MANUAL DE BIOESTADÍSTICA Y DEMOGRAFÍA

Autor
Mauricio Pérez Flórez

Pontificia Universidad
JAVERIANA
Cali

MANUAL DE
BIOESTADÍSTICA Y DEMOGRAFÍA

Pontificia Universidad JAVERIANA Cali

Santiago de Cali, 2021

MANUAL DE
BIOESTADÍSTICA Y DEMOGRAFÍA

Autor:

Mauricio Pérez Flórez

Pontificia Universidad JAVERIANA Cali

Santiago de Cali, 2021

Pérez Flórez, Mauricio
Manual de bioestadística y demografía / Mauricio Pérez Flórez. -- Santiago de Cali : Pontificia Universidad Javeriana, Sello Editorial Javeriano, 2021.

186 páginas: ilustraciones algunas a color, figuras, tablas ; 24 cm
Incluye referencias bibliográficas.
ISBN: 978-958-5177-95-6
ISBN(e): 978-958-5177-96-3

1. Bioestadística 2. Estadística vital 3. Salud pública -- Estadísticas 4. Demografía 5. Medicina -- Investigaciones -- Métodos estadísticos 6. Estudios demográficos I. Pérez Flórez, Mauricio II. Pontificia Universidad Javeriana Cali. Facultad de Ciencias de la Salud. Departamento de Salud Pública y Epidemiología.
SCDD 610.727 ed. 23 CO-CaPUJ
lmc/2021

Pontificia Universidad
JAVERIANA
Cali

Facultad de Ciencias de la Salud
Departamento de Salud Pública y Epidemiología
Manual de Bioestadística y Demografía

Autores:
© Mauricio Pérez Flórez

Colaboradores en la creación de contenidos:
© Jeinny Corrales

ISBN: 978-958-5177-95-6
ISBN(e): 978-958-5177-96-3

Formato: 17 x 24 cms
Coordinación editorial: Claudia Lorena González González
Asistente editorial: Jennifer Ramírez Martínez
Diagramación: Kevin Nieto Vallejo
Portada: Kevin Nieto Vallejo
Corrección de estilo: Comunicaciones Creativas
Impresión: Carvajal Soluciones de Comunicación S.A.S.

Pontificia Universidad Javeriana Cali
Calle 18 N°118-250
Teléfonos (57-2) 3218200

Santiago de Cali, Colombia, 2021.

Contenido

1. Introducción

Este manual universitario fue elaborado para que estudiantes universitarios de pre y posgrado de carreras del área de la salud consigan un entendimiento claro y sencillo de la bioestadística en cuanto a sus usos, alcances y herramientas más importantes para el manejo, análisis e interpretación de la información. También se abarcan los fundamentos de la demografía para el estudio de las dinámicas poblacionales en sus componentes de nacimientos, mortalidad y migración.

2. Bioestadística descriptiva

Se puede comenzar afirmando que la **estadística** es una ciencia relacionada con la recolección, organización, procesamiento, resumen, análisis e interpretación de **datos**, que además permite hacer inferencias[1] hacia toda una población a partir de los datos de una muestra (1). Las herramientas estadísticas (tablas, gráficas e indicadores) y sus técnicas de análisis inferencial (intervalos de confianza, pruebas de hipótesis, modelos de regresión, entre otros) se utilizan en varios campos del conocimiento incluyendo la salud, agricultura, industria, mercadeo, psicología, economía, entre otros.

La **bioestadística** es básicamente la aplicación de la estadística en el campo de las ciencias biológicas y de la salud. En otras palabras, la estadística recibe el nombre de bioestadística cuando los datos que se analizan provienen de seres vivos o fenómenos biológicos (1).

La importancia del entendimiento de la estadística radica en que la información (los datos) está en todas partes y un buen análisis de ellos, con técnicas estadísticas apropiadas, permitirá tomar decisiones bien informadas.

Hoy día los profesionales de la salud (médicos, salubristas, enfermeros, nutricionistas, etc.) necesitan estar actualizados, requieren leer y entender artículos científicos, interpretar tablas, gráficas y resultados. Estos profesionales deben ser críticos frente a la literatura que lean y deben estar en la capacidad de emitir juicios sobre su calidad (2). Por ello, requieren una formación científica y manejar conceptos de la bioestadística y demografía, además de la epidemiología.

La materia prima de la estadística son **los datos** resultados de mediciones realizadas en personas, animales, muestras biológicas o cualquier otra unidad de estudio. Estos suelen ser almacenados en bases de datos como se muestra en la Figura 1, donde el esquema general es que cada fila (registro) representa a una unidad de estudio y cada columna corresponde a la medición específica de una variable.

1 Inferir es deducir algo u obtener una conclusión con base a otra cosa.

Variables

Registros

	Variable1	Variable2	Variable3	Variable4	.....
Registro1					
Registro2					
Registro3					
Registro4					
Registro5					
Registro6					
Registro7					
Registro8					
Registro9					
Registro10					
.....					

Figura 1. Esquema general de una base de datos. Elaboración propia.

La estadística se encarga de analizar los datos contenidos en una base de datos para obtener información de fácil comprensión y que ayude a la toma de decisiones (Figura 2). Como veremos a continuación una base de datos puede generarse a partir de un proceso de investigación o ser parte de un proceso de recolección rutinario a partir de diversas instituciones.

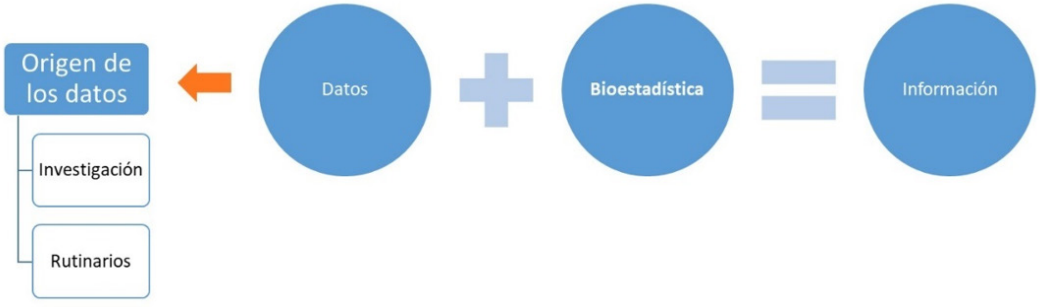


Figura 2. Esquema mostrando la utilidad de la Bioestadística para convertir los datos en información. Elaboración propia.

2.1 Bioestadística en la salud

En la Figura 3 se quiere resaltar la utilidad de la estadística en el campo de la salud poblacional y los servicios de salud. La epidemiología (clínica y poblacional) y la salud pública son ciencias poblacionales que utilizan información de las comunidades (o pacientes en el caso de la clínica) para cumplir con sus objetivos. Las fuentes de datos que utiliza la epidemiología para cumplir su función, pueden provenir de fuentes "primarias" o "secundarias", como se explicará a continuación, y su análisis se hace usando las herramientas de la estadística. Por esto es importante que los profesionales de la salud adquieran habilidades en la manipulación de bases de datos e interpretación de la información.

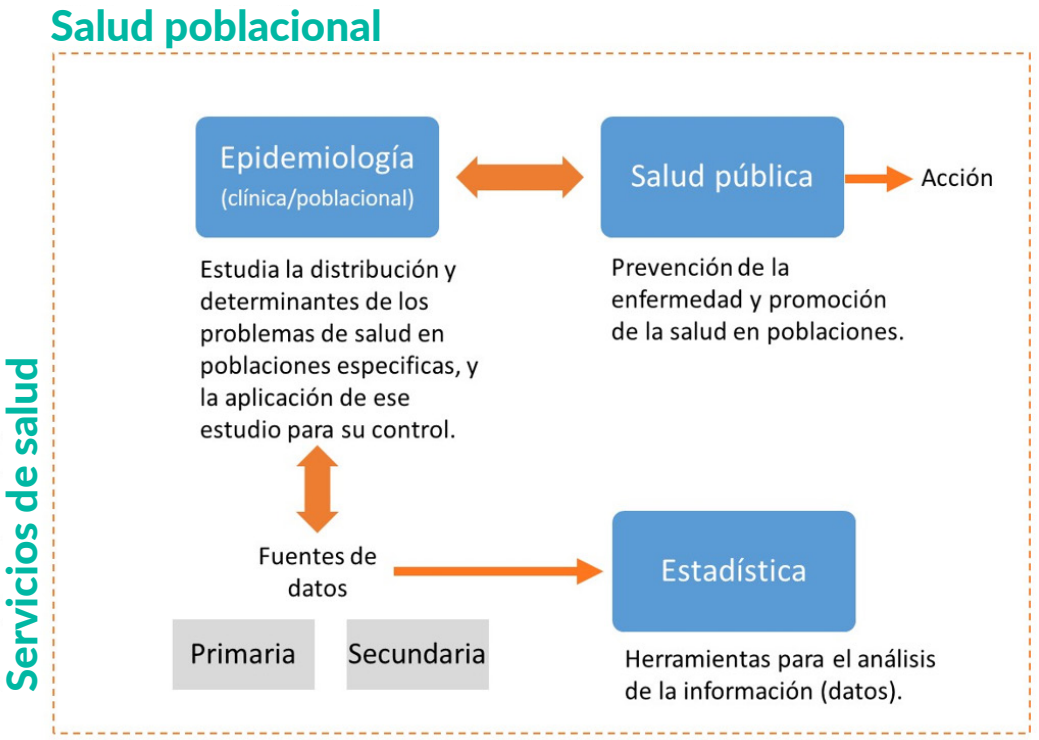


Figura 3. Papel de la estadística en el área de la salud. Elaboración propia.

En general, puede decirse que los datos que utilizan los profesionales de la salud o investigadores provienen de dos grandes fuentes:

1. Datos resultantes de un estudio de investigación (descriptivo o analítico) → Datos primarios

2. Datos generados rutinariamente por instituciones → Datos secundarios

Utilizar *fuentes de datos primarias* (datos primarios) significa que un investigador recolectó sus propios datos a partir de un proceso de investigación para fines de su estudio. En otras palabras; son datos recolectados para un estudio o investigación en particular.

Usar *fuentes de datos secundarias* (datos secundarios) se refiere al uso de información existente que ha sido generada por otros investigadores o instituciones.

En todo proceso de investigación con un enfoque cuantitativo se genera (datos primarios) o utiliza (datos secundarios) información para responder alguna pregunta de investigación.

2.1.1 El proceso de investigación

La ciencia tiene como objetivo principal generar conocimiento científico, no solo describe eventos y fenómenos, sino que también se interesa en explicar cómo y porqué suceden. El conocimiento científico para resolver un problema (pregunta de investigación) o evaluar una hipótesis (3) se genera al poner en marcha el método científico[2] por medio de una investigación, la cual puede ser vista como un proceso sistemático, ordenado y objetivo .

En las ciencias de la salud se generan investigaciones enfocadas a entender procesos biológicos, describir enfermedades o eventos de interés en tiempo, lugar y persona, evaluar métodos diagnósticos, tratamientos o intervenciones, determinar factores de riesgo que afectan la salud de la población, entre otros. La manera de abordar estos u otros interrogantes se resume en la Figura 4 (4).

2 El método científico es una estrategia estandarizada o proceso sistemático cuya finalidad es obtener conocimientos válidos del mundo real. Este método pretende explicar fenómenos, establecer relaciones entre variables y enunciar leyes que expliquen la realidad. Los científicos que emplean el método científico siguen las siguientes fases: observación, formulación de hipótesis, experimentación, emisión de conclusiones.

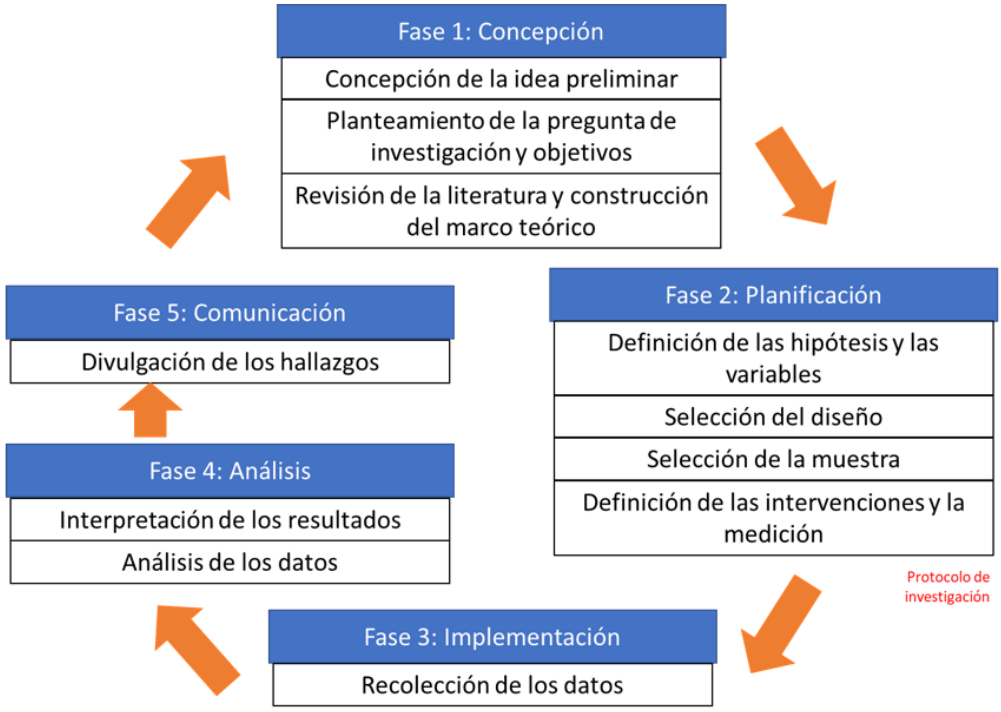


Figura 4. Esquema general del proceso de investigación. Elaboración propia, adaptado de Ruiz & Morillo. Epidemiología Clínica. Primera edición. Editorial Médica Internacional, 2004 (4).

En la fase inicial (Fase 1) el investigador concibe la idea de su investigación, plantea el problema de investigación, lo delimita y justifica el estudio, plantea la pregunta de investigación y/o hipótesis, objetivos del estudio, hace una revisión de literatura y búsqueda de información que guiarán el proceso.

En la fase de planificación del estudio (Fase 2) se propone la metodología a seguir para responder a los objetivos del estudio. El investigador define las variables de interés en su estudio, selecciona el diseño más adecuado (investigación cualitativa[3], cuantitativa[4] o mixta), define su población de estudio, la muestra, las intervenciones y la manera en que se hará la medición de las variables. El resultado final de esta fase es el protocolo del estudio.

La fase de implementación (Fase 3) se refiere a la recolección de la información planeada en la Fase 2. De acuerdo con el diseño del estudio se utilizan metodologías como la observación directa, grupos focales, entrevista en profundidad, encuestas, diarios de campo, o también dispositivos, exámenes médicos, pruebas diagnósticas, entre otros.

3 Por ejemplo un análisis documental, análisis de casos, etnografía, teoría fundamentada, fenomenología, investigación-acción, entre otros.
4 Por ejemplo un estudio transversal, cohortes, casos y controles, ensayo clínico, ecológico, entre otros.

En la fase de análisis (Fase 4) se hace una descripción de los hallazgos e interpretación de los resultados. En la investigación cuantitativa el análisis de los datos involucra metodologías y técnicas estadísticas que analizan objetivamente la información.

La de comunicación (Fase 5) hace referencia a la divulgación de los hallazgos (resultados, conclusiones y recomendaciones), que dan respuesta a los objetivos del estudio. Se espera llenar vacíos de información, aportar nuevo conocimiento y/o generar nuevas hipótesis que den continuidad al proceso de investigación. El producto final de esta fase, y de la investigación como tal, regularmente es un artículo o manuscrito que se publica en una revista especializada en el tema. Otras opciones para la difusión de los hallazgos del estudio son reuniones académicas como congresos, seminarios o simposios.

En este proceso de investigación, la epidemiología y la estadística son herramientas indispensables. La epidemiología se reconoce como una ciencia básica de la clínica y la salud pública, aportando protocolos y criterios para hacer una investigación rigurosa. La estadística aporta métodos para recolectar, presentar, resumir y analizar objetivamente la información, además que permite el estudio de muestras y hacer inferencias hacia la población de estudio (3).

La epidemiología y la estadística aportan elementos metodológicos importantes en el desarrollo de una investigación con enfoque cuantitativo. La Figura 5 pretende resaltar que una vez definidos los objetivos de cualquier estudio de investigación (considerados el pilar de una investigación), el marco metodológico define el proceso que se debe seguir para alcanzar tales objetivos. Se debe seleccionar un diseño epidemiológico (observacional o experimental; descriptivo o analítico; transversal o longitudinal) acorde a la pregunta de investigación. Se recomienda revisar los aspectos generales de los principales diseños epidemiológicos (5,6) para entender los fundamentos, usos y aplicaciones de un estudio transversal (7), cohortes (8), casos y controles (9), ensayos clínicos (10), ecológicos (11), entre otros.

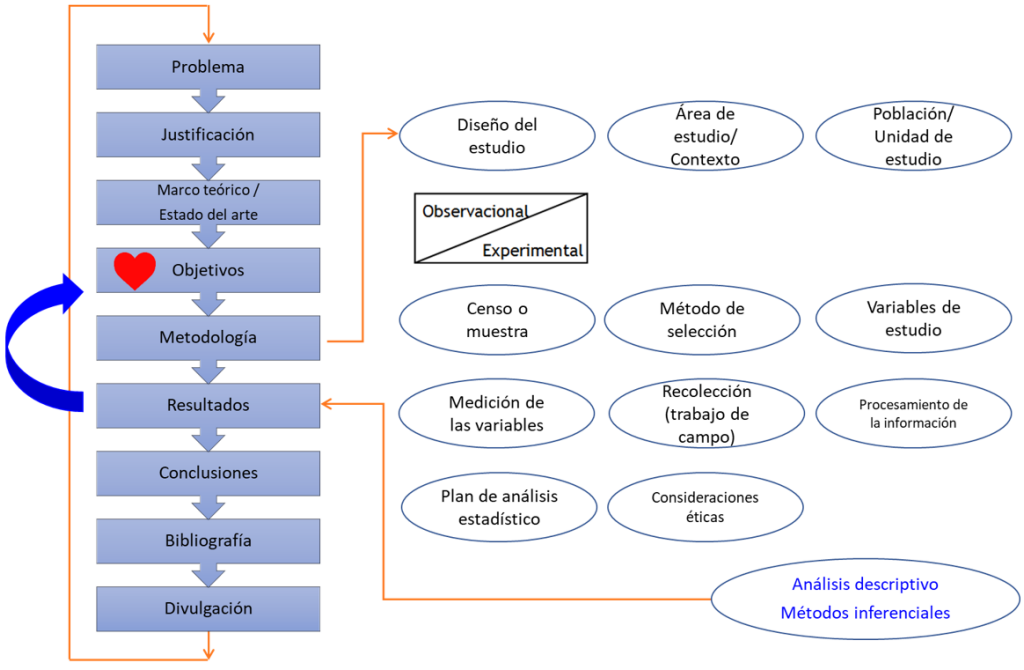


Figura 5. Esquema general del proceso de investigación resaltando los elementos metodológicos donde aportan la epidemiología y la estadística. Elaboración propia.

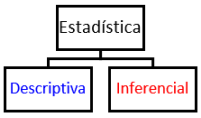
Los otros componentes conceptuales del marco metodológico mostrados en la Figura 5 serán presentados a continuación, incluyendo la población de estudio, realización de un censo o un muestreo, definición de variables de estudio y su clasificación, procesamiento de la información y plan de análisis estadístico acorde al diseño del estudio.

2.1.2 Conceptos generales

Como se mencionó previamente la **estadística** es una ciencia relacionada con la recolección, organización, procesamiento, resumen, análisis e interpretación de **datos**, que además permite hacer inferencias a una población a partir de los datos de una muestra (1).

En esta definición están bien diferenciadas dos ramas bien conocidas y usadas de la estadística tradicional (también llamada estadística clásica)[5]:

5 Según la forma de abordar los problemas de la estadística, esta se clasifica en estadística clásica y estadística bayesiana. Este manual aborda los temas de la estadística clásica.



la estadística es una ciencia relacionada con la recolección, organización, procesamiento, resumen, análisis e interpretación de datos.

que además permite hacer inferencias a una población a partir de los datos de una muestra.

Luego, el estudio de la estadística clásica se divide en dos grandes ramas: la estadística descriptiva y la estadística inferencial (Figura 6).

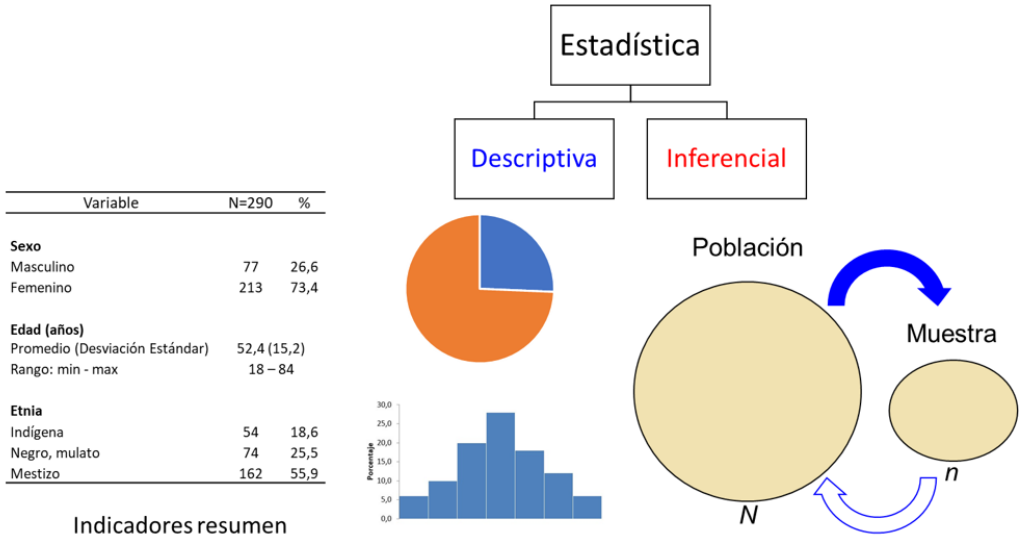


Figura 6. Esquema general de las ramas de la Estadística[6]. Elaboración propia.

La **estadística descriptiva** se usa para describir las características de un conjunto de datos (de una muestra o de una población) que puede servir para tomar decisiones. La estadística descriptiva comprende un conjunto de métodos para organizar, resumir y presentar los datos de manera informativa. Esta rama de la estadística utiliza tablas, gráficas e indicadores resumen para su finalidad.

La **estadística inferencial** abarca un conjunto de métodos utilizados para descubrir características de una población a partir de una muestra tomada de ella. En un lenguaje más estadístico, esta rama de la estadística es utilizada para estimar "parámetros"[7]

6 *N* simboliza el número de elementos de la población y *n* el número de elementos de una muestra
7 Un parámetro en estadística es una característica que se mide en la población.

de la población a partir de "estimadores"[8] con base en una muestra. Las principales herramientas de la estadística inferencial aplicadas en el campo de la salud son:

1. La estimación de parámetros por medio de intervalos de confianza y pruebas de hipótesis.

2. Las técnicas de regresión y correlación.

El módulo 2 de este manual se enfocará en el alcance, objetivos y herramientas de la estadística inferencial. Por ahora, es necesario definir algunos conceptos importantes.

En estadística el concepto **población** o **población de estudio** hace referencia al conjunto de todos los elementos (individuos, objetos, unidades, mediciones, etc.) de interés en un estudio, a partir del cual se obtendrá información y hacia el que se extenderán las conclusiones del estudio (1). Este concepto no debe relacionarse exclusivamente con seres humanos, debido a que -por ejemplo- en la investigación en salud la población de estudio puede ser personas (pacientes con determinado evento, embarazadas, usuarios del sistema de salud, personal de salud, médicos, familias, etc..), entidades (centros de salud, hospitales, clínicas, instituciones, etc.), unidades (animales, plantas, células, muestras biológicas de sangre, orina, etc.) o registros/ expedientes (defunciones, casos de malaria, etc.), entre otros. Tiene sentido decir que la población de estudio se determina o define según el estudio y campo de interés. Es importante resaltar que en un estudio la población debe estar adecuadamente definida en tiempo y espacio para ser capaces de decidir si un elemento pertenece o no a la población y poder ser incluido en el estudio. La letra "*N*" simboliza el número de elementos en la población de estudio.

Ejemplo 1. Población de estudio

a. Suponga que el interés de un investigador es obtener información sobre el estado de salud en una comunidad ABC. Más específicamente desea conocer la prevalencia de desnutrición en niños menores de 5 años. → En esta investigación la población de estudio son todos los niños menores de 5 años de la comunidad ABC. Además, el periodo de estudio definirá temporalmente la población: por ejemplo, todos los niños menores de 5 años de la comunidad ABC en marzo de 2018.

b. Ahora, si el interés del investigador hubiese sido determinar los niveles de plomo en sangre en las mujeres embarazadas en esa misma comunidad, la población de estudio hubiese sido diferente → Mujeres en estado de embarazo residentes de la comunidad ABC en marzo de 2018.

8 Un estimador es una característica que se mide en la muestra.

Cada elemento de la población se conoce como **unidad de estudio**, de la cual se necesita información; es decir, en donde se harán las mediciones para obtener el dato. La unidad de estudio es el sujeto de interés en una investigación.

En la investigación cuantitativa el investigador tiene la opción de hacer un censo o tomar una muestra de la población de estudio. Un **censo** es el estudio de todos y cada uno de los elementos que conforman su población de estudio. Cuando es posible el estudio de toda la población se obtendrá resultados exactos de ella; es decir, se podrán conocer los valores exactos de los parámetros. Sin embargo, en muchas ocasiones no es factible la realización de un censo por diferentes razones:

1. Regularmente el estudio de todos los elementos de una población resulta costoso, pues consume tiempo, dinero, recurso humano, etc.

2. En ocasiones la población es tan grande que excede las posibilidades del equipo de investigación.

3. Cuando las pruebas o mediciones que se hacen a los sujetos causan su destrucción.

Cuando no es posible hacer un censo, lo que ocurre muy frecuentemente, se toma una muestra, la cual es definida como una parte o un subconjunto de la población de estudio que se selecciona para su estudio. La razón de ello es que se hace difícil el estudio de todos los elementos de la población, principalmente por limitación de recursos, capacidad logística y tiempo, entre otros. No obstante, se espera que los resultados obtenidos de una muestra sean un reflejo de la población, y así poder usar esta información para hacer inferencias acerca de esta[9]. Por lo anterior, se requiere asegurar la selección de una "buena muestra" -lo cual no es tarea fácil-. La letra "*n*" simboliza el tamaño de una muestra.

Toda muestra debe cumplir con unos requerimientos mínimos que la hagan apropiada en un estudio y permita extrapolar los resultados a la población. Una "buena muestra" debe ser representativa y adecuada.

a. **Representativa**: se espera que una muestra refleje las características de la población. Por ejemplo, si en la población el 50 % son hombres y el 50 % son mujeres, no debe seleccionarse una muestra de solo hombres o solo mujeres. Para asegurar que una muestra es representativa se utiliza el "azar" o el recurso "aleatorio", porque la teoría dice que en promedio se obtendría una muestra representativa (12). En estadística los métodos de muestreo probabilísticos[10] utilizan el azar y por lo tanto aseguran una muestra representativa. Algunos procedimientos utilizados para la selección de las unidades que conformarán la muestra son la generación de números aleatorios en computador, calculadoras científicas, tablas de números aleatorios, etc.

9 Esta es una de las principales características de la investigación cuantitativa
10 Muestreo aleatorio simple, estratificado, sistemático, conglomerado y polietápico.

b. **Adecuada**: esta característica hace referencia al tamaño de la muestra. Por ejemplo, si en una comunidad hay 5 000 habitantes, no sería muy confiable en los resultados de una muestra de 5 o 10 personas. En una muestra aleatoria entre mayor sea la muestra, mejor es la inferencia que se puede hacer. El cálculo del tamaño de muestra se determina según el estudio, el parámetro de interés, la precisión deseada, la confiabilidad, la variación de los datos, el tipo de muestreo, entre otros. Este paso de la investigación cuantitativa determina -entre otros- la factibilidad del estudio, su duración y costo. Esta tarea pertenece al estadístico vinculado al estudio y debe ser realizada en la etapa de planeación.

Una de las características de la investigación cuantitativa es que permite examinar los datos de manera científica, matemática y objetiva con ayuda de herramientas de la estadística. En la investigación cuantitativa se definen un conjunto de variables que ayudarán a responder los objetivos del estudio. El investigador debe hacer una definición clara y precisa de las variables, y estandarizar la forma en que serán medidas.

Una **variable** es una característica observable o medible en las unidades de estudio de la población. Dicha característica se define como variable cuando cambia de un individuo a otro, o en el mismo individuo según el momento en que se mida. Por ejemplo; el sexo, el nivel de estudios, la estatura, el peso, la talla, la temperatura, etc. son variables porque toman diferentes valores en las unidades de estudio.

Se han propuesto diferentes clasificaciones de las variables según su naturaleza, nivel de medición y relación que guardan unas con otras, según como se observa en la Tabla 1:

Tabla 1. Clasificación de variables.

Naturaleza de la variable	Nivel de medición	Relación de unas con otras (cuando aplica)
	Nominal	
Cuantitativas o numéricas	Ordinal	Dependiente o respuesta
Cualitativas o categóricas	De intervalo	Independiente o explicativa
	De razón	

Nota. Elaboración propia.

Naturaleza de las variables:

Cuando la característica que se estudia es *no numérica* (atributos o cualidades) se conoce como variable cualitativa o categórica, mientras que si esta se puede reportar en forma numérica se dice que la variable es *cuantitativa* o *numérica*. La clasificación de las variables según su naturaleza se muestra en la Figura 7 y algunos ejemplos en la Tabla 2.

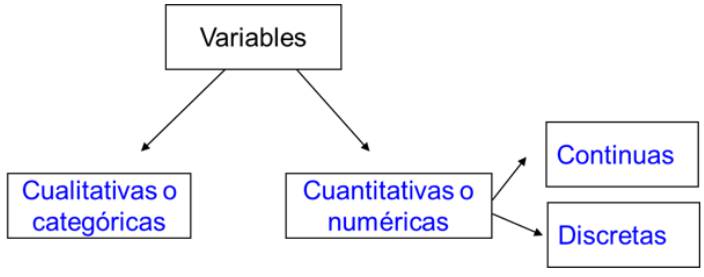


Figura 7. Esquema general de clasificación de las variables según su naturaleza. Elaboración propia.

Tabla 2. Definición y ejemplos de "variables" según su naturaleza.

<table>
<tr><th>Tipo de variable</th><th colspan="2">Descripción</th><th>Ejemplos</th><th>Valores de la variable</th></tr>
<tr><td rowspan="4">Cualitativa o categórica</td><td colspan="2" rowspan="4">Son variables que describen atributos o cualidades (no numéricos) en las unidades de estudio</td><td>Sexo</td><td>Masculino
Femenino</td></tr>
<tr><td>Estado del paciente al egreso</td><td>Vivo
Muerto</td></tr>
<tr><td>Nivel de estudios</td><td>Ninguno
Primaría
Secundaria
Técnico o tecnológico
Universitario
Posgrado</td></tr>
<tr><td>Clasificación de una enfermedad</td><td>Leve
Moderada
Severa</td></tr>
<tr><td rowspan="4">Cuantitativas o numérica</td><td colspan="4">Son aquellas variables que describen atributos en las unidades de estudios asociados a datos numéricos. Estas pueden ser continuas o discretas.</td></tr>
<tr><td rowspan="2">Continuas</td><td rowspan="2">Cuando el dato numérico puede tomar infinitos valores dentro de un rango. La precisión en la medición depende del instrumento de medición</td><td>Peso</td><td>Un valor asociado a los reales. Por ejemplo:
66 kilos; 55,4 kilos; 48,75 kilos; 7,7524 kilos</td></tr>
<tr><td>Temperatura</td><td>Por ejemplo: 37 °C, 38.5 °C, 39.27 °C</td></tr>
<tr><td rowspan="2">Discretas</td><td rowspan="2">Cuando el dato que toma la variable es un número entero. Regularmente son variables asociadas con conteos</td><td>Número de dientes con caries</td><td>Valores enteros. Por ejemplo: 0, 1, 2, 3, 4, …</td></tr>
<tr><td>Número de infartos</td><td>Valores enteros. Por ejemplo: 0, 1, 2, 3, 4, …</td></tr>
</table>

Nota. Elaboración propia.

Nivel de medición o escala de medición de las variables:

De acuerdo con el nivel de medición de las variables, estas se clasifican en nominal, ordinal, de intervalo y de razón (Figura 8). Los niveles de medición nominal y ordinal pertenecen a las variables cualitativas, y los niveles de intervalo y de razón son exclusivos de las variables cuantitativas. Algunos ejemplos se presentan en la Tabla 3.

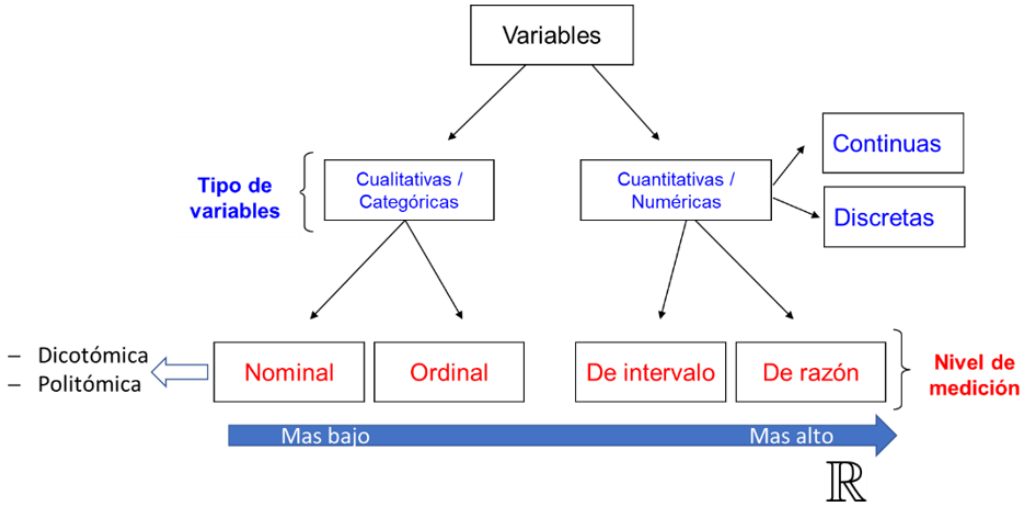


Figura 8. Esquema general de clasificación de las variables según su naturaleza y nivel o escala de medición. Elaboración propia.

Tabla 3. Definición y ejemplos de "variables" según su nivel de medición.

Nivel	Descripción	Ejemplos	Valores de la variable[11]	Comentarios
Nominal	Es el nivel de medición "más bajo". Aquí los valores que puede tener una variable cualitativa son categorías que no tienen ningún orden en particular. Estas variables pueden ser dicotómicas (dos categorías) o politómicas (más de dos categorías)	Sexo	1. Masculino 2. Femenino	"Masculino" y "Femenino" son 2 atributos que no tienen un orden. Sexo es una variable dicotómica.
		Estado del paciente al egreso	0. Vivo 1. Muerto	"Vivo" y "Muerto" son 2 atributos que no tienen un orden. Los números 0 y 1 son solo etiquetas. Esta es una variable dicotómica.
		Área de atención de un medico	1. UCI 2. Consulta externa 3. Cirugía 4. Pediatría	Esta es una variable politómica porque las categorías de la variable son más de dos.
Ordinal	Cuando las categorías de la variable tienen un orden intrínseco. Aquí es importante el orden en que se presentan las categorías	Nivel de estudios	1. Ninguno 2. Primaría 3. Secundaria 4. Técnico o tecnológico 5. Universitario 6. Postgrado	A pesar de que la variable es cualitativa, las categorías de "Nivel de estudios" tienen un orden intrínseco, de menos a más estudios. Los códigos de 0 a 6 solo son etiquetas
		Clasificación de una enfermedad	1. Leve 2. Moderada 3. Severa	Entre leve, moderada y severa hay un orden importante. Los códigos de 1 a 3 solo son etiquetas

11 Con el fin de facilitar su manipulación, es común codificar las variables; esto es asignarles códigos a sus valores; por ejemplo para sexo se puede usar 1 "Masculino" y 2 "Femenino" o viceversa. Sin embargo, a pesar de codificar la variable sexo, su naturaleza sigue siendo cualitativa o categórica.

De intervalo	Son valores cuantitativos usando una escala de medición que NO corresponde a los números reales. La escala de medición tiene estas características: 1. Incluye un cero arbitrario que no indica ausencia de la condición 2. No tienen sentido las comparaciones	Temperatura	Por ejemplo: 20 °C 40 °C	Una temperatura de 0 °C no significa ausencia de temperatura. Además no podemos afirmar que 40 °C es dos veces más caliente que 20 °C
		Coeficiente intelectual (CI)	Por ejemplo: 70 140	Un CI de 0 no significa ausencia de inteligencia. Además, no podemos decir que una persona con un CI de 140 sea 2 veces más inteligente que una persona con un CI de 70
De razón	Es la escala o nivel de medición "más alta". Se asociada con los números reales. Se caracteriza porque: 1. El punto cero es significativo, e indica ausencia de la característica 2. Se pueden hacer comparaciones válidas	Número de dientes con caries	Por ejemplo: 0, 1, 2, 3, 4, …	Aquí el 0 significa ausencia de dientes con caries. Además, una persona con 4 dientes con caries tiene el doble que una persona con 2 dientes cariados
		Peso	Por ejemplo: 50 100	Una persona que pesa 100 kilos tiene el doble de peso que otra persona de 50 kilos.

Nota. La importancia de definir el tipo de variable y su nivel de medición para cada una de las variables de interés en una investigación radica en que los métodos estadísticos que se utilizan para el análisis (descriptivo o inferencial) dependen del tipo de datos. Elaboración propia.

Relación que guardan las variables unas con otras:

En algunos estudios las variables se pueden clasificar según la relación que guardan unas con otras; en variables *dependientes* o *independientes*.

- Variables dependientes, también llamadas variables respuesta.
- Variables independientes, también llamadas variables explicativas.

Esta clasificación se usa cuando los estudios pretenden explorar, identificar o determinar factores asociados a un evento de interés. En este caso el evento de interés es la *variable respuesta* y los factores que pueden modificar esa respuesta son las *variables independientes*.

Ejemplo 2. Variables dependientes e independientes

El siguiente ejemplo es adecuado para identificar y clasificar las variables de un estudio como *dependiente* e *independientes* (cuando aplica):

Tabla 4. Ejemplo Variables dependientes e independientes.

Factores relacionados con una *enfermedad* o *evento* de interés	
Los factores que pueden modificar la enfermedad son las variables independientes	La enfermedad o evento es la variable respuesta (variable dependiente)
Por ejemplo, el estudio de "Framingham Heart Study"[12] se propuso identificar las causas de la enfermedad cardiovascular, para lo cual midieron variables como el colesterol y la presión arterial. En este estudio la variable respuesta era la enfermedad cardiovascular y las variables independientes fueron el nivel del colesterol y la presión arterial.	

Nota. Elaboración propia

Generalmente en los estudios epidemiológicos la enfermedad o evento relacionado con la salud es la variable dependiente, mientras que los factores que inciden o modifican su aparición, magnitud y distribución se consideran las variables independientes (13).

12 Framingham heart study [Internet]. Framinghamheartstudy.org. [citado el 16 de diciembre de 2021]. Disponible en: https://framinghamheartstudy.org/

2.1.3 Ejercicios propuestos

Ejercicio 1.

Para cada una de las siguientes variables, indicar el tipo de variable y su escala o nivel de medición:

Tabla 5. Ejercicio 1.

Variable	Tipo de variable			Escala o nivel de medición
	Cualitativa	Cuantitativa		
		Discreta	Continua	
Gravedad del edema de miembros inferiores (Leve/Moderado/Grave)				
Temperatura diaria de un refrigerador (°C)				
Número de dientes con caries en un paciente				
Tiempo de hospitalización en UCI de un paciente medido desde la hora de ingreso hasta la hora de salida				
Diagnóstico clínico del paciente				
Respiraciones por minuto				
Sexo del paciente				
Edad del paciente				
Peso de un recién nacido				
Número de muertes en Cali en 2012				
Grupo sanguíneo				
Nivel socioeconómico				
Puntaje de un examen de Estadística				
Existencia de un programa de control del dengue				
Nivel de satisfacción del paciente con la atención recibida				
Edad del paciente (años cumplidos)				

Ejercicio 2.

Utiliza los siguientes conceptos para rellenar los espacios en blanco:

Censo	Cualitativas	Cuantitativas continuas
Descriptiva	Parámetro	Inferencial
Muestra	Población de estudio	Representativa, adecuada
Variable	Unidad de estudio	Estadística
Estimador		

a. La _____ proporciona métodos para organizar y resumir datos, y para sacar conclusiones válidas y confiables con base en la información que contienen los datos.
b. Una investigación por lo común tiene que ver con una colección muy bien definida de individuos, pacientes, unidades u objetos que forman la _____ de una investigación.
c. Cuando se cuenta con la misma información para todos los individuos u objetos de la población, se tiene lo que se llama un _____.
d. Las limitaciones de tiempo, dinero y demás recursos insuficientes, hacen impráctico o imposible hacer investigaciones con toda la población blanco, es decir se hace difícil levantar un censo. En vez de eso se selecciona una _____.
e. En una investigación la población de estudio fueron los recién nacidos en el Hospital Metropolitano de la ciudad durante enero de 2013. El peso promedio de estos bebés fue de 3 500 g. Este valor se refiere a un _____.
f. Una _____ es cualquier característica cuyo valor cambia entre los individuos u objetos de la población.
g. Las _____ son las entidades que serán objeto de medición en una investigación
h. La estatura de una persona, su peso, talla, nivel de glucemia, temperatura, entre otros, son ejemplos de variables de tipo _____.
i. El sexo, nivel de educación, diagnóstico del paciente, resultado de una prueba (+/-), son algunos ejemplos de variables de tipo _____
j. Las técnicas para hacer una generalización de toda la población a partir de una muestra se ubican dentro de la rama de la estadística llamada _____
k. La rama de la estadística que utiliza métodos gráficos (diagrama de barras, gráfico de sectores, histogramas, etc.) y también medidas numéricas de resumen como promedio, mediana, desviación estándar, rango, entre otras es la _____
l. Una muestra es una parte o subconjunto de la población de estudio que se selecciona para ser parte de una investigación. Las dos características que debe cumplir una muestra es que esta debe ser _____ y _____. Reflexiona al respecto.
m. El peso promedio de una muestra de 40 recién nacidos en el Hospital Metropolitano de la ciudad durante enero de 2013 fue de 3 600 g. Este valor se refiere a un _____.

2.2 Herramientas de la bioestadística descriptiva

Como se mencionó, la estadística descriptiva es la rama de la estadística que tiene como finalidad organizar, presentar y resumir grandes cantidades de datos por medio de tablas, gráficas e indicadores resumen.

Con fines académicos se pueden agrupar las herramientas de la bioestadística descriptiva en dos grandes categorías: 1) tablas y gráficas; e 2) indicadores resumen.

2.2.1 Presentación de datos en tablas y gráficas

La información que se recolecta en una investigación cuantitativa o mixta debe ser organizada y presentada en el *capítulo de resultados* de una investigación. En general, los artículos científicos que resultan de la investigación cuantitativa se caracterizan porque la primera tabla de resultados ("Tabla 1") contiene un análisis descriptivo de la población de estudio, mostrando sus principales características sociodemográficas.

El análisis descriptivo de un conjunto de datos debe comenzar con la organización y presentación de los datos en tablas de frecuencias o de manera visual mediante gráficas adecuadas.

Una de las aplicaciones donde mejor se refleja la utilidad de la estadística descriptiva, es en la organización de conjuntos de datos, o bases de datos, por medio de **tablas** y **gráficos** que muestren fácilmente la información de interés y permitan su interpretación.

A continuación, se mencionan algunas herramientas (tablas y gráficos) de la estadística descriptiva para la organización y presentación de los datos, según la naturaleza de las variables:

Tabla 6. Herramientas de la estadística descriptiva.

Tipo de variables	Tablas	Gráficas
Variables cualitativas	Tabla de frecuencias (una variable) Tabla cruzada o tabla de contingencia (dos variables)	Gráfico de sectores o pastel Gráfico de barras (una o dos variables)
Variables cuantitativas	Tabla de distribución de frecuencias	Histograma Diagrama de cajas o boxplot* Gráfico de dispersión (dos variables numéricas) Gráfico de líneas (tendencia)*

Nota. Estos gráficos se usan para representar variables numéricas, aunque su representación puede considerar grupos definidos por alguna(s) variable(s) cualitativa(s). Elaboración propia.

Para mostrar la utilidad de la Estadística en la organización y presentación de información, se tomará como ejemplo una "pequeña" base de datos. Sin embargo, su proceso se puede generalizar a bases de datos más complejas (con mayor número de registros y/o variables).

Ejemplo 3. Pacientes con diagnóstico de asma

Suponga que un médico registra la información de los pacientes que asisten a consulta. En el último año se atendieron 20 pacientes con diagnóstico de asma. La información recolectada se presenta en la base de datos de la Tabla 7.

Tabla 7. Base de datos conteniendo información de 20 pacientes con diagnóstico de asma.

Código	Sexo	Antecedentes	Hospitalización	Edad
10001	2	1	2	15
10002	1	1	1	16
10003	2	0	3	17
10004	1	0	1	17
10005	2	0	0	17
10006	1	1	1	18
10007	2	0	4	18
10008	1	1	2	19
10009	2	0	1	19
10010	2	0	4	19
10011	1	1	0	19
10012	2	0	2	20
10013	1	1	0	20
10014	2	0	1	20

10015	2	0	5	20
10016	1	1	3	21
10017	2	0	0	21
10018	2	1	2	22
10019	1	1	1	22
10020	2	0	2	23

Nota. Elaboración propia.

El diccionario de datos[13] (Codebook) de esta base de datos se presenta en la Tabla 8 .

Tabla 8. Diccionario de datos de la base de datos anterior.

#	Nombre de la variable	Descripción	Valores
1	Código	Código del paciente	Códigos
2	Sexo	Sexo del encuestado	1=Masculino 2=Femenino
3	Antecedentes	Antecedentes de familiares con asma	0=No 1=Sí
4	Hospitalización	Número de hospitalizaciones en el último año	0+
5	Edad	Edad en años cumplidos	0+

Nota. Elaboración propia.

