

Libros de **Cátedra**

Análisis estadístico en Educación Física

Medidas, usos e interpretaciones

Stella Maris Doná - Román Fornessi
(coordinadores)

FACULTAD DE
HUMANIDADES Y CIENCIAS DE LA EDUCACIÓN

S
sociales


EDITORIAL DE LA UNLP



UNIVERSIDAD
NACIONAL
DE LA PLATA

ANÁLISIS ESTADÍSTICO EN EDUCACIÓN FÍSICA

MEDIDAS, USOS E INTERPRETACIONES

Stella Maris Doná

Román Fornessi

(coordinadores)

Facultad de Humanidades y Ciencias de la Educación



A las y los estudiantes de las carreras de Educación Física, de la Facultad de Humanidades y Ciencias de la Educación de la Universidad Nacional de La Plata.

Índice

Agradecimientos	6
------------------------	----------

Introducción

La importancia de la estadística en la Educación Física	7
<i>Stella Maris Doná, Mariela Cotignola, Martina Díaz y María Agustina Gandini</i>	

Capítulo 1

Estadística e investigación en Educación Física	12
<i>Stella Maris Doná, Mariela Cotignola, Martina Díaz y María Agustina Gandini</i>	

Capítulo 2

Sobre las unidades de análisis y sus características	25
<i>Lucía Girardi y Rocío Belén Villafañe</i>	

Capítulo 3

Organización, presentación y comunicación de los datos	39
<i>Martina Díaz y Román Fornessi</i>	

Capítulo 4

Medidas estadísticas descriptivas de posición, centrales y no centrales	58
<i>Lucía Girardi y Rocío Belén Villafañe</i>	

Capítulo 5

Medidas de variabilidad y de forma	76
<i>Román Fornessi y María Agustina Gandini</i>	

Capítulo 6

Probabilidad y distribuciones probabilísticas. La distribución normal	97
<i>Stella Maris Doná y Martina Díaz</i>	

Capítulo 7

Asociación entre variables cualitativas	127
<i>Martina Díaz y Mariela Cotignola</i>	

Capítulo 8

Correlación bivariada y regresión entre variables numéricas	143
<i>Stella Maris Doná, Mariela Cotignola y Gisela Marcos Castillo</i>	

Capítulo 9

Muestreo y estimación paramétrica. Nociones de muestreo, estimación de parámetros
y prueba de hipótesis _____ 172

Stella Maris Doná y Mariela Cotignola

Quiénes escribimos este libro _____ 188

Agradecimientos

A la Universidad Nacional de La Plata y a la Facultad de Humanidades y Ciencias de la Educación.

Educación pública, gratuita y de calidad, siempre.

INTRODUCCIÓN

La importancia de la estadística en la Educación Física

Stella Maris Doná, Mariela Cotignola, Martina Díaz y María Agustina Gandini

¿Por qué estudiar estadística?

Este libro surge de la necesidad que advertimos en la cátedra de Estadística Aplicada a la Educación Física de elaborar material bibliográfico que vincule estas dos disciplinas, que revele, de alguna manera, la importancia, la utilidad y la pertinencia de estudiar estadística en la formación de profesionales de la Educación Física. También, surge como producto colectivo de un equipo de trabajo interdisciplinario, que convoca a personas con formación en Educación Física, Psicología, y Sociología a pensar y elaborar conjuntamente una herramienta didáctica para el aula y el estudio.

La asignatura Estadística Aplicada a la Educación Física está ubicada en el segundo año del Plan de Estudios del año 2000 de las carreras de Licenciatura y Profesorado en Educación Física¹. Es de carácter obligatorio y corresponde al Trayecto de la Formación en Investigación, junto con la asignatura Metodología de Investigación en Educación Física, de cuarto año. Tiene como materias correlativas a Educación Física I y Teoría de la Educación Física I, al tiempo que es correlativa de Metodología de Investigación en Educación Física y articula con contenidos de otras asignaturas.

El desarrollo de los contenidos de la asignatura no está pensado de forma aislada, sino en consonancia con las materias del mismo año, como así también en relación a materias de años posteriores. Asignaturas como: Capacitación en informática, Fisiología Aplicada a la Educación Física, Metodología de la investigación en Educación Física, Observaciones y Prácticas de la enseñanza en Educación Física, entre otras, utilizan como saber previo, los conocimientos que ofrece la materia.

