de estas fases. Es decir, su misión es ordenar, describir y sintetizar la información recogida. En este proceso será necesario establecer medidas cuantitativas que reduzcan a un número manejable de parámetros el conjunto (en general grande) de datos obtenidos.

La realización de graficas (visualización de los datos en diagramas) también forma parte de la estadística descriptiva dado que proporciona una manera visual directa de organizar la información. La finalidad de la estadística descriptiva no es, entonces, extraer conclusiones generales sobre el fenómeno que ha producido los datos bajo estudio, sino solamente su descripción.

Las técnicas que se utilizan para obtener una valoración numérica de la manifestación de una variable dentro de un conjunto de individuos están dentro de lo que se denomina estadística descriptiva

Estadística inferencial

Es la rama estadística que se ocupa de los procesos de estimación (puntual y por intervalos), análisis y pruebas hipótesis. La finalidad de la estadística

inferencial es llegar a conclusiones que brinden una adecuada base científica para la toma de decisiones, considerando la información muestral recolectada. Las afirmaciones o resultados que se obtienen son acerca de la población de la cual proviene el conjunto de datos analizado.

En otras palabras, la estadística inferencial se ocupa del análisis, interpretación de los resultados y de las conclusiones a las que se puede llegar a partir de la información obtenida de una muestra con el fin de extender sus resultados a la población bajo estudio.

> Las técnicas de lo que se conoce como estadística inferencial permiten, mediante el procesamiento numérico de los datos registrados en una muestra, realizar inferencias sobre la forma que asume la variable de interés en la población respectiva

La generalización de las conclusiones obtenidas en una muestra a toda la población está sujeta a riesgo por cuanto los elementos de la muestra son obtenidos mediante un muestreo probabilístico. La estadística inferencial provee los procedimientos para efectuar la inferencia inductiva y medir la incertidumbre de las conclusiones que se van a generalizar. Los problemas más importantes en este proceso son:

- Estimación puntual: Es la estimación del valor del parámetro por medio de un único valor obtenido mediante el cálculo o evaluación de un estimado para una muestra específica
- Estimación por intervalos: Es la estimación del valor de un parámetro mediante un conjunto de valores contenidos en un intervalo. Para la obtención de intervalos de confianza se debe considerar el coeficiente de confianza que es la probabilidad de que el intervalo contenga al parámetro poblacional
- Prueba de Hipótesis: Es el procedimiento estadístico de comprobación de una afirmación y se realiza a través de las observaciones de una muestra aleatoria.

El objetivo de la inferencia estadística es hacer inferencias acerca de una población basada en la información contenida en una muestra. Ahora considerando que las poblaciones están caracterizadas por medidas descriptivas numéricas llamadas parámetros., a la inferencia estadística le corresponde hacer inferencias acerca de los parámetros poblacionales.

Conceptos básicos

Población

La población es la totalidad de sujetos de una condición que se está observando, es el conjunto de todos los individuos u objetos (de aquí en adelante sólo llamados individuos), cuyas características comunes se han de analizar, que es el objeto del estudio. Esta definición incluye, por ejemplo, a todos los sucesos en que podría concretarse un fenómeno o experimento cualesquiera. Una población puede ser finita o infinita.

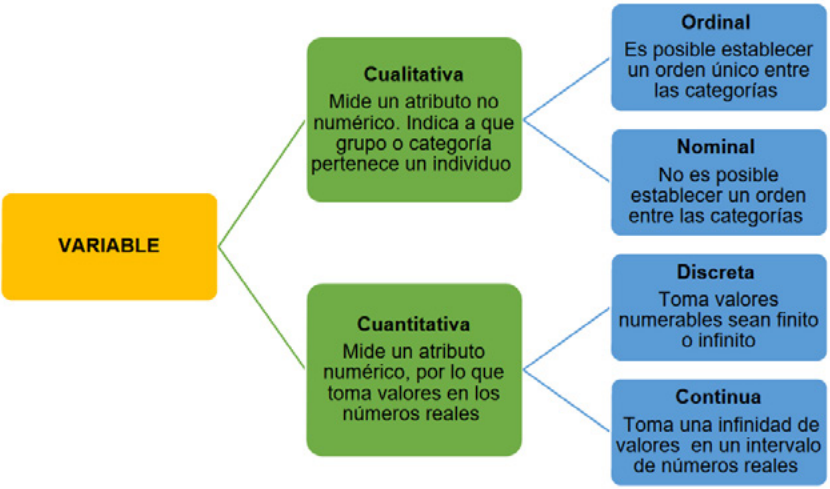
Parámetros y estadígrafos

Los parámetros son medidas cuantitativas que describen una característica de la población, entre ellas están: media aritmética, varianza y coeficiente de variación. En una población se presentan muchas características y, en consecuencia, tendrá varios parámetros. Los estadígrafos o estadísticas son medidas cuantitativas que describen una característica de la muestra y se consideran estimadores para la población

Variable

Es una característica de la población que es común a todos los individuos. Puede tomar valores diferentes para distintos individuos. Dentro de las variables se puede establecer la siguiente clasificación (figura 2).

Figura 2. Tipo de variables



Muestra

Una muestra es un subconjunto de la población; seleccionado al azar (esto es lo ideal), donde todos los miembros de la población tienen la misma probabilidad de formar parte de ella. En la práctica, este subconjunto es el que realmente se analiza para obtener información sobre la población. Se pueden distinguir dos tipos de muestras:

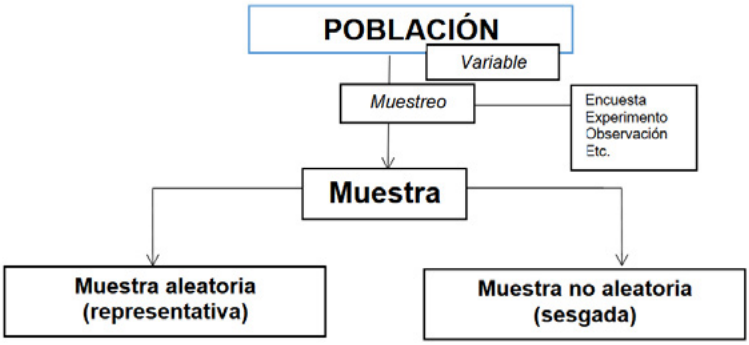
- Muestra aleatoria (representativa): la cual refleja en la medida de lo posible las características de la población de la cual proviene
- Muestra no aleatoria (sesgada): en la cual conjuntos importantes de la población no están representados.

Además de la aleatoriedad, la representatividad de la muestra también depende de su tamaño. En este sentido, si la variabilidad de la población es grande, se requiere de una muestra grande; si la variabilidad es poca, con una muestra pequeña basta.

El proceso mediante el cual se obtiene una muestra se conoce como muestreo, en el cual se pueden utilizar diversos los instrumentos para la recolección de datos: encuestas, experimentos, la simple observación, etc. Incluso, hay bases de datos proporcionadas por instituciones nacionales e internacionales que pueden ser consultadas en la red.

Esquemáticamente, los conceptos mencionados se pueden representar de la siguiente manera (figura 3).

Figura 3. Conceptos estadísticos básicos



Tipo de muestreo

Una muestra puede ser obtenida de dos tipos:

- Probabilística: Las técnicas de muestreo probabilísticas, permiten conocer la probabilidad que cada individuo a estudio tiene de ser incluido en la muestra a través de una selección al azar.
    - Aleatorio simple: Garantiza que todos los individuos que componen la población blanca tienen la misma oportunidad de ser incluidos en la muestra. Esto significa que la probabilidad de selección de un sujeto a estudio x es

independiente de la probabilidad que tienen el resto de los sujetos que integran forman parte de la población blanco.

- Muestreo aleatorio estratificado: En este método se divide a la población en estratos o subgrupos menores, parecidos internamente respecto a una característica, pero heterogéneos entre ellos, diferenciándolos por una variable que resulte de interés para la investigación.
- Aleatorio sistemático: El muestreo sistemático es un tipo de muestreo probabilístico que requiere tener un control preciso del marco muestral de individuos seleccionables. Para este tipo de muestreo se debe conocer la población y de igual forma se deben numerar todos los elementos.
- Muestreo por conglomerados El muestreo aleatorio por conglomerados es una técnica utilizada cuando no es posible obtener una lista completa de todos los elementos de una población. Consiste en dividir la población en grupos o conglomerados homogéneos, seleccionar aleatoriamente algunos de estos conglomerados y luego tomar una muestra aleatoria de cada uno de ellos. Esta técnica se emplea en situaciones donde la población es grande y dispersa, y los conglomerados representan unidades naturales o artificiales.

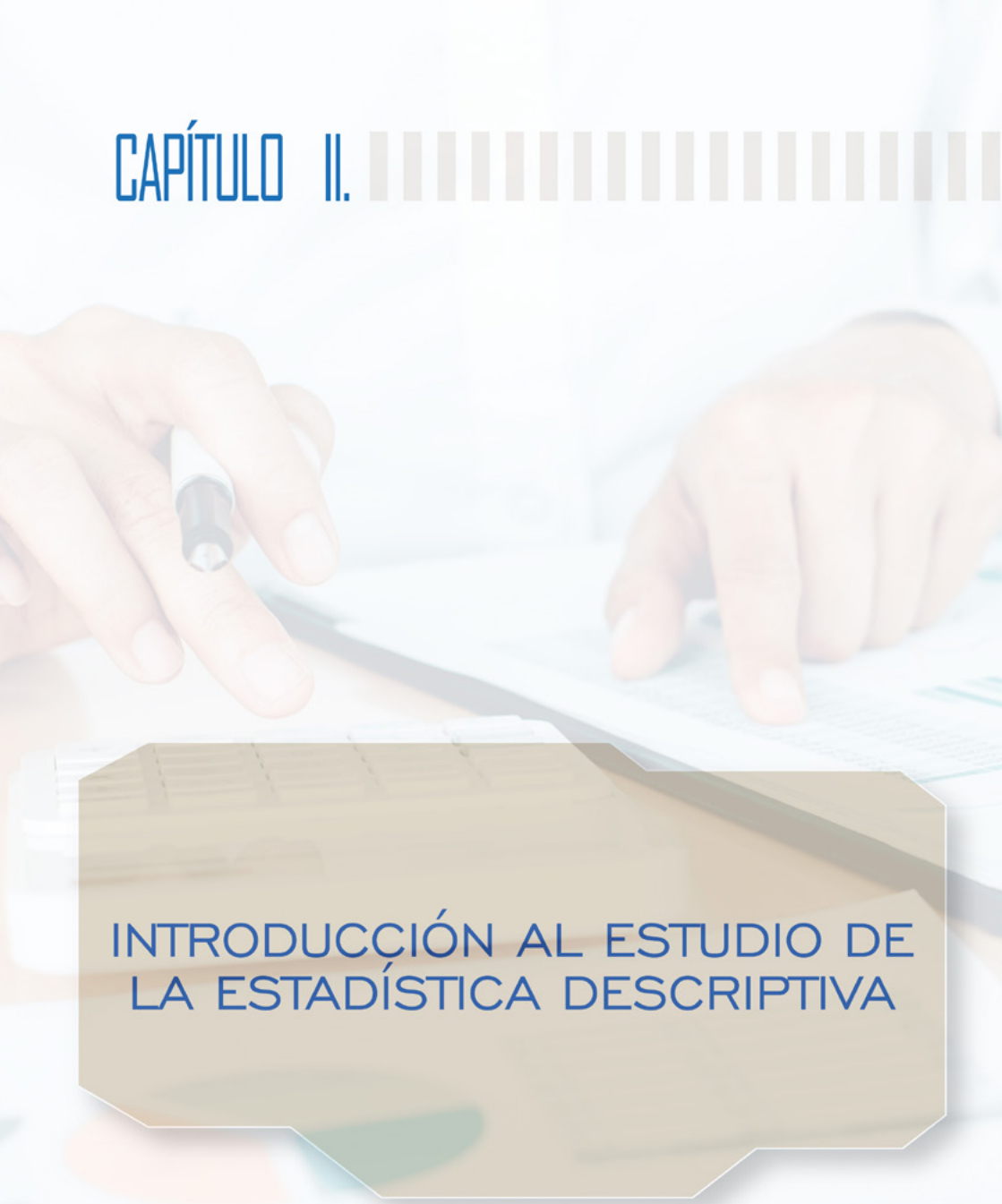
- No probabilística: Las técnicas de muestreo de tipo no probabilísticas la selección de los sujetos a estudio dependerá de ciertas características, criterios, etc. que el investigador considere en ese momento; por lo que pueden ser poco válidos y confiables o reproducibles; debido a que este tipo de muestras no se ajustan a un fundamento probabilístico, es decir, no dan certeza que cada sujeto a estudio represente a la población blanco

- Intencional: Permite seleccionar casos característicos de una población limitando la muestra sólo a estos casos. Se utiliza en escenarios en las que la población es muy variable y consiguientemente la muestra es muy pequeña.
- Por conveniencia: Permite seleccionar aquellos casos accesibles que acepten ser incluidos. Esto, fundamentado en la conveniente accesibilidad y proximidad de los sujetos para el investigador.
- Accidental o consecutivo: Se fundamenta en reclutar casos hasta que se completa el número de sujetos necesario para completar el tamaño de muestra deseado. Estos, se eligen de manera casual, de tal modo que quienes realizan el estudio eligen un lugar, a partir del cual reclutan los sujetos a estudio de la población que accidentalmente se encuentren a su disposición

Distribución

Describe el comportamiento de una variable asociada a un conjunto de datos; consiste de todos los valores diferentes que toma la variable, e incluye las frecuencias con las que se repite cada valor.

CAPÍTULO II.



INTRODUCCIÓN AL ESTUDIO DE LA ESTADÍSTICA DESCRIPTIVA

CAPÍTULO II. INTRODUCCIÓN AL ESTUDIO DE LA ESTADÍSTICA DESCRIPTIVA

La estadística descriptiva es un conjunto de técnicas numéricas y gráficas para describir y analizar un grupo de datos, sin extraer conclusiones (inferencias) sobre la población a la que pertenecen. En un sentido muy amplio, la estadística descriptiva es la parte o fase de la estadística dedicada a la descripción de un conjunto de n datos, entendiendo por descripción la clasificación, representación gráfica y resumen de los mismos. En un contexto más general esos n datos constituirán una muestra de tamaño n extraída de una población, y la descripción de dicha muestra habrá de completarse posteriormente con una inferencia o generalización al total de la población.

Métodos numéricos. Medidas de tendencia central

Los métodos tabulares y gráficos tienen algunas limitaciones para describir y analizar un conjunto de datos. Por ejemplo, si es necesario realizar la descripción de un fenómeno ante un grupo de personas, se estaría en seria desventaja si no se cuenta con el material y equipo necesario para elaborar tabulaciones o gráficas. Ante esta situación, se acude a otras herramientas proporcionadas por la estadística descriptiva: las medidas de tendencia central y de dispersión.

Las medidas de tendencia central son medidas descriptivas que señalan hacia dónde tienden a concentrarse los valores contenidos en un conjunto de datos, es decir, proporcionan información sobre la posición o localización de los datos observado. Su resultado debe ser un valor típico o representativo de la muestra o población, el cual es utilizado para describir o analizar un fenómeno. Al ser una idea abstracta y representativa

del conjunto de datos, las medidas de tendencia central tienen la ventaja de poder ser transmitidas de manera verbal.

Las medidas de tendencia central ofrecen información acerca de valores típicos o promedio de un conjunto de datos.

Medidas de tendencia central en datos no agrupado

Media aritmética (ẍ)

Esta medida suele recibir también el nombre de media o promedio, y es el valor estadístico de tendencia central más utilizado, su confiabilidad depende de la forma de su distribución y de la existencia o no de valores extremos. Por lo general, es una buena representación de un conjunto de datos y se le puede considerar como el punto de equilibrio (o centro de gravedad) de un conjunto de mediciones o puntuaciones, en el caso de que no se encuentren agrupadas en intervalos, se define como la suma de todas ellas, dividida entre el total de casos.

Más formalmente, la media de una muestra se define como

$$\ddot{x} = \frac{\sum x}{n}$$

Mientras que su equivalente para la población está dado por

$$\mu = \frac{\sum x}{N}$$

Las formas de los estadísticos y de los parámetros difieren únicamente en el símbolo que se utiliza en la parte izquierda de la ecuación, y el uso de n minúscula y mayúscula, que es una norma común para denotar el número de observaciones en una muestra y en una población, respectivamente. La media ***ẍ***, única para un conjunto de datos, se sitúa en el centro de gra-

vedad de la distribución de los mismos reforzando su papel de medida de tendencia central. Sin embargo, debe tenerse en cuenta que la media es una medida sensible a observaciones atípicas o extremas. Un valor alejado del resto tendría un efecto importante sobre el valor de la media. Por ejemplo, si se trata de un valor considerablemente mayor que el conjunto de las observaciones, la media se desplazará hacia la derecha (aumentará su valor), pudiendo situarse en un lugar poco representativo del conjunto de datos. Existen alternativas ajustadas del cálculo de la media (medias robustas) que tratan de corregir este problema otorgando un menor peso a las observaciones alejadas.

Ejemplo:

Calcule la media de los números 9, 8, 7, 6,5

$$\bar{x}¿\frac{9+8+7+6+5}{5}=7$$

Propiedades de la media

- Se define de modo inequívoco en tanto que su método de cálculo es reconocido en forma general.
- Es única, ya que un conjunto de datos tiene una y sólo una media.
- Su valor está influido por todas las observaciones en el conjunto de datos.

Mediana (Me)

Es el valor que divide a un conjunto de datos en dos partes iguales, de manera que el número de valores mayores que o iguales a la mediana es el mismo que el número de valores menores que o iguales a la mediana. El valor de la mediana, para un conjunto de datos, se obtiene de forma que deja el mismo número de observaciones a su izquierda que a su derecha.

La forma más común de calcular la mediana cuando el número de valores es impar es ordenar las observaciones en términos de su magnitud y luego elegir el valor intermedio como la mediana. Una manera más formal de expresar lo anterior está dado por:

$$Mediana(n\,impar)=\frac{X_{n+1}}{2}$$

donde n es el número de observaciones y $\frac{n+1}{2}$ es el subíndice de x

Cuando el número de observaciones en un conjunto de datos es par, no hay un valor medio a elegir como la mediana. En este caso, la mediana se calcula como la media de los dos valores intermedios. Una manera más formal de expresar esto está dado por:

$$Mediana(n\,par)=\frac{\frac{Xn}{2}+\frac{Xn+1}{2}}{2}$$

Donde $\frac{n}{2}$ y $\frac{n}{2}$ +1 son los subíndices para identificar los dos valores intermedios

Ejemplo 1:

Calcule la mediana de:

54, 55,64, 64,69, 71,73, 74,75, 75,76, 78, 80, 82, 90

Dado que, en este caso, el número de datos es impar, solo hay un valor que se sitúa en el centro, dejando el mismo número de datos a izquierda y a derecha, que es el que ocupa la posición 8 (deja siete datos a su izquierda y siete a su derecha). Si el número total de datos fuera par, se calcularía la semisuma entre los dos datos centrales. En general, se calculará en primer lugar el rango de la mediana, que informará sobre la posición que

debe ocupar esta, una vez ordenados los datos de menor a mayor, de la siguiente forma:

$$Me = \frac{n+1}{2} = \frac{15+1}{2} = 8$$

En este caso, la mediana es el dato que ocupa la posición 8 y sería Md = 74

Ejemplo 2:

Si se considerara un conjunto de n = 16 observaciones (se añade la observación 93 al grupo anterior), se tendrá que:

:54, 55,64, 64,69, 71,73, 74,75, 75,76, 78, 80, 82, 90, 93

Donde:

$$Me = \frac{74+75}{2} = 74,5$$

La mediana sería un valor entre el dato que ocupa la posición 8 y el dato que ocupa la posición 9, que en este caso corresponde a los valores 74 y 75. La mediana se obtendrá entonces:

Moda (Mo)

La moda se define, para un conjunto de datos, como el valor más frecuente, es decir, el valor que más veces se repite. Si todos los puntos en un conjunto ocurren con la misma frecuencia, no hay moda. Por otro lado, si dos o más puntuaciones ocurren con igual frecuencia y esa frecuencia es mayor que la de las otras puntuaciones en el conjunto, entonces habrá más de una moda. En el caso de datos de nivel nominal u ordinal se puede calcular la categoría modal. La categoría modal es aquella que tiene mayor frecuencia. Si dos o más categorías tienen la misma frecuencia y ésta es mayor que la de todas las demás categorías, entonces hay más de una categoría modal.

Siguiendo con el ejemplo 2 se observa que dos datos se repiten exactamente el mismo número de veces y representan la mayor frecuencia observada y son: 64 y 75. Por tanto, se dispondría de dos valores para la moda: Mo = {64,75}

Este resultado evidencia que la moda no tiene por qué ser única para un conjunto de datos. Por otra parte, basta que un dato se repita más veces que el resto para considerarse moda, aunque no sea una buena medida resumen de los datos, siendo, por tanto, la medida más débil de las estudiadas hasta el momento. Una alternativa para el cálculo de la moda en el caso de variables cuantitativas continuas, donde es habitual observar frecuencias bajas en la mayoría de los valores observados, es agrupar en intervalos y detectar el intervalo o intervalos con mayor frecuencia absoluta lo que podría definirse como intervalo modal.

Media geométrica

La media geométrica de un conjunto de observaciones es la raíz n ésima de su producto. El cálculo de la media geométrica exige que todas las observaciones sean positivas. Esta es una medida que puede aplicarse al crecimiento exponencial o interés compuesto, pues obtiene la raíz enésima de un grupo de n datos multiplicados entre sí, por ejemplo, la raíz cúbica del producto de 3 datos, o la raíz octava del producto de 8 datos. El resultado obtenido, al elevarse a la potencia enésima, produce el producto de todos los datos multiplicados entre sí.

Para una población

$$G=\sqrt[n]{\sum x_1}$$

Para una muestra

$$G=\sqrt[N]{\sum x_1}$$

Características de la media geométrica

1. El cálculo de la media geométrica está basado en todos los elementos de un conjunto de datos. El valor de cada elemento de dicho conjunto afecta así el valor de la media geométrica
2. Si uno de los valores es cero, el valor de G es cero.
3. Si uno de los valores es negativo y el número de datos es par, el valor de G es imaginario y no tiene interpretación. Si uno de los valores es negativo y el número de datos es impar, aunque G existe, su valor no es representativo
4. La media geométrica es afectada por valores extremos en una menor cantidad que lo es la media aritmética. Por ejemplo, la media geométrica de los valores 1, 4 y 16 es 4, mientras que la media aritmética de los mismos valores es 7. El valor 7 es más cercano al valor alto 16 que el valor 4 lo es de 16. El valor de G es siempre menor que el valor de la media de los mismos datos, excepto cuando todos los valores en una serie son iguales, tales como la media geométrica y la media aritmética para los valores 4, 4 y 4 que son ambas 4
5. La media geométrica da igual ponderación a las tasas de cambio iguales. En otras palabras, al promediar tasas de cambio geométricamente, la tasa que muestra el doble de su base es compensada por la otra que muestra la mitad de su base; la tasa que muestra un quinto de su base; y así sucesivamente. Las tasas de cambio son ordinariamente expresadas en porcentajes. Puesto que la base de cada proporción expresada en porciento es siempre igual a 100%, el promedio de dos proporciones las cuales se compensan deberá ser 100% también
6. La media geométrica de las proporciones de los valores individuales con respecto a cada valor precedente en una secuencia de valores es la única medida de tendencia central apropiada para las proporcio-

nes. La media aritmética de las proporciones no dará un resultado consistente

Media armónica

La media armónica (H ó Mh) es una medida o estadígrafo de tendencia central similar al promedio o media aritmética; salvo por que se construye a partir de los valores recíprocos de la variable. Por tanto, es el valor recíproco de la media aritmética. La media armónica se calcula a partir del valor inverso de la sumatoria del inverso de los valores de la variable.

Para la población

$$MH = \frac{n}{\sum\left(\frac{1}{x_i}\right)}$$

Para la muestra

$$MH = \frac{N}{\sum\left(\frac{1}{x_i}\right)}$$

A diferencia de la media aritmética que resulta altamente sensible a los valores extremos inferiores o superiores y requiere de distribuciones más o menos simétricas; la media armónica muestra menos sensibilidad a los valores altos y mantiene su representatividad en distribuciones asimétricas o discontinuas. Esta medida utiliza todos los datos de la distribución y se emplea para promediar variaciones con respecto a la misma variable; como pueden ser la productividad, el precio, el tiempo, la velocidad, el rendimiento, etc.

Media ponderada

En ocasiones, los datos que se brindan, no tienen el mismo peso porcentual, o tienen una frecuencia distinta entre sí, motivo por el cual se utiliza el promedio ponderado. Cuando los valores por promediar tienen diferentes grados de importancia entre sí, debe utilizarse el promedio ponderado, el cual aplica un factor de ponderación (o importancia relativa) a cada uno de los valores que se van a promediar. Si los datos están dados porcentualmente, la fórmula está dada por:

Para una población:

$$\mu = \sum w \frac{X}{\sum w}$$

Para una muestra:

$$X = \sum w \frac{x}{\sum w}$$

Donde $\sum wX$ es la suma de todos los pesos (w) multiplicada por los valores observados (X), en tanto que es igual a N (el número de observaciones de la población) o n (e número de observaciones de la muestra).

Media cuadrática

Se define esta como la raíz cuadrada de la media aritmética de los cuadrados de los valore

$$XQ = \sqrt{\frac{\sum x^2}{N}}$$

Esta media tiene su utilidad con frecuencia en la aplicación a fenómenos físicos.

Medidas de tendencia central y de dispersión en datos agrupados

Se identifica como datos agrupados a los datos dispuestos en una distribución de frecuencia. En tal caso las fórmulas para el cálculo de la media, mediana, moda, varianza y desviación estándar deben incluir una leve modificación. A continuación, se entregan los detalles para cada una de las medidas.

Media en datos agrupados

La fórmula es la siguiente

$$\bar{x} \frac{\sum X_{in_i}}{n}$$

Donde ni representa cada una de las frecuencias correspondientes a los diferentes valores de x_i.

Ejemplo:

Una distribución de frecuencia de madres que asisten a un programa de lactancia materna, clasificadas según el número de partos (tabla 1)

Tabla 1. Número de parto

N° de parto (X)	n_i	$X_i n_i$	Frecuencia acumulada
1	4	4	4
2	13	26	17
3	16	48	33
4	6	24	39
5	3	15	42
Total	**42**	**117**	

$$\bar{x} \frac{117}{42}=2,78$$

Entonces las 42 madres han tenido, en promedio, 2,78 partos

Mediana en datos agrupados

Si la variable es de tipo discreto la mediana será el valor de la variable que corresponda a la frecuencia acumulada que supere inmediatamente a n/2. En los datos de la tabla 1 Me=3, ya que 42/2 es igual a 21 y la frecuencia acumulada que supera inmediatamente a 21 es 33, que corresponde a un valor de variable (X_i) igual a 3.

Moda en datos agrupados

Si la variable es de tipo discreto la moda será al valor de la variable (X_i) que tenga la mayor frecuencia absoluta. En los datos de la tabla 1 el valor de la moda es 3 ya que este valor de variable corresponde a la mayor frecuencia absoluta =16.

Varianza en datos agrupados

Para el cálculo de varianza en datos agrupados se utiliza la fórmula

$$S^2 = \sum ¿¿¿$$

Con los datos de la tabla 1 y recordando que el promedio ($\bar{X}$) resultó ser 2,78 partos por madre,

N° de parto (X)	n_i	$x_i n_i$	$(x_i - \bar{X})^2$	$(x_i \bar{X})^2 n_i$
1	4	4	3,1684	12,67
2	13	26	0,6084	7,9
3	16	48	0,0484	0,7747
4	6	24	1,4884	8,93
5	3	15	4,9284	14,7852
Total	**42**	**117**		**45,06**

$$S^2 = \sum ¿¿¿ = \frac{45,06}{42-1} = \frac{45,06}{41} 1,1$$

Medidas de variabilidad

Las medidas de tendencia central proporcionan información sobre la localización de los datos, pero no sobre la dispersión o variabilidad con la que se sitúan en torno a dichas medidas. Sería incorrecto concluir que dos conjuntos de datos son iguales sólo porque tienen las mismas medidas de tendencia central.

Es decir, en caso de que el valor de la media aritmética sea el mismo para ambos conjuntos, la mediana y la moda también podrían ser iguales, pero la distribución de tales datos forma una curva completamente diferente. Esto ocurre porque las distancias de los datos tienen diferentes concentraciones respecto del punto de equilibrio, que está representado por la media aritmética. Para medir la concentración de los datos, se emplean las medidas de variación o dispersión, también conocidas como medidas de variabilidad.

Rango

El rango está en función únicamente de las puntuaciones más grande y más pequeña de un conjunto de datos. La medida más sencilla y visualmente intuitiva para cuantificar la dispersión de los datos es el rango y se obtendrá calculando la distancia entre el mayor y el menor valor observado. Si se trabaja con los datos del ejemplo A se tendrá que:

$$R= X_{máx} - X_{mín} = 90-54=36$$

Luego el rango de valores observados muestra una distancia de 36 entre el menor y el mayor valor observado. La obtención del rango es sencilla, sin embargo, en su construcción solo intervienen dos de los datos observados, que, además, son los más extremos. Esto tiene como consecuencia que el rango será una medida extremadamente sensible a observaciones extremas y que no tiene en cuenta gran parte de la información disponible.

Desviación media

Otra manera de estimar la dispersión de los valores de la muestra es comparar cada uno de estos con el valor de una medida de centralización. Una de las medidas de dispersión más usada es la desviación media, también llamada con más precisión desviación media respecto a la media aritmética. Se define como la media aritmética de las diferencias absolutas entre los valores de la variable y la media aritmética de la muestra.

Al igual que el rango, la desviación media es una medida de variabilidad altamente intuitiva. Sin embargo, a diferencia del rango, la desviación media toma en cuenta todos los datos para el cálculo de la variabilidad, por lo que se trata de un estadístico más estable.

Diversas medidas de variabilidad están basadas en las diferencias entre los valores de una distribución y algún punto central en la distribución. Por ejemplo, suponga que se calcula la diferencia x - ẍ para cada puntuación de un conjunto de datos. Este valor, llamado puntuación de desviación o simplemente una desviación, indica el número de unidades entre la puntuación y la media. Cuando los datos están agrupados de forma estrecha alrededor de la media, las desviaciones tienden a ser pequeñas. Para datos que están más dispersos, las desviaciones son más gran des. Es posible que una representación razonable de la variabilidad se base en el promedio de estas desviaciones.

Cuando los datos se encuentran más esparcidos, el promedio de las desviaciones es mayor que para los datos con menor dispersión. La dificultad del uso de las desviaciones de esta forma es que siempre suman cero. Esto, entonces, es el fundamento de la desviación media. La desviación media (MD) es el promedio de los valores absolutos de las desviaciones en un conjunto de puntuaciones. La expresión formal es:

$$MD = \frac{\sum x - \ddot{x}}{n}$$

Ejemplo 3:

Tabla 2. Cálculo de la desviación media y la varianza de la distribución A

	(1)	(2)	(3)	(4)	(5)
	x	X^2	x-ẍ		
	3	9	-1	1	1
	3	9	-1	1	1
	3	9	-1	1	1
	4	16	0	0	0
	4	16	0	0	0
	4	16	0	0	0
	4	16	0	0	0
	4	16	0	0	0
	5	25	1	1	1
	5	25	1	1	1
	5	25	1	1	1
Σ	**44**	**182**	**0**	**6**	**6**

Fuente: González (4)

Tabla 3. Cálculo de la desviación media y la varianza de la distribución B

(1)	(2)	(3)	(4)	(5)
x	X^2	x-ẍ		
1	1	-3	3	9
2	4	-2	2	4
2	4	-2	2	4
3	9	-1	1	1
4	16	0	0	0
4	16	0	0	0
4	16	0	0	0
5	25	1	1	1
6	36	2	2	4

	6	36	2	2	4
	7	49	3	3	9
∑	**44**	**212**	**0**	**16**	**36**

Fuente: González (4)

La columna (1) de la tabla 2 muestra las puntuaciones que forman la distribución A. Las columnas (3) y (4) de esta tabla muestran, respectivamente, las puntuaciones de desviación y los valores absolutos de las puntuaciones de desviación para los datos de la columna (1). Utilizando la suma de la columna (4) obtenemos:

$$MD=\frac{6}{11}=.55$$

La tabla 3 presenta los resultados para la distribución B. Utilizando la suma de la columna (4) de esta tabla obtenemos

$$MD=\frac{16}{11}=1,45$$

De esta manera, el promedio de la desviación de las puntuaciones de la distribución A fue de .55 unidades, mientras que en la distribución B fue de 1.45 unidades; por lo tanto, se confirma que la distribución B tiene mayor variabilidad que la distribución A.