Presentación de datos con variables cualitativas o categóricas

Como se mencionó previamente, existen muchas formas para presentar los datos, sin embargo, los más utilizados son las tablas y las gráficas.

Para organizar y presentar **variables categóricas** se pueden usar tablas de frecuencias, tablas de contingencia (tablas cruzadas), gráficas de sectores (pastel) o gráficos de barras.

Utilizando algunas variables categóricas del Ejemplo 3, las opciones para presentar algunos resultados son:

13 El diccionario de datos es una tabla que describe cada una de las variables recolectadas en un estudio. A cada una de las variables se le asigna un nombre de una palabra (o más de una palabra separadas por guion bajo, ejemplo: "edad_madre"), se hace una descripción de la variable y se definen los valores de las variables que fueron recodificadas. Este diccionario de datos es obtenido de la "Tabla de variables" del estudio.

Tabla 9. Opciones de presentación de variables.

Tabla de frecuencias (una variable)

Tabla x. **Tabla de frecuencias** del sexo de los pacientes

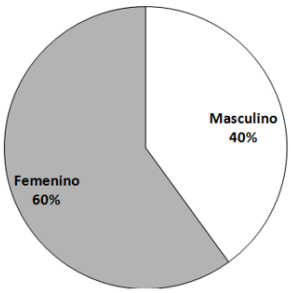
Sexo	Número de Personas	Porcentaje
Masculino	8	40
Femenino	12	60
Total	20	100

Tabla de contingencia o cruzada (dos variables)

Tabla x. **Tabla cruzada** del sexo del paciente según los antecedentes de familiares con asma

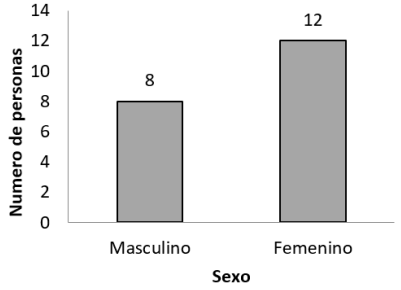
Sexo	Antecedentes n (%)		Total
	Sí	No	n (%)
Masculino	7 (78)	1 (9)	8 (40)
Femenino	2 (22)	10 (91)	12 (60)
Total	9 (100)	11 (100)	20 (100)

Diagrama de sectores o pastel



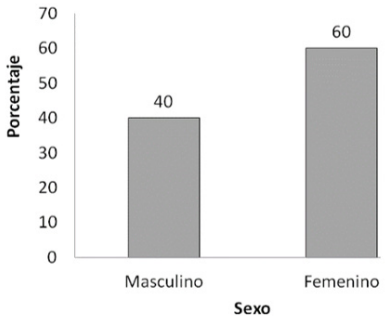
Gráfica 1. **Diagrama de sectores** del sexo de los pacientes con asma (*n*=20). Elaboración propia.

Diagrama de barras (1 variable)



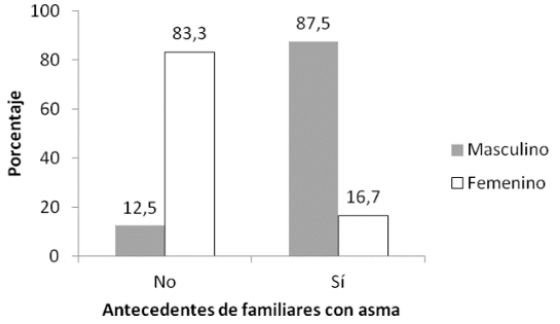
Gráfica 2. **Diagrama de barras** del sexo de los pacientes (*n*=20). Elaboración propia.

Diagrama de barras (una variable)



Gráfica 3. **Diagrama de barras** del sexo de los 20 pacientes con asma. Elaboración propia.

Diagrama de barras (dos variables)



Gráfica 4. **Diagrama barras** del sexo del paciente según los antecedentes de familiares con asma (*n*=20). Elaboración propia.

Nota. La Gráfica 2 y la Gráfica 3 muestran la misma información, la única diferencia es que la primera muestra la frecuencia absoluta (número de personas) y la segunda el porcentaje. En general, se prefieren las gráficas en porcentajes, ya que son más informativas.Elaboración propia.

Una **tabla de frecuencias**, también llamada tabla de distribución de frecuencias, corresponde a la clasificación del grupo de datos de "una variable" de clasificación.

Por ejemplo, la siguiente tabla de frecuencias muestra la distribución del "sexo" de los 20 pacientes con asma:

Tabla 10. Tabla de frecuencias del sexo de los pacientes (n=20).

(1)	(2)	(3)
Sexo	Número de personas	Porcentaje
Masculino	8	40
Femenino	12	60
Total	20	100

Nota. Elaboración propia.

- La columna 1 de la tabla contiene la variable en estudio, en este caso es el "Sexo" seguido de las categorías de la variable "Masculino" y "Femenino". Al final de la tabla siempre se coloca el "Total".
- La columna 2 contiene la frecuencia absoluta, esto es el conteo de elementos en cada clase o categoría. La frecuencia absoluta se denota con f_i.

  Por ejemplo, $f_1 = 8$ indica que en el grupo de pacientes con asma había 8 hombres.

  Así mismo $f_2 = 12$ indica que en el grupo de pacientes con asma había 12 mujeres.

- La tercera columna es el porcentaje que se obtiene a partir de la frecuencia relativa (h_i), la cual se obtiene así: $h_i = f_i / n$

$$h_1 = \frac{f_1}{n} = \frac{8}{20} = 0{,}40 \qquad h_2 = \frac{f_2}{n} = \frac{12}{20} = 0{,}60$$

- Cuando se multiplica la frecuencia relativa por 100 se obtiene el porcentaje. Luego:

  La frecuencia relativa 1 es 0,40 (h_1=0,40) indica que, del total de pacientes con asma, el 40 % son hombres.

  De igual manera la frecuencia relativa 2 es 0,60 (h_2=0,60) se interpreta que el 60 % del total de pacientes con asma son de sexo femenino.

Una **tabla de contingencia** o **tabla cruzada** es la extensión de una tabla de frecuencias para el caso de dos variables categóricas. La tabla "**Tabla cruzada** del sexo del paciente según los antecedentes de familiares con asma" en la Tabla 9 es un ejemplo de una tabla

cruzada para las variables sexo (masculino/femenino) y antecedentes de familiares con asma (si/no) siendo ambas variables cualitativas.

Las tablas de contingencia siempre deben mostrar la frecuencia absoluta y su porcentaje. Por la naturaleza bivariada de estas tablas se pueden calcular tres tipos diferentes de porcentajes llamados *porcentaje fila*, *porcentaje columna* o *porcentaje total*.

Tabla 11. Tipos de porcentaje en las tablas de contingencia.

<table>
<tr><th>Tipo de porcentaje</th><th>Comentarios</th></tr>
<tr><td>Porcentaje fila
<table>
<tr><th rowspan="2">Sexo</th><th colspan="2">Antecedentes n (%)</th><th rowspan="2">Total</th></tr>
<tr><th>Sí</th><th>No</th></tr>
<tr><td>Masculino</td><td>7 (87,5)</td><td>1 (12,5)</td><td>8 (100)</td></tr>
<tr><td>Femenino</td><td>2 (16,7)</td><td>10 (83,3)</td><td>12 (100)</td></tr>
</table></td><td>

Cada fila suma el 100 %.
Para los hombres:

Sí antecedentes: 7/8×100 = 87,5 %.
No antecedentes: 1/8×100 = 12,5 %.

Para las mujeres:

Sí antecedentes: 2/12×100 = 16,7 %.
No antecedentes: 10/12×100 = 83,3 %.

Interpretación:

Entre los hombres el 87,5 % tenía antecedentes de familiares con asma.
Entre las mujeres el 16,7 % tenía antecedentes de familiares con asma.

</td></tr>
<tr><td>Porcentaje columna
<table>
<tr><th rowspan="2">Sexo</th><th colspan="2">Antecedentes n (%)</th></tr>
<tr><th>Sí</th><th>No</th></tr>
<tr><td>Masculino</td><td>7 (77,8)</td><td>1 (9,1)</td></tr>
<tr><td>Femenino</td><td>2 (22,2)</td><td>10 (90,9)</td></tr>
<tr><td>Total</td><td>9 (100)</td><td>11 (100)</td></tr>
</table></td><td>

Cada columna suma el 100 %.
Antecedentes de asma (Sí):

Masculino: 7/9×100 = 77,8 %.
Femenino: 2/9×100 = 22,2 %.

Antecedentes de asma (No):

Masculino: 1/11×100 = 9,1 %.
Femenino: 10/11×100 = 90,9 %.

Interpretación:

Entre las personas con antecedentes de familiares con asma el 77,8 % son hombres y el 22,2 son mujeres.
Entre las personas sin antecedentes de familiares con asma el 9,1 % son hombres y el 90,9 % son mujeres.

</td></tr>
</table>

Porcentaje total

Sexo	Antecedentes n (%)		Total
	Sí	No	
Masculino	7 (35)	1 (5)	8 (40)
Femenino	2 (10)	10 (50)	12 (60)
Total	9 (45)	11 (55)	20 (100)

- El total de la tabla suma el 100 %.
  - Hombres con antecedentes de familiares con asma: 7/20×100 = 35 %.
  - Hombres sin antecedentes de familiares con asma: 1/20×100 = 5 %.
  - Mujeres con antecedentes de familiares con asma: 2/20×100 = 10 %.
  - Mujeres sin antecedentes de familiares con asma: 10/20×100 = 50 %.
- Interpretación:
  - De la **población en estudio el 35 %** son hombres que si tienen familiares con antecedentes de asma.
  - De la **población en estudio el 5 %** son hombres que no tienen familiares con antecedentes de asma.
  - De la **población en estudio el 10 %** son mujeres que sí tienen familiares con antecedentes de asma.
  - De la **población en estudio el 50 %** son mujeres que no tienen familiares con antecedentes de asma.

Nota. Elaboración propia.

Presentación de datos con variables cuantitativas o numéricas

Si la variable en estudio es cuantitativa (por ejemplo; edad, peso o talla, entre otras), los datos pueden presentarse en una tabla de distribución de frecuencias, o utilizarse gráficas como el histograma o el diagrama de cajas (*boxplot*).

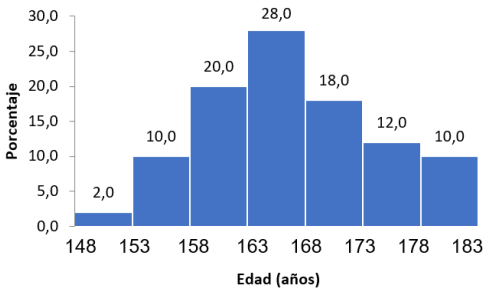
Tabla 12. Tabla de distribución de frecuencias.

Tabla X. Tabla de distribuciones de las estaturas de un grupo de 50 estudiantes (*n*=50).

k	Estatura (cm)	Número de estudiantes	%	Frecuencia absoluta acumulada	% acumulado
1	148 – 153	1	2	1	2
2	153 – 158	5	10	6	12
3	158 – 163	10	20	16	32
4	163 – 168	14	28	30	60
5	168 – 173	9	18	39	78
6	173 – 178	6	12	45	90
7	178 – 183	5	10	50	100
	Total	50	100		

Nota. Elaboración propia.

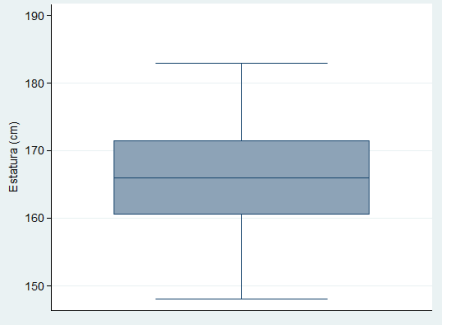
Histograma



Gráfica 5. **Histograma** de las estaturas de un grupo de 50 estudiantes (*n*=50).

Nota. Elaboración propia.

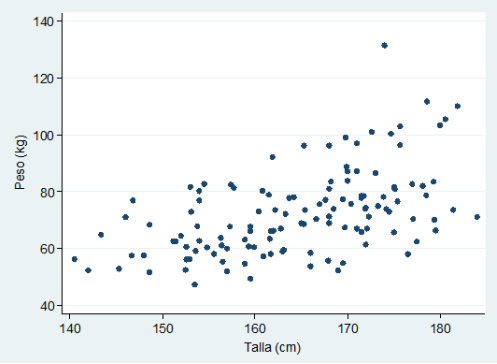
Diagrama de cajas



Gráfica 6. **Diagrama de cajas** de las estaturas de un grupo de 50 estudiantes (*n*=50).

Nota. Elaboración propia.

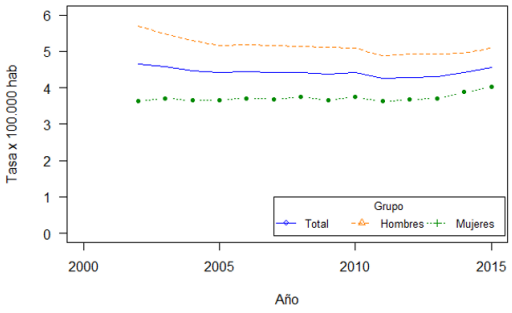
Gráfico de dispersión



Gráfica 7. **Gráfica de dispersión** de la estatura y el peso de un grupo de estudiantes (*n*=126).

Nota. Elaboración propia.

Gráfico de líneas



Gráfica 8. **Gráfico de líneas** mostrando la tasa de mortalidad en Colombia (total y por sexo), 2002-2015.

Nota. Elaboración propia.

Variables cuantitativas discretas:

Si la variable es discreta y los diferentes valores que toma la variable no son muchos (alrededor de seis), se utilizan tablas de frecuencias similares a las variables categóricas. Utilizando la variable *Número de hospitalizaciones* del Ejemplo 3, la tabla de distribución de frecuencias y gráfica son:

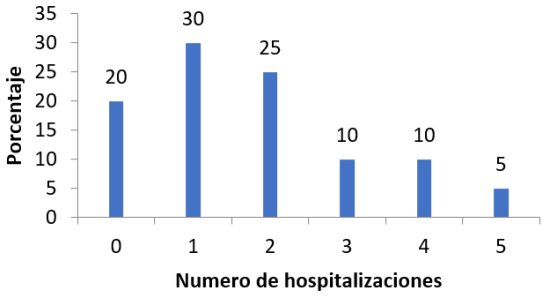
Tabla 13. Tabla de frecuencias.

Tabla de frecuencias	Gráfica de barras para variables discretas

Tabla X. Distribución del número de hospitalizaciones en el último año de los 20 pacientes con asma

Nº hospitalizaciones	N° de pacientes	%
0	4	20
1	6	30
2	5	25
3	2	10
4	2	10
5	1	5
Total	20	100

Nota. Elaboración propia.



Gráfica 9. **Diagrama barras** para el número de hospitalizaciones de los pacientes con asma en el último año (*n*=20). Elaboración propia.

Variables cuantitativas continuas

Con variables cuantitativas continuas, o discretas con muchos valores, se prefieren agrupar los datos en tablas de distribución de frecuencias.

Una distribución de frecuencias es la agrupación de los datos en clases (o intervalos) mutuamente excluyentes mostrando el número de observaciones en cada una.

A continuación, se explica el procedimiento a seguir con un ejemplo.

Ejemplo 4. Estatura de 60 estudiantes de un curso

Los siguientes datos corresponden a la estatura (en centímetros) de un grupo de 60 estudiantes (n=60).

154	162	172	166	178	160	179	165	166	156
179	166	166	176	160	155	169	173	169	173
164	168	148	183	165	167	150	174	164	161
182	160	173	161	178	171	172	166	169	170
161	172	162	166	160	164	170	159	167	160
153	168	165	155	151	168	162	175	165	156

Cuando se tienen cada uno de los datos, es decir los datos originales, se dice que son datos en bruto o crudos (*raw data*, por sus siglas en inglés). Sin embargo, estos datos en bruto no brindan mucha información. Es necesario hacer uso de la estadística descriptiva para organizarlos en tablas o gráficas.

A continuación, se describen los pasos a seguir para construir una la tabla de distribución de frecuencias. Aunque existen programas estadísticos que hacen estas tablas y gráficas, es útil entender su metodología de construcción.

El objetivo de esta actividad es utilizar los datos del Ejemplo 4 y llenar cada una de las 6 columnas de la Tabla 14.

Tabla 14. Tabla de distribuciones de las estaturas de un grupo de 60 estudiantes (*n*=60).

(1)	(2)	(3)	(4)	(5)	(6)
K	Estatura (cm)	Frecuencia absoluta f_i	Frecuencia relativa h_i	Frecuencia absoluta acumulada F_i	Frecuencia relativa acumulada H_i
	–				
	–				
	–				
	–				
	–				
	–				
	–				
	Total				

Nota. Elaboración propia.

Para mayor facilidad, se organizan los datos en orden ascendente, es decir de menor a mayor.

148	150	151	153	154	155	155	156	156	159
160	160	160	160	160	161	161	161	162	162
162	164	164	164	165	165	165	165	165	166
166	166	166	166	166	167	167	168	168	169
169	169	170	170	171	172	172	173	173	173
174	175	175	176	178	178	179	179	182	183

1. Identificar el dato mínimo: $X_{min} = 148$
2. Identificar el dato máximo: $X_{max} = 183$
3. Calcular el rango: *R*

   Rango = $X_{max} - X_{min} = 183 - 148 = 35$
4. Decidir el número de clases o número de intervalos: *k*

   No hay una norma general para elegir *k*. La elección de *k* puede ser mediante:

   - Regla empírica → *k* entre 6 y 15.
   - $k=\sqrt{n}$ Para el ejemplo $\sqrt{60} = 7,7$ luego *k* puede ser *k*=7 o *k*=8.
   - Formula de Sturges: $k = 1 + [3,322 \log_{10} n]$ Para el ejemplo $k = 6,9$; luego *k* puede ser 7 (*k*=7).

Se decide k= 7.

En la Tabla 14, se llenan los datos de la columna (1), con los números del 1 al 7.

5. Determinar el ancho de clase: w

$$w = \frac{\text{rango}}{k} = \frac{35}{7} = 5 \text{ cm}$$

6. Establecer los límites de cada clase. Se comienza con el dato mínimo (148) y se va sumando el ancho de clase definido (w=5). Los intervalos son:

K	Estatura (cm)
1	148 – 153
2	153 – 158
3	158 – 163
4	163 – 168
5	168 – 173
6	173 – 178
7	178 – 183

En la Tabla 14 se llenan los datos de la columna (2).

Hay que tener en cuenta que los intervalos son cerrados a la izquierda y abiertos a la derecha, es decir, el intervalo 1 que va de 148 a 153 es:

[148 – 153)

Significa que incluye el 148 y abarca hasta 152,9999…., es decir no incluye el 153.

7. La frecuencia absoluta (f_i), columna 3, se llena al contar el número de elementos de cada clase, es decir, contando cuántos de los datos en bruto caen en cada intervalo.

Por ejemplo, los primeros 3 valores (148, 150 y 151) pertenecen al intervalo 1: [148 – 153), luego f_1=3.

Los siguientes seis valores (153, 154, 155, 155, 156, 156) pertenecen al intervalo 2: [153 – 158), luego f_2=6.

Y así sucesivamente. En la Tabla 14 se llenan los datos de la columna (3).

8. La columna 4 hace referencia a la frecuencia relativa, calculada para cada intervalo como:

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

$h_1 = 3 / 60 = 0{,}050$

$h_2 = 6 / 60 = 0{,}100$

Y así sucesivamente. En la Tabla 14 se llenan los datos de la columna (4).

9. La siguiente columna (5) hace referencia a la frecuencia absoluta acumulada (F_i) y va acumulando las frecuencias absolutas (f_i) hasta cada intervalo.

$F_1 = 3$, ya que solo 3 datos son menores que 153.

$F_2 = 9$, ya que 9 datos son menores que 158 (3+6)

$F_3 = 21$, ya que 21 datos son menores que 163 (3+6+12)

$F_4 = 37$, ya que 37 datos son menores que 168 (3+6+12+16)

Y así sucesivamente.

En la Tabla 14 se llenan los datos de la columna (5).

10. La última columna (6) es la frecuencia relativa acumulada (H_i), y se obtiene de manera similar a la frecuencia absoluta acumulada de la columna 5, pero esta vez se utilizan las frecuencias relativas de la columna 4.

$H_1 = 0{,}050$, ya que el 5% de los datos son menores que 153.

$H_2 = 0{,}150$, ya que el 15% datos son menores que 158 (0,050 + 0,100)

$H_3 = 0{,}350$, ya que el 35% de los datos son menores que 163 (0,050 + 0,100 + 0,200)

$H_4 = 0{,}617$, ya que el 61,7% de los datos son menores que 168 (0,050 + 0,100 + 0,200 + 0,267)

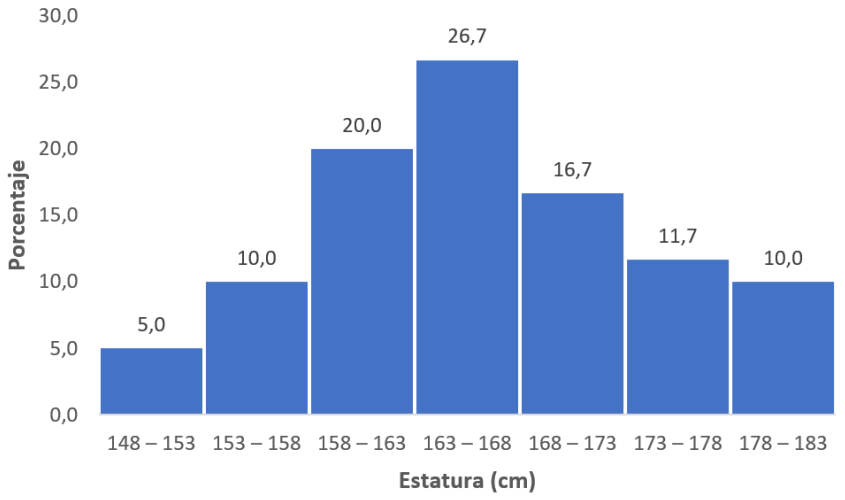
Y así sucesivamente. En la Tabla 14 se llenan los datos de la columna (6). Finalmente, la organización de los datos del Ejemplo 4 se presenta en la Tabla 15.

Tabla 15. Tabla de distribuciones de las estaturas de un grupo de 60 estudiantes (n=60).

(1)	(2)	(3)	(4)	(5)	(6)
K	Estatura (cm)	Frecuencia absoluta f_i	Frecuencia relativa h_i	Frecuencia absoluta acumulada F_i	Frecuencia relativa acumulada H_i
1	148 – 153	3	0,050	3	0,050
2	153 – 158	6	0,100	9	0,150
3	158 – 163	12	0,200	21	0,350
4	163 – 168	16	0,267	37	0,617
5	168 – 173	10	0,167	47	0,783
6	173 – 178	7	0,117	54	0,900
7	178 – 183	6	0,100	60	1,000
	Total	60	1		

Nota. Aunque en la tabla anterior se presentan las frecuencias relativas (columnas 4 y 6) se recomienda multiplicarlas por 100 y reportar el porcentaje. Elaboración propia.

El **histograma** es una gráfica en la que las clases se marcan en el eje horizontal y las frecuencias absolutas, relativas o el porcentaje en el eje vertical. Usando los datos de la Tabla 15 se presenta el histograma en porcentaje (Gráfica 10).



Gráfica 10. Histograma de las estaturas de un grupo de 60 estudiantes (*n*=60). Elaboración propia.

2.2.2 Ejercicios propuestos

Ejercicio 3.

Utilice los datos de la Tabla 15 para determinar e interpretar:

Cantidad	Interpretación
$f_4 =$	
$F_5 =$	
$H_4 =$	
$h_7 =$	
$H_5 =$	
$f_6 =$	
$F_4 =$	
$h_2 =$	

Ejercicio 4.

Con base en los datos de la Tabla 15 y/o el histograma de la Gráfica 10, responda las siguientes preguntas:

1. ¿Qué porcentaje de los estudiantes miden más de 1,68 metros?
2. ¿Qué porcentaje de los estudiantes miden menos de 1,60 metros?
3. ¿Qué porcentaje de los estudiantes miden entre 1,60 y 1,70 metros?
4. ¿Qué porcentaje de los estudiantes miden más de 1,70 metros?

Forma de la distribución de una variable numérica

Gráficos para las variables numéricas como el histograma o el diagrama de cajas (*boxplot*) permiten explorar visualmente la **forma de la distribución** de una variable numérica, la cual puede ser simétrica o asimétrica (Figura 9).

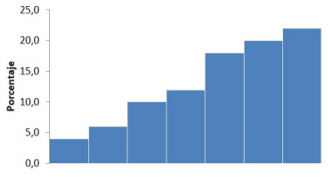
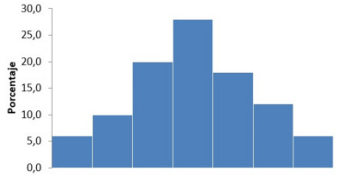
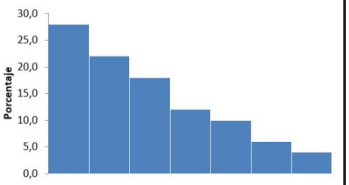
Sesgada a la izquierda	Simétrica	Sesgada a la derecha
		
Un conjunto de datos es sesgado a la izquierda cuando el histograma presenta una cola larga a la izquierda. Este es un fenómeno que se presenta cuando existen datos atípicos[14] muy pequeños.	Un conjunto de datos es simétrico cuando el histograma tiene forma de campana. En este tipo de distribuciones los datos tienden a un valor central, con datos más grandes y pequeños en igual proporción.	Una distribución es sesgada hacia la derecha cuando los datos se concentran en el lado izquierdo del histograma y existe una cola larga hacia la derecha, como resultado de unos pocos datos atípicos que son muy grandes.

Figura **9.** Forma de distribución de una variable numérica. Elaboración propia.

La forma de la distribución de una variable numérica juega un papel importante en el resumen de los datos numéricos. Cuando una variable numérica es simétrica el centro de la distribución (valor central o promedio) será un buen indicador para representar a todo el conjunto de datos. En caso contrario, cuando las distribuciones son asimétricas (sesgada a la derecha o a la izquierda) se debe buscar otro valor que realmente represente el centro de la distribución, en estos casos la mediana.

Para explorar la forma de la distribución de una variable numérica existen las siguientes alternativas:

- Visualmente mediante gráficos como el histograma o el diagrama de cajas (*boxplot*).
- Comparando los indicadores de centramiento (media, mediana y moda) que se verán a continuación.
- Calculando coeficientes de asimetría y curtosis.

14 Datos atípicos son datos muy diferentes con respecto al conjunto de datos, ya sea porque son muy pequeños o muy grandes

2.2.3 Indicadores resumen

La estadística descriptiva se encarga de organizar, presentar y resumir información almacenada en bases de datos para facilitar su interpretación. Previamente se mostraron alternativas visuales (tablas y gráficas) para presentar información de variables categóricas y numéricas.

En la bioestadística, para **resumir** las variables de una base de datos, se usan indicadores de acuerdo con su naturaleza (cualitativa o cuantitativa). La finalidad de estos indicadores es condensar la información de las variables para conseguir una mejor interpretación.

- Cuando las variables son cualitativas (categóricas) el resumen se hace mediante **porcentajes**.
- Cuando son variables cuantitativas (numéricas) hay diversos **indicadores** que resumen algunas características que nos interesan:

    La **forma** de la distribución.

    La **tendencia** central.

    La **dispersión** o variabilidad.

    La **posición** de algunos estadísticos de interés.

La forma de la distribución incluye dos elementos: 1) La simetría; pudiendo encontrar distribuciones simétricas o también distribuciones sesgadas, y 2) La curtosis que mide que tan puntiaguda es una distribución. La tendencia central se refiere al punto medio (central) de una distribución. La dispersión se refiere a la variabilidad o grado de dispersión de los datos en una distribución. Finalmente, la posición brinda información de algunos puntos de corte de la distribución de los datos.

Algunos gráficos propios de las variables numéricas, como el histograma o el diagrama de cajas, permiten la exploración visual de la forma de la distribución del conjunto de datos, su tendencia y variabilidad. Sin embargo, la estadística descriptiva también ofrece algunos indicadores para analizar estas características.

Indicadores de forma

La forma de la distribución de una variable numérica incluye dos características importantes que se pueden medir mediante los Coeficientes de Asimetría y Curtosis.

a. La simetría se refiere a si los datos se distribuyen alrededor de un valor central y la forma es similar a la izquierda y derecha; es decir, en espejo (Figura 10). El Coeficiente de Asimetría tiene la siguiente interpretación:

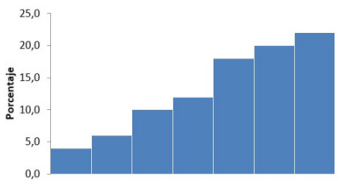
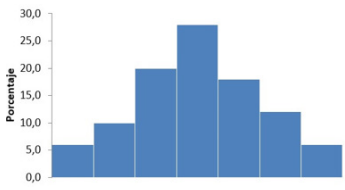
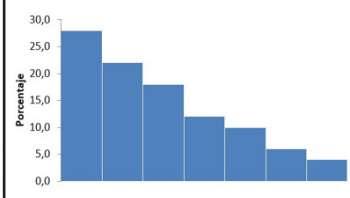
Distribución sesgada a la izquierda	Distribución simétrica	Distribución sesgada a la derecha
		
Coeficiente de Asimetría < 0	**Coeficiente de Asimetría = 0**	**Coeficiente de Asimetría > 0**
Una distribución sesgada a la izquierda tiene coeficiente de asimetría menor que cero.	Una distribución simétrica tiene coeficiente de asimetría de cero.	Una distribución sesgada a la izquierda tiene coeficiente de asimetría mayor que cero.

Figura **10.** Forma de la simetría en la distribución de una variable numérica. Elaboración propia.

b. La curtosis se refiere al grado de concentración de los datos numéricos alrededor de su valor central (Figura 11). El Coeficiente de Curtosis mide qué tan puntiaguda es una distribución.

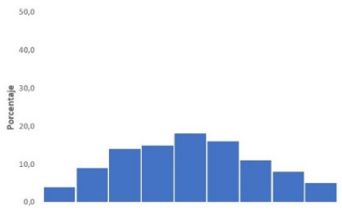
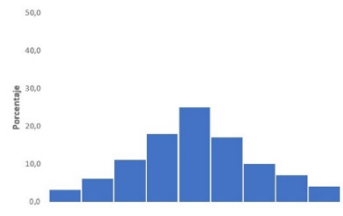
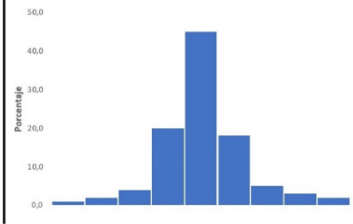
Distribución platicúrtica	Distribución mesocúrtica	Distribución leptocúrtica
		
Coeficiente de Curtosis < 0	**Coeficiente de Curtosis = 0**	**Coeficiente de Curtosis > 0**
Una distribución platicúrtica presenta una baja concentración de los datos alrededor del valor central.	Una distribución mesocúrtica tiene concentración media alrededor del valor central.	Una distribución leptocúrtica presenta una alta concentración de datos alrededor del valor central.

Figura 11. Forma de la curtosis en la distribución de una variable numérica. Elaboración propia.

En esta parte del manual se abordará principalmente el cálculo de algunos indicadores: medidas de tendencia central, dispersión y posición (Tabla 16).

Tabla 16. Resumen de indicadores más importantes de tendencia central, dispersión y posición.

Medidas de tendencia central	Medidas de dispersión	Medidas de posición
Promedio Mediana Moda	Rango Rango intercuartil Desviación estándar y varianza Coeficiente de variación	Cuartiles Deciles Percentiles

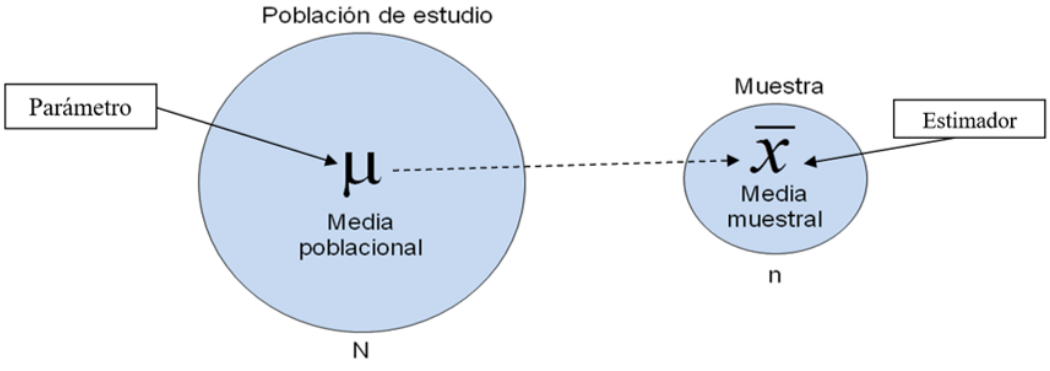
Nota. Elaboración propia

Medidas de tendencia central

Promedio, media aritmética o media

Las medidas de tendencia central (llamadas también medidas o indicadores de centramiento) tienen como objetivo "resumir" o señalar el centro de un conjunto de datos. Las medidas de tendencia central que se abordan en este manual son la *media*, la *mediana* y la *moda*.

Los estudiantes están familiarizados con el concepto de **promedio**, conocido también como **media aritmética** o simplemente "**media**". La media se puede calcular con los datos de una población (μ como un parámetro)[15] o de una muestra (como un estimador)[16].



15 Recuerde que un parámetro es una característica medible en la población, luego μ es un parámetro.
16 Un estimador se calcula con los datos de una muestra, luego $\bar{x}$ es un estimador.

Media poblacional	Media muestral
$\mu = \frac{\sum x_i}{N}$	$\bar{x} = \frac{\sum x_i}{n}$
Para obtener la media poblacional (μ) se suman todos los datos y se divide sobre el tamaño de la población (N).	Para obtener la media muestral ($\bar{x}$) se suman todos los datos y se divide sobre el tamaño de la muestra (n).

La media es la medida de tendencia central más conocida y utilizada, y su papel es mostrar el valor central de la información o del conjunto de datos. Esta medida es utilizada cuando el conjunto de datos es simétrico. Sin embargo, cuando existen datos atípicos en el conjunto de datos, la distribución es sesgada y entonces la media aritmética no es representativa.

Ejemplo 5. Peso de bebés recién nacidos en el hospital A - Promedio

Un día cualquiera en un hospital nacieron 10 bebés (n=10). Los pesos de los recién nacidos (en gramos) fueron los siguientes:

3 250	3 350	3 700	3 150	3 750	3 400	3 600	3 500	3 850	3 450

El peso promedio de los bebes nacidos en el hospital A es 3 500 gramos.

$$\bar{x} = \frac{\sum x_i}{n} = \frac{3250 + 3350 + 3700 + \cdots . + 3850 + 3450}{10} = \frac{35000}{10} = 3500$$

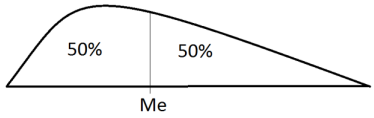
Ejercicio 5

Suponga que los siguientes datos corresponden a los pesos (gramos) de 9 bebés nacidos en un hospital B. ¿Cuál es el peso promedio de los recién nacidos en el hospital B?

2 200	2 600	2 800	2 950	3 000	3 050	3 100	3 700	3 750

Mediana

Cuando se tienen distribuciones sesgadas (a la derecha o a la izquierda) el centro de la información se describe mejor con una medida de ubicación llamada la mediana. La ***mediana*** divide un conjunto de datos en dos partes iguales, por tanto, se encuentra en la posición central del conjunto de datos. El 50 % de los datos son menores o iguales que la mediana y el otro 50 % de los datos son mayores o iguales a ella.



Para calcular la mediana con datos en bruto se requiere que el conjunto de datos se encuentre ordenado de manera ascendente; es decir, en orden de menor a mayor. Cuando el número de datos es impar, la mediana es el dato de la posición central (Ver Ejemplo 6). Cuando el número de datos es par, la mediana es el promedio de los dos valores centrales (Ver Ejemplo 7).

Ejemplo 6. Peso de bebés recién nacidos en el hospital B - Mediana

Suponga que los siguientes datos corresponden a los pesos (gramos) de 9 bebés nacidos en un hospital (n=9). Los datos se encuentran ordenados de menor a mayor. La mediana del peso de los bebés es 3 000 gramos.

2 200	2 600	2 800	2 950	3 000	3 050	3 100	3 700	3 750
				↓				
				Mediana				

Como se tienen 9 datos (n impar), solo hay un dato que se encuentra en la posición central. La mediana obtenida nos indica que la mitad de los bebés pesaron 3 000 gramos o menos y la otra mitad pesaron más de 3 000 gramos.

Ejemplo 7. Peso de bebés recién nacidos en el hospital A - Mediana

Usando los datos del Ejemplo 5. Primero se deben ordenar los datos de menor a mayor.

3 150	3 250	3 350	3 400	3 450	3 500	3 600	3 700	3 750	3 850
				↓	↓				

En este ejemplo se tienen 10 datos (n par). Cuando el número de datos es par, la mediana es el promedio de los 2 valores centrales. En este caso la mediana es 3 475 gramos.

$$Me = \frac{3\,450 + 3\,500}{2} = 3\,475$$

Moda

La **moda** es el dato que más se repite o que aparece con mayor frecuencia. Una distribución puede tener una moda, varias modas o no tener moda.

Ejercicio 6

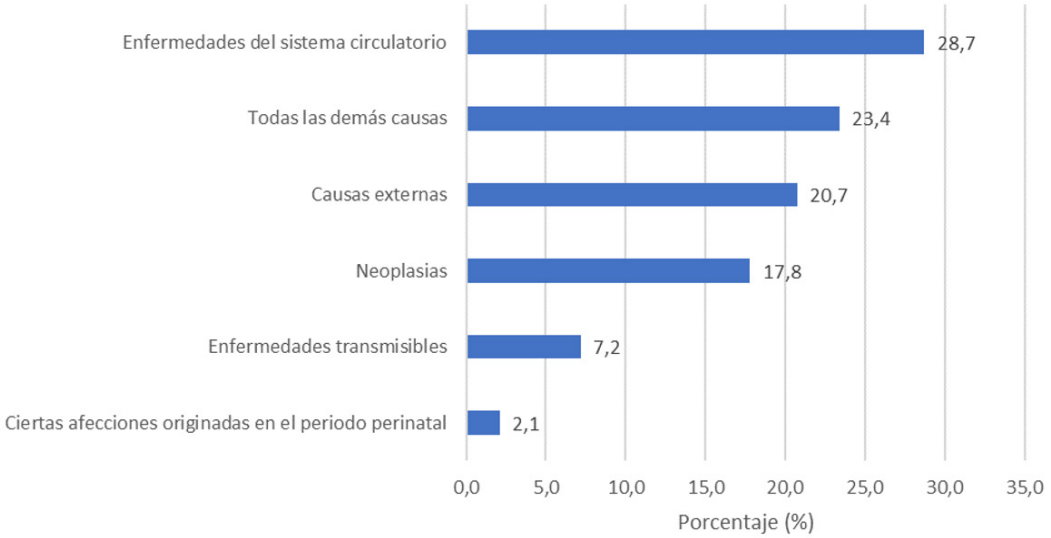
Para cada uno de los siguientes conjuntos de datos calcule la moda:

1. {150 200 250 225 200 155 160}
2. {150 200 250 225 200 155 160 250}
3. {150 200 250 225 180 155 160}
4. {150 200 150 225 200 225 160 225}

La moda es el único indicador de centramiento que puede ser utilizado con variables categóricas. En estos casos la moda será la categoría de respuesta que obtuvo mayor frecuencia.

Ejercicio 7

¿Cuál es la moda de la causa de muerte en la región pacífico de Colombia durante el periodo 2002-2015 de acuerdo con la información de la Gráfica 11? y ¿Cuál fue su porcentaje?



Gráfica 11. Grupos de causas de muerte en la región pacífico de Colombia, 2002-2015. Elaboración propia.

Medidas de dispersión o variabilidad:

Las medidas de tendencia central (media, mediana, moda) se utilizan para señalar la ubicación central del conjunto de datos. Sin embargo, si se consideran solo las medidas de ubicación o si se comparan varios conjuntos de datos utilizando valores centrales, es probable que se llegue a una conclusión errónea.

Además de las medidas de ubicación central es importante considerar la dispersión o variabilidad en los datos, y para evaluarlas se utilizarán los siguientes indicadores: rango, rango intercuartil, desviación estándar y coeficiente de variación.

Rango

El **rango** (R) hace referencia a la distancia entre el valor más grande y el valor más pequeño del conjunto de datos. Para reportarlo suelen presentarse los valores mínimo y máximo de los datos.

$$R = X_{max} - X_{min}$$

El rango es la medida de dispersión más sencilla de calcular, pero un defecto del rango es que se basa solo en dos valores; el dato mínimo y el máximo, y no utiliza todo el conjunto de datos. Luego, con la presencia de un solo dato atípico el rango puede llegar a ser muy grande.

Ejemplo 8. Peso de bebés recién nacidos en los hospitales A y B - Rango

A continuación, los datos de pesos de los recién nacidos en los dos hospitales (A y B). Para el hospital A se tiene que el rango es de 700 gramos. ¿Cuál es el rango del peso de los bebés para el hospital B?

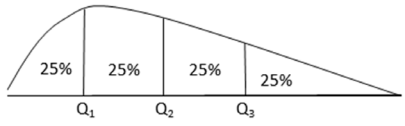
Indicador	Hospital A	Hospital B
Dato máximo (X_{max})	3850	
Dato minino (X_{min})	3150	
Rango	700	

¿En cuál de los dos hospitales hay mayor variación en el peso de los bebés?

El rango es una de las medidas de dispersión que suele acompañar a la mediana. La otra medida de dispersión que suele acompañar la mediana es el rango intercuartil.

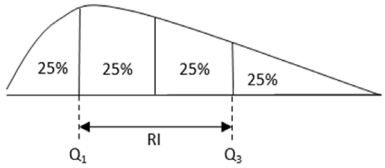
Rango intercuartil

El **rango intercuartil** (RI) indica la amplitud del 50 % de los datos alrededor de la mediana, y para obtenerlo es necesario introducir el concepto de cuartil. Los cuartiles son medidas de posición que dividen un conjunto de datos en cuatro partes iguales:



- El cuartil 1 (Q_1) es el valor del conjunto de datos que es mayor que el 25 % de los datos. Se puede decir que el 25 % de los datos son menores que Q_1 y por lo tanto el 75 % son mayores que Q_1.
- El cuartil 2 (Q_2) es el valor del conjunto de datos que es mayor que el 50 % de los datos. El cuartil 2 coincide con la mediana, luego el 50 % de los datos son menores que Q_2 y el otro 50 % son menores que Q_2.
- El cuartil 3 (Q_3) es el valor del conjunto de datos que es mayor que el 75 % de los datos. Luego el 25 % de los datos son mayores que Q_3.

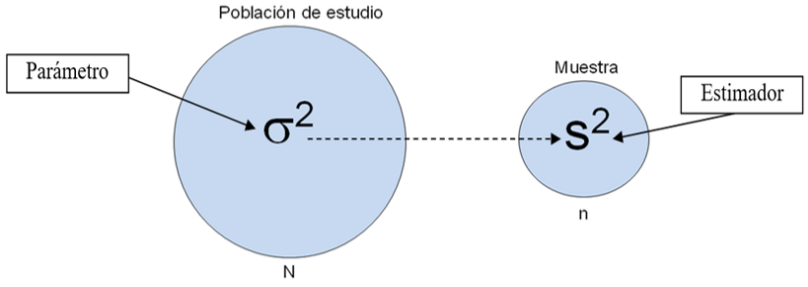
El rango intercuartil es la distancia entre el primer y el tercer cuartil ($RI = Q_3 - Q_1$)



Este indicador será retomado **más adelante** con las medidas de posición.

Varianza y desviación estándar

La **varianza** se puede medir con los datos de la población o de una muestra. En el primer caso se habla de un parámetro y en el segundo de un estimador. La varianza poblacional se simboliza con σ^2 (sigma cuadrado) y la varianza muestral con S^2 ("ese" al cuadrado). La raíz cuadrada de la varianza se llama desviación estándar. Por lo tanto, la desviación estándar poblacional es σ y la desviación estándar muestral es S.



Varianza poblacional	Varianza muestral
$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{N}$	$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$
Desviación estándar poblacional	Desviación estándar muestral
$\sigma = \sqrt{\sigma^2}$	$s = \sqrt{s^2}$

La varianza no es muy utilizada ya que trabaja con unidades al cuadrado. La **desviación estándar** es la medida de dispersión que acompaña la media. Se interpreta como la dispersión promedio de los datos con respecto a la media. Suponga que se tienen 5 datos representados por X_1, X_2, X_3, X_4 y X_5 como se muestra en la Figura 12. Las flechas representan la distancia o desviación de cada uno de los datos con respecto al promedio. Si se obtiene un promedio de los valores asociados a esas flechas se habrá calculado la desviación estándar.

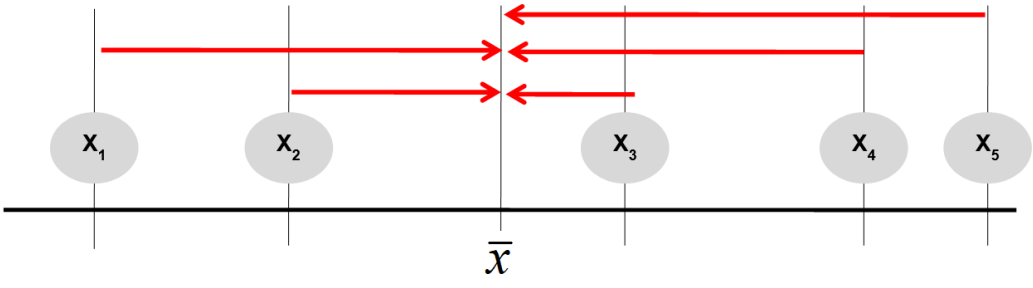


Figura **12.** Esquema para representar la desviación estándar. Elaboración propia.

Ejemplo 9. Peso de bebés recién nacidos en el hospital A - Desviación estándar

A seguir, la desviación estándar de los pesos de los recién nacidos utilizando los datos del Ejemplo 5. Los datos del peso (en gramos) son:

3 250	3 350	3 700	3 150	3 750	3 400	3 600	3 500	3 850	3 450

Primero se obtiene la varianza mediante la fórmula: $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$

Se debe recordar que el peso promedio de los pesos de los bebés fue de 3 500 gramos.