En el trayecto de formación del tronco común y el trayecto de formación pedagógica, la estadística se convierte en una herramienta útil en el marco de las prácticas sociales y

¹Profesorado y Licenciatura en Educación Física – Plan de estudios 2000, U.N.L.P. Facultad de Humanidades y Ciencias de la Educación, disponible en <https://www.memoria.fahce.unlp.edu.ar/planes/pl.2/pl.2.pdf>

educativas, como el juego, la gimnasia, los deportes, la vida en la naturaleza y al aire libre, y la natación. Su uso facilita el análisis de la relación entre estas prácticas y los procesos de enseñanza y aprendizaje, considerando la diversidad de marcos institucionales y político-contextuales en los que se desarrollan, ya sean escuelas, clubes, colonias de vacaciones, etc.

En el trayecto de formación biológica, su aplicación resulta clave para el análisis biomecánico de los movimientos corporales, el estudio de variables fisiológicas en reposo y en ejercicio, entre otros. Estos estudios permiten optimizar el rendimiento, prevenir lesiones y diseñar programas de entrenamiento más efectivos; todo se realiza e interpreta con técnicas estadísticas.

En lo que respecta al trayecto de formación general la estadística puede ser útil en el análisis de procesos sociales, educativos y culturales, muchos de los cuales pueden medirse y estudiarse con datos cualitativos como cuantitativos.

Si bien, históricamente, la estadística se la ha vinculado únicamente a las áreas de investigación, reduciéndose sólo a ello, el saber que aportan los contenidos de la materia son mucho más amplios e indispensables para el buen desempeño de las acciones más elementales de nuestra práctica profesional cotidiana.

Al realizar el registro de nuestras clases y balances, las comparaciones entre escuelas, entre cursos, el promedio de notas, el seguimiento particular de cada estudiante y la comparación con su grupo, las evaluaciones periódicas, el seguimiento de los jugadores en un club, de aquello que están bajo nuestra instrucción, en un gimnasio, plaza o colonia; para leer y comprender documentos, estudios, datos publicados.

Ya sea como docentes o instructores en el ámbito deportivo, o recreativo, o de salud, como profesionales en instituciones de gestión pública o privada, o cualquier otro espacio del ejercicio profesional, todos ellos requieren de este tramo de nuestra formación profesional: el del aprendizaje de las técnicas estadísticas básicas.

En el campo de la Educación Física nos encontramos con una gran variedad de estudios de investigación que nos aportan conocimientos y robustecen nuestro ejercicio profesional. La diversidad de temas de nuestra profesión es muy rica y aún hay mucho por descubrir. Aunque la mayoría de que los/as ingresantes eligen en su mayoría la carrera de profesorado (es un dato obtenido con técnicas estadísticas)² debe señalarse que muchos de ellos optan también por la licenciatura, rama cuyo aporte a la investigación en educación física, es sumamente importante y urgente.

La inclusión de la materia en el actual Plan de Estudios refiere a la formación necesaria en ese campo, en el que aparecen conjugados aspectos formativos e instrumentales. Es formativa en tanto desarrolla una forma especial de pensamiento y es instrumental ya que proporciona herramientas de trabajo al futuro profesional de la Educación Física: de alguna u otra, todo Profesor o Licenciado en Educación Física necesita del uso de las técnicas estadísticas.

Debe señalarse también que la estadística es una metodología utilizada por distintas profesiones; luego, se trata de un saber que permite el vínculo interdisciplinario, el trabajo y la

² Información proporcionada por el Departamento de Educación Física de la FaHCE. UNLP.

interacción con otras miradas y perspectivas, permitiendo imprimir la mirada de la Educación Física y la riqueza de su comprensión de la realidad, a las otras áreas de trabajo (Vilchez Guizado y Ramón Ortiz, 2018).

Los/as estudiantes de la carrera reconocen que las estadísticas están más cerca de nuestra profesión de lo que creemos, por lo tanto, nos referimos a un tipo de conocimiento que se considera fundamental para el perfil de quien egresa. Su uso permite comprender dinámicas sociales, adaptar estrategias pedagógicas para optimizar los procesos de enseñanza y aprendizaje, evaluar el impacto de programas educativos, recreativos y deportivos, en distintos contextos (Fornessi, 2022).

En tal sentido, hay un aspecto realmente clave que queremos remarcar: el carácter político, social, cultural e histórico que atraviesa la enseñanza de la estadística para nosotras/os, al tiempo que complejiza su relación con la Educación Física interpelando a las/os estudiantes en su cotidianidad y sus modos de estar en el mundo sobre el cual les toca trabajar.

La aplicación de herramientas estadísticas resulta de utilidad tanto en los procesos de investigación científica como siendo soporte técnico de la propia práctica docente y profesional. Casi todas las investigaciones aplicadas requieren algún tipo de análisis estadístico para que sea posible evaluar sus resultados. En algunos casos, para resolver un problema de carácter empírico, es preciso llevar a cabo análisis de características complejas; otras veces, basta con efectuar un análisis muy simple y directo.