Varianza

La varianza es una medida de variabilidad menos intuitiva pero generalmente más útil que el rango o la desviación media. Como estadístico descriptivo, la varianza es menos interesante que la desviación media, pero en general resulta más útil en virtud de su papel en la inferencia.

La varianza está expresada, por tanto, en unidades al cuadrado de la variable. Para conseguir una medida en las mismas unidades que la variable

original se extrae la raíz cuadrada, obteniéndose la denominada desviación típica o estándar.

$$\sigma^2 = \frac{\sum (x-\mu)^2}{N}$$

De esta manera, el parámetro de la varianza es el promedio del cuadrado de las desviaciones de las puntuaciones que conforman la población. El estadístico es

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

Para calcular la varianza de la tabla 2 la columna (5) muestra el cuadrado de las desviaciones de los datos de la distribución A. Utilizando la suma de esta columna se obtiene:

$$s^2 = \frac{6}{10} = .60$$

Al utilizar la suma de la misma columna en la tabla 3 se obtiene

$$s^2 = \frac{36}{10} = 3{,}60$$

Desviación estándar

Sin lugar a dudas la medida más usada para estimar la dispersión de los datos es la desviación estándar. Esta es especialmente aconsejable cuando se usa la media aritmética como medida de tendencia central. Se basa en un valor promedio de las desviaciones respecto a la media. En este caso, en vez de tomar valores absolutos de las desviaciones, para evitar así que se compensen desviaciones positivas y negativas, se usan los cuadrados de las desviaciones. Esto hace además que los datos con desviaciones grandes influyan mucho en el resultado final. Se define entonces la varianza de una muestra con datos repetidos.

el parámetro está dado por

$$\sigma = \sqrt{\sum ¿¿¿¿}$$

$$\sigma = \sqrt{\frac{\sum x^2 - \frac{\left(\sum x\right)^2}{N}}{N}}$$

el estadístico es

$$s = \sqrt{\sum ¿¿¿¿}$$

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

Posición de un dato respecto de la media

Al conocer el valor de la desviación estándar de una población se hace posible establecer qué desviación tiene un dato en particular respecto de la media aritmética, en valores de esa medida de la dispersión.

Por ejemplo, supóngase que en una población (y en este caso se está considerando una población de un gran tamaño, como los habitantes de una nación o los pacientes que padecen una afección determinada) la media aritmética para una variable evaluada en forma de datos numéricos es 195 y la desviación estándar es 6; si el dato para un integrante de esa población es 204, puede decirse que ese dato está 9 unidades (de las utilizadas para la evaluación de la variable) por encima de la media, que significa 1,5 desviaciones estándar, según surge del siguiente cálculo:

$$(x - \mu) / \sigma = (204 - 195) / 6 = 1{,}5$$

Al resultado de la ecuación se lo designa habitualmente con el símbolo z e indica la ubicación de un dato respecto de la media de la población a la que pertenece en términos de desviaciones estándar.

Medidas de posición relativa

Las medidas de posición son métodos que localizan las posiciones relativas de las observaciones en una distribución. Por ejemplo, tal vez desee encontrar un punto en la escala de medición debajo del cual se localiza el 25% de las observaciones de la distribución. Los puntos caracterizados en esta forma se llaman percentiles. Por otra parte, quizás desee calcular el porcentaje de observaciones que se localizan por debajo de un punto específico en la escala. Estos porcentajes se llaman rangos percentiles. Por último, tal vez desee determinar la posición de una observación en relación con la media de una distribución por medio de las llamadas puntuaciones z.

Percentiles

Los percentiles son los valores de la variable que dividen a un conjunto de datos ordenados en cien partes iguales. Un percentil es un punto en la escala de medición debajo del cual se localiza un porcentaje específico de las observaciones. Con base en esta definición, la mediana se define como el percentil 50.

$$P_p = LRL + (w)\left[\frac{(pr)(n) - cf}{f}\right]$$

donde

- P_p representa el p-ésimo percentil
- LRL es el límite real inferior del intervalo que contiene el p-ésimo percentil

- w es la anchura del intervalo, calculada como la diferencia entre los límites reales superior e inferior de ese intervalo
- pr es p expresada como una proporción (esto es, p/100)
- n es el número total de observaciones
- cf es la frecuencia acumulativa hasta el intervalo del percentil
- f es la frecuencia de ese intervalo

Tabla 4. Distribución de frecuencias de las presiones sanguíneas de 144 adolescentes moderadamente obesos

PS	Frec	PS	Frec	PS	Frec	PS	Frec
143	2	128	3	113	0	98	2
142	0	127	3	112	0	97	2
141	0	126	7	111	3	96	2
140	4	125	4	110	3	95	3
139	6	124	4	109	1	94	0
138	3	123	2	108	0	93	1
137	11	122	3	107	2	92	2
136	3	121	1	106	1	91	0
135	8	120	3	105	2	90	1
134	5	119	2	104	0	89	0
133	8	118	2	103	1	88	0
132	4	117	1	102	1	87	0
131	3	116	3	101	0	86	1
130	5	115	6	100	4		
129	3	114	2	99	1		

Fuente: González (4)

Ejemplo 4: calcular los percentiles 25, 60 y 75 de los datos de la tabla 4

Mediante la construcción de una distribución de frecuencias acumulativas y observando que (.25) (144) =36, el intervalo del percentil 25 es 114.5 a 115.5. Esto se deduce del hecho de que la frecuencia acumulativa hasta el límite real inferior de 114.5 es 35 y para 115.5 es 41. El punto en la escala

debajo del cual hay 36 observaciones debe estar entre estos dos límites. Utilizando esta información con la ecuación

$$P_p = LRL + (w)\left[\frac{(pr)(n) - cf}{f}\right]$$

$$P_{25} = 114{,}5 + (1.0)\left[\frac{(25)(144) - 35}{6}\right] = 114.67$$

Dado que (6) (144) 86.4, y que las frecuencias acumulativas hasta 129.5 y 130.5 son 82 y 87 respectivamente, P_{60} debe localizarse en este intervalo. Nuevamente, aplicando la ecuación y se obtiene

$$P_{60} = 129{,}5 + (1.0)\left[\frac{(.6)(144) - 82}{5}\right] = 130.38$$

Utilizando el mismo método, P_{75} se obtiene de la siguiente manera

$$P_{75} = 134{,}5 + (1.0)\left[\frac{(.75)(144) - 107}{8}\right] = 134{,}63$$

Cuartiles

Se llaman cuartiles a tres valores que dividen la distribución en cuatro partes iguales. Se representan y designan como cuartil primero (Q1), segundo (Q2) y tercero (Q3). Cada parte agrupa, por tanto, al 25%, al 50% y al 75% de los datos de la distribución. El primero deja a su izquierda (o debajo, según se prefiera) el 25 % de los datos; el segundo deja a la izquierda el 50 %, por lo que se trata de la propia mediana; el tercero deja a la derecha el 25 %.

$$Q_3 = L_i + c\,\frac{\frac{3N}{4} - F_{i-1}}{F_i}$$

Deciles

Análogamente, se llaman deciles a nueve valores de la variable que dividen a la distribución en diez partes iguales. Es decir, los deciles agrupan

a los datos en diez partes correspondientes cada una con el 10% de la distribución. Se representan por D1, D2, ..., D9 y la expresión que permite calcularlos es:

$$D_k = L_i + c\,\frac{\frac{Kn}{10} - F_{i-1}}{f_i}$$

Estudio de las distribuciones de frecuencias

Distribución de frecuencia

También es conocido como tablas de frecuencia el cual indica cómo un conjunto de datos se divide en varias categorías (o clases) al alistar todas las categorías junto con el número de valores de los datos que hay en cada una.

Las frecuencias asociadas a valores o rango de valores de una variable indican la cantidad de veces que el valor fue observado en el conjunto de unidades en análisis (5). Las frecuencias sirven para conocer cómo se distribuyen los datos o valores de la variable, permitiendo aproximar la distribución de frecuencias a alguna función o modelo teórico para posteriores análisis y cálculos probabilísticos.

Analizando las frecuencias es factible identificar datos extremos (es decir poco frecuentes por ser muy pequeños o muy grandes), y valores, o conjuntos de valores, que aparecen con mayor frecuencia. Las frecuencias en que se presentan los valores de una variable se pueden tabular o graficar.

Tablas de distribuciones de frecuencias

Una tabla de frecuencias organiza los datos de manera tal que en una columna de la tabla aparecen los valores de la variable, según el tipo de variable, y en sucesivas columnas se muestran diferentes tipos de frecuen-

cias asociadas a esos valores (frecuencias absolutas, frecuencias relativas, frecuencias absolutas acumuladas y frecuencias relativas acumuladas). Las tablas de frecuencias son una de las técnicas básicas para el resumen de información a partir de una muestra de datos.

Los principales elementos de una tabla estadística son: Título, unidades, encabezado, cuerpo o contenido, nota de pie valores y referencias. Se elabora colocando en la primera columna los datos diferentes o subgrupos de datos (llamados clases o intervalos de clase) y en la columna siguiente el número de observaciones que corresponden a cada dato o a cada grupo de datos llamada frecuencia).

Una tabla de este tipo dará, en forma abreviada, una información completa acerca de la distribución de los valores observados. Estas tablas facilitan el uso de los métodos gráficos y aritméticos. La presentación de los datos en forma ordenada, por medio de una tabla, dependerá de los datos de que se trate, y si estos son cualitativos o cuantitativos.

Las tablas de frecuencias se utilizan para representar la información contenida en una muestra de tamaño n extraída de una población, (x1, . . . , xn).

Modalidades:

Cada uno de los valores que puede tomar una variable (cualitativa o cuantitativa discreta). Se denotan como:

$$c_i , i = 1, \ldots , k.$$

El número de individuos de la muestra en cada modalidad ci se denota por ni

- Frecuencia absoluta: para cada modalidad ci, la frecuencia absoluta es

$$n_i , i = 1, \ldots , k.$$

- Frecuencia relativa: para cada modalidad ci , la frecuencia relativa es

$$fi = ni /n, i = 1, \ldots, k.$$

- Frecuencia absoluta acumulada: la frecuencia absoluta acumulada de una modalidad ci es

$$¿=\sum_{i=1}^{i} n_{i=n1+\ldots\ni,i=1\ldots.k}$$

- Frecuencia relativa acumulada: la frecuencia relativa acumulada de una modalidad ci es

$$Fi=\sum_{i=1}^{i} f_{i=f\,1+\ldots fi=\frac{N_i}{n},i=1\ldots\ldots\ldots,k}$$

Distribuciones de frecuencias relativas

La frecuencia relativa se obtiene dividiendo cada frecuencia entre el número total de respuestas. La frecuencia relativa, entonces, es la proporción de respuestas de cada tipo, es decir, representan el porciento de veces en que ocurre un dato.

Frecuencia relativa fi: Cociente entre la frecuencia absoluta y el número de observaciones realizadas N. es decir:

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

Cumpliéndose las propiedades

$$0\leq fi\leq 1;\sum_{i=1}^{k} fi=\sum_{i=1}^{k}\frac{n_i}{N}\frac{\sum_{i=1}^{k} n_i}{N}=1$$

Esta frecuencia relativa se puede expresar también en tantos por cientos del tamaño de la muestra, para lo cual basta con multiplicar por 100.

En la tabla 1 se muestra el resultado de respuesta de pacientes que calificaron su percepción del dolor en una escala ordinal de cuatro puntos. En dicha tabla están los datos ordenados en distribuciones de frecuencia, frecuencias relativas, frecuencias acumulativas y frecuencias relativas acumulativas. La primera columna lista las categorías de la escala de menor a mayor. La segunda muestra la frecuencia de respuesta para cada las categorías que se obtiene mediante el conteo del número de veces que ocurre cada respuesta en el conjunto de datos. La frecuencia, entonces, es el número de respuestas de cada tipo.

La tabla 5 muestra la frecuencia relativa de respuesta, la cual se obtiene dividiendo cada frecuencia entre el número total de respuestas (en este caso 60). La frecuencia relativa, entonces, es la proporción de respuestas de cada tipo.

Tabla 5. Distribución de mediciones de dolor percibido

Categoría del dolor	Frecuencia	Frecuencia relativa	Frecuencia acumulativa	Frecuencia relativa acumulativa
Severo	4	0,07	60	1
Moderado	8	0,13	56	0,93
Leve	17	0,28	48	0,8
Ninguno	31	0,52	31	0,52

Fuente: Clifford y Taylor (6)

A partir de las dos primeras columnas, se observa que el mayor número de pacientes (31) indicó no haber tenido dolor. Este número representa .52 (o 52%) del total de la muestra. El dolor severo fue menos común, pues únicamente 4 personas (.07 de la muestra) eligieron esta categoría.

Distribuciones de frecuencias acumulativas

La frecuencia acumulativa se obtiene mediante la suma de la frecuencia en una categoría dada con las categorías que indican un nivel menor de la variable medida. Por ejemplo, en la tabla 4 sobre la distribución de mediciones del dolor percibido, la columna de la frecuencia acumulativa muestra el número de pacientes que indicaron que su dolor era menor o igual al nivel representado. Por ejemplo, 48 pacientes (31+17) clasificaron su dolor como leve o menor que leve, mientras que 56 pacientes (31+17+8) percibieron su dolor como moderado o menor que moderado (6).

Distribuciones de frecuencias relativas acumulativas

La frecuencia relativa acumulativa se calcula al dividir cada frecuencia acumulativa entre el número total o población total. Siguiendo con el ejemplo de la tabla 1, se observa que .80 de los pacientes creyeron que su dolor era leve o de menor intensidad, mientras que .93 sintieron que su dolor era moderado o de menor intensidad. La columna de la frecuencia relativa acumulativa, entonces, muestra la proporción de los pacientes que indicaron que su dolor fue menor que o igual que el nivel representado (6).

Las distribuciones de frecuencias, frecuencias relativas, frecuencias acumulativas y frecuencias relativas acumulativas que se muestran en la tabla 1 fueron calculadas para una variable de nivel ordinal. Las primeras dos distribuciones también pueden utilizarse para una variable de nivel nominal.

Distribuciones agrupadas

La distribución de frecuencias agrupadas o tabla con datos agrupados se emplea si las variables toman un número grande de valores o la variable es continua. Se agrupan los valores en intervalos que tengan la misma

amplitud denominados clases. A cada clase se le asigna su frecuencia correspondiente.

Intervalo de clase

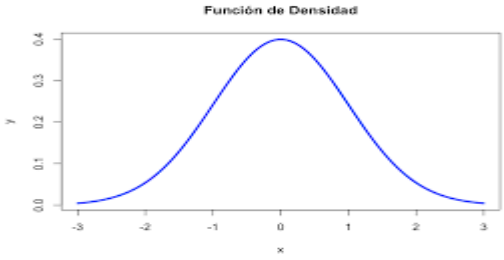
Son los intervalos en los que se agrupan y ordenan los valores observados. Cada uno de estos intervalos está delimitado (acotado) por dos valores extremos llamados límites. El tamaño de los intervalos dependerá del número de intervalos utilizados y viceversa. No existen reglas rígidas y rápidas al respecto. Al construir tablas de este tipo se deben responder dos preguntas relacionadas. ¿En cuántos intervalos se deben agrupar los valores y qué tan grandes deberán ser los intervalos? Muy pocos intervalos provocan la pérdida de mucha información, mientras que muchos intervalos hacen fracasar el propósito de resumir los datos En esencia, se deseará presentar los datos dándoles el mayor significado posible. Sin embargo, hay algunas reglas generales que sirven como guía. Sugerencias:

- no debe haber menos de seis ni más de 15 intervalos. Generalmente, el número de observaciones determina el de intervalos: mientras más observaciones más intervalos (7).
- cuando sea posible, se debe usar una anchura en los intervalos de clase de 5 unidades, de 10 unidades o de algún múltiplo de 10 para que el resumen de los datos sea más comprensible.
- Siempre que sea posible, los intervalos de clase también deben tener la misma longitud, donde el extremo inferior del primer intervalo sea menor que o igual a la medida más pequeña del conjunto de datos, y el extremo superior del último intervalo sea mayor que o igual a la medida más grande. También debemos señalar que los intervalos deben ser contiguos, pero sin traslaparse.

Forma de distribución

La forma que asumen las representaciones gráficas de las distribuciones de frecuencias puede ser diversa. Sin embargo, una gran cantidad de datos numéricos con los que se valoran las variables de interés en el campo de las ciencias de la salud lleva al empleo de gráficos de distribución de frecuencias similares al que se muestra en la figura 4.

Figura 4. Gráfico de la distribución gaussiana (normal)



También es de interés tener presente que esa misma forma de distribución de frecuencias se observa cuando se representa la distribución de los errores que se cometen al registrar los datos. Con menos frecuencia se registran datos que se alejan de ese valor central y más común, y esta situación es la que aparece representada gráficamente.

A esta forma de distribución de datos numéricos se la conoce como distribución normal. Algunas de sus características se deducen de la observación de la figura 4:

- La distribución normal es simétrica respecto de un valor que corresponde a la media aritmética de los datos considerados. Esto significa que los valores de la media aritmética y de la mediana son coincidentes.

- El dato con mayor frecuencia, es decir, la moda representada por el punto más alto de la línea corresponde al valor de la media aritmética y de la mediana. Las tres medidas de tendencia central más comunes son coincidentes en esta forma de distribución (8).
- La forma de la línea puede ser semejante a la del corte de una campana con dos puntos ubicados en forma simétrica respecto de la media, en los cuales la línea cambia de dirección. Esos dos puntos de inflexión corresponden a los representados con los datos ubicados a una desviación estándar por abajo y por arriba de la media aritmética.

CAPÍTULO III.

DATOS: TIPOS Y CARACTERÍSTICAS

CAPÍTULO III. DATOS: TIPOS Y CARACTERÍSTICAS

Variable

Se considera como variable cualquier característica o propiedad general de una población que sea posible medir con distintos valores o describir con diferentes modalidades, por ejemplo: el coeficiente intelectual de los estudiantes de un grupo puede tener diferentes valores, o el estado civil de los empleados de una organización puede estar caracterizado como soltero, casado, separado, entre otros. Así, estas dos características se consideran variables porque, como el término lo indica, varían al medirse o caracterizarse de una unidad de análisis a otra.

En algunos casos, las características de las unidades de análisis pueden ser medidas, mientras que en otros solo es posible describirlas. Para el ejemplo anterior, el coeficiente intelectual es posible medirse en los estudiantes, lo cual obedece a una característica cuantitativa, y el estado civil en los empleados solo se puede describir (no medir), por ser una característica cualitativa. En este sentido, las variables pueden diferenciarse en dos grupos: cualitativas y cuantitativas (9).

Variables cualitativas

Las variables cualitativas son aquellas que representan atributos de los elementos y no permiten una representación numérica definida. Sin embargo, algunas cualidades pueden ser representadas por códigos numéricos que, en el fondo, generan categorías de orden cualitativo. Entre las variables cualitativas están: el estrato socioeconómico, el estado civil, la profesión, el color de una flor, entre otras.

Variables cuantitativas

Estas variables permiten una escala numérica y las características de los elementos son observados cuantitativamente a través de una medida y una escala definidas. Entre las variables cuantitativas se encuentran: el salario de los empleados, la talla de una persona, el peso, el número de hijos en una familia, entre otros.

Las variables cualitativas y cuantitativas se representan con letras mayúsculas del alfabeto (X, Y, Z...) y los atributos de cada variable se simbolizan con letras minúsculas en compañía de subíndices. Por ejemplo, la variable estado civil de los empleados en una empresa puede ser representada por la letra X y sus posibles atributos de "soltero, casado, separado" se representan por, x1: soltero, x2: casado, x3: separado.

Las variables cuantitativas pueden ser clasificadas en dos grupos: continuas y discretas (10).

- variable cuantitativa continua: si a lo largo de un intervalo puede tomar cualquier valor; es decir, entre uno y otro valor de la variable siempre puede existir otro valor intermedio. Son variables cuantitativas continuas la talla o altura de personas, el peso de objetos, el salario de empleados, el tiempo dedicado a realizar una actividad, la temperatura de un lugar, entre otras.
- variable es cuantitativa discreta: si solo puede tomar un valor de un conjunto de números, existen separaciones entre dos valores sucesivos que no pueden llenarse con valores intermedios, en este caso la variable toma valores aislados. Por ejemplo, los empleados de una organización, artículos vendidos en un almacén, instituciones educativas de un sector; en estos casos, solo es posible medir la variable con valores como 15, 16, 17 u otro número entero y no con valores intermedios, tales como 15,7 o 16,8.

Escalas de medición

La escala de medición es considerada como un sistema que asigna valores numéricos a características susceptibles de medir. Normalmente, las escalas pueden ser de cuatro tipos: nominal, ordinal, de intervalos y de razón.

Escala nominal

La escala nominal se utiliza para representar a las variables cualitativas (también llamadas categóricas) y determina múltiples categorías identificadas por un nombre, que bien pudieron estar fijadas previamente o ser precisadas por el investigador según sus necesidades, manteniendo rigurosidad en su definición y convirtiéndolas en elementos mutuamente excluyentes, pues las categorías son exclusivas y solo existe una para **cada elemento de la población,** algunos ejemplos: color del cabello (negro, rubio, castaño, otro); estado de un artículo (bueno, imperfecto); género de los estudiantes (masculino, femenino).

Escala ordinal

Esta escala se caracteriza por present**ar niveles con un rango determi**nado, lo que facilita la comparación entre ellos y es posible diferenciarlos como "mayor que" o "menor que". Es importante resaltar que la distancia entre un nivel y otro adyacente no es la misma (1).

Ejemplos:

- estado de salud de una persona:
  - ✓ sano
  - ✓ ligeramente afectado
  - ✓ enfermo
  - ✓ muy enfermo
- Producción en una empresa:

- ✓ Alta
- ✓ Media
- ✓ baja

A pesar de que los niveles se pueden representar por un número, éste cumple la función de etiqueta y no es posible usarlo como cantidad numérica en operaciones matemáticas, tal es el caso del estrato socioeconómico, donde los números representan un nivel satisfacción de necesidades y un orden definido en los sectores a los cuales pertenece: estrato $1 < 2 < 3 < 4 < 5 < 6$.

Escala de intervalo

La escala de intervalo presenta mayor información que las escalas nominal y ordinal. Se caracteriza por establecer de forma ordenada los niveles y si la distancia entre uno y otro es la misma, lo cual conlleva a usar una unidad de distancia de referencia. Por esta razón, esta escala permite relacionar intervalos y de esta forma se puede observar, por ejemplo, que la distancia entre 5 y 6 es la misma que existe entre 23 y 24. En esta escala se asigna el punto cero como una medida arbitraria y no implica ausencia de la característica que se está midiendo Un ejemplo típico para esta escala es la medición de la temperatura; para ello se pueden usar varios sistemas: el Celsius, Kelvin o Fahrenheit. En cualquiera de estos sistemas se observa que la distancia entre un grado y el consecutivo es la misma; además, el cero en cualquiera de ellos no implica ausencia de temperatura (9).

Escala de razones

La escala de razones es aquella que posee más información en relación a las escalas anteriores. Toma un cero no arbitrario (absoluto) que significa ausencia del atributo o la característica; esto facilita la comparación, tanto en intervalos como en razones, en cualquier sistema de medición

que se utilice. Por ejemplo, si un elemento posee una longitud de 4,6 cm tendrá el doble de extensión al compararse con otro elemento que mide 2,3 cm, en cualquier tipo de sistema en que se registre la longitud. A esta escala pertenecen todas aquellas mediciones que están relacionadas con el tiempo, longitud, superficie (áreas), capacidad (volúmenes), peso, dinero, entre otras.

En términos generales, se denomina para cualquier escala de medición como débil si contiene poca información, razón por la cual restringe la aplicación de los métodos estadísticos. Mientras que las escalas con mayor información son consideradas como fuertes y es posible analizarlas por medio de métodos específicos diseñados para explicar su comportamiento.

Si se clasifica de una manera más amplia, las variables cualitativas utilizan la escala nominal, mientras que las cuantitativas emplean las escalas de razones o de intervalo. Las variables que usan la escala ordinal se ubican en una transición entre cualitativas y cuantitativas.

Características de los conjuntos de datos

Tipos de datos

Se puede definir los datos como aquella información extraída de la realidad que tiene que ser registrada en algún soporte físico o simbólico que implica una elaboración conceptual y además que se pueda expresar a través de alguna forma de lenguaje.

La estructura de los datos está compuesta por tres elementos

Unidad de análisis: Son los elementos menores y no divisibles que componen el universo de estudio de una investigación. El mismo puede ser una persona, una familia, un país, una región, una institución o en general, cualquier objeto.

Variable: algún aspecto y/o magnitud de un elemento o unidad de análisis capaz de asumir diferentes cualidades y/o valores, en otras palabras, cualquier característica de la unidad de observación que interese registrar, la que en el momento de ser registrada pueden ser transformados en un número.

Valor de una variable, observación o medición, al número que describe a la característica de interés en una unidad de observación particular.

Datos categóricos o cualitativos

Las variables categóricas resultan de registrar la presencia de un atributo. Las categorías de una variable cualitativa deben ser definidas claramente durante la etapa de diseño de la investigación y deben ser mutuamente excluyentes y exhaustivas. Esto significa que cada unidad de observación debe ser clasificada sin ambigüedad en una y solo una de las categorías posibles y que existe una categoría para clasificar a todo individuo. En este sentido, es importante contemplar todas las posibilidades cuando se construyen variables categóricas, incluyendo una categoría tal como No sabe / No contesta, o No registrado u Otras, que asegura que todos los individuos observados serán clasificados con el criterio que define la variable. Los datos categóricos se clasifican en dicotómicos, nominales y ordinales (11).

a. Dicotómicos (Dos categorías): El individuo o la unidad de observación puede ser asignada a solo una de dos categorías. En general, se trata de presencia - ausencia del atributo y es ventajoso asignar código 0 a la ausencia y 1 a la presencia. Ejemplos:

1. varón -mujer
2. embarazada - no embarazada
3. fumador -no fumador
4. hipertenso -normotenso

Debe notarse que los ejemplos 1) y 2) definitivamente cubren todas las categorías, mientras que 3) y 4) son simplificaciones de categorías más complejas. En 3) no está claro donde se asignan los ex-fumadores, en tanto que en 4) fue necesario establecer un criterio de corte para armar una variable categórica a partir de una variable numérica

b. Más de dos categorías

- Categorías nominales: No existe orden obvio entre las categorías. Ejemplos: país de origen, estado civil, diagnóstico.
- Categorías ordinales: Existe un orden natural entre las categorías. Ejemplos: 1) Tabaquismo: No fuma / exfumador / fuma ≤ 10 cigarrillos diarios / fuma > 10 cigarrillos diarios 2) Severidad de la patología: Ausente / leve / moderado / severo.

Aun cuando los datos ordinales puedan ser codificados como números como en el caso de estadios de cáncer de mama de I a IV, no se puedes decir que una paciente en el estadio IV tiene un pronóstico dos veces más grave que una paciente en estadio II, ni que la diferencia entre estadio I y II es la misma que entre estadio III y IV.

En cambio, cuando se considera la edad de una persona, 40 años es el doble de 20 y una diferencia de 1 año es la misma a través de todo el rango de valores. Por esta razón, se debe ser cuidadoso al tratar variables cualitativas, especialmente cuando se han codificado numéricamente, ya que no pueden ser analizadas como números, sino que deben ser analizados como categorías. Es incorrecto presentar, por ejemplo, el estadio promedio de cáncer en un grupo de pacientes. En la práctica clínica se usan escalas

para definir grados de un síntoma o de una enfermedad, tales como 0, +, ++, +++. Es importante definir operativamente este tipo de variables y estudiar su confiabilidad de modo de asegurar que dos observadores puestos frente al mismo paciente lo clasificarán en la misma categoría.

Datos numéricos

Una variable es numérica cuando el resultado de la observación o medición es un número. Se clasifican en:

- Discretos. La variable sólo puede tomar un cierto conjunto de valores posibles. En general, aparecen por conteo. Ejemplo: número de miembros del hogar, número de intervenciones quirúrgicas, número de casos notificados de una cierta patología.
- Continuos. Generalmente son el resultado de una medición que se expresa en unidades. Las mediciones pueden tomar teóricamente un conjunto infinito de valores posibles dentro de un rango. En la práctica los valores posibles de la variable están limitados por la precisión del método de medición o por el modo de registro. Ejemplos: altura, peso, pH, nivel de colesterol en sangre.

La distinción entre datos discretos y continuos es importante para decidir qué método de análisis estadístico utilizar, ya que hay métodos que suponen que los datos son continuos (12).

Consideremos, por ejemplo, la variable edad. Edad es continua, pero si se la registra en años resulta ser discreta. En estudios con adultos, en que la edad va de 20 a 70 años, por ejemplo, no hay problemas en tratarla como continua, ya que el número de valores posibles es muy grande. Pero en el caso de niños en edad preescolar, si la edad se registra en años debe tratarse como discreta, en tanto que si se la registra en meses puede tratarse como continua.

Del mismo modo, la variable número de pulsaciones/min es una variable discreta, pero se la trata como continúa debido al gran número de valores posibles. Los datos numéricos (discretos o continuos) pueden ser transformados en categóricos y ser tratados como tales. Aunque esto es correcto no necesariamente es eficiente y siempre es preferible registrar el valor numérico de la medición, ya que esto permite:

- Analizar la variable como numérica ⇒ Análisis estadístico más simple y más potente
- Armar nuevas categorías usando criterios diferentes.

Sólo en casos especiales es preferible registrar datos numéricos como categóricos, por ejemplo, cuando se sabe que la medición es poco precisa (número de cigarrillos diarios, número de tazas de café en una semana).

Representaciones gráficas

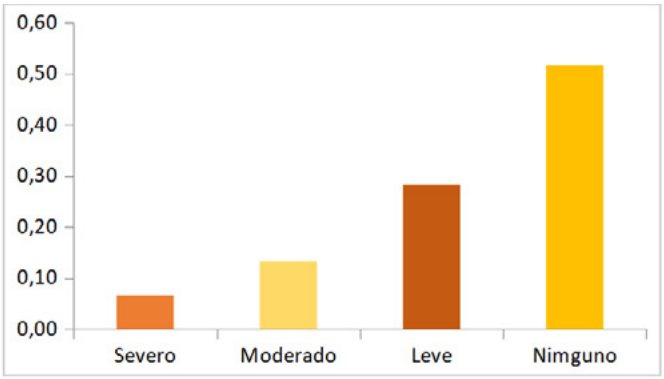
Si las tablas de distribución de frecuencias proporcionan información sobre el comportamiento de la variable a través del estudio de la distribución de las observaciones entre las diferentes categorías de clasificación, las representaciones gráficas la complementan de forma eficaz, proporcionando una imagen que permite extraer conclusiones de forma rápida acerca de la misma. Dependiendo del tipo de variable será más oportuno utilizar un tipo de representación u otra.

Gráficas o diagramas de barras

Los diagramas de barras son adecuados para representar variables cualitativas y cuantitativas discretas. En estos diagramas se representan las categorías de la variable en el eje horizontal y sus frecuencias (absolutas o relativas) en el eje vertical (3). Para cada categoría de la variable se construye un rectángulo de anchura constante y altura proporcional a la frecuencia.

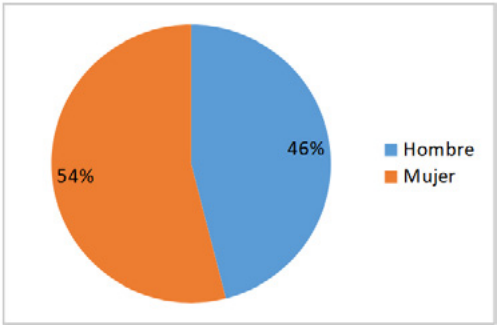
En la figura 5 se muestra la distribución de frecuencias relativas de la tabla 4 relativa a las mediciones de dolor percibido.

Figura 5. Gráfica de barras de las frecuencias relativas



Fuente: Clifford y Taylor (6)

Figura 6. Diagrama de sectores para la variable sexo



Fuente: Elaboración propia

El diagrama de sectores (figura 6) y el diagrama de barras (figura 5) suelen representar frecuencias absolutas o relativas y están especialmente indicados para variables cualitativas. Sin embargo, en el caso de variables

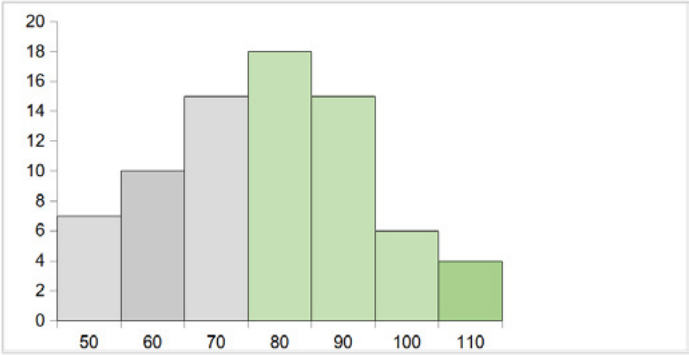
cuantitativas discretas de pocos valores, podrían ser utilizados con la misma eficacia. Incluso en el caso de variables cuantitativas continuas podría utilizarse el diagrama de sectores si previamente la variable ha sido agrupada en un número relativamente reducido de intervalos.