El procedimiento para obtener la varianza se presenta en la Tabla 17. La columna uno (1) muestra cada uno de los datos. La columna dos (2) resta el promedio a cada uno de los datos, y la columna tres (3) muestra este resultado. La columna cuatro (4) eleva la columna tres al cuadrado.

Tabla 17. Cálculo de la desviación estándar.

(1)	(2)	(3)	(4)
X_i	$X_i - \bar{x}$	$X_i - \bar{x}$	$(X_i - \bar{x})^2$
3 250	3 250 – 3 500 = -250	-250	62 500
3 350	3 350 – 3 500 = -150	-150	22 500
3 700	3 700 – 3 500 = 200	200	40 000
3 150	3 150 – 3 500 = -350	-350	122 500
3 750	3 750 – 3 500 = 250	250	62 500
3 400	3 400 – 3 500 = -100	-100	10 000
3 600	3 600 – 3 500 = 100	100	10 000
3 500	3 500 – 3 500 = 0	0	0
3 850	3 850 – 3 500 = 350	350	122 500
3 450	3 450 – 3 500 = -50	-50	2 500
Total	**0**	**0**	**455 000**

Nota. Elaboración propia.

Luego, la varianza es:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{455.000}{10-1} = 50.555{,}6\ gramos^2$$

La desviación estándar es $s = \sqrt{s^2} = \sqrt{50.555{,}6} = 224{,}8\ gramos$

Significa que, en promedio, los pesos de los bebés varían 225 gramos con respecto a su media.

Ejercicio 8

Obtenga la desviación estándar de los pesos de los recién nacidos utilizando los datos del hospital B:

2 200	2 600	2 800	2 950	3 000	3 050	3 100	3 700	3 750

Realice la operación en la siguiente tabla:

X_i	$X_i - \bar{x}$	$(X_i - \bar{x})^2$
Total		

Coeficiente de variación

El **coeficiente de variación** (CV) es una medida de dispersión relativa. Reporta la variación en relación con la media. Lo anterior permite comparar distribuciones con diferente promedio y dispersión, o también conjuntos de datos con diferentes unidades de medición.

$$CV = \frac{s}{\bar{x}} \times 100$$

Cuando el promedio es positivo, el CV varía desde cero hasta infinito:

$$0 \leq CV < \infty$$

Entre más cercano a cero, los datos son más homogéneos, y entre más alejado de cero los datos son más heterogéneos. En ciencias de la salud se considera que un conjunto de datos es homogéneo cuando el CV es menor al 20 %.

Ejemplo 10. Peso de bebes recién nacidos en el hospital A – Coeficiente de variación

Utilizando los datos del hospital A, se concluye que los pesos de los bebes son muy homogéneos ya que el CV es del 6,4 %.

$$CV = \frac{s}{\bar{x}} \times 100 = \frac{224{,}6}{3500} \times 100 = 6{,}4\%$$

Ejercicio 9

Calcule e interprete el coeficiente de variación de los pesos de los recién nacidos utilizando los datos del hospital B:

2 200	2 600	2 800	2 950	3 000	3 050	3 100	3 700	3 750

Medidas de posición

Además del análisis descriptivo de un conjunto de datos por medio de indicadores de centramiento y dispersión, también hay otras formas de describir la variación o extensión en un conjunto de datos. Por ejemplo, determinando la ubicación de los valores que dividen un grupo de observaciones en partes iguales. Nos encontramos entonces con las **medidas de posición** (cuartil, decil y percentil) (ver Figura 13).

- Los cuartiles (Q_1, Q_2 y Q_3) dividen un grupo de observaciones en 4 partes iguales.
- Los deciles (D_1 a D_9) dividen los datos en 10 partes iguales.
- Los percentiles (P_1 a P_{99}) dividen el conjunto de datos en 100 partes iguales.

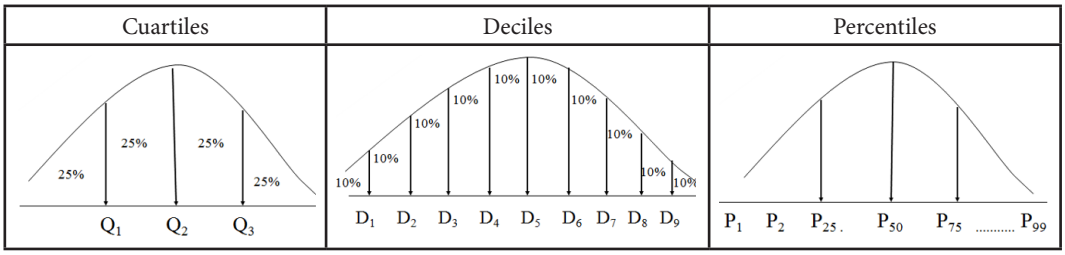


Figura **13.** Representación de las medidas de posición: cuartiles, deciles y percentiles. Elaboración propia.

Con los datos en bruto, la obtención de cuartiles, deciles y percentiles se hacen con la fórmula de los percentiles. Esta fórmula indica la posición o localización (L) del estadístico de interés, una vez los datos están organizados de manera ascendente:

$$L_P = \frac{(n+1)P}{100}$$

Tenga en cuenta que esta fórmula se utiliza para obtener la posición de cuartiles, deciles y percentiles. La letra "P" es el número del percentil de interés. Por ejemplo, el cuartil 1 se calcula con el percentil 25 (P=25), el cuartil 2 con el percentil 50 (P=50) y el cuartil 3 con el percentil 75 (P=75). Así mismo el decil 1 se calcula con el percentil 10 (P=10), el decil 2 con el percentil 20 (P=20), etc.

Ejemplo 11. Calificaciones de un examen para una beca

Considere los siguientes datos correspondientes a las calificaciones de un examen de 27 estudiantes (*n*=27) para la obtención de una beca:

79 78 78 67 76 87 85 73 66 99 84 72 66 57

94 84 72 63 51 48 50 61 71 82 93 100 89

Calculemos e interpretemos los cuartiles de la distribución.

Cuartil 1: Q_1

El primer paso es organizar los datos en orden ascendente:

48 50 51 57 61 63 66 66 67 71 72 72 73 76

78 78 79 82 84 84 85 87 89 93 94 99 100

El cuartil 1 (Q_1) es equivalente al percentil 25 (P_{25}). La localización del percentil 25 (P=25) se obtiene así:

$$L_p = \frac{(n+1)P}{100} = L_{25} = \frac{(27+1)25}{100} = 7$$

Significa que el cuartil 1 es el dato que se encuentra en la posición 7:

48 50 51 57 61 63 **66** 66 67 71 72 72 73 76

78 78 79 82 84 84 85 87 89 93 94 99 100

El percentil 25 o cuartil 1 es 66. Significa que el 25 % de las calificaciones fueron menores o iguales a 66 puntos.

Cuartil 2: Q_2

El cuartil 2 (Q_2) es el percentil 50 (P_{50}) (o la misma mediana). En la formula (P=50).

$$L_p = \frac{(n+1)P}{100} = L_{50} = \frac{(27+1)50}{100} = 14$$

Significa que el cuartil 2 es el dato que se encuentra en la posición 14:

48 50 51 57 61 63 66 66 67 71 72 72 73 **76**

78 78 79 82 84 84 85 87 89 93 94 99 100

El percentil 50 o cuartil 2 es 76. Entonces el 50% de los aspirantes obtuvieron 76 puntos o menos, y el otro 50% 76 puntos o más.

Cuartil 3: Q_3

El cuartil 3 (Q_3) es el percentil 75 (P_{75}) (P=75).

$$L_p = \frac{(n+1)P}{100} = L_{75} = \frac{(27+1)75}{100} = 21$$

Significa que el cuartil 3 es el dato que se encuentra en la posición 21:

48 50 51 57 61 63 66 66 67 71 72 72 73 76

78 78 79 82 84 84 **85** 87 89 93 94 99 100

El percentil 75 o cuartil 3 es 85. Es decir que el 75 % de los aspirantes obtuvieron una calificación menor o igual a 85 puntos. Por lo tanto, un 25 % obtuvo más de 85 puntos.

En el cálculo de la localización (L) se pueden presentar dos situaciones:

a. L_p resulta en un valor entero, en cuyo caso el percentil "P" se encuentra ubicado en una posición exacta (L_p). Por ejemplo, al calcular el cuartil 3 del ejemplo anterior, L_p dio un valor entero (posición L_p =21).

b. L_p tiene una parte decimal que designada "d". En ese caso el percentil "P" se ubica después de la posición L_p (se podría decir entre los datos de las posiciones x_i y x_{i+1}).

$$P = x_i + d\,(x_{i+1} - x_i)$$

Si en el ejemplo anterior el tamaño de la muestra hubiese sido n=28, la localización del cuartil 3 hubiese sido:

$$L_{75} = \frac{(28+1)75}{100} = 21{,}75$$

En este caso la parte decimal es *d=0,75*, y el cuartil tres se calcula así:

El dato de la posición 21 es: x_i=85

El dato de la posición 22 es x_{i+1}= 87

El cuartil tres se puede aproximar así:

$$Q_3 = 85 + 0{,}75\ (87 - 85)$$
$$Q_3 = 85 + 0{,}75(2)$$
$$Q_3 = 85 + 1{,}5$$
$$Q_3 = 86{,}5$$

Diagrama de cajas

El **diagrama de cajas** o boxplot también es conocido como gráfica con valores extremos, *Box and Whisker*, o gráfica de cajas y bigotes. El *boxplot* es una herramienta visual muy útil para comunicar información contenida en variables numéricas, por ejemplo tendencias, dispersión, asimetría y datos atípicos de la distribución.

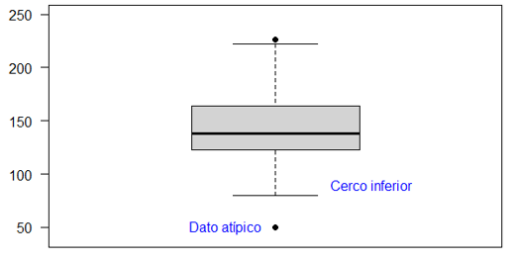
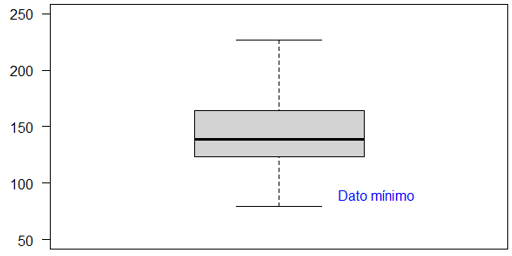
Un diagrama de caja es la representación gráfica, basada en los cuartiles. Para construir un diagrama de cajas se necesitan 5 estadísticos:

1. $X_{mín}$ o Cerco inferior= Q_1-[1.5*RI]
2. Primer cuartil: Q_1
3. Mediana o segundo cuartil: Q_2
4. Tercer cuartil: Q_3
5. X_{max} o Cerco superior= Q_3+[1.5*RI]

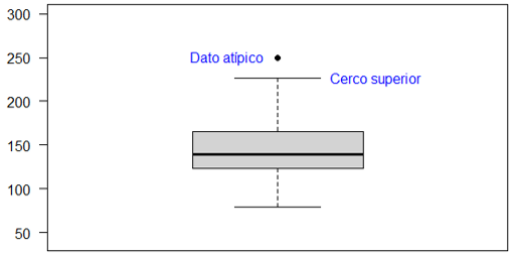
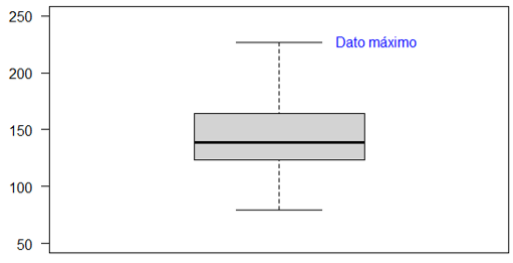
Valores inferiores al cerco inferior o mayores que el cerco superior se consideran **datos atípicos**. En otras palabras:

- Datos atípicos (muy pequeños) son aquellos menores al cerco inferior: Q_1-[1.5*RI]
- Datos atípicos (muy grandes) son aquellos mayores que el cerco superior: Q_3+[1.5*RI]

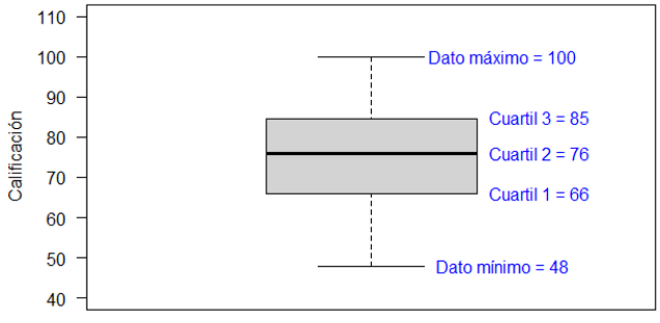
Parte inferior del gráfico:

Si hay datos atípicos (datos muy pequeños) el diagrama comienza en el cerco inferior. Los datos atípicos se grafican como puntos.	Si no hay datos atípicos (muy pequeños) el gráfico comienza en el dato mínimo.
	

Parte superior del gráfico:

Si hay datos atípicos (muy grandes) el diagrama finaliza en el cerco superior. Los datos atípicos se grafican como puntos.	Si no hay datos atípicos (muy grandes) el gráfico finaliza en el dato máximo.
	

El diagrama de cajas con los datos del Ejemplo 11 se muestra en la Gráfica 12.



Gráfica 12. Diagrama de cajas de las calificaciones de un grupo de 27 estudiantes. Elaboración propia.

Observe que la Gráfica 12 muestra una distribución muy simétrica para los datos.

3. Estadística inferencial

En el módulo 1 (Figura 6) se mencionan dos ramas de la Estadística: descriptiva e inferencial. Se dijo que la **estadística inferencial** abarcaba un conjunto de métodos utilizados para descubrir características de una población a partir de una muestra tomada de ella, y que sus principales herramientas son:

1. La estimación de parámetros por medio de intervalos de confianza y pruebas de hipótesis.

2. Las técnicas de regresión y correlación.

En este módulo se abordarán los intervalos de confianza y las pruebas de hipótesis para estimar parámetros de una población a partir de los estimadores de una muestra.

Como se muestra en la Figura 14, en la Estadística Inferencial, de la población de estudio se selecciona una "buena muestra" (❶), se estudia adecuadamente la muestra y sus características, para posteriormente generalizar los resultados de la muestra hacia la población (❷). En ese proceso de "generalización" hacia la población de estudio, la "probabilidad" juega un papel muy importante, ya que es la base de la Estadística Inferencial. Sin embargo, serán retomados los conceptos más importantes.

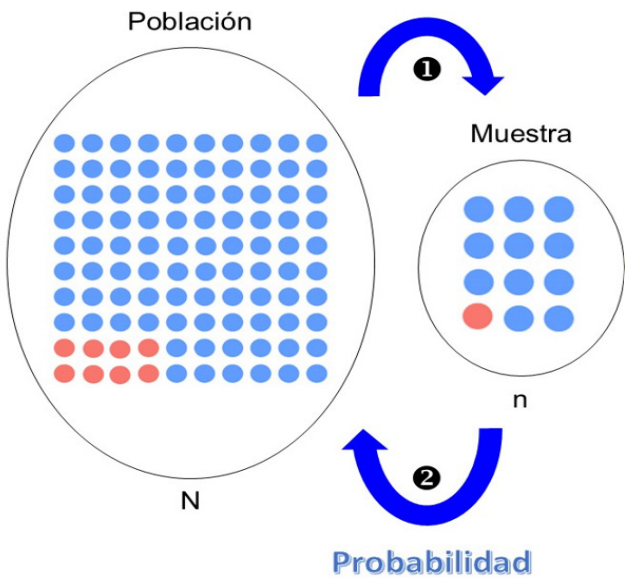


Figura 14. Esquema de la inferencia estadística representando la selección de una muestra (❶) y su posterior generalización (❷) con base a la probabilidad. Fuente: elaboración propia.

3.1 Fundamentos de probabilidad

La importancia del estudio de la probabilidad es que la Inferencia Estadística sienta sus bases en muchos conceptos de la Teoría de la Probabilidad.

3.1.1 Conceptos de probabilidad

La **probabilidad** hace parte del diario vivir en la práctica de un profesional de la salud donde está presente la incertidumbre. Un profesional de la salud puede observar la frecuencia con la que suceden ciertos eventos (1). Por ejemplo:

1. En un hospital se observa que aproximadamente 8 de cada 10 cirugías del corazón son exitosas. En términos de probabilidad significa que un paciente tiene una probabilidad de 0.80 o del 80 % de tener una operación exitosa.
2. Un médico le dirá a su paciente, después de examinarlo y comprobar sus síntomas, que está 95 % seguro de que padece una enfermedad específica.

3. Una enfermera dice que 9 de cada 10 pacientes suspenderán su cita.
4. Adicionalmente la lectura científica sobre investigación en salud a menudo reporta un concepto estadístico muy utilizado, el valor P, que corresponde a una probabilidad.

La probabilidad es una cantidad numérica (entre 0 y 1, inclusive) asociada con la posibilidad de que suceda un evento.

La probabilidad de un evento ***A*** es simbolizada con ***P(A)***, y por tanto: ***0*** ≤ ***P(A)*** ≤ ***1***. La probabilidad cuantifica la incertidumbre.

- Un evento que nunca sucederá (o que es imposible) tiene una probabilidad nula, ***P(A) = 0***.
- Un evento certero (o seguro) tiene la probabilidad máxima, ***P(A) = 1***.
- La probabilidad del evento ***A*** (obtener "cara" cuando lanzamos una moneda (no cargada o legal) es ***P(A) = ½ = 0,5.***

Es importante mencionar dos requerimientos básicos de una probabilidad:

1. Cualquier probabilidad siempre debe estar entre 0 y 1:

$$0 \leq P(A_i) \leq 1$$

2. Si un conjunto de eventos es colectivamente exhaustivo y los eventos son mutuamente excluyentes, la suma de las probabilidades siempre debe ser igual a 1:

$$\sum_{i=1}^{n} P(A_i) = 1$$

¿Cómo se calcula una probabilidad?

En términos generales hay 2 enfoques para asignar probabilidades a los eventos simples[17] . Sin embargo, ambos enfoques son muy similares:

Enfoque clásico

- El enfoque clásico parte del supuesto de que todos los resultados de un experimento son igualmente probables.
- Este enfoque proviene de juegos de azar.

Enfoque empírico

- El enfoque empírico determina la probabilidad de un evento al observar la frecuencia de ocurrencia del evento en el pasado.
- Este enfoque se usa analizando la frecuencia de datos históricos.

Las probabilidades se calculan así:

Enfoque clásico

$$P(A) = \frac{\#\ de\ resultados\ favorables}{\#\ total\ de\ posibles\ resultados}$$

Enfoque empírico

$$P(A) = \frac{\#\ de\ veces\ que\ el\ evento\ ocurre\ en\ el\ pasado}{\#\ total\ de\ observaciones}$$

17 Un evento simple involucra un solo evento (Por ejemplo, el evento A), mientras que los eventos compuestos involucran dos o más eventos (Por ejemplo, los eventos A y B)

Ejemplo 12. Cálculo de probabilidades bajo el enfoque clásico

La mejor forma de entender este enfoque es usando un ejemplo proveniente de los juegos de azar. Si se lanza un dado al aire, ¿Cuál es la probabilidad de observar un "numero par"?

Solución:

Definiendo el evento de interés "A: Observar un número par", la probabilidad se calcula así:

$$P(A) = \frac{\#\ de\ resultados\ favorables}{\#\ total\ de\ posibles\ resulados}$$

Donde:

El total de posibles resultados está dado por el conjunto:

$$S = \{ 1, 2, 3, 4, 5, 6 \}$$

Luego, existen seis (6) diferentes resultados.

El número de resultados que son favorables para el evento A son tres:

$$A = \{ 2, 4, 6 \}$$

Luego, si se lanza un dado al aire la probabilidad de observar un "número par" es de 0,5 o 50 %.

$$P(A) = \frac{\#\ de\ resultados\ favorables}{\#\ total\ de\ posibles\ resulados} = \frac{3}{6} = 0{,}5$$

Ejemplo 13. Cálculo de probabilidades bajo el enfoque empírico

El administrador de una entidad de salud desea estimar la probabilidad que una mujer embarazada requiera una cesárea al momento del parto. Para ello solicita el reporte de los nacimientos en el último año encontrando la siguiente información:

Tipo de parto	Frecuencia absoluta
Espontáneo	246
Por cesárea	200
Otros	4
Total	**450**

Solución:

Definiendo el evento de interés "A: Parto por cesárea", la probabilidad se calcula así:

$$P(A) = \frac{\#\ de\ veces\ que\ el\ evento\ ocurre\ en\ el\ pasado}{\#\ total\ de\ observaciones}$$

Usando la información disponible del último año:

El número total de observaciones fueron 450.

El número de veces que ocurrió el "Parto por cesárea" en el pasado fue 200.

Luego, la probabilidad estimada que una mujer embarazada requiera una cesárea al momento del parto es de 0,444 o del 44,4 %.

$$P(A) = \frac{\#\ de\ veces\ que\ el\ evento\ ocurre\ en\ el\ pasado}{\#\ total\ de\ observaciones} = \frac{200}{450} = 0{,}444$$

Importante: El estudiante debe reconocer que el enfoque empírico para calcular probabilidades usa el mismo concepto de la "frecuencia relativa" estudiado en el módulo de estadística descriptiva.

3.1.2 Variables aleatorias

En la Estadística Descriptiva se mencionó el concepto de "**variables**" que son aquellas características o atributos que se pueden medir en las unidades de estudio. Por ejemplo, si la unidad de estudio es un "paciente", en este se pueden medir algunas variables cualitativas (sexo, estado civil, resultado de una prueba diagnóstica, entre otras), variables cuantitativas discretas (número de hijos, número de lesiones, número de consultas el último año, etc.) o variables cuantitativas continuas (peso, talla, nivel de glucosa, etc.).

Ahora una "**variable aleatoria**" (VA), simbolizada con letras mayúsculas como X, Y o Z, es una cantidad numérica cuyo valor es el resultado de un experimento o evento aleatorio, luego, su valor es aleatorio.

Las variables aleatorias pueden ser discretas o continuas. Una VA discreta es aquella que puede tomar un conjunto finito de valores; es decir, toma números o valores enteros. Una VA continua puede tomar infinitos valores dentro de un rango; es decir, toma valores continuos.

Veamos algunos ejemplos de variables aleatorias:

Tabla 18. Ejemplos tipos de variables.

VA discretas	VA continuas
Y: Número de episodios de otitis en el primer año de vida. ***Y*** es una VA discreta que puede tomar valores {0, 1, 2, 3,….}	***X:*** Nivel de exposición a radiación anual (rem[18]). ***X*** es una VA continua que puede tomar valores mayores a cero.
Z: Número de bebes de sexo masculino en los primeros cinco partos del día. ***Z*** es una VA discreta que puede tomar valores {0, 1, 2, 3, 4, 5}	***Y:*** Estatura (cm) de un bebe recién nacido. ***Y*** es una VA continua que puede tomar valores mayores a cero.
X: Número de caras al lanzar una moneda legal dos veces. ***X*** es una VA discreta que puede tomar valores {0, 1, 2}	***Z:*** Tiempo (minutos) de recuperación de la anestesia. ***Z*** es una VA continua que puede tomar valores mayores a cero.

Nota. Elaboración propia.

3.1.3 Distribuciones de probabilidad

La Teoría de la Probabilidad ha estudiado el comportamiento de las variables aleatorias de tal forma que los diferentes valores de ellas tienen asignada una probabilidad.

Ejemplo 14. Asignación de probabilidades a las variables aleatorias

Una forma fácil de entender la asignación de probabilidades a los valores de una VA es usando la variable aleatoria ***X*** que expresa el "número de caras al lanzar una moneda legal dos veces". Usaremos el enfoque clásico.

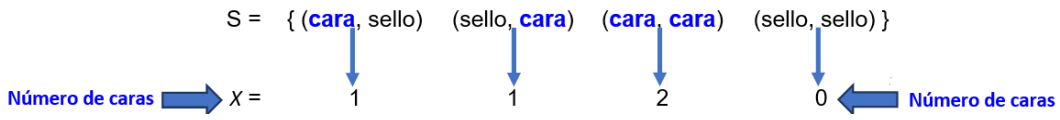
Los posibles resultados de este experimento son:

$$S = \{(\text{cara, sello}), (\text{sello, cara}), (\text{cara, cara}), (\text{sello, sello})\}$$

donde "S" representa el espacio muestral o conjunto de todos los posibles resultados de un experimento.

Luego, los valores de la variable aleatoria ***X "número de caras"*** son:

18 Es la unidad física para indicar la peligrosidad de una radiación (Roentgen Equivalent Man o rem). Sus dimensiones son Joules sobre kilogramo (J/kg).



es decir que **X** puede tomar los valores {0, 1, 2}.

La probabilidad permite asignar a cada valor de **X** una probabilidad.

- El valor de ***X=0*** se observó en uno de los cuatro resultados. Esto es, ***P(X=0)=1/4.***
- El valor de ***X=1*** se observó en dos de los cuatro resultados. Esto es, ***P(X=1)=2/4.***
- El valor de ***X=2*** se observó en uno de los cuatro resultados. Esto es, ***P(X=2)=1/4.***

Observe que cada valor de **X** tiene asociada una probabilidad. Así se obtiene la "**distribución de probabilidad de X**" que en este caso puede resumirse en una tabla que muestra todos los posibles valores de **X** con su respectiva probabilidad (Ver Tabla 19).

Tabla 19. Distribución de probabilidad de la variable aleatoria X "número de caras al lanzar una moneda legal dos veces".

X	P(X)
0	0.25
1	0.50
2	0.25
Total	1

Nota. Elaboración propia.

Note que esta distribución de probabilidad cumple con los dos requisitos básicos de toda probabilidad, es decir:

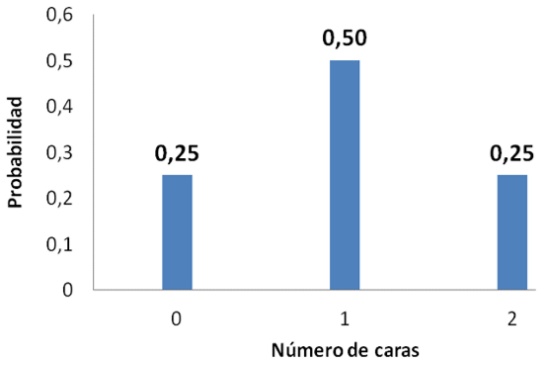
1. Todas las probabilidades están entre 0 y 1:

$$0 \leq P(A_i) \leq 1$$

2. La suma de todas las probabilidades es igual a 1:

$$\sum_{i=1}^{n} P(A_i) = 1$$

La Gráfica 13 muestra la distribución de probabilidad de **X** donde se observa que lo más probable, cuando se lancen dos monedas, es observar una cara.



Gráfica 13. Distribución de probabilidad de la variable aleatoria X "número de caras al lanzar una moneda legal dos veces". Elaboración propia.

Ejemplo 15. Lanzar una moneda legal cuatro veces

De la misma forma se puede construir la distribución de probabilidad de la variable aleatoria **X** que expresa el "**número de caras al lanzar una moneda legal cuatro veces**". Los 16 posibles resultados de este experimento son:

Id	Resultados posibles			
1	sello	sello	sello	sello
2	**cara**	sello	sello	sello
3	sello	**cara**	sello	sello
4	sello	sello	**cara**	sello
5	sello	sello	sello	**cara**
6	**cara**	**cara**	sello	sello
7	**cara**	sello	**cara**	sello
8	**cara**	sello	sello	**cara**
9	sello	sello	**cara**	**cara**
10	sello	**cara**	sello	**cara**
11	sello	**cara**	**cara**	sello
12	sello	**cara**	**cara**	**cara**
13	**cara**	sello	**cara**	**cara**
14	**cara**	**cara**	sello	**cara**
15	**cara**	**cara**	**cara**	sello
16	**cara**	**cara**	**cara**	**cara**

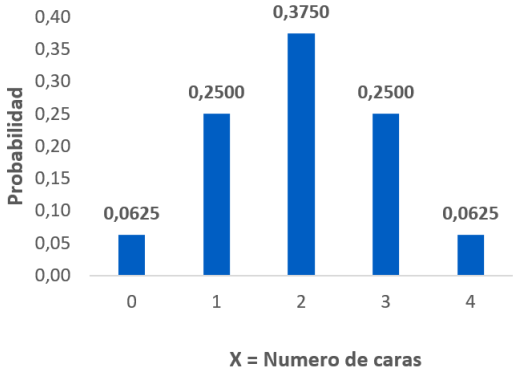
Ahora observamos la variable aleatoria X (número de caras al lanzar una moneda legal cuatro veces). La última columna entrega los resultados de esta variable aleatoria.

Id	Resultados posibles				X
1	sello	sello	sello	sello	**0**
2	**cara**	sello	sello	sello	**1**
3	sello	**cara**	sello	sello	**1**
4	sello	sello	**cara**	sello	**1**
5	sello	sello	sello	**cara**	**1**
6	**cara**	**cara**	sello	sello	**2**
7	**cara**	sello	**cara**	sello	**2**
8	**cara**	sello	sello	**cara**	**2**
9	sello	sello	**cara**	**cara**	**2**
10	sello	**cara**	sello	**cara**	**2**
11	sello	**cara**	**cara**	sello	**2**
12	sello	**cara**	**cara**	**cara**	**3**
13	**cara**	sello	**cara**	**cara**	**3**
14	**cara**	**cara**	sello	**cara**	**3**
15	**cara**	**cara**	**cara**	sello	**3**
16	**cara**	**cara**	**cara**	**cara**	**4**

Tabla 20. Distribución de probabilidad de la variable aleatoria *X* "número de caras al lanzar una moneda legal cuatro veces".

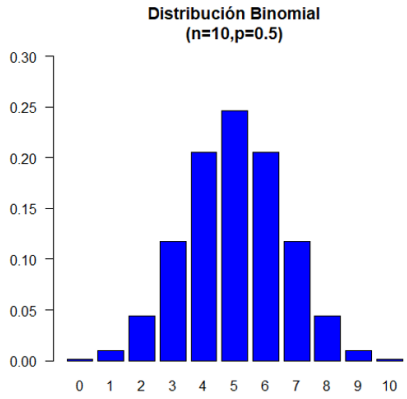
X	*N*	*P(X)*
0	1	0,0625
1	4	0,2500
2	6	0,3750
3	4	0,2500
4	1	0,0625
Total	16	1

Nota. Elaboración propia.



Experimentos o situaciones similares a los Ejemplo 14 y Ejemplo 15 se ajustan a una **distribución de probabilidad binomial** donde se realizan ***n*** ensayos independientes, cada uno con probabilidad de éxito ***p*** y probabilidad de fracaso ***(1-p)***. Usando la fórmula de la distribución binomial[19] se puede encontrar la distribución de probabilidad de ***X*** con ***n=10*** y ***p=0,05*** (Gráfica 14).

19 Fórmula de la distribución binomial que calcula la probabilidad de observar exactamente *k* éxitos en *n* pruebas cuando la probabilidad de éxito es p: $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$; $k = 0,1, \dots, n$



Gráfica 14. Distribución de probabilidad de una variable aleatoria binomial *X* "número de éxitos en *n* pruebas" con *n*=10 y *p*=0,05. Elaboración propia.

De manera similar a esta distribución de probabilidad, se han definido otros **modelos de probabilidad** para variables aleatorias discretas y continuas. Estos modelos asignan probabilidades a los diferentes valores de las variables aleatorias de forma similar a los Ejemplo 14 y Ejemplo 15.

Los modelos más importantes para asignar probabilidades se presentan en la Figura 15.

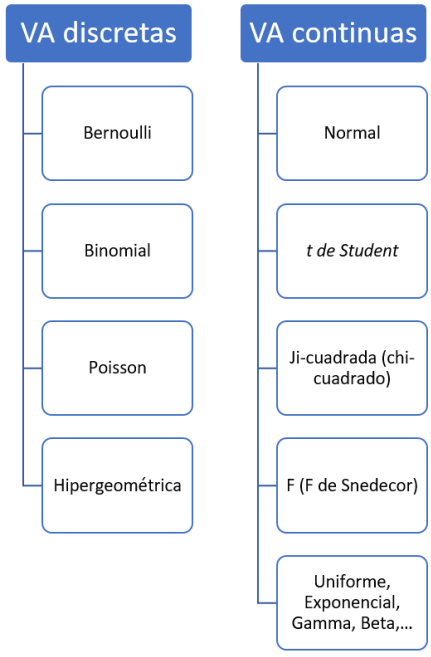


Figura 15. Distribuciones de probabilidad más importantes para variables aleatorias discretas y continuas. Elaboración propia.

En este manual se empezó estudiando la distribución normal que es una de las más utilizadas en bioestadística (Figura 16).

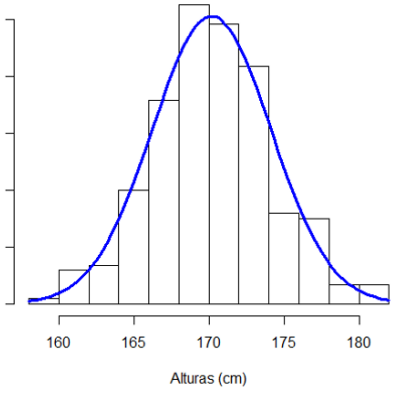


Figura 16. Distribución de probabilidad normal. Elaboración propia.

Cabe resaltar que toda distribución de probabilidad debe cumplir con dos requisitos básicos de toda probabilidad, es decir:

$$0 \leq P(A_i) \leq 1 \qquad \sum_{i=1}^{n} P(A_i) = 1$$

Cuando las variables aleatorias son continuas no podemos hablar de sumatorias porque tenemos infinitos valores. Los requerimientos se convierten en:

1. La función de densidad de probabilidad es mayor que cero: $f(x) \geq 0$.
2. El área bajo la curva representa una probabilidad y su área debe ser uno[20]:

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

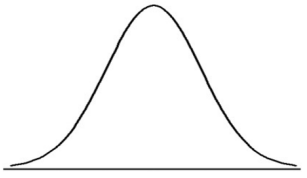
20 En términos matemáticos se dice que la integral de la función, f(x), es igual a 1.

3.1.4 La distribución normal

La **distribución normal** es quizás la distribución más conocida y usada en estadística. Fue reconocida por primera vez por el matemático francés Abraham de Moivre (1667-1754). Posteriormente, el matemático y físico alemán, Carl Friedrich Gauss (1777-1855) realizó desarrollos más profundos y formuló su ecuación; de ahí que también se la conozca, más comúnmente, como la "**campana de Gauss**" (14).

Otros nombres con los que se conoce esta distribución son:

- Distribución normal.
- Distribución de Gauss.
- Distribución gaussiana.



La importancia de la *distribución normal* se debe principalmente a que diversas variables asociadas a fenómenos naturales y estadísticos que se usan en la estadística inferencial siguen, aproximadamente, esta distribución. Características antropométricas (como el peso o la talla), o psicológicos (como el cociente intelectual) son algunos ejemplos de variables que frecuentemente siguen una distribución normal (14).

Cuando el conjunto de datos que se estudia tiene una distribución simétrica, lo cual se puede visualizar en un histograma o un diagrama de cajas *(boxplot)*, es un indicio que la variable sigue una distribución normal.

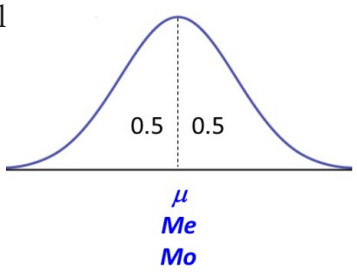
Por ejemplo, la VA continua X (estatura) que se muestra en la Figura 16 tiene forma de campana o de montaña, por lo tanto se dice que la variable X sigue una distribución normal.

La función de densidad de X, simbolizada por $f(x)$, es la función matemática que genera la campana y está dada por (1):

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\left[\frac{(x-\mu)^2}{2\sigma^2}\right]} \quad -\infty < x < \infty$$

donde μ (miu) es la media, σ (sigma) es la desviación estándar y σ^2 es la varianza de X.

Las principales características de la distribución normal son (1):



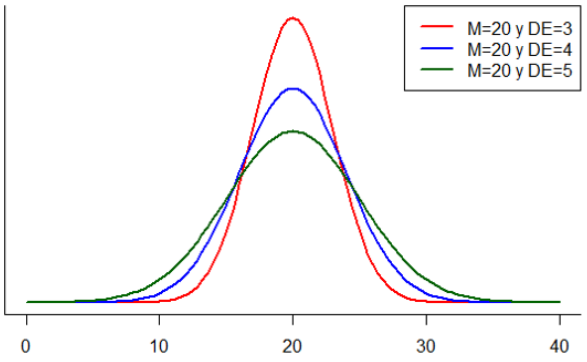
- Tiene forma de campana.
- Es unimodal.
- Es simétrica con respecto a la media.
- Coinciden en un mismo punto la media, la mediana y la moda.
- El área bajo la curva representa una probabilidad, por tanto el área bajo toda la curva es uno (1) o el 100 %.
- Al ser simétrica respecto a la media (μ) deja un área igual a 0.5 a la izquierda y otra igual a 0.5 a la derecha.
- Es asintótica, es decir su rango va desde menos infinito a más infinito ($-\infty < x < \infty$).
- Los dos parámetros que definen la curva son μ y σ.

Siendo la media (μ) y la desviación estándar (σ) los dos parámetros que definen una curva normal existen infinitas curvas según los valores que tomen estos parámetros.

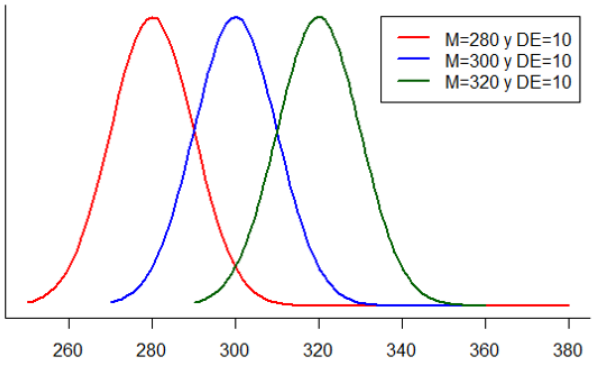
La Figura 17 (a, b y c) (15) muestran tres ejemplos de distribuciones normales con estas características:

- **a)** Distribuciones con igual promedio y diferente variabilidad.
- **b)** Distribuciones con igual variabilidad y diferente promedio.
- **c)** Distribuciones con diferente promedio y diferente variabilidad.

a)



b)



c)

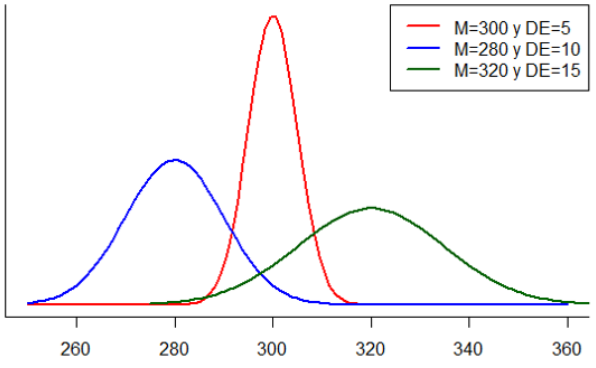


Figura 17. Distribuciones de probabilidad normal con diferentes parámetros de media (M) y desviación estándar (DE): a) Igual promedio y diferente variabilidad; b) Diferente promedio e igual variabilidad; c) Diferente promedio y variabilidad. Elaboración propia.

Por ejemplo, si tenemos una variable aleatoria continua:

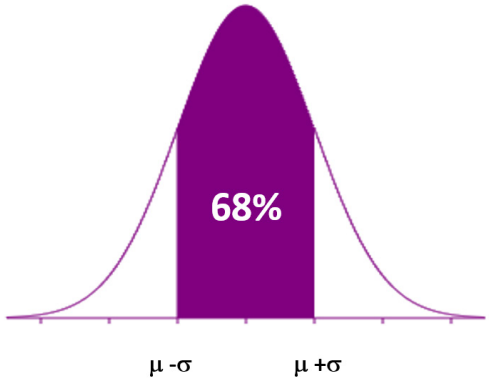
X: tensión arterial sistólica (mmHg)

Para simbolizar que la variable sigue una distribución normal con media μ y varianza σ^2 se denota:

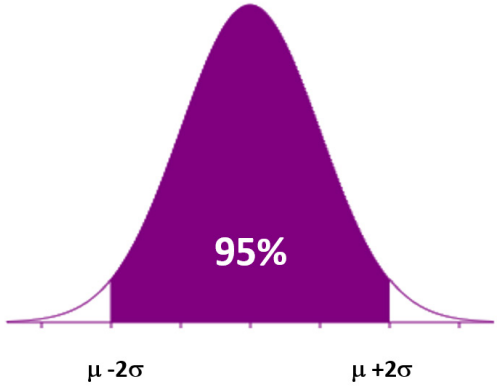
$$X \sim N(\mu,\sigma^2)$$

Otra característica importante de una distribución normal es que puede ser resumida en la regla empírica (Figura 18) la cual establece que en cualquier distribución de este tipo, aproximadamente el 68 % de las observaciones se encuentran a menos de una desviación estándar de la media (μ±σ) **(a)**; cerca del 95 % de los datos se encuentran a dos desviaciones estándares de la media (μ±2σ) **(b)**; y el 99 % estarán a menos de tres desviaciones estándares de la media (μ±3σ) **(c)** (1).

En la normal, aproximadamente el 68 % de las observaciones se encuentran a una desviación estándar de la media.



Cerca del 95 % de las observaciones se encuentran a una desviación estándar de la media.



Aproximadamente el 99 % de las observaciones se encuentran a una desviación estándar de la media.

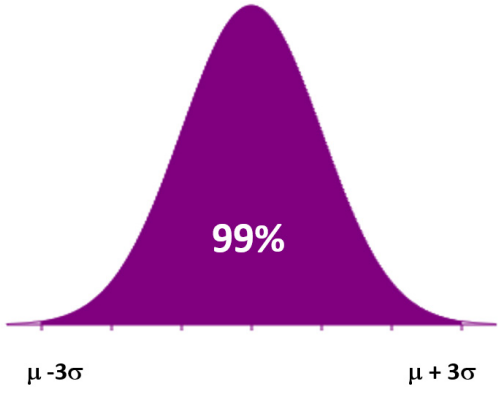
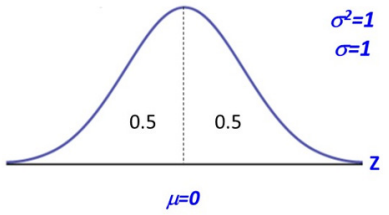


Figura 18. Regla empírica de la distribución normal. Gráficas realizadas en Calculadoras Online [Internet]. Calculadoras Online. 2018 [citado el 15 de diciembre de 2021]. Disponible en: https://calculadorasonline.com/

3.1.5 La distribución normal estándar

La *distribución normal estándar* es la más importante de las distribuciones normales (1).



Tiene una media igual a cero (***μ=0***) y una varianza igual a 1 (***σ²=1***). Luego su desviación estándar también es igual a uno (***σ=1***).

Una variable que tenga distribución normal estándar se denota así: $Z \sim N(0, 1)$.

Su función densidad *f(z)* es (1):

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\left[\frac{z^2}{2}\right]} \qquad ; -\infty < z < \infty$$

Esta distribución se puede obtener a partir de la ecuación principal, creando una variable aleatoria *Z*:

$$Z = \frac{X - \mu}{\sigma}$$

Esto significa que cualquier variable aleatoria ***X*** con distribución normal, $X \sim Normal(\mu, \sigma^2)$, se puede convertir en una distribución normal estándar al restarle el promedio (μ) y dividiendo sobre su desviación estándar (σ) (1). A este proceso se le llama "estandarizar", y será abordado más adelante.

Ejemplo 16. Distribución normal estándar

Muchas herramientas de la Estadística Inferencial asumen una distribución normal, siendo estas las bases matemáticas para las estimaciones de varios parámetros usando intervalos de confianza o pruebas de hipótesis.

Asumiendo una distribución normal estándar, calcular las siguientes probabilidades:

a. Probabilidad de observar un valor de ***Z*** menor que cero.

b. Probabilidad de observar un valor de ***Z*** mayor que uno.

c. Probabilidad de observar un valor de ***Z*** entre -1,96 y +1,96.

d. Probabilidad de observar un valor de ***Z*** mayor que 1,96.

e. Probabilidad de observar un valor de ***Z*** menor que -1,96.

f. Probabilidad de observar un valor de ***Z*** menor que -1,96 y mayor que 1,96.

Formalmente se denota así:

g. Probabilidad de observar un valor de ***Z*** menor que cero → P(Z < 0).

h. Probabilidad de observar un valor de ***Z*** mayor que uno → P(Z > 1).

i. Probabilidad de observar un valor de ***Z*** entre -1,96 y +1,96 → P(-1,96 < Z < 1,96).

j. Probabilidad de observar un valor de ***Z*** mayor que 1,96 → P(Z > 1,96).

k. Probabilidad de observar un valor de ***Z*** menor que -1,96 → P(Z < -1,96).

l. Probabilidad de observar un valor de ***Z*** menor que -1,96 y mayor que 1,96→ P(Z < -1,96) + P(Z > 1,96).