El conocimiento estadístico alienta el espíritu crítico frente a los datos publicados y problematiza sobre su construcción e interpretación. Permite, en cualquier instancia del proceso educativo o de la práctica profesional, darle un sentido a los datos disponibles al poder ordenarlos, sistematizarlos, resumirlos, analizarlos, compartirlos y repensarlos.

Tal como señala Batanero (2002) al hablar de cultura estadística, ésta no sólo sirve como herramienta clave en la formación profesional del Profesor/a y Licenciado/a en Educación Física, sino que busca desarrollar la capacidad para, reflexionar y comunicar una posición crítica respecto de los resultados y conclusiones a los que se arriben basados en el uso de la estadística.

Somos sujetos sociales, atravesados por nuestra formación. El estudio de la estadística debe despertar en nosotros la necesidad ética de ser rigurosos en aquello que afirmamos, críticos de los conocimientos disponibles y creativos en los aportes al campo propio de la educación física.

Si bien está basada en la matemática, la estadística es teoría y aplicación; si bien trabaja con razonamientos complejos, esta materia nos hace incorporar un tipo de razonamiento de índole lógica deductiva/inductiva. Cabe destacar, que el propósito es que los/as estudiantes se apropien de nociones que les permita pensar que la estadística más allá de cálculos y operaciones algebraicas, sino que contribuye a la formación de profesionales críticos, éticos, cuyo modo de entender la realidad que nos circunda cuenta con útiles herramientas que aportan fundamento a diversas áreas dentro de la disciplina que nos ocupa. Por ello, nuestra perspectiva se propone superar la mirada de un modelo biologicista o uno basado

exclusivamente en hechos y resultados; buscamos enfatizar que la enseñanza-aprendizaje de análisis e investigación con métodos estadísticos se encuentra dentro del paradigma de las ciencias sociales.

Por todo lo dicho, el desarrollo de los contenidos de este libro de cátedra busca proporcionar a las/os estudiantes conocimientos acerca de los procedimientos que permiten resumir e interpretar información con técnicas estadísticas descriptivas e introducirlas/os en el uso de técnicas estadísticas inferenciales que permitan obtener conclusiones a partir de bases rigurosas y que ayuden a tomar decisiones bajo incertidumbre.

Ello permitirá desarrollar las competencias básicas para la lectura comprensiva y el análisis crítico de trabajos de investigación que utilicen técnicas estadísticas, vincular y aplicar los saberes a la práctica en Educación Física, aplicar los principios éticos del ejercicio profesional a la construcción e interpretación de los datos, posicionándose como profesionales ancladas/os históricamente, en el contexto socio-político y cultural actual.

Referencias

- Batanero, C. (2002). “Los retos de la cultura estadística”, Conferencia Inaugural de las Jornadas Interamericanas de Enseñanza de la Estadística. Buenos Aires, Argentina. Disponible en: <https://www.ugr.es/~batanero/pages/ARTICULOS/CULTURA.pdf>
- Fornessi, R. (2022). *Caracterización y análisis de la experiencia de estudiantes de Educación Física en Estadística Aplicada a la Educación Física (2017-2020)*. Revista Educación Física y Ciencia, 24 (4), e237, <https://doi.org/10.24215/23142561e237>. Disponible en: http://www.scielo.org.ar/scielo.php?script=sci_arttext&pid=S2314-25612022000400237&lng=es&nrm=iso
- Vilchez Guizado, J.; Ramón Ortiz, J. (2018). Elementos de estadística y probabilidades. Editorial Académica Española.

Bibliografía

- Bologna, E. y Morales, M. (2011) “Presentación: ¿Estadística en psicología y educación?”, en Estadística para psicología y educación. Editorial Brujas.
- Castro Nogueira, L.; Castro Nogueira M.; Morales Navarro, J. (2005). “Concepto de medición”, en *Metodología de las Ciencias sociales, una introducción crítica*. Editorial Tecnos.
- Johnson, R. y Kuby P. (1999) Estadística elemental. Cap. 3 México: ITP. Disponible en: [file:///C:/Users/Usuario/Downloads/Estadística%20Elemental%20-%20Johnson%20Kuby%2011ed%20\(1\)-1.pdf](file:///C:/Users/Usuario/Downloads/Estadística%20Elemental%20-%20Johnson%20Kuby%2011ed%20(1)-1.pdf)

Sosa Escudero, W. (2020). Mentiras verdaderas. A modo de introducción, en *¿Qué es (y qué no es) la estadística? Usos y abusos de una disciplina clave en la vida de los países y las personas*. Editorial Siglo XXI.