Histograma y polígonos de frecuencia

El histograma es el principal método gráfico para la representación de variables cuantitativas continuas. Los histogramas pueden representar a las frecuencias absolutas o relativas, dependiendo de la ubicación que se les dé a éstas sobre el eje vertical del plano cartesiano. De esta manera se obtiene el histograma de frecuencias absolutas o el histograma de frecuencias relativas.

El histograma muestra, en este caso en la figura 7, un comportamiento de la variable nivel de colesterol, en el que la mayoría de las observaciones se concentran en la zona central de la distribución, disminuyendo de forma progresiva la frecuencia de observaciones a uno y otro extremo (6).

Figura 7. Histograma de frecuencia



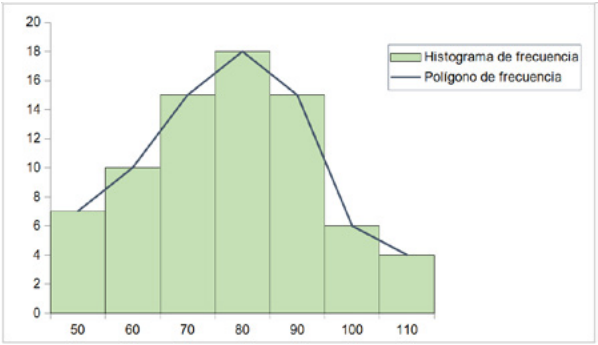
Fuente: Clifford y Taylor (6)

Polígono de frecuencia

El polígono de frecuencias es un conjunto de líneas sobre un plano cartesiano que representan el comportamiento de la característica en la población. Al igual que el histograma, el polígono se aplica a la variable cuantitativa continua.

Los polígonos son particularmente convenientes cuando se desea comparar dos o más distribuciones, se construyen en forma similar a los histogramas, excepto que, en lugar de poner una barra sobre cada intervalo, se coloca un punto a la altura apropiada del eje y. En el caso de los polígonos de frecuencias y de frecuencias relativas, el punto se coloca en el punto medio del intervalo, en tanto que en las distribuciones acumulativas el punto se coloca en el límite real superior del intervalo. Estos puntos se conectan luego con líneas rectas que se unen al eje x en el extremo inferior, y en los polígonos de frecuencias y de frecuencias relativas, con los extremos superiores de la distribución. Para los polígonos de frecuencias y de frecuencias relativas, los puntos en los que la línea hace contacto con el eje x corresponden a los que serían los puntos medios de un intervalo adicional en cada extremo de la distribución (figura 8).

Figura 8. Polígono de frecuencia



Fuente: Clifford y Taylor (6)

Ojivas o polígonos de frecuencias acumuladas

La ojiva representa el comportamiento acumulado de las unidades de investigación en relación a la variable analizada. Al igual que en los polígonos, las ojivas pueden ser construidas con las frecuencias absolutas o relativas (figura 9)

Figura 9. Polígono de frecuencias relativas acumuladas

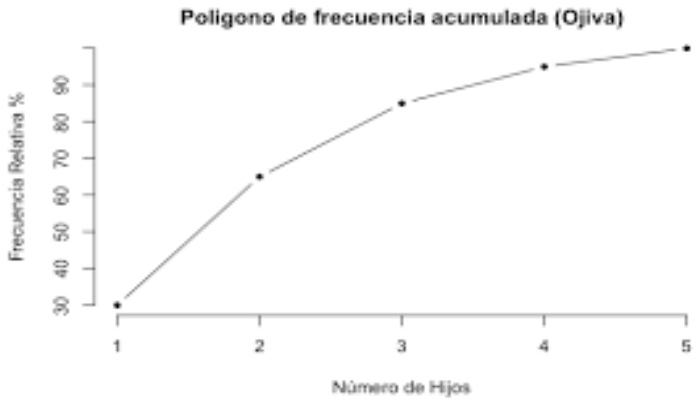
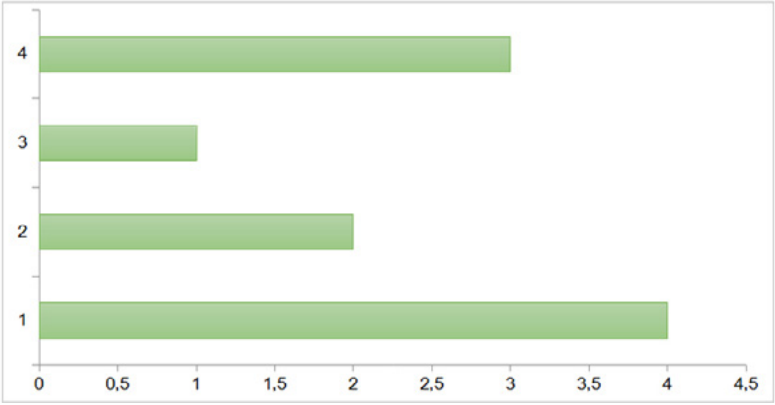


Diagrama de barras

El diagrama de barras es de las gráficas más utilizadas en los diferentes tipos de informes debido a que dan a conocer de forma fácil y sencilla las características de un grupo de elementos de una muestra o una población, especialmente cuando están asociadas a variables cualitativas o cuantitativas discretas. El diagrama de barras consiste en líneas gruesas que constituyen rectángulos de anchura variable que representan los valores que toma la variable, y de longitud definida por las frecuencias absolutas o relativas. Las barras se construyen de forma horizontal o vertical y cada una puede ser representada con frecuencias absolutas o relativas

Figura 10. Diagrama de barras horizontal



Fuente: Clifford y Taylor (6)

Gráficas de tallo y hojas

Este gráfico tiene la ventaja de reflejar los datos originales de la muestra, a la vez que permite visualizar la distribución de frecuencias. Su principal ventaja sobre el histograma es que conserva los valores de la variable mostrada.

En primer lugar, para cada observación de la variable, se separa el último dígito significativo (hoja) de los restantes dígitos del valor de la variable (tallo). A continuación, todos los posibles tallos se colocan ordenados en una misma columna. Finalmente, para cada valor de la variable, se coloca su hoja a la derecha del tallo correspondiente. Las hojas de un mismo tallo suelen colocarse en orden creciente. El resultado se conoce con el nombre de gráfico de tallo y hojas. Los siguientes pasos describen su construcción. A continuación, se muestra los pasos para su construcción:

3. Divida cada observación en un componente de “tallo” y “hoja” como se describe a continuación.

4. Elabore una lista de los componentes de tallo del valor más pequeño al más alto, como se haría en el eje x de un histograma
5. Coloque los componentes de hoja asociados con cada tallo encima del tallo en orden ascendente

- Ejemplo: La figura 11 muestra el gráfico de tallo y hojas del colesterol HDL en los 100 controles de un estudio con datos para esta variable. Los 2 valores más bajos del colesterol HDL son 0,21 y 0,26 mmol/l, cuyo tallo común es 0,2 y sus respectivas hojas son 1 y 6, que aparecen a la derecha de la primera línea del gráfico. El siguiente tallo es 0,3, que no tiene ninguna hoja ya que no hay valores entre 0,30 y 0,39 mmol/l, y lo mismo sucede con el tallo 0,4. En el tallo 0,5 hay una hoja igual a 7, que corresponde al valor 0,57 mmol/l. En el tallo 0,6 hay 5 hojas (35558), que corresponden a los 5 valores del colesterol HDL entre 0,60 y 0,69 mmol/l y que son 0,63, 0,65, 0,65, 0,65 y 0,68 mmol/l. El resto de los tallos se interpreta de la misma manera. A partir de este gráfico resulta sencillo calcular los cuantiles; así, por ejemplo, la mediana se obtendría como la media de los valores ordenados en las posiciones 50 y 51, (1,10 + 1,12) /2 = 1,11 mmol/l.

Figura 11. Gráfico de tallo y hoja

Frecuencia	Tallo	Hoja
2	0,2	16
0	0,3	
0	0,4	
1	0,5	7
5	0,6	35558
3	0,7	467
2	0,8	002344455579
3	0,9	0013334566779
3	1	0111123455559
9	1,1	023456789

5	1,2	000023356689999
7	1,3	1223778
6	1,4	345789
6	1,5	133689
2	1,6	44
2	1,7	34
2	1,8	36
1	1,9	0
1	2	9

Fuente: Clifford y Taylor (6)

Para finalizar, es importante mencionar lo siguiente:

1. las gráficas de tallo y hojas son más efectivas con conjuntos de datos relativamente pequeños
2. en general, las gráficas de este tipo no se utilizan para mensajes difundidos masivamente, como la publicación de informes de investigación. Más bien, los investigadores los emplean de manera informal para comprender sus datos
3. las gráficas de tallo y hojas pueden ser más complejas que la mostrada aquí. Por ejemplo, quizá las hojas consistan en dos o más dígitos y los tallos podrían estar agrupados en forma similar a los histogramas agrupados.
4. debe señalarse que los tallos por lo general se colocan verticalmente, con las hojas formando renglones. Esta norma no se siguió aquí para enfatizar la semejanza del gráfico con el histograma.

Tabulación de datos binarios o cruzados

En la mayoría de los estudios estadísticos se emplea el análisis unidimensional para interpretar su comportamiento de forma aislada o individualmente. Sin embargo, los vínculos que tienen las diferentes personas,

objetos o fenómenos, facultan el establecimiento de relaciones entre las características o variables que ellas presentan. Estas relaciones permiten analizar simultáneamente el comportamiento de dos variables, ya sean cualitativas o cuantitativas, usando para ello la tabulación cruzada o tablas de contingencia

Tabla de contingencia

Analizar la distribución de una variable con relación a otra u otras es una tarea corriente en salud pública, vinculada, la mayoría de las veces, a la búsqueda de un patrón que indique la relación, (o la falta de ella) entre las variables estudiadas. Este es un proceso clave en la identificación de las posibles causas de los problemas de salud, y también de factores que, aun cuando no puedan ser finalmente considerados causales, resulten estar asociados a estos daños y constituyan importantes elementos prácticos para la identificación de grupos con mayores riesgos de padecer determinado daño.

El estudio de la influencia de una variable (variable independiente) sobre la forma en que se modifica otra (variable dependiente) es conocido como análisis bivariado; y será multivariado cuando el estudio evalúe de forma simultánea el efecto sobre una variable dependiente de dos o más variables independientes.

Las tablas de contingencia (tablas de doble entrada) son una herramienta fundamental para este tipo de análisis. Están compuestas por filas (horizontales), para la información de una variable y columnas (verticales) para la información de otra variable. Estas filas y columnas delimitan celdas donde se vuelcan las frecuencias de cada combinación de las variables analizadas.

Ejemplo: investigaciones recientes han determinado que es posible que un índice cintura-cadera (ICC), definido como el cociente entre el perímetro de la cintura y el de la cadera, elevado se asocie a la aparición de ciertas patologías, como la diabetes y enfermedades cardiovasculares, de una manera más clara que el índice de masa corporal (IMC) elevado. Supongamos que, con el objeto de apoyar esa teoría, se analiza una muestra de n = 252 varones de más de 40 años que son clasificados en función de su ICC como normales, si ICC ≤ 0.94, o con cuerpo de manzana, si ICC > 0.94. Por otra parte, son también valorados médicamente distinguiendo entre sanos, diabéticos y enfermos cardiovasculares. Ambas clasificaciones se recogen de manera simultánea la siguiente tabla de contingencia (tabla 6).

Tabla 6. Tabla de contingencia para las variables tipo de ICC y estado de salud

		Estado de Salud			
	2 x 3	**Sano**	**Cardio**	**Diabetes**	**Total**
Tipo de ICC	Normal	114	22	20	156
	Manzana	52	28	16	96
	Total	**166**	**50**	**36**	**252**

Fuente: Clifford y Taylor (6)

En este caso se distinguen r = 2 categorías (filas) diferentes en la variable tipo de ICC y s = 3 categorías (columnas) diferentes en la valoración médica, por lo que decimos que se trata de una tabla tipo 2 × 3. En los márgenes derechos e inferior de la tabla aparecen las frecuencias que denominaremos marginales, que corresponderían a un estudio por separado de las variables ICC y valoración, respectivamente. Las 6 frecuencias (2 × 3) que aparecen en el interior de la tabla pueden denominarse conjuntas o, también, observadas. Se denotan mediante Oij , donde el subíndice i hace referencia a las filas y el j a las columnas. Así, por ejemplo, O12 se entiende

como la frecuencia observada en la fila 1 y columna 2, es decir, con los datos del ejemplo, se está hablando del número de individuos con ICC normal y enfermedad cardiaca. Es obvio que la suma de frecuencias observadas de una misma fila es la frecuencia marginal que aparece en la columna derecha, y que la suma de frecuencias observadas en una misma columna es la frecuencia marginal que aparece en la fila de abajo. La suma total de las frecuencias conjuntas coincide con las de las marginales, tanto por filas como por columnas, y es el tamaño de muestra n = 252.

Tablas de asociación: exposición–enfermedad

Tabla 2x2

Las tablas 2x2 simples (de un único estrato) permiten el análisis de 2 variables dicotómicas: típicamente, una variable independiente (exposición) y una variable dependiente (enfermedad). Debe advertirse que esta es la situación más común y que es por ello que se usan las denominaciones exposición y enfermedad, pero podría tratarse de otra situación como la de un ensayo clínico, por ejemplo, en la cual, en lugar de dos niveles de exposición se tuviera dos tratamientos y en lugar de enfermedad tuviéramos dos posibles desenlaces. Hay cuatro opciones de tablas 2x2 destinadas a cuatro diseños de estudios epidemiológicos:

- **Estudio transversal**

Los estudios transversales examinan la prevalencia de enfermedades y problemas de salud y también de conocidos o potenciales factores de riesgo y/o protección. Se tratan básicamente de una imagen fotográfica de la población, o de una muestra de ella, en la que se explora, a nivel individual y en forma simultánea, la presencia o ausencia (o niveles) de una o más variables independientes y de una o más variables potencialmente dependientes de las primeras. Si bien la imagen de una fotografía da la

sensación de que en un estudio de este tipo la información se recolecta en un instante (un día o pocos días), la recolección de datos puede ser más prolongada (semanas o meses). Sin embargo, la información de cada individuo seguirá siendo una foto.

- **Estudio de cohortes**

Los estudios de cohortes sustentan su estrategia de análisis en el seguimiento en el tiempo de dos o más grupos de individuos que han sido divididos según el grado de exposición a un determinado factor (corrientemente en 2 grupos: expuestos y no expuestos). Al inicio, ninguno de los individuos incluidos en ambos grupos tiene la enfermedad o daño en estudio y para responder a la pregunta acerca de si la exposición influye en el desenlace habrá de compararse la incidencia de "nuevos casos" entre ambos grupos. Estas incidencias pueden ser calculadas de dos formas:

✓ Como número de casos nuevos en relación a la población que integra la cohorte (incidencia acumulada)
✓ Considerando el período que cada individuo permaneció en el grupo (tasa de incidencia o densidad de incidencia)

La incidencia acumulada es más sencilla de calcular porque como denominador solo se requiere el número de individuos que se incluyó en cada grupo. Sin embargo, la tasa de incidencia es una medida más precisa, ya que considera el momento en que se producen los casos y los períodos de seguimiento de los individuos, que típicamente no son iguales para todos los sujetos.

- **Estudio de casos y controles**

En los estudios de casos y controles los sujetos incluidos proceden típicamente de dos grupos, según sean casos (con la enfermedad o daño en estudio) o controles (sin el daño en cuestión). La idea básica

es comparar los antecedentes de los enfermos de una población con los de los sanos de la misma población. Se trata de poner de manifiesto posibles diferencias en las exposiciones que expliquen, al menos parcialmente, la razón por la que unos enfermaron y otros no. En el análisis se comparan las exposiciones de los casos con las de los controles, y los resultados son presentados usando los llamados odds (cociente entre la probabilidad de enfermar y la probabilidad de no enfermar) y la razón de odds de adquirir una enfermedad entre expuestos y entre no expuestos (odds ratio, OR).

Tablas 2x2 estratificadas

La relación entre un factor de riesgo (supuesto o real) y un daño es en ocasiones modificada por la presencia de un tercer factor. Esta situación, conocida como efecto de confusión, podría definirse como la que producen aquellos factores que, estando relacionados con el factor de riesgo en estudio, condicionan la aparición del daño (siempre que no se trate de un factor que se halle en el trayecto causal que va del factor de riesgo al daño). Así, por ejemplo, la relación directa del consumo diario de comprimidos de β-carotenos y la prevención de las enfermedades coronarias será usualmente distorsionada por la presencia de otros factores que se encuentran vinculados a la actitud preventiva de quien toma suplementos. Seguramente, entre quienes toman esta medicación, habrá una menor proporción de fumadores y desarrollarán mayor actividad física que los que no la toman. Como estos factores tienen un efecto protector sobre la enfermedad coronaria, el efecto en la reducción del daño será resultado de la acción combinada de estos factores. De no repararse en esto, se estaría atribuyendo solo al consumo de β-carotenos una acción protectora mayor a la real.

Tablas 2XN simples

Las tablas 2xN simples (de un único estrato) permiten el análisis de una variable categórica (variable independiente que mide los niveles de exposición) y una variable dicotómica (variable dependiente que señala la presencia o no del daño). Como en el caso de las tablas 2x2, se podrá optar por tres formatos de tablas según se esté analizando:

- ✓ estudio transversal
- ✓ estudio cohorte
  - Para tasas de incidencia
  - Para incidencia acumulada
- ✓ Estudio de casos y controles

Este tipo de tablas permite calcular las prevalencias, incidencia u odds (según el tipo de estudio) para cada nivel de exposición y calcula la razón de las prevalencias, tasas de incidencia u odds ratio, utilizando por defecto como valor de referencia el nivel 1 de exposición. El nivel de referencia puede ser seleccionado y, si bien en general la elección es "natural", se deberá considerar que es más fácil analizar las razones y las tasas cuando se utiliza como nivel de referencia al nivel con menor prevalencia o incidencia.

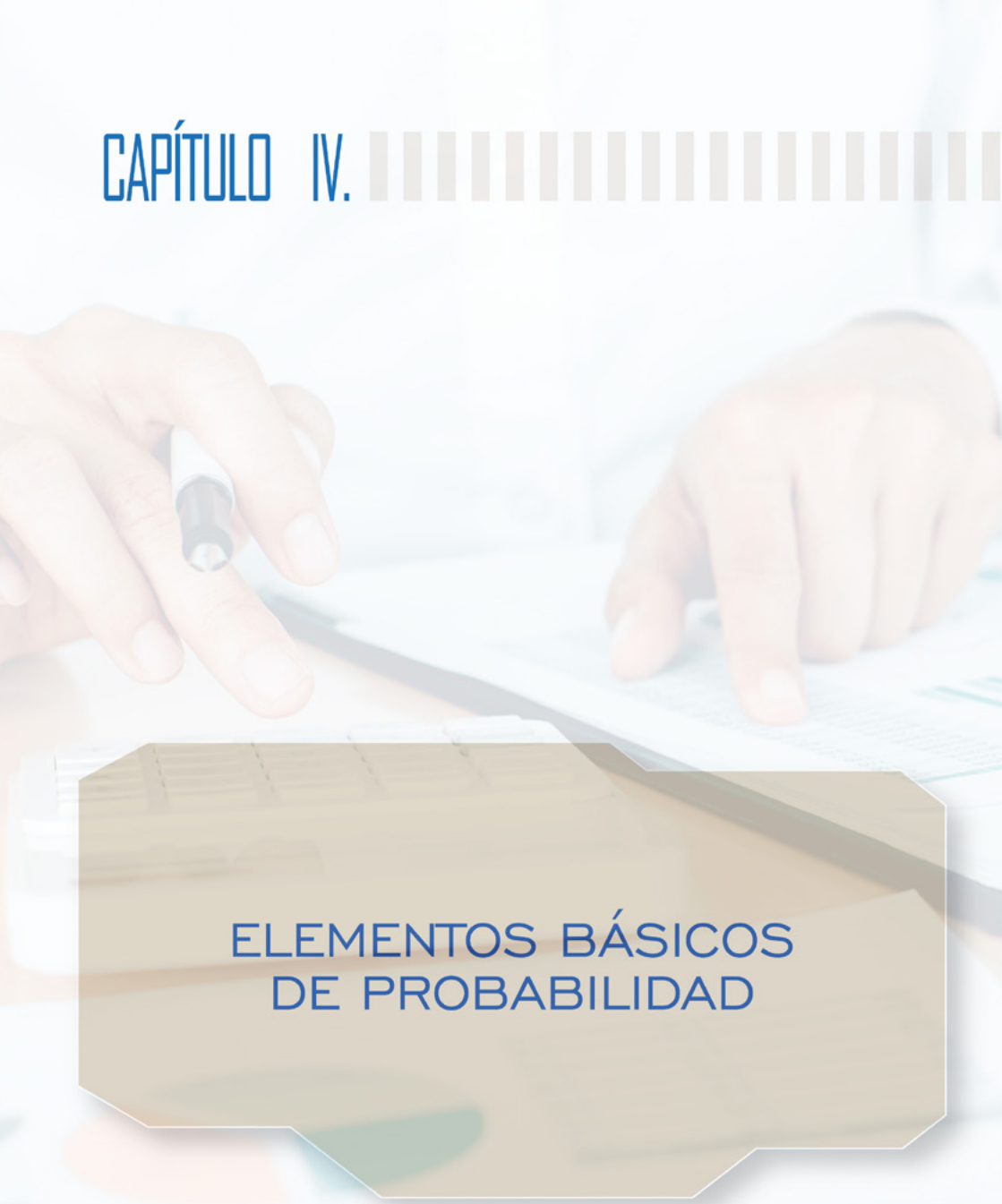
Tablas 2XN estratificadas

La estratificación de las tablas 2xN permite incorporar otra variable o factor para analizar si la relación entre la exposición y el daño cambia según las diferentes categorías de la variable por la que se está estratificando. También aquí se podrá optar por tres formatos de tablas según se esté analizando un estudio transversal, de cohortes, o de casos y controles, y deberá definirse un nivel de referencia para el cálculo de las razones de prevalencia, riesgos relativos u odds ratio, respectivamente.

Tablas de frecuencias relativas

Para el cálculo de las frecuencias relativas de cada celda, se divide el valor de cada celda entre el número total de datos (n) y multiplicar el resultado por cien para expresarlo en porcentaje, para obtener los porcentajes de fila, se divide cada frecuencia entre su respectivo total de la fila. Los porcentajes de columna se obtienen de forma similar, es decir, dividiendo cada frecuencia de la columna entre el total de cada una de ellas.

CAPÍTULO IV.



ELEMENTOS BÁSICOS DE PROBABILIDAD

CAPÍTULO IV: ELEMENTOS BÁSICOS DE PROBABILIDAD

Introducción

En este capítulo del libro se introduce al estudiante en la comprensión de la idea de azar y manejar por ende el concepto de probabilidad. El término probabilidad se refiere al estudio del azar y la incertidumbre en cualquier situación en la que varios posibles sucesos pueden ocurrir. Podríamos decir que toda experiencia cuyo resultado dependa del azar, es decir, que no podamos predecir con exactitud su resultado, es una experiencia aleatoria.

La probabilidad es la base de la estadística inferencial. La disciplina de la probabilidad proporciona métodos confiables para cuantificar las oportunidades o probabilidades de ocurrencia asociadas con los sucesos. Dicho de otra manera, es el mecanismo por medio del cual se hacen inferencias. De allí la necesidad de adquirir conocimientos básicos sobre probabilidad antes de comenzar los estudios relativos a la inferencia estadística que se estudiara en el capítulo siguiente del libro. El propósito es, entonces, familiarizar al estudiante con este tema, que resulta, en algunas ocasiones difícil de asimilar y comprender.

Conceptos básicos de la teoría de la probabilidad

Se definen, seguidamente los conceptos básicos de probabilidad y se mencionan los elementos fundamentales de la teoría de conjuntos.

Experimento

Un experimento es cualquier acción o proceso cuyo resultado está sujeto a la incertidumbre. Aunque la palabra experimento en general sugiere una situación de prueba cuidadosamente controlada en un laboratorio, aquí se usa en un sentido mucho más amplio. Por tanto, experimentos que

pueden ser interesantes incluyen lanzar al aire una moneda una o varias veces, seleccionar una carta o más de un mazo, pesar una hogaza de pan, medir el tiempo del recorrido entre la casa y el trabajo en una mañana particular, obtener tipos de sangre de un grupo de individuos, o medir las resistencias a la compresión de diferentes vigas de acero (13).

Fenómenos determinísticos y aleatorios

Los fenómenos que se presentan en la vida cotidiana se pueden clasificar en determinísticos y no determinísticos o aleatorios. Un experimento o fenómeno es determinista si se obtiene el mismo resultado cuando se repite el experimento en las mismas condiciones. Por el contrario, un experimento o fenómeno es aleatorio (o estocástico) cuando al repetir el experimento en igualdad de condiciones los resultados varían, a pesar de mantener constantes las condiciones con las que se realiza el experimento.

- Ejemplo 1:

Si lanzamos una piedra al aire, podemos afirmar con certeza que volverá a caer a la superficie de la tierra, pero no podemos saber con precisión el punto en el que caerá. Así, la caída de la piedra a la superficie de la tierra es un fenómeno determinístico, mientras que el lugar en que se producirá dicha caída es aleatorio, ya que existe incertidumbre respecto del punto preciso en el la piedra caerá.

Fenómeno o experimento aleatorio

Un fenómeno o experimento aleatorio, representado por la letra del alfabeto griego Épsilon (), es el que satisface las siguientes características:

- El experimento puede repetirse indefinidamente bajo idénticas o parecidas condiciones.

- Cualquier modificación en las condiciones iniciales de la repetición modifica completamente el resultado final del experimento.
- Se pueden conocer a priori el conjunto de los posibles resultados del experimento, pero no se puede predecir un resultado particular.
- Si el experimento se repite un gran número de veces, la proporción con que cada resultado aparece tiende a estabilizarse.

Espacio muestral de un experimento

Se entiende por espacio muestral al conjunto de todos los posibles resultados simples de una característica aleatoria. Los espacios muestrales podrán ser finitos o infinitos, según la característica aleatoria posea un número finito o infinito de posibles resultados simples y el conjunto de resultados simples que forma el espacio muestral debe ser exhaustivo y exclusivo, es decir debe contener todos los posibles resultados y sólo uno de ellos se producirá.

El espacio de probabilidades o espacio muestral es el conjunto de todos los resultados que se pueden obtener al realizar un experimento aleatorio y es frecuente representarlo por la letra Omega [Ω].

Ejemplo: El espacio muestral del experimento lanzar una moneda al aire es Ω = {C, X} y el de lanzar dos monedas Ω.= {CC, CX, XC, XX}, donde "C" representa el lado caro de la moneda y la "X" la ceca (o reverso de la moneda). El Espacio Muestral del lanzamiento de un dado equilibrado sería: Ω = {1, 2, 3, 4, 5, 6}.

El espacio muestral puede ser de distintos ***tipos*** (14):

Discreto: Si es factible contabilizar (contar) los posibles resultados:

- ***Finito:*** Número finito de resultados (Resultados al lanzar un dado, número de piezas defectuosas en un conjunto de 20 unidades...)

- *Infinito:* Número infinito de resultados (Número de veces que lanzamos una moneda hasta salir cara, número de defectos que encontramos en una muestra de 20 unidades).
- *Exhaustivo y exclusivo:* El conjunto de resultados simples que forma el espacio muestral debe ser exhaustivo y exclusivo, es decir debe contener todos los posibles resultados y sólo uno de ellos se producirá.
- *Continuo:* Pueden salir cualquier valor de la recta real (La altura de los pacientes que acuden a consulta en un hospital, la longitud de una pieza, su densidad, su resistencia a la rotura, y en general cualquier aspecto del experimento que sea "medible").

Además, los espacios muestrales pueden ser univariantes, cuando los resultados simples proceden de una única característica aleatoria, o multivariantes, cuando se refieren a dos o más características consideradas simultáneamente (15).

Suceso aleatorio

Cada subconjunto del espacio muestral. Por ejemplo, en el caso del lanzamiento del dado, el suceso "A sale un número par", sería A = {2, 4, 6}. Existen distintos tipos de sucesos:

- Suceso elemental o simple: Cada uno de los posibles resultados de realizar el experimento.
- Suceso seguro: Suceso que siempre se verifica (ocurre o se presenta), es decir, es el espacio muestral.
- **Suceso imposible: Suceso que nunca se verifica, se representa por el conjunto vacío (Ø).**
- Sucesos compatibles y sucesos incompatibles: Cuando dos sucesos A y B tienen algún suceso elemental común, se les llaman sucesos compatibles. Si por el contrario no lo tienen se les denomina sucesos incompatibles.

- Ejemplo 2:

Considere un experimento en el cual cada tres vehículos que toman la salida de una autopista particular viran ya sea a la izquierda (L) o a la derecha (R) al final de la rampa de salida. Los ocho posibles resultados que constituyen el espacio muestral son: {LLL, RLL, LRL, LLR, LRR, RLR, RRL y RRR}. Así pues, existen ocho eventos simples, entre los cuales están E_1 = {LLL}, E_5 = {LRR}.

Algunos eventos compuestos incluyen:

- A = {RLL, LRL, LLR}, el evento en que sólo uno de los tres vehículos vira a la derecha,
- B = {LLL, RLL, LRL, LLR}, el evento en que a lo más uno de los vehículos vira a la derecha.
- C = {LLL, RRR}, el evento en que los tres vehículos viran en la misma dirección.

Suponga que cuando se realiza el experimento, el resultado es {LLL}. Entonces ha ocurrido el evento simple E1 y, por tanto, también comprende los eventos B y C (pero no A).

Definición de probabilidad

La probabilidad es la rama matemática que intenta determinar qué tan posible es que ocurra un resultado (suceso o evento) de un experimento aleatorio. Así pues, cuando se dice que «***la probabilidad de que ocurra un evento (***o suceso) de interés es del 45%», significa que se ha calculado que un resultado sucederá 45 veces por cada 100 experimentos que se hagan. Luego, para averiguar si el cálculo es correcto o no, se deberían realizar cien experimentos iguales y contar las veces que ocurre dicho resultado.

Definición clásica

Si preguntamos a cualquier persona que nos diga cuál es la probabilidad de obtener ***ceca*** al lanzar una moneda al aire, casi con seguridad nos contestará "un 50%". Asimismo, si consultamos cuál es la probabilidad de obtener ***el número 6*** al lanzar un dado, es muy posible que la respuesta sea "un sexto"; mientras que si preguntamos cuál es la probabilidad de obtener ***un número par***, la respuesta será un 50%. Estas respuestas intuitivas están ligadas a la definición clásica de probabilidad:

Sea Ω un espacio muestral finito que contiene N eventos simples, y sea A un evento que puede darse de n maneras distintas; es decir, que al realizar un experimento hay N resultados posibles de los cuales n son favorables al evento A. La probabilidad de que ocurra el evento A está dada por:

$$P(A) = \frac{resultados\ favorables}{resultados\ posibles} = \frac{n}{N}$$

Si relacionamos la definición precedente con la teoría de conjuntos, podemos afirmar que la probabilidad de que se dé el evento A está dada por el cociente entre la cantidad de elementos del conjunto favorables al evento A y el número de elementos del conjunto, siendo estos últimos igualmente probables.

- Ejemplo 3:

Si queremos calcular la probabilidad de obtener el número 5 en el lanzamiento de un dado, tenemos que dividir el número de resultados en los que se cumple este evento (solo una cara del dado tiene el número 5) entre el número total de posibles resultados (un dado tiene seis caras así que podemos obtener seis resultados diferentes):

$$P\ (\text{obtener el número 5}) = \frac{1}{6} = 0{,}1667 = 16{,}7\%$$

En este problema, las probabilidades de sacar cualquier cara del dado son iguales, es decir, todos los sucesos del espacio muestral son equiprobables.

- Ejemplo 4

Un individuo está por jugar a un juego en el que se lanzan dos dados equilibrados; El jugador gana un dólar ($ 1) si el resultado de la suma de los números obtenidos en ambos dados es siete. ¿Qué probabilidad tiene de ganar?

La cantidad de resultados posibles cuando se lanzan dos dados es 36 (estos resultados son igualmente probables): si el resultado del primer dado es 1, el segundo puede arrojar cualquiera de los números del 1 al 6, con lo cual ya tenemos seis resultados posibles; si el primer dado es 2, el segundo nuevamente podrá arrojar cualquier valor del 1 al 6, con lo cual ya sumamos doce resultados; y así sucesivamente hasta completar 6^2 = 36 resultados posibles.