El área morada de la Figura 19 representan las probabilidades requeridas:

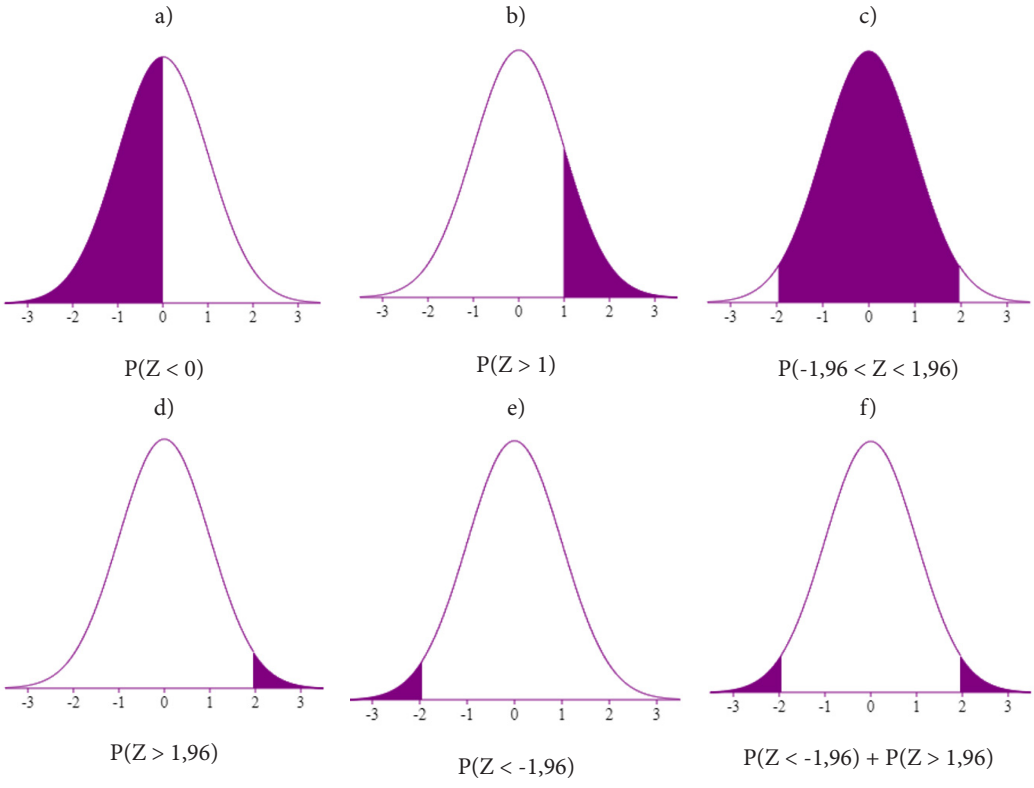


Figura 19. Probabilidades requeridas del Ejemplo 16. Gráficas realizadas en Calculadoras Online [Internet]. Calculadoras Online. 2018 [citado el 15 de diciembre de 2021]. Disponible en: https://calculadorasonline.com/

Para obtener estas probabilidades se tienen varias opciones:

1. Calcular integrales y obtener el área bajo la curva que representa la "probabilidad" a partir de la función de densidad *f(z)*.

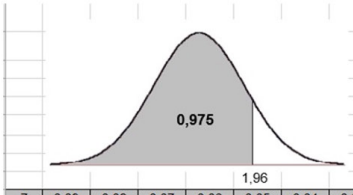
2. Usar los valores tabulados en una tabla de distribución de la normal estándar (ver anexo 1).

3. Usar algún programa como Excel u otros como R o Epidat.

4. Usar alguna calculadora online en Internet (recomendado).

La tabla de la distribución normal estándar (anexo 1) entrega áreas o probabilidades menores que un valor de Z (Figura 20). Por ejemplo:

$$P(Z < 1{,}96) = 0{,}975$$



DISTRIBUCION NORMAL ESTÁNDAR
Áreas bajo la curva

$P(Z < 1{,}96) = 0{,}9750$

$Z_{0{,}9750} = 1{,}96$

$P(Z > 1{,}96) = 1 - P(Z < 1{,}96) = 1 - 0{,}9750 = 0{,}025$

Z	0,09	0,08	0,07	0,06	0,05	0,04	0,03	0,02	0,01	0,00
-3,80	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
-3,70	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001
-3,60	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0001	0,0002	0,0002
-3,50	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002
-3,40	0,0002	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003
-3,30	0,0003	0,0004	0,0004	0,0004	0,0004	0,0004	0,0004	0,0005	0,0005	0,0005
-3,20	0,0005	0,0005	0,0005	0,0006	0,0006	0,0006	0,0006	0,0006	0,0007	0,0007
-3,10	0,0007	0,0007	0,0008	0,0008	0,0008	0,0008	0,0009	0,0009	0,0009	0,0010
-3,00	0,0010	0,0010	0,0011	0,0011	0,0011	0,0012	0,0012	0,0013	0,0013	0,0013
-2,90	0,0014	0,0014	0,0015	0,0015	0,0016	0,0016	0,0017	0,0018	0,0018	0,0019
-2,80	0,0019	0,0020	0,0021	0,0021	0,0022	0,0023	0,0023	0,0024	0,0025	0,0026
-2,70	0,0026	0,0027	0,0028	0,0029	0,0030	0,0031	0,0032	0,0033	0,0034	0,0035
-2,60	0,0036	0,0037	0,0038	0,0039	0,0040	0,0041	0,0043	0,0044	0,0045	0,0047
-2,50	0,0048	0,0049	0,0051	0,0052	0,0054	0,0055	0,0057	0,0059	0,0060	0,0062
-2,40	0,0064	0,0066	0,0068	0,0069	0,0071	0,0073	0,0075	0,0078	0,0080	0,0082
-2,30	0,0084	0,0087	0,0089	0,0091	0,0094	0,0096	0,0099	0,0102	0,0104	0,0107
-2,20	0,0110	0,0113	0,0116	0,0119	0,0122	0,0125	0,0129	0,0132	0,0136	0,0139
-2,10	0,0143	0,0146	0,0150	0,0154	0,0158	0,0162	0,0166	0,0170	0,0174	0,0179
-2,00	0,0183	0,0188	0,0192	0,0197	0,0202	0,0207	0,0212	0,0217	0,0222	0,0228
-1,90	0,0233	0,0239	0,0244	0,0250	0,0256	0,0262	0,0268	0,0274	0,0281	0,0287
-1,80	0,0294	0,0301	0,0307	0,0314	0,0322	0,0329	0,0336	0,0344	0,0351	0,0359
-1,70	0,0367	0,0375	0,0384	0,0392	0,0401	0,0409	0,0418	0,0427	0,0436	0,0446
-1,60	0,0455	0,0465	0,0475	0,0485	0,0495	0,0505	0,0516	0,0526	0,0537	0,0548

Z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,00	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,10	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,20	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,30	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,40	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,50	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,60	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,70	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,80	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,90	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,00	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,10	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,20	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,30	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,40	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,50	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,60	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,70	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,80	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,90	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,00	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,10	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,20	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890

Figura 20. Tabla de la distribución normal estándar obteniendo la probabilidad de que Z sea menor que 1,96 → P(Z < 1,96) = 0,9750.

Para las siguientes distribuciones de probabilidad continuas que serán estudiadas (*normal*, *t de Student* y *Ji-cuadrada*) sugerimos el uso de calculadoras en línea para calcular áreas bajo la curva o puntos de corte de sus distribuciones.

Calculadora en línea distribución normal:

https://calculadorasonline.com/calculadora-de-distribucion-nomal-campana-de-gauss/

https://homepage.divms.uiowa.edu/~mbognar/applets/normal.html

https://davidmlane.com/hyperstat/z_table.html

Calculadora en línea distribución *t* de *Student*:

https://homepage.divms.uiowa.edu/~mbognar/applets/t.html

Calculadora en línea distribución *Ji-cuadrada*:

https://homepage.divms.uiowa.edu/~mbognar/applets/chisq.html

Usando la calculadora en línea sugerida se obtienen los siguientes resultados (Figura 21):

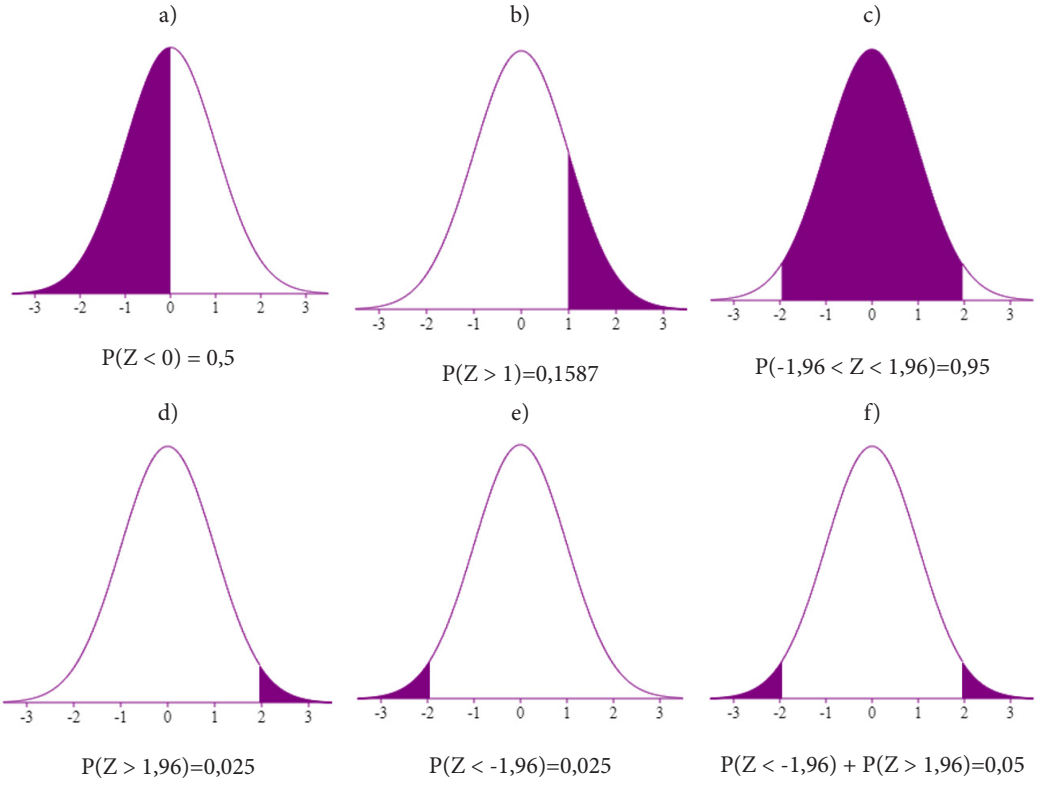


Figura 21. Probabilidades requeridas del Ejemplo 16. Gráficas realizadas en Calculadora de Distribucion Normal o Distribucion Gausiana de probabilidad - Campana de Gauss [Internet]. Calculadoras Online. 2019 [citado el 21 de diciembre de 2021]. Disponible en: https://calculadorasonline.com/calculadora-de-distribucion-nomal-campana-de-gauss/

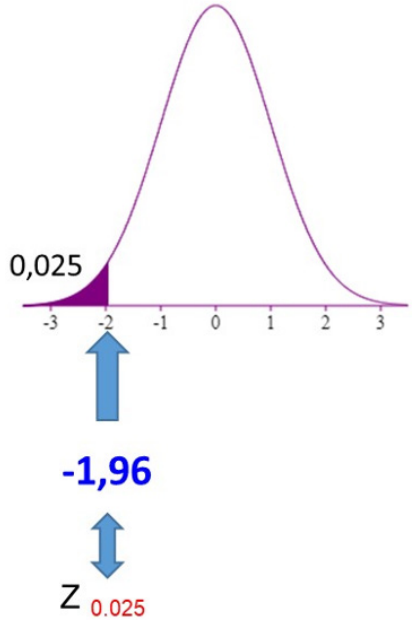
Notación importante: la Figura 22 muestra una notación importante que se requiere para futuros temas que se trabajarán en la estadística inferencial.

La probabilidad de encontrar un valor de ***Z*** menor que -1,96 es del 2,5 %:

$$\mathbf{P\ (Z < -1{,}96) = 0{,}025}$$

Entonces, el valor de ***Z*** que deja por debajo el 2,5 % de los datos es -1,96:

$$Z_{0.025} = -1{,}96$$



De igual manera, la probabilidad de encontrar un valor de ***Z*** menor que 1,96 es del 97,5 %:

$$\mathbf{P\ (Z < 1{,}96) = 0{,}9750}$$

Entonces, el valor de ***Z*** que deja por debajo el 97,5 % de los datos es 1,96:

$$Z_{0.9750} = 1{,}96$$

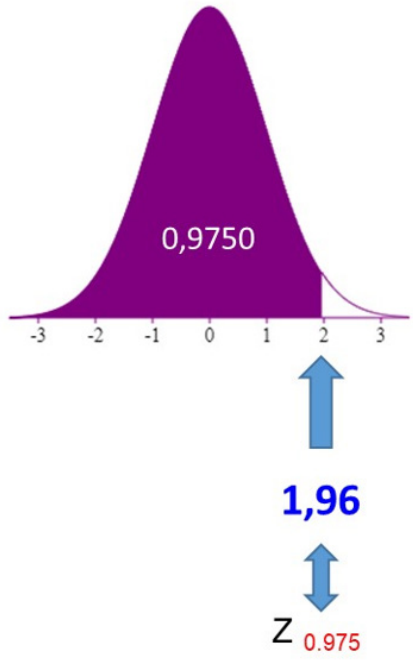


Figura 22. Notación importante en el manejo de la distribución normal estándar. Gráficas realizadas en https://calculadorasonline.com/

Entonces, en el próximo módulo cuando necesitemos valores de $Z_{0.05}$, $Z_{0.95}$, $Z_{0.025}$ y $Z_{0.975}$ debemos tener clara esta notación (Figura 23).

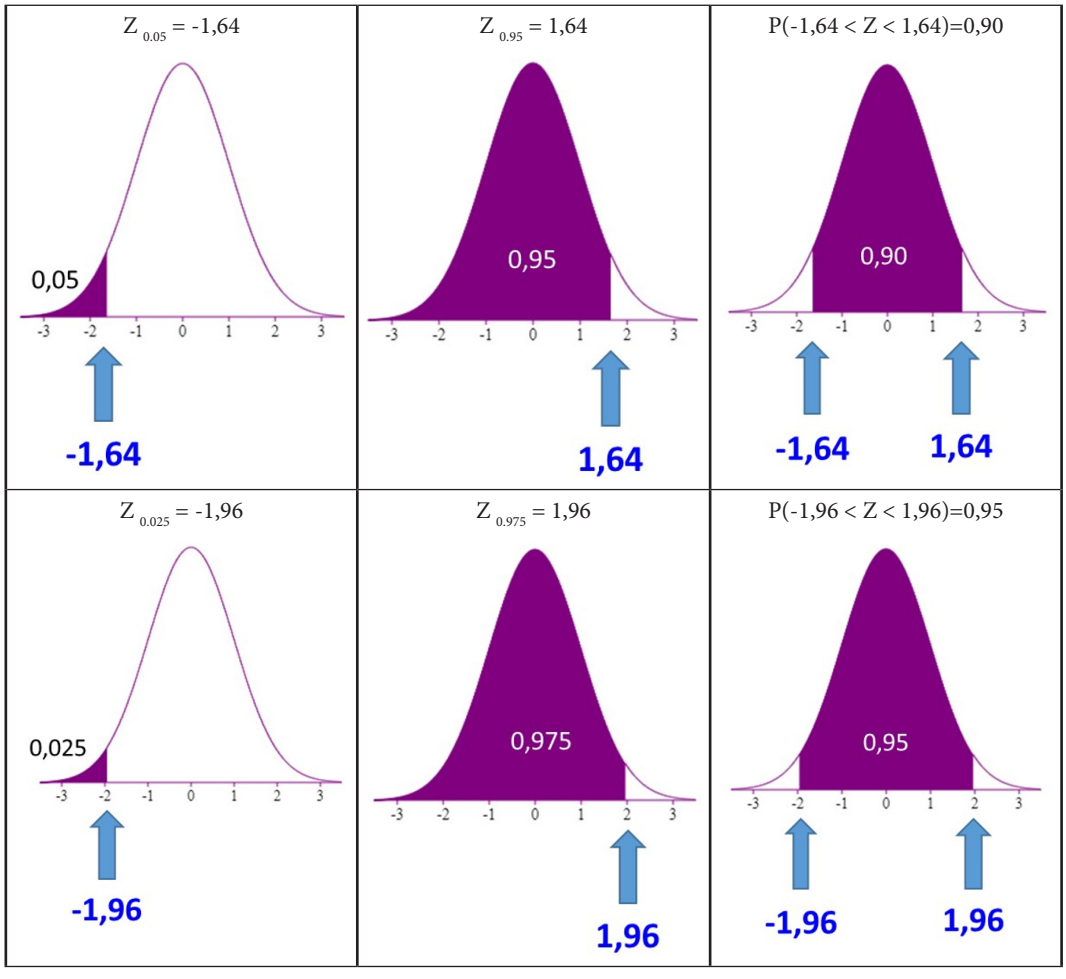


Figura 23. Valores importantes de la distribución normal estándar asociados a valores de $Z_{0.05}$, $Z_{0.95}$, $Z_{0.025}$ y $Z_{0.975}$. Gráficas realizadas en Calculadoras Online [Internet]. Calculadoras Online. 2018 [citado el 15 de diciembre de 2021]. Disponible en: https://calculadorasonline.com/

Estandarización

En la vida real ninguna variable tendrá una distribución normal estándar, es decir con media cero y varianza uno. Sin embargo, para hallar probabilidades se hace uso de un proceso muy sencillo llamado estandarización (1).

La estandarización se utiliza para convertir de una variable aleatoria con distribución normal X (con cualquier media y cualquier varianza) a una normal estándar (Z). Básicamente al valor de interés X se le resta la media (μ) y se divide sobre la desviación estándar (σ).

$$Z = \frac{X - \mu}{\sigma}$$

Ejemplo 17. Aplicación de la estandarización

El Instituto Nacional de Salud tiene un "Protocolo de Vigilancia y Control del Bajo Peso al Nacer a Término[21]" con el fin de identificar los recién nacidos a término con bajo peso al nacer[22]. Lo anterior con el propósito de evaluar intervenciones inmediatas que minimicen los riesgos de morbilidad o mortalidad asociada a esta condición y establecer la distribución del evento para priorizar su atención como medida de control en salud pública.

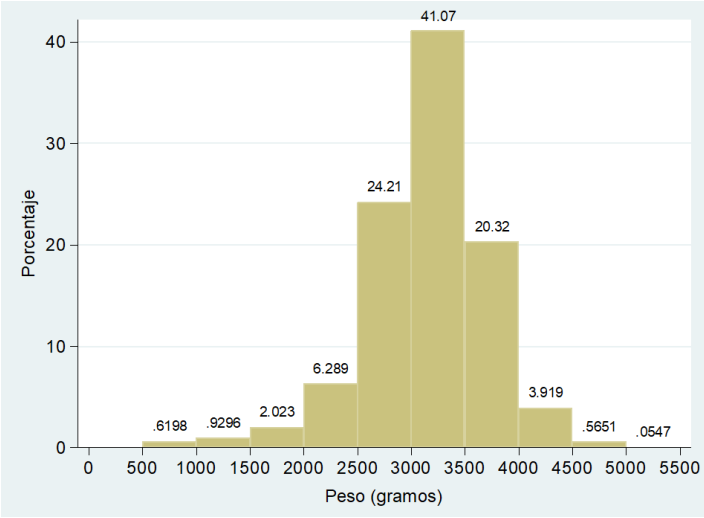
La Gráfica 15 muestra la distribución del peso al nacer (gramos) de 2 000 niños nacidos en un hospital de alto nivel en Cali-Colombia durante el 2012.

Aquí hay una variable aleatoria continua (X: peso al nacer en gramos) con una distribución normal.

$$X \sim N\,(\mu=3.094, \sigma^2=564^2)$$

21 Se considera un embarazo a término al cabo de 40 semanas (280 días), con un rango normal entre 37 y 42 semanas.

22 Se considera bajo peso al nacer (BPN) al neonato cuyo peso es igual o menor a 2 499 gramos, independiente de la edad gestacional y cualquiera que sea la causa.



Gráfica 15. Histograma de la distribución del peso al nacer (gramos) de 2 000 niños nacidos en un hospital de alto nivel en Cali-Colombia durante el 2012. Peso promedio=3 094 gramos; desviación estándar=564 gramos. Elaboración propia.

Asumiendo que los pesos al nacer siguen una distribución normal:

a. ¿Qué porcentaje de los bebes nacidos en el hospital de Cali presentaron bajo peso al nacer?

b. ¿Cuántos niños presentaron bajo peso al nacer?

c. ¿Qué porcentaje presentaron un peso normal?

d. ¿Qué porcentaje presentaron un peso mayor a 4 000 gramos?

e. ¿Qué porcentaje presentaron un peso menor a 3 750 gramos?

f. ¿En cuál intervalo se concentraron el 95% del peso de los recién nacidos?

Solución:

a ¿Qué porcentaje de los bebes nacidos en el hospital de Cali presentaron bajo peso al nacer?

Sea ***X*** una variable aleatoria continua que representa el "peso al nacer (gramos)" de los bebes. De la Gráfica 15 se sabe que X ~Normal (3 094 ,564²). Esto es:

$$\mu = 3\,094 \text{ gramos}$$

$$\sigma = 564 \text{ gramos}$$

El bajo peso al nacer se presenta cuando es menor a 2 500 gramos. Luego se debe calcular:

$$P(X < 2.500) = ?$$

Las opciones para hallar la respuesta son dos: 1) Proceso de estandarización; y 2) Calculadora en línea.

Para estandarizar se sabe que:

$$Z = \frac{X - \mu}{\sigma}$$

Luego:

$$P\left[\frac{\boldsymbol{X} - \mu}{\sigma} < \frac{\mathbf{2.500} - \mu}{\sigma}\right]$$

$$P\underbrace{\left[\frac{\boldsymbol{X} - 3.094}{564}\right.}_{\boldsymbol{Z}} < \left.\frac{\mathbf{2.500} - 3.094}{564}\right]$$

Entonces,

$$P\left[Z < \frac{\mathbf{2.500} - 3.094}{564}\right]$$

$$P[Z < -1{,}053]$$

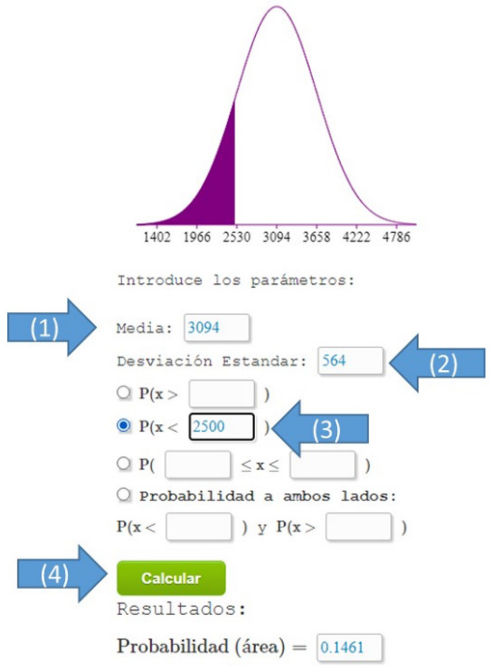
De la tabla de la distribución normal estándar, $P(Z < -1{,}05) = 0{,}1469$

Usando la calculadora en línea sugerida[23] se deben seguir estos pasos:

1. Especificar el promedio de 3 094
2. Indicar el valor de la desviación estándar (564)
3. Seleccionar la opción P(X<2 500)
4. Clic en "Calcular"

Luego, P(X<2.500) = 0,1461[24]

En resumen, el **14,6 %** de los bebés nacidos en Cali en 2012 presentaron bajo peso al nacer.



b ¿Cuántos niños presentaron bajo peso al nacer?

Del punto a) se encontró que el 14,61 % de los bebes presentaron bajo peso al nacer.

Si la información de la Gráfica 15 informaba que ese año en Cali ocurrieron 2 000 nacimientos, entonces:

$$2000 \times 0{,}1461 = 292{,}2$$

Es decir que en Cali en 2012 ocurrieron aproximadamente 292 nacimientos con bajo peso al nacer.

c ¿Qué porcentaje presentaron un peso normal?

Se debe calcular el porcentaje de bebés que pesaron más de 2 500 gramos → **P(X > 2.500).**

Hay varias alternativas para obtener la respuesta. La primera opción es calcular **P(X > 2.500)** por el complemento, siendo quizás la opción más sencilla.

23 Calculadora de Distribucion Normal o Distribucion Gausiana de probabilidad - Campana de Gauss [Internet]. Calculadoras Online. 2019 [citado el 21 de diciembre de 2021]. Disponible en: https://calculadorasonline.com/calculadora-de-distribucion-nomal-campana-de-gauss/

24 Las diferencias encontradas en las probabilidades usando la tabla de la normal estándar (0,1469) y la calculadora online (0,1461) es explicado por el redondeo de las cifras. Recuerde que la tabla de la normal estándar solo entrega dos cifras decimales para los valores de Z.

En el punto a) se encontró que el 14,61 % de los bebés presentaron bajo peso al nacer (X<2.500) luego los bebés que NO presentaron bajo peso al nacer (X>2.500) es el resto (85,39 %).

$$P(X>2.500) = 1 - P(X<2.500)$$

$$P(X>2.500) = 1 - 0,1461$$

$$P(X>2.500) = 0,8539$$

La segunda opción es usar el proceso de estandarización:

$$P(X > 2.500) = ?$$

$$P(X > 2.500) = 1 - P(X < 2.500)$$

Se sabe que:

$$Z = \frac{X - \mu}{\sigma}$$

Luego:

$$1 - P\left[\frac{\mathbf{X} - \mu}{\sigma} < \frac{\mathbf{2.500} - \mu}{\sigma}\right]$$

$$1 - P\left[\underbrace{\frac{\mathbf{X} - 3.094}{564}}_{\boldsymbol{Z}} < \frac{\mathbf{2.500} - 3.094}{564}\right]$$

$$1 - P\left[Z < \frac{\mathbf{2.500} - 3.094}{564}\right]$$

$$1 - P[Z < -1,053]$$

De la tabla de la distribución normal estándar, P(Z < -1,05) = 0,1469. Luego,

$$1 - 0,1469 = 0,8531$$

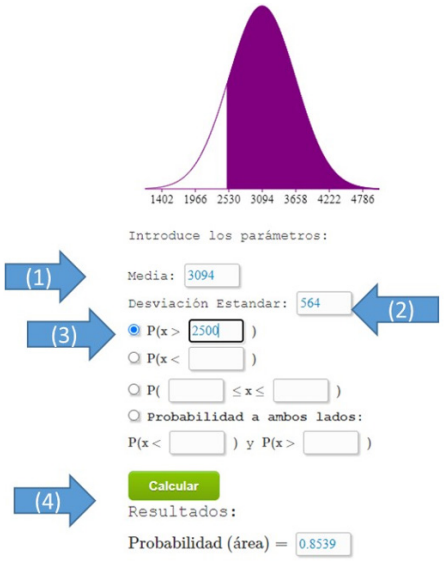
Y la tercera opción es usar la calculadora en línea sugerida:

Usando la calculadora en línea se deben seguir estos pasos:

1. Especificar el promedio de 3 094.
2. Indicar el valor de la desviación estándar (564).
3. Seleccionar la opción P(X>2 500).
4. Clic en "Calcular".

Luego, P(X>2 500) = 0,8539

En resumen, el **85,4 %** de los bebes nacidos en Cali en 2012 presentaron peso normal al nacer (>2 500 gr).



d ¿Qué porcentaje presentaron un peso mayor a 4 000 gramos?

Se debe calcular el porcentaje de bebés que pesaron más de 4 000 gramos:

$$P(X > 4.000) = ?$$

$$P(X > 4.000) = 1 - P(X < 4.000)$$

Usando el proceso de estandarización se sabe que:

$$Z = \frac{X - \mu}{\sigma}$$

Luego:

$$1 - P\left[\frac{\mathbf{X} - \mu}{\sigma} < \frac{\mathbf{4.000} - \mu}{\sigma}\right]$$

$$1 - P\left[\underbrace{\frac{\mathbf{X} - 3.094}{564}}_{\boldsymbol{Z}} < \frac{\mathbf{4.000} - 3.094}{564}\right]$$

$$1 - P\left[Z < \frac{\mathbf{4.000} - 3.094}{564}\right]$$

$$1 - P[Z < 1{,}606]$$

De la tabla de la distribución normal estándar, P(Z < 1,61) = 0,9463. Luego,

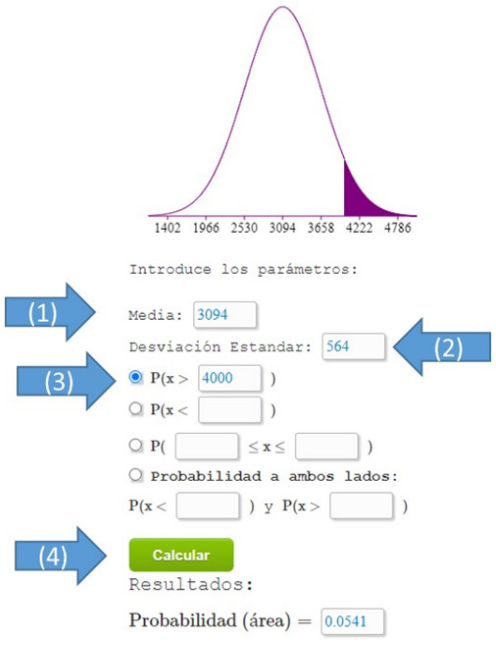
$$1 - 0{,}9463 = 0{,}0537$$

Usando la calculadora en línea se deben seguir estos pasos:

1. Especificar el promedio de 3 094.
2. Indicar el valor de la desviación estándar (564).
3. Seleccionar la opción P(X>4 000).
4. Clic en "Calcular".

Luego, P(X>4 000) = 0,0541

En resumen, el **5,4 %** de los bebés nacidos en Cali en 2012 presentaron un peso mayor a 4 000 gramos.



e ¿Qué porcentaje presentaron un peso menor a 3 750 gramos?

$$P(X < 3.750) = ?$$

Para hallar P(X < 3.750) las opciones son dos: 1) Proceso de estandarización; y 2) Calculadora en línea.

Usando el proceso de estandarización se sabe que:

$$Z = \frac{X - \mu}{\sigma}$$

Luego:

$$P\left[\frac{\mathbf{X} - \mu}{\sigma} < \frac{\mathbf{3.750} - \mu}{\sigma}\right]$$

$$P\left[\frac{\mathbf{X} - 3.094}{564} < \frac{\mathbf{3.750} - 3.094}{564}\right]$$

$$\underbrace{\qquad}_{Z}$$

$$P\left[Z < \frac{\mathbf{3.750} - 3.094}{564}\right]$$

$$P[Z < 1{,}163]$$

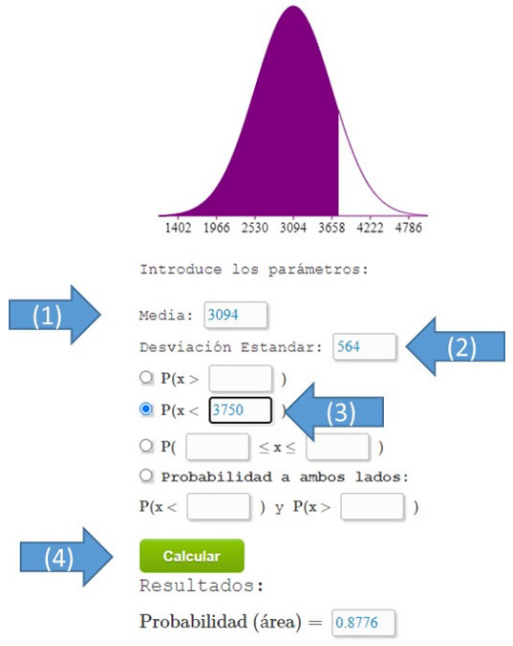
De la tabla de la distribución normal estándar, P(Z < 1,16) = 0,8770.

Usando la calculadora en línea se deben seguir estos pasos:

1. Especificar el promedio de 3 094.
2. Indicar el valor de la desviación estándar (564).
3. Seleccionar la opción P(X<3 750).
4. Clic en "Calcular".

Luego, P(X<3.750) = 0,8776

En resumen, el **87,8 %** de los bebes nacidos en Cali en 2012 presentaron un peso al nacer mayor a 3 750 gr.



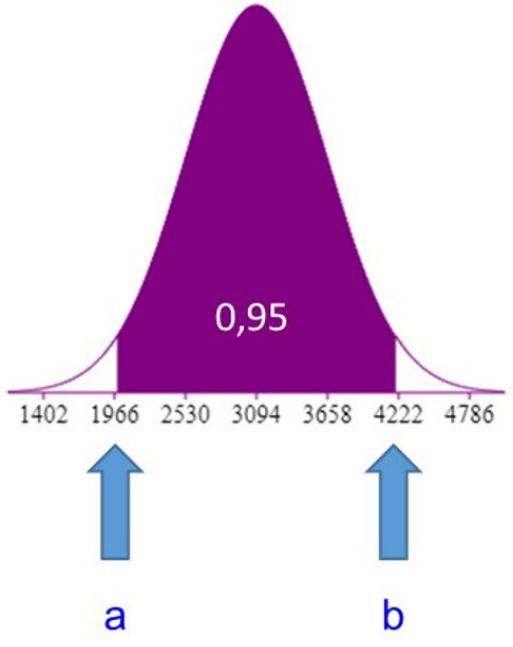
f ¿En cuál intervalo se concentraron el 95 % del peso de los recién nacidos?

Se quiere saber cuál es el peso tal que el 95 % de los bebés se encuentran en la parte central de la distribución.

Se puede escribir así:

$$P(a < X < b) = 0{,}95$$

Se deben encontrar los valores de **a** y **b**, sabiendo que el 2,5 % de los valores son menores que **a** y el 97,5 % son menores que **b**.



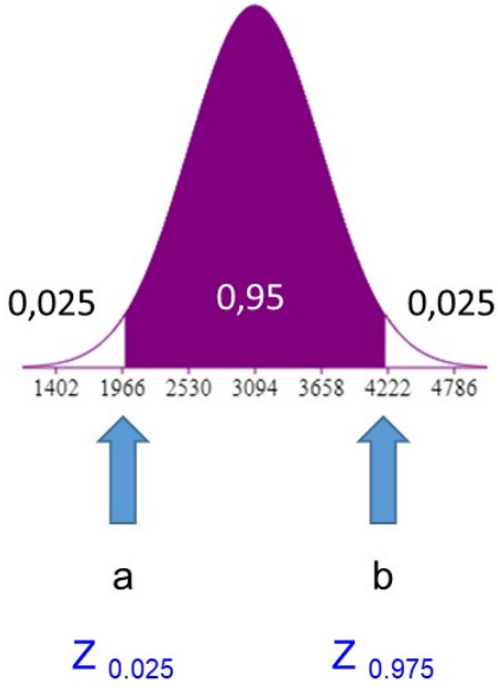
Sabemos que el 2,5 % de los pesos son menores que **a**.

Entonces, usando la fórmula de estandarización:

$$Z = \frac{X - \mu}{\sigma}$$

Se tiene que:

$$Z_{0,025} = \frac{a - \mu}{\sigma}$$



Se sabe que:

$Z_{0.025} = -1{,}96$

$\mu = 3\,094$ gramos

$\sigma = 564$ gramos

Luego,

$$Z_{0,025} = \frac{a - \mu}{\sigma}$$

$$-1.96 = \frac{a - 3094}{564}$$

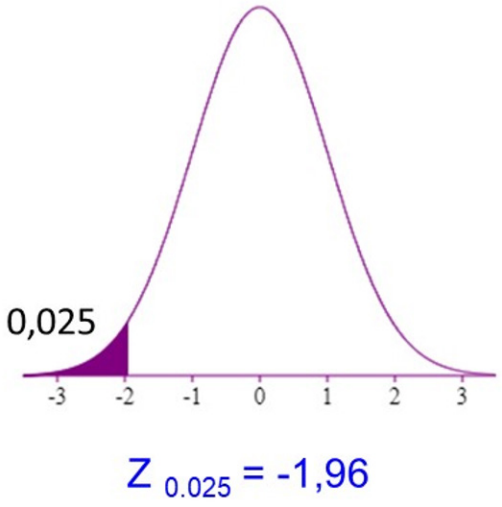
Despejando **a** se tiene que:

$$(-1.96)(564) = a - 3094$$

$$(-1.96)(564) + 3094 = a$$

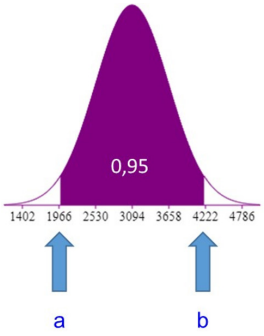
$$a = (-1.96)(564) + 3094$$

$$a = 1.988{,}56 \text{ gramos}$$



Se sigue el mismo procedimiento para encontrar **b**.

Ahora para encontrar **b** se sabe que el 97,5 % de los pesos son menores que **b**.



Entonces, usando la fórmula de estandarización:

$$Z = \frac{X - \mu}{\sigma}$$

Se tiene que:

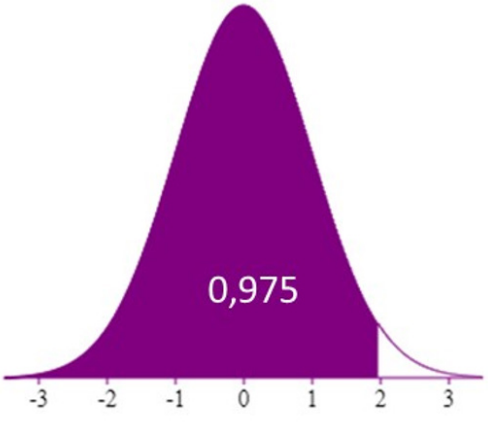
$$Z_{0,975} = \frac{b - \mu}{\sigma}$$

Se sabe que:

$$Z_{0.975} = 1{,}96$$

$$\mu = 3\,094 \text{ gramos}$$

$$\sigma = 564 \text{ gramos}$$



Luego,

$$Z_{0,975} = \frac{b - \mu}{\sigma}$$

$$1.96 = \frac{b - 3094}{564}$$

Despejando **b** se tiene que:

$$(1.96)(564) = b - 3094$$

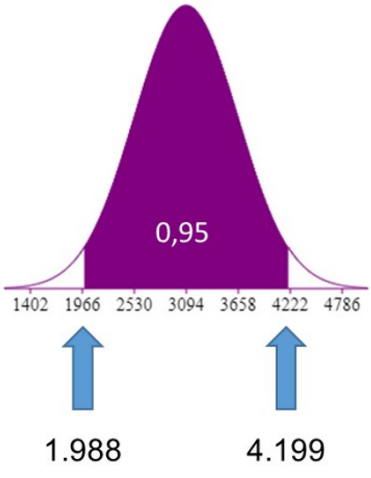
$$(1.96)(564) + 3094 = b$$

$$b = (1.96)(564) + 3094$$

$$b = 4.199{,}44 \text{ gramos}$$

Se concluye que:

- El 2,5 % de los pesos fueron menores a 1 988 gramos.
- El 2,5 % de los pesos fueron mayores a 4 199 gramos.
- El 95 % de los bebés pesaron entre 1 988 y 4 199 gramos.



3.1.6 La distribución *t de Student*

Otra distribución muy usada en la Estadística Inferencial es la *t de Student*. La *distribución t de Student* también hace parte de las distribuciones continuas de probabilidad donde ya estudiamos la distribución normal y además estudiaremos la *distribución Ji-cuadrada*.

William Sealy Gosset (1876-1937), bajo el seudónimo *de Student*, desarrolló la *prueba t* y la *distribución t* (16). Esta distribución surge de la necesidad de estimar la media de una población (o conjunto de datos) con distribución normal, con tamaño de muestra pequeño y varianza poblacional (σ^2) desconocida.

El uso más importante de la distribución *t de Student* es para la construcción de intervalos de confianza y pruebas de hipótesis relacionada con el parámetro "media" y "diferencia de medias".

Al igual que la distribución normal que contiene una familia de curvas según los diferentes valores de los parámetros μ y σ, la distribución *t de Student* también son una familia de distribuciones de probabilidad continúa determinadas por un solo parámetro conocido como "grados de libertad" (df, del inglés *degrees of freedom*).

El significado de grados de libertad (denotado como $\boldsymbol{k}$) representa el número de observaciones "independientes" en un conjunto de datos.

Existe una distribución *t* diferente para cada tamaño de muestra (n) por lo que decimos que la distribución *t* tiene $k\ (k=n-1)$ grados de libertad (Ver Figura 24).

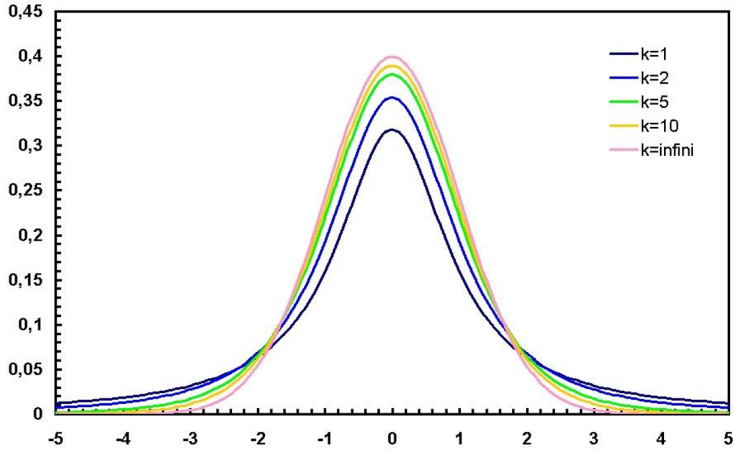


Figura 24. Distribución de probabilidad *t de Student* para varios grados de libertad (k=1,2,5,10,∞). De The original uploader was Thorin de Wikipedia en francés. - Transferido desde fr.wikipedia a Commons., CC BY-SA 1.0, https://commons.wikimedia.org/w/index.php?curid=1878902

Una VA con distribución ***t de Student*** con ***k*** grados de libertad se denota como:

$$t \sim t\ de\ Student(k)$$

La distribución *t de Student*[25] presenta estas características:

- Las curvas son simétricas y tienen forma de campana (similar a distribución normal estándar).
- Las medias de todas las curvas son 0 y sus varianzas son mayores que 1. Esto hace que las colas sean más pesadas que la distribución normal estándar, es decir, las colas disminuyen más lentamente hacia el eje *x*.
- A medida que aumentan los grados de libertad la curva *t de Student* se asemeja a la normal estándar (Figura 25).

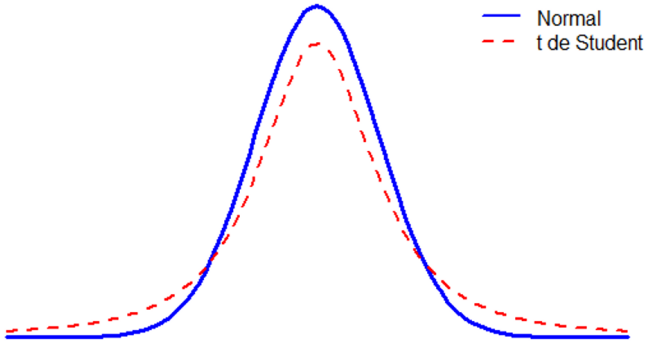


Figura 25. Distribución de probabilidad *t de Student* y normal estándar. Elaboración propia.

Al igual que en la distribución normal, para encontrar probabilidades o cuantiles (valores de t) de la distribución se tienen estas opciones:

1. Hallar la probabilidad por medio de una integral de la función de densidad *f(t)*.
2. Usar los valores tabulados en una tabla de distribución *t de Student* (ver anexo 2).
3. Usar algún programa como Excel u otros como R o Epidat.
4. Usar alguna calculadora online en Internet (recomendado).

Ejemplo 18. Distribución *t de Student*

25 La función de densidad es:

$$f(t) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)\sqrt{\pi k}}\left[1+\frac{t^2}{k}\right]^{-\frac{(k+1)}{2}} \qquad ; -\infty < t < \infty$$

donde Γ es la función gamma.

Las siguientes gráficas y resultados de la distribución *t de Student* se obtuvieron de una calculadora en línea[26] donde grados de libertad se denotan con $\boldsymbol{\nu}$ ($\nu = n-1$) que es equivalente a lo que previamente se había denotado con $\boldsymbol{k}$ ($k=n-1$).

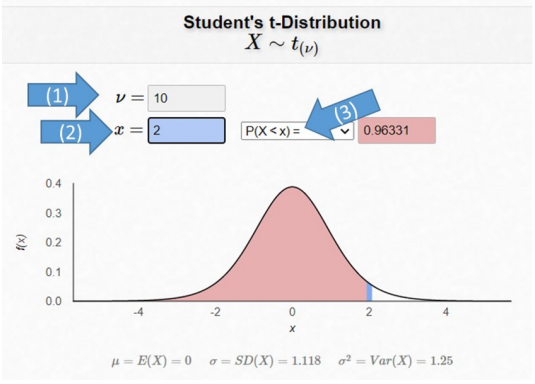
a) ¿Cuál es la probabilidad de observar un valor de $\boldsymbol{t}$ menor que 2 cuando el tamaño de muestra es 11?

$$P(t_{10} < 2) = ¿?$$

Como el tamaño de muestra es 11 (n=11) entonces los grados de libertad son $\boldsymbol{\nu}$=$n-1$ ($\boldsymbol{\nu}$=10).

Para usar la calculadora en línea siga estos pasos como lo muestra la imagen:

1. Especificar los grados de libertad (v=10).
2. Hay que indicar que el valor de $\boldsymbol{X}$ es 2.
3. Seleccionar el área menor (<).

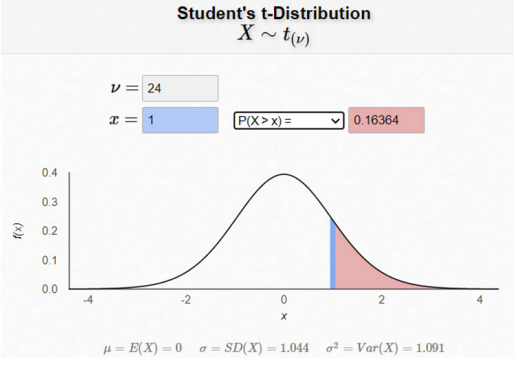


Automáticamente en la celda rosada se calcula la probabilidad. Luego:

$$P(t_{10} < 2)=0.96331$$

b) Probabilidad de observar un valor de t mayor que 1 cuando el tamaño de muestra es 25:

$$P(t_{24} > 1)=0.1636$$



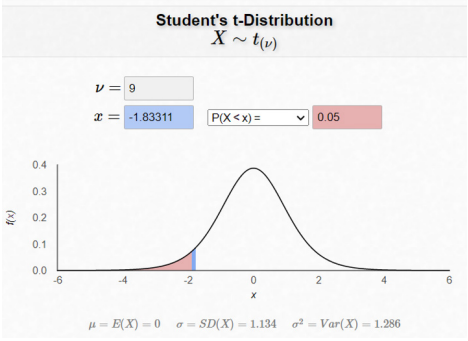
26 Bognar M. Student's t-Distribution. 2021 [citado el 12 de 2021]. Disponible en: https://homepage.divms.uiowa.edu/~mbognar/applets/t.html

c) El valor de *t* para un nivel de significancia del 5 % y un tamaño de muestra de 10.

$$t_{9;0.05} = -1.833$$

En la calculadora indique los grados de libertad (ν=9), seleccione P(X<x) y coloque 0.05 en la celda rosada. El valor de *t* se muestra en la celda azul.