CAPÍTULO 1

Estadística e investigación en Educación Física

Stella Maris Doná, Mariela Cotignola, Martina Díaz y María Agustina Gandini

Educación Física y estadística

Hablar de estadística es hablar de investigación, de indagación, de la búsqueda de un conocimiento. Es una metodología posible de ser aplicada para analizar e interpretar información que se ha recogido a fin de abordar algún tema de nuestro interés.

Por ejemplo, en el campo de la Educación Física, es habitual que un/a Profesor/a tenga un grupo de personas a su cargo y realice comparaciones entre los rendimientos iniciales, medios y finales. Para ello debe utilizar técnicas estadísticas: recoge los datos, los organiza, analiza y presenta los resultados observados que caracterizan al grupo. De igual manera, si se desea demostrar que existe una relación entre dos sucesos como realizar ejercicios de resistencia y una mejora en la capacidad aeróbica, se aplican técnicas estadísticas para probar hipótesis; o si se desea conocer cuál será la talla promedio, ya no de un grupo de los adultos argentinos, sino de todos los habitantes adultos de la Argentina, se realizan con técnicas estadísticas denominadas de estimación de parámetros.

Como vemos, el principal aporte que hace la estadística como disciplina es proporcionar métodos de análisis que brindan una medida de precisión a las conclusiones a las que se arriban.

Vilchez Guizado y Ramón Ortiz (2018), define a la **estadística** de la siguiente manera:

Es la disciplina que tiene por objeto recolectar, organizar, procesar, presentar, analizar e interpretar datos, ya sea para ayudar en la toma de decisiones o para explicar condiciones regulares o irregulares de algún fenómeno o estudio aplicado, de ocurrencia en forma aleatoria o condicional, a fin de obtener leyes que describan su comportamiento y expliquen con pertinencia las observaciones. De tal forma que los errores que se puedan cometer al sacar conclusiones y hacer estimaciones sobre la base de los datos observados, a partir de un razonamiento inductivo fundamentada en la teoría de probabilidad (pp. 8-9).

La **estadística** es un conjunto de herramientas para organizar, resumir, interpretar y comunicar la información de manera efectiva. Es una disciplina que se ocupa del análisis y la interpretación de datos para extraer conclusiones, hacer inferencias y tomar decisiones informadas. En otras palabras, la estadística no sólo es un conjunto de técnicas matemáticas, sino también un enfoque fundamental para entender fenómenos y tomar decisiones en situaciones de incertidumbre, decisiones que impactan sobre las personas (Sosa Escudero, 2020).

En palabras de Moreira Cid (2017, p. 25), “la **estadística** es una rama de las matemáticas que reúne y clasifica los datos numéricos para poder, a partir de ellos, generar conclusiones e inferencias”.

La **estadística** es un conjunto de métodos que permiten recolectar datos, resumirlos con precisión y derivar de ellos, alcanzar conclusiones significativas y poder tomar decisiones bajo condiciones de incertidumbre (Johnson y Kubly, 1999).

Más allá de que hay autores que ubican a la estadística como una parte de las matemáticas (Maureira Cid, 2017 y otros, la entienden como una disciplina transversal a todas las actividades de investigación que se realicen (Vilchez Guizado y Ramón Ortiz, 2018), lo importante es entender que el desarrollo de la disciplina supone una teoría y una aplicación de esa teoría a la resolución de problemas concretos que favorecen el avance del conocimiento. La **estadística** brinda procedimientos para poder poner en correspondencia los postulados de la teoría con los hechos observados, aportando rigor y un fundamento, al ejercicio de la práctica profesional (Bologna, 2011).

Algo fundamental a tener en cuenta es que las técnicas estadísticas se aplican a **conjuntos** de objetos a un gran número de sujetos. Su objetivo fundamental es proporcionar una síntesis de las características más sobresalientes de un grupo de unidades sobre las que recae el estudio. Asimismo, esos aspectos que definen al grupo se deben al azar, es decir, no se pueden predecir, se desconoce el resultado que se busca conocer y, justamente por ese motivo, se utiliza la estadística. Por ejemplo, si queremos conocer cuál es el área laboral de mayor preferencia de los estudiantes de la Carrera, podríamos preguntar a un grupo de alumnas/os; las respuestas seguramente serán variadas (podrían ser: docencia, deportes, rehabilitación, etc.), pero se puede saber de antemano, por eso aplicamos las técnicas estadísticas.

A continuación, desarrollaremos conceptos básicos con los que trabaja la estadística: población, muestra, medición. De igual manera, haremos a las dos grandes áreas de aplicación, sus dos funciones: descriptiva e inferencial. Los capítulos de este Libro de Cátedra continuarán con el desarrollo de otros conceptos fundamentales de la estadística y las técnicas más relevantes para el campo de la Educación Física.