Luego, deberíamos determinar la cantidad de resultados favorables al evento "la suma de los dados es 7": éste puede darse de seis maneras distintas (1 y 6; 2 y 5; 3 y 4; 4 y 3; 5 y 2; 6 y 1).

En la tabla 7 se resumen todos los resultados posibles, y aparecen sombreados los resultados favorables al evento:

Tabla 7. Resultados posible del lanzamiento de 2 dados

		DADO 2					
		1	2	3	4	5	6
DADO 1	1	2	3	4	5	6	7
	2	3	4	5	6	7	8
	3	4	5	6	7	8	9
	4	5	6	7	8	9	10
	5	6	7	8	9	10	11
	6	7	8	9	10	11	12

Fuente: Bacchini et.al (16)

Así, la probabilidad de que el apostador gane, está dada por el cociente entre el número de resultados favorables al suceso y el número de resultados posibles:

$$P(A)= \frac{6}{36} = \frac{1}{6} = 0,1667 = 16,67\%$$

Definición frecuentista

El concepto frecuentista de probabilidad surge debido a la existencia de fenómenos aleatorios en los cuales no se puede determinar con precisión la probabilidad clásica de cada evento simple, es decir, que no podemos precisar cuántos resultados favorables a un evento existen y/o cuántos resultados posibles hay.

Consideremos algunos ejemplos en los cuales no se puede determinar con precisión los casos favorables y los casos posibles: un jefe de control de calidad desea determinar la probabilidad de que un artículo sea defectuoso, un fanático está interesado en la probabilidad de que su equipo de fútbol gane o un profesor que quiere saber la probabilidad de que sus alumnos aprueben.

Para estimar la probabilidad de cada uno de esos eventos, se recurre a la segunda manera de definir a la probabilidad, utilizando la frecuencia relativa de ocurrencia de los mismos.

Sea K el número de veces que se observa un fenómeno determinado, y sea k el número de veces en que ocurre un resultado favorable al evento ***A***. La probabilidad de ocurrencia del evento ***A*** es la frecuencia relativa observada cuando el número total de observaciones crece indefinidamente:

$P(A)= \lim_{k \to \infty} \frac{k}{K}$ (La probabilidad se define como el límite de la frecuencia relativa cuando el número de repeticiones de un experimento tiende al infinito).

La gran mayoría de los fenómenos aleatorios con que nos enfrentaremos en la práctica son de este tipo, por lo cual esta definición de probabilidad será muy utilizada a lo largo de la presente obra.

- Ejemplo 5:

Consideremos un control de calidad de una empresa, en el cual se desea saber la probabilidad de que un determinado artefacto tenga una vida útil superior a las 1200 horas (hs). Para ello, el departamento de control de calidad separa 500 unidades de la producción y mide la vida útil de cada unidad. Los resultados se observan en la tabla 8:

Tabla 8. Resultado del experiemnto de la vida util

Duración en horas (en hs)	frec.abs.	frec. reL.
menos de 800	10	2%
800 a 899	40	8%
900 a 999	55	11%
1000 a 1099	70	14%
1100 a 1199	85	17%
1200 a 1299	115	23%
1300 a 1399	84	17%
1400 o más	41	8%
	500	**100%**

Fuente: Bacchini et.al (16)

$$P(A) = \frac{115+84+41}{500} = 0{,}23+0{,}17+0{,}08 = 0{,}38$$

Esta definición de probabilidad da lugar a las «pruebas de hipótesis», que serán tratadas en el próximo capítulo del libro. Consideremos el lanzamiento de un dado y supongamos que queremos detectar si el mismo está cargado. Para ello, podríamos lanzar el dado un gran número de veces y observar la frecuencia relativa de ocurrencia de cada resultado;

por ejemplo, si lanzamos el dado 600 veces, deberíamos esperar que 100 veces se dé cada uno de los resultados posibles. Sin embargo, difícilmente esto ocurra, y supongamos que el resultado 2 se dio 140 veces. Lo que se pretende al realizar un test de hipótesis, es probar si la evidencia empírica es suficiente como para afirmar que el dado está efectivamente cargado a favor del número 2, o si la observación de una cantidad elevada de dicho resultado se debió simplemente al azar propio del experimento. Continuaremos con este tema en el capítulo correspondiente a la prueba de hipótesis.

Definición subjetiva de probabilidad

La definición subjetiva de probabilidad está relacionada con el grado de creencia que tiene quien lleva a cabo un experimento respecto de la probabilidad de ocurrencia del mismo.

Así, por ejemplo, al lanzar un nuevo producto al mercado, un gerente de ventas puede creer que el mismo tendrá un 70% de aceptación en el público, es decir, que la probabilidad (subjetiva) de que un individuo acepte el producto es de 0,7. Esta probabilidad suele llamarse también probabilidad a priori ya que refleja el grado de creencia antes de que se realice cualquier prueba empírica. Las probabilidades a priori suelen modificarse luego mediante algún tipo de experimento como, por ejemplo, una encuesta para ver la aceptación que podría tener el producto. Una vez que el experimento se realiza, se modifican las probabilidades a priori para obtener las probabilidades a posteriori, las cuales serán utilizadas para tomar decisiones.

Teoría de conjuntos, probabilidad y operaciones con sucesos

La teoría de conjuntos es una herramienta útil en el estudio del azar. Su estudio facilita la comprensión de los resultados aleatorios. Su uso en la

teoría de la probabilidad demanda el conocimiento previo de algunos conceptos frecuentes:

Se denomina conjunto a la colección de observaciones o elementos definidos, diferentes y mutuamente exclusivos.

Un conjunto se clasifica como finito o infinito según el número de sus elementos. Cuando el conjunto es finito se pueden enlistar todos los elementos del mismo. Sin embargo, algunos universos finitos son tan grandes que llegan a ser incontables.

Para designar un conjunto se utiliza una letra mayúscula, mientras que para los elementos se acostumbran letras minúsculas o números agrupados entre llaves. Por ejemplo: el conjunto A está formado por los elementos 1, 2, 3, 4, 5, o $A = \{1, 2, 3, 4, 5\}$, mientras que el conjunto, B está compuesto por los elementos a, e, i, o, u, o $B = \{a, e, i, o, u\}$.

Un suceso o evento es simplemente un conjunto, así que las relaciones y los resultados de la teoría elemental de conjuntos pueden ser utilizados para estudiar sucesos. Se utilizarán las siguientes operaciones para crear sucesos nuevos a partir de los ya dados.

Probabilidad de la unión de dos sucesos

La unión de dos sucesos A y B es el conjunto de sucesos que están en A, en B o en ambos. La unión de dos sucesos se expresa con el símbolo $\cup$, así pues, la unión de los sucesos A y B se escribe $A \cup B$.

- Ejemplo 6:

Si se considera el lanzamiento de un dado y se definen los eventos $A = \{1, 2, 3\}$ y $B = \{2, 4, 6\}$, entonces $A \cup B = \{1, 2, 3, 4, 6\}$.

La probabilidad de la unión de dos sucesos es igual a la probabilidad del primer suceso, más la probabilidad del segundo suceso, menos la probabilidad de la intersección de los sucesos.

Es decir, la fórmula de la probabilidad de la unión de dos sucesos es

$$P(A \cup B) = P(A)+P(B)-P(A \cap B)$$

No obstante, si los dos sucesos son incompatibles, la intersección entre los dos sucesos es nula. En consecuencia, la probabilidad de la unión de dos sucesos incompatibles se calcula sumando la probabilidad de ocurrencia de cada suceso.

A y B son incompatibles → $P(A \cap B) = 0$. En consecuencia:

$$P(A \cup B) = P(A)+P(B)$$

Probabilidad de intersección de los sucesos:

La intersección de los sucesos A y B está formada por todos los sucesos que son de A y de B a la vez, se expresa mediante el símbolo ⋂. Así pues, la intersección de los sucesos A y B se escribe $A \cap B$.

- Ejemplo 7: Si se considera el lanzamiento de un dado y se definen los eventos A = {1, 2, 3} y B = {2, 4, 6}, entonces $A \cap B = \{2\}$.

La probabilidad de la intersección de dos sucesos es igual a la probabilidad de que ocurra un suceso multiplicado por la probabilidad condicional de que ocurra el otro suceso dado el primer suceso. Por lo tanto, la fórmula de la probabilidad de la intersección de dos sucesos es:

$$P(A \cap B) = P(A) \cdot P(B|A) = P(B) \cdot P(A|B)$$

No obstante, si los dos sucesos son independientes, significa que la probabilidad de ocurrencia de un suceso no depende de si ocurre el otro suceso.

En consecuencia, la fórmula de la probabilidad de la intersección de los dos sucesos independientes es la siguiente:

$$P(A \cap B) = P(A) \cdot P(B)$$

En ocasiones A y B no tienen resultados en común, por lo que la intersección de A y B no contiene resultados.

Sea que ϕ denote el evento nulo (el evento sin resultados). Cuando $A \cap B = \phi$, se dice que A y B son eventos mutuamente excluyentes o disjuntos.

- Ejemplo 8:

En una pequeña ciudad hay tres distribuidores de automóviles: un distribuidor GM que vende Chevrolet y Buick, un distribuidor Ford que vende Ford y Lincoln, y un distribuidor Toyota que vende Yaris y Corolla. Si un experimento consiste en observar la marca del siguiente automóvil vendido, entonces los eventos {Chevrolet, Buick}, {Ford, Lincoln} y {Yaris, Corolla}, son mutuamente excluyentes porque el siguiente automóvil vendido no puede ser a la vez un producto GM, un producto Ford y un producto Toyota).

Probabilidad de la diferencia de dos sucesos

La probabilidad de la diferencia de dos sucesos se refiere a la probabilidad de que ocurra un suceso sin que el otro suceso ocurra a la vez.

Así pues, la probabilidad de la diferencia de los sucesos A-B es igual a la probabilidad del suceso A menos la probabilidad de la intersección entre el suceso A y el suceso B. De modo que la fórmula de la probabilidad de la diferencia de dos sucesos es la siguiente:

$$P(A - B) = P(A) - P(A \cap B)$$

Complemento de un conjunto (Suceso contrario)

Complemento de un conjunto (A^C), es el conjunto de todos los elementos del espacio muestral que no pertenecen al evento A.

► Ejemplo 9:

En el experimento de lanzar un dado, si A = {6} el suceso contrario de A, es no obtener seis, es decir, obtener 1, 2, 3, 4 o 5. Como el contrario de A, incluye todos los sucesos elementales que no están en el suceso A, la suma de ambos da lugar a todo el espacio muestral. Y eso significa, por tanto, que la probabilidad del contrario de, es decir A^C, es uno menos la probabilidad de A. Esta sencilla fórmula, también llamada regla del complemento, convierte problemas complicados en otros más sencillos.

► Ejemplo 10:

En el experimento de lanzar un dado, si A= obtener un par. Para calcular P (A^C), tenemos: **A** = {2, 4, 6}, su complemento es A^C = {1, 3, 5}.

Probabilidad del suceso contrario

La fórmula de probabilidad del suceso contrario es: $P(A^C) = 1- P(A)$.

► Ejemplo 11:

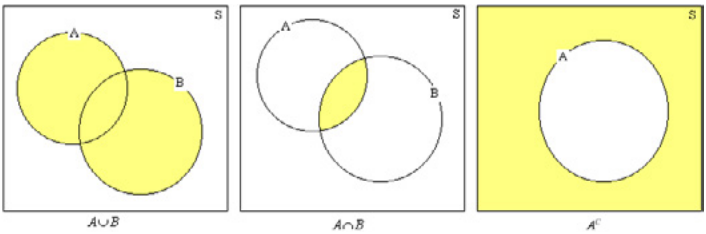
Una urna contiene ocho bolas rojas, cinco amarillas y siete verdes. Si se extrae una bola al azar: 1) calcular la probabilidad de que sea roja: P(A = Bola roja) = $\frac{8}{20}$ = 0,4

► Ejemplo 12:

Una urna contiene ocho bolas rojas, cinco amarillas y siete verdes. Si se extrae una bola al azar: 2) calcular la probabilidad de que no sea roja: $P(A^C) = 1- \frac{8}{20} = 1 - 0,4 = 0, 6$

En el diagrama de Venn (figura 12), se puede observar algunas de las operaciones con los sucesos y definiciones expuestas anteriormente: a) unión, b) intersección, y c) suceso complementario.

Figura 12. Diagrama de Venn



Fuente: Bacchini et.al (16)

Probabilidad condicional e independencia

En este apartado se analiza la influencia que tiene sobre la ocurrencia de un evento determinado la información que se posee sobre la ocurrencia de otro evento relacionado con el mismo, si es que existe tal influencia.

Cuando se trabaja con fenómenos aleatorios, muchas veces podemos contar con cierta información que modificaría nuestra estimación de la probabilidad del mismo. En estos casos, se dice que la probabilidad del evento en cuestión está condicionada a la ocurrencia de otro evento.

Probabilidad condicionada

La probabilidad condicionada, también llamada probabilidad condicional, es una medida estadística que indica la probabilidad de que ocurra un evento A si otro evento B ha sucedido. La probabilidad condicionada P (A|B) se refiere a cuánto de probable es que suceda el evento A una vez que ya se ha producido el evento B.

La probabilidad condicional del evento A dado el evento B es igual a la probabilidad de la intersección entre el evento A y el evento B dividido por la probabilidad del evento B. Por lo tanto, la fórmula de la probabilidad condicionada es la siguiente:

$P(\frac{A}{B}) = \frac{p(A \cap B)}{P(B)}$ siendo la $P(B) > 0$

- Ejemplo 13:

Considérese el lanzamiento de dos dados, de forma no simultánea. El resultado del primero de ellos se denotará por d_1 y el resultado del segundo por d_2. La probabilidad de que la suma sea 3 está dada por:

$P (d_1 + d_2 = 3 / d_1 = 2) = 2/36 = 1/18 = 0{,}0555 = 5{,}5\%$

Sin embargo, si sabemos que el resultado del primer dado es 2, la única manera de que la suma sea 3 es que el resultado del segundo sea 1, por lo tanto, la probabilidad será: 1/6 = 0,1666 = 16,6%

El ejemplo anterior se resolvió de manera directa utilizando la definición clásica de probabilidad, podría resolverse utilizando la fórmula de la probabilidad condicional de la manera siguiente:

$P (d1 + d2 = 3 / d1 = 2 = ¿ = (1/36)/(1/6) = 0{,}1666 = 16{,}6\%$

Puede observarse que el condicionamiento es equivalente a "recortar" el espacio muestral: se eliminan del espacio muestral aquellos eventos que resultan imposibles de acuerdo a la información con la que contamos.

Eventos estadísticamente independientes

Lógicamente, puede suceder que tengamos información sobre la ocurrencia de un evento determinado B, y sin embargo la probabilidad de ocurrencia del evento A no se vea alterada. Esto quiere decir, que la ocurrencia de B no tiene ninguna influencia sobre el evento A, es decir, que los eventos son estadísticamente independientes.

Dos eventos A y B son estadísticamente independientes, si la ocurrencia de uno no afecta la probabilidad de ocurrencia del otro, es decir que: $P(A/B) = P(A)$

De las definiciones de probabilidad condicional y eventos independientes, se desprende la regla del producto de probabilidades de eventos independientes.

Si ***A*** y ***B*** son dos eventos estadísticamente independientes, entonces la probabilidad conjunta es igual al producto de las probabilidades marginales: $P(A \cap B) = P(A).\ P(B)$

Se destaca que **la independencia es una relación simétrica** entre eventos, esto quiere decir que, si ***A*** es independiente de ***B***, entonces ***B*** es independiente de ***A***.

- Ejemplo 14:

Considérese el lanzamiento de dos dados y los siguientes eventos: A_1= *"el resultado del primer dado es dos"* y A_2= *"el resultado del segundo es tres"*. La probabilidad marginal de cada uno de ellos es:

$P(A1) = P(d1=2) = 1/6$ y $P(A2) = P(d=3) = 1/6 = 0{,}166,$

entonces:

La probabilidad conjunta es: $P(A1 \cap A2) = 1/36 = 0{,}027$

Tablas de contingencias y cálculo de probabilidad

Imaginemos una población conformada por 20 individuos con dos características asociadas a todas ellas. Cada una de las 20 personas se caracteriza por ser fumador (F), por no ser fumador (F^C), por tener alguna enfermedad en particular (E), o por no tener la enfermedad (E^C).

Las sumas utilizadas para los cálculos de probabilidad pueden resumirse fácilmente en una tabla de contingencia, como se muestra en la tabla 9.

Tabla 9. Tabla de contingencia que muestra las frecuencias de población

	Enfermo	No Enfermo	
Fuma	9	3	12
No Fuma	2	6	8
	11	9	

Fuente: Clifford y Taylor (6)

Los números en las cuatro celdas son recuentos de las personas que cumplen con los dos criterios indicados. Eso es, había nueve personas que fumaban y tenían la enfermedad, tres que fumaban y no tenían la enfermedad, dos que no fumaban y tenían la enfermedad y seis que no fumaban y no tenían la enfermedad. Los valores en los márgenes de la tabla dan el conteo total de la característica indicada. Así, 12 personas fumaban, ocho no fumaban, 11 tenían la enfermedad y nueve no tenían la enfermedad.

- Ejemplo 15:

Utilice los datos de la tabla de contingencia para calcular las siguientes probabilidades.

P (F^c), P (E), P (FE), P (FE^c), P (F^c E^c), P (FỤE), P (E^c / F), y P (F^c / E^c).

P (F^c) es la probabilidad de seleccionar a una persona que no fuma. Si utilizamos el conteo apropiado de la tabla de contingencia nos da 8/20 = 40.

De igual forma, ***P*** (E) es la probabilidad de seleccionar a alguien con la enfermedad y es 11/20 = .55.

P (FE^c) es la probabilidad de seleccionar a alguien que fuma ***y*** no tiene la enfermedad, en tanto que ***P*** (F^cE^c) es la probabilidad de seleccionar a un no fumador que no tiene la enfermedad. Éstas son, respectivamente, 3/20 0 = 0,15 y 6/20 = 0, 30.

P (F∪E^C) es la probabilidad de seleccionar a alguien que fuma ***o*** que no tiene la enfermedad. El número de personas que cumplen con este requerimiento es 9 + 3 + 6 = 18, lo que nos da una probabilidad de 18/20 = 0,90.

Dado que 9 + 2 + 6 = 17 personas son no fumadores ***o*** tienen la enfermedad, P (F^C ∪ E) = 17/20 =0, 85.

La probabilidad condicional P (E^C/F) es la probabilidad de seleccionar a alguien que no tiene la enfermedad, siendo que él o ella es un fumador. Dicho de otra forma, es la probabilidad de seleccionar a alguien que no tiene la enfermedad si la selección se hace deliberadamente entre los fumadores. Puesto que hay 12 fumadores, tres de los cuales no tienen la enfermedad, la probabilidad de seleccionar a una persona no enferma entre los fumadores es 3/12 = 0,25

Finalmente, P (F^C ∪ E^C) es la probabilidad de un no fumador, de aquellos que no tienen la enfermedad. Otra vez, si utilizamos los datos de la tabla, hay un total de nueve personas sin enfermedad. Ya que seis de ellas son no fumadores, la probabilidad de seleccionar a un no fumador entre las personas sin enfermedad es 6/9 = 0,67.

Teorema de la probabilidad total

El teorema de la probabilidad total es una ley que permite calcular la probabilidad de un suceso que no forma parte de un espacio muestral a partir de las probabilidades condicionales de todos los sucesos de dicho espacio muestral.

El enunciado del teorema de la probabilidad total dice que dado un conjunto de sucesos $\{A_1, A_2, \ldots, A_n\}$ que forman una partición sobre el espacio muestral, la probabilidad del evento B es igual al sumatorio de los productos de la probabilidad de cada suceso $P(A_i)$ por la probabilidad condicional $P(B|A_i)$.

Por lo tanto, la fórmula del teorema de la probabilidad total es la siguiente:

$$P(B)=\sum_{i=1}^{n} P(B/¿\,Ai).P(Ai)¿$$

Enunciado del Teorema de Bayes

El teorema de Bayes dice que dado un espacio muestral formado por un conjunto de sucesos mutuamente excluyentes $\{A_1, A_2,\ldots, A_i,\ldots, A_n\}$ cuyas probabilidades no son nulas y otro evento B, se puede relacionar matemáticamente la probabilidad condicional de A_i dado el evento B con la probabilidad condicional de B dado A_i.

Así pues, la fórmula del teorema de Bayes es la siguiente:

$$P\left(\frac{A_i}{B}\right)=\frac{P\left(\frac{B}{A_i}\right).P(A_i)}{\sum_{k=1}^{n} P\left(\frac{B}{A_k}\right).P(A_k)}$$

El teorema de Bayes se usa con frecuencia en la evaluación de pruebas diagnósticas. Cuando se desarrolla una prueba diagnóstica y se comparan sus resultados con los de un patrón oro (método de referencia en el diagnóstico de la enfermedad), suelen determinarse los siguientes parámetros o características propias de la prueba diagnóstica: Sensibilidad, especificidad, o el valor predictivo positivo o el valor predictivo negativo de un test diagnóstico (17).

Axiomas y propiedades de la probabilidad

Dados un experimento y un espacio muestral Ω, el objetivo de la probabilidad es asignar a cada evento A un número P(A), llamado la probabilidad del evento A, que dará una medida precisa de la oportunidad de que A ocurra. Para garantizar que las asignaciones serán consistentes con las nociones intuitivas de la probabilidad, todas las asignaciones deberán satisfacer los siguientes axiomas (propiedades básicas) de probabilidad.

Axioma 1: Para cualquier evento A, $P(A) \geq 0$

Axioma 2: $P(\Omega) = 1$

Axioma 3: Si A_1, A_2, A_3, es un conjunto de eventos mutuamente excluyentes, entonces $P(A_1 \cup A_2 \cup A_3 \cup \ldots) = \sum_{i=1}^{n} P(Ai)$

Se podría preguntar por qué el tercer axioma no contiene ninguna referencia a un conjunto finito de eventos disjuntos. Es porque la propiedad correspondiente para un conjunto finito puede ser derivada de los tres axiomas. Se pretende que la lista de axiomas sea tan corta como sea posible y que no contenga ninguna propiedad que pueda deducirse a partir de las demás que aparecen en la lista. El axioma 1 refleja la noción intuitiva de que la probabilidad de que A ocurra sea no negativa. El espacio muestral es por definición el evento que debe ocurrir cuando se realiza el experimento (Ω contiene todos los posibles resultados), así que el axioma 2 dice que la máxima probabilidad posible de 1 está asignada a Ω. El tercer axioma formaliza la idea que si se desea la probabilidad de que al menos uno de varios eventos ocurra y dado que dos eventos no pueden ocurrir al mismo tiempo, la probabilidad de que al menos ocurra uno es la suma de las probabilidades de los eventos individuales (13).

Proposición: $P(\phi) = 0$, donde ϕ es el evento nulo (el evento que no contiene resultados en absoluto). Esto a su vez implica que la propiedad contenida en el axioma 3 es válida para un conjunto finito de eventos disjuntos.

Aplicación de la teoría de probabilidad a la ciencia médica

Valor predictivo de pruebas diagnósticas: sensibilidad y especificidad

Las pruebas diseñadas para establecer la presencia o ausencia de alguna enfermedad rara vez son perfectas. Por ejemplo, nos gustaría que una prueba médica para establecer la presencia o ausencia de alguna enfer-

medad en particular fuera positiva para quienes tengan la enfermedad y negativa para quienes no la tengan. Por desgracia, en algunas ocasiones una persona enferma recibe un resultado negativo o una persona sana obtiene un resultado positivo.

El buen o mal desempeño de una prueba en este aspecto puede evaluarse a través del cálculo de su sensibilidad, especificidad, valor predictivo positivo y valor predictivo negativo.

La sensibilidad y especificidad de una prueba diagnóstica constituyen los dos indicadores clásicos para evaluar su capacidad demarcatoria. Su definición exige distinguir entre el criterio que define teóricamente la posesión del estado patológico y los procedimientos empleados para evaluar si dicho criterio se cumple. Supongamos que se puede establecer si un sujeto posee o no cierta condición patológica (está enfermo o sano), situaciones que se denotarán por E y S, respectivamente. Simultáneamente, supondremos que existe una prueba que, aplicada sobre cierto sujeto, puede dar lugar a solo uno de dos resultados posibles: positivo o negativo, que se representarán respectivamente con las letras R+ y R-.

En principio, el resultado (R+) sobre un sujeto dado constituye un indicio de que éste tiene la condición E (es decir, de que está enfermo) y el resultado (R-) induciría a pensar que el sujeto en cuestión tiene la condición complementaria S (o sea, que no tiene la enfermedad). El grado de eficiencia inherente a una prueba diagnóstica se resume en los dos parámetros mencionados, conocidos como sensibilidad y especificidad.

La sensibilidad mide la capacidad de la prueba para detectar a un sujeto enfermo; expresa cuán «sensible» es dicho recurso diagnóstico a la presencia de la enfermedad y viene definido por la probabilidad condicional siguiente: Sensibilidad = P(R+/E). La sensibilidad es entonces la probabilidad de que la prueba identifique como enfermo a aquel que realmente lo es.

El otro parámetro, es decir, la especificidad, mide la capacidad que tiene la prueba de diagnosticar como sanos a los que efectivamente lo son. La especificidad se define como la probabilidad condicional: Especificidad = P(R-/S).

Procede sin embargo subrayar que, desde el punto de vista operativo, los conceptos que realmente interesan en relación con las pruebas diagnósticas no son la sensibilidad y la especificidad. El clínico procede en la dirección opuesta: partiendo del resultado de la prueba intenta deducir la condición del paciente.

Lo que este reclama de una prueba es que, si el resultado de la prueba es positivo, la probabilidad de que el sujeto esté efectivamente enfermo sea muy alta y, análogamente, que sea muy alta la de que el individuo esté sano, supuesto que la prueba arroje un resultado negativo. En términos formales, lo ideal es que sean muy altos los valores P(E/R+) y P(S/R-) que son probabilidades condicionales a las que ha dado en llamarse valores predictivos de la prueba. Es bien conocido que, si bien sensibilidad y especificidad son números inherentes a la prueba (en el sentido de que no dependen de cuál sea la población o el sujeto específico a la que se aplique), no ocurre lo mismo con sus valores predictivos.

Si llamamos P(E) a la probabilidad a priori de que el sujeto esté enfermo, es decir su prevalencia**, y** P(S) =1- PE, a su complemento, aplicando la fórmula del Teorema de Bayes se obtienen de inmediato las siguientes relaciones, que expresan la forma concreta que alcanzan estos valores cuando se ponen en función de los tres parámetros (la formula puede reescribirse de la manera siguiente):

$$P(E/R+) = \frac{(Sensibilidad).(Prevalencia)}{(Sensibilidad).(Prevalencia)+(1-Especificidad)(1-Prevalencia)}$$ = Valor predictivo positivo (VPP), y:

$P(S/R-) = \frac{(Especificidad).(1-Prevalencia)}{(Especificidad).(1-Prevalencia)+(1-Sensibilidad).(Prevalencia)}$ = Que representa el valor predictivo negativo (VPN).

A partir de estos valores, obviamente, se pueden obtener las probabilidades de estar sano a pesar de que el resultado haya sido positivo (ser un falso positivo) y de estar enfermo, aunque la prueba haya dado negativo (ser un falso negativo):

P(S/R+) = 1– P(E/R+) y P(E/R-) =1– P(S/R-)

- Ejemplo 16: Supongamos que, en una comunidad, la prevalencia de padecer una enfermedad coronaria (EC) entre sujetos mayores de 50 años es 5%. Esto quiere decir que, de cada 100 sujetos que están en esa franja de edad, 5 tendrán la dolencia; o, dicho de otro modo, que la probabilidad a priori de que un sujeto elegido al azar se halle en ese caso es P(E)=0,05.
- Si a cierto individuo se le practica una angiografía (una prueba invasiva y, por ende, peligrosa), hay dos resultados posibles: (R+ y R-). Después de hacerlo, ¿cuál es la probabilidad de que el sujeto esté enfermo en cada caso? Teniendo en cuenta que la sensibilidad y especificidad de una angiografía para el diagnóstico de una EC son, según Austin (18), sensibilidad igual a 0,87 y especificidad igual a 0,54. Aplicando las fórmulas anteriores se tiene que:

$$P(E/R+) = \frac{(0,87).(0,05)}{(0,87).(0,05)+(0,95)(0,46)} = 0,09 = 9\%.$$ Y

$$P(S/R-) = \frac{(0,95).(0,54)}{(0,95).(0,54)+(0,05).(0,13)} = 0,99 = 99\%$$

De este modo, se puede sostener que la probabilidad de que el individuo esté enfermo, si la prueba fue positiva, se eleva a 0,09 y si fue negativa, tal probabilidad se reduce a 0,01 (1-0,99). Los valores 0,09 y 0,01 son las llamadas probabilidades a posteriori de estar enfermo.

- Ejemplo17:

Considérese la información suministrada en la tabla 10. Habiendo aplicado la prueba en la detección precoz de la enfermedad en una población sobre la que se sabe que la enfermedad se produce con frecuencia del 0,5% de los individuos, se desea cuantificar la sensibilidad, especificidad, falsos positivos y negativos de la prueba, y los valores predictivos positivo y negativo.

Tabla 10. Resultados de la enfermedad

		Resultados de la enfermedad		
		Enfermo (E)	**No enfermo (S)**	
Resultados de la prueba diagnóstica	R+	190	120	310
	R-	10	680	690
		200	800	1000

Fuente: Nolasco y Moncho (15)

Al realizar las operaciones tenemos:

Sensibilidad = *P*(+¿*E*)= 190/200 = 0,95

Especificidad = P(—¿*S*¿ = 680/800 = 0,85

Falso positivo = *P*¿)=0,15

Falso negativo = P(—¿*S*)= $\frac{10}{200}$ = 0,05

Los valores predictivos dependerán de la prevalencia, que en este caso se estima a partir del dato frecuencial 0,5%, como ***P(E)*** = 0,005, y aplicando la formula vista anteriormente, tenemos:

Valor positivo predictivo = P(*E*/+¿) = $\frac{(0{,}95).(0{,}005)}{(0{,}95).(0{,}005)+(0{,}15).(0{,}995)}$ = 0,0308

$$\text{Valor negativo predictivo} = P(S/\text{—¿}) = \frac{(0{,}85).(0{,}995)}{(0{,}85).(0{,}995)+(0{,}5).(0{,}005)} = 0{,}9997$$

Como se observa, el valor predictivo negativo (99,97%) resulta altamente satisfactorio, pues la probabilidad de «acierto» cuando el resultado es negativo es muy elevada. Sin embargo, el valor predictivo positivo (3,08%) no resulta aparentemente satisfactorio, puesto que la probabilidad de acierto es muy baja (15).

Prevalencia e incidencia

Los valores estadísticos generados a partir de datos nominales se emplean en las ciencias de la salud para describir diferentes situaciones.

- Dos porcentajes, cuyos usos son muy habituales para evaluar la situación en cuanto a un estado patológico, son las tasas de prevalencia y de incidencia.

Como proporciones expresadas en porcentajes, ambas tasas se calculan al dividir la frecuencia de datos en una categoría por la cantidad total de datos y, por lo general, al multiplicar la proporción así obtenida por un valor constante, generalmente 100.

La diferencia entre ambas tasas radica en cuáles son los datos que se toman en cuenta para obtener la frecuencia. En la tasa de prevalencia se cuenta la cantidad de datos en la categoría en un momento determinado, mientras que en la tasa de incidencia se cuenta la cantidad de datos que aparecieron en la categoría durante un lapso determinado; por ejemplo, un año.

Esto significa que en la tasa de incidencia no se tienen en cuenta los datos existentes en la categoría desarrollados en períodos anteriores. La situación puede determinar que en el caso de enfermedades crónicas (el paciente

no se cura ni se muere) la tasa de prevalencia aumente a pesar de que a partir de medidas preventivas se logre disminuir la tasa de incidencia.

Valoración del riesgo

En general, se entiende por riesgo a la probabilidad de que se presente un daño, como resultado de la exposición a un agente, sea este químico, físico (por ejemplo, radiaciones y calor) o biológico (virus, bacterias) (19).

En las situaciones más frecuentes estudiadas en las ciencias de la salud, valorar el riesgo significa evaluar si la presencia de una situación o un factor determinado, como el hábito de fumar o ejercer una determinada profesión, significa una posibilidad definida de desarrollar una afección específica, por ejemplo, enfermedad pulmonar o alteraciones en la columna vertebral, respectivamente.

La evaluación de ese riesgo puede realizarse al comparar los hechos que se producen en conjuntos de individuos o unidades experimentales (en los que el factor está presente) respecto de los que se producen en conjuntos de individuos o unidades experimentales en donde no lo está, como fumadores y no fumadores, por ejemplo.