Nota: este valor también lo puede obtener de la tabla de distribución *t de Student* (ver anexo 2).

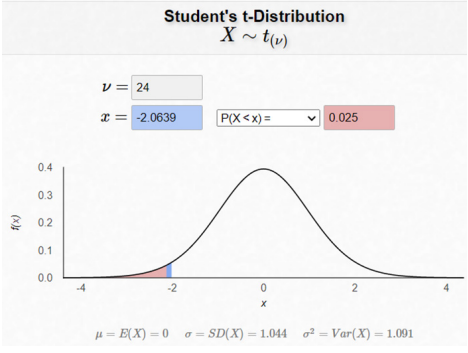


d) El valor de *t* para un nivel de significancia del 2.5 % y un tamaño de muestra de 25.

$$t_{24;0.025} = -2.063$$

En la calculadora indique los grados de libertad (ν=24), seleccione P(X<x) y coloque 0.025 en la celda rosada. El valor de *t* se muestra en la celda azul.

Nota: este valor también lo puede obtener de la tabla de distribución *t de Student* (ver anexo 2).

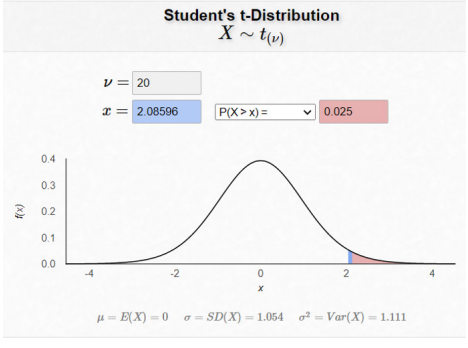


e) El valor de *t* que deja por encima el 2,5 % de los datos; con 20 grados de libertad.

En la calculadora indique los grados de libertad (ν=20), seleccione P(X>x) y coloque 0.025 en la celda rosada. El valor de *t* se muestra en la celda azul.

Nota: este valor también lo puede obtener de la tabla de distribución *t de Student* (ver anexo 2). Basta buscar el área menor:

$$t_{20;0.975} = 2.0860$$

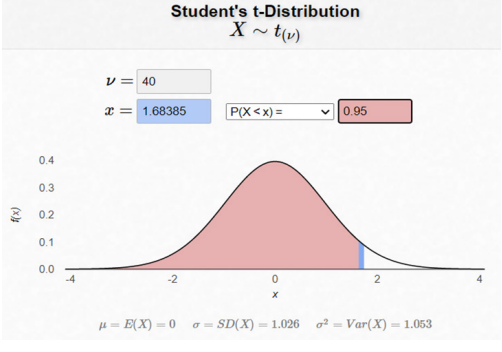


f) El valor de *t* para un nivel de significancia del 95 % y un tamaño de muestra de 41.

$$t_{40;0.95} = 1.684$$

En la calculadora indique los grados de libertad (ν=40), seleccione P(X<x) y coloque 0.95 en la celda rosada. El valor de *t* se muestra en la celda azul.

Nota: este valor también lo puede obtener de la tabla de distribución *t de Student* (ver anexo 2).



3.1.7 La distribución *Ji-cuadrado*

Otra distribución también usada con frecuencia en la Estadística Inferencial es la *Ji-cuadrada(o)* llamada también chi-cuadrada(o) *(χ^2)*. Esta distribución también pertenece a las distribuciones de probabilidad continuas al igual que la normal y la *t de Student* estudiadas previamente.

La distribución de probabilidad *Ji-cuadrada* tiene un parámetro (***k***) que representa los grados de libertad de la variable aleatoria.

La distribución χ^2 fue propuesta por Carl Pearson y tiene múltiples usos, aunque su uso más común es la prueba chi-cuadrado:

- Es usada en el análisis de tablas de contingencia (para realizar prueba de hipótesis de independencia o de homogeneidad).
- En pruebas de bondad de ajuste para comparar la distribución teórica y empírica de un conjunto de datos.
- Para realizar inferencias sobre la varianza poblacional (σ^2).

Esta distribución proviene de la sumatoria de ***k*** variables normales estándar (*Z*) independientes:

$$\chi^2_{(k)} = Z_1^2 + Z_2^2 + \ \ldots + Z_k^2 = \sum Z_i^2$$

Una VA ***X*** con *distribución Ji-cuadrada* con ***k*** grados de libertad se denota como:

$$X \sim \chi^2_{(k)}$$

La Figura 26 muestra funciones de densidad[27] de una distribución *Ji-cuadrada* para diferentes grados de libertad.

27 Su función de densidad es:

$$f(x;k) = \frac{1}{2^{k/2}\Gamma(k/2)\sqrt{\pi k}} x^{k/2} e^{-x/2} \qquad ; x > 0$$

donde Γ es la función gamma.

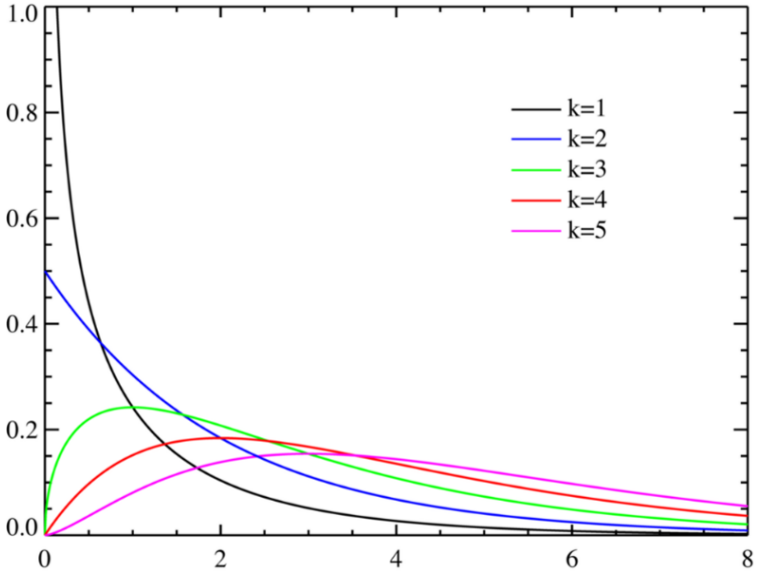


Figura 26. Distribución de probabilidad *Ji-cuadrada* para diversos grados de libertad (k=1,2,3,4,5). Wikipedia contributors. Distribución ?2 [Internet]. Wikipedia, The Free Encyclopedia. Disponible en: https://es.wikipedia.org/w/index.php?title=Distribuci%C3%B3n_%CF%87%C2%B2&oldid=140038719

La distribución χ^2 presenta estas características:

- Son una familia de distribuciones cuya curva varía de acuerdo con los grados de libertad (***k***).
- Solo toma valores positivos (desde cero hasta infinito). Esto porque proviene de VA elevadas al cuadrado.
- No es simétrica.
- Observe que valores grandes de la distribución son los menos frecuentes.

Al igual que en la distribución normal y *t de Student*, para encontrar probabilidades o cuantiles (valores de χ^2) de la distribución se tienen estas opciones:

1. Hallar la probabilidad por medio de una integral de la función de densidad *f(x;k)*.
2. Usar los valores tabulados en una tabla de *distribución Ji-cuadrado* (ver anexo 3).
3. Usar algún programa como Excel u otros como R o Epidat.
4. Usar alguna calculadora online en Internet (recomendado).

Ejemplo 19. Distribución *Ji-cuadrada*

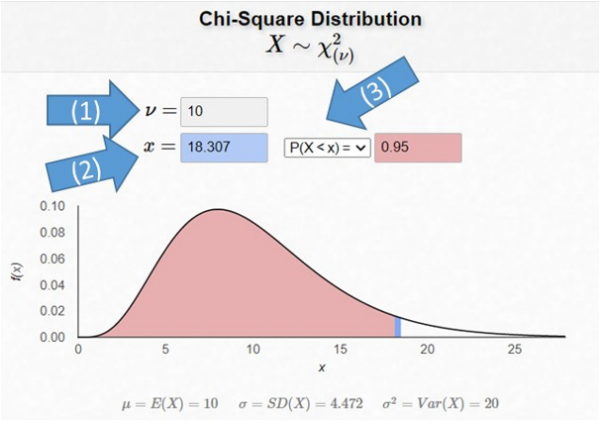
Las siguientes gráficas de la distribución χ^2 se obtuvieron de una calculadora en línea[28]:

a) ¿Cuál es la probabilidad de observar un valor de χ^2 menor que 18,307 cuando se tienen 10 grados de libertad?

$$P(\chi^2_{10} < 18{,}307) = ¿?$$

Para usar la calculadora en línea siga estos pasos como lo muestra la imagen:

1. Especificar los grados de libertad (ν=10)
2. Se debe indicar que el valor de $\boldsymbol{X}$ es 18,307
3. Seleccionar el área menor (<), es decir, P(X<x)

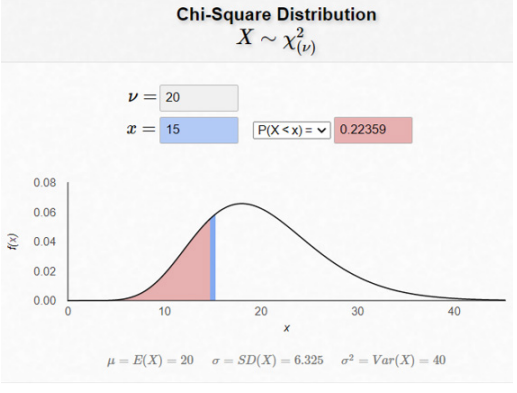


Automáticamente en la celda rosada se calcula la probabilidad. Luego:

$$P(\chi^2_{10} < 18{,}307) = 0.95$$

b) Probabilidad de observar un valor de $\boldsymbol{X}$ menor que 15 en una distribución Ji-cuadrada con 20 grados de libertad.

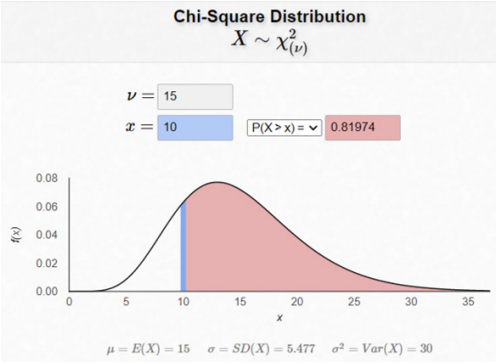
$$P(\chi^2_{20} < 15) = 0.2236$$



28 Chi-Square Distribution Applet/Calculator [Internet]. Uiowa.edu. [citado el 18 de diciembre de 2021]. Disponible en: https://homepage.divms.uiowa.edu/~mbognar/applets/chisq.html

c) Probabilidad de observar un valor de ***X*** mayor que 10 en una distribución *Ji-cuadrada* con 15 grados de libertad.

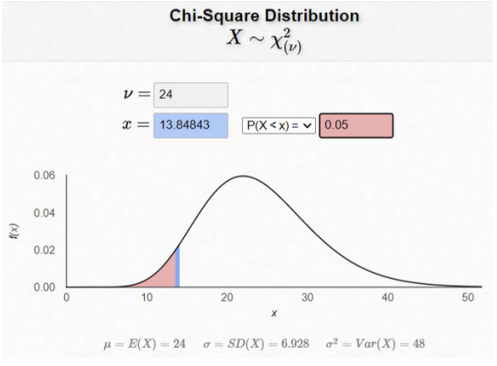
$$P(\chi^2_{15} > 10) = 0.8197$$



d) El valor de ***X*** para un nivel de significancia del 5 % en una distribución *Ji-cuadrada* con 24 grados de libertad.

$$\chi^2_{24;0.05} = 13.8484$$

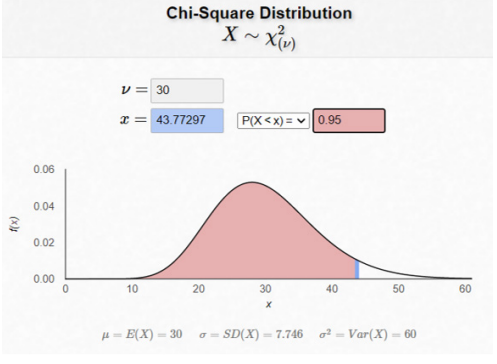
Nota: este valor también lo puede obtener de la tabla de distribución χ^2 (ver anexo 3).



e) El valor de ***X*** para un nivel de significancia del 95 % en una distribución *Ji-cuadrada* con 30 grados de libertad.

$$\chi^2_{30;0.95} = 43.7730$$

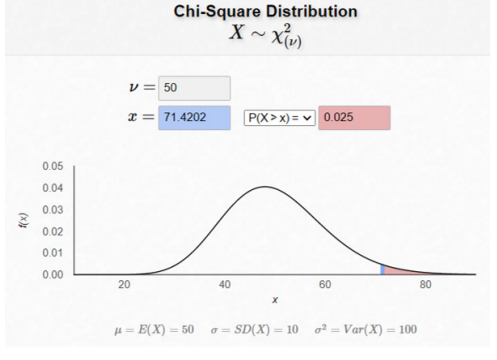
Nota: este valor también lo puede obtener de la tabla de distribución χ^2 (ver anexo 3).



f) En una distribución *Ji-cuadrada* con 50 grados de libertad el valor de ***X*** que deja por encima el 2.5 % de los datos:

$$\chi^2_{50;0.975} = 71.4202$$

Nota: este valor también lo puede obtener de la tabla de distribución χ^2 (ver anexo 3).



Ejemplo 20. Obteniendo probabilidades y cuantiles

Para las siguientes probabilidades y/o cuantiles de las distribuciones de probabilidad continuas (normal estándar, *t de Student* y *Ji-cuadrada*) verificar los valores.

Distribución normal estándar:

a. $P(Z<-1) = 0,1587$
b. $P(Z < 1,64) = 0,9495$
c. $P(Z < 1,96) = 0,975$
d. $P(Z > 1) = 0,1587$
e. $P(Z > 1,64) = 0,0505$
f. $P(-1 < Z < 1) = 0,6827$
g. $P(Z < -1) + P(Z > 1) = 0,3173$

Distribución normal estándar:

a. Ejemplo: $P(Z<-1,08) = 0,14 \rightarrow Z_{0.14} = -1,08$
b. $Z_{0.05} = -1,64$
c. $Z_{0.01} = -2,32$
d. $Z_{0.10} = -1,28$
e. $Z_{0.5} = 0$
f. $Z_{0.90} = 1,28$
g. $Z_{0.95} = 1,64$
h. $Z_{0.975} = 1,96$

Distribución *t de Student*:

a. Ejemplo: $P(t_{10} < 2{,}228) = 0{,}975 \rightarrow t_{10;\,0{,}975} = 2{,}228$

b. $t_{20;\,0.975} = 2{,}086$

c. $t_{20;\,0.025} = -2{,}086$

d. $t_{60;\,0.95} = 1{,}67$

e. $t_{60;\,0.05} = -1{,}67$

f. $t_{100;\,0.95} = 1{,}66$

g. $t_{100;\,0.975} = 1{,}95$

h. $t_{60;\,0.025} = -2$

Distribución *Ji-cuadrado*:

a. Ejemplo: $P\left(\chi^2_{(10)} < 18{,}307\right) = 0{,}95 \rightarrow \chi^2_{10;\,0.95} = 18{,}307$

b. $\chi^2_{20;\,0.95} = 31{,}41$

c. $\chi^2_{10;\,0.975} = 20{,}48$

d. $\chi^2_{30;\,0.90-} = 40{,}26$

e. $\chi^2_{40;\,0.95} = 55{,}76$

f. $\chi^2_{60;\,0.95} = 79{,}08$

g. $\chi^2_{60;\,0.05} = 43{,}18$

h. $\chi^2_{75;\,0.975} = 100{,}84$

3.2 Inferencia sobre parámetros

3.2.1 Parámetros y estimadores

Como se mencionó, la Estadística Inferencial intenta estimar parámetros de la población con base en los estimadores de una muestra.

Un **parámetro** es una característica que se puede medir o cuantificar en una población de estudio. Un **estimador** es una característica que se puede calcular con la ayuda de una muestra. Cada parámetro tiene su estimador y algunos de ellos se presentan en la Tabla 21.

El estimador de la media poblacional μ es la media muestral $\bar{x}$; el estimador de la proporción poblacional P es la proporción muestral $\hat{p}$; el estimador de la diferencia de medias poblacionales ($\mu_1 - \mu_2$) es la diferencia de medias muestrales ($\bar{x}_1 - \bar{x}_2$) y el estimador de la diferencia de proporciones poblacionales ($P_1 - P_2$) es la diferencia de proporciones muestrales ($\hat{P}_1 - \hat{P}_2$), entre otros.

Tabla 21. Resumen de algunos parámetros y estimadores.

Parámetro (símbolo)	Estimador (símbolo)	Comentario
Media poblacional (μ)	$\bar{x}$	Se habla de media o promedio cuando se quiere resumir datos numéricos o cuantitativos en una sola población (estudios descriptivos). Por ejemplo, nivel de plomo, presión sanguínea, peso, talla, etc.
Proporción poblacional (P)	$\hat{p}$	Se habla de proporción o porcentaje cuando se quieren resumir datos cualitativos o categóricos en una sola población (estudios descriptivos). Por ejemplo, la prevalencia de hipertensión o de tabaquismo, etc.
Diferencia de medias ($\mu_1 - \mu_2$)	$\bar{x}_1 - \bar{x}_2$	Cuando se quiere comparar una variable numérica en dos poblaciones independientes (estudios analíticos). Por ejemplo, comparar el nivel promedio de plomo en sangre en mujeres de estratos bajo (μ_1) y alto (μ_2).
Diferencia de proporciones ($P_1 - P_2$)	$\hat{p}_1 - \hat{p}_2$	Cuando se quiere comparar una variable categórica en dos poblaciones independientes (estudios analíticos). Por ejemplo, la proporción de pacientes que se curan tomando un medicamento A (P_1) comparados con los que toman un medicamento B (P_2).

Nota. Para poder conocer un parámetro será necesario hacer un censo de la población de estudio. Para conocer un estimador basta con tomar una "buena muestra" de la población de estudio. Elaboración propia.

En la investigación cuantitativa algunos estudios son realizados con el fin de estimar parámetros de una, dos o más poblaciones, para lo cual se toman muestras que permitan hacer inferencias o generalizaciones a la población de estudio. Es importante identificar el parámetro de interés en una investigación, pues ello determina las herramientas de la estadística descriptiva e inferencial que se deben utilizar durante el análisis estadístico de los datos.

Ejemplo 21. Parámetros y estimadores

a. En el "Ejemplo 1 a)" se deseaba conocer la prevalencia de desnutrición en niños menores de 5 años en la comunidad ABC. Si se mide "Desnutrición" como una variable cualitativa nominal → "Desnutrición: 1. Sí o 0. No" el parámetro de interés en este estudio es una proporción. Si se realiza un censo de los niños menores de 5 años en esta comunidad y se encuentra que el 10 % de ellos están desnutridos, se habría estimado el parámetro proporción poblacional (P=0.10). Si por el contrario, se toma una muestra aleatoria de niños de esta comunidad, por ejemplo 100 niños (n=100) y se encuentra que 9 (x=9) de ellos estaban desnutridos, lo que se ha hecho es una estimación de la proporción poblacional (P) por medio de una proporción muestral ($\hat{p} = x/n$). En este caso $\hat{p}$ =0.09, lo que sugiere que, si se tomara una "buena muestra", el valor del estimador estaría muy cercano al valor del parámetro.

b. En el "Ejemplo 1 b) el interés del investigador era determinar los niveles de plomo en sangre en las mujeres embarazadas en esa misma comunidad. Como esta variable es cuantitativa continúa, el parámetro de interés es una media poblacional (μ). Nuevamente, se realiza un censo de las mujeres embarazadas de esa comunidad y se encuentra que el nivel promedio de plomo en sangre es de 10 $\mu g/dl$ (microgramo por decilitro) se habría estimado el parámetro media poblacional (μ=10 $\mu g/dl$). Si se toma una muestra aleatoria de 50 embarazadas (n=50) y se encuentra que en la muestra el nivel medio de plomo en sangre es de 10,2 $\mu g/dl$, se habría obtenido el estimador de la media poblacional; es decir, $\bar{x}$ = 10,2 $\mu g/dl$. Nuevamente, si la muestra es "buena" el estimador estará cercano al parámetro.

3.2.2 Fundamentos de estadística inferencial

Se debe recordar que un parámetro es una característica que se puede medir o cuantificar en una población de estudio, mientras que un estimador es una característica que se puede medir en una muestra; la finalidad de este último es estimar el parámetro desconocido (Figura 27).

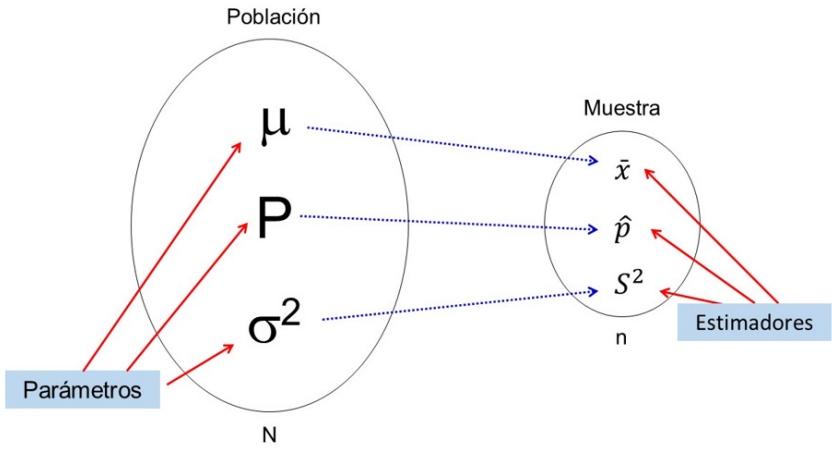


Figura 27. Esquema de la inferencia estadística representando algunos parámetros y estimadores en estudios que envuelven una sola población. Elaboración propia.

En la Figura 27 se representa una población de estudio y una muestra, y además se muestran los parámetros *media poblacional* (μ), *proporción poblacional* (P) y *varianza poblacional* (σ^2), y sus respectivos estimadores: *media muestral* ($\bar{\mathbf{X}}$), *proporción muestral* ($\hat{\mathbf{p}}$) y *varianza muestral* (s^2). Algunos ejemplos de estos parámetros y estimadores se presentaron en la Tabla 21 y en el Ejemplo 21. Además, cada parámetro tiene su estimador como se muestra en la Tabla 22.

Tabla 22. Ejemplos de algunos parámetros y sus estimadores según el tipo de estudio (descriptivo o analítico).

Estudios descriptivos	El estimador de la media poblacional $\boldsymbol{\mu}$ es la media muestral El estimador de la proporción poblacional $\boldsymbol{P}$ es la proporción muestral El estimador del coeficiente de correlación de Pearson ($\boldsymbol{\rho}$) es el coeficiente de correlación muestral ($\boldsymbol{R}$)
Estudios analíticos	El estimador de la diferencia de medias poblacionales ($\boldsymbol{\mu_1 - \mu_2}$) es la diferencia de medias muestrales ($\bar{x}_1 - \bar{x}_2$) El estimador de la diferencia de proporciones poblacionales ($\boldsymbol{P_1 - P_2}$) es la diferencia de proporciones muestrales ($\hat{p}_1 - \hat{p}_2$) El estimador del Riesgo Relativo poblacional ($\boldsymbol{RR}$) es el riesgo relativo estimado ($\widehat{\mathrm{RR}}$) El estimador del Odds Ratio poblacional ($\boldsymbol{OR}$) es el Odds ratio estimado ($\widehat{\mathrm{OR}}$) Entre otros …..

Nota. Elaboración propia.

En la investigación cuantitativa los estudios se realizan con dos finalidades:

a. Estimar uno o varios parámetros de una población.

b. Evaluar hipótesis de investigación que involucran parámetros de uno o más grupos de estudio.

Para esto se toma una "buena muestra" que luego permita hacer inferencias o generalizaciones hacia la población de estudio.

Es importante identificar el parámetro de interés en una investigación, pues ello determina las herramientas inferenciales que se deben utilizar, inicialmente para calcular el tamaño de la muestra y posteriormente para el análisis de los datos.

En la Tabla 21 se presentaron algunos ejemplos de parámetros y estimadores. Otros ejemplos se presentan en la Tabla 23.

Tabla 23. Resumen de algunos parámetros y estimadores.

Parámetro (símbolo)	Estimador (símbolo)	Comentario
Riesgo relativo (RR)	$\widehat{RR}$	El riesgo relativo (RR) es una medida de asociación usada principalmente en estudios longitudinales. Por ejemplo, en un estudio de cohortes para evaluar la asociación entre exposición al humo de tabaco y bajo peso al nacer.
Odds ratio (OR)	$\widehat{OR}$	El Odds Ratio (OR) es una medida de asociación usada en estudios de casos-control o transversales. Por ejemplo, en un estudio de caso-control para evaluar la asociación entre antecedente de migraña y síndrome hipertensivo del embarazo.
Coeficiente de correlación (ρ)	R	El Coeficiente de Correlación de Pearson es un indicador de asociación lineal entre dos variables numéricas. Por ejemplo, en un estudio para evaluar la correlación entre la conciencia sobre el riesgo de cáncer colorrectal (escala) y la edad o el IMC.

Nota. Elaboración propia.

Se usará esta notación (Figura 28):

- El símbolo θ (theta) para denotar cualquier parámetro. Por ejemplo, la media poblacional (μ), la proporción poblacional (P) o la diferencia de medias ($\mu_1 - \mu_2$), etc.
- El símbolo $\hat{\theta}$ (theta estimado) para denotar un estimador. Por ejemplo, la media muestral ($\bar{x}$), la proporción muestral ($\hat{p}$) o la diferencia de medias muestrales ($\bar{x}_1 - \bar{x}_2$), etc.

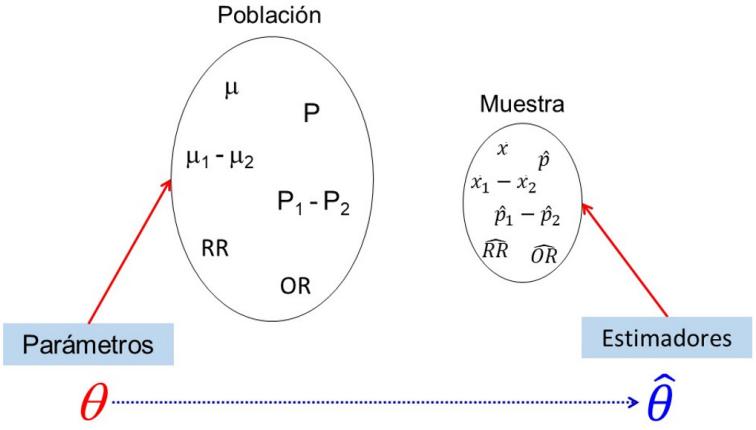


Figura 28. Notación para representar un parámetro (θ) o un estimador ($\hat{\theta}$). Elaboración propia.

Se empezarán las inferencias de los parámetros (θ) recordando que hay dos grandes herramientas de la estadística inferencial, ambas con un objetivo similar y conduciendo a las mismas conclusiones:

1. La estimación.
2. Las pruebas de hipótesis.

Dentro de la estimación del parámetro (θ) se habla una estimación puntual y otra por intervalos (Figura 29).

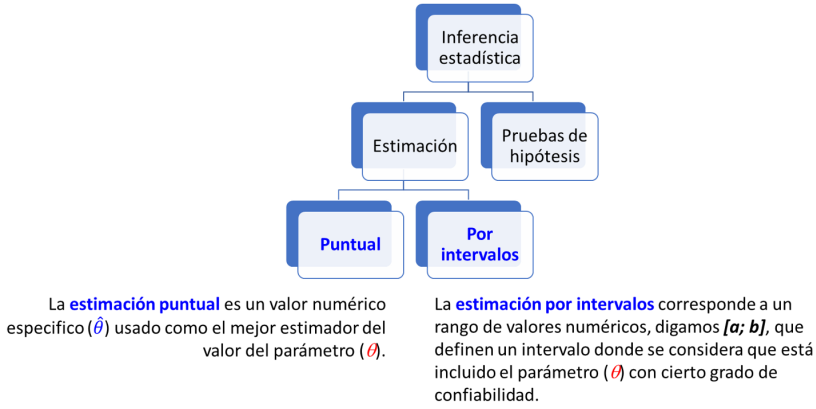


Figura 29. Tipos de estimación en estadística inferencial: puntual y por intervalos. Elaboración propia.

La Tabla 24 muestra un resumen de los parámetros más utilizados en investigación en salud y sus estimadores puntuales.

Tabla 24. Resumen de algunos parámetros y estimadores puntuales.

Definición del parámetro	Parámetro (θ)	Estimador puntual ($\hat{\theta}$)
Media	μ	$\bar{x}$
Diferencia de medias	$\mu_1 - \mu_2$	$\bar{x}_1 - \bar{x}_2$
Proporción	P	$\hat{p}$
Diferencia de proporciones	$P_1 - P_2$	$\hat{p}_1 - \hat{p}_2$
Riesgo relativo	RR	$\widehat{RR}$
Odds ratio	OR	$\widehat{OR}$
Coeficiente de correlación	ρ	R

Nota. Elaboración propia.

La estimación por intervalos, llamados intervalos de confianza (IC) para θ, entregan un rango de valores [a; b] donde se espera que esté contenido el parámetro (θ) con alto grado de confiabilidad (1-α) (Ver Figura 30). Significa que, si se repitiera el muestreo muchas veces, aproximadamente el (1-α)% de los intervalos construidos incluirían el parámetro (θ). Note que (1-α) está asociado al nivel de confianza en la estimación de un parámetro (o probabilidad que el intervalo contenga el parámetro desconocido), mientras que α se relaciona con la probabilidad de equivocarnos en la estimación. Es muy común construir intervalos de confianza del 95 % (lo que conduce a una probabilidad de error del 5 %, α=0,05), aunque también se pueden usar otros niveles de confianza como del 90 % (α=0,10) o del 99 % (α=0,01), entre otros.

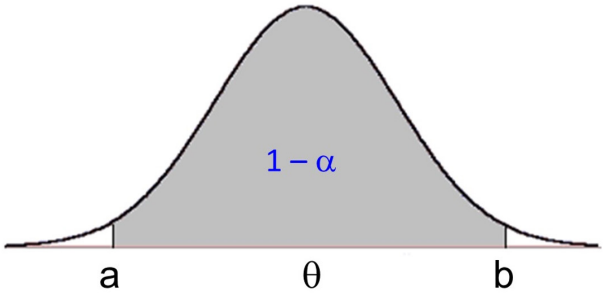


Figura 30. Representación de un intervalo de confianza [a; b] donde se tiene una confiabilidad del (1-α)% que en este rango de valores está contenido el parámetro (θ). Elaboración propia.

3.2.3 Distribuciones muestrales

La construcción de los intervalos de confianza para los parámetros (θ) requiere el estudio de la distribución muestral de los estimadores ($\hat{\theta}$)[29].

Por ejemplo, para la estimación de la media poblacional (μ) por medio de intervalos, se requiere conocer la distribución muestral de la media muestral ($\bar{X}$).

Vamos a explicar la distribución muestral de $\bar{X}$ y después generalizamos el concepto para la distribución muestral de los otros estimadores.

Distribución muestral de $\bar{X}$

La distribución muestral de una estadística, en este caso la media muestral ($\bar{X}$), se refiere a la distribución de todos los valores posibles que puede asumir una estadística, calculados a partir de muestras del mismo tamaño y extraídas aleatoriamente de la misma población.

Ejemplo 22. Distribución muestral de la media muestral

Suponga una población de estudio conformada por cinco niños (N=5). Suponga que las edades (en años) de los cinco niños son las siguientes:

$$x_1 = 6 \qquad x_2 = 8 \qquad x_3 = 10 \qquad x_4 = 12 \qquad x_5 = 14$$

La Figura 31 muestra la distribución de la edad, la cual está lejos de tener una distribución normal.

29 Sin embargo, también existen alternativas computacionales basadas en simulación y métodos de remuestro (por ejemplo, el *bootstrapping* o *bootstrap*) que permiten obtener intervalos de confianza sin necesidad de tener explicita su distribución.

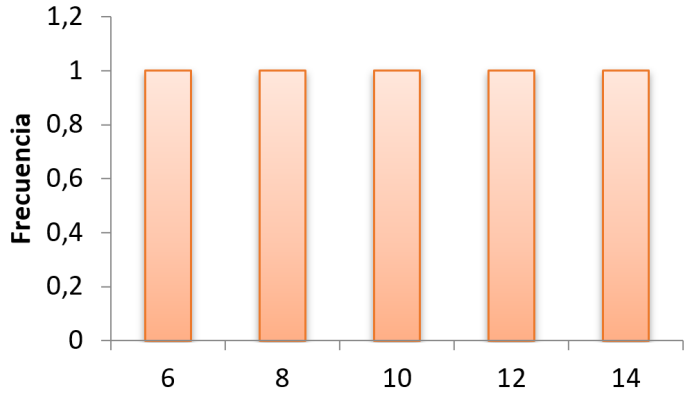


Figura 31. Distribución de las edades de cinco niños (*N*=5). Elaboración propia.

La media (μ), la varianza (σ^2) y la desviación estándar (σ) de la edad calculados a partir de estos datos corresponden a parámetros (porque son obtenidos de toda la población).

La media poblacional es:

$$\mu = \frac{\sum_{i=1}^{n} x_i}{N} = \frac{6 + 8 + 10 + 12 + 14}{5} = 10 \text{ años}$$

La varianza poblacional es:

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{N} = \frac{40}{5} = 8 \text{ años}^2$$

Y la desviación estándar poblacional es: $\sigma = \sqrt{\sigma^2} = \sqrt{8} = 2.83$ años

Para construir la distribución muestral de la media muestral ($\bar{X}$) se deben extraer todas las muestras posibles; por ejemplo, de tamaño ***n=2***, a partir de esta población.

Si el muestreo se realiza con reemplazo (es decir que los valores se pueden repetir) se tendrían (**N**×**N**), es decir, 5x5=25 diferentes muestras.

Las 25 diferentes muestras, junto con sus medias muestrales ($\bar{X}$), se muestran en la Tabla 25.

Tabla 25. Todas las posibles muestras de tamaño *n*=2 y sus medias muestrales.

# muestra	Muestra		$\bar{X}$
1	6	6	6
2	6	8	7
3	6	10	8
4	6	12	9
5	6	14	10
6	8	6	7
7	8	8	8

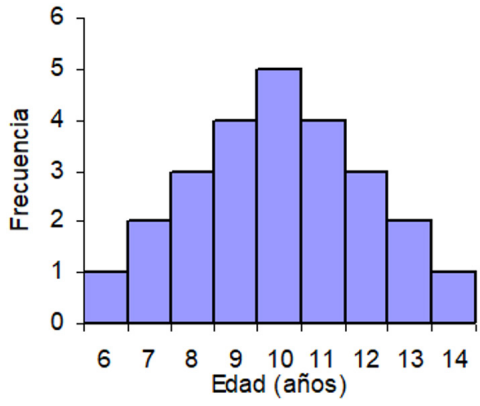
8	8	10	9
9	8	12	10
10	8	14	11
11	10	6	8
12	10	8	9
13	10	10	10
14	10	12	11
15	10	14	12
16	12	6	9
17	12	8	10
18	12	10	11
19	12	12	12
20	12	14	13
21	14	6	10
22	14	8	11
23	14	10	12
24	14	12	13
25	14	14	14

Nota. Elaboración propia.

Para observar la forma de distribución de $\bar{X}$ se construye su distribución de frecuencia y/o su gráfica. La Tabla 26 muestra en la primera columna los diferentes valores que tomó la media muestral, luego la frecuencia con la que ocurrió ese valor y finalmente su frecuencia relativa. La figura del lado muestra la distribución muestral de $\bar{X}$ que se asemeja bastante a una distribución normal.

Tabla 26 . Distribución muestral de para muestras de tamaño *n*=2.

$\bar{X}$	Frecuencia	Frecuencia relativa
6	1	1/25
7	2	2/25
8	3	3/25
9	4	4/25
10	5	5/25
11	4	4/25
12	3	3/25
13	2	2/25
14	1	1/25
Total	25	25/25



Nota. Elaboración propia.

Cuando se calcula la media de todas las posibles muestras de la Tabla 25 se obtiene:

$$\mu_{\bar{x}} = \frac{\sum_{i=1}^{25} \bar{x}_i}{N^n} = \frac{6 + 7 + 7 + \cdots .. + 14}{25} = 10 \text{ años}$$

La varianza de todas las posibles muestras de la Tabla 25 es:

$$\sigma^2{}_{\bar{x}} = \frac{\sum_{i=1}^{25} (\bar{x}_i - \mu_{\bar{x}})^2}{N^n} = \frac{(6-10)^2 + (7-10)^2 + \cdots . + (14-10)^2}{25} = \frac{100}{25} = 4$$

Y la desviación estándar de todas las posibles muestras (conocido como error estándar de la media) es 2 años ($\sigma_{\bar{x}} = \sqrt{\sigma^2{}_{\bar{x}}} = \sqrt{4} = 2$).

Resultados importantes sobre la distribución muestral de $\bar{X}$:

1. La media de todas las medias es igual a la media poblacional: $\mu_{\bar{x}} = \mu$
2. La varianza de la distribución muestral de $\bar{X}$ es igual a la varianza de la población dividida entre el tamaño de la muestra:

$$\sigma^2{}_{\bar{x}} = \frac{\sigma^2}{n} = \frac{8}{2} = 4$$

3. La desviación estándar de la distribución muestral de $\bar{X}$ es:

$$\sigma_{\bar{x}} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$$

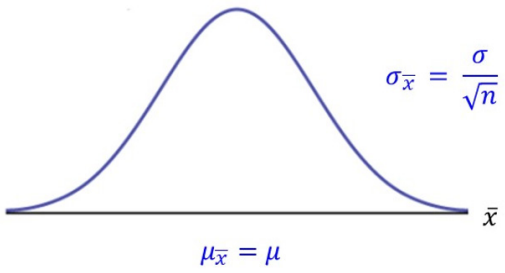
Error estándar de la media

En resumen, la distribución muestral de $\bar{X}$ es aproximadamente normal con:

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

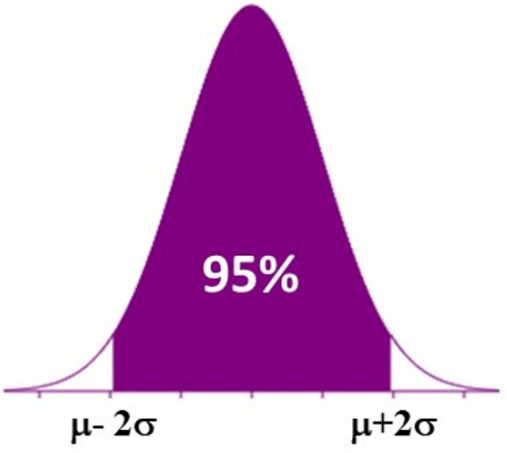
Desviación estándar (error estándar): $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

Esto significa que la distribución muestral de $\bar{X}$ cumple con las características de una curva normal.



Conclusión:

Se debe recordar que la regla empírica mencionada previamente establece que, en cualquier distribución normal, aproximadamente el 95 % de las observaciones se encuentran a dos desviaciones estándares de la media ($\mu \pm 2\sigma$).

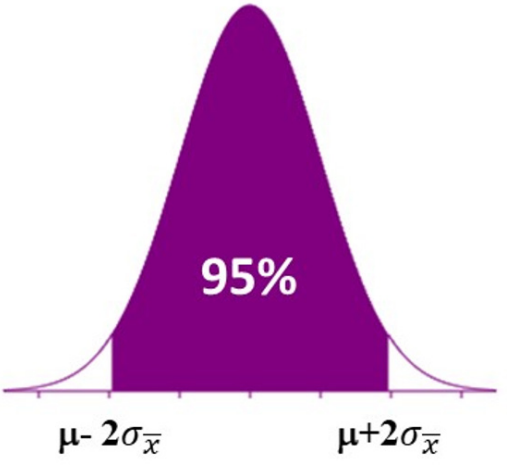


Luego, para la distribución muestral de $\bar{X}$ el intervalo que contiene el 95 % (IC95 %) de las observaciones es: $\mu_{\bar{x}} \pm 2\ \sigma_{\bar{x}}$.

Estos resultados son útiles para la construcción de los intervalos de confianza de los parámetros.

La estructura general de los IC95 % es:

$$\text{Estimador puntual} \pm 2 \times \text{Error estándar}$$



De igual forma, como se obtuvo la distribución muestral de $\bar{X}$ se puede construir la distribución muestral para otros estimadores (Tabla 27).

Tabla 27. Resumen de las distribuciones muestrales de algunos parámetros.

Definición del parámetro	Parámetro (θ)	Estimador puntual ($\hat{\theta}$)	Distribución muestral aproximadamente normal	
			Media	Desviación estándar (error estándar)
Media	μ	$\bar{x}$	$\mu_{\bar{x}} = \mu$	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$
Proporción	P	$\hat{p}$	$\mu_{\hat{p}} = P$	$\sigma_{\hat{p}} = \sqrt{\frac{P(1-P)}{n}}$
Diferencia de medias	$\mu_1 - \mu_2$	$\bar{x}_1 - \bar{x}_2$	$\mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2$	$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
Diferencia de proporciones	$P_1 - P_2$	$\hat{p}_1 - \hat{p}_2$	$\mu_{\hat{p}_1 - \hat{p}_2} = P_1 - P_2$	$\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{P_1(1-P_1)}{n_1} + \frac{P_2(1-P_2)}{n_2}}$

Nota. Bioestadística: Base para el análisis de las ciencias de la salud. Cuarta edición. Editorial Limusa; Wayne D. 2002 (1).

3.2.4 Estimación por intervalos de confianza

Conociendo la distribución muestral de algunos estimadores (Tabla 27), entonces el intervalo de confianza (IC) para un parámetro θ corresponden al rango de valores **[a; b]** donde se espera que esté contenido el parámetro (θ) con alta probabilidad (1-α) (Figura 32).

Denotaremos:

1-α: Nivel de confianza, nivel de confiabilidad.

α: Nivel de significancia, nivel de significación, error tipo I.

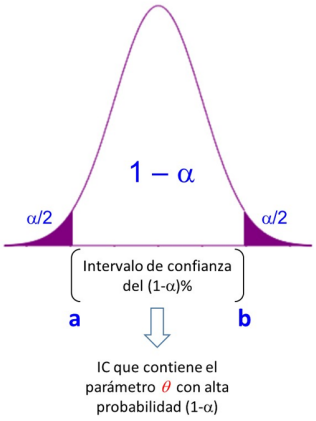


Figura 32. Concepto y estructura de un intervalo de confianza para θ. Gráficas realizadas en Calculadoras Online [Internet]. Calculadoras Online. 2018 [citado el 18 diciembre 2021]. Disponible en: https://calculadorasonline.com/

Estructura de los intervalos de confianza

Usando el resultado de la distribución muestral de la media ($\mu_{\bar{x}} \pm 2\ \sigma_{\bar{x}}$) se encuentra que, en general, todos los intervalos de confianza se construyen partiendo del estimador puntual, al cual se le suma y luego se le resta un margen de error (coeficiente de confiabilidad × el error estándar del estimador).

A continuación (Tabla 28) se presenta un resumen de las fórmulas para **9 casos** diferentes.

Tabla 28. Resumen de intervalos de confianza de algunos parámetros.

Caso	Parámetro (θ)	Intervalo de confianza del (1-α)% para θ	Supuestos
1	Media (μ)	$\bar{x} \pm Z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$	Variable con distribución normal y σ conocida; o tamaño de muestra grande y σ desconocida ($s=\sigma$)
2	Media (μ)	$\bar{x} \pm t_{1-\alpha/2;n-1}\frac{s}{\sqrt{n}}$	Variable con distribución normal, σ desconocida (usar "s") y n (pequeño)
3	Proporción (P)	$\hat{p} \pm Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$	La distribución de $\hat{p}$ será aproximadamente normal cuando $n\hat{p} > 5$ y $n(1-\hat{p}) > 5$
4	Diferencia de medias ($\mu_1 - \mu_2$)	$(\bar{x}_1 - \bar{x}_2) \pm Z_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$	Poblaciones con distribución normal y varianzas poblacionales conocidas
5	Diferencia de medias ($\mu_1 - \mu_2$)	$(\bar{x}_1 - \bar{x}_2) \pm t_{1-\frac{\alpha}{2};n_1+n_2-2}\sqrt{\frac{s_p^2}{n_1}+\frac{s_p^2}{n_2}}$ donde: $s_p^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$	Poblaciones con distribución normal, varianzas poblacionales desconocidas pero iguales
6	Diferencia de medias ($\mu_1 - \mu_2$)	$(\bar{x}_1 - \bar{x}_2) \pm t'_{1-\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$ donde: $t'_{1-\frac{\alpha}{2}} = \frac{w_1 t_1 + w_2 t_2}{w_1 + w_2}$ $w_1 = \frac{s_1^2}{n_1} \quad ; \quad w_2 = \frac{s_2^2}{n_2}$	Poblaciones con distribución normal, varianzas poblacionales desconocidas pero diferentes
7	Diferencia de proporciones ($P_1 - P_2$)	$(\hat{P}_1 - \hat{P}_2) \pm Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{P}_1(1-\hat{P}_1)}{n_1}+\frac{\hat{P}_2(1-\hat{P}_2)}{n_2}}$	n_1 y n_2 (grande) P_1 y P_2 no cercano a (0,1)

<table>
<tr>
<td>8</td>
<td>Riesgo relativo (RR)</td>
<td>$\left[\widehat{RR}e^{\left[Z_{\frac{\alpha}{2}} * EE\right]}; \widehat{RR}e^{\left[Z_{1-\frac{\alpha}{2}} * EE\right]}\right]$</td>
<td>$EE_{\widehat{RR}} = \sqrt{\frac{1}{a} - \frac{1}{a+b} + \frac{1}{c} - \frac{1}{c+d}}$
donde:
<table>
<tr><th>Factor</th><th>Enfermos</th><th>Sanos</th><th>Total</th></tr>
<tr><td>Expuestos</td><td>a</td><td>b</td><td>a+b</td></tr>
<tr><td>No expuestos</td><td>c</td><td>d</td><td>c+d</td></tr>
<tr><td>Total</td><td>a+c</td><td>b+d</td><td>n</td></tr>
</table>
</td>
</tr>
<tr>
<td>9</td>
<td>Odds ratio (OR)</td>
<td>$\left[\widehat{OR}e^{\left[Z_{\frac{\alpha}{2}} * EE\right]}; \widehat{OR}e^{\left[Z_{1-\frac{\alpha}{2}} * EE\right]}\right]$</td>
<td>$EE_{\widehat{OR}} = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$
donde:
<table>
<tr><th>Factor</th><th>Casos</th><th>Controles</th><th>Total</th></tr>
<tr><td>Expuestos</td><td>a</td><td>b</td><td>a+b</td></tr>
<tr><td>No expuestos</td><td>c</td><td>d</td><td>c+d</td></tr>
<tr><td>Total</td><td>a+c</td><td>b+d</td><td>n</td></tr>
</table>
</td>
</tr>
</table>

Nota. Bioestadística: Base para el análisis de las ciencias de la salud. Cuarta edición. Editorial Limusa; Wayne D. 2002 (1).