Primeras nociones estadísticas

Población

A excepción del censo, las técnicas estadísticas se aplican a muestras obtenidas de una población.

Llamaremos **población** al conjunto total de sujetos o de unidades de análisis sobre los que recae nuestro estudio (Maureira Cid, 2017). Es la colección de todos los individuos, objetos o eventos sobre los que se analizan las propiedades que la describen (Johnson y Kuby, 1999).

Una población es el conjunto total de los elementos sobre los que se intenta sacar conclusiones; dichos elementos forman parte de ese conjunto población en la medida que presenten una o varias características comunes, definidas en tiempo y espacio, en un momento dado y un lugar determinado; poseer o no esas cualidades hace que, taxativamente, una unidad pertenezca o no a esa población (Vilchez Guizado y Ramón Ortiz, 2018).

Por ejemplo, el INTI³ publicó en el año 2022, un estudio antropométrico de las medidas corporales de los habitantes de nuestro país. El trabajo describe que los elementos que conformaron la población de estudio fueron: ser argentino, mayor de 12 años, sin exclusión de género. Según el censo nacional realizado en 2022⁴, del total de 45.892.285 habitantes, sólo pertenecen a la población del estudio antropométrico, un número de 33.377.173 elementos, que son los que poseen las características definidas para ser parte de la población en estudio.

De igual manera, un estudio sobre los estudiantes de las carreras de Educación Física debe definir sus características como, por ejemplo, ser alumnos de la UNLP, inscriptos en la Licenciatura o el Profesorado, ser activo, en el curso del ciclo lectivo 2025.

Un estudio sobre el grupo de personas con las que trabajamos como instructor/ra, estará conformada por el número total de sujetos que se encuentran inscriptos, en un determinado club o espacio, en un determinado año, mes, categoría, etc.

Es muy importante definir la población en estudio. Deben especificarse de manera exhaustiva todas las condiciones que deben tener sus elementos, como así también, las condiciones que excluyen a un elemento de la población. Nótese que, en estadística, el concepto de población no es sinónimo de habitantes. Las unidades que la conforman pueden ser tanto personas como objetos o eventos. Por ejemplo, un estudio sobre instituciones deportivas tiene como unidad de análisis cada una de las instituciones deportivas cuyas características se especifiquen, tanto de inclusión como de exclusión para formar parte de la población; una investigación sobre la concepción del cuerpo podría tener como elementos integrantes de la población en estudio, cada uno de los autores que se refieren al término, durante un período específico de la historia.

³ INTI: Instituto Nacional de Tecnología Industrial, de la República Argentina. Estudio disponible en https://www.argentina.gob.ar/sites/default/files/2023/10/estudio_antropometrico_ar_-_informe_tecnico.pdf

⁴ Resultados del censo nacional, disponible en: <https://censo.gob.ar/>

Las poblaciones pueden ser finitas o infinitas. Nos referiremos a una población finita cuando se conoce el número de elementos que la integran, es decir, podemos contar con un listado detallado de absolutamente todas las unidades de nuestro análisis (Johnson. y Kuby, 1999).

La clasificación de infinita no se refiere a tan sólo a que son incontables los elementos que la integran. Se relaciona con un número muy grande de elementos que, aunque se pueda tener acceso a cada uno de los elementos de una población, superan las 100.000 unidades de análisis. En el mismo sentido, serán consideradas infinitas aquellas poblaciones hipotéticas, por ejemplo, un estudio sobre *los adolescentes*. En rigor, una población es considerada infinita cuando la muestra con la que se trabaja representa menos del 5% de los casos de dicha población. Es la relación entre la cantidad de casos de la población y de la muestra la que hace considerar a una población como infinita (Maureira Cid, 2017).

A la cantidad de casos que integran una población, la designaremos con la letra **N** (ene mayúscula). A todas las medidas e índices estadísticos que describen a una población se denominan **parámetros**.

Muestra

Una vez definida la población en estudio, la cantidad de elementos que la integran es, en general, un número grande. Resulta muy costoso en tiempo y recursos trabajar con todas las unidades de análisis para extraer conclusiones válidas que la describan. Además, analizar un gran volumen de datos puede traer aparejados errores involuntarios en su recolección y organización.

Estos problemas se resuelven extrayendo lo que se denomina una **muestra**, un subconjunto de elementos de la población (Vilchez Guizado y Ramón Ortiz, 2018).