Los procedimientos numéricos que se emplean varían según si los datos son obtenidos a partir de diseños experimentales prospectivos (cohorte) o retrospectivos (casos y controles).

Riesgo relativo

Considérese los siguientes ejemplos:

- Ejemplo 18:

En un diseño prospectivo se conforman dos grupos de individuos, según la presencia del posible factor de riesgo o no. Ambos grupos se siguen a

través del tiempo y en cada uno de sus integrantes se registra la aparición del desenlace o no, desarrollo de la enfermedad o no.

Al cabo del lapso previsto para la experiencia, se pueden haber recolectado datos como los que se muestran en la tabla 11.

Tabla 11. Evaluación de los factores de Riesgo (Diseño Prospectivo)

	Con Enfermedad	**Sin Enfermedad**	**Total**
Con factor de Riesgo	40	160	200
Sin factor de riesgo	20	180	200

Fuente: Macchi (1)

A partir de estos datos se puede evaluar en cada grupo el riesgo, la relación porcentual entre la frecuencia de enfermedad y el total de integrantes del grupo. Los cálculos realizados son:

- Riesgo con factor: 40 / 200 = 0,20 (20%);
- Riesgo sin factor 20 / 200 = 0,10 (10%);
- Riesgo relativo: 0,20 / 0,10 = 2.

En el ejemplo, esos valores son 20 y 10% para los grupos con factor de riesgo y sin él, respectivamente. Estos valores indican la probabilidad de contraer la condición indeseable en presencia o ausencia del factor de interés.

La relación entre ambas proporciones –o entre los porcentajes (40 / 20)–, que en este caso es 2, se denomina riesgo relativo.

Un valor 1 en el riesgo relativo indica que el factor no constituye un riesgo; un valor mayor de 1, como en el ejemplo, indica que el riesgo es mayor con la presencia del factor; y un valor menor de 1 indicaría que el factor no solo no es un riesgo, sino que podría ser un factor beneficioso para disminuir la posibilidad de desarrollo de la enfermedad.

Odds ratio o razón de productos cruzados

En los diseños retrospectivos, los grupos se conforman según se haya producido el desenlace o no, presencia de enfermedad o su ausencia. Luego, se valora la exposición de los integrantes de esos grupos al factor de riesgo en el pasado.

- Ejemplo 19:

Considérese los datos de la tabla 12 Diseño Retrospectivo.

Nótese que en este caso no se conoce el total de individuos expuestos al factor de riesgo, ya que ellos fueron seleccionados una vez producido el desenlace o no.

Tabla 12. Evaluación de los factores de Riesgo (Diseño Retrospectivo)

	Con Enfermedad	**Sin Enfermedad**
Con factor de Riesgo	40	160
Sin factor de riesgo	20	180
Total	60	340

Fuente: Macchi (1)

Por este motivo, no es posible calcular la incidencia que indica el riesgo (obsérvese que, en este caso, el denominador es la cantidad total de individuos del conjunto). En cambio, es posible calcular razones al relacionar las frecuencias de la presencia del factor de riesgo en los grupos de enfermos y no enfermos. Los cálculos son:

- Chance (odds) con enfermedad: 40 / 20 = 2;
- Chance (odds) sin enfermedad: 160 / 180 = 0,89;
- Odds ratio: 2 / 0,89 = 2,25.

En el ejemplo, esa razón, que se describe como chance u odds en inglés, es 2 (40 / 20) y 0,89 (160 / 180) en los grupos con enfermedad y sin ella, respectivamente.

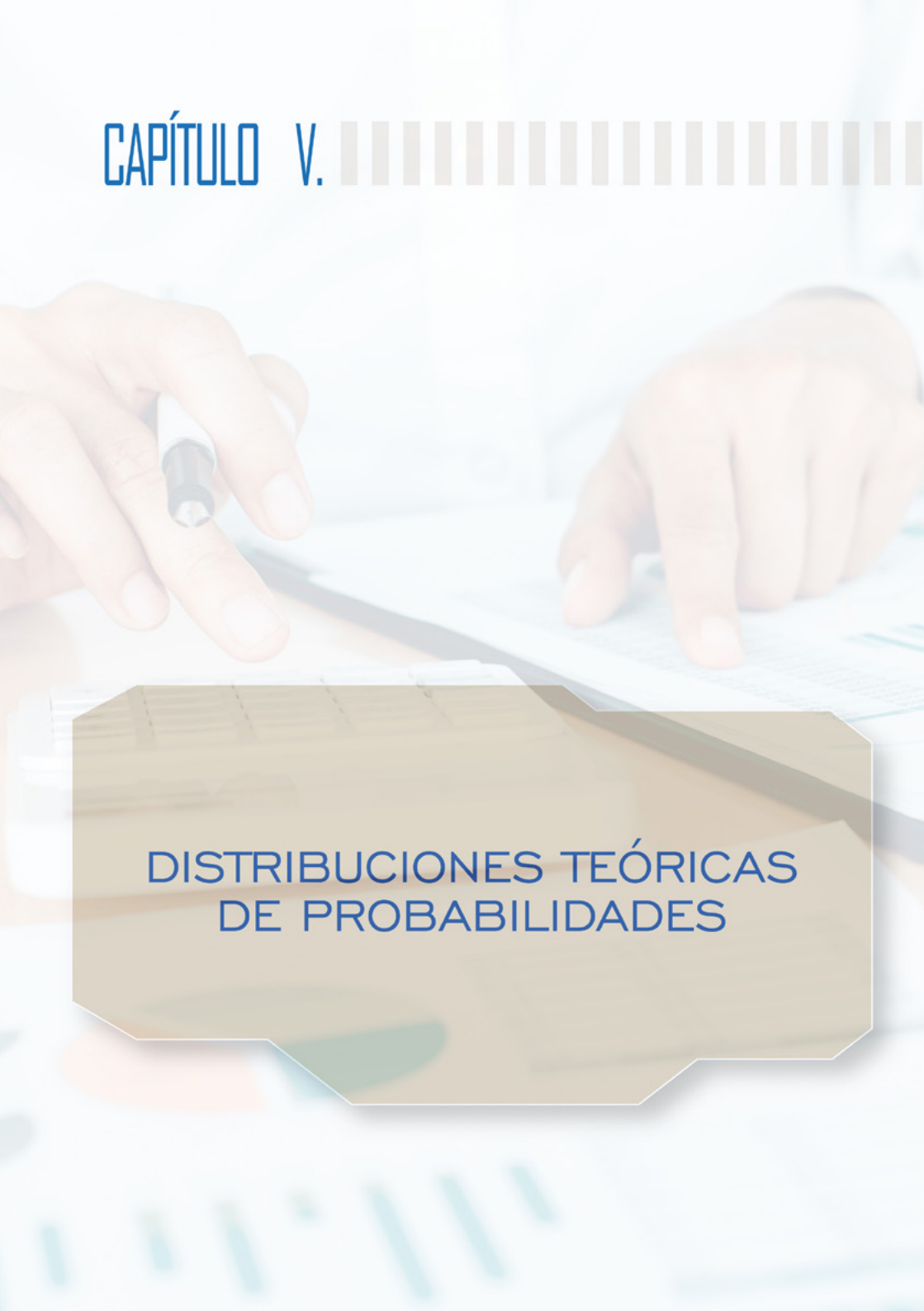
Para valorar el factor de riesgo, se establece la razón entre las dos razones, que en este caso es 2,25 (2 / 0,89) y se la designa con el nombre de razón de chances, razón de productos cruzados o, con mucha asiduidad, con las palabras inglesas odds ratio (OR). Un valor mayor de 1 (2,25 en el ejemplo) indica una mayor frecuencia de individuos con el factor de riesgo en el grupo con enfermedad y, por ende, la posible contribución que este tiene en su desarrollo.

Al igual que con lo que sucede en la evaluación de pruebas diagnósticas, debe tenerse presente que, si los cálculos de riesgo relativo o de odds ratio se realizan a partir de muestras, solo deben servir de base para la aplicación de la estadística inferencial en la estimación de la situación en las correspondientes poblaciones.

En síntesis: En las ciencias de la salud, las razones o proporciones se usan de manera habitual para el cálculo de porcentajes a fin de establecer las tasas de prevalencia y de incidencia de una patología, así como para la evaluación de pruebas diagnósticas mediante el cálculo de porcentajes de sensibilidad, especificidad y valor predictivo.

Las proporciones y razones también permiten evaluar el riesgo que representa una determinada condición para que aparezca un hecho definido y, por lo general, no deseado, mediante los valores de riesgo relativo y de odds ratio.

CAPÍTULO V.



DISTRIBUCIONES TEÓRICAS DE PROBABILIDADES

CAPÍTULO V: DISTRIBUCIONES TEÓRICAS DE PROBABILIDADES

Introducción

El presente capítulo se centra en describir algunos modelos teóricos de probabilidad que permiten caracterizar la distribución poblacional de determinadas variables y que, a su vez, son aplicables a múltiples situaciones prácticas. Cuando se realiza un estudio o un experimento aleatorio, es frecuente asignar a los resultados del mismo una cantidad numérica. A la función que asocia un número real a cada resultado de un experimento se le denomina variable aleatoria.

Concepto de variable aleatoria

Aunque el concepto de variable se ha introducido con anterioridad, una definición más formal de variable aleatoria es, por tanto, la de una función definida sobre el espacio muestral Ω que asigna a cada posible resultado de un experimento un valor numérico (10). Aunque en general pueden definirse múltiples variables aleatorias para un mismo experimento, es aconsejable seleccionar en cada caso aquellas variables que recojan las características fundamentales del experimento. Las variables aleatorias suelen denotarse por letras mayúsculas del final del alfabeto, tales como X, Y o Z, mientras que los valores que pueden tomar se representan por sus correspondientes letras minúsculas, x, y o z.

Desde un punto de vista formal, que una variable aleatoria se define como una función que asigna a cada uno de los posibles resultados de un fenómeno aleatorio un valor numérico (11). Por ejemplo, en el caso de variables cuantitativas como el nivel de colesterol, nivel de ácido úrico o número de ingresos en un servicio de urgencias, la variable aleatoria vendría definida por cada una de las posibilidades de las variables consideradas, puesto

que en este caso ya son valores numéricos. Cuando las variables son de tipo cualitativo como el sexo o el nivel de ***estudios,*** será necesario asignar valores numéricos a cada una de las posibilidades (p. ej., 1 Hombre, 2 Mujer, para el sexo; 1 Sin estudios, 2 Primaria, 3 Secundaria y 4 Universitarios, para el nivel de estudios).

Una consecuencia inmediata derivada de este proceso de asignación de valores numéricos es que las variables aleatorias se clasifican únicamente en: variables aleatorias «***discretas***» y variables aleatorias continuas. Las variables aleatorias discretas pueden tomar un número finito o infinito numerable de valores, mientras que las continuas pueden alcanzar un número infinito no numerable de posibles valores, es decir, pueden tomar cualquier valor en un intervalo.

- Ejemplo 1:

A continuación, se definen algunas variables aleatorias para el experimento consistente en observar la supervivencia a los 6 meses de 4 pacientes con enfermedad catastrófica sometidos a tratamiento, una variable aleatoria "***X"*** podría ser el número de supervivientes, que tomaría los valores ***X*** = 0, 1, 2, 3 ó 4 en función del número de pacientes que hayan sobrevivido a los 6 meses. Alternativamente, podría definirse otra variable aleatoria ***Y*** como el número de muertes, cuyos valores serían "***Y"*** = 0, 1, 2, 3 ó 4 en función del número de muertes observadas. Para el experimento de medir el colesterol HDL de una persona, la variable aleatoria ***X*** más natural sería el nivel de colesterol HDL en mmol/l, que podría tomar cualquier valor positivo. Si el interés se centra en saber si los niveles de colesterol HDL son superiores o inferiores al umbral de 0,90 mmol/l, otra variable aleatoria ***Y*** podría definirse como ***Y*** = 0 si el nivel observado es inferior a 0,90 mmol/l y 1 en caso contrario. La elección de los valores 0 y 1 es arbitraria, bastaría con asignar dos valores distintos para diferenciar ambos tipos de resultados.

Como las variables aleatorias son funciones definidas sobre el espacio muestral, sus posibles valores tendrán asociada una probabilidad, que corresponderá a la probabilidad del suceso constituido por aquellos resultados del experimento que toman dichos valores. Los diferentes valores de una variable aleatoria y las probabilidades asociadas constituyen la distribución de probabilidad de la variable.

- Ejemplo 2:

En el primer experimento del ejemplo anterior, el número de supervivientes es una variable aleatoria que toma los valores $X = 0, 1, 2, 3$ ó 4. La probabilidad asociada al valor 0 $P(X = 0)$ sería la probabilidad del suceso "ninguno de los 4 pacientes sobrevive a los 6 meses", la probabilidad asociada al valor 1 $P(X = 1)$ sería la probabilidad del suceso "sólo 1 de los 4 pacientes sobrevive a los 6 meses", y así sucesivamente. En el segundo experimento, el nivel de colesterol HDL es una variable aleatoria X que puede tomar cualquier valor en el intervalo $(0, \infty)$. En este caso no tiene sentido preguntarse, por ejemplo, cuál es la probabilidad de tener exactamente un nivel de colesterol HDL de 1 mmol/l, ya que, si esta variable se pudiera determinar con una precisión infinita, la probabilidad $P(X = 1) = 0$. En tal caso, deberíamos preguntarnos por la probabilidad de un determinado intervalo de valores. Así, por ejemplo, la probabilidad $P(X \leq 1)$ sería la probabilidad del suceso "tener niveles de colesterol HDL menores o iguales a 1 mmol/l".

En general, se distinguen dos grandes grupos de variables aleatorias:

- **Variables aleatorias discretas** son aquellas que tan sólo puede tomar un número discreto (finito o infinito) de valores. Cada uno de estos valores lleva asociada una probabilidad positiva, mientras que la probabilidad de los restantes valores es 0.

- **Variables aleatorias continuas** son aquellas que pueden tomar cualquier valor dentro de un intervalo. En este caso, la probabilidad de obtener un valor concreto es 0, por lo que las probabilidades se asignan a intervalos de valores.

A continuación, se describen las principales características de las variables aleatorias discretas y continuas, así como algunas distribuciones teóricas de probabilidad que serán aplicables a muchas de las variables aleatorias utilizadas en la práctica.

Distribuciones de probabilidad discretas

Las variables aleatorias discretas toman un número discreto de valores con probabilidad no nula y, en consecuencia, estarán completamente caracterizadas si se conoce la probabilidad asociada a cada uno de estos valores. La función que asigna a cada posible valor X_i, i = 1, 2, ..., de la variable discreta X su probabilidad $P(X = Xi)$ se conoce como función de masa de probabilidad. Esta función debe cumplir las siguientes propiedades: la probabilidad de cada valor ha de estar entre 0 y 1, $0 < P(X = xi) \leq 1$, y la suma de las probabilidades para todos los valores debe ser igual a 1,

$$\sum_{i \geq 1} P(X = xi) = 1$$

Una vez conocida la función de masa de probabilidad, la probabilidad de que una variable aleatoria discreta X esté comprendida en cualquier subconjunto A se calcula como la suma de las probabilidades de aquellos valores xi incluidos dentro de ese subconjunto,

$$P(X \in A) = \sum_{Xi \in A} P(X = xi)$$

En particular, la función de distribución F(x) de una variable aleatoria x se define como la probabilidad de observar un valor menor o igual a x.

$$P(x)=P(X\leq x)=\sum_{xi\leq x} P(X=xi)$$

La función de distribución de una variable discreta será una función escalonada creciente con saltos en valores xi con probabilidad no nula.

- Ejemplo 3:

  Supongamos que por estudios previos se estima que, después de 6 meses de tratamiento en 4 pacientes con una enfermedad catastrófica, la probabilidad de que sobrevivan 0, 1, 2, 3 ó 4 pacientes viene determinada por la segunda columna de la Tabla 3. Estos valores y sus probabilidades constituyen la función de masa de probabilidad de la variable número de supervivientes. Los valores de la función de distribución en 0, 1, 2, 3 y 4 aparecen en la tercera columna de la tabla 13; así, al calcular la función de distribución en 1 tenemos F (1) = P(X ≤ 1) = P(X = 0) + P(X = 1) = 0,1296 + 0,3456 = 0,4752. Note que F(x) está definida sobre cualquier número real, aun cuando la variable tome sólo los valores 0, 1, 2, 3 y 4 con probabilidad no nula.

Tabla 13. Función de masa de probabilidad y función de distribución del número de supervivientes a los seis meses de cuatro pacientes con enfermedad catastrófica sometidos a tratamiento

Numero de supervivientes (X)	Función de masa P(X=x)	Función de distribución P(x) = P(X≤x)
0	0,1296	0,1296
1	0,3456	0,4752
2	0,3456	0,8208

3	0,1536	0,9744
4	0,0256	1

Fuente: Pastor-Barriuso (10)

Distribución binomial

La distribución binomial es un modelo teórico de distribución de probabilidad discreta aplicable a aquellos experimentos en los que se realizan n pruebas independientes, cada una de ellas con sólo dos resultados posibles (éxito o fracaso) y la misma probabilidad de éxito π. En tal caso, se dice que la variable aleatoria X "número de éxitos en las n pruebas" sigue una distribución binomial con parámetros n y π. A partir de los resultados del tema probabilidad (véase capitulo anterior), puede probarse que la distribución binomial toma valores en $k = 0, 1, ..., n$ con probabilidad $P(X=k) = \binom{n}{k} \pi^k (1-\pi)^{n-k}$, donde $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ es el número de combinaciones de n elementos tomados de k en k, con $n! = n(n-1) \cdot \ldots \cdot 1$ y $0! = 1$. Por supuesto, estas probabilidades constituyen una función de masa de probabilidad ya que, para cualquier n y π, su suma es exactamente igual a 1. En la práctica, resulta tedioso calcular las probabilidades de una distribución binomial mediante la fórmula anterior. Por esta razón, existen tablas que facilitan el cálculo de las probabilidades binomiales para $n = 2, 3, ..., 20$ y $\pi = 0{,}05, 0{,}10, ..., 0{,}50$.

En general, la distribución binomial se aplica al estudio de observaciones repetidas e independientes de una misma variable dicotómica (con sólo dos resultados posibles), tal como el resultado de un tratamiento (éxito o fracaso) en pacientes de similares características sometidos a una misma terapia.

- Ejemplo 4:

En los ejemplos anteriores, se ha considerado el experimento de observar la supervivencia (o muerte) en pacientes con una enfermedad catastrófica

sometidos al mismo tratamiento. Si por estudios previos se sabe que la supervivencia a los 6 meses en dichos pacientes es del 40%, el número de supervivientes a los 6 meses en una muestra de 4 pacientes seguirá una distribución binomial ***X*** de parámetros ***n*** = 4 y π = 0,4.

Esta probabilidad está constituida por la unión de tantos sucesos como posibles combinaciones de 4 pacientes tomados de 2 en 2; es decir, $\binom{4}{2} = \frac{4!}{2!(4-2)!} = \frac{24}{4} = 6$ sucesos. Además, estos sucesos son mutuamente excluyentes y todos ellos tienen una misma probabilidad de ocurrir de $(0{,}4)^2(1-0{,}4)^2$. En consecuencia, la probabilidad de que sobrevivan 2 pacientes cualesquiera es: $P(X=2) = \binom{4}{2} \cdot 0{,}4)^2 (1-0{,}4)^2 = 0{,}3456$, que corresponde a la probabilidad binomial de parámetros n=4 y $\pi = 0{,}4$ para $k = 2$. Aplicando esta fórmula, las probabilidades para $k = 0, 1, 2, 3$ ó 4 supervivientes aparecen en la tabla 12.

Los resultados de la función de probabilidad Binomial se pueden determinar con cualquier paquete estadístico, como MINITAB o SPSS (para cualquier valor de ***n*** y ***π***) y con las tablas de la función de probabilidad (para algunos valores de n y π) (20).

Distribución de Poisson

La distribución de Poisson es otro modelo teórico de distribución discreta particularmente útil para el estudio epidemiológico de la ocurrencia de determinadas enfermedades. Se dice que la variable aleatoria ***X*** "número de casos de una determinada enfermedad a lo largo de un periodo de tiempo ***t***", donde ***t*** es un intervalo de tiempo arbitrariamente largo, tal como 1 ó 10 años, sigue una distribución de Poisson si se cumplen las siguientes hipótesis respecto a la incidencia acumulada (IA) de la enfermedad (esto es, la probabilidad de desarrollar un nuevo caso en un periodo de tiempo determinado):

- Proporcionalidad: La probabilidad de observar un caso es aproximadamente proporcional al tiempo transcurrido, de tal forma que, en un intervalo de tiempo arbitrariamente corto, la probabilidad de observar un caso es muy pequeña y la probabilidad de observar más de un caso es esencialmente nula.
- Estacionaridad: El número de casos por unidad de tiempo permanece aproximadamente constante a lo largo de todo el periodo de tiempo ***t***. Notar que, si se produjera un cambio substancial de la incidencia de la enfermedad en el tiempo, esta asunción no sería aplicable.
- Independencia: La ocurrencia de un caso en un determinado instante no afecta a la probabilidad de observar nuevos casos en periodos posteriores. Así, por ejemplo, esta hipótesis de independencia no se cumplirá en brotes epidémicos.

Aunque la distribución de Poisson se emplea habitualmente en el estudio de la morbi-mortalidad debida a determinadas enfermedades, esta distribución es en general aplicable a la ocurrencia en el tiempo de aquellos sucesos aleatorios que satisfagan las hipótesis anteriores (por ejemplo, los accidentes de tráfico).

Bajo estas asunciones, se establece que la probabilidad de que ocurran k sucesos, k = 0, 1, 2, ..., en un periodo de tiempo t para una variable aleatoria X que sigue una distribución de Poisson es: $P(X=k)=\frac{e^{-\mu}\mu^{k}}{k!}$, donde el parámetro μ es el número esperado de sucesos en el periodo de tiempo t. A diferencia de la distribución binomial, donde el número de éxitos k no puede exceder el número finito de pruebas realizadas, en la distribución de Poisson el número de pruebas se considera infinito y el número de sucesos k puede ser arbitrariamente grande, aunque la probabilidad P(X = k) decrecerá al aumentar k hasta hacerse esencialmente nula. Para cualquier parámetro μ > 0, estas probabilidades son positivas y suman 1, constituyendo una función de masa de probabilidad. Una característica

importante de la distribución de Poisson es que tanto su media como su varianza son iguales al parámetro μ. Al igual que en la función de distribución Binomial, existen tablas que facilitan el cálculo de las probabilidades de Poisson para μ de 0,5 a 20 en intervalos de 0,5. (21)

- Ejemplo 6:

Según estudios realizados la tasa de mortalidad por cáncer de vesícula en hombres es de I = 1,80 casos por 100.000 personas año. Partiendo de esta información, se pretende determinar la distribución del número de muertes por cáncer de vesícula en un periodo de 1 ó 2 años en una población de 140.000 hombres. Las asunciones de estacionaridad e independencia parecen razonables por tratarse de casos de mortalidad por cáncer en periodos cortos de tiempo. Además, como la tasa de mortalidad (I) es baja y se asume constante en el tiempo, puede probarse que la incidencia acumulada en un periodo de tiempo t es:

$$IAt = 1 \exp(It) \approx It;$$

es decir, la probabilidad de que un individuo de esta población muera por cáncer de vesícula es aproximadamente proporcional al tiempo transcurrido, cumpliéndose así la hipótesis de proporcionalidad. La incidencia acumulada en 1 año es:

$$IA1 = 0{,}000018 \text{ y en 2 años } IA2 = 0{,}000018 \cdot 2 = 0{,}000036.$$

En consecuencia, el número de muertes por cáncer de vesícula en un periodo de tiempo t seguirá una distribución de Poisson con un número esperado de casos igual al producto del tamaño poblacional por la probabilidad individual de muerte en dicho periodo, $\mu = 140.000 \cdot 0{,}000018 = 2{,}52$ muertes esperadas en 1 año y $140.000 \cdot 0{,}000036 = 5{,}04$ en 2 años.

Estas distribuciones de probabilidad se muestran en la tabla 14. Por ejemplo, la probabilidad de que no se produzca ninguna muerte por cáncer

de vesícula durante 1 año en esta población se calcula a partir de la distribución de Poisson de parámetro $\mu = 2{,}52$ como $P(X = 0) = e^{-\mu}\ \mu^{0}/0! = e^{-2{,}52} = 0{,}0805$.

Tabla 14. Distribución de probabilidad del número de muertes por cáncer de vesícula en periodos de 1 y 2 años en una población de 140.000 hombres

Número de muertes (k)	P(X = k)	
	1 año	2 años
0	0,0805	0,0065
1	0,2028	0,0326
2	0,2555	0,0822
3	0,2146	0,1381
4	0,1352	0,174
5	0,0681	0,1754
6	0,0286	0,1474
7	0,0103	0,1061
8	0,0032	0,0668
9	0,0009	0,0374
10	0,0002	0,0189
11	0,0001	0,0086
12	0,0000	0,0036
13	0,0000	0,0014
14	0,0000	0,0005
15	0,0000	0,0002
16	0,0000	0,0001
17	0,0000	0,0000

Fuente: Pastor-Barriuso (10)

Distribuciones de probabilidad continuas

Las variables aleatorias continuas son aquellas que pueden tomar cualquier valor dentro de un intervalo. La probabilidad de que estas variables

tomen exactamente un valor determinado es 0 y, en consecuencia, carece de sentido definir una función de masa de probabilidad. Para las variables aleatorias continuas, las probabilidades se asignan a intervalos de valores mediante una función de densidad de probabilidad, denotada por f(x). Esta función ha de ser no negativa para cualquier valor x, $f(x) \geq 0$, y el área total bajo la curva definida por esta función de densidad debe ser igual a 1 (10).

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

A partir de la función de densidad, la probabilidad de que una variable aleatoria continua X tome valores dentro de cualquier intervalo (a, b) puede calcularse como el área bajo la función de densidad entre los puntos a y b,

$$P(a < X < b) = \int_{a}^{b} f(x)dx$$

Así, aun cuando la probabilidad de obtener un valor concreto es 0, la función de densidad tomará valores elevados en regiones de alta probabilidad y valores pequeños. Así, aun cuando la probabilidad de obtener un valor concreto es 0, la función de densidad tomará valores elevados en regiones de alta probabilidad y valores pequeños en regiones de baja probabilidad. La función de distribución F(x) corresponde a la probabilidad de que la variable tome un valor igual o inferior a x y, en el caso de una variable aleatoria continua, se calcula como el área bajo de la curva de la función de densidad a la izquierda de x,

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt$$

La función de distribución de una variable aleatoria continua es una función que, partiendo de 0, crece en forma continua hasta alcanzar el valor de 1.

Distribución normal

La distribución normal fue reconocida por primera vez por el Francés Abraham de Moivre (1667-1754), Posteriormente Carl Friedrich Gauss (1777-1855)

elaboró desarrollos más profundos y elaboró la ecuación de la curva de ahí que también se le conozca como: Campana de Gauss. La distribución Gaussiana, es el modelo teórico de distribución continua más utilizado en la práctica. Muchas mediciones epidemiológicas y clínicas presentan distribuciones similares al modelo teórico normal (presión arterial, colesterol sérico, índice de masa corporal) o bien pueden transformarse para conseguir distribuciones aproximadamente normales (típicamente mediante transformaciones logarítmicas de los datos originales). No obstante, como se verá en apartados posteriores, la utilidad fundamental de la distribución normal surge dentro de las técnicas de inferencia estadística: incluso cuando la distribución poblacional de una variable diste mucho de ser normal, puede probarse que, bajo ciertas condiciones, la distribución de los valores medios de dicha variable seguirá un modelo aproximadamente normal.

La distribución normal es la distribución de probabilidad más importante en estadística, debido a tres razones fundamentales (22):

- Desde un punto de vista matemático resulta conveniente suponer que la distribución de una población de donde se ha extraído una muestra aleatoria sigue una distribución normal, ya que entonces se pueden obtener las distribuciones de varias funciones importantes de las observaciones muestrales, que además resultan tener una forma sencilla.
- Desde un punto de vista científico, la distribución normal aproxima en muchas ocasiones los valores obtenidos para variables que se miden sin errores sistemáticos. Por ejemplo, se ha observado que muchos experimentos físicos frecuentemente tienen distribuciones que son aproximadamente normales, como estaturas o pesos de los individuos, beneficios medios de las empresas, la duración de

un producto perecedero, el tiempo necesario para llevar a cabo un trabajo, etcétera.

- La última razón es la existencia del Teorema Central del Límite, establece que cuando se dispone de una muestra aleatoria grande, aunque presente una distribución no normal e incluso distribuciones típicas de variables aleatorias discretas, pueden tratarse como aproximadamente distribuciones normales.

Algunos ejemplos típicos de la distribución normal son: Estatura de las personas, peso de los recién nacidos en un hospital, nivel de colesterol HDL en sangre, valores de tensión arterial sistólica, etcétera.

La distribución normal y su función de densidad

Una variable aleatoria continua X sigue una distribución normal si su función de densidad es:

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right),$$

para cualquier valor x en la recta real, $-\infty < x < \infty$. Esta función de densidad depende de los parámetros μ y σ, donde:

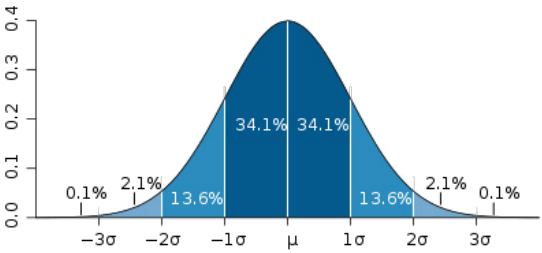
- μ representa la esperanza o media poblacional de la distribución, y
- σ corresponde a su desviación típica poblacional.
- El símbolo (π) es una constante matemática aproximada por 3.1416.

La Distribución Normal (μ, σ2)

La distribución normal o Gaussiana con media μ y varianza σ^2 se denota abreviadamente por $N(\mu, \sigma^2)$. Para cualquier μ y $\sigma > 0$, la función de densidad normal es positiva y el área total bajo la curva es igual a 1. Esta función de densidad, que aparece representada en la figura 13, tiene forma de campana, es simétrica alrededor de la media μ y tiene dos puntos de inflexión en $\mu + \sigma$ y $\mu - \sigma$. Al tratarse de una distribución simétrica, la media

y la mediana coinciden. El valor más frecuente 1/(π 2σ) se alcanza en la media μ y su dispersión alrededor del valor medio aumenta al aumentar la desviación típica σ. Así, puede probarse que el 68,27% del área bajo una función de densidad normal está comprendido entre $\mu \pm \sigma$, el 95,45% entre $\mu \pm 2\sigma$ y el 99,73% entre $\mu \pm 3\sigma$.

Figura 13. Función de densidad de una distribución normal con media μ y desviación típica σ.



Fuente: Char (23)

En resumen, las propiedades de la Distribución Normal son (24):

- La campana de Gauss presenta simetría y la misma distribución de áreas, independientemente de cuáles sean sus párametros media y desviación típica.
- Tiene una única moda, que coincide con su media y su mediana.
- La curva normal es asintótica al eje de abscisas. Por ello, cualquier valor entre -∞ y ∞ es teóricamente posible. El área total bajo la curva es, por tanto, igual a 1.
- Es simétrica con respecto a su media μ. Según esto, para este tipo de variables existe una probabilidad de un 50% de observar un dato mayor que la media, y un 50% de observar un dato menor.
- La distancia entre la línea trazada en la media y el punto de inflexión de la curva es igual a una desviación típica (δ).

- El área bajo la curva aproximadamente: A una desviación estándar de la media es igual 0.68. A dos desviaciones estándar de la media es igual a 0.95. A tres desviaciones estándar de la media es igual a 0.99.
- Las distribuciones que siguen una binomial o una distribución de Poisson se pueden aproximar a una distribución normal.

Distribución normal estandarizada N (0, 1)

La distribución normal con media 0 y desviación típica 1 se denomina «distribución normal estandarizada», y suele denotarse por *Z* o *N* (0, 1). La función de densidad de una distribución normal estandarizada se reduce a:

$$f(Z)=\frac{1}{\sqrt{2\pi}}\exp\left(\frac{-1}{2}Z^2\right)$$

para cualquier valor comprendido entre: $-\infty < z < \infty$

Esta función es simétrica con respecto de 0, que es la moda de la distribución. La densidad disminuye paulatinamente en los extremos (figura 14). La simetría se utiliza, a menudo, en los cálculos de probabilidad.

Para obtener las probabilidades bajo la función de densidad normal estandarizada, no se recurre al cálculo integral, ya que estas probabilidades están tabuladas y son fácilmente accesibles. En general, estas tablas facilitan la función de distribución; es decir, la probabilidad de que la variable normal estandarizada tome un valor igual o inferior a Z. La función de distribución normal estandarizada se denota por $F(z) = P(Z \leq z)$, y se ilustra en el documento Web «Tabla de distribución normal estandarizada» (25), que utilizamos como referencia para el cálculo de probabilidades en los ejemplos que se presentan en este apartado.