Observe que las fórmulas para los casos ❶ y ❷ se usan para estimar el intervalo de confianza de un promedio (μ) y la fórmula ❸ en el caso de una proporción (P).

La diferencia fundamental entre las fórmulas ❶ y ❷ son los supuestos que se deben verificar. En la práctica para estimar el parámetro *media* (μ) se utiliza la fórmula ❷ principalmente porque la desviación estándar poblacional (σ) pocas veces es conocida.

De igual forma para los intervalos de confianza de la diferencia de medias ($\mu_1 - \mu_2$) se tienen las fórmulas de los casos ❹, ❺ y ❻. Su elección dependerá del cumplimiento de los supuestos en cada caso.

Ejemplo 23. Intervalo de confianza para la media poblacional

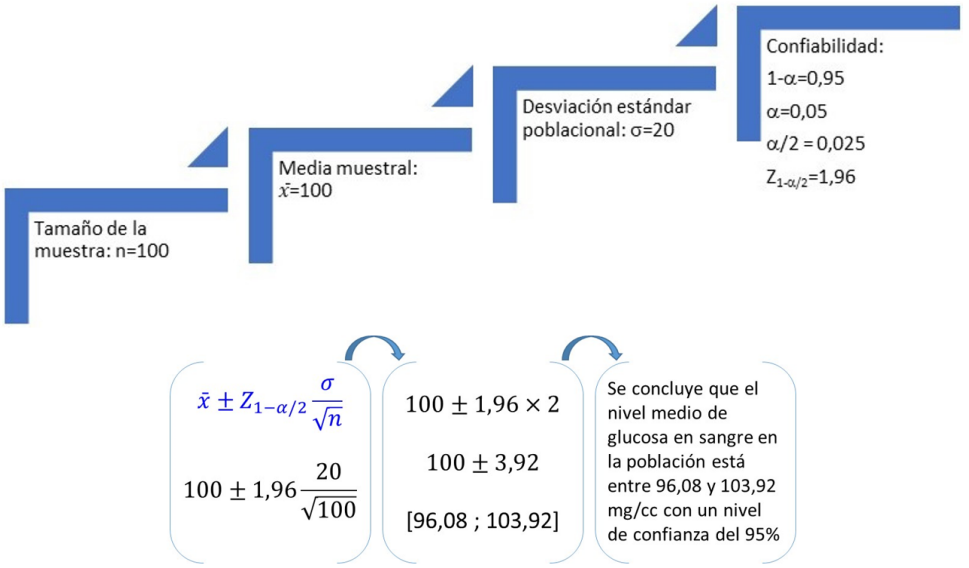
Suponga que se toma una muestra aleatoria de 100 personas a quienes se les mide el nivel de glucosa en sangre, encontrando una media muestral de 100 mg/cc. Si la desviación estándar del nivel de glucosa en sangre en la población es de 20 mg/cc, calcule el intervalo de confianza del 95 % para el nivel de glucosa en sangre en la población.

Solución:

Siempre es importante identificar el parámetro en estudio. En este caso corresponde a la media poblacional (μ). Sabiendo que la desviación estándar es conocida (σ=20 mg/cc) se utiliza la fórmula del caso 1:

$$\bar{x} \pm Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

Resumen de la información disponible:



Ejemplo 24. Intervalo de confianza para la proporción poblacional

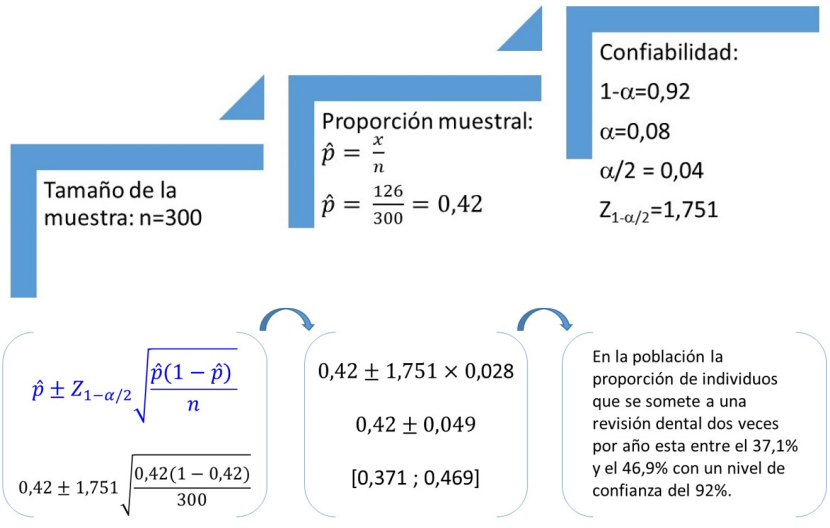
Se llevó a cabo una encuesta para estudiar los hábitos y actitud hacia la salud dental de cierta población urbana de adultos. De los 300 adultos entrevistados, 126 de ellos dijeron que se sometían regularmente a una revisión dental dos veces por año. Se desea construir un intervalo de confianza del 92 % (IC92 %) para la proporción de individuos de la población que se somete a una revisión dental dos veces por año.

Solución:

En este ejemplo el parámetro en estudio es una proporción poblacional (***P***). Para obtener el IC92 % para ***P*** se utiliza la formula del caso ❸:

$$\hat{p} \pm Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Resumen de la información disponible:



3.2.5 Ejercicios propuestos

Ejercicio 10.

Suponga que un investigador está interesado en obtener una estimación del nivel promedio de alguna enzima en cierta población de seres humanos (1). El investigador toma una muestra de 60 individuos, determina el nivel de la enzima en cada uno de ellos y calcula una media muestral de 22. Además, sabe que la variable de interés sigue una distribución aproximadamente normal con una varianza de 45. Calcule e interprete el intervalo de confianza del 90 % para el nivel promedio de la enzima en la población.

Ejercicio 11.

Suponga que se realizó un estudio con el fin de estimar la prevalencia de tabaquismo y explorar factores que influyen en el uso de tabaco en escolares de cinco ciudades de Colombia. La Encuesta Mundial de Tabaquismo en Jóvenes fue implementada por el Instituto Nacional de Cancerología entre junio y diciembre de 2007, aplicada en los

colegios públicos y privados de cinco ciudades de Colombia (Bogotá, Bucaramanga, Cali, Manizales y Valledupar), en los grados 7 a 10, e incluyó estudiantes de 13 a 15 años. La selección de la muestra en cada ciudad se hizo mediante un muestreo probabilístico (bietápico y por conglomerados). Los colegios (conglomerados) seleccionados aleatoriamente fueron visitados para presentar el estudio y solicitar la autorización. En cada colegio aleatoriamente se seleccionaron los grados de interés y los estudiantes para aplicar la encuesta que fue anónima, voluntaria y autodiligenciada por los participantes. La muestra final quedó conformada por 150 estudiantes de los cuales 42 reportaron haber fumado cigarrillo alguna vez en su vida. Construya el intervalo de confianza del 96 % para la prevalencia de tabaquismo en la población.

3.2.6 Pruebas de hipótesis

Cabe resaltar que la inferencia estadística clásica tiene dos grandes herramientas para hacer inferencias sobre los parámetros (θ):

1. La estimación.
2. Las pruebas de hipótesis.

Luego de haber estudiado la estimación (puntual y por intervalos) se introducirán las pruebas de hipótesis sobre un parámetro (θ).

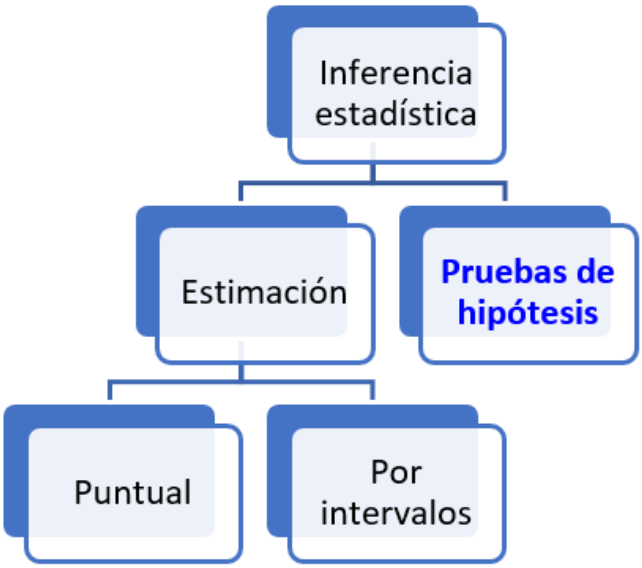


Tabla 29. Parámetros de la población.

Una **hipótesis** es una afirmación o proposición acerca de la población; es decir, **acerca de los parámetros** de la población.

Definición del parámetro	Parámetro (θ)
Media	μ
Diferencia de medias	$\mu_1 - \mu_2$
Proporción	P
Diferencia de proporciones	$P_1 - P_2$
Riesgo relativo	RR
Odds ratio	OR
Coeficiente de correlación	ρ

Ejemplo 25. Hipótesis sobre algunos parámetros

Tabla 30. Hipótesis sobre algunos parámetros.

Parámetro	Hipótesis
Media (μ)	El tiempo medio de alivio del dolor de un medicamento es mayor a 240 minutos.
Proporción (P)	La prevalencia de complicaciones en pacientes sometidas a cesárea en el último año es superior al 5%.
Diferencia de medias ($\mu_1 - \mu_2$)	Un nuevo tratamiento permite un mayor control de la presión arterial sistólica que el medicamento estándar.
Diferencia de proporciones ($P_1 - P_2$)	La eficacia de cura de un nuevo medicamento puede ser mayor que la del tratamiento estándar.
Odds ratio (OR)	La oportunidad de cáncer de cuello uterino es mayor en las mujeres que consumen anticonceptivos orales en comparación con las que no los usan.

La estadística inferencial tiene una amplia gama de pruebas de hipótesis para diferentes objetivos (Figura 33).

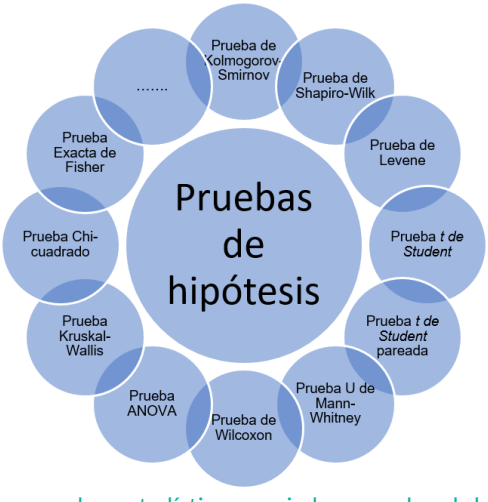


Figura 33. Resumen de algunas pruebas estadísticas asociadas a pruebas de hipótesis. Elaboración propia.

Por ahora, es necesario familiarizarse con el concepto de "**prueba de hipótesis**" y sus elementos.

Siempre que usted vea "**valor de P o valor P**" es porque se ha realizado una prueba de hipótesis.

Elementos de una prueba de hipótesis

Toda prueba de hipótesis tiene estos seis elementos:

1. Las hipótesis sobre el parámetro (θ):
   - *Ho*: → Es la hipótesis nula.
   - *Ha*: → Es la hipótesis alterna o alternativa.
2. Un nivel de significancia (α).
3. Un estadístico de prueba.
4. Una región de rechazo o un valor de P.
5. Una decisión.
6. Una conclusión.

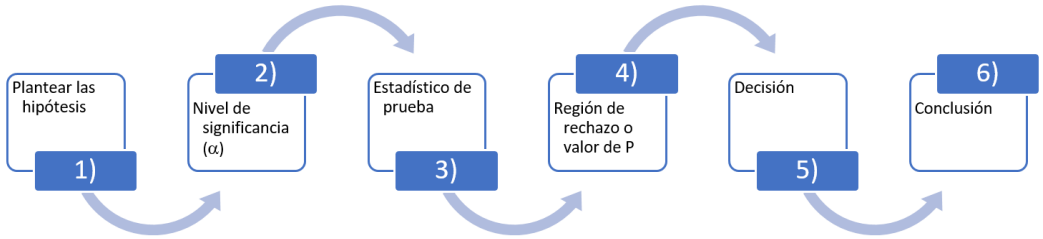
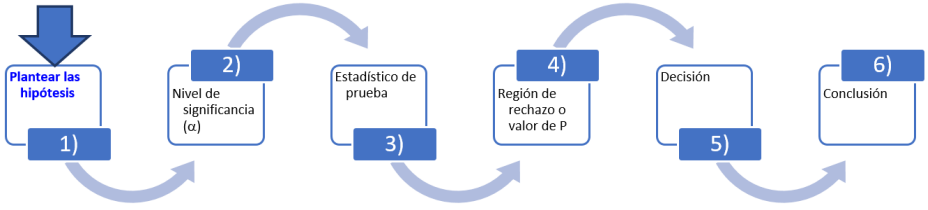


Figura 34. Elementos de una prueba de hipótesis.Elaboración propia.

Las pruebas de hipótesis son herramientas de investigación científica. Al inicio de un estudio el investigador plantea unas hipótesis de investigación (que pueden ser o no ser verdaderas), toma una muestra de datos para contrastar la realidad con sus hipótesis, compara los resultados observados de la muestra con sus hipótesis y toma una decisión en cuanto a sus hipótesis planteadas.



1. Plantear las hipótesis sobre el parámetro (θ):

- *Ho*: → Es la hipótesis nula.
- *Ha*: → Es la hipótesis alterna o alternativa.

Lo primero que se plantea es la **hipótesis de investigación (*Ha*).** Por ejemplo, si el investigador sospecha que el tiempo medio de alivio del dolor de un medicamento es mayor a 240 minutos, entones la hipótesis alterna es:

$$Ha: \mu > 240$$

La hipótesis alterna (*Ha*) es la hipótesis de investigación.

La **hipótesis nula (*Ho*)** es el complemento de *Ha*, es decir que *Ho* queda definida inmediatamente se plantee la hipótesis de investigación (*Ha*). Para el ejemplo, las hipótesis son:

$$Ho: \mu \leq 240$$

$$Ha: \mu > 240$$

La hipótesis nula se mantiene hasta que la evidencia sea contundente y permita su rechazo.

En el paso 5 la hipótesis nula (*Ho*) es la que se rechaza o no se rechaza.

La hipótesis nula (*Ho*) es una hipótesis de igualdad, no diferencia o no asociación.

Por ejemplo, si en un estudio analítico se cree que un nuevo medicamento B es más efectivo (mayor cura) que el estándar (A), las hipótesis serían:

$$Ho: P_B \leq P_A$$

$$Ha: P_B > P_A$$

Por esto se dice que la hipótesis nula es de igualdad, de no diferencia o de no asociación. La hipótesis nula siempre lleva implícito el igual (=, ≤ o ≥).

De acuerdo con el interés del investigador, las hipótesis se pueden plantear en un solo sentido (unilateral) o en ambos sentidos (bilateral).

Por ejemplo:

Bilateral (two-tailed)	Unilateral (one-tailed)	Unilateral (one-tailed)
•*Ho*: $\mu=\mu_0$ •*Ha*: $\mu\neq\mu_0$	•*Ho*: $\mu\geq\mu_0$ •*Ha*: $\mu<\mu_0$	•*Ho*: $\mu\leq\mu_0$ •*Ha*: $\mu>\mu_0$
En este caso el investigador está interesado en comprobar si un promedio es igual o no a un valor especifico (μ_0). Por ejemplo que el promedio de alivio del dolor es igual o no a 240 minutos. No le interesa el sentido. *Ho:* $\mu = 240$ *Ha:* $\mu \neq 240$	Aquí el investigador desea probar que un promedio es menor a un valor especifico (μ_0). Por ejemplo que el promedio de alivio del dolor es menor a 240 minutos. *Ho:* $\mu \geq 240$ *Ha:* $\mu < 240$	Y en este caso el interés es probar que un promedio es mayor a un valor especifico (μ_0). Por ejemplo que el promedio de alivio del dolor es mayor a 240 minutos. *Ho:* $\mu \leq 240$ *Ha:* $\mu > 240$

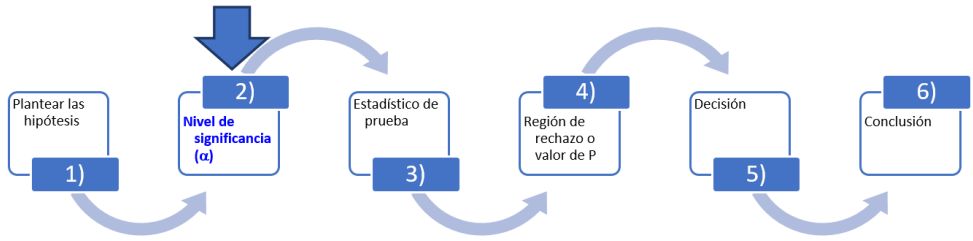
Figura 35. Planteamiento de una hipótesis. Elaboración propia

Algunos ejemplos de pruebas de hipótesis se presentan en la Tabla 31.

Tabla 31. Ejemplos de pruebas de hipótesis bilaterales y unilaterales para algunos parámetros.

Parámetro	Hipótesis bilateral	Hipótesis unilateral (Ha: <)	Hipótesis unilateral (Ha: >)
Media (μ)	*Ho:* $\mu = \mu_0$ *Ha:* $\mu \neq \mu_0$	*Ho:* $\mu \geq \mu_0$ *Ha:* $\mu < \mu_0$	*Ho:* $\mu \leq \mu_0$ *Ha:* $\mu > \mu_0$
Proporción (P)	*Ho:* $P = P_0$ *Ha:* $P \neq P_0$	*Ho:* $P \geq P_0$ *Ha:* $P < P_0$	*Ho:* $P \leq P_0$ *Ha:* $P > P_0$
Diferencia de medias ($\mu_1 - \mu_2$)	*Ho:* $\mu_1 = \mu_2$ *Ha:* $\mu_1 \neq \mu_2$	*Ho:* $\mu_1 \geq \mu_2$ *Ha:* $\mu_1 < \mu_2$	*Ho:* $\mu_1 \leq \mu_2$ *Ha:* $\mu_1 > \mu_2$
Diferencia de proporciones ($P_1 - P_2$)	*Ho:* $P_1 = P_2$ *Ha:* $P_1 \neq P_2$	*Ho:* $P_1 \geq P_2$ *Ha:* $P_1 < P_2$	*Ho:* $P_1 \leq P_2$ *Ha:* $P_1 > P_2$
Riesgo relativo (RR)	*Ho:* $RR = 1$ *Ha:* $RR \neq 1$	*Ho:* $RR \geq 1$ *Ha:* $RR < 1$	*Ho:* $RR \leq 1$ *Ha:* $RR > 1$
Odds ratio (OR)	*Ho:* $OR = 1$ *Ha:* $OR \neq 1$	*Ho:* $OR \geq 1$ *Ha:* $OR < 1$	*Ho:* $OR \leq 1$ *Ha:* $OR > 1$

Nota. Elaboración propia.



El **nivel se significancia (*α*)** corresponde al nivel de error que se aceptaría cometer en la prueba. Se debe reconocer que se están trabajando con muestras y es factible que se puede cometer un error en el proceso de pruebas de hipótesis.

El nivel de significancia (α) se establece desde que se calcula el tamaño de muestra. Regularmente es de 0,05, pero puede tomar cualquier valor menor o igual a 0,10.

Recuerde que el nivel de confianza y el nivel de significancia suma 1 o 100 % (Figura 34).

1-*α*: Nivel de confianza, nivel de confiabilidad

α: Nivel de significancia, nivel de significación, error tipo I

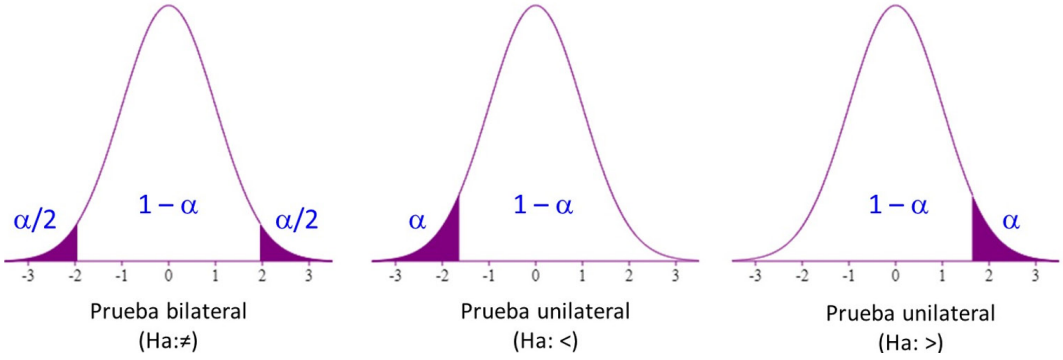
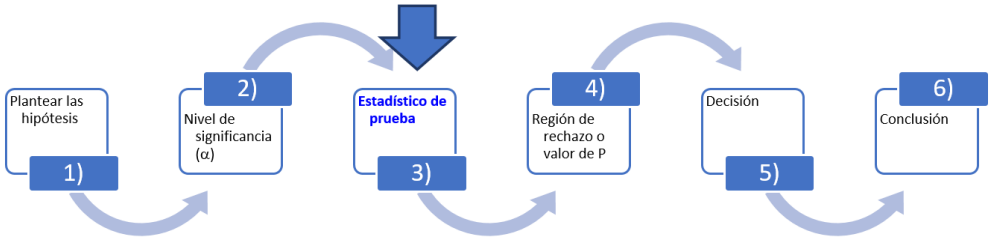


Figura 36. Representación del nivel de significancia para pruebas de hipótesis bilaterales y unilaterales. Gráficas realizadas en Calculadoras Online [Internet]. Calculadoras Online. 2018 [citado el 15 de diciembre de 2021]. Disponible en: https://calculadorasonline.com/



El **estadístico de prueba** juega un papel muy importante en una prueba de hipótesis. Este se calcula con los datos de la muestra y es el que brinda información sobre la concordancia entre los datos observados y la hipótesis nula.

Cada prueba de hipótesis tiene su propio estadístico de prueba (por ejemplo, z_c o t_c, léase z calculado y t calculado, respectivamente) y ellos siguen una distribución de probabilidad conocida. En la Tabla 32 se muestra un resumen de los estadísticos de prueba.

Tabla 32. Resumen de estadísticos de prueba de algunos parámetros.

Caso	Parámetro (θ)	Estadístico de prueba	Supuestos
1	Media (μ)	$z_c = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$	Variable con distribución normal y σ conocida; o tamaño de muestra grande y σ desconocida (s=σ)
2	Media (μ)	$t_c = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim t_{(n-1)}$	Variable con distribución normal, σ desconocida (usar "s") y n (pequeño)
3	Proporción (P)	$z_c = \frac{\hat{p} - P_0}{\sqrt{\frac{P_0(1-P_0)}{n}}} \sim N(0,1)$	La distribución de $\hat{p}$ será aproximadamente normal cuando $n\hat{p} > 5$ y $n(1-\hat{p}) > 5$
4	Diferencia de medias ($\mu_1 - \mu_2$)	$z_c = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$	Poblaciones con distribución normal y varianzas poblacionales conocidas
5	Diferencia de medias ($\mu_1 - \mu_2$)	$t_c = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} \sim t_{(n_1+n_2-2)}$ donde: $s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}$	Poblaciones con distribución normal, varianzas poblacionales desconocidas pero iguales. Esta prueba estadística se conoce como la *t de Student* para muestras independientes asumiendo varianzas iguales.
6	Diferencia de medias ($\mu_1 - \mu_2$)	$t'_c = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t_{(gl)}$ donde: $gl = \frac{\left[\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right]^2}{\frac{1}{n_1-1}\left[\frac{s_1^2}{n_1}\right]^2 + \frac{1}{n_2-1}\left[\frac{s_2^2}{n_2}\right]^2}$	Poblaciones con distribución normal, varianzas poblacionales desconocidas pero diferentes. Esta prueba estadística se conoce como la *t de Student* para muestras independientes asumiendo varianzas diferentes.

7	Diferencia de proporciones $(P_1 - P_2)$	$z_c = \frac{(\hat{p}_1 - \hat{p}_2) - (P_1 - P_2)}{\sqrt{\frac{\bar{P}\bar{Q}}{n_1} + \frac{\bar{P}\bar{Q}}{n_2}}} \sim N(0,1)$ donde, $\bar{P} = \frac{x_1 + x_2}{n_1 + n_2}$ $\bar{Q} = 1 - \bar{P}$	n_1 y n_2 (grande) P_1 y P_2 no cercano a (0,1)

Nota. Bioestadística: Base para el análisis de las ciencias de la salud. Cuarta edición. Editorial Limusa; Wayne D. 2002 (1).

El estadístico de prueba es el que proporciona la evidencia y permitirá rechazar o no rechazar *Ho*.

Por ejemplo, con las hipótesis planteadas previamente:

$$Ho: \mu \leq 240$$

$$Ha: \mu > 240$$

¿Qué decisión debería tomarse si en una muestra de n=64 pacientes, se encuentra que el promedio es 240 minutos? ¿Rechazaríamos *Ho*?

¿Y si la media muestral fuera 242 minutos? ¿O 244 minutos? ¿O 246 minutos?

Para probar las hipótesis en cuestión:

$$Ho: \mu \leq 240$$

$$Ha: \mu > 240$$

Y asumiendo que la desviación estándar poblacional es σ=24 minutos, se usa el siguiente estadístico de prueba (caso 1), el cual se sabe que sigue una distribución normal estándar:

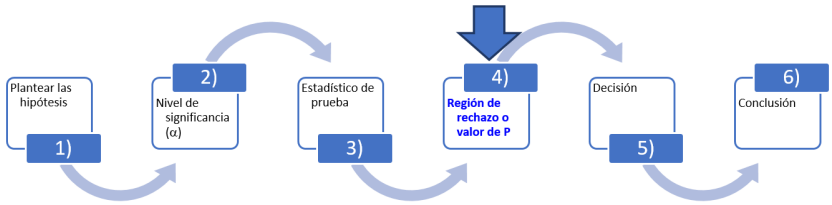
$$z_c = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

El estadístico de prueba se calcula con los datos muestrales asumiendo que *Ho* es verdadera. Los valores del estadístico de prueba (Z_c) para los cuatro escenarios propuestos se presentan en la Tabla 33.

Tabla 33. Estadísticos de prueba para cuatro escenarios posibles (*Ho*: μ ≤ 240 vs *Ha*: μ > 240).

Escenario	Media muestral	Estadístico de prueba	Valor del estadístico de prueba
1	$\bar{X} = 240$	$z_c = \frac{240 - 240}{24/\sqrt{64}}$	$z_c = 0$
2	$\bar{X} = 242$	$z_c = \frac{242 - 240}{24/\sqrt{64}}$	$z_c = 0{,}67$
3	$\bar{X} = 244$	$z_c = \frac{244 - 240}{24/\sqrt{64}}$	$z_c = 1{,}33$
4	$\bar{X} = 246$	$z_c = \frac{246 - 240}{24/\sqrt{64}}$	$z_c = 2$

Nota. Cuanto mayor es la diferencia entre el promedio muestral y el valor propuesto de la hipótesis nula, el valor del estadístico de prueba (Z_c) se hace mayor. Elaboración propia.



En el paso 5 la decisión de rechazar o no *Ho* depende de qué tanta compatibilidad muestran los datos observados en la muestra con la hipótesis nula. Esto se puede hacer de dos formas:

1. Comparando el estadístico de prueba con un valor teórico, asumiendo que la hipótesis nula es verdadera → Determinar la **región de rechazo** (RR).

2. Calculando la probabilidad de haber observado un estadístico de prueba igual o más extremo al que se observó con los datos cuando la hipótesis nula es verdadera → Calcular el **valor P**.

La **región de rechazo** la determina el nivel de significancia (α) del estudio. La hipótesis nula (*Ho*) se rechaza cuando el estadístico de prueba cae en la región de rechazo, indicando poca compatibilidad de los datos con la hipótesis nula. La Figura 37 muestra el establecimiento de la región de rechazo cuando el estadístico de prueba tiene una distribución normal estándar.

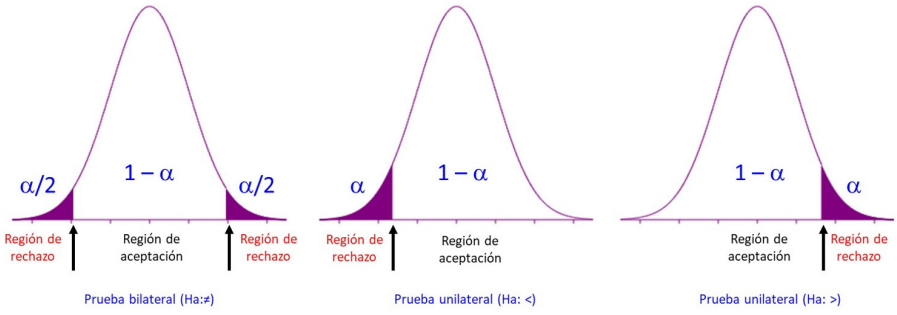


Figura 37. Representación de la región de rechazo en pruebas de hipótesis bilaterales y unilaterales. Gráficas realizadas en Calculadoras Online [Internet]. Calculadoras Online. 2018 [citado el 15 de diciembre de 2021]. Disponible en: https://calculadorasonline.com/

En el ejemplo que se está desarrollando:

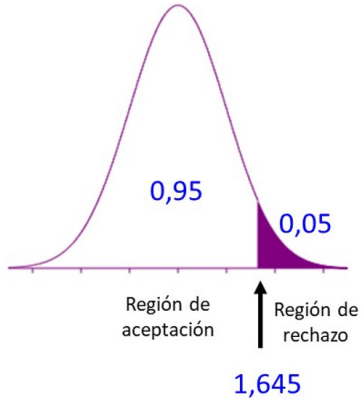
Ho: $\mu \leq 240$

Ha: $\mu > 240$

El nivel de significancia fue α=0,05.

Entonces la región de rechazo está dada por:

$Z_{1-\alpha} = Z_{0,95} = 1,645$



Luego solo en el escenario 4 se puede rechazar *Ho*.

El **valor P** responde a la pregunta: ¿Qué tan probable es observar un estadístico de prueba como el observado ($Z_c = 0$; $Z_c = 0,67$; $Z_c = 1,33$; $Z_c = 2$) o más extremo si la media poblacional es realmente 240?

La respuesta a esta pregunta es dada por el **valor P**. El **valor P** indica la probabilidad de observar un estadístico de prueba (Z_c) igual o mayor al observado si la hipótesis nula es verdadera.

Prueba bilateral (Ha:≠)	Prueba unilateral (Ha: <)	Prueba unilateral (Ha: >)
Valor P = P(Z < -Zc) + P(Z > Zc)	Valor P = P(Z < Zc)	Valor P = P(Z > Zc)

Las siguientes gráficas muestran el **valor P** para los cuatro escenarios planteados (Figura 38). El **valor P** corresponde al área morada. Observe que a medida que Z_c se hace más grande el **valor P** se hace menor, sugiriendo el rechazo de *Ho*.

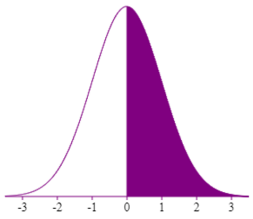
Escenario 1:

$\bar{x} = 240$

$Z_c = 0$

Valor p $= P(Z > 0)$

Valor p $= 0{,}5$



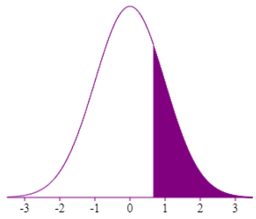
Escenario 2:

$\bar{x} = 242$

$Z_c = 0{,}67$

Valor p $= P(Z > 0{,}67)$

Valor p $= 0{,}252$



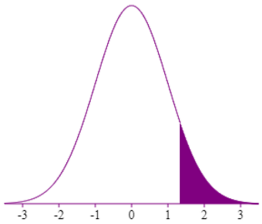
Escenario 3:

$\bar{x} = 244$

$Z_c = 1{,}33$

Valor p $= P(Z > 1{,}33)$

Valor p $= 0{,}091$



Escenario 4:

$\bar{x} = 246$

$Z_c = 2$

Valor p $= P(Z > 2)$

Valor p $= 0{,}023$

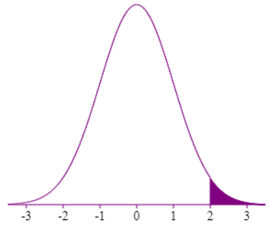
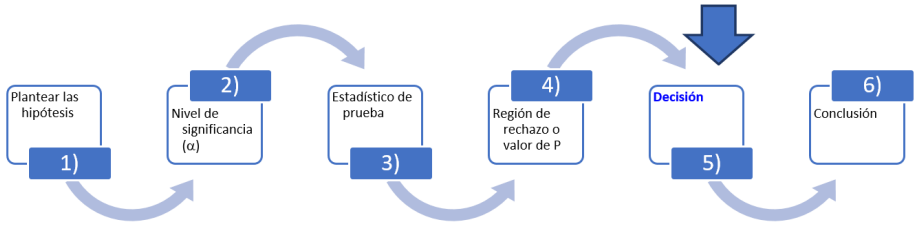


Figura 38. Representación del valor P (área morada) para cuatro escenarios posibles (*Ho: μ ≤ 240* vs *Ha: μ > 240*). Gráficas realizadas en Calculadoras Online [Internet]. Calculadoras Online. 2018 [citado el 15 de diciembre de 2021]. Disponible en: https://calculadorasonline.com/



La **decisión** de una prueba de hipótesis se refiere a rechazar o no rechazar la hipótesis nula (*Ho*). Esta decisión sobre *Ho* puede tomarse usando la región de rechazo o el valor P.

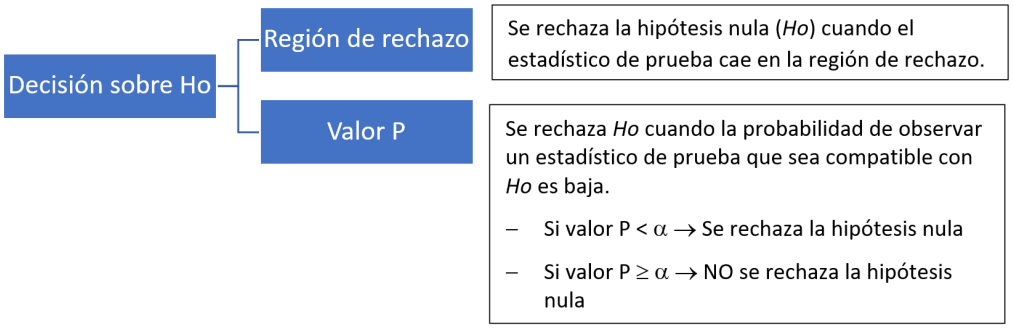
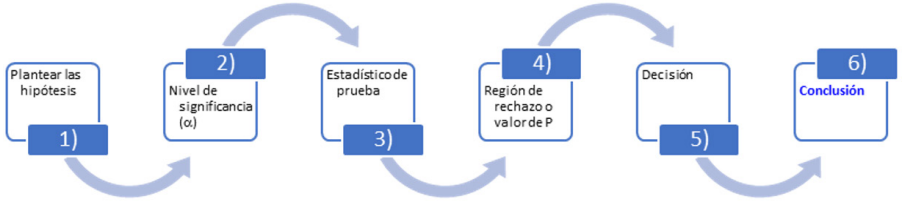


Figura 39. Decisión sobre Ho. Elaboración propia.



La **conclusión** de una prueba de hipótesis conduce a una interpretación.

$$Ho: \mu \leq 240$$

$$Ha: \mu > 240$$

Bajo los tres primeros escenarios planteados, *Ho* no se hubiera rechazado (porque valor P ≥ 0,05) y se hubiese concluido que el promedio de alivio del dolor es menor o igual a 240 minutos (no es mayor a 240 minutos) con un nivel de significancia de 0,05.

En el escenario cuatro, la decisión hubiera sido el rechazo de *Ho* (porque valor P=0,023 < 0,05) y se hubiese concluido que el promedio de alivio del dolor es mayor a 240 minutos con un nivel de significancia de 0,05.

Errores en una prueba de hipótesis

Por el hecho de tomar decisiones sobre los parámetros (θ) con base en los resultados de una muestra debe reconocerse que se pueden cometer errores.

Por ejemplo:

- Rechazar *Ho* cuando esta es verdadera.
- No rechazar *Ho* cuando es falsa.

El siguiente cuadro muestra las decisiones correctas y los posibles errores (tipo I y II) que se pueden cometer en un proceso de prueba de hipótesis.

Tabla 34. Aciertos y desaciertos en un proceso de prueba de hipótesis.

		Condición verdadera	
		H_0 es verdadera	H_0 es falsa
Conclusiones de la muestra	No rechaza H_0	"Decisión correcta" Nivel de confianza **(1- α)**	Error tipo II **β**
	Rechaza H_0	Error tipo I Nivel de significación **α**	"Decisión correcta" Poder de la prueba **(1 – β)**

Nota. Elaboración propia.

- Error I: Se comete cuando se rechaza *Ho* que es verdadera (α).
- Error tipo II: Se comete cuando no se echaza la *Ho* que es falsa (β).

Ejemplo 26. Prueba de hipótesis para la media poblacional

Un grupo de investigadores está interesado en conocer la concentración media de una enzima en cierta población. Ellos se preguntan si es posible concluir que el nivel medio de la enzima en esta población es diferente de 25. Los datos de una muestra de 64 individuos extraída de la población dieron una media de 22. Supóngase que la muestra proviene de una población normal y que la población tiene una varianza conocida de 45 (σ^2=45).

Solución:

El parámetro en estudio es un promedio (μ). Se tiene que n=64; $\bar{x}$=22; $\sigma = \sqrt{45}$. Por la información disponible (variable con distribución normal y varianza poblacional conocida) se usa el estadístico de prueba del caso ❶ de la Tabla 32. Las hipótesis de la prueba son:

$$Ho: \mu = 25$$

$$Ha: \mu \neq 25$$

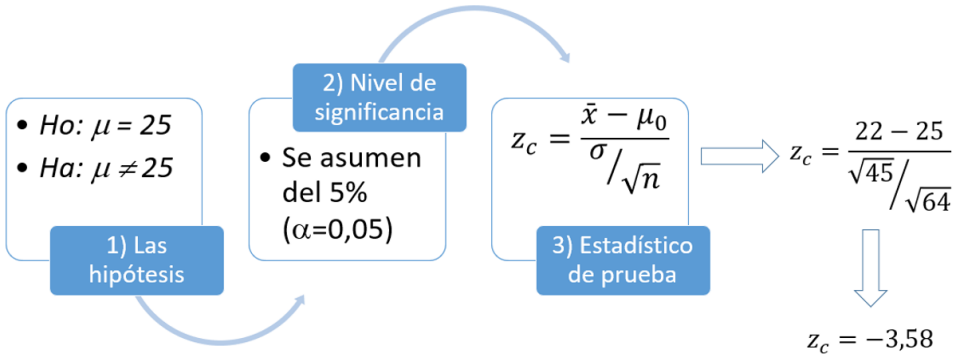
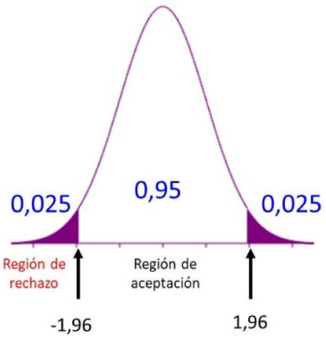


Figura 40. Pasos de la prueba de hipótesis. Elaboración propia.

Región de rechazo:

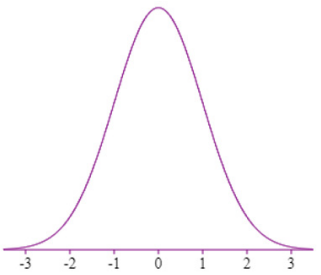


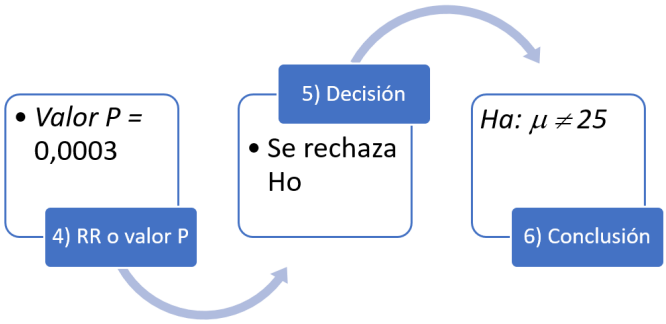
Valor P:

Valor P = $P(Z < -Z_c) + P(Z > Z_c)$

Valor P = $P(Z < -3{,}58) + P(Z > 3{,}58)$

Valor P = 0,0003





Paso 4: El estadístico de prueba (Z_c = -3,58) cae en la región de rechazo (Z_c < -1,96) y el valor P es menor que el nivel de significancia (valor P= 0,0003 < α = 0,05).

Paso 5: Se puede rechazar la hipótesis nula (*Ho*).

Paso 6: Sí es posible concluir que el nivel medio de la enzima en esta población es diferente de 25, con un nivel de significancia del 5 % (α=0,05).

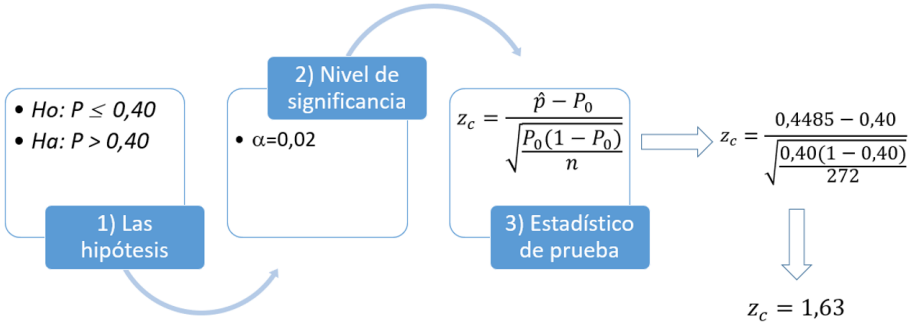
Ejemplo 27. Prueba de hipótesis para la proporción poblacional

La obesidad es un problema de salud pública en Colombia y en el mundo, con diversas consecuencias para la salud cardiovascular, calidad de vida y mortalidad. El objetivo de un estudio fue determinar la prevalencia de sobrepeso y obesidad en adultos de una comunidad. Se midieron 272 personas (182 mujeres y 90 hombres) de 18 a 64 años en variables como estatura, peso y perímetro de cintura. Usando los criterios de obesidad general de la Organización Mundial de la Salud (OMS) se encontraron 122 personas con sobrepeso/obesidad. ¿Se puede concluir que más del 40 % de la población en esa comunidad presenta sobrepeso/obesidad? Use un nivel de significancia del 2 %.

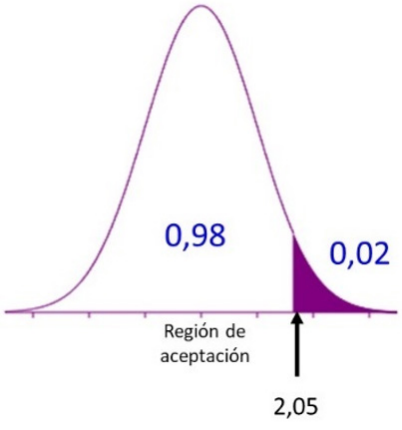
Solución:

El parámetro en estudio es una proporción (*P*). Se tiene que n=272; $\hat{p} = {}^{122}/_{272} = 0{,}4485$; α=*0,02*. Se debe usar el estadístico de prueba del caso ❸.

Pasos de la prueba de hipótesis:



Región de rechazo:

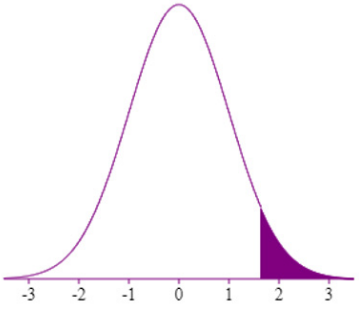


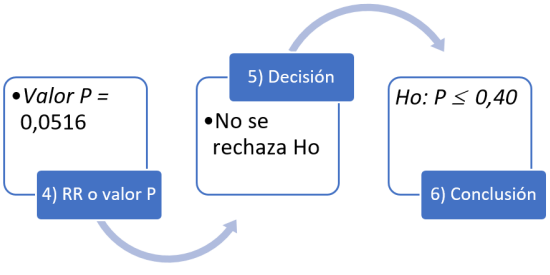
Valor P:

Valor P = $P(Z > Z_c)$

Valor P = $P(Z > 1{,}63)$

Valor P = 0,0516





Paso 4: El estadístico de prueba (Z_c = 1,63) cae en la región de aceptación (Z_c < 2,05) y el valor P es mayor que el nivel de significancia (valor P= 0,0519 > α = 0,02).

Paso 5: No se puede rechazar la hipótesis nula (*Ho*).

Paso 6: No se puede concluir que más del 40 % de la población en esa comunidad presenta sobrepeso/obesidad. Con un nivel de significancia del 2 % (α=0,02) la prevalencia de sobrepeso/obesidad es menor o igual al 40 %.

3.2.7 Ejercicios propuestos

Ejercicio 12.

Un grupo de investigadores reunió las concentraciones de amilasa en el suero de una muestra aleatoria de 50 personas aparentemente sanas. La media muestral fue de 96 unidades/100 ml. Se sabe que la variable presenta una distribución normal en la población general con una desviación estándar de 36 unidades/100 ml.

Los investigadores desean saber si es posible concluir que la media de la población de la cual se extrajo la muestra es:

a. Distinta de 108.

b. Menor de 108.

Ejercicio 13.

El director de una secretaría de salud sospecha que menos de la mitad de la población de conductores de la comunidad utiliza con regularidad el cinturón de seguridad del asiento.