Al definir una población, se han especificado de manera exhaustiva las características que deben cumplir los elementos para pertenecer a ella. Las unidades de análisis pasarán a ser unidades de observación e integrarán la muestra de elementos con los que trabajaremos sólo si son obtenidas de dicha población y, por lo tanto, poseen esas mismas características; una muestra es una porción de la población. Siempre que trabajamos con muestras, debe explicitarse cuál es la población de referencia a la cual será válida extender los resultados que se obtienen del estudio muestral (Bologna, 2011).

Básicamente, los elementos que integran una muestra pueden ser escogidos de dos maneras. Decimos que una **muestra es probabilística o aleatorio** cuando se puede garantizar la representatividad de cada uno de los elementos de la población que la integrarán y ello se consigue cuando todos ellos tienen la misma probabilidad de ser seleccionado para integrar la muestra; se llama principio de equiprobabilidad (Vilchez Guizado y Ramón Ortiz, 2018).

Una **muestra es no probabilística o no aleatoria** cuando las unidades de observación no han sido seleccionadas respetando aquel principio de representatividad (aunque lo llegaron a

ser por casualidad) sino que la facilidad que se tenga a su acceso es lo que se privilegia al momento de su elección para que conforme la muestra (Maureira Cid, 2017).

Hay que aclarar que no hay una única muestra posible de ser extraída de una población; habrá tantas muestras posibles de obtener como se desee, sean probabilísticas o no. Las técnicas estadísticas específicas para la selección de dichos elementos, se denomina muestreo.

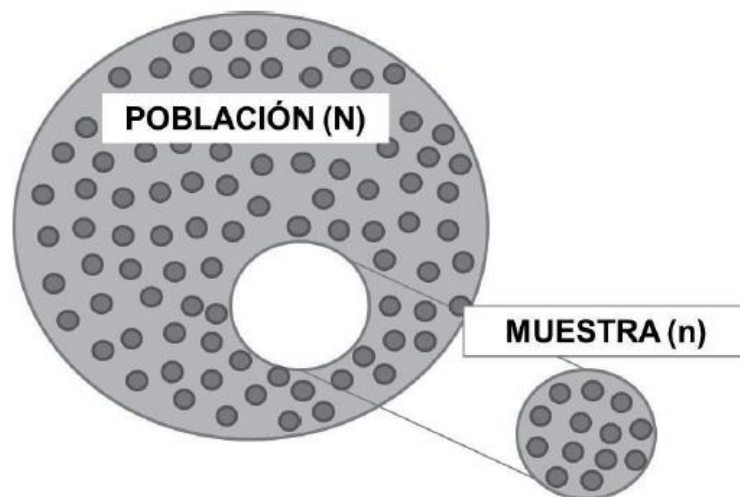
A la cantidad de casos que integran una muestra, la designaremos con la letra n (ene minúscula). A todas las medidas e índices estadísticos que describen a una población se denominan estadísticos.

Siguiendo el ejemplo del estudio realizado por el INTI, habiéndose definido la población, se extrajo una muestra probabilística conformada por 13.276 personas.

La siguiente figura muestra la relación entre población y muestra.

Figura 1.1

Población y muestra



Nota. Fuente: La figura se obtuvo de Maureira Cid (2017, p. 82).

Funciones de la estadística: descripción e inferencia

La estadística describe las características de una población o de una muestra; al tiempo que también proporciona técnicas que permiten inferir los resultados obtenidos sobre una muestra y la población, o probar hipótesis sobre el comportamiento de ciertos fenómenos.

La **estadística descriptiva** provee una serie de procedimientos destinados a resumir, a sintetizar información, a volverla inteligible para que podamos interpretarla y extraer conclusiones del análisis de ese conjunto de datos (Bologna, 2011).

Esta parte de la estadística se ocupa de describir, analizar y representar un conjunto de datos utilizando indicadores numéricos, tablas y gráficos que resumen y presentan la información contenida en ellos (Vilchez Guizado y Ramón Ortíz, 2018).

Las medidas fundamentales de la estadística descriptiva son las denominadas de Tendencia Central, de Variabilidad, y de Forma (asimetría y curtosis). No son las únicas: también lo son las tasas, razones, las medidas de orden, coeficientes de asociación, independencia y correlación entre variables, entre otros. Durante el transcurso de los Capítulos de este Libro, las iremos desarrollando.

En nuestro ejemplo del estudio antropométrico del INTI, se realizaron tablas, gráficos y se tomaron diversas medidas estadísticas que permitieron la descripción de la muestra. Por ejemplo, se ha observado que en lo respecta al género, hay un 36,5% de masculinidades y 63,5% de feminidades; el 70% de la muestra corresponde a personas dentro del rango etario de 20-49 años; la talla promedio en cuerpos de mujer es de 161,1 cm y 173,8 cm, en cuerpos de varones.