Uso de tabla normal estándar

Cuando la media de la distribución es 0 y la varianza es 1 se denomina “distribución normal tipificada”, y su ventaja reside en que hay tablas

donde se recoge la probabilidad acumulada para cada punto de la curva de esta distribución.

La tabla nos da la probabilidad acumulada, es decir, la que va desde ∞ hasta un valor. No nos da la probabilidad concreta en ese punto. En una distribución continua en el que la variable puede tomar infinitos valores, la probabilidad en un punto concreto es cero.

El área bajo la curva de cualquier distribución normal se puede encontrar utilizando una tabla normal estándar (25), y cambiando a unidades estándar la escala de unidades reales. La media de la distribución sirve como punto de referencia y la desviación estándar como la unidad que mide distancias relativas a partir de la media.

La tabla normal estándar fue ideada de manera que se pueda leer en unidades z y muestra el área bajo la curva, es decir, la probabilidad de que un valor quede en ese intervalo, entre la media y los valores seleccionados.

Antes de utilizar la tabla considérese lo siguiente:

- Los valores de probabilidad, las áreas bajo la curva, que nos proporciona la tabla son valores calculados a partir de la media hasta el valor z seleccionado.
- La tabla normal también se puede utilizar para calcular áreas bajo la curva más allá de un valor dado de z. La clave, en este caso, es que la mitad del área es 50% y, por lo tanto, el área de un valor más allá de z es igual a 50% menos el valor de z en tablas.
- La media de la distribución siempre toma el valor de cero, es decir, se encuentra a cero desviaciones de sí misma.
- Como la distribución normal es simétrica con respecto a su media, el lado izquierdo de la curva es una imagen idéntica de su lado derecho. Debido a esta simetría en la tabla sólo se proporcionan los valores para la mitad derecha de la distribución.

- Valores de z mayores a 4 se aproximan a un resultado de 0.5000 o 50%.

La tabla se encuentra ordenada en términos de valores de z, hasta dos decimales, como, por ejemplo: 2.78, 1.04 y 2.45. Una particularidad de una tabla normal típica es que los valores z se presentan en dos partes. Los valores del entero y el primer decimal (2.7, 1.0 y 2.4) se enumeran hacia abajo en el lado izquierdo de la tabla, es decir, primera columna, mientras que el último dígito aparece en la parte superior. Veamos cómo se calcula el área bajo la curva entre la media y un valor Z:

▶ Ejemplo 6:

Suponga que se quiere obtener el área entre la media y un valor Z, cuando Z = 1.25. (véase la tabla en la referencia electrónica (25).

Primero localizamos el valor de 1.2, en el lado izquierdo de la tabla.

Luego en la parte superior, el valor de 0.05 (5 es el último dígito).

El área bajo la curva se puede encontrar (leer) en la intersección de la fila z = 1.2 y la columna 0.05.

El valor es 0,8944, que se refiere al porcentaje del área bajo la curva normal entre la media y un valor z= 1.25.

El porcentaje equivale a la probabilidad de que una variable aleatoria distribuida normalmente tenga un valor entre la media y un valor real equivalente a 1.25 desviaciones estándar sobre la media.

▶ Ejemplo 7:

La probabilidad de obtener un valor inferior a 0,50 en una distribución normal estandarizada se obtiene directamente de la tabla de distribución normal antes referenciada (25), como el valor de la función de distribución en 0,50; es decir, $P(Z \leq 0{,}50) = F(0{,}50) = 0{,}6915$. Asimismo, aunque en la tabla referenciada no aparecen las probabilidades acumuladas para

valores negativos, la probabilidad de obtener un valor inferior a – 0,25 en una distribución normal estandarizada puede calcularse fácilmente a partir de dicha tabla. Como la distribución normal estandarizada es simétrica alrededor de 0, la probabilidad a la izquierda de – 0,25 es igual a la probabilidad a la derecha de 0,25 y, en consecuencia, $P(Z \leq -0,25) = P(Z \geq 0,25) = 1 - P(Z \leq 0,25) = 1 - F(0,25) = 1 - 0,5987 = 0,4013$. A partir de los resultados anteriores, la probabilidad de que un valor de la distribución normal estandarizada se encuentre entre – 0,25 y 0,50 viene dada por $P(-0,25 \leq Z \leq 0,50) = P(Z \leq 0,50) - P(Z \leq -0,25) = 0,6915 - 0,4013 = 0,2902$.

También, el percentil 97,5 de una distribución normal estandarizada se denota por $Z_{0,975}$ y corresponde al valor Z que deja por debajo una probabilidad del 0,975. Con apoyo en la tabla ya citada, se tiene que $F(1,96) = 0,9750$ y, por tanto, $Z_{0,975} = 1,96$. Por tratarse de una distribución simétrica en 0, el percentil 2,5 corresponde al percentil 97,5 con signo opuesto; es decir, el percentil 2,5 es $Z_{0,025} = -Z_{0,975} = -1,96$. Así, los valores ± 1,96 abarcan el 95% central de la distribución normal estandarizada. Este resultado será particularmente útil en los temas de inferencia estadística.

El cálculo de probabilidades para cualquier distribución normal con media μ y varianza σ^2 no requiere de tablas específicas, sino que puede realizarse a partir de las tablas de la distribución normal estandarizada. Para ello, se hace uso del siguiente resultado sobre la estandarización de una distribución normal: si una variable aleatoria X sigue una distribución normal con media μ y varianza σ^2, $X \sim N(\mu, \sigma^2)$, entonces la variable aleatoria $Z = (X - \mu) / \sigma$ sigue una distribución normal estandarizada,

$$Z = \frac{X - \mu}{\sigma} \sim (N, 1),$$

donde el símbolo ~ significa “estar distribuido como”. Como es sabido, al restar a los valores de una variable su media y dividirlos por su desviación típica, la variable resultante tiene media 0 y desviación típica 1. El resulta-

do anterior garantiza además que la variable estandarizada conserva la distribución normal. Este procedimiento de estandarización de variables normales permite utilizar las tablas correspondientes a la distribución normal estandarizada.

- Ejemplo 8:

Supongamos que el colesterol HDL en una población de hombres adultos sigue una distribución normal X con media μ = 1,10 mmol/l y desviación típica σ = 0,30 mmol/l. Utilizando la estandarización de variables normales, el porcentaje de hombres de esta población que tienen niveles de colesterol HDL entre 0,90 y 1,20 mmol/l corresponde a:

$$P(0{,}90 \leq X \leq 1{,}20) = P[0{,}9—1{,}10/0{,}30 \leq X —1{,}10/0{,}30 \leq 1{,}20—1{,}10/0{,}30]$$

$$= P(0{,}67 \leq Z \leq 0{,}33) = P(Z \leq 0{,}33)\ P(Z \leq 0{,}67)$$

Utilizando la tabla de distribución de probabilidad (véase referencia (25)), se obtiene que $P(Z \leq 0{,}33) = F(0{,}33) = 0{,}6293$ y $P(Z \leq -0{,}67) = F(-0{,}67) = 1 - F(0{,}67) = 1 - 0{,}7486 = 0{,}2514$. Así, resulta que $P(0{,}90 \leq X \leq 1{,}20) = 0{,}6293 - 0{,}2514 = 0{,}3779$; es decir, el 37,79% de los hombres de esta población tienen niveles de colesterol HDL entre 0,90 y 1,20 mmol/l. Para obtener el percentil 90 de la distribución del colesterol HDL en esta población, se calcula primero el percentil 90 en la distribución normal estandarizada, que corresponde a $Z_{0,90} = 1{,}28$, ya que $F(1{,}28) \approx 0{,}90$. Para pasar este percentil estandarizado al correspondiente percentil del colesterol HDL basta resolver $Z_{0,90} = (X_{0,90} - \mu)/\sigma$. Por tanto, el percentil 90 del colesterol HDL es $X_{0,90} = \mu + Z_{0,90}\,\sigma = 1{,}10 + 1{,}28 \cdot 0{,}30 = 1{,}484$ mmol/l.

CAPÍTULO VI.

MUESTREO Y ESTIMACIÓN ESTADÍSTICA

CAPÍTULO VI. MUESTREO Y ESTIMACIÓN ESTADÍSTICA

Introducción

Un primer paso en la realización de un estudio o proyecto de investigación es definir la población de la cual se desea conocer una determinada característica o parámetro. Ocasionalmente, resulta factible obtener información para todos los elementos de la población mediante registros o censos. Sin embargo, en la mayoría de los estudios no es posible obtener información de toda la población, por lo que debemos limitarnos a la recogida de datos en una pequeña fracción del total o muestra. La utilización de muestras presenta varias ventajas con respecto a la enumeración completa de la población:

- Coste reducido. Si los datos se obtienen de una pequeña fracción del total, los gastos se reducen. Incluso si la obtención de información en toda la población es factible, suele ser mucho más eficiente la utilización de técnicas de muestreo.
- Mayor rapidez. Los datos pueden ser más fácilmente recolectados y estudiados si se utiliza una muestra que si se emplean todos los elementos de la población. Por tanto, el uso de técnicas de muestreo es especialmente importante cuando se necesita la información con carácter urgente.
- Mayor flexibilidad y mayores posibilidades de estudio. La disponibilidad de registros completos es limitada. Muy a menudo, la única alternativa posible para la realización de un estudio es la obtención de datos por muestreo.
- Mayor control de calidad del proceso de recogida de datos. Al recoger datos en un número menor de efectivos, resulta más fácil recoger un número mayor de variables por individuo, así como tener un mejor control de la calidad del proceso de recogida de datos.

Si se dispone de información para todas las unidades de la población, el parámetro poblacional de interés quedará determinado con total precisión. Sin embargo, si se emplea únicamente una fracción del total, el parámetro poblacional desconocido ha de estimarse a partir de la muestra, con el consiguiente error derivado tanto por el carácter parcial de la muestra como por su posible falta de representatividad poblacional. La teoría de muestreo persigue un doble objetivo.

1. Por un lado, estudia las técnicas que permiten obtener muestras representativas de la población de forma eficiente.
2. Por otro lado, la teoría de muestreo indica cómo utilizar los resultados del muestreo para estimar los parámetros poblacionales, conociendo a la vez el grado de incertidumbre de las estimaciones. Así, la teoría de muestreo pretende dar respuesta a varias preguntas de interés:

- ¿Cómo se eligen a los individuos que componen la muestra?
- ¿Cuántos individuos formarán parte de la muestra?
- ¿Cómo se cuantifican las diferencias existentes entre los resultados obtenidos en la muestra y los que hubiéramos obtenido si el estudio se hubiera llevado a cabo en toda la población?

Estas cuestiones están estrechamente relacionadas entre sí. Así, por ejemplo, al aumentar el tamaño muestral aumenta la exactitud en las estimaciones. En el presente apartado se discuten los principales tipos de muestreo probabilístico, así como la estimación en el muestreo aleatorio simple. Antes de ello, es conveniente revisar la definición de algunos conceptos que se utilizan de forma repetida a lo largo del capítulo:

- Población o universo muestral es la colección de elementos o unidades de análisis acerca de los cuales se desea información. Con frecuencia, no se puede obtener información de toda la población, sino tan sólo de unidades que cumplen una serie de características

(criterios de inclusión/exclusión). La población marco es aquella sobre la que es posible obtener información. La muestra se obtiene de la población marco, por lo que debe recordarse que las conclusiones extraídas de la muestra son generalizables a la población marco y no necesariamente a la población de inicio o universo.

- Dentro del proceso de selección de una muestra, la población suele dividirse en unidades de muestreo, que deben constituir una partición de toda la población. Estas unidades de muestreo pueden coincidir con las unidades de análisis, pero también pueden estar constituidas por un conjunto de distintas unidades de análisis.

▶ Ejemplo 1.

Supongamos que se desea estudiar la capacidad funcional de una población de ancianos institucionalizados. Para ello, se dispone de una lista de residencias, algunas de las cuales se seleccionan para el estudio. Dentro de cada residencia seleccionada, se eligen a su vez algunos ancianos que formarán parte de la muestra definitiva. En tal caso, la selección de la muestra se habría realizado en dos etapas: las residencias constituirían las unidades de muestreo de primera etapa y los ancianos (unidades de análisis) serían las unidades de muestreo de segunda etapa.

- Muestreo probabilístico es aquel en que todas las unidades de la población tienen una probabilidad conocida y no nula de ser seleccionadas para la muestra. El muestreo probabilístico minimiza la probabilidad de sesgos (si el tamaño muestral no es muy limitado, la muestra será muy probablemente representativa de la población) y permite cuantificar el error cometido en las estimaciones como consecuencia de la variabilidad aleatoria. La teoría del muestreo se basa fundamentalmente en el muestreo probabilístico, ya que otros tipos de muestreo (de conveniencia, por cuotas) están sujetos

a una mayor probabilidad de sesgos y es más difícil extrapolar los resultados a la población.

- En el muestreo con reposición, cada vez que se elige un nuevo elemento muestral se dispone de toda la población para realizar la selección, mientras que en el muestreo sin reposición los elementos que ya han aparecido en la muestra no están disponibles para ser elegidos de nuevo. En el muestreo con reposición, por tanto, una unidad poblacional puede aparecer más de una vez en la muestra. En la práctica, el muestreo suele realizarse sin reposición. No obstante, si el tamaño de la población es muy grande con respecto al tamaño muestral, la probabilidad de que un elemento de la población sea elegido más de una vez en la muestra es tan pequeña que ambos tipos de muestreo son similares.

Principales tipos de muestreo probabilístico

En este apartado se describen brevemente los principales procedimientos probabilísticos de selección de muestras, tales como los muestreos aleatorios simple, sistemático, estratificado, por conglomerados y polietápico. Un tratamiento más extenso de estos procedimientos puede encontrarse en los libros de muestreo referenciados al final del tema.

Muestreo aleatorio simple

El muestreo aleatorio simple es el más sencillo y conocido de los distintos tipos de muestreo probabilístico. Supongamos que se pretende seleccionar una muestra de tamaño n a partir de una población de N unidades. Un muestreo aleatorio simple es aquel en el que cualquier subconjunto de tamaño n tiene la misma probabilidad de ser seleccionado. Puede probarse que el muestreo aleatorio simple es un procedimiento equiprobabilístico;

es decir, todas las unidades de la población tienen la misma probabilidad n/N de ser elegidas en la muestra.

Para la selección de una muestra aleatoria simple, se enumeran previamente las unidades del universo o población de 1 a N y a continuación se seleccionan n números distintos entre 1 y N utilizando algún procedimiento aleatorio, típicamente mediante una tabla de números aleatorios o un generador de números aleatorios por ordenador.

- Las tablas de números aleatorios (26) son tablas con los dígitos 0, 1, 2, ..., 9, donde cada dígito tiene la misma probabilidad de ocurrir y el valor de un dígito concreto es independiente del valor de cualquier otro dígito de la tabla.
- La mayoría de los programas de análisis estadístico contienen generadores de números aleatorios. Estos generadores producen grandes secuencias de dígitos pseudoaleatorios, que satisfacen aproximadamente las mismas propiedades de aleatoriedad enunciadas anteriormente.

▶ Ejemplo 2:

Supongamos que, en el ejemplo anterior, se dispone de una lista completa de los $N = 875$ ancianos institucionalizados en dicha población, de los cuales se desean seleccionar $n = 10$. La selección de una muestra aleatoria simple de este tamaño puede realizarse a partir de la Tabla de números aleatorios (26), como sigue. Comenzando en cualquier lugar de esta tabla y leyendo grupos de 3 dígitos en cualquier dirección, seleccionar los 10 primeros números distintos entre 1 y 875. Por ejemplo, empezando en el primer dígito de la tercera fila y de izquierda a derecha, estos números son: 339, 117, 619, 68, 440, 788, 696, 716, 183 y 546. Notar que los números 897 y 898 han sido descartados por ser superiores a $N = 875$. La muestra alea-

toria simple estaría así constituida por aquellos ancianos de la población numerados previamente por estos 10 valores.

Puede probarse que, como el muestreo aleatorio simple es un procedimiento equiprobabilístico, una media o una proporción poblacional se estiman simplemente mediante la media o proporción muestral. La estimación de parámetros poblacionales a partir de una muestra aleatoria simple, así como la varianza o error de las estimaciones, se discutirá en detalle al final de este apartado.

Muestreo sistemático

En ocasiones, la numeración consecutiva de las unidades de la población y la posterior selección de una muestra aleatoria simple resultan muy laboriosas. En tales circunstancias, un procedimiento alternativo más sencillo es el llamado muestreo sistemático. Bajo este procedimiento, no siempre es necesario numerar previamente los elementos de la población, sino que basta con disponer de alguna ordenación explícita (por ejemplo, orden de archivo de historias clínicas o visitas sucesivas de pacientes a una consulta médica).

Para la selección de una muestra sistemática de tamaño n de una población de N unidades, se elige aleatoriamente un número de arranque r entre 1 y k, donde k es la parte entera de N/n, y a partir del elemento que ocupa el lugar r, se toman los restantes elementos en intervalos de amplitud k hasta completar la muestra deseada. Así, la muestra estará constituida por los elementos ordenados en los lugares r, $r + k$, $r + 2k$, ..., $r + (n - 1)k$. Como en general N no es múltiplo de n, este método de selección no es necesariamente equiprobabilístico (si N/n no es un número entero, las unidades comprendidas entre los lugares $nk + 1$ y N nunca podrán formar parte de la muestra). Una modificación a este procedimiento, que garantiza la obtención de una muestra equiprobabilística, consiste en seleccionar el número

aleatorio de arranque r entre 1 y N, y tomar cada k-ésima unidad a partir de ahí, continuando en el primer elemento al alcanzar el final de la lista

- Ejemplo 3:

Para seleccionar una muestra sistemática de tamaño $n = 10$ de la población de $N = 875$ ancianos institucionalizados, se calcula primero la amplitud del intervalo de selección como la parte entera de $N/n = 875/10 = 87{,}5$; es decir, $k = 87$. Si se seleccionara el número de arranque r entre 1 y 87, el último anciano seleccionado ocuparía en el lugar $r + (n - 1)k = r + (10 - 1)87 = r + 783$, que sería siempre inferior o igual a 870 (dado que $r \leq 87$). En consecuencia, los ancianos en los lugares 871 a 875 nunca podrían formar parte de la muestra. Para asegurar un muestreo equiprobabilístico, el número de arranque se selecciona aleatoriamente entre 1 y 875. Suponiendo que este número de arranque fue $r = 427$ y tomando intervalos de amplitud $k = 87$, la muestra sistemática quedaría integrada por aquellos ancianos en los lugares 427, 514, 601, 688, 775, 862, 74, 161, 248 y 335.

En el muestreo sistemático, la ordenación de los elementos de la población determinará las posibles muestras. En consecuencia, este orden ha de estar exento de cualquier periodicidad relacionada con las variables a estudio. Así, por ejemplo, si para estimar el nivel de contaminación atmosférica en una ciudad se toma una muestra sistemática de días con $k = 7$, la muestra estará formada por los mismos días de la semana y presentará un claro sesgo por falta de representatividad. No obstante, estas periodicidades son muy infrecuentes en la práctica y pueden solventarse con facilidad (en el ejemplo anterior, bastaría con utilizar un intervalo de selección distinto de 7). En general, si la ordenación de las unidades de la población es esencialmente aleatoria, la estimación de parámetros y sus correspondientes errores en un muestreo sistemático se realiza igual que en un muestreo aleatorio simple.

Muestreo estratificado

En los muestreos anteriores, las muestras se seleccionan por procedimientos puramente aleatorios. Así, si el tamaño muestral es suficientemente grande, la muestra será muy probablemente representativa de la población. Sin embargo, no existe una garantía absoluta de que la muestra finalmente seleccionada sea representativa para cualquier variable de interés. Cuando se desea asegurar la representatividad de determinados subgrupos o estratos de la población, la alternativa más sencilla es seleccionar por separado distintas submuestras dentro de cada estrato. Este procedimiento de selección se conoce como muestreo estratificado.

Los estratos han de definir subgrupos de población que sean internamente homogéneos con respecto a la característica o parámetro de interés y, por tanto, heterogéneos entre sí. En la práctica, los estratos se definen en función de variables fáciles de medir previamente y relevantes para el tema objeto de estudio (por ejemplo, edad, sexo, raza o área geográfica de residencia). En general, el número de estratos ha de ser reducido (rara vez resulta eficiente utilizar más de 5 estratos) y el tamaño por estrato no debe ser muy pequeño.

Para la selección de una muestra estratificada de tamaño n, la población de N unidades se divide en K estratos de tamaños $N1, N2, ..., NK$, cuya suma es igual a N. Los estratos son mutuamente excluyentes y exhaustivos, de tal forma que cada elemento de la población pertenece a uno y sólo a uno de los estratos. Una vez determinados estos estratos, se selecciona por separado una muestra de cada estrato de tamaño $n1, n2, ..., nK$, respectivamente, cuya suma será igual al tamaño total n de la muestra. La selección dentro de cada estrato suele realizarse por muestreo aleatorio simple o sistemático, y el procedimiento se denomina entonces muestreo aleatorio estratificado. En el muestreo estratificado, es necesario determinar cómo se distribuye el tamaño muestral total n entre los distintos estratos;

es decir, la asignación de los tamaños muestrales $n1, n2, ..., nK$. Aunque existen distintos tipos de asignación en función del tamaño y varianza por estrato (véase referencias al final del tema), nos limitaremos aquí a la asignación proporcional, que es el procedimiento utilizado con mayor frecuencia. En la asignación proporcional, la muestra total se reparte entre los estratos de forma proporcional al tamaño de cada estrato en la población. Así, como la proporción poblacional en cada estrato es Nk/N, el tamaño muestral del estrato k-ésimo será:

$$\mathrm{Nk} = \mathrm{n}\,\frac{Nk}{N}$$

Resulta inmediato probar que esta asignación da lugar a una muestra equiprobabilística.

- Ejemplo 4:

La capacidad funcional de los ancianos disminuye en gran medida con la edad. Supongamos que, de los $N = 875$ ancianos institucionalizados, se sabe que el 60% tienen menos de 75 años ($N1 = 525$) y el restante 40% tienen 75 o más años ($N2 = 350$). Para simplificar la exposición, supongamos además que los ancianos menores de 75 años corresponden a los primeros 525 números de la lista. Así, de los $n = 10$ ancianos seleccionados por muestreo aleatorio simple en el Ejemplo 2, la mitad resultaron ser mayores de 75 años. Esto es, por simple variabilidad aleatoria, los mayores de 75 años están ligeramente sobrerrepresentados en la muestra y, en consecuencia, la capacidad funcional media obtenida de esta muestra podría infraestimar la verdadera capacidad funcional de los ancianos institucionalizados. Para asegurar una mejor representatividad muestral por edad, podría realizarse un muestreo estratificado con asignación proporcional a ambos estratos de edad. Es decir, de la muestra de tamaño $n = 10$, seleccionaríamos 6 ancianos menores de 75 años ($n1 = nN1/N = 10{\cdot}0{,}6 = 6$) y 4 mayores de 75 años ($n2 = nN2/N = 10{\cdot}0{,}4 = 4$). Utilizando un muestreo aleatorio simple

dentro de cada estrato, los 6 números seleccionados entre 1 y 525 fueron 505, 493, 24, 402, 371 y 265, y los 4 números seleccionados entre 526 y 875 fueron 851, 820, 717 y 696. La muestra estratificada proporcional estaría formada por los 10 ancianos correspondientes a dichos números.

Cabe reseñar aquí dos características importantes del muestreo estratificado. Por un lado, la asignación proporcional es la única que produce muestras equiprobabilísticas y, en consecuencia, la media y proporción poblacional se estiman mediante la media y la proporción muestral. Para cualquier otra asignación, la estimación de parámetros poblacionales requiere de la inclusión de pesos para cada observación muestral (típicamente, el inverso de la probabilidad de selección). Por otra parte, para un mismo tamaño muestral, el muestreo estratificado facilita estimaciones ligeramente más precisas (con menor error) que el muestreo aleatorio simple. Este resultado es debido a que, cuanto más homogéneos sean los estratos, más precisas serán las estimaciones en dichos estratos y esto redundará en una mayor precisión de las estimaciones para toda la población.

Muestreo por conglomerados

La aplicación de los diseños muestrales anteriores requiere de la enumeración u ordenación de todos los elementos de la población. Sin embargo, a menudo no se dispone de una lista completa o, aun disponiendo de tal lista, resulta muy costoso obtener información de las unidades muestreadas. Por ejemplo, si se seleccionara una muestra aleatoria simple de 1000 individuos de una gran ciudad, los individuos seleccionados estarían muy dispersos y la recogida de información sería extraordinariamente laboriosa. En tales circunstancias, una alternativa consiste en clasificar a la población en grupos o conglomerados, para así seleccionar una muestra de estos conglomerados y después tomar a todas o a una parte de las unidades

incluidas dentro de los conglomerados seleccionados. Este método de selección se denomina muestreo por conglomerados y presenta dos ventajas fundamentales:

- Este muestreo es la única alternativa posible cuando no se dispone de una lista con todas las unidades de la población. En el muestreo por conglomerados, únicamente es necesario contar con listas de las unidades que integran los conglomerados seleccionados.
- Aun cuando otras técnicas de muestreo sean posibles, con frecuencia el muestreo por conglomerados resulta más económico, ya que las unidades muestrales están concentradas en los conglomerados seleccionados.

Se debe observar que, a diferencia de la estratificación, donde interesa que los estratos sean lo más homogéneos posible, los conglomerados deben ser heterogéneos: en cada conglomerado debe haber unidades representativas de toda la población, de lo contrario se perdería información al seleccionar únicamente algunos de ellos. El número de conglomerados es típicamente elevado, de los cuales suele seleccionarse un número relativamente pequeño para resolver el problema de la dispersión muestral. Supongamos que se pretende extraer una muestra de tamaño n a partir de una población de N unidades agrupadas en M conglomerados de tamaños N1, N2, ..., NM. Entre los distintos métodos de selección por conglomerados, el muestreo por conglomerados con probabilidad proporcional a su tamaño resulta particularmente útil en la práctica. Para llevar a cabo este muestreo, se procede como sigue:

A. Ordenar arbitrariamente los conglomerados y calcular los tamaños acumulados. Estos tamaños acumulados delimitarán, para cada conglomerado, un rango de valores de amplitud igual a su tamaño poblacional.

B. Si se pretende seleccionar m conglomerados, extraer una muestra sistemática de tamaño m entre 1 y N. Los conglomerados seleccionados serán aquellos cuyo rango incluya alguno de los valores muestreados.

C. Dentro de cada conglomerado seleccionado, obtener una muestra aleatoria simple o sistemática de tamaño n/m.

- Ejemplo 5:

Con cualquiera de las técnicas de muestreo utilizadas en los ejemplos anteriores, la muestra incluiría muy probablemente ancianos institucionalizados en múltiples residencias, con el consiguiente inconveniente en la recogida de información. Supongamos que los $N = 875$ ancianos institucionalizados se encuentran distribuidos en $M = 15$ residencias con los tamaños especificados en la tabla 15. Para optimizar el trabajo de campo, se decide extraer la muestra de tamaño $n = 10$ a partir de $m = 2$ residencias (conglomerados) seleccionadas con probabilidades proporcionales a sus tamaños.

Tabla 15. Distribución del número de ancianos institucionalizados por residencia

Residencias	Tamaño (Ni)	Tamaño acumulado	Rango asignado
1	50	50	ene-50
2	30	80	51-80
3	35	125	81-115
4	70	185	116-485
5	55	240	186-240
6	45	285	241-285
7	125	410	286-410
8	80	490	411-490
9	20	510	491-510
10	100	610	511-610

11	65	675	611.675
12	35	710	576-710
13	400	750	711-750
14	75	825	751-825
15	50	875	826-875

Fuente: Pastor-Barrioso (10)

En primer lugar, se asigna a cada residencia un rango de valores de amplitud igual a su tamaño (tabla 15). A continuación, se extrae una muestra sistemática de tamaño 2 entre 1 y 875: si el número de arranque resultó ser 316, los valores muestreados son 316 y 753 (ver apartado de muestreo sistemático). Así, como el valor 316 está incluido dentro del rango asignado a la residencia 7 y el valor 753 en el rango de la residencia 14, resultan seleccionadas las residencias 7 y 14. Para completar la muestra de $n = 10$ ancianos, se extraen finalmente muestras aleatorias simples de tamaño $n/m = 10/2 = 5$ de las residencias 7 y 14. De los 125 ancianos institucionalizados en la residencia 7, se seleccionaron los números 74, 23, 104, 111 y 57; y de los 75 ancianos de la residencia 14, los números 38, 51, 25, 34 y 41. En conclusión, la muestra total estará formada por los ancianos listados en los lugares 74, 23, 104, 111 y 57 de la residencia número 7, más aquellos que ocupan los lugares 38, 51, 25, 34 y 41 de la residencia número 14.

El muestreo por conglomerados con probabilidades proporcionales a sus tamaños facilita muestras equiprobabilísticas, así la media y la proporción poblacional pueden estimarse mediante sus correspondientes funciones muestrales. En general, para un tamaño muestral constante, la precisión de las estimaciones en un muestreo por conglomerados es menor que en un muestreo aleatorio simple. Las unidades de un mismo conglomerado suelen estar correlacionadas y, en consecuencia, aportan menos información que los elementos seleccionados de forma más dispersa mediante un muestreo aleatorio simple.

Muestreo polietápico

Los diseños muestrales empleados en la práctica se realizan combinando las técnicas descritas anteriormente. En muchas situaciones, resulta más apropiado obtener la muestra final en diferentes etapas o pasos. En un muestreo polietápico, la población se divide en grupos exhaustivos y mutuamente excluyentes, que constituyen las llamadas unidades de primera etapa; cada una de ellas se desagrega a su vez en subgrupos o unidades de segunda etapa, y así sucesivamente, hasta llegar en una última etapa a los elementos o unidades de análisis. La selección de unidades en cada una de las etapas se realiza mediante una técnica de muestreo diferente y la muestra final será la resultante de aplicar sucesivamente cada una de estas técnicas.

- Ejemplo 6:

En el ejemplo anterior se seleccionaron 2 de las 15 residencias y, dentro de cada residencia seleccionada, se eligieron a su vez 5 ancianos para formar la muestra definitiva. Este procedimiento de selección es, de hecho, un muestreo bietápico: las residencias constituirían las unidades de muestreo de primera etapa y los ancianos serían las unidades de muestreo de segunda etapa.

Determinación del tamaño muestral

Las inferencias poblacionales derivadas a partir de una muestra conllevan indefectiblemente un margen de error. Así, en el diseño de un estudio epidemiológico o clínico, es necesario plantearse de antemano el número de sujetos que deben ser estudiados para responder a la pregunta de investigación con un grado razonable de certidumbre. La determinación a priori del tamaño muestral es una parte importante del diseño de un estudio por distintos motivos:

- Permite concretar la hipótesis de trabajo. El investigador ha de precisar la hipótesis principal del estudio y, en función de su experiencia, investigaciones previas o estudios piloto, especificar la magnitud de efecto clínica o biológicamente relevante que se pretende detectar.
- Permite evaluar la factibilidad del estudio. Una de las limitaciones más frecuentes en los estudios epidemiológicos es la imposibilidad de reclutar un número suficiente de pacientes, bien sea por limitaciones en los recursos económicos, en el número de pacientes disponibles o en el tiempo de duración del estudio.
- Previene la obtención de resultados no concluyentes.

La precisión de una estimación y la potencia estadística de un contraste de hipótesis aumentan conforme aumenta el tamaño muestral, de tal forma que una muestra insuficiente dará lugar a estimaciones imprecisas y contrastes de baja potencia. Desde un punto de vista puramente teórico, basta con aumentar el tamaño muestral para obtener estimaciones arbitrariamente precisas o para detectar como estadísticamente significativo cualquier efecto por pequeño que sea. Aun cuando esto sea posible en la práctica, la utilización de muestras excesivamente grandes es ineficiente, ya que la posible detección de efectos trivialmente pequeños y de escasa utilidad práctica no justificaría los recursos empleados. En último término, el objetivo de la determinación a priori del tamaño muestral consiste en estimar la muestra mínima necesaria para asegurar estimaciones razonablemente precisas o para tener una potencia suficiente en la detección de efectos clínicamente relevantes.

El cálculo del tamaño muestral requiere de información previa a la realización del estudio. Estos datos suelen proceder de investigaciones previas relacionadas y, en la medida de lo posible, han de ajustarse a unas hipótesis de trabajo verosímiles. En cualquier caso, las asunciones realizadas en el cálculo del tamaño muestral pueden diferir de los resultados posteriores

del estudio y, en consecuencia, estas determinaciones deben servir como guía orientativa más que como norma rígida para la estimación del tamaño muestral. Conviene apuntar también que la muestra resultante se refiere al número de sujetos necesarios para el análisis y no a los inicialmente incluidos. Así, la muestra estimada ha de incrementarse en previsión de las posibles pérdidas de sujetos que pudieran ocurrir en el estudio.