En una encuesta a 400 conductores adultos de automóviles, 180 respondieron que regularmente utilizaban el cinturón de seguridad del asiento.

Usando un nivel de significancia del 5 % ¿Qué se puede concluir?

3.2.8 Pruebas paramétricas y no paramétricas

La estadística inferencial ofrece diversas herramientas de análisis que son seleccionadas de acuerdo con:

- El objetivo del estudio (describir, estimar, comparar, asociar, correlacionar, entre otros).
- El diseño del estudio (descriptivo, analítico, transversal, longitudinal, entre otros).
- La naturaleza de las variables (cualitativas dicotómicas, cualitativas politómicas, discretas, continuas) y el nivel de medición (nominal, ordinal, de razón).
- La distribución de las variables (normal o asimétrica).
- El tipo de muestras a analizar (independientes o relacionadas).
- El número de grupos a comparar.
- El cumplimiento de los supuestos de las mismas pruebas estadísticas.

Importante:

- Cuando el objetivo de un estudio es meramente descriptivo se hace uso de las herramientas estudiadas en el módulo de la estadística descriptiva.
- Cuando un estudio pretende estimar el parámetro de una población (media o proporción) se usan los intervalos de confianza.
- Cuando el estudio es analítico se usan intervalos de confianza (diferencia de medias, diferencia de proporciones, riesgo relativo u odds ratio) y/o alguna de las pruebas estadísticas que se discutirán a continuación.

Muestras independientes vs Muestras relacionadas:

Un concepto importante en la selección de la prueba estadística adecuada para analizar un conjunto de datos tiene que ver con el tipo de datos que se van a comparar. Los datos pueden provenir de muestras independientes o de muestras relacionadas.

Las **muestras independientes** se caracterizan porque no guardan ninguna relación entre ellas. Por ejemplo, estudios analíticos que usan poblaciones diferentes (diferentes sitios o diferentes pacientes). Un ejemplo son los ensayos aleatorizados donde se asignan aleatoriamente los pacientes a diferentes brazos del estudio para garantizar grupos independientes.

Las **muestras relacionadas** se refieren a mediciones repetidas en el tiempo en una misma unidad de análisis. Por ejemplo, en estudios longitudinales los pacientes son observados y medidos en varios momentos en el tiempo, generando datos relacionados (unidades analizadas antes-después). También se tienen muestras relacionadas cuando una misma muestra biológica es analizada por dos observadores o en aquellos diseños de caso-control emparejados.

Pruebas paramétricas vs Pruebas no paramétricas:

Aquellas pruebas estadísticas que asumen una distribución teórica para los parámetros en estudio se llaman pruebas paramétricas. Por ejemplo, la prueba *t de Student* (casos 5 y 6 de la Tabla 32) asume que las muestras son independientes y que la variable en estudio presenta distribución normal (1). Cuando la variable en estudio no presenta una distribución teórica conocida (por ejemplo, no tiene una distribución normal) se usan las pruebas no paramétricas. En otras palabras, las pruebas no paramétricas no hacen afirmaciones de los parámetros poblacionales ni asumen una distribución especifica de la variable en estudio (1). Sin embargo, el uso de pruebas no paramétricas conduce, en gran medida, a la pérdida de información como se explica en el Ejemplo 28.

A continuación, se presentan las pruebas estadísticas más comunes, de acuerdo con el tipo de variables involucradas en los análisis (numérica o categórica), el tipo de prueba (paramétrica o no paramétrica) y el tipo de muestras a comparar (independientes o relacionadas). Estas pruebas, como cualquier prueba de hipótesis, siguen los seis pasos mencionados previamente.

Tabla 35. Pruebas estadísticas para comparar dos grupos.

Tipo de variables	Tipo de prueba	Muestras independientes	Muestras relacionadas o pareadas
Numéricas	Paramétrica	Prueba *t de Student* • Varianzas iguales • Varianzas diferentes	Prueba *t de Student* pareada
	No paramétrica	Prueba U de Mann-Whitney	Prueba de Wilcoxon
Categóricas	No paramétrica	Prueba Ji-cuadrado Prueba Exacta de Fisher	Prueba de McNemar

Tabla 36. Pruebas estadísticas para comparar k grupos (k≥3).

Tipo de variables	Tipo de prueba	Muestras independientes	Muestras relacionadas o pareadas
Numéricas	Paramétrica	Análisis de varianza (ANOVA) de una vía	ANOVA de medidas repetidas
	No paramétrica	Prueba Kruskal-Wallis	Prueba de Friedman
Categóricas	No paramétrica	Prueba Ji-cuadrada para comparaciones múltiples	Prueba Q de Cochran

Tabla 37. Herramientas estadísticas para asociar dos variables numéricas.

Tipo de variables	Tipo de coeficiente	Coeficiente
Numéricas	Paramétrico	Coeficiente de correlación de Pearson
Numéricas u Ordinales	No paramétrico	Coeficiente de correlación de Spearman

Con fines prácticos, a continuación, se presenta un corto resumen de algunas de las pruebas estadísticas más usadas en investigación en salud (Tabla 38). Esta tabla no

pretende ser exhaustiva y tampoco presenta los estadísticos de prueba; su finalidad es guiar la elección y aplicabilidad de las pruebas de hipótesis más usadas.

Tabla 38. Resumen de algunas herramientas de la estadística inferencial.

Objetivo/Nombre	Supuestos	Las hipótesis
Evaluar la distribución normal de una variable numérica: • Prueba de Shapiro-Wilk	• Variable numérica	• *Ho:* La variable en estudio presenta una distribución normal • *Ha:* La variable no presenta una distribución normal
Comprobar la igualdad de las varianzas en varios grupos: • Prueba de Levene, Prueba de Barlett, F-test.	• Variables numéricas	• *Ho:* Varianzas iguales en todos los grupos • *Ha:* Varianzas no iguales en todos los grupos
Comparar el promedio entre dos muestras independientes: • Prueba *t de Student* (caso 5)*	• Muestras independientes • Variable numérica con distribución normal en cada grupo • Varianzas iguales	• *Ho:* $\mu_1 = \mu_2$ • *Ha:* $\mu_1 \neq \mu_2$ O también pruebas unilaterales.
Comparar el promedio entre dos muestras independientes: • Prueba *t de Student* (caso 6)*	• Muestras independientes • Variable numérica con distribución normal en cada grupo • Varianzas diferentes	• *Ho:* $\mu_1 = \mu_2$ • *Ha:* $\mu_1 \neq \mu_2$ O también pruebas unilaterales.
Comparar la mediana entre dos muestras independientes: • Prueba U de Mann-Whitney	• Muestras independientes • Variable numérica u ordinal	• *Ho:* Las dos muestras se distribuyen iguales ($Me_1 = Me_2$) • *Ha:* Las dos muestras no se distribuyen iguales ($Me_1 \neq Me_2$)
Determinar diferencias en el promedio en dos muestras relacionadas: • Prueba *t de Student* pareada*	• Muestras relacionadas • Variables numéricas con observación pareada: (X_i, Y_i), donde: $d_i = X_i - Y_i$ • Distribución normal de la diferencia ($d_i = X_i - Y_i$)	• *Ho:* $\mu_d = 0$ • *Ha:* $\mu_d \neq 0$ O también pruebas unilaterales.
Determinar cambios en una variable en dos muestras relacionadas: • Prueba de Wilcoxon	• Muestras relacionadas • Variables numéricas u ordinales con observación pareada: (X_i, Y_i), donde: $d_i = X_i - Y_i$	• *Ho:* No hay diferencias entre las muestras • *Ha:* Hay diferencias entre las muestras

Comparar el promedio entre *k* muestras independientes: • Análisis de varianza (ANOVA) de un factor*	• Muestras independientes • Variable numérica con distribución normal en cada grupo • Varianzas iguales en todos los grupos	• *Ho:* $\mu_1 = \mu_2 = \mu_3 = ... = \mu_k$ • *Ha:* No todas las medias son iguales Cuando se rechaza *Ho*, posteriormente se deben usar pruebas de comparaciones múltiples para determinar cuáles grupos difieren. Algunas pruebas para comparaciones múltiples son Bonferroni, TukeyHSD, Duncan, Dunnet. Corrección de Holm, entre otros
Comparar la mediana entre *k* muestras independientes: • Kruskal-Wallis	• Muestras independientes • Variable numérica u ordinal	• *Ho:* $Me_1 = Me_2 = Me_3 = ... = Me_k$ • *Ha:* No todas son las medianas son iguales Cuando se rechaza *Ho* las comparaciones múltiples se hacen con el Procedimiento de Dunn.
Determinar asociación entre dos variables categóricas de una tabla de contingencia: Prueba Ji-cuadrado	• Variables categóricas en una tabla de contingencia • No más del 20 % de las celdas deberían tener valores esperados menores que 5 • Ninguna celda debería tener valor esperado menor que 1	Cuando se tiene una sola muestra: • *Ho:* Las variables son independientes (no hay asociación entre las dos variables) • *Ha:* Las variables no son independientes (hay asociación entre las dos variables) Cuando se tienen *k* muestras independientes: • *Ho:* Las *k* poblaciones son homogéneas en el criterio de clasificación • *Ha:* Las *k* poblaciones no son homogéneas en el criterio de clasificación
Determinar asociación entre dos variables categóricas de una tabla de contingencia: Prueba Ji-cuadrado con corrección de Yates	• Variables categóricas en una tabla de contingencia • Tablas 2×2 • Ninguna de las cuatro celdas tiene valores esperados menores que 5	
Determinar asociación entre dos variables categóricas de una tabla de contingencia: Prueba Exacta de Fisher	• Variables categóricas en una tabla de contingencia • Valores esperados < 5 • Regularmente cuando n < 20	
Determinar si hay correlación lineal entre dos variables numéricas: Coeficiente de correlación de Pearson*	• Variables numéricas (*x* y *y*) con distribución normal • Las mediciones son en una misma unidad: (x_i, y_i)	• *Ho:* $\rho = 0$ (no hay ninguna correlación entre las dos variables numéricas) • *Ha:* $\rho \neq 0$ (hay alguna correlación entre las dos variables numéricas)

Determinar si hay correlación entre dos variables numéricas u ordinales: Coeficiente de correlación de Spearman	• Variables numéricas u ordinales • Las mediciones son en una misma unidad: (x_i, y_i)	• *Ho:* No hay ninguna correlación entre las dos variables • *Ha:* Hay correlación entre las dos variables

Nota. Prueba paramétrica. Elaboración propia.

Ejemplo 28. Prueba no paramétrica U de Mann-Whitney

Suponga que se diseñó un experimento para estimar los efectos de la inhalación prolongada de óxido de cadmio (1). La variable respuesta fue la concentración de hemoglobina (gramos) después de realizar el experimento. Para ello, se tomó una muestra de 25 animales de laboratorio, de los cuales 15 fueron sujetos para el experimento mientras que los otros 10 sirvieron de grupo control. En este experimento se quería saber si es posible concluir que la inhalación prolongada de óxido de cadmio disminuye los niveles de hemoglobina. Los resultados del experimento se presentan en la Tabla 39. Usar α=0,05.

Tabla 39. Concentraciones de hemoglobina (gramos) en los animales del experimento.

Grupo 1: Experimental (n=15)	Grupo 2: Control (n=10)
14,4	17,4
14,2	16,2
13,8	17,1
16,5	17,5
14,1	15,0
16,6	16,0
15,9	16,9
15,6	15,0
14,1	16,3
15,3	16,8
15,7	
16,7	
13,7	
15,3	
14,0	

Solución:

Las concentraciones de hemoglobina se muestran en la Figura 37, donde aprecian menores niveles en el grupo experimental. Esta tendencia se confirma con el resumen presentado en la Tabla 40.

Tabla 40. Resumen de las concentraciones de hemoglobina (gramos) por grupo.

Estadístico	Grupo 1: Experimental	Grupo 2: Control
Tamaño de muestra	15	10
Promedio	15,06	16,42
Desviación estándar	1,077	0,897
Mínimo	13,7	15
Cuartil 1	14,1	16
Mediana	15,3	16,55
Cuartil 2	15,9	17,1
Máximo	16,7	17,5

Se asume que la distribución de la hemoglobina es aproximadamente normal en el grupo control, pero no en el experimental. En este experimento para comparar una variable numérica (hemoglobina) en dos grupos independientes (experimental y control), y asumiendo normalidad solo en el grupo control, se debe usar la prueba no paramétrica U de Mann-Whitney.

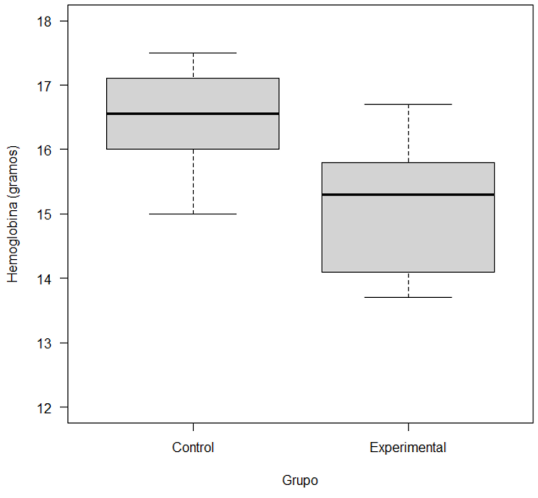


Figura 41. Concentraciones de hemoglobina (gramos) en animales expuestos (*n*=15) y no expuestos (*n*=10) a inhalación de óxido de cadmio. Elaboración propia.

Recuerde que en este experimento se quería saber si era posible concluir que la inhalación prolongada de óxido de cadmio disminuye los niveles de hemoglobina. En otras palabras, si después del experimento el nivel de hemoglobina en el grupo experimental (Grupo 1) es menor que en el grupo control (Grupo 2).

Las hipótesis de interés son:

$$Ho: Me_1 \geq Me_2$$

$$Ha: Me_1 < Me_2$$

Me_1 representa la mediana de la hemoglobina en el Grupo 1 (animales expuestos al óxido de cadmio) y Me_2 es la mediana en el Grupo 2 (control).

Anteriormente se mencionó que una de las desventajas en el uso de las pruebas no paramétricas es la pérdida de información. Para calcular el estadístico de prueba (*T*) de la prueba U de Mann-Whitney se debe combinar la información de ambos grupos (experimental y control). Para ello, primero se ordenan todas las observaciones de menor a mayor, pero considerando el grupo al que pertenece cada observación (Tabla 41). Luego se asignan "rangos" a cada dato de acuerdo con su posición: uno ("1") al menor valor, dos ("2") al segundo menor valor, y así sucesivamente. Note que en la Tabla 41 los valores 14,1 están repetidos. A estos datos les corresponderían los rangos "4" y "5", pero al estar empatados, se les asigna el "rango promedio", es decir, "4,5".

Tabla 41. Asignación de rangos a las muestran combinadas del experimento de acuerdo con su posición una vez ordenados los datos de menor a mayor.

Grupo	Hemoglobina	Rango asignado al grupo 1	Rango asignado al grupo 2
Experimental	13,7	1	
Experimental	13,8	2	
Experimental	14	3	
Experimental	14,1	4,5	
Experimental	14,1	4,5	
Experimental	14,2	6	
Experimental	14,4	7	
Control	15		8,5
Control	15		8,5
Experimental	15,3	10,5	
Experimental	15,3	10,5	
Experimental	15,6	12	
Experimental	15,7	13	
Experimental	15,9	14	
Control	16		15
Control	16,2		16

Control	16,3		17
Experimental	16,5	18	
Experimental	16,6	19	
Experimental	16,7	20	
Control	16,8		21
Control	16,9		22
Control	17,1		23
Control	17,4		24
Control	17,5		25
	Suma de rangos (S)	**145**	**180**

Nota. Elaboración propia

El estadístico de prueba[30] del ejemplo es $T=25$ y su valor p asociado es 0,002[31]. Considerando que el valor P=0,002 es menor que el nivel de significancia (α=0,05), se puede rechazar la hipótesis nula y concluir que la inhalación prolongada de óxido de cadmio disminuye los niveles de hemoglobina, con un nivel de confianza del 95 %.

Ejemplo 29. Prueba no paramétrica Ji-cuadrada

Para estudiar la relación entre el tipo sanguíneo y la severidad de una enfermedad en una población, se tomó una muestra de 1 500 personas de una población (1). Los datos del estudio se presentan en la Tabla 42. Usar α=0,05.

Tabla 42. Tabla de contingencia* clasificando a las personas por tipo sanguíneo y la severidad de la enfermedad.

Grado de la afección	Tipo sanguíneo				
	A	B	AB	O	Total
Ninguno	543	211	90	476	1 320
Moderado	44	22	8	31	105
Severo	28	9	7	31	75
Total	615	242	105	538	1 500

Nota. Esta tabla de contingencia se considera de tamaño 3×4, porque la variable de las filas (grado de afección) tiene tres categorías de respuesta (i=3) y la variable de las columnas (tipo sanguíneo) tiene cuatro categorías de respuesta (j=4).

30 La estadística de prueba es: $T = S - \frac{n(n+1)}{2}$; donde S es la suma de los rangos del grupo 1 y n es el tamaño de la muestra de ese mismo grupo. La elección del grupo 1 es arbitraria. Con los datos del ejemplo el estadístico de prueba calculado es $T = 145 - \frac{15(15+1)}{2} = 25$.

31 Valor calculado en un paquete estadístico.

Solución:

Para determinar asociación entre dos variables categóricas se puede usar la prueba Ji-cuadrado. Se plantean las siguientes hipótesis:

- *Ho*: No hay asociación entre el tipo sanguíneo y la severidad de la enfermedad
- *Ha*: Sí hay asociación entre el tipo sanguíneo y la severidad de la enfermedad

El estadístico de prueba se obtiene a partir de los datos observados (O_{ij}, de la Tabla 42) y los datos esperados (E_{ij})[32] asumiendo que las variables son independientes. Usando un paquete estadístico se encuentra que el estadístico de prueba[33] es de $\chi^2_c = 5{,}12$ y el valor P =0,529.

Como el valor P=0,529 es mayor que el nivel de significancia (α=0,05), no se puede rechazar la hipótesis nula, y se concluye que no hay asociación entre el tipo sanguíneo y la severidad de la enfermedad, con un nivel de confianza del 95 %.

32 $E_{ij} = \frac{r_i c_j}{n} = \frac{total\ fila\ i * total\ columna\ j}{Tamaño\ de\ la\ muestra}$

33 $\chi^2_c = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$;donde Oi son los valores observados y Ei son los valores esperados, en cada una de las celdas de la tabla de contingencia.

4. Demografía

4.1 Fundamentos de demografía

De acuerdo con la definición del Diccionario Demográfico Multilingüe[34] "La *demografía* es la ciencia que tiene por objeto el estudio de las poblaciones humanas tratando, desde un punto de vista principalmente cuantitativo, su dimensión, su estructura, su evolución y sus características generales". En otras palabras, la demografía se interesa en el estudio del movimiento, evolución y cambios que se presentan en las poblaciones humanas. Luego el interés de la demografía son las poblaciones y no los individuos (17).

El concepto de **dimensión** se refiere al tamaño de una población (número de habitantes). La **estructura** hace referencia al estudio de la población en subgrupos de acuerdo con las variables que afectan los fenómenos demográficos como sexo, edad, lugar de residencia, estado civil, nivel educativo, etc. La **evolución** corresponde a los cambios en tamaño y estructura de la población, producto de diversas dinámicas sociales, económicas y políticas en la historia de un territorio.

El concepto de "población" en el contexto de la bioestadística e investigación hace referencia al conjunto de todos los elementos (individuos, objetos, unidades, mediciones, etc.) de interés en un estudio, sobre los cuales se obtendrá la información y hacia los cuales se extenderán las conclusiones de un estudio. Sin embargo, en demografía el concepto de "**población**" hace referencia a un conjunto de habitantes en un espacio geográfico, o territorio determinado, y que naturalmente experimenta continuos procesos de cambio.

El estudio de la demografía es importante para las áreas relacionadas con estudios poblacionales (epidemiología, salud pública, geografía, sociología, entre otras) permitiendo comprender la evolución de las poblaciones (18).

4.1.1 Fenómenos demográficos básicos

Reconociendo que toda población está en continuo cambio (17), por procesos de entrada y salida de personas en la población, la demografía estudia: 1) el estado; y 2) la dinámica de las poblaciones a lo largo del tiempo. El estado de la población se refiere al tamaño, la distribución geográfica y estructura poblacional por sexo, edad u otros subgrupos de interés. La dinámica se refiere a todo aquello que hace que una población cambie en el tiempo y se distinga de otras, siendo su principal interés la interrelación de tres fenómenos demográficos básicos: la **fecundidad** (natalidad),

34 IUSSP. Diccionario demográfico multilingüe. Liège: Union Internationale por l'Étude Scientiphique de la Population; 1985. Disponible en: http://www.demopaedia.org/

la **mortalidad** (defunciones) y la **migración** (inmigrantes y emigrantes). Los dos primeros (los nacimientos y las defunciones) son los principales determinantes del cambio o dinámica demográfica.

En la dinámica demográfica cada uno de estos fenómenos demográficos interviene de modo diferente: unos componentes suman individuos y otros restan, siendo además su aporte diferencial por las variables mencionadas previamente: sexo, edad, etc. La interrelación de la natalidad, la mortalidad y la migración determinan la dimensión y la estructura de una población y dieron origen a la llamada **ecuación compensadora** (17) o balance poblacional:

$$P_f - P_0 = N - D + I - E$$

donde:

- P_f : Población final.
- P_0 : Población inicial.
- N : Nacimientos.
- D : Defunciones.
- I: Inmigrantes.
- E: Emigrantes.

La Figura 42 muestra los componentes de la ecuación compensadora sobre una pirámide poblacional.

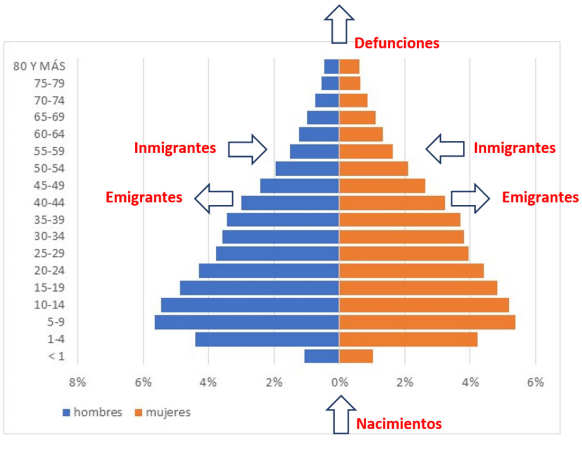


Figura 42. Componentes de la ecuación compensadora sobre una pirámide poblacional. Elaboración propia.

Así, el estudio de la fecundidad y mortalidad ocupa gran parte de los análisis demográficos por medio de sus indicadores, los cuales serán explicados más adelante. Por ahora, para comprender mejor los cambios de las dinámicas poblacionales en la historia de la humanidad, se introducen dos teorías que explican de manera muy general, esas transformaciones.

4.1.2 Transición demográfica y transición epidemiológica

En la historia de la humanidad el estado de la población ha mostrado diversos patrones. Los conceptos de transición epidemiológica y transición demográfica surgen como teorías que intentaban explicar los cambios demográficos en Europa a finales del siglo XIX e inicios del XX (19). La transición epidemiológica surgió para explicar el descenso en la mortalidad, encontrándose con una transición demográfica como resultado de la observación descendente de los fenómenos de mortalidad y fecundidad a causa de transformaciones sociales y económicas luego de la modernización industrial.

La transición demográfica se refiere a la evolución desde perfiles poblacionales con altas tasas de natalidad, fecundidad y mortalidad general hacia poblaciones con bajos indicadores en estos fenómenos, aumento de la esperanza de vida al nacer y un envejecimiento poblacional (20). Los indicadores demográficos mostraban altos niveles tanto de fecundidad como de mortalidad, produciendo una compensación entre nacimientos y defunciones, esta última explicada por epidemias mortales como la peste, periodos de hambruna o guerras. En estos tiempos la dinámica demográfica reflejaba un crecimiento poblacional lento e inestable.

Los cambios sociales y culturales, que condujeron a una transformación importante en el crecimiento poblacional y la disminución de la mortalidad, se relacionan principalmente con la revolución industrial en el siglo XVIII y el desarrollo económico, adelantos en medicina y el mejoramiento en la higiene de las viviendas y la alimentación.

Otro cambio importante se vio reflejado en las costumbres de las familias y la disminución de la fecundidad. El cambio en los patrones de mortalidad y fecundidad, pero a diferentes ritmos, condujo a una transición demográfica. Por ejemplo, en Europa la población se cuadruplicó entre 1750 y 1950 y la población mundial se duplicó entre 1950 y 1987 pasando de 2500 a 5000 millones de habitantes. En Colombia la población se duplicó entre 1950 y 1972 pasando de 11 a 22 millones de habitantes, y nuevamente se duplicó en 2010 a 44 millones de personas.

La preocupación por el crecimiento poblacional impulsó la demografía a nivel mundial, reconociendo la importancia del estudio de las poblaciones y la búsqueda de estrategias para controlar su crecimiento. Los estudios demográficos son fundamentales para entender las características de la población, para la toma de decisiones y la elaboración de políticas públicas.

4.1.3 Pirámide poblacional

La pirámide poblacional (Figura 42) es quizás una de las herramientas más representativas de la demografía, siendo uno de los pilares para el análisis de la estructura poblacional. Esta gráfica resume, para una población y en un momento del tiempo, el resultado de la interacción de la natalidad, mortalidad y migración.

Las pirámides poblacionales son herramientas visuales propias de la demografía, las cuales permiten un análisis general de la distribución por sexo y edad de la población en un momento determinado. Si se analiza la forma de una pirámide poblacional (Figura 42) se puede ver que básicamente está constituida por dos histogramas en espejo (uno para cada sexo). Se ha definido por convención mostrar el sexo masculino a la izquierda y el femenino a la derecha.

En una pirámide poblacional el eje *X* representa el número de habitantes o el porcentaje, mientras que el eje *Y* contiene las edades de la población, ya sea edades simples o agrupadas.

Las pirámides de población son adecuadas para analizar la dinámica de una población o para comparar varias poblaciones en un momento dado. La Figura 43 muestra la distribución de la población en Colombia para los años 2000 y 2020, construida desde el portal https://www.populationpyramid.net/, donde se aprecia una reducción importante en el peso de las edades más jóvenes, y con esto, un envejecimiento de su población.

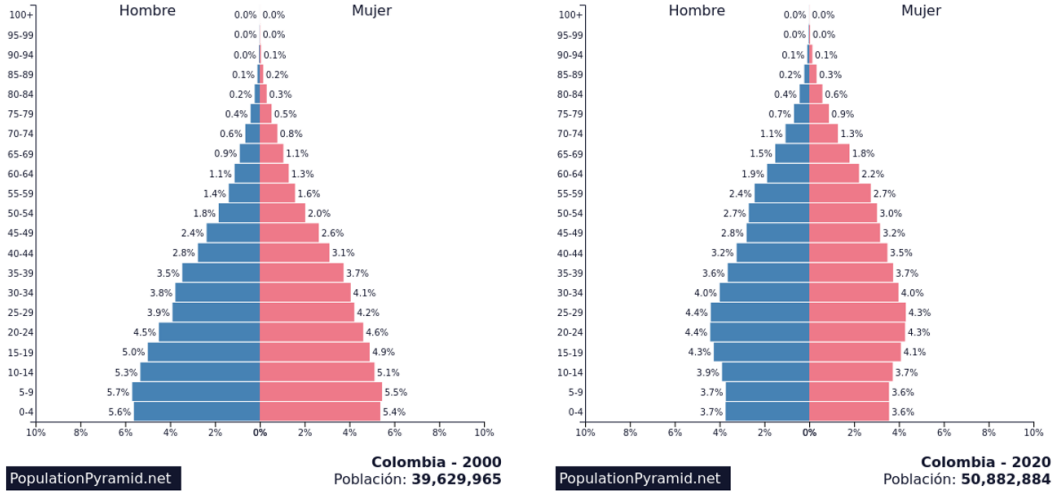


Figura 43. Pirámide poblacional de Colombia para los años 2000 y 2020. Population pyramids of the world from 1950 to 2100 [Internet]. PopulationPyramid.net. [citado el 21 de diciembre de 2021]. Disponible en: https://www.populationpyramid.net/

Como se mencionó previamente, en las pirámides el eje *Y* contiene las edades que pueden ser agrupadas (Figura 43) o simples (Figura 44). La Figura 44 contiene las pirámides de población con edades simples para Colombia en los años 2000 y 2020, construida usando el módulo de Demografía del programa Epidat 4.2.

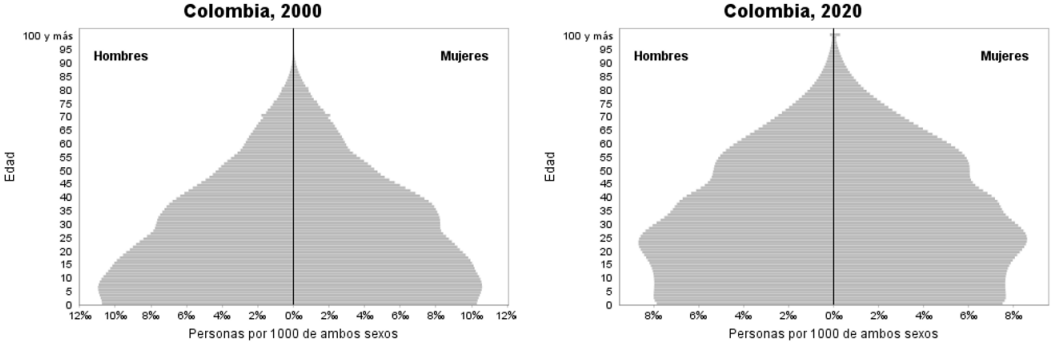


Figura 44. Pirámide poblacional de Colombia con edades simples para los años 2000 y 2020. DANE. Serie nacional de población por edad, sexo y área para el periodo 1950-2070 con base en el CNPV 2018. Pirámides elaboradas en Epidat 4.2

Aunque una pirámide construida con edades simples permite un análisis más preciso, tiene la desventaja que se ve afectada por la calidad de los datos y será inestable en poblaciones pequeñas, mostrando gran variabilidad. Para evitar esto las pirámides suelen construirse con edades quinquenales como se presentó en la Figura 43.

De acuerdo con la forma, existen tres perfiles clásicos de las pirámides poblacionales (ver Figura 45) (21):

1. **Expansiva:** este tipo se asemeja bastante a una verdadera pirámide, con una base ancha y un rápido descenso al aumentar la edad. Este tipo de población es característica de países del tercer mundo y regiones rezagadas, donde la población es predominantemente joven y con altas tasas de natalidad.

2. **Constrictiva:** esta pirámide muestra una base más estrecha como resultado de un descenso de la natalidad y la mortalidad, además de un envejecimiento poblacional. Este tipo de pirámide es frecuente en países desarrollados que están finalizando su transición demográfica y con un crecimiento natural reducido.

3. **Estacionaria:** este tipo de pirámide es muy vertical, con igual proporción de población joven y adulta. Su población es muy envejecida, con bajas tasas de natalidad y crecimiento poblacional muy bajo.

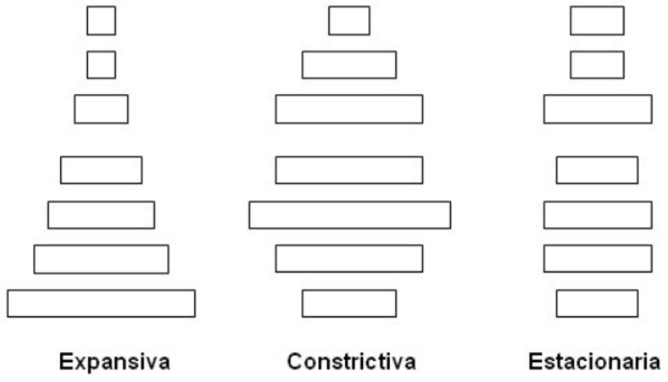


Figura 45. Tipos de pirámides poblacionales. ARCIA, Leonel. Demografía y Salud: Apuntes para una Conferencia. Rev haban cienc méd [online]. 2009, vol.8, n.4 (21)

Pero para aportar los elementos que explican las dinámicas de población se requieren de indicadores poblacionales relacionados con la fecundidad, mortalidad y migración. A continuación, se introducen las principales fuentes de información que permiten la construcción de tales indicadores.

4.1.4 Fuentes de información

La demografía, al igual que la epidemiología, requiere de información para poder estudiar el estado y la dinámica de las poblaciones. En general, el registro y la recolección de eventos y variables de interés puede realizarse de dos formas: estática (transversal) o dinámica (secuencial).

La recolección de información demográfica estática es la que se realiza de forma transversal; es decir, una sola vez. En esta categoría se encuentran los censos de población y las encuestas poblacionales por muestreo. La recolección de información dinámica o secuencial se recolecta en la medida que van ocurriendo los eventos. A esta categoría pertenece el registro de estadísticas vitales (nacimientos y defunciones).

Las fuentes de datos pueden también clasificarse como primarias o secundarias:

- Fuentes de información primaria (datos primarios): es aquella información que es recolectada por un estudio o investigador para fines propios, es decir para cumplir los objetivos de un estudio específico.
- Fuentes de información secundaria (datos secundarios): es la información recolectada fuera del contexto de una investigación concreta.

En conclusión, las principales fuentes de información demográfica provienen de tres fuentes básicas: los censos poblacionales, las estadísticas vitales y las encuestas poblacionales.

Censos de población

De acuerdo con las Naciones Unidas, un censo es "un conjunto de operaciones que consiste en reunir, elaborar y publicar datos demográficos, económicos y sociales, correspondientes a todos los habitantes de un país o territorio definido y referido a un momento determinado o a ciertos períodos de tiempo dados" (18).

Los temas que incluye usualmente un censo de población son:

- Geográficos.
- Demográficos.
- Educación.
- Actividad económica.
- Otros.

En la historia de Colombia se han realizado 18 censos de población (22) siendo el ultimo realizado el Censo Nacional de Población y Vivienda 2018 con acceso a su información desde el portal del Departamento Administrativo Nacional de Estadística (DANE): https://www.dane.gov.co/.

La información disponible del Censo Nacional de Población y Vivienda 2018 (Figura 46) corresponde a: 1) Proyecciones de población; 2) Índice de Necesidades Básicas Insatisfechas (NBI) 2018; 3) Microdatos censales anonimizados; 4) Población ajustada por cobertura; y 5) Principales resultados de viviendas, hogares y personas.

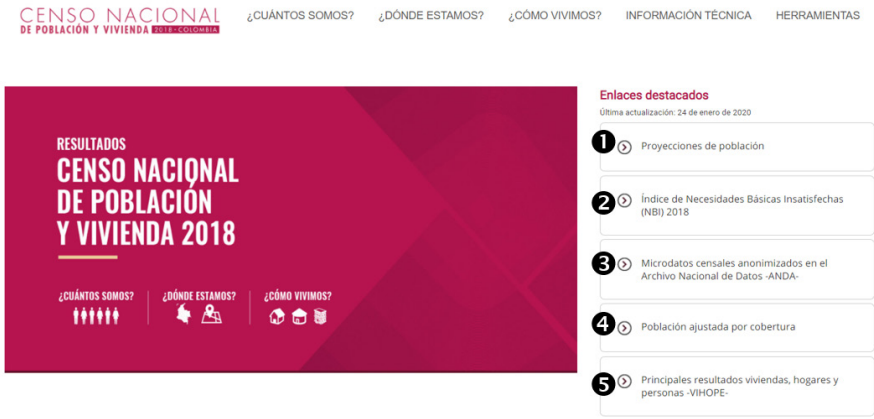


Figura 46. Información disponible en el DANE sobre el Censo Nacional de Población y Vivienda 2018 (imagen capturada en Sep-09-2020).

La opción ❶ de la Figura 46 (Proyecciones de población) permite acceder a la información de proyecciones de población para Colombia para diferentes periodos (1950-1984, 1985-1992, 1993-2004, 2005-2017 y 2018-2070), niveles de agregación (nacional, departamental y municipal) y variables demográficas (área geográfica, sexo y edad). Un resumen de la información disponible se presenta en la Tabla 43.

Tabla 43. Resumen de la información disponible sobre proyecciones y retroproyecciones de población calculadas con base en los resultados del Censo Nacional de Población y Vivienda -CNPV- 2018.

Nivel de detalle	Descripción del contenido
Proyecciones y retroproyecciones de población **nacional**- para el periodo 1950-2017 y 2018-2070 con base en el CNPV 2018	Serie nacional de población por área, para el periodo: 2018-2070 / 2005-2017 / 1993-2004 / 1985-1992 / 1950-1984 Serie nacional de población por sexo, para el periodo: 2018-2070 / 2005-2017 / 1993-2004 / 1985-1992 / 1950-1984 Serie nacional de población por área, sexo y edad para el periodo: 2018-2070 / 2005-2017 / 1993-2004 / 1985-1992 / 1950-1984
Proyecciones y retroproyecciones de población **departamental**- para el periodo 1950-2017 y 2018-2070 con base en el CNPV 2018	Serie nacional de población por área, para el periodo: 2018-2070 / 2005-2017 / 1993-2004 / 1985-1992 / 1950-1984 Serie nacional de población por sexo, para el periodo: 2018-2070 / 2005-2017 / 1993-2004 / 1985-1992 / 1950-1984 Serie nacional de población por área, sexo y edad para el periodo: 2018-2070 / 2005-2017 / 1993-2004 / 1985-1992 / 1950-1984
Proyecciones y retroproyecciones de población **municipal**- para el periodo 1950-2017 y 2018-2070 con base en el CNPV 2018	Serie nacional de población por área, para el periodo: 2018-2070 / 2005-2017 / 1993-2004 / 1985-1992 / 1950-1984 Serie nacional de población por sexo, para el periodo: 2018-2070 / 2005-2017 / 1993-2004 / 1985-1992 / 1950-1984 Serie nacional de población por área, sexo y edad para el periodo: 2018-2070 / 2005-2017 / 1993-2004 / 1985-1992 / 1950-1984

Nota. Esta información es importante al momento de calcular indicadores demográficos y epidemiológicos. Le solicitamos familiarizarse con esta información para el momento de acción donde deberá construir indicadores demográficos y epidemiológicos. Elaboración propia

Estadísticas vitales

Algunos eventos suelen registrarse en la medida que ocurren, siendo relevante la información del tiempo y lugar de ocurrencia del hecho, además del lugar de residencia de la persona que presentó el evento. Los más importantes son los nacimientos, las defunciones, los matrimonios y los divorcios. Otros registros de información importantes son los registros educativos, registros poblacionales de cáncer, registros de seguridad social, entre otros.

Los registros vitales son estadísticas rutinarias que se recolectan sobre nacimientos y defunciones (fetales y no fetales) los cuales permiten observar cambios en los niveles de mortalidad y fecundidad de la población.

- La información sobre nacimientos es necesaria para calcular diversos indicadores como tasa de mortalidad infantil, tasa de natalidad, razón de mortalidad materna, entre otros.
- La información sobre mortalidad permite calcular tasas de mortalidad general o especificas por sexo, grupos etarios o causas de defunción, además de la mortalidad proporcional.

El DANE pone a disposición del público la información de las estadísticas vitales desde su portal principal (Figura 47), accediendo por la ruta: Estadísticas por tema (1) / Salud (2).

Figura 47. Acceso a estadísticas vitales del DANE (imagen capturada en Sep-09-2020).

Las estadísticas vitales se encuentran disponibles por "Nacimientos y defunciones" (3) como se muestra en la Figura 48.

Figura 48. Información de Salud disponible en el DANE (imagen capturada en Sep-09-2020).

En el componente de "Nacimientos y defunciones" (Figura 49) se tiene acceso a los microdatos (4) y a un sistema de consulta (5), este último para obtener tablas resumen o tablas cruzadas de los nacimientos y las defunciones de acuerdo con las variables contenidas en los certificados de nacimiento y defunción, respectivamente.

Figura 49. Información disponible de estadísticas vitales en el DANE (imagen capturada en May-05-2021).

Encuestas poblacionales

Las encuestas demográficas, también llamadas *encuestas poblacionales*, obtienen información de una muestra representativa de la población. En este tipo de estudios se evidencia el uso de la estadística inferencial, con especial énfasis en el diseño del muestreo, para la obtención de una muestra representativa de la población.

En Colombia el Ministerio de Salud y Protección Social es la entidad encargada de realizar los estudios y encuestas poblacionales para la obtención de información sobre diversos aspectos de interés para la salud pública, epidemiología y demografía, entre otros. Esta información es de vital importancia ya que permite caracterizar diversas dimensiones de la salud poblacional y además la elaboración de informes como el Análisis de Situación de Salud (ASIS).

Para ello el Ministerio de Salud y Protección Social, responsable de la investigación para la salud, estableció un Sistema Nacional de Estudios y Encuestas Poblacionales para la Salud con una agenda que prioriza los estudios requeridos por el país. Un resumen de los principales estudios y encuestas relacionadas con la salud, su periodicidad y última fecha de realización se presenta en la Tabla 44.

Tabla 44. Resumen de estudios y encuestas en salud en Colombia.

Encuesta/Estudio	Última	Periodicidad
Estudio Nacional de Salud Mental – ENSM	2015	Cada 10 años
Encuesta Nacional de Salud – ENS	2017	Cada 5 años
Encuesta Nacional de Parasitismo Intestinal en Población Escolar	2015	Cada 10 años
Encuesta Nacional de Salud Bucal IV - ENSAB IV	2014	Cada 10 años
Encuesta Mundial de Tabaquismo en Adultos – GATS	2016	Cada 5 años
Encuesta Nacional de Salud, Bienestar y Envejecimiento – SABE	2015	Cada 10 años
Encuesta de Salud Escolar - ENSE y Encuesta Nacional de Tabaquismo en Jóvenes - ENTJ	2016	Cada 3 años
Encuesta Nacional de Situación Alimentaria y Nutricional de los Pueblos Indígenas de Colombia – ENSIN INDIGENA	2013	Cada 5 años
Encuesta Nacional de Consumo de Sodio, Yodo y Flúor	2013	Cada 10 años
Estudio Consumo de Sustancias Psicoactivas en Hogares	2014	Cada 5 años
Encuesta Nacional de Demografía y Salud – ENDS	2016	Cada 5 años
Encuesta Nacional de Situación Nutricional en Colombia – ENSIN	2016	Cada 5 años
Análisis de la Situación de Salud - ASIS	2019	Anual
Estudio de Carga de Enfermedad	2010	Cada 10 años

Nota. SISPRO - Ministerio de Salud y Protección Social, Dirección de Epidemiología y Demografía, Cali, Octubre 9 de 2013.

El Sistema Nacional de Estudios y Encuestas Poblacionales para la Salud también estableció una muestra maestra en salud siendo representativa de la población nacional, departamental, regional, subregional o municipal.

En la Figura 48 se puede observar que el DANE brinda acceso a la "Encuesta nacional de consumo de sustancias psicoactivas (ENCSPA)" y la "Encuesta nacional de calidad de vida (ECV)".

Errores de las fuentes de información

Es importante mencionar que el uso de cualquier fuente de información (censos, estadísticas vitales o encuestas poblacionales) siempre están sujetas a dos tipos de errores: cobertura (subregistro) y calidad de la información.

Otras fuentes de información: Nacionales

Las siguientes fuentes de información también pueden serle útiles en su vida profesional:

Tabla 45. Fuentes de información.

Entidad	Breve descripción
DATOS ABIERTOS GOBIERNO DIGITAL COLOMBIA https://www.datos.gov.co/	La iniciativa de Datos Abiertos del Gobierno Nacional busca que todas las entidades del sector público pongan a disposición la información pública. La información se comparte en formatos digitales estructurados para que puedan ser usados por la sociedad general con el fin de generar informes, reportes, estadísticas, investigaciones, control social, entre otros temas.
INSTITUTO NACIONAL DE SALUD https://www.ins.gov.co/	El Instituto Nacional de Salud (INS) es una entidad pública cuya finalidad es la investigación y vigilancia de problemas de salud pública prioritarios para el país. Esta entidad es responsable de proponer políticas y normas, además de promover y desarrollar investigación y brindar servicios de salud en los campos de salud pública, control de enfermedades transmisibles y no transmisibles, alimentación y nutrición, producción de biológicos, control de calidad de alimentos, productos farmacéuticos y afines, salud ocupacional, protección del medio ambiente y salud intercultural, para contribuir a mejorar la calidad de vida de la población. A través del INS se tiene acceso a la información de todos los eventos de notificación obligatoria a través del Sivigila (Sistema Nacional de Vigilancia en Salud Pública).

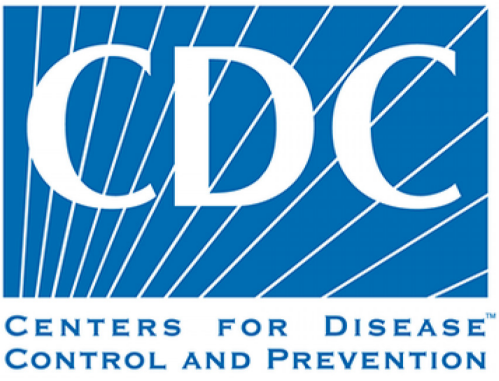
SISPRO Sistema Integrado de Información de la Protección Social https://www.sispro.gov.co	Sispro es un sistema de información conformado por bases de datos de cuatro sectores: salud, pensión, riesgo laboral y promoción social. Cada componente está basado en módulos, que en el caso de la salud, son epidemiología, información ambiental, recursos para la salud y cuentas de salud.
Así Vamos en Salud® https://www.asivamosensalud.org/	Es una página gubernamental que hace seguimiento a varios indicadores del sector salud que son relevantes de monitoreo. "Así vamos en salud" realiza el análisis de indicadores e índices con base en fuentes oficiales y dispone de información a nivel nacional, departamental y en algunos casos a nivel municipal.

Otras fuentes de información: Internacionales

Tabla 46. Fuentes de información Internacional.