La **estadística inferencial** proporciona técnicas para dos grandes operaciones: inferir parámetros y probar hipótesis de investigación. La estimación de parámetros se realiza aplicando dos posibles métodos, una estimación puntual o una estimación por intervalos, del valor descriptivo encontrado en la muestra, a la población. La prueba o contraste de hipótesis trata de establecer si resulta verosímil un valor de un parámetro de la población; se evalúa con la información obtenida en la muestra y la teoría de la probabilidad (Maureira Cid, 2017).

Los estadísticos, los valores calculados sobre los datos muestrales, son los que se usarán para realizar aproximaciones a los valores poblacionales denominados parámetros. A los estadísticos que estiman parámetros se los denomina estimadores puntuales, o simplemente **estimadores** (Bologna, 2011)

En el ejemplo del INTI, dado que la muestra seleccionada es probabilística, es decir, se ha garantizado un número de elementos suficiente y la representatividad de la población como para poder afirmar la cintura promedio de un cuerpo de hombre argentino mayor de 18 años es de 93,4 cm (estimación puntual) y el de cuerpos de mujer se encuentra dentro del rango entre 74-86 cm (estimación por intervalos).

Generalmente se cuenta sólo con un pequeño conjunto de datos, los que integran la muestra seleccionada, pero -a partir de ellos- es posible inferir las características de conjuntos más grandes de datos, aun cuando sólo evaluemos a algunos de ellos. (Maureira Cid, 2017).

Una de las pretensiones del estudio sobre las muestras, es poder hacer proyecciones de la descripción observada en la muestra, sobre la población de la cual se extrajo. Para realizar el pasaje de valores muestrales a poblacionales, la estadística se basa en el cálculo de probabilidades. La teoría de las probabilidades es el puente que permite pasar de la estadística descriptiva a inferencial (Vilchez Guizado y Ramón Ortíz, 2018).

La estadística inferencial se basa en hallazgos encontrados en las muestras, a partir de la aplicación de técnicas descriptivas de la estadística descriptiva y de la teoría de las

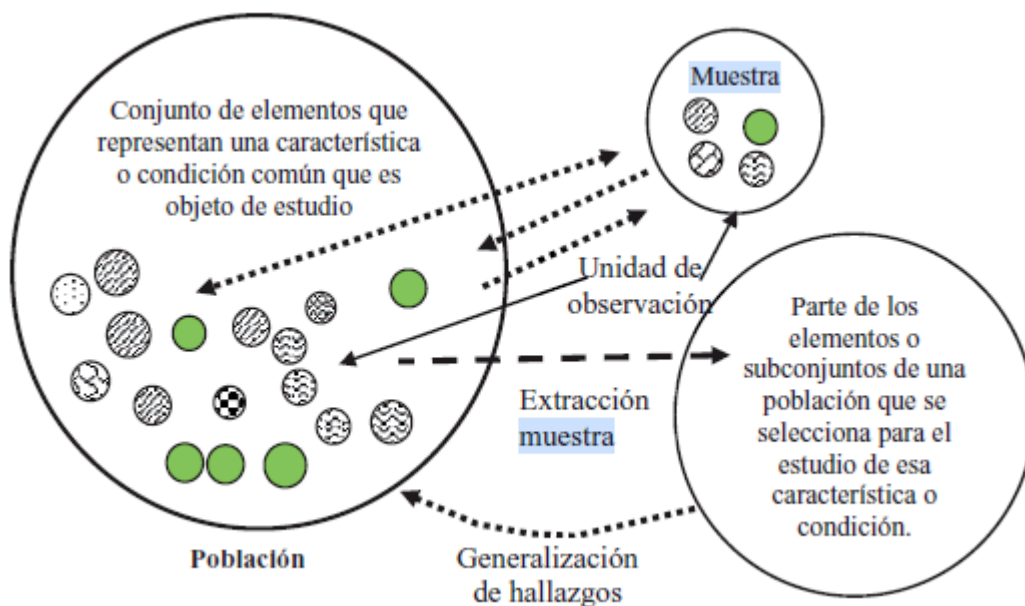
probabilidades. Estos dos elementos se conjugan a fin de inferir a la población, de cual es parte la muestra, conclusiones válidas (Bologna, 2011).

Es importante entender que, para poder aplicar las técnicas inferenciales, las muestras deben cumplir ciertas condiciones: haber sido lesionada de manera probabilística y tener un número suficiente de casos, al tiempo que conocer cómo se describe la población de la cual que fue extraída.