Tamaño muestral para la estimación de una media

En este apartado se presentan las fórmulas para determinar el tamaño muestral necesario para obtener estimaciones fiables de un parámetro poblacional (típicamente la media de una variable continua o la proporción de sujetos con una determinada característica) a partir de una única muestra. Esta situación concierne esencialmente a los estudios descriptivos o transversales. El objetivo se centra en calcular el tamaño muestral mínimo necesario para estimar el parámetro poblacional con un determinado grado de precisión, que suele cuantificarse mediante la amplitud del intervalo de confianza

A partir de la aproximación normal N(μ, σ2/n) a la distribución de una media muestral ẍ puede construirse un intervalo de confianza al 100(1 – α)% para la media poblacional μ como ẍ ± Z1- α /2 σ/$\sqrt{n}$. Obsérvese que este intervalo incluye la desviación típica poblacional σ en lugar de su estimación muestral, ya que la determinación del tamaño de una muestra precede a su selección y, en consecuencia, no se dispone de información muestral. La precisión de la estimación queda entonces determinada por la amplitud del intervalo de confianza o, más concretamente, por la distancia del centro a los límites del intervalo:

$$\delta = ¿Z1- \alpha /2 \frac{\sigma}{\sqrt{n}}$$

de donde puede despejarse el tamaño muestral n para obtener:

$$n = \frac{Z1-\alpha/2\,\sigma^2}{\delta^2}$$

De esta expresión se desprende que el tamaño muestral para la estimación de una media poblacional depende de tres elementos, que deben ser determinados de antemano para poder aplicar la fórmula:

- El nivel de confianza 100(1 – α) %. Cuanto mayor sea este nivel de confianza, mayor será el tamaño muestral. En la práctica, suele utilizarse una confianza del 95% (α = 0,05), de tal forma que el percentil de la distribución normal estandarizada es Z1- α /2 = Z 0,975 = 1,96
- La varianza poblacional σ2. Cuanto más dispersa sea una variable, mayor será la muestra necesaria para describirla aceptablemente. Se requiere, por tanto, de un valor aproximado de la varianza de la variable a estudio, que suele obtenerse a partir de trabajos similares ya realizados o de un estudio piloto.
- La precisión deseada . El tamaño muestral será tanto mayor cuanto mayor sea la precisión exigida a la estimación (es decir, cuanto menor sea). El criterio para establecer la precisión de una estimación ha de fundamentarse en el conocimiento previo sobre la magnitud aproximada del parámetro. Así, por ejemplo, una precisión de un kilogramo puede ser aceptable para estimar el peso medio en personas adultas, pero resulta claramente insuficiente en recién nacidos.

▶ Ejemplo 7:

En un pequeño estudio piloto realizado en personas adultas de una determinada población, la media y la desviación típica de la presión arterial sistólica resultaron ser 130 y 20 mm Hg, respectivamente. Utilizando esta información preliminar, se planea obtener una muestra aleatoria simple de mayor tamaño para estimar el nivel medio de presión arterial sistólica

con una precisión de ±2 mm Hg. Asumiendo un nivel de confianza del 95% y una desviación típica similar a la del estudio piloto.

Se tiene:

$$n = \frac{1{,}96^2 20^2}{2^2} = 384{,}16 \approx 385 \text{ personas;}$$

es decir, se requerirían aproximadamente 385 sujetos para estimar la presión arterial sistólica media de esta población con una precisión de ±2 mm Hg. Obsérvese que el tamaño muestral aumenta de forma cuadrática con la precisión deseada, de tal forma que para el doble de precisión $\delta = 1$ mm Hg, el tamaño muestral mínimo necesario sería cuatro veces mayor, esto es:

$$n = \frac{1{,}96^2 20^2}{1^2} = 1536{,}64 \approx 1.537 \text{ personas}$$

El teorema central del límite

Sea X una variable aleatoria cualquiera con media μ y varianza σ^2; entonces, si el tamaño muestral es lo suficientemente grande (habitualmente $n \geq 30$), la media muestral $\ddot{X}$ se distribuirá asintóticamente según un modelo de probabilidad normal con la misma media que la variable original μ y con varianza $\frac{\sigma^2}{n}$. Este resultado puede expresarse de las siguientes formas equivalentes:

$$\ddot{X}\, N\left(\mu, \frac{\sigma}{\sqrt{n}}\right); \frac{\ddot{X} - \mu}{s/\sqrt{n}} \sim N(0,1)$$

La mayoría de los autores aceptan que con n ≥ 30 es suficiente para que la aproximación a la normal sea adecuada, lo cual ha sido comprobado en múltiples experimentos de simulación con ordenador. La aproximación a la normal mejora según aumenta n (27).

Estimación en el muestreo aleatorio simple

Una vez descritas las principales técnicas de muestreo probabilístico, nos ocuparemos a continuación de la «estimación de parámetros poblacionales». En adelante, se asume que la muestra se ha obtenido mediante un

muestreo aleatorio simple a partir de una población de tamaño grande. El cálculo del valor exacto de un parámetro poblacional requiere de conocimientos del valor de la variable objeto de estudio para todos y cada uno de los elementos de la población. Como se ha comentado anteriormente, en la mayoría de las ocasiones no se dispone de esta información, sino que se cuenta tan sólo con una muestra. A la función de los valores de una muestra que permite hacerse una idea acerca del valor del parámetro poblacional se le denomina **«**estimador**»**, y al resultado de aplicar dicha función a una determinada muestra se le llama estimación.

Cuando lo que se pretende es estimar o cuantificar el valor de un parámetro desconocido de la población podría procederse de dos maneras diferentes:

1. Proporcionando un único valor para el parámetro.
2. Proporcionando un intervalo que contendrá al verdadero valor del parámetro de la población con una determinada seguridad o confianza.

La diferencia entre estas dos aproximaciones radicaría en el hecho de que, mientras que en el primero de los casos «estimación puntual» no se proporciona ningún tipo de información sobre la magnitud probable del parámetro objeto de estudio ni del error que pudiera cometerse (es evidente que si se utiliza la información de una pequeña parte de la población para estimar el valor del parámetro correspondiente a toda la población, el valor de la estimación será aproximado y, por tanto, sujeto a error), mediante la aproximación por intervalos (estimación ***por*** intervalos o intervalos ***de*** confianza) sí se responde a estas cuestiones.

Algunos parámetros y sus estimadores puntuales

Algunos parámetros y sus respectivos estimadores puntuales son:

- μ representa la media poblacional de una variable cuantitativa y su estimador puntual es la media muestral $\ddot{x}$

- **σ** representa la desviación típica poblacional de una variable cuantitativa y su estimador puntual es la desviación típica muestra *S* (de la misma forma, el estimador de la varianza poblacional σ^2 es la varianza muestral S^2).
- P representa el porcentaje de valores de una categoría de interés en una variable categórica y su estimador puntual es el porcentaje de esta característica en la muestra *P^*

Estimación puntual de una media poblacional

Supongamos que *x*1, *x*2, ..., *xn* son los valores obtenidos en una muestra de tamaño *n* para una variable con media poblacional *μ* y varianza σ^2 desconocidas. Un estimador natural de la media poblacional *μ* es la media muestral Ẍ:

$$\ddot{X} = \frac{1}{n}\sum_{i=1}^{n} xi$$

Esta media muestral quedará completamente determinada una vez obtenida la muestra, pero el valor de la estimación variará en función de la muestra seleccionada. Así, la media muestral puede considerarse como una variable aleatoria, cuyo valor dependerá de la muestra finalmente seleccionada de entre todas las posibles muestras de tamaño n de la población de referencia. A la distribución de los valores de la Ẍ sobre todas las posibles muestras del mismo tamaño se le denomina distribución muestral de la Ẍ. Las razones teóricas que justifican la utilización de la media muestral como estimador de la media poblacional, frente a otros posibles estimadores, se basan en esta distribución muestral.

Considérese los ejemplos siguientes:

► Ejemplo 8**:**

Se pretende estimar el tiempo medio de estancia en el hospital de los pacientes afectados por una determinada patología. Se cuenta con infor-

mación correspondiente a 450 pacientes, procedentes de varios hospitales de la geografía del país, a partir de los cuales se obtuvo un tiempo medio de estancia de 8 días con una desviación típica de 1,5 días.

- Ejemplo 9.

Una nueva técnica quirúrgica fue practicada con éxito en 40 de los 50 pacientes intervenidos y seleccionados al azar entre los afectados por una determinada patología. ¿Cuál sería la proporción de éxito en la intervención en la población de afectados por dicha patología?

A partir de los datos del ejemplo 8, podría utilizarse la media del tiempo de estancia en el hospital de los 450 pacientes estudiados ($\ddot{X}$) para estimar la media de estancia de todos los individuos afectados por esa patología (μ). El estimador puntual de la media poblacional sería 8 días de estancia ($\ddot{x}$=8).

En el ejemplo 9 se pretende estimar el valor de la proporción poblacional de éxito de una determinada intervención quirúrgica a partir de la información contenida en una muestra de 50 individuos. La proporción de éxito tras la intervención observada se calcularía de la siguiente forma:

$$\hat{\rho} = \frac{r}{n} = \frac{40}{50} = 0{,}80 \approx 80\ \%,$$

Donde ***r,*** es el número de individuos de la muestra en los que la intervención ha sido un éxito y ***n*** el tamaño de la muestra. Esto significa que en el 80% de los pacientes la intervención (nueva técnica quirúrgica) ha sido un éxito. ¿Sería factible utilizar la proporción muestral como estimador puntual de la proporción poblacional de éxito en la intervención?

Estimación mediante intervalos de confianza

El proceso de inferencia es aquel mediante el cual se pretende estimar el valor de un parámetro a partir del valor de un estadístico. Esta estimación puede ser puntual o bien por intervalo. La mejor estimación puntual de un

parámetro es simplemente el valor del estadístico correspondiente, pero es poco informativa porque la probabilidad de no dar con el valor correcto es muy elevada, es por eso que se acostumbra a dar una estimación por intervalo, en el que se espera encontrar el valor del parámetro con una elevada probabilidad. Esta estimación recibe el nombre de estimación mediante intervalos de confianza.

La estimación por intervalos de confianza consiste en determinar un posible rango de valores o intervalo (***a; b***), en el que, con una determinada probabilidad, sus límites contendrán el valor del parámetro poblacional que andamos buscando. Para cada muestra obtendremos un intervalo distinto que, para el ***X*** % de ellas, contendrá el verdadero valor del parámetro. A este intervalo se le denomina *intervalo de confianza*.

Evidentemente esta técnica no tiene por qué dar siempre un resultado correcto, tal y como hemos comentado para algunas muestras el intervalo correspondiente contendrá el verdadero valor del parámetro y para otras no. A la probabilidad de que hayamos acertado al decir que el intervalo contiene al parámetro se la denomina «nivel de confianza» (o simplemente confianza).

En este apartado se estudia la estimación por intervalos de confianza para la media poblacional de una variable cuantitativa (μ) y para una proporción o porcentaje (***P***) de una característica de interés en la población a partir de una variable categórica. En general siempre buscaremos estimar cantidades poblacionales (μ, P) y no sus equivalentes muéstrales (**ẍ; *P*** ^), ya que de estos últimos conoceremos sus valores exactos y en consecuencia no necesitan ser estimados (se conocen sin ningún tipo de ambigüedad). El reto que nos proponemos es, a partir de los valores muéstrales, conocer tanto como sea posible los valores poblacionales. Para ello, utilizaremos las distribuciones de los correspondientes estimadores:

- Intervalo de confianza para una media poblacional μ:

Utilizaremos la distribución en el muestreo del estadístico ẍ.

1. Si la desviación típica poblacional $\boldsymbol{\sigma}$ es conocida podemos utilizar la expresión:

$$\ddot{x} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \rightarrow \frac{\ddot{X}-\mu}{s/\sqrt{n}} \sim N(0,1)$$

2. Si la desviación típica poblacional $\boldsymbol{\sigma}$ es desconocida (que es lo habitual), y por tanto a lo sumo conoceremos S que es un estimador de $\boldsymbol{\sigma}$. En ese caso, debemos introducir una nueva distribución llamada «*Distribución t de Student*», pues la distribución del estadístico ẍ cuando usamos la desviación típica muestral S es:

$$\frac{\ddot{X}-\mu}{s/\sqrt{n}} \sim t_{n-1,}$$

donde t n-1 representa la distribución t de Student con n−1 grados de libertad. Esta distribución se estudiará en el siguiente punto de este tema.

- Intervalo de confianza para un porcentaje poblacional P:

Utilizaremos la distribución en el muestreo del estadístico P^.

$$P\hat{} \sim N\left(P\hat{}\, \sqrt{P^{\square}} \rightarrow \frac{P^{-P}}{\sqrt{P^{\square}}} \sim N(0,1)\right.$$

- Distribución t-Student

Cuando nos disponemos a hacer inferencia sobre la media poblacional (μ) a partir de la media muestral (ẍ), resulta lógico utilizar el Teorema Central del Límite, es decir, que: $\frac{\ddot{X}-\mu}{s/\sqrt{n}} \sim N(0,1)$

En esta expresión $\boldsymbol{\sigma}$ representa la desviación típica poblacional, de la que habitualmente no tendremos información sobre ella, es decir será un valor desconocido. Si tenemos una muestra de tamaño ***suficientemente***

grande, podemos estimar el valor de la desviación típica poblacional σ, a partir de la desviación típica muestral *S* con una precisión *aceptable*. Por tanto, la expresión anterior seguirá siendo válida. Pero si la muestra que tenemos no es *suficientemente grande*, la estimación que tendremos de σ a partir de *S* no será lo suficientemente precisa, y por tanto la expresión anterior no será válida. En consecuencia, si σ no es conocida y el tamaño muestral que disponemos no es *suficientemente grande*, la expresión , que es la que realmente usaremos para calcular el intervalo de confianza que pretendemos obtener, no seguirá una distribución $N(0; 1)$ sino otra distribución similar (pero diferente), una distribución *t de Student*.

La distribución *t de Student* es una distribución con las siguientes características:

- Forma de campana.
- La máxima probabilidad se concentra alrededor del valor 0 (que es su media, moda y mediana) y disminuye a medida que nos alejamos de este valor central.
- Su forma se define por un parámetro "*g*" llamado *grados de libertad*, y que modula la mayor o menor variabilidad de los valores de esta distribución.

Como consecuencia de las características anteriores resulta que la distribución t tiene una forma muy similar a la distribución Normal Estándar, pero en función de los grados de libertad cambia su forma. A medida que aumentan los grados de libertad, la distribución *t* se va aproximando a la distribución Normal estándar.

Tablas de probabilidad para la distribución t de Student

Al igual que para la distribución Normal Estándar, para la distribución *t de Student* existen tablas de probabilidad que facilitan sus cálculos (28). Cada

fila de esta tabla se refiere a un número de grados de libertad diferente, que aparecen en la primera columna. A su vez cada una de las columnas de la tabla corresponde a un valor concreto de probabilidad. Para cada combinación de la fila y columna la tabla reproduce aquel valor que para los grados de libertad correspondientes deja a su izquierda la probabilidad determinada por la columna a la que pertenece.

Ejemplo de estimación de un intervalo de confianza para la media poblacional μ

En el caso del ejemplo 7, podría construirse un intervalo al 95% de confianza para la media del tiempo de estancia hospitalario de la forma siguiente:

$$I_{0,95(\mu)} = \left[\ddot{x} - 1{,}96\frac{S}{\sqrt{n}}; \ddot{x} + 1{,}96\frac{S}{\sqrt{n}}\right] = \left[8 - 1{,}96\frac{1.5}{\sqrt{450}}; 8 + 1{,}96\frac{1.5}{\sqrt{450}}\right] = [7{,}86; 8{,}14]$$ donde

$\ddot{x}$ representa la media muestra (8 días) y S es la desviación estándar muestral, es decir, 1,5 días. ($S = \sqrt{\frac{\Sigma(x-\ddot{x})^2}{n-1}}$).

En este caso se concluiría que el tiempo medio de estancia en un hospital de los pacientes afectados por esa patología se situaría entre los 7,86 y los 8,14 días con una confianza o seguridad del 95%, es decir, con una posibilidad de error del 5%. Si bien la información proporcionada por el intervalo de confianza es evidentemente mayor, es importante destacar que para su construcción ha sido necesario contar con la media muestral como estimador puntual del parámetro (media de la población). En general, en el proceso de construcción de intervalos de confianza para un parámetro poblacional será necesario disponer del estimador puntual de dicho parámetro.

Intervalo de confianza para un porcentaje poblacional

En el ejemplo 8 se desea estimar un intervalo de confianza del 95% para el valor de la proporción poblacional de éxito de una determinada intervención quirúrgica a partir de la información contenida en una muestra de 50

individuos, donde 40 de las intervenciones han sido exitosas. La proporción de éxito tras la intervención observada se calcularía de la forma siguiente:

El estimador puntual, $\hat{P} = \frac{r}{n} = \frac{40}{50} = 0{,}80$,

El intervalo de confianza para la proporción poblacional está centrado en la proporción muestral $(\rho^{\square})$, y sus límites superior e inferior son:

$[\rho^{\square} - Z\alpha/2\sqrt{\frac{\rho^{(1-\rho^{\square})}}{n}}$; $\rho^{+¿¿} Z\alpha/2\sqrt{\frac{\rho^{(1-\rho^{\square})}}{n}}$, donde $Z\alpha/2$ representa el valor crítico correspondiente al grado de confianza 1-α de la distribución normal tipificada.

Así, para calcular un intervalo de confianza al 95%, se procede de la manera siguiente: Para $1-\alpha = 0{,}95 \rightarrow Z\alpha/2 = 1.96$

$$\left[0{,}80 - 1.96\sqrt{\frac{0{,}80(1-0{,}80)}{50}}\right]; \left[0{,}80 + 1.96\sqrt{\frac{0{,}80(1-0{,}80)}{50}}\right] = 0{,}6891;\ 0{,}9108$$

Luego, la proporción de éxito en la intervención quirúrgica en la población de afectados por la mencionada patología estará comprendida entre 0,6891 (68,91%) y 0.9108 (91,08%), con una confianza de 0,95 o del 95%. El error que se puede cometer en esta estimación es a = 0,05 o del 5%.

CAPÍTULO VII.

CONTRASTE DE HIPÓTESIS

CAPÍTULO VII. CONTRASTE DE HIPÓTESIS

Estadística inferencial

La estadística inferencial incluye tres grandes áreas: muestreo, estimación de parámetros y pruebas de hipótesis. La estadística inferencial es inferir resultados de una muestra. Se utiliza fundamentalmente cuando se trabaja con el contraste de hipótesis, mediante el análisis de correlación entre variables o de comparación entre grupos o mediciones (29).

Hipótesis estadística

Una hipótesis estadística es una afirmación respecto a una característica de una población. Las hipótesis se formularán normalmente como una declaración sobre una o más poblaciones, especialmente sobre sus parámetros. Ejemplos de hipótesis pueden ser: La edad media de los individuos con la enfermedad X es 50 años. El tratamiento X tiene el mismo efecto que un placebo (por tanto, no tiene valor terapéutico).

Tipos de hipótesis

Una hipótesis estadística puede ser:

- Relativa a una única población, sobre la que se observa una variable, X, y se toma una muestra (X1, X2, . . . , Xn).
- Relativa a más de una población, por ejemplo, relativa a dos poblaciones sobre las que se observan dos variables, X e Y , y se toman dos muestras (X1, X2, . . . , Xn) e (Y1, Y2, . . . , Ym).
- Paramétrica: Se asume un modelo paramétrico sobre la población y se plantea una hipótesis sobre uno o más parámetros, por ejemplo, se asume que $X \sim N(\mu, \sigma 2)$ y se plantea la hipótesis $\mu < 2$.

- No paramétrica: No se asume ningún modelo paramétrico sobre la población y se plantea una hipótesis sobre una característica de la misma, por ejemplo, $Pr(X \leq 7) = 0,5$.
- Simple: La hipótesis tiene un solo elemento, por ejemplo, $\mu = 7$.
- Compuesta: La hipótesis tiene más de un elemento, por ejemplo, $\mu \geq 7$.

Contrastes de hipótesis

Al igual que ocurre con la estimación de parámetros desconocidos de la población mediante intervalos de confianza, los contrastes de hipótesis también constituyen una herramienta para la realización de inferencias sobre determinados parámetros de la población objeto de estudio, a partir de una muestra observada de dicha población. El funcionamiento de esta segunda técnica inferencial se basa en la realización de una afirmación acerca de un parámetro de una o más poblaciones «hipótesis» y en el estudio de la compatibilidad entre esta afirmación y lo observado en la muestra.

En principio, cuanto mayor sea la discrepancia entre la hipótesis realizada y la información proporcionada por la muestra observada, mayor será la evidencia en contra de dicha hipótesis. Considérese los siguientes ejemplos:

► Ejemplo 1:

En un estudio se recabó información sobre el nivel de colesterol de un grupo de 46 pacientes seleccionados al azar de entre los afectados por una determinada patología. El promedio de nivel de colesterol fue de 235 mg/100 ml con una desviación típica de 28 mg/100 ml. ¿Proporcionan estos datos evidencia suficiente que indique que el nivel promedio de colesterol en este tipo de pacientes es superior a 220 mg/100 ml?

- Ejemplo 2:

De los 250 individuos a los que se les administró un determinado tratamiento, 180 respondieron de forma positiva. ¿Puede considerarse que la proporción de éxito del tratamiento es del 80%?

En el primer ejemplo, la población objeto de estudio está constituida por toda la población de pacientes afectados por la patología en cuestión. El parámetro sobre el que se realiza la afirmación que se pretende contrastar sería la media de nivel de colesterol en mg/100 ml. La afirmación que se realiza sobre el parámetro «hipótesis» sería que la media es igual o superior a 220 mg/100 ml.

En el segundo ejemplo, la población objeto de estudio la conformarían todos los individuos afectados o que pudieran estar afectados por el suministro de un determinado tratamiento. La afirmación que se realiza sobre el parámetro (hipótesis) es que la proporción poblacional es de 0,8 (o del 80%).

En ambos casos el resultado del contraste debe conducir a la confirmación o a la negación de la afirmación realizada.

Contraste de hipótesis sobre la media de una población

Para ilustrar mejor el procedimiento de realización de un contraste de hipótesis, así como los elementos que intervienen en el mismo, se utilizarán los datos del primero de los ejemplos anteriores.

Se cuenta con información sobre el nivel de colesterol de 46 pacientes en los que el promedio es de 235 mg/100 ml y la desviación típica de S = 28 mg/100 ml. A efectos de simplicidad se supondrá, en un primer momento, que lo que el investigador pretende demostrar es si el promedio de colesterol poblacional es igual o distinto de 220 mg/100 ml.

Los pasos para la realización del contraste de hipótesis serían los siguientes:

1. **Definición de las hipótesis del contraste: Hipótesis Nula e hipótesis alternativa**

El parámetro sobre el que se pretende realizar un contraste de hipótesis es, en este caso, una media poblacional. Se trata de decidir si puede considerarse que la media poblacional es 220 o, por el contrario, la media poblacional es distinta de 220. Estas dos posibles decisiones se expresan mediante las siguientes hipótesis:

H_0: $\mu = 220$ $\qquad\qquad$ H_1: $\mu \neq 220$

Donde H_0 se denomina «hipótesis nula» y H_1 «hipótesis alternativa». La dinámica del contraste establece que la hipótesis nula se mantendrá a no ser que los datos muestren una fuerte evidencia en contra de la misma, en cuyo caso, se optará por la hipótesis alternativa.

2. **Definición de una medida de discrepancia o estadístico de contraste entre lo que se afirma en la hipótesis nula y la información que proporcionan los datos de la muestra observada**

En el caso del contraste de una media, cuando se desconoce la desviación típica poblacional, el estadístico de contraste utilizado es:

$EC = \frac{\ddot{X} - \mu}{s/\sqrt{n}}$, donde μ es el valor de la media que se especifica en la hipótesis nula, $\ddot{X}$ es la media muestral, S el estimador puntual de la desviación típica poblacional y n el tamaño de la muestra. En este caso se tendrá:

$$EC = \frac{\ddot{X} - \mu}{s/\sqrt{n}} = \frac{235 - 220}{28/\sqrt{46}} = 3{,}63$$

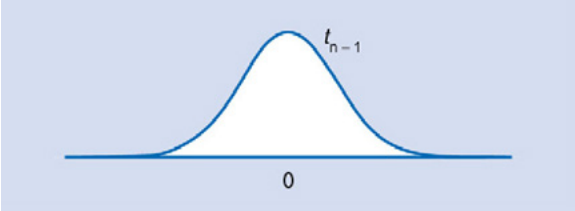
Puede observarse que, dado que la desviación típica es siempre una cantidad positiva, la medida de discrepancia o estadístico de contraste (EC) tomará el valor 0 únicamente cuando la media observada en la muestra coincida exactamente con la que se propone en la hipótesis nula. Además, cuanto

mayor sea la discrepancia entre la media observada y la media a la que se refiere la hipótesis nula, mayor será el numerador (en términos absolutos) y, por tanto, el valor del estadístico de contraste. De este modo puede concluirse que valores del estadístico de contraste próximos a 0 favorecerían a la hipótesis nula, mientras que valores muy distantes de 0 indicarían la existencia de evidencia en contra de dicha hipótesis. Será necesario decidir cuándo la discrepancia es lo suficientemente grande como para rechazar la hipótesis nula. En este sentido será muy útil la consideración del siguiente paso.

3. **Conocer la distribución de probabilidad asociada a la medida de discrepancia o estadístico de contraste**

El conocimiento de la distribución de probabilidad que gobierna el comportamiento del estadístico de contraste será vital para el desarrollo final del contraste de hipótesis. En el ejemplo se cuenta con 46 datos. Si la variable nivel de colesterol sigue un modelo de distribución normal o por aplicación del teorema central del límite, puede establecerse que la distribución de probabilidad asociada al estadístico de contraste es, una t de Student con n − 1 grados de libertad, tal y como se refleja en la figura 14. En este caso se trataría de una distribución t de Student con n − 1 = 46 − 1 = 45 grados de libertad. La distribución t 45 es simétrica y está centrada en 0. Si la hipótesis nula fuera cierta, el estadístico de contraste debería tomar valores en la zona central de la distribución, siendo muy improbable observar valores en cualquiera de sus dos extremos.

Figura 14. Distribución muestral del estadístico de contraste



4. **Establecimiento del nivel de significación del contraste**

Para decidir exactamente qué valores del estadístico de contraste tendrían una probabilidad de observarse prácticamente despreciable si la hipótesis nula fuera cierta, se establece el denominado «nivel ***de*** significación» del contraste o nivel de significación ***a*** priori α. Habitualmente se considera el valor α = 0,05, aunque dependiendo del caso puede utilizarse el nivel 0,01 o 0,001. En este ejemplo se decide utilizar un nivel a = 0,05.

5. **Construcción de la regla de decisión**

Si el nivel de significación elegido es α = 0,05, deberían despreciarse valores del estadístico de contraste con una probabilidad inferior a 0,05. Los valores menos probables del estadístico de contraste si la hipótesis nula fuera cierta se concentran en ambos extremos de la distribución, por tanto, se define como región crítica de contraste o región de rechazo de la hipótesis nula a la región $-\infty, t0{,}025[\cup]t0{,}975 + \infty$, que se muestra sombreada en la figura 15.

De forma complementaria, se define como ***región de aceptación*** de la hipótesis nula a la región comprendida entre $t_{0,025}$ y $t_{0,975}$: $[t_{0,025}, t_{0,975}]$.

Los valores $t_{0,025}$ y $t_{0,975}$ sobre una distribución t_{45} que determinan la región de aceptación y de rechazo de la hipótesis nula en este caso son:

$t_{0,025} = -2{,}014$ y $t_{0,975} = 2{,}014$. La regla de decisión quedará, por tanto:

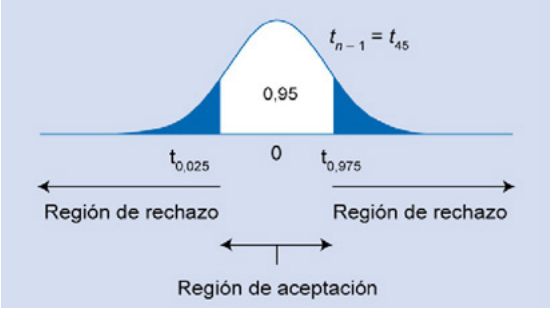
- Si EC > 2,014 o EC < − 2,014 → Se rechaza la hipótesis nula H_0.
- Si −2,014 ≤ EC ≤ 2,014 → Se acepta la hipótesis nula H_0.

6. **Aplicación de la regla de decisión**

Por último, será necesario aplicar la regla de decisión del contraste de hipótesis. Para ello habrá que establecer la región en la que se sitúa el estadístico de contraste calculado con anterioridad en el paso 2.

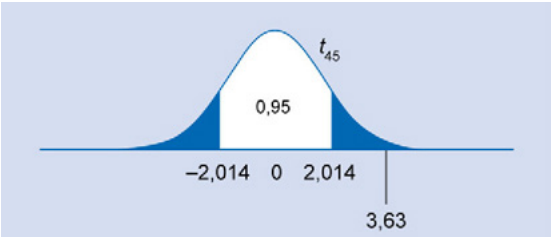
Como puede apreciarse en la figura 16, el valor del estadístico de contraste EC = 3,63 se encuentra situado en la región de rechazo de la hipótesis nula (3,63 > 2,014) y la regla de decisión conducirá a rechazarla, considerándose que existe evidencia suficiente que indica que el promedio de nivel de colesterol es significativamente distinto de 220 mg/100 ml.

Figura 15. Región critica de contraste y de aceptación de H_0



Fuente: Moncho-Vasallo (11)

Figura 16. Situación del estadístico de contraste



Fuente: Moncho-Vasallo (11)

Errores en un contraste de hipótesis

Dado que en un contraste de hipótesis son dos los posibles resultados (rechazar la hipótesis nula o no rechazarla), también serán dos los posi-

bles errores que puedan cometerse, consecuencia de cada una de las dos decisiones posibles. En la tabla 16 se refleja esta situación.

Tabla 16. Tipos de error en un contraste de hipótesis

Decisión del contraste	**Realidad**	
	H_0 cierta	H_1 cierta (H_0 falsa)
Se acepta H0	Ausencia de error	Error tipo II
Se rechaza H0	Error tipo I	Ausencia de error

Fuente: Moncho-Vasallo (11)

El error de tipo I se define como el error que se comete al rechazar la hipótesis nula cuando esta es cierta. El contraste conducirá al rechazo de la hipótesis nula únicamente cuando el estadístico de contraste se sitúe en la región crítica y esto solo puede ocurrir con una probabilidad α.

$$\alpha = P(\text{Rechazar } H_0 \mid H_0 \text{ cierta})$$

Luego, al establecer un nivel de significación para el contraste se está controlando la probabilidad de cometer un error de tipo I. Por otra parte, el error de tipo II es el que se comete cuando no se rechaza la hipótesis nula a pesar de que es falsa y suele denominarse β.

$$\beta = P(\text{No rechazar } H_0 \mid H_0 \text{ falsa})$$

Hipótesis nula e hipótesis alternativa

Los contrastes de hipótesis se basan en la definición de dos hipótesis enfrentadas, la hipótesis nula H_0 y la hipótesis alternativa H_1. La hipótesis nula es la hipótesis que se pretende contrastar y será mantenida a menos que los datos observados en la muestra indiquen una fuerte evidencia de que no es cierta. Si en el procedimiento de realización del contraste de hipótesis únicamente se ha controlado el error de tipo I a través del establecimiento del nivel de significación a priori a y se desconoce la

probabilidad de cometer un error de tipo II, la aceptación de la hipótesis nula no implica evidencia de su certeza sino, simplemente, que no se ha encontrado evidencia de que no lo sea. Por otra parte, si el contraste de hipótesis se decide por la hipótesis alternativa y, por tanto, rechaza la hipótesis nula, será porque la evidencia en contra de dicha hipótesis es manifiesta. Por este motivo, algunos autores prefieren afirmar que una hipótesis nula, contrastada en dichas condiciones, nunca puede ser aceptada, sino simplemente rechazada o no rechazada.