Entidad	Breve descripción
World Health Organization Health statistics and information systems Health data and statistics https://www.who.int/healthinfo/statistics/en/	Sitio de la Organización Mundial de la Salud (OMS) que brinda acceso a varias consultas de indicadores a nivel de país. Desde aquí se puede acceder a: • THE GLOBAL HEALTH OBSERVATORY • Global Health Estimates (GHE) • WHO Mortality Database
THE GLOBAL HEALTH OBSERVATORY https://www.who.int/data/gho	Sitio de la Organización Mundial de la Salud (OMS) con una amplia gama de indicadores para los países del mundo.
Health statistics and information systems Global Health Estimates (GHE) https://www.who.int/healthinfo/global_burden_disease/en/	Sitio de la OMS con datos demográficos como expectativa de vida, mortalidad infantil, causas de muerte, entre otros.

 https://apps.who.int/healthinfo/statistics/mortality/whodpms/	Sistema de consulta de la OMS sobre datos de mortalidad por países.
 https://www.paho.org/data/index.php/es/	Sitio de la Organización Panamericana de la Salud (OPS) que brinda acceso a PLISA (Plataforma de Información en Salud para las Américas)
 https://www.prb.org/	"Population Reference Bureau" (PRB) es una organización privada sin ánimo de lucro que se especializa en recopilar y suministrar estadísticas necesarias para fines de investigación y/o académicos centrados en el medio ambiente, la salud y la estructura de las poblaciones. En esta página se encuentran indicadores demográficos de los países del mundo.
 https://gco.iarc.fr/	El Observatorio Global del Cáncer (GCO) es una plataforma web interactiva que presenta estadísticas globales del cáncer para informar el control y la investigación del cáncer. Aquí se pueden consultar estadísticas de incidencia, mortalidad y prevalencia (casos, tasas crudas y estandarizadas) para diferentes tipos de cáncer y por sexo y edad.
 http://www.healthdata.org/	El Institute for Health Metrics and Evaluation (IHME) es un instituto de investigación de la Universidad de Washington en Seattle que trabaja en el área de estadísticas de salud global y evaluación de impacto. Aquí se pueden consultar diversos indicadores a nivel de país.

 https://www.cdc.gov/DataStatistics/	El CDC (centro para el control y prevención de enfermedades) brinda acceso a una gran variedad de datos y estadísticas de diversos temas relacionado con la salud.
 https://www.cdc.gov/tobacco/global/gtss/gtssdata/index.html	GTSSData alberga y muestra datos de cuatro encuestas relacionadas con el tabaco realizadas en todo el mundo. El propósito de GTSSData es mejorar la capacidad de los países para monitorear el consumo de tabaco, orientar los programas nacionales de prevención y control del tabaco y facilitar la comparación de datos relacionados con el tabaco a nivel nacional, regional y mundial.

4.2 Indicadores demográficos

Recordemos que la Demografía estudia los movimientos que ocurren en las poblaciones humanas. En el estudio de esas poblaciones se utiliza una serie de indicadores para cuantificar su comportamiento en relación con su estado y dinámica en el tiempo, además que estos permiten la comparación entre varias poblaciones.

4.2.1 Principales indicadores: demográficos, mortalidad y morbilidad

Como se mencionó, el estado de la población se refiere a su tamaño, distribución geográfica y estructura por edad, sexo u otras agrupaciones de interés. Por otra parte,

la dinámica de las poblaciones estudia tres componentes determinantes del cambio poblacional: la fecundidad, la mortalidad y la migración.

En la Demografía y en la Epidemiología se utilizan diversos indicadores que se pueden clasificar de la siguiente forma:

1. Desde la Demografía, en indicadores de estado y de movimiento (ver Tabla 47).
2. Desde la Epidemiología, en indicadores de mortalidad y morbilidad.

Tabla 47. Principales indicadores demográficos: de estado y de movimiento.

Indicadores de estado	Indicadores de movimiento
1. Tamaño de la población	1. Tasa de natalidad
2. Composición por sexo y edad	2. Tasa bruta de mortalidad
3. Composición:	3. Tasa de crecimiento
• Menores de 15 años	4. Tiempo de duplicación de la población
• De 15-19 años	5. Tasa de mortalidad infantil
• Mayores de 65 años	6. Tasa de fecundidad general
• Mujeres en edad fértil (15-49 años)	7. Tasa global de fecundidad
4. Composición por zona: urbana/rural	8. Esperanza de vida al nacer
5. Densidad poblacional	
6. Razón de masculinidad	
7. Razón de niños-mujeres	
8. Índices de infancia, juventud, vejez, envejecimiento	
9. Índices de dependencia infantil, por vejez, general	
10. Índice de Friz	
11. Tasa de alfabetismo	

Nota. Elaboración propia.

La forma más simple de obtener medidas poblacionales son los datos absolutos (conteos o frecuencias absolutas). Sin embargo, estas cifras pierden su utilidad cuando se quieren realizar comparaciones entre diferentes grupos poblacionales, regiones o países con diferentes tamaños de población. Por ejemplo, se puede afirmar que en una comunidad se presentaron 2 000 muertes, pero el valor de 2 000 no dice nada si se desconoce la población y su tamaño.

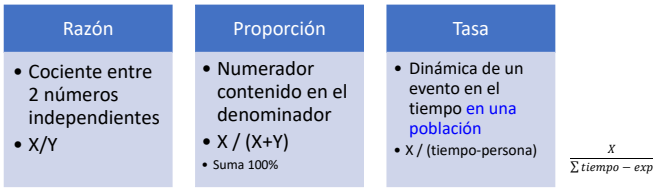
Para solucionar esto se han generado medidas relativas que no dependen del tamaño poblacional y permiten comparaciones más sensatas.

La demografía y la epidemiología utilizan indicadores básicos para medir respectivamente la dinámica de las poblaciones y el estado de salud (frecuencia de la mortalidad y la enfermedad o morbilidad). La mayoría de estos indicadores se

construyen como el cociente o división de dos cantidades que en matemáticas se llaman: numerador y denominador:

$$\frac{\text{Numerador}}{\text{Denominador}}$$

La relación que guardan el numerador y el denominador definen tres conceptos matemáticos conocidos como *razón, proporción* y *tasa*.



Razón

Una razón es una medida que compara dos cantidades ***X*** y ***Y*** por medio de una división. Es el cociente entre dos cantidades numéricas (X/Y) generalmente pertenecientes a diferentes grupos o de diversa naturaleza. En otras palabras, es el cociente entre dos números o dos cantidades diferentes:

$$\text{Razón} = \frac{\text{X}}{\text{Y}}$$

Su resultado se interpreta como el número de unidades de ***X*** por cada unidad de ***Y***.

Ejemplo 30:

$$\text{Índice de masculinidad} = \frac{\text{Total hombres}}{\text{Total mujeres}}$$

Las estimaciones en Colombia para el 2020 reportan un total de 24 594 882 hombres y 25 777 542 mujeres. El Índice de masculinidad:

$$\text{Índice de masculinidad} = \frac{24.594.882}{25.777.542} = 0{,}954$$

Indica que por cada mujer hay 0,954 hombres. Si se multiplica por 100 su interpretación es más sencilla indicando que por cada 100 mujeres hay aproximadamente 95 hombres en Colombia 2020. Esta razón hombre: mujer también puede reportarse como 95,4:100, indicando que existen 95,4 hombres por cada 100 mujeres.

Ejemplo 31:

En Colombia durante el año 2018 se registraron 649 115 nacimientos y 236 932 defunciones. La relación:

$$\text{Relación}\frac{\text{nacimientos}}{\text{defunciones}} = \frac{649.115}{236.932} = 2{,}74$$

Indica que por cada defunción se presentan 2,74 nacimientos. Si se multiplica por 100 se dice que por cada 100 defunciones ocurren 274 nacimientos.

Proporción

Una proporción es el cociente de dos magnitudes de un mismo evento. Se caracteriza porque el numerador está contenido en el denominador.

$$\text{Proporción} = \frac{X}{X+Y}$$

La frecuencia relativa (h_i) que se vio en el módulo de Bioestadística Descriptiva ($h_i = f_i / n$) hace referencia a una proporción, representando el peso relativo de $\boldsymbol{X}$ dentro del total ($\boldsymbol{X}$+$\boldsymbol{Y}$).

Una proporción es una medida que expresa la frecuencia de un evento con relación al total observado. Nótese que el denominador es el número de sujetos observados y NO incluye "el tiempo". Una proporción está entre cero y uno o, si multiplicamos por 100, entre el 0 % al 100%.

Una proporción indica cuánto representa una categoría dentro de un gran total. Es el mismo concepto de una frecuencia relativa o porcentaje.

Ejemplo 32:

Tabla **48**. Distribución del hábito de cigarrillo en los participantes del Estudio del Corazón de Framingham.

¿Fuma?	Número	%
Si	2 181	49,2
No	2 253	50,8
Total	4 434	100

La información de la Tabla 48 indica que el 49,2 % de los participantes del "Estudio del Corazón de Framingham" manifestaron fumar, mientras que el 50,8 % dijo que no. Para las personas que sí fuman su cálculo se hace como 2 181/4 434. Como el denominador (4 434) incluye al numerador (2 181) lo convierte en una proporción.

Ejemplo 33:

En Colombia durante el año 2018 se registraron 649 115 nacimientos de los cuales el tipo de parto fue espontáneo en 354 133 de ellos, por cesárea 288 000 nacimientos y

otros (6 982). La proporción de nacimientos por parto espontáneo fue 354 133/649 115 = 0,546 lo que significa que el 54,6 % de los nacimientos fue por parto espontáneo.

Tasa

Una tasa es un tipo de proporción que considera el tiempo.

$$Tasa = \frac{X}{X+Y}\ (tiempo) \qquad \text{o} \qquad Tasa = \frac{X}{\sum tiempo-exposición}$$

Es el cociente entre el número de eventos en un área y periodo específico (X) sobre el tiempo que estuvo la población expuesta al riesgo de presentar el evento (∑tiempo-persona). Se caracteriza porque tanto el numerador como el denominador tienen implícito el tiempo. Es una medida que expresa la frecuencia (velocidad) con la cual ocurre un evento en el tiempo.

$$Tasa = \frac{Nº\ eventos\ en\ área\ y\ período}{Tiempo\ que\ las\ personas\ sanas\ estuvieron\ expuestas\ al\ riesgo}$$

A nivel poblacional nunca se conocerá el denominador, el cual es reemplazado por la población a mitad del periodo, representando ∑tiempo-persona (persona-año).

La tasa es una medida relativa de fenómenos como la *mortalidad* y la *morbilidad* con la ventaja que permite hacer comparaciones en el tiempo y entre regiones.

Una tasa se interpreta como la frecuencia relativa con la que sucede el evento en una población y tiempo específico. Algunas veces el valor de la tasa es un valor muy pequeño: entonces, para facilitar su interpretación, es común expresar una tasa en múltiplos de 10, 100, 1 000, 10 000 o 100 000, entre otros.

La tasa es una de las medidas más usadas en el estudio de poblaciones en áreas como la demografía, epidemiología y salud pública.

Ejemplo 34:

En Colombia durante el año 2018 se registraron 649 115 nacimientos y 236 932 defunciones. La población a mitad de periodo de ese año fue de 48 258 494 personas.

a. La tasa de mortalidad general (TMG):

$$\text{TMG} = \frac{\text{Nº defunciones del periodo}}{\text{Total población a mitad de periodo}} = \frac{236.932}{48.258.494} = 0{,}00491$$

La TMG suele expandirse por 1 000 (0,00491 × 1 000 = 4,91). Luego en Colombia para el año 2018 ocurrieron 4,91 (aproximadamente 5) muertes por cada mil personas.

b. La tasa de natalidad (TN):

$$TN = \frac{\text{Nº nacimientos}}{\text{Total población a mitad de periodo}} = \frac{649.115}{48.258.494} = 0{,}01345$$

La TN puede expandirse por 1 000 (0,01345 × 1 000 = 13,45). Luego en Colombia para el año 2018 ocurrieron13,45 nacimientos por cada mil personas.

Ejemplo 35:

Durante el 2018 en una comunidad se presentaron 80 casos de malaria en una población de 12 000 habitantes.

$$\text{Tasa} = \frac{80}{12.000} = 0{,}0067$$

Las siguientes interpretaciones son todas equivalentes:

- 80 casos de malaria entre 12 000 habitantes durante 2018.
- 0,0067 casos de malaria por cada habitante durante 2018.
- 6,7 casos de malaria × 1 000 habitantes durante 2018.
- 66,7 casos de malaria × 10 000 habitantes durante 2018.
- 666,7 casos de malaria × 100 000 habitantes durante 2018.

4.2.2 Indicadores demográficos importantes

En la Tabla 47 se presentaron algunos de los principales indicadores demográficos (de estado y de movimiento). A continuación, se describen los más importantes.

1. Estructura poblacional: Es el análisis de la distribución de la población por sexo y edad. Su análisis se muestra usando pirámides poblacionales ya sea con edades simples o agrupadas.
2. Población total o tamaño de la población: Es el número de habitantes en una determinada comunidad, región o país.
3. Composición por zona: Indica el número absoluto y porcentaje de:
    - Personas que residen en zonas urbanas.
    - Personas que residen en zonas rurales.

4. Densidad poblacional: Es la razón entre el número de personas de una comunidad y su superficie total. Representa el número de personas por kilómetro cuadrado (km^2).

5. Población por sexo: Indica el número absoluto y porcentaje de:

   - Población femenina.
   - Población masculina.

6. Composición por sexo y edad de grupos poblacionales de interés: Indica el número absoluto y porcentaje de población:

   - Menores de 15 años (índice de infancia).
   - Adolescente (15-19 años).
   - Mujeres en edad fértil (15-49 años).
   - Mayores de 65 años (índice de vejez).

7. Razón de masculinidad, índice de masculinidad o relación hombre/mujer: Es la relación entre el número de hombres y de mujeres en la población. Se puede expresar como el número de varones por cada 100 mujeres.

$$\text{Razón hombre} - \text{mujer} = \frac{\text{Total hombres}}{\text{Total mujeres}} \times 100$$

8. Razón de niños-mujeres: Expresa el número de niños menores de 5 años por 100 mujeres de 15 a 49 años de la población.

$$\text{Razón niños} - \text{mujeres} = \frac{\#\ \text{niños} < 5\ \text{años}}{\text{Nº mujeres edad reproductiva } (15 - 49\text{a})} \times 100$$

9. Índice de infancia: Representa la relación entre los menores de 15 años y la población total. Se interpreta como el porcentaje de la población que es menor a 15 años.

$$\text{Índice de infancia} = \frac{\#\ \text{niños} < 15\ \text{años}}{\text{Población total}} \times 100$$

10. Índice de juventud: Representa la relación entre la cantidad de personas entre 15 y 29 años y la población total. Se interpreta como el porcentaje de la población que se encuentra en el rango de 15 a 29 años.

$$\text{Índice de juventud} = \frac{\#\ \text{personas } 15 - 29\ \text{años}}{\text{Poblacion total}} \times 100$$

11. Índice de vejez: Cociente entre el número de personas de 65 y más años y la población total. Se interpreta como el porcentaje de la población que es mayor a 65 años.

$$\text{Índice de vejez} = \frac{\text{Población} > 65 \text{ años}}{\text{Poblacion total}} \times 100$$

12. Índice de envejecimiento: Cociente entre el número de personas de 65 años y más y el número de personas entre 0 y 14 años.

$$\text{Índice de envejecimiento} = \frac{\text{Población} > 65 \text{ años}}{\text{Poblacion } 0 - 14 \text{ años}} \times 100$$

13. Índice de dependencia infantil: Cociente entre el número de personas de 0 a 14 años y el número de personas entre 15 y 64 años.

$$\text{Índice de dependencia infantil} = \frac{\text{Pobl.} < 15 \text{ años}}{\text{Pobl. } 15 - 64 \text{ años}} \times 100$$

14. Índice de dependencia por vejez también llamado índice de dependencia en mayores o del adulto mayor. Es el cociente entre el número de personas de 65 años y más y el número de personas entre 15 y 64 años.

$$\text{Índice de dependencia en mayores} = \frac{\text{Pobl.} > 65 \text{ años}}{\text{Pobl. } 15 - 64 \text{ años}} \times 100$$

15. Índice de dependencia económica también llamado índice demográfico de dependencia, índice global de dependencia o razón de dependencia. Es la relación entre personas dependientes (menores de 15 años y mayores de 64 años) y personas económicamente activas (15-64 años) dentro de la población.

$$\text{Índice de dependencia} = \frac{\text{Pobl.} < 15 + \text{Pobl.} > 65 \text{ años}}{\text{Pobl. } 15 - 64 \text{ años}} \times 100$$

16. Índice de Friz: Cociente entre el número de personas de 0 a 19 años y el número de personas de 30 a 49 años.

$$\text{Índice de Friz} = \frac{\text{Pobl. } 0 - 19 \text{ años}}{\text{Pobl. } 30 - 49 \text{ años}} \times 100$$

Interpretación: Si el Índice de Friz es mayor a 160 se considera que la población es joven; valores entre 60-160 indican una población madura; y valores menores a 60 indican una población vieja.

17. Tasa de alfabetismo: Indicador del nivel educativo de una población. Expresa el porcentaje de personas mayores a 10 años que son alfabetos.

$$\text{TA} = \frac{\text{Poblacion alfabeta} \geq 10a}{\text{Poblacion} \geq 10a}$$

18. Tasa de natalidad (TN): Expresa el número de nacimientos registrados por cada mil habitantes en una región y periodo determinado (normalmente un año).

$$\text{TN} = \frac{\text{Nº nacidos vivos del período}}{\text{Población total a mitad de período}} \times 1.000$$

19. Tasa bruta de mortalidad (TBM) o tasa de mortalidad general (TMG): Indica la relación entre el número de defunciones en un periodo y la población media del mismo periodo. La TMG expresa el número de muertes por cada mil habitantes.

$$\text{TMG} = \frac{\text{Nº defunciones del período}}{\text{Total población a mitad de período}} \times 1.000$$

20. Tasa de mortalidad infantil (TMI): Indica el número de defunciones en niños menores de un año por cada mil nacidos vivos en un periodo de tiempo.

$$\text{TMI} = \frac{\text{Nº defunciones} < 1 \text{ año del período}}{\text{Total nacidos vivos del período}} \times 1.000$$

21. Tasa de fecundidad general (TFG): Es un indicador de fecundidad que expresa la relación entre el número de nacimientos ocurridos en un periodo y la población femenina en edad fértil. En otras palabras, es el número de nacidos vivos por cada 1 000 mujeres en edad reproductiva (15 a 49 años).

$$\text{TFG} = \frac{\text{Nº nacimientos del período}}{\text{Nº mujeres edad reproductiva } (15 - 49a)} \times 1.000$$

22. Tasa global de fecundidad por mujer: Número promedio esperado de hijos que habría de tener una mujer durante su vida, si en el transcurso de sus años reproductivos experimentase las tasas de fecundidad específicas por edad prevalentes en un determinado año o periodo, para un determinado país, territorio o área geográfica.

23. Esperanza de vida al nacer o expectativa de vida al nacer (EVN): Este es un indicador que expresa la media de cantidad de años que vivirá un recién nacido si los patrones de mortalidad vigentes se mantienen constantes en dicha población.

Ejemplo 36:

La información de la Tabla 49 se obtuvo del DANE, con la finalidad de calcular algunos indicadores demográficos:

Tabla 49. Algunos datos demográficos de los departamentos de Valle del Cauca y Chocó para los años 2015 y 2019.

Descripción	Valle del Cauca		Chocó	
	2015	2019	2015	2019
Población total	4 397 194	4 506 768	509 240	539 933
Hombres	2 105 067	2 138 822	251 669	266 585
Mujeres	2 292 127	2 367 946	257 571	273 348
Menores de 15 años	1 074 088	1 049 759	179 823	181 487
Entre 15-64 años	2 915 317	2 978 528	302 951	326 357
Mayores de 65 años	407 789	478 481	26 466	32 089
Niños menores de 5 años	345 407	342 009	61 818	60 277
Adolescente (15-19 años)	374 362	368 293	50 862	54 708
Mujeres en edad fértil (15-49 años)	1 158 071	1 170 112	128 958	137 629
Personas de 15-29 años	374 362	368 293	50 862	54 708
Superficie (km^2)	22 195		46 530	
Número de nacidos vivos	53 478	48 735	5 906	5 956
Número de defunciones	27 133	28 443	1 080	1 210
Número de defunciones en menores de 1 año	617	521	136	128

Nota. Elaboración propia a partir de información del DANE en sus componentes: Proyecciones de población y Estadísticas Vitales.

Solución:

Utilizando la información de la tabla 49 se calcularon algunos indicadores demográficos que se presentan en la Tabla 50. El Valle del Cauca presentó más habitantes por kilómetro cuadrado; para el año 2019 la densidad poblacional fue de 203,1 en el Valle del Cauca y de 11,6 en Chocó.

En 2019, el 47,5 % de la población del Valle del Cauca era masculina y el 52,5 % femenina, mientras que en Chocó los porcentajes fueron más similares (49,4 %

hombres y 50,6 % mujeres). De aquí, la razón de masculinidad (hombres por cada 100 mujeres) fue de 90,3 en Valle del Cauca y 97,5 en Chocó.

Por medio del índice de infancia y el índice de vejez se aprecia que Chocó es un departamento con su población más joven, mientras que en el Valle del Cauca es más envejecida. En 2019 el índice de infancia fue del 33,6 % en Chocó y 23,3 % en Valle del Cauca, mientras que el índice de envejecimiento fue de 5,9 % y 10,6 %, respectivamente. Así mismo, en ambos departamentos los índices de infancia mostraron un descenso, mientras que los índices de vejez y envejecimiento aumentaron durante el periodo 2015-2019. La razón de niños-mujeres (por 100) fue mayor en Chocó (43,8) que en el Valle del Cauca (29,2) para el año 2019.

Coherente con la estructura poblacional de ambos departamentos, en 2019 el Chocó mostró mayor índice de dependencia infantil (55,6 %) que el Valle del Cauca (35,2 %), así como mayor índice de dependencia económica (65,4 % vs 51,3 %, respectivamente).

Finalmente, se observó que las tasas de natalidad de ambos departamentos mostraron un descenso durante el periodo 2015-2019. En 2019 las tasas brutas de mortalidad (por 1 000) fueron mayores en el Valle del Cauca (6,3) que en Chocó (2,2), mientras que la mortalidad infantil fue claramente más alta en Chocó (21,5) que en el Valle del Cauca (10,7).

Tabla 50. Indicadores demográficos de los departamentos de Valle del Cauca y Chocó para los años 2015 y 2019

Descripción	Valle del Cauca		Chocó	
	2015	2019	2015	2019
Densidad poblacional (habitantes por km^2)	198,1	203,1	10,9	11,6
Hombres (%)	47,9	47,5	49,4	49,4
Mujeres (%)	52,1	52,5	50,6	50,6
Índice de infancia (% de menores de 15 años)	24,4	23,3	35,3	33,6
Entre 15-64 años (%)	66,3	66,1	59,5	60,4
Índice de vejez (% de mayores de 65 años)	9,3	10,6	5,2	5,9
Razón de masculinidad (%)	91,8	90,3	97,7	97,5
Razón de niños-mujeres (por 100)	29,8	29,2	47,9	43,8
Índice de juventud (%)	8,5	8,2	10,0	10,1
Índice de envejecimiento (%)	38,0	45,6	14,7	17,7
Índice de dependencia infantil (%)	36,8	35,2	59,4	55,6
Índice de dependencia por vejez (%)	14,0	16,1	8,7	9,8
Índice de dependencia económica (%)	50,8	51,3	68,1	65,4
Tasa de natalidad (por 1.000)	12,2	10,8	11,6	11,0
Tasa bruta de mortalidad (por 1.000)	6,2	6,3	2,1	2,2
Tasa de mortalidad infantil (por 1.000)	11,5	10,7	23,0	21,5

Nota. Elaboración propia a partir de información del DANE en sus componentes: Proyecciones de población y Estadísticas Vitales.

4.2.3 Análisis de la mortalidad

La mortalidad es quizás uno de los eventos más estudiados y de mayor interés mundialmente donde se materializan condiciones biológicas, culturales, sociales y estructurales de una sociedad. Su análisis, según variables como causa de muerte, sexo y edad es vital para comprender el perfil demográfico y epidemiológico de una región y sus tendencias, o para evidenciar desigualdades en salud entre grupos poblacionales cuando se comparan por características sociales, económicas y geográficas.

El análisis de la mortalidad a través de indicadores, involucra la construcción de un grupo de tasas (general, especificas, crudas o estandarizadas) que objetivamente dan cuenta de las muertes ocurridas en un área, periodo y causa determinada, ya sea en la población general o en grupos de esta (13).

Tasa de mortalidad general y especificas

La medida más simple y general de la mortalidad en una población es la *tasa de mortalidad general* (TMG), también llamada *tasa bruta de mortalidad* (TBM), que mide la frecuencia relativa de defunciones, con respecto al tamaño poblacional, en un área y tiempo determinado (regularmente en un periodo de un año) (23).

$$\text{TMG} = \frac{\text{Nº defunciones del período}}{\text{Total población a mitad de período}} \times 1.000$$

De allí su cálculo se realiza dividiendo el número de muertes del periodo y área, entre el tamaño medio de la población en ese periodo y área, por cada mil personas (×1 000). Para que ambos elementos de la tasa sean comparables, y las tasas reflejen la realidad, el componente de área debe corresponder al lugar de residencia. Así está implícito en las normas de la Comisión de Estadística de las Naciones Unidad donde se recomienda enumerar las muertes según el lugar de residencia del fallecido o de la madre para muertes fetales o en menores de un año (23). Sin embargo, en el caso de muertes por causas externas se debe usar el lugar de ocurrencia.

En el Ejemplo 34 se calculó la tasa de mortalidad general para Colombia en 2018. Su cálculo, al utilizar como denominador a la población total de Colombia, hace que la TMG sea una medida gruesa de la mortalidad. Por esta razón la TMG también se conoce como *tasa bruta de mortalidad*.

Sin embargo, es lógico pensar que la población general no siempre está expuesta al mismo riesgo, y muchos eventos varían con el sexo o la edad. Por ejemplo, los adultos mayores tienen un mayor riesgo de mortalidad, solo las mujeres están a riesgo de mortalidad materna o solo los hombres presentan riesgo de cáncer de próstata. Luego, se requieren de tasas más refinadas que consideren grupos más específicos.

Esas tasas son llamadas **tasas específicas de mortalidad** cuyos denominadores solo incluyen a grupos poblacionales y no a la población total. Mientras que la mortalidad general involucra el volumen total de muertes (por todas las causas, todos los grupos de edad y ambos sexos) la mortalidad específica se puede calcular por subgrupos. Así, es común calcular tasas específicas de mortalidad por sexo, grupos etarios, o causas de muerte.

Por ejemplo, la tasa de mortalidad especifica por sexo:

$$\text{Tasa de mortalidad masculina} = \frac{\text{N° defunciones masculinas en un año}}{\text{Población masculina en ese año}} \times 1.000$$

$$\text{Tasa de mortalidad femenina} = \frac{\text{N° defunciones femeninas en un año}}{\text{Población femenina en ese año}} \times 1.000$$

4.2.4 Métodos de estandarización de tasas

No es adecuado usar indicadores generales (por ejemplo, tasa de mortalidad general o tasa de morbilidad general) para comparar poblaciones, básicamente porque las poblaciones son diferentes en muchos aspectos: estructura poblacional, nivel de desarrollo económico y social, condiciones de vida, entre otros. Las variables que afectan directa o indirectamente a los fenómenos demográficos, o incluso las relaciones causales en epidemiología, reciben el nombre de *variables confusoras*. Luego los investigadores deben considerar su ajuste y considerar estrategias para eliminar su efecto.

Son muchas las variables que afectan los fenómenos demográficos dentro de las cuales se pueden considerar la estructura por sexo y edad, índice de desarrollo humano, nivel educativo, nivel socioeconómico, entre otras. Sin embargo, se reconoce que la variable que ejerce mayor efecto sobre los indicadores demográficos y de salud es la estructura por edad, la cual debe controlarse para eliminar su efecto.

En demografía existen varias alternativas para controlar el efecto de las variables confusoras sobre un indicador. Una de las estrategias más sencillas y utilizadas son los métodos de estandarización de tasas: directo e indirecto. Luego, las tasas de mortalidad general se pueden estandarizar y obtener tasas de mortalidad estandarizadas por edad.

Estandarización directa

Sus bases matemáticas son las mismas de un promedio ponderado, donde $\bar{x}_w$ sería la tasa estandarizada, x_i correspondería a las tasas específicas para cada grupo de edad y w_i el peso relativo de cada grupo etario proveniente de una población referencia. La siguiente fórmula se usa para obtener un promedio ponderado de un conjunto de datos x_i donde cada uno de ellos tiene un peso relativo diferente w_i.

$$\bar{x}_w = \frac{\sum w_i x_i}{\sum w_i}$$

Para estandarizar una tasa de mortalidad general por el método directo se requiere:

1. Desagregar la información por grupos etarios: mortalidad y población.
2. Calcular tasas de mortalidad especificas por grupo de edad (TE_i).
3. Determinar la distribución etaria de una población referencia (W_i).
4. Calcular la tasa de mortalidad estandarizada (TME), así:

$$TME = \sum TE_i w_i$$

Se recomienda que la población estándar tenga una estructura etaria intermedia de las poblaciones bajo estudio. Generalmente la población de referencia se elige según las comparaciones que se van a realizar. Por ejemplo, si las tasas se van a comparar entre países del mundo se debería usar la población mundial; si se comparan países de América Latina se usa la población de esta región; o si las comparaciones son entre departamentos de Colombia se usa como población estándar la de Colombia. Otra alternativa seria usar como población referencia a una de las poblaciones bajo estudio o incluso la estructura poblacional promedio de ellas.

También se recomienda que los grupos etarios que se usen sean lo más pequeños posible. Como la mortalidad en los primeros años de vida es muy variable se sugiere separar los menores de 1 año, de 1-4 años y luego grupos quinquenales (5-9; 10-14; 15-19, …, 75-79 y 80 y más).

Ejemplo 37:

Datos de mortalidad, población y tasas de mortalidad general (TMG) en Chocó y Valle del Cauca, en el año 2015:

Dato	Chocó	Valle del Cauca
Defunciones	1 380	26 152
Población	500 093	4 613 684
TMG (× 1.000)	2,76	5,67

a. Paso 1: Desagregar el total de muertes (1 380 para Chocó y 26 152 para Valle del Cauca) y la población (500 093 para Chocó y 4 613 684 para Valle del Cauca) por grupos etarios. Ver Tabla 51.

b. Paso 2: Calcular las tasas de mortalidad especificas por grupo etario (TE_i). Ver

Tabla 51.

Tabla 51. Datos de mortalidad, población y TE para Choco y Valle del Cauca, año 2015.

Chocó	Muertes	Población	Tasa especifica (× 1000)
0-4	247	65 738	3,76
5-9	15	63 392	0,24
10-19	60	114 775	0,52
20-29	111	98 929	1,12
30-39	91	53 065	1,71
40-49	89	37 425	2,38
50-59	147	32 089	4,58
60-69	151	19 198	7,87
70-79	186	10 868	17,11
80 y más	283	4 614	61,34
TOTAL	**1 380**	**500 093**	**2,76**

Valle	Muertes	Población	Tasa especifica (× 1000)
0-4	613	363 944	1,68
5-9	76	358 356	0,21
10-19	831	754 291	1,10
20-29	1 701	788 145	2,16
30-39	1 379	677 893	2,03
40-49	1 465	592 252	2,47
50-59	2 469	512 454	4,82
60-69	3 654	318 078	11,49
70-79	5 265	173 151	30,41
80 y más	8 699	75 120	115,80
TOTAL	**26 152**	**4 613 684**	**5,67**

c. Paso 3: Se usa como población de referencia a Colombia para el año 2005. El ponderador (w_i) corresponde a la frecuencia relativa o peso relativo de cada grupo etario para Colombia.

Colombia, 2005	Población	w_i
0-4	4 343 774	0,101
5-9	4 465 233	0,104
10-19	8 670 098	0,202
20-29	7 094 902	0,165
30-39	5 992 989	0,140
40-49	5 142 532	0,120
50-59	3 363 611	0,078
60-69	2 076 637	0,048
70-79	1 237 739	0,029
80 y más	501 077	0,012
TOTAL	**42 888 592**	**1,000**

d. Paso 4: Se multiplica la tasa especifica por el ponderador ($TE_i w_i$) y la suma de esta columna es la tasa de mortalidad estandarizada por edad.

Chocó	TE_i (× 1000)	$TE_i w_i$	Wi	Valle	TE_i (× 1000)	$TE_i w_i$

0-4	3,76	0,38	0,101	0-4	1,68	0,17
5-9	0,24	0,02	0,104	5-9	0,21	0,02
10-19	0,52	0,11	0,202	10-19	1,10	0,22
20-29	1,12	0,19	0,165	20-29	2,16	0,36
30-39	1,71	0,24	0,140	30-39	2,03	0,28
40-49	2,38	0,29	0,120	40-49	2,47	0,30
50-59	4,58	0,36	0,078	50-59	4,82	0,38
60-69	7,87	0,38	0,048	60-69	11,49	0,56
70-79	17,11	0,49	0,029	70-79	30,41	0,88
80 y más	61,34	0,72	0,012	80 y más	115,80	1,35
TOTAL	**2,76**	**3,17**	**1,000**	**TOTAL**	**5,67**	**4,52**
	Tasa cruda	Tasa estandarizada			Tasa cruda	Tasa estandarizada

Luego, si el departamento de Chocó tuviera la misma estructura poblacional de Colombia en 2005, se esperarían 3,17 defunciones por cada mil personas, mientras que en el Valle del Cauca se presentarían 4,52 por cada mil personas. Note que la estandarización de las tasas mostró que en Chocó aumentó el riesgo mientras en el Valle del Cauca disminuyó.

Estandarización indirecta

Este método compara las muertes observadas (O) con las esperadas (E) si la población tuviera el mismo riesgo de una población de referencia.

En este método se siguen los siguientes pasos:

1. Desagregar la población por grupos etarios. Note que NO es necesario desagregar las defunciones por grupos etarios.

2. Definir el riesgo de muerte (tasas específicas de mortalidad) de una población de referencia.

3. Usar las tasas específicas de mortalidad de la población referencias para obtener el número de muertes esperadas (E_i) en cada grupo etario de cada población bajo estudio.

4. Se calcula la Razón de Mortalidad Estandarizada (REM), comparando las defunciones observadas y las esperadas.

$$\text{REM} = \frac{\text{Casos observados}}{\text{Casos esperados}} = \frac{\text{O}}{\text{E}}$$

Ejemplo 38:

Se usa la misma información del Ejemplo 37 con los datos de mortalidad, población y tasas de mortalidad general (TMG) en Chocó y Valle del Cauca, en el año 2015:

Dato	Chocó	Valle del Cauca
Muertes	1 380	26 152
Población	500 093	4 613 684
TMG (× 1.000)	2,76	5,67

a. Paso 1: Desagregar la población por grupos etarios:

Chocó	Muertes observadas (O)	Población
0-4		65 738
5-9		63 392
10-19		114 775
20-29		98 929
30-39		53 065
40-49		37 425
50-59		32 089
60-69		19 198
70-79		10 868
80 y más		4 614
TOTAL	**1 380**	**500 093**

Valle	Muertes observadas (O)	Población
0-4		363 944
5-9		358 356
10-19		754 291
20-29		788 145
30-39		677 893
40-49		592 252
50-59		512 454
60-69		318 078
70-79		173 151
80 y más		75 120
TOTAL	**26.152**	**4 613 684**

b. Paso 2: Se define el riesgo de muerte de la región Pacífico en el 2015 (población de referencia). Estas tasas de mortalidad especificas son las que se usan en el paso 3.

Región Pacífico	Muertes	Población	Tasa (× 1000)
0-4	1 416	727 338	1,947
5-9	143	711 710	0,201
10-19	1 262	1 467 262	0,860
20-29	2 628	1 420 208	1,850
30-39	2 187	1 180 573	1,852
40-49	2 288	978 693	2,338
50-59	3 628	815 552	4,449
60-69	5 382	518 085	10,388
70-79	7 828	291 068	26,894
80 y más	12 994	126 685	102,569
TOTAL	**39 756**	**8 237 174**	**4,826**

c. Paso 3: Se usa el riesgo de muerte de la población de referencia para obtener el número de muertes que se esperarían (E_i) en los grupos etarios de Chocó y Valle del Cauca:

		Chocó		Valle del Cauca	
Referencia	**Tasa (× 1000)**	**Población**	E_i	**Población**	E_i
0-4	1,947	65 738	128	363 944	709
5-9	0,201	63 392	13	358 356	72
10-19	0,860	114 775	99	754 291	649
20-29	1,850	98 929	183	788 145	1.458
30-39	1,852	53 065	98	677 893	1.256
40-49	2,338	37 425	87	592 252	1.385
50-59	4,449	32 089	143	512 454	2.280
60-69	10,388	19 198	199	318 078	3.304
70-79	26,894	10 868	292	173 151	4.657
80 y más	102,569	4 614	473	75 120	7.705
TOTAL	**4,826**	**500 093**	**1 716**	**4 613 684**	**23 474**

Luego en Chocó se esperaban 1 716 defunciones y en el Valle del Cauca 23 474 si estos departamentos tuvieran el mismo riesgo de la región Pacífico.

d. Paso 4: La RME se obtiene con la fórmula:

$$\text{REM} = \frac{\text{Casos observados}}{\text{Casos esperados}} = \frac{\text{O}}{\text{E}}$$

	Chocó	**Valle del Cauca**
Muertes observadas (O)	1 380	26 152
Muertes esperadas (E)	1 716	23 474
REM	0,804	1,114

El valor de la RME para Chocó fue del 80,4 % indicando que el riesgo de morir es 19,6 % menor en comparación con la región Pacífico controlando por la edad. Por el contrario, en el Valle del Cauca el riego fue 11,4 % mayor del esperado en la región Pacífico.

Otros indicadores importantes de mortalidad

Diferentes tasas de mortalidad suelen ser utilizados como indicadores de desarrollo y bienestar o como un reflejo del contexto social y económico de un país o región. Por ejemplo, las tasas de mortalidad general, tasa de mortalidad infantil (< 1 año), tasa de mortalidad de la niñez (<5 años), tasa de mortalidad materna, entre otros, son usados para evaluar la situación de salud. Además, otras tasas de mortalidad como las de homicidios o suicidios se consideran indicadores proxy del conflicto armado interno, del nivel de desintegración, de la falta de cohesión social y de condiciones

de vida deficientes de un país o región (24). Finalmente el análisis de la mortalidad puede ser complementado mediante el análisis de la esperanza de vida al nacer (23).

1. Letalidad

   - Tasa de letalidad:

$$\text{Tasa de letalidad} = \frac{\text{N° muertes de una enfermedad X en un período}}{\text{N° casos Dx de la enfermedad X en un período}} \times 1000$$

   - Letalidad (%):

$$\text{Letalidad (\%)} = \frac{\text{N° muertes de una enfermedad X}}{\text{N° casos Dx de la enfermedad X}} \times 100$$

2. Razón de mortalidad materna (RMM):

$$\text{RMM} = \frac{\text{N° muertes maternas durante un año}}{\text{Total nacidos vivos durante ese año}} \times 100.000$$

3. Tasa de mortalidad materna (TMM):

$$\text{TMM} = \frac{\text{N° muertes maternas durante un año}}{\text{Total mujeres edad reproductiva } (15 - 49a)} \times 100.000$$

La Figura 50 muestra los tiempos definidos para los mortinatos, la mortalidad perinatal, infantil, neonatal, neonatal temprana y tardía, y mortalidad post neonatal.

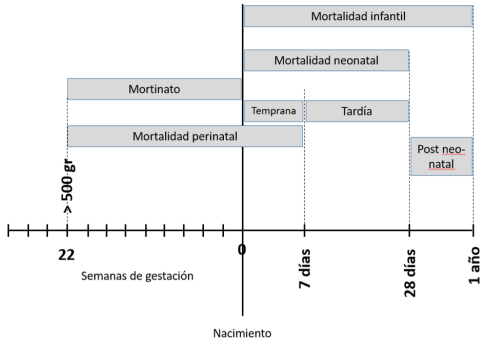


Figura 50. Tiempos útiles para la mortalidad infantil y los mortinatos. Elaboración propia

4. Tasa de mortalidad perinatal (TMP):

$$\text{TMP} = \frac{\text{N° mortinatos} + \text{muertes} < 2\ 8 \text{ días, en un año}}{\text{Total nacidos vivos}} \times 1.000$$

5. Tasa de mortalidad neonatal (TMN) (< 28 días):

$$\text{TMN} = \frac{\text{N° muertes} < 28 \text{ días durante un año}}{\text{Total nacidos vivos en ese año}} \times 1.000$$

- TMN temprana:

$$\text{TMN temprana} = \frac{\text{N° muertes} < 7 \text{ días durante un año}}{\text{Total nacidos vivos en ese año}} \times 1.000$$

- TMN tardía:

$$\text{TMN tardía} = \frac{\text{N° muertes entre 7 y 28 días durante un año}}{\text{Total nacidos vivos en ese año}} \times 1.000$$

6. Tasa de mortalidad post neonatal:

$$\text{TM post neonatal} = \frac{\text{N° muertes} > 28 \text{ días \& } < 1 \text{ año, durante un año}}{\text{Total nacidos vivos en ese año}} \times 1.000$$

7. Tasa de mortalidad infantil (< 1 año):

$$\text{TMI} = \frac{\text{N° muertes} < 1 \text{ año durante un año}}{\text{Total nacidos vivos en ese año}} \times 1.000$$

8. Tasa de mortalidad de la niñez o en menores de cinco años:

$$\text{Tasa mortalidad niñez} = \frac{\text{\# defunciones} < 5 \text{ años}}{\text{Total de nacimientos vivos}} \times 1.000$$

5. Bibliografía

1. Daniel WW. *BIOESTADISTÍCA. Base para el análisis de las diferencias de la salud*. 4a Edición. México: Limusa Wiley; 2007. 755 p.

2. Díaz RML. *Importancia de La Bioestadistica para investigacion en salud*. Rev Cuba Hematol Inmunol y Hemoter [Internet]. 2018;34(3):8–11. Available from: http://revhematologia.sld.cu/index.php/hih/article/view/872/804}

3. Calvache JA, Barón FJ, Shoemaker, RG. *La bioestadística y su aplicación a la investigación en salud*. Rev Fac Cienc Salud Univ Cauca [Internet]. 2006;8(3):56–9. Available from: https://revistas.unicauca.edu.co/index.php/rfcs/article/view/919

4. Ruiz A, Morillo LE. *Epidemiología Clínica. Investigación aplicada*. Primera ed. Bogotá D.C.: Editorial Médica Internacional; 2004. 576 p.

5. Hernández-Avila M, Garrido-Latorre F, López-Moreno S. *Diseño de estudios epidemiológicos*. Salud Publica Mex. 2000;42(2, marzo-abril).

6. Donis H JH. *Tipos de diseños de los estudios clínicos y epidemiológicos*. Av en Biomed [Internet]. 2013;2(2):76–99. Available from: https://www.redalyc.org/articulo.oa?id=331327989005

7. Hernández, B.; Velasco-Mondragón H. *Encuestas transversales*. Salud Publica Mex [Internet]. 2007;42(5):447–55. Available from: http://bvs.insp.mx/rsp/articulos/articulo.php?id=000640

8. Lazcano-Ponce EC, Fernández E, Salazar-Martínez E, Hernández-Avila M. *Estudios de cohorte. Metodologia, sesgos y aplicacion*. Salud Publica Mex. 2000;42(3):230–41.

9. Lazcano-Ponce E, Salazar-Martínez E, Hernández-Avila M. *Estudios epidemiológicos de casos y controles. Fundamento teórico, variantes y aplicaciones*. Salud Publica Mex [Internet]. 2001;43(2):135–50. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11381843

10. Calva-mercado JJ, Sc M, Jj C. *Estudios clínicos experimentales*. Salud Publica Mex. 2004;42(Volume 4):349–58.

11. Borja-Aburto VH. *Estudios ecológicos*. Salud Publica Mex. 2000;42(6):533–8.

12. Otzen T, Manterola C. *Técnicas de Muestreo sobre una Población a Estudio*. Int J Morphol. 2017;35(1):227–32.

13. Moreno-Altamirano A, López-Moreno S, Corcho-Berdugo A. *Principales medidas en epidemiología*. Salud Publica Mex. 2000;42(4):337–48.

14. Pértegas Díaz S, Pita Fernández S. *La distribución normal*. CAD ATEN PRIMARIA [Internet]. 2001;8:268–74. Available from: https://www.fisterra.com/mbe/investiga/distr_normal/distr_normal.asp

15. Webster AL. *Estadística aplicada a los negocios y la economía*. Tercera ed. Santafé de Bogotá: McGraw-HIll; 2000. 640 p.

16. Sánchez Turcios RA. t-Student. *Usos y abusos*. Rev Mex Cardiol. 2015;26(1):59–61.

17. Servizo Galego de Saúde. *Epidat 4: Ayuda de Demografía*. Octubre 2014. [Internet]. Octubre. 2014. Available from: https://www.sergas.es/Saude-publica/Documents/1896/Ayuda_Epidat_4_Demografia_Octubre2014.pdf

18. Centro Centroamericano de Población. Capacitación a distancia. *Curso Análisis Demográfico* [Internet]. Universidad de Costa Rica. [cited 2021 May 6]. Available from: https://ccp.ucr.ac.cr/cursos/demografia_03/

19. Carlos RE. *Transición epidemiológica en Colombia: de las enfermedades infecciosas a las no transmisibles*. Rev Ciencias Biomédicas. 2012;3(2).

20. Grajales A IC, Cardona A D. *La segunda transición demográfica y el nivel de desarrollo de los departamentos de Colombia*, 2005. Rev Fac Nac Salud Pública [Internet]. 2010;28(3):209–20. Available from: http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0120-386X2010000300002&lang=pt%0Ahttp://www.scielo.org.co/pdf/rfnsp/v28n3/v28n3a02.pdf

21. Arcia LA. *Demografia y salud. Apuntes para una conferencia*. Rev Habanera Ciencias Medicas. 2009;8(4).

22. Meisel-Roca A. *La no reversión de la fortuna en el largo plazo : geografía y persistencia espacial de la prosperidad en Colombia*, *1500-2005*. Cuad Hist Económica y Empres ; No 35 [Internet]. 2014; Available from: http://repositorio.banrep.gov.co/handle/20.500.12134/1980

23. Elizaga JC. *Métodos demográficos para el estudio de la mortalidad* [Internet]. Segunda ed. Santiago de Chile: Centro Latinoamericano de Demografía; 1972. 204 p. Available from: https://repositorio.cepal.org/handle/11362/7414

24. Dávila CA, Pardo AM, Pardo AM. *Mortalidad por suicidios en Colombia y México: tendencias e impacto entre 2000 y 2013. Biomédica* [Internet]. 2016;36(3):415–22. Available from: http://www.revistabiomedica.org/index.php/biomedica/article/view/3224

6. Anexos

Anexo 1. Tabla de distribución de la normal estándar.

Anexo 2. Tabla de distribución *t de Student*.

Anexo 3. Tabla de distribución *Ji-cuadrado*.