La siguiente figura muestra la relación entre una población y una muestra que permite la operación inferencial de estadísticos a parámetros.

Figura 1.2

Relación entre población y muestra



Nota. La figura se obtuvo de Vilchez Guizado y Ramón Ortiz (2018, p. 26)

Medición

Medir es una acción empleada habitualmente en nuestro medio. Medimos la distancia entre un punto y otro de un recorrido, el tiempo que se emplea en realizarlo, la cantidad de vueltas que se completan al nadar durante un cierto tiempo; se miden los conocimientos con una nota asignada a una evaluación, etc.

La estadística mide, toma medidas que describen las características principales de las unidades de análisis que conforman una muestra; mide si lo observado en una muestra dista mucho de lo que debería observarse en su población, etc.

Luego, resulta metodológicamente crucial definir qué entendemos por medición.

Medir “consiste en comparar una magnitud con otra de su misma especie que se toma como unidad, o con otra magnitud adecuada al caso para conocer su extensión o cantidad” (Castro Nogueira, L.; Castro Nogueira M. y Morales Navarro, 2005, p. 129).

La operación de medir parece bastante simple para ciencias como la química, la física, pero muy compleja para las ciencias sociales pues resulta difícil establecer homogeneidad entre las magnitudes que se desean comparar. Por ejemplo, puede establecerse un procedimiento y un instrumento válido para medir un tiempo. Pero, cómo medir el grado de satisfacción con el actual Plan de Estudio.

La mediación busca establecer una relación que intente emparejar el mundo de los números de la matemática y el empírico de las observaciones. Si puede establecerse esa relación, podría asignarse un número a cierto objeto empírico.

La clave radica en descubrir cuáles son las reglas que rigen el mundo de los números para poder verificar su hallazgo, o no, en las propiedades de la realidad.

Para poder medir, es necesario tener en claro qué aspecto de la realidad se va a medir y definirla teóricamente. Una vez establecido esto, se debe disponer de algún instrumento y escala para hacerlo.

Por ejemplo, si se desea describir el peso de un grupo de estudiantes, debe definirse la característica peso; no es lo mismo definirlo en términos de cantidad de kg y tomar una balanza para hacerlo, o definirla en términos de nivel de peso: bajo, normal, sobrepeso, obesidad, asociado al concepto de talla e índice de masa corporal⁵.

La **definición teórica** de la característica a estudiar y el instrumento para observarla posibilitan la medición al mismo tiempo que también la condicionan pues dicen cómo se está concibiendo ese suceso, dependiendo de su **definición operacional** (Castro Nogueira, L.; Castro Nogueira M. y Morales Navarro, 2005).

Lo que se mide son propiedades de los objetos, son constructos observables de un hecho, es la definición operacional de aquella definición teórica. Si el peso es la masa corporal, tomo una balanza y peso a cada una de las personas que integran la muestra. Si el peso es entendido como un nivel, debo tomar un metro y medir en cm la talla de cada persona para luego calcular el índice de masa corporal y definir el nivel de peso de cada sujeto que conforma la muestra. Una misma característica del grupo en estudio, se puede definir de maneras diversas y, por lo tanto, la comprensión de la realidad y todas las conclusiones que realicemos sobre los datos disponibles, es también diferente. Ni buena, ni mala, ni peor ni mejor, será distinta.

Fue Stevens quien en 1946 publicó un artículo científico que sentó las bases del concepto de **medición** abriéndolo desde las ciencias como la física y la química, hacia las ciencias sociales. Afirma que medir, “en sentido amplio, se define como la asignación de numerales a objetos o eventos de acuerdo a ciertas reglas” (p. 677).

⁵ Ver clasificación del peso que realiza la OMS, en <https://www-who-int.translate.goog/news-room/factsheets/detail/obesity-and-overweight? x tr sl=en& x tr tl=es& x tr hl=es& x tr pto=tc>

Medir es una operación que busca el **isomorfismo** entre las propiedades de los números y la presencia de esas propiedades matemáticas en la característica que se observa en una unidad de análisis. Los números se caracterizan por tener un orden, distancias iguales entre las unidades de medida de una escala y un punto de origen (arbitrario o real).

Las diferentes reglas que se apliquen para asignar un número de manera que represente a una realidad (ejemplo: un número del 1 al 10, asignado como nota de un examen rendido) es la clave para entender cómo se mide. Steves señala que “hay diferentes tipos de medición” (1946, p. 677).

No es posible hablar de un único proceso matemático-estadístico de medición. Como se estableció, de acuerdo con las propiedades de los números que se puedan observar en las propiedades de los eventos a estudiar, será el nivel de la medición. Hay hechos de la realidad que sólo pueden clasificarse, como la característica de