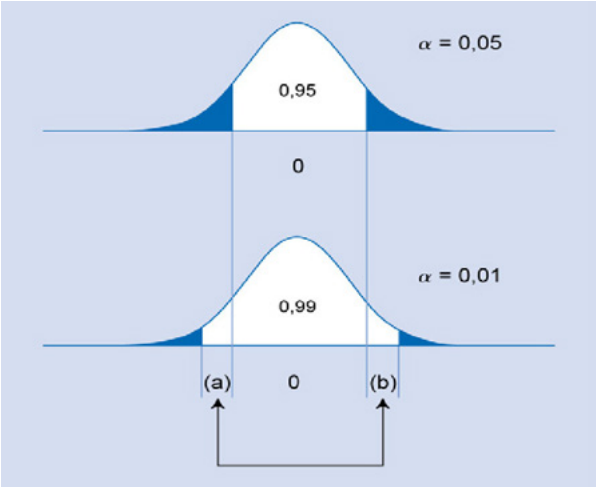
Debe tenerse en cuenta que puede determinarse el tamaño muestral necesario para contrastar una hipótesis controlando simultáneamente el error de tipo I y de tipo II, de forma que las decisiones por la hipótesis nula tengan asociado un error tan pequeño como se requiera.

Contraste y nivel de significación

La posibilidad de rechazar una hipótesis nula depende en gran parte de la magnitud de la región crítica de contraste. El tamaño de esta región crítica está determinado por el valor del nivel de significación del contraste a. Por tanto, el resultado del contraste de hipótesis será dependiente del nivel de significación elegido, pudiéndose dar el caso, en que se rechace la hipótesis nula trabajando con un nivel de significación, por ejemplo, $\alpha = 0{,}05$ y no rechazarse al nivel $\alpha = 0{,}01$.

En la figura 17 puede observarse que para todos los valores de la región (a) y (b) se rechazaría la hipótesis nula al nivel $\alpha = 0{,}05$, pero no al nivel $\alpha = 0{,}01$. Por otra parte, la aplicación de la regla de decisión del contraste tan solo permite concluir si se consigue o no se consigue rechazar la hipótesis nula pero no informa sobre la magnitud de la evidencia en contra de dicha hipótesis. Ambos problemas pueden solucionarse definiendo el que se denomina nivel crítico p, p-valor o nivel de significación a posteriori.

Figura 17. Diferencia en la región crítica de contraste según el nivel de significación.



Fuente: Moncho-Vasallo (11)

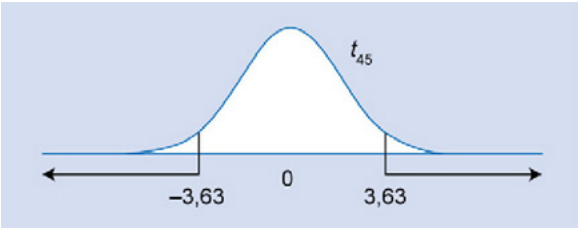
Nivel de significación a posteriori o p-valor

El *p-valor* se define como la probabilidad de observar, bajo la suposición de que la hipótesis nula es cierta, un valor del estadístico de contraste o medida de discrepancia igual o más extremo que el observado en la muestra. Por tanto, el valor de *p* no se fija a priori, sino que es calculado a partir de los datos de la muestra *(a* posteriori*).* Valores muy pequeños de *p* estarían indicando que el estadístico de contraste se encuentra situado en cualquiera de los dos extremos de la distribución y la evidencia en contra de la hipótesis nula sería patente. Además, cuanto más pequeño sea el valor de *p* mayor será la evidencia en contra de la hipótesis nula.

En el ejemplo 1 (estudio sobre el nivel de colesterol de un grupo de 46 pacientes seleccionados al azar) el valor del estadístico de contraste era EC = 3,63.

El valor de *p* será la probabilidad de observar un valor del estadístico o medida de discrepancia igual o más extremo al observado, tal y como puede observarse en la figura 18.

Figura 18. Región que determina el valor de p



Fuente: Moncho-Vasallo (11)

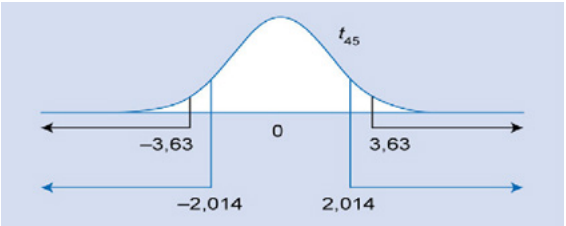
Dado que la región crítica de contraste tiene dos colas (se rechaza tanto con valores en el extremo derecho de la distribución como en el extremo izquierdo), el valor de la *p* del contraste debe calcularse teniendo en cuenta las dos regiones. Así, el valor de *p* viene determinado por:

$p = P(EC - 3,63) + P(EC \leq -3,63) = 0,00036 + 0,00036 = 0,00072$

Un valor de *p* = 0,00072 indica que el EC se ha situado en el extremo de la distribución. Recuérdese que cuanto más pequeño sea este valor más evidencia en contra de la hipótesis nula. ¿Qué hubiera ocurrido si se hubiera trabajado al nivel de significación $\alpha = 0,05$?

Como puede observarse, en la figura 19 al nivel de significación $\alpha = 0,05$ se rechazaría la hipótesis nula, ya que el EC se situaría dentro de la región de rechazo.

Figura 19. Valor p del contraste y nivel de significación α



Fuente: Moncho-Vasallo (11)

En general puede establecerse que, para un nivel de significación del contraste cualquiera a, se verificará:

- Si $p < \alpha \rightarrow$ Se rechaza H_0 al nivel α.
- Si $p \geq \alpha \rightarrow$ Se acepta H_0 al nivel α.

Obsérvese que en el ejemplo se rechazaría también la hipótesis nula trabajando al nivel de significación $\alpha = 0{,}01$ e incluso al nivel $\alpha = 0{,}001$, ya que el valor de la p del contraste es inferior en todos los casos.

Contrastes bilaterales y unilaterales

Los contrastes de hipótesis pueden ser ***unilaterales*** o ***de una cola*** y ***bilaterales*** o ***de dos colas*** dependiendo de la forma en que se planteen las hipótesis.

Así, podría considerarse que cuando el interés se centra en contrastar la hipótesis de que un determinado parámetro de la población tome exactamente un valor dado frente a la hipótesis de que el parámetro tome un valor distinto al propuesto, el contraste será bilateral o de dos colas.

$H_{0:}\ \theta = \theta_0$

$-H_{1:}\ \theta \neq \theta_0$

Obsérvese que la región de rechazo definida por un contraste de este tipo quedaría situada a ambos extremos de la distribución, puesto que debería rechazarse la hipótesis nula tanto cuando $\theta > \theta_0$ como cuando $\theta < \theta_0$. Por otra parte, si la hipótesis se plantea de forma que se atiende únicamente al hecho de que ese mismo parámetro de la población tome un valor superior (análogamente inferior) a un valor dado, el contraste será unilateral o de una cola.

$H_{0:} \theta \leq \theta_0 \quad H_{0:} \theta \geq \theta_0$

$H_{0:} \theta > \theta_0 \quad H_{0:} \theta < \theta_0$

En este caso la región de rechazo definida por el contraste se situaría bien a la derecha de la distribución (se rechaza H_0 cuando $\theta > \theta_0$), o bien a la izquierda de la distribución (se rechaza H0 cuando $\theta < \theta_0$).

Potencia de un contraste

La potencia de un contraste de hipótesis se define como la probabilidad de rechazar H0 cuando es falsa o, equivalentemente, la probabilidad de decidirse por la hipótesis alternativa cuando esta es cierta.

$$\text{Potencia} = P\ (\text{rechazar } H_0 \mid H_0 \text{ falsa})$$

En algunos estudios, las características de la muestra seleccionada pueden impedir la detección de evidencia significativa en contra de la hipótesis nula que se plantea, aunque esta sea falsa. En este sentido, la potencia del contraste podría interpretarse como la probabilidad de encontrar en el estudio evidencia significativa en contra de la hipótesis nula, en el caso de que efectivamente la hipótesis nula fuera falsa.

Cuando la hipótesis alternativa es una hipótesis simple (especifica un único valor para el parámetro) el valor de la potencia es único. Sin embargo, cuando la hipótesis alternativa es compuesta existirá un valor de potencia

asociado a cada una de las posibilidades. Esto sugiere la definición de la que se denomina *función de potencia* de un contraste de la forma:

$$\text{Pot}(\theta) = P(\text{Rechazar } H_0 \mid \theta)$$

Donde θ representa todas las posibilidades del parámetro sobre el que se pretende contrastar alguna hipótesis.

Cuando sustituimos u por el valor que se especifica en la hipótesis nula u0, la función de potencia toma el valor a.

$$\text{Pot}(\theta_0) = P(\text{Rechazar } H_0 \mid \theta_0) = \alpha = P(\text{Error tipo I})$$

Por otra parte, cuando θ toma cualquier otro valor, la función de potencia quedará:

$$\text{Pot}(\theta_0) = P(\text{Rechazar } H_0 \mid \theta_0) = 1 - P(\text{Aceptar } H_0 \mid \theta) = 1 - \beta(\theta) = 1\, P(\text{Error tipo II})$$

Lo deseable es que tanto el error de tipo I como el error de tipo II sean lo más bajos posible, siendo habitual trabajar al nivel de significación α = 0,05 y potencia no superior a 0,2.

Contraste de hipótesis sobre una proporción

El ejemplo 2 planteaba que de los 250 individuos a los que se les administró un determinado tipo de tratamiento, 180 respondieron de forma positiva. ¿Puede considerarse que la proporción de éxito del tratamiento es del 0,8 o del 80%? Para responder a esta cuestión será necesario realizar un contraste de hipótesis sobre la proporción poblacional de respuesta positiva al tratamiento. Si se siguen los pasos para la realización del contraste, establecidos con anterioridad se tendrá que:

H_0: p = 0, 8 0

H_0: p ≠ 0, 8

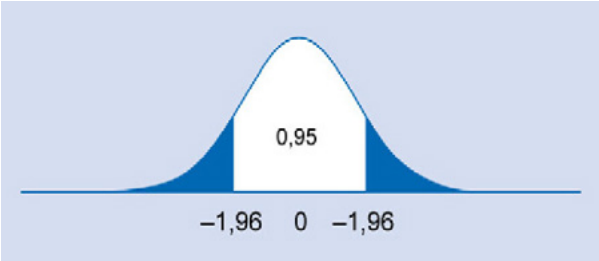
La medida de discrepancia o estadístico de contraste en el caso de una proporción poblacional será:

$$EC = \frac{\hat{P} - P_0}{\sqrt{\frac{P_0 - (1 - P_0)}{n}}} = \frac{0{,}72 - 0{,}8}{\sqrt{\frac{0{,}72 - (1 - 0{,}72)}{250}}} = 2{,}81$$

donde p_0 es el valor de la proporción poblacional que se especifica en la hipótesis nula, $\hat{p}$ la proporción calculada a partir de los datos de la muestra y n el tamaño de la muestra.

Debe observarse que en este caso se verifican las condiciones para la aproximación normal. En caso contrario sería necesario utilizar la distribución binomial exacta. Si se trabaja al nivel habitual a = 0,05, como el valor del estadístico de contraste (EC = 2,81) es mayor que 1,96, se situará en la región crítica de contraste (**véase la figura** 20) y, por tanto, se procederá a rechazar la hipótesis nula. Podrá concluirse, por tanto, que la proporción poblacional de individuos con respuesta positiva al tratamiento es significativamente distinta de 0,8 (o del 80%). Dado que el rechazo se ha producido a la derecha de la distribución, puede concluirse que la proporción poblacional es significativamente superior a 0,8.

Figura 20. Región crítica de contraste. Contraste de una proporción.



Fuente: Moncho-Vasallo (11)

Comparación de dos medias poblacionales

Para la realización de un contraste sobre la diferencia de dos medias poblacionales es necesario tener en cuenta:

- Si se trata de muestras independientes o apareadas (relacionadas).
- Si las varianzas son iguales o distintas.

Adicionalmente, será necesario contar con un número mínimo de datos para utilizar las pruebas paramétricas correspondientes. En otro caso, deberá recurrirse a pruebas no paramétricas, cuya discusión desborda el alcance del presente libro.

Prueba t de comparación de medias para muestras independientes y varianzas iguales

Para la utilización de la prueba ***t*** de comparación de medias se requiere que las muestras sean independientes y que la variable siga una distribución normal. En la práctica suele utilizarse cuando el tamaño de cada una de las muestras es superior o igual a 30. Considérese el siguiente ejemplo:

► Ejemplo 3:

Supóngase que se cuenta con información sobre la edad en dos grupos de pacientes seleccionados al azar. El promedio de edad en el primer grupo formado por 40 individuos fue de 56 años, con una desviación típica de S_1 = 12 años, mientras que en el segundo de 35 individuos el promedio de edad fue de 62 años y una desviación típica de S_2 = 14 años. Se desea saber si las medias de edad son significativamente distintas entre los dos grupos de pacientes.

Se dispone de la siguiente información:

$$\ddot{X}_{1} = 56 \;\; ; \;\; \ddot{X}\,2 = 62$$

$$S_1 = 12 \;\; ; S_2 = 14$$

$$N = 40 \;\; ; \;\; n = 35$$

Puede observarse que el tamaño de las muestras es superior a 30 en los dos casos. Por otra parte, suponemos que las varianzas son iguales, es decir, se cumple la condición de la igualdad de varianzas. Por tanto, estará indicada la utilización de la prueba ***t*** para muestras independientes y varianzas poblacionales iguales.

El estadístico de contraste será:

$$EC = \frac{\ddot{X}1 - \ddot{X}2}{\sqrt{S_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}, \text{ donde:}$$

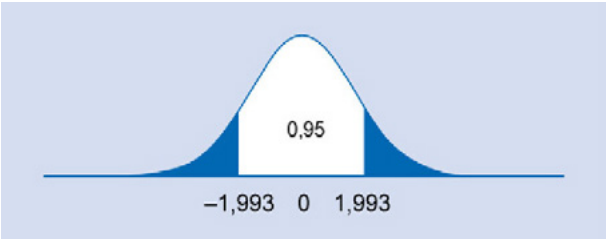
$$S_p^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2} = \frac{(40-1)12^2+(35-1)14^2}{40+35-2} = 168{,}22$$

Téngase en cuenta que, dado que se considera que las varianzas poblacionales son iguales y se cuenta con dos varianzas (una para cada muestra), se construye una varianza común ponderándolas adecuadamente y obteniendo la cantidad S^2_p El valor del estadístico quedará:

$$EC = \frac{\ddot{X}1 - \ddot{X}2}{\sqrt{S_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{56-62}{\sqrt{168{,}22\left(\frac{1}{40}+\frac{1}{35}\right)}} = 1{,}999$$

Que se distribuirá según una ***t*** de Student con n1 + n2 − 2 = 40 + 35 − 2 = 73 grados de libertad. Si se trabaja al nivel de significación a = 0,05 se tendrá que él percentil 0,975 será 1,993. Así, la región crítica de contraste quedará definida como se describe en la figura 21.

Figura 21. Región crítica de contraste. Comparación de medias. Varianzas iguales



Fuente: Moncho-Vasallo (11)

Como el estadístico de contraste EC = −1,999 es menor que −1,993 se situará en la región crítica de contraste (cola de la izquierda), pudiéndose concluir que existe evidencia de que las medias de edad son significativamente distintas entre los dos grupos de pacientes.

Programas de cómputo y análisis estadístico

Es importante dejar bien claro que la computadora es una herramienta muy útil en el análisis estadístico de los datos, pero que no piensa ni sustituye el buen juicio del investigador. La computadora sigue instrucciones en lenguaje binario, y las cumple a una velocidad que no deja de ser impresionante. Si se le alimenta correctamente y las instrucciones que se le dan son las adecuadas, los resultados serán sorprendentes. Pero si esto se hace incorrectamente, los resultados también serán sorprendentes por la magnitud de los errores. Vale la pena insistir: la computadora es una herramienta y en ningún momento sustituye la labor del investigador en el análisis estadístico (7).

Con la ayuda de la computadora, se agiliza la tabulación y también las operaciones, pero es el investigador quien tiene que decidir qué análisis es el más adecuado para sus datos, y es él quien tiene que interpretar los

resultados. Si lo anterior ha quedado claro, entonces se puede avanzar con el tema de los programas de cómputo que pueden ser de gran ayuda en el análisis estadístico.

Durante el análisis estadístico, el investigador tiene que realizar varias tareas:

- Presentación de la propuesta o proyecto.
- Búsqueda y registro de datos.
- Captura y transformación de datos.
- Revisión de la captura.
- Tabulación de datos.
- Cálculo de estadísticos.
- Interpretación de resultados.
- Reporte final del trabajo.

Para realizar estas tareas, el investigador usa varios programas de cómputo. Algunos son exclusivos del trabajo estadístico, otros son menos específicos. En términos generales, entre los programas de cómputo que se utilizan se incluyen los procesadores de texto, los administradores de bases de datos, las hojas de cálculo, los programas para presentaciones y los programas estadísticos. A esta lista también podríamos agregar algunas utilerías y páginas de la web que funcionan como calculadoras estadísticas/epidemiológicas. El uso que se hace de ellos difiere de muchas maneras, pero con frecuencia realizan tareas comunes.

A continuación, sin ánimo de ser exhaustivos, se harán comentarios sobre algunos programas de cómputo:

Procesadores de texto

Estos programas son los más genéricos de todos los que se utilizan. Básicamente sirven para escribir las propuestas iniciales, reportes finales del trabajo y formas para captar datos (cuestionarios, cédulas de captura).

Algunos incluyen utilerías capaces de elaborar cuadros y gráficos. Otros incluso permiten escribir fórmulas como las que se presentan en este libro. Entre estos programas destaca Word para Windows.

Administradores de bases de datos

Cumplen una función muy importante durante el trabajo estadístico: ayudan a capturar datos en los archivos de cómputo en los que se almacenan. Además de generar la estructura de la base y de permitir capturar los datos, estos programas también permiten editar y transformar datos, así como generar nuevas variables y asignarles valores a partir de los existentes.

Hojas de cálculo

Su estructura de celdas, construidas a partir de columnas y renglones, además de las funciones que incluyen, las hace muy útiles para el trabajo estadístico. Entre estos programas, los más conocidos son Excel y Lotus. Su primera aplicación suele ser la captura de datos; es mucho más sencilla que la que se puede hacer en los administradores de bases de datos, pero menos versátil. Para capturar datos en una hoja de cálculo, se procede a identificar las columnas con las variables y los renglones con los registros. En las celdas del primer renglón se anotan los nombres de las variables y, a partir del segundo renglón, se capturan los datos que corresponden a cada elemento del grupo. Por brevedad, en la captura suelen utilizarse códigos y no etiquetas. Por ejemplo, en vez de escribir "masculino" en la columna de sexo se puede anotar "1" y en lugar de "femenino", "2". De esta manera se ahorra mucho tiempo y se reduce el número de errores.

Las hojas de cálculo incluyen funciones que, a partir de un bloque de datos, permiten realizar varias operaciones, entre las que se encuentran los cálculos de la media, mediana, moda, varianza, desviación estándar,

coeficiente de correlación, intercepción y pendiente de la regresión lineal. También pueden mostrar los valores de varias distribuciones de probabilidad, como la normal, binomial, Poisson, ***t***, ***F***, _2, así como realizar las pruebas de chi-cuadrada, ***t*** de Student, ***F*** y ***z***.

Las hojas de cálculo también son de gran ayuda para realizar los gráficos necesarios en el análisis estadístico. La facilidad con la cual se puede realizar un gráfico en estos programas permite ensayar con varias formas diferentes hasta que se encuentre la que mejor presenta los resultados.

Además de todas las facilidades que brindan las hojas de cálculo, también se debe mencionar la gran capacidad que tienen para importar y exportar archivos generados en otros formatos, lo cual los hace muy útiles cuando se trabaja en varias plataformas de datos.

Programas para presentaciones

Estos programas ayudan a elaborar la presentación de los datos, principalmente mediante proyecciones o carteles. Para hacerlo, generalmente se le concede preferencia al uso de gráficos o cuadros. Entre estos programas se encuentran Power Point y Harvard Graphics.

Programas estadísticos

De todos los programas que se comentan en este apartado, éstos son los que más han revolucionado el trabajo en la estadística.

Entre ellos existe una gran variedad de funciones y costos. Los hay genéricos o especializados y también gratuitos o muy caros. Es difícil decir cuál es el mejor, pero no cabe duda de que se debe tener el que mejor se conozca y que permita realizar los análisis estadísticos que se requieren para el trabajo. La referencia que aquí se hace se concentra en los programas: Epi Info, OpenEpi, SPSS y R. Pero existen otros, igualmente versátiles.

Epi Info

Este programa ha sido desarrollado y distribuido por el CDC de Atlanta. En un principio se concibió como una herramienta auxiliar de los epidemiólogos de campo para usarse en equipos portátiles de cómputo, pero con el tiempo ha encontrado su lugar en la mayoría de los grupos de trabajo que laboran en el campo de la salud, entre los que sin duda es uno de los programas más populares.

Varias de sus características han contribuido a brindarle ese lugar privilegiado, entre las que destaca su facilidad de manejo; pero ninguna le ha dado tanto impulso como el hecho de que el programa se distribuye libremente y sin costo a través de Internet desde el CDC de Atlanta (www.cdc.gov) y otros sitios de la red. Esta gran difusión ha facilitado la traducción del programa y sus manuales al español, entre otros idiomas. Epi Info se desarrolló para ejecutarse en dos plataformas diferentes: MSDOS y Windows. Su versión en español se puede descargar en:

https://www.cdc.gov/epiinfo/esp/es_index.html

La versión Epi Info 7+, corre en ambiente Windows, y es una gran ventaja, porque se maneja de manera semejante a otros programas de este ambiente gráfico; así, si ya se conoce uno, todos los demás resultan familiares y más fáciles de aprender. De manera general, la interfaz del Menú nos muestra las utilerías que ofrece el programa: crear formas de captura (Create Forms), capturar datos (Enter Data), analizar datos (Analyze Data) y crear mapas (Create Maps). A estas mismas opciones se puede llegar a través de “Tools” en el menú que se encuentra en el borde superior de la ventana. En el mismo menú superior se encuentra “StatCalc”; esta opción incluye una serie de calculadoras epidemiológicas. El análisis de datos (en Analyze Data) se puede realizar tanto en archivos propios de Epi Info 7, como archivos con formatos de Access, Excel, SQL y ASCII.

SPSS (Statistical Package for Social Science)

Este programa tiene una larga historia en el análisis estadístico. Las primeras versiones se hicieron para correr en equipos muy grandes, y fue uno de los primeros programas de estadística disponibles en las computadoras personales. En la actualidad, el programa corre en varias plataformas, entre las que se encuentra Windows.

En las primeras versiones de SPSS para PC, el usuario tenía que saber mucho de programación, porque cada comando se escribía en una pantalla negra en la que no se veían ayudas. Ahora el ambiente gráfico facilita el manejo a tal punto que el usuario puede aprender a utilizarlo en horas (o en minutos si se tiene alguna experiencia en otros programas de cómputo). En la Web, se pueden descargar diversos manuales y guías de usuarios que facilitan su uso (30)

Al entrar al programa se ve una pantalla cuadriculada muy semejante a una hoja de cálculo. En esta pantalla se puede empezar a capturar datos de la misma manera como se señaló para las hojas de cálculo, con las mismas dificultades, pero sin la facilidad de poder realizar operaciones en las celdas. Los datos capturados de esta forma pueden guardarse en un archivo de SPSS y después pueden llamarse para continuar la captura o iniciar el análisis.

El programa también permite leer bases de datos generadas por otros programas, como Excel o Fox, por ejemplo.

Programa R

El programa R es un ambiente de programación para realizar gráficos y cálculos estadísticos. Su gran ventaja es que es un programa de acceso abierto y gratuito en constante actualización, el cual puede descargarse (31) y encontrar temas de ayuda en https://www.r-project.org/. Es un

proyecto de colaboración en el cual los colaboradores donan códigos de acceso libre, actualizan el programa y sus paquetes, corrigen errores de programación y documentan las distintas funciones de R. Entre la variedad de cálculos estadísticos que pueden realizarse en R, se encuentran los análisis estadísticos clásicos, modelación lineal y no lineal, análisis de series de tiempo, análisis de clasificación y estadística multivariada, por mencionar algunos. Los usuarios o personas que conocen el lenguaje de programación pueden generar sus propios códigos para realizar cálculos específicos (p. ej., simulaciones Monte Carlo). Otra ventaja es la calidad del diseño de gráficos para su publicación. En R pueden importarse bases de datos guardadas en formato de texto (*.txt) o formato CVS (*.csv) con columnas delimitadas por comas o tabulaciones; también es posible importar y utilizar los archivos creados en Excel (*.xls). La gran desventaja de este programa radica en que la consola de R trabaja con un lenguaje de programación y no cuenta con los menús de selección a los que estamos habituados los usuarios de la plataforma de Windows; es necesario conocer su lenguaje, basado en códigos de programación específicos, para poder comenzar a utilizar la consola blanca. Otro de los inconvenientes es que los códigos son extremadamente sensibles a los errores tipográficos, por lo que una simple coma (,) fuera de lugar interrumpe el proceso del análisis generando mensajes de error.

Actualmente existen programas gratuitos que funcionan como editores de código R y que trabajan bajo la plataforma de Windows. Estos programas facilitan la escritura de los comandos en un documento llamado “script”, el cual puede ser archivado como texto (*.txt) o código R (*.r) para un subsecuente uso y edición. Estos programas se vinculan con la consola de R, con lo cual al tiempo que se escribe el código se pueden enviar las instrucciones a R y observar los resultados.

La consola de R puede requerir paquetes de comandos para realizar análisis específicos, los cuales deben ser instalados en R antes de usarse. El paquete necesario para realizar análisis estadísticos clásicos se instala automáticamente al instalar el programa R ("stats"). El paquete ODBC Database Access ("RODBC") debe ser instalado por el usuario si desea importar a R las bases de datos creadas en formato Excel (*.xls).

R Commander ("Rcmdr") es un paquete que funciona como un programa de análisis estadístico dentro del programa R. La ventaja de trabajar con R Commander radica en que éste aporta todas las ventajas de R (p. ej., gráficos) por medio de una interfaz mucho más amigable para el usuario, basada en menús y ventanas de selección bajo el perfil de Windows.

Es importante señalar que la aplicabilidad de R Commander no se limita a los modelos generalizados. Los lectores pueden iniciarse en el lenguaje de R utilizando dicho paquete como otra herramienta estadística (p. ej., análisis de varianza, análisis de regresión), entre otros.

Open Epi

OpenEpi (32) es una página en la web que se puede consultar a través de un browser (en http://www.openepi.com/Menu/OE_Menu.htm), pero que también se puede utilizar a nivel local si previamente se descargó el programa en la computadora.

Está conformado por una serie de calculadoras epidemiológicas y con enlaces a muchas páginas especializadas en análisis estadístico y epidemiológico.

REFERENCIAS

1. Macchi R. Introducción a la Estadística en Ciencias de la Salud Buenos Aires: Editorial Médica Panamericana; 2019.
2. Montanero- Fernández J, Minuesa- Abril C. Estadística básica para Ciencias de la Salud: Cáceres; 2018.
3. Pérez-Tejada H. Estadística para las ciencias sociales, del comportamiento y de la salud. 3a. edición México, D.F: Cengage Learning Editores, S.A; 2008.
4. González- Delgado M. Bioestadística y Vigilancia Epidemiológica Bogotá D.C: Fundación Universitaria del Área Andina; 2017.
5. Deanza – College B, Deanza –College S. Introducción a la estadística Houston, Texas: Rice University; 2022.
6. Clifford- Blair R, Taylor R. Bioestadística México: Pearson Educación; 2008.
7. Celis de la Rosa A, Labrada- Martagón V. Bioestadística: Editorial El Manual Moderno, S.A. de C.V; 2014.
8. Hernández- Hidalgo C, Terrés - Sandoval A, Valdez -Monroy. Estadística y Probabilidad I. Cuaderno de Trabajo; 2019.
9. Posada -Hernández G. Elementos básicos de estadística descriptiva para el análisis de datos Medellín: Funlam; 2016.
10. Pastor-Barriuso R. Bioestadística: Centro Nacional de Epidemiología; 2012.
11. Moncho- Vasallo J. Estadística aplicada a las Ciencias de la Salud: Gea Consultoría Editorial, s. l; 2015.

12. Gorgas -García J, Cardiel- López N, Zamorano -Calvo J. Estadística básica para estudiantes de ciencias: Universidad Complutense de Madrid; 2009.

13. Devore J. Fundamentos de probabilidad y estadística. Novena Edición ed. México: Cengage Learning; 2016.

14. Carot Sánchez T. Introducción a la estadística y a las probabilidades. [Online].; 2014. Acceso 18 de noviembre de 2023. Disponible en: https://www.etsii.upv.es/conbuenpie/documentos/11398-Estadistica_Apuntes_Previos.pdf.

15. Nolasco-Bonmatí A, Moncho-Vasallo J. Estadística básica en Ciencias de la Salud Alicante, España: Universidad de Alicante; 2016.

16. Bacchini R, Vázquez V, Bianco M, García J. Introducción a la probabilidad y la estadística. [Online].; 2018. Acceso 21 de noviembre de 2023. Disponible en: http://bibliotecadigital.econ.uba.ar/download/libros/Bacchini_Introduccion-a-la-probabilidad-y-a-la-estadistica-2018.pdf.

17. Rodríguez Benot A, Crespo Montero R. Introducción a la estadística básica para enfermería nefrológica. Seden. 1998; 99(7): p. 20-34. Disponible en: https://revistaseden.org/files/art319_1.pdf.

18. Austin E, Cobb F, Coleman R, Jones R. Prospective evaluation of radionuclide angiocardiography for the diagnosis of coronary artery disease. Am J Cardiol. 1982; 50(6): p. 1212-6. doi: 10.1016/0002-9149(82)90451-9. PMID: 7148693.

19. Martín-Olmedo P, Carroquino-Saltó M, Ordóñez.Iriarte J, Moya J. La Evaluación de Riesgos en Salud. Guía metodológica. (SESA) SEdSA, editor. Madrd, España: Serie “De aeribus, aquis et locis” nº3; 2016.

20. Bioestadística:Tablas de distribucion de probabilidades. Distribución Binomial. [Online].; 2023. Acceso 26 de noviembre de 2023. Disponible en: https://www.ugr.es/~bioestad/_private/Tablas.pdf.

21. Cátedra Probabilidad y Estadistica. Tabla estadistica: Distribución de Poisson, F(x). [Online].; 2023. Acceso 26 de noviembre de 2023. Disponible en: https://www.uv.es/montes/NHD/PoissonAcumulada.pdf.

22. DeGroot M. Probabilidad y Estadística. 2nd ed. Madrid. España: Addison-Wesley Iberoamericana.; 1988.

23. Tadhkira. Teoría de la campana de gauss. [Online]; 2007. Acceso 26 de noviembrede 2023. Disponible en: https://farid.austrinus.com/2007/12/19/teoria-de-la-campana-de-gauss-parte-i/.

24. Universidad Autónoma del estado de México. Distribución normal estadistica. [Online].; 2019. Acceso 25 de noviembre de 2023. Disponible en: http://ri.uaemex.mx/bitstream/handle/20.500.11799/106113/Distribuci%C3%B3n%20Normal.pdf?sequence=1&isAllowed=y.

25. Sitio Web Tuveras.com. Tabla de distribución normal N (0,1). [Online].; 2023. Acceso 25 de noviembre de 2023. Disponible en: http://www.tuveras.com/estadistica/normal/tabla.htm.

26. Development Core Team. Tabla de Número Aleatorios. [Online].; 2008. Acceso 27 de noviembre de 2023. Disponible en: file:///C:/Users/JAG/Downloads/Tabla_de_Numeros_Aleatorios.pdf.

27. Álvarez-Cáceres R. Estadística aplicada a las ciencias de la salud Madrid España: Ediciones Díaz de Santos; 2007.

28. Tabla t-Student. Tabla de la distribución t de Student. [Online].; 2023. Acceso 3 de diciembre de 2023. Disponible en: https://cms.

dm.uba.ar/academico/materias/verano2022/probabilidades_y_estadistica_C/tablas/tabla_tstudent.pdf.

29. González F, Escoto M, Chávez J. Estadística aplicada en Psicología y Ciencias de la salud Ciudad de México, México: Editorial El Manual Moderno, S.A. de C.V.; 2017.

30. International Business Machines Corporation (IBM). IBM SPSS Statistics 29 Guía breve. [Online].; 2021. Acceso 5 de diciembre de 2023. Disponible en: https://www.ibm.com/docs/en/SSLVMB_29.0.0/nl/es/pdf/IBM_SPSS_Statistics_Brief_Guide.pdf.

31. Fundación R. El Proyecto R para Computación Estadística. [Online]; 2023. Acceso 5 de diciembrede 2023. Disponible en: https://www.r-project.org/.

32. OpenEpi. Estadísticas epidemiológicas de código abierto para Salud Pública. [Online].; 2023. Acceso 5 de diciembre de 2023. Disponible en: http://www.openepi.com/Menu/OE_Menu.htm.

ISBN: 978-9942-609-27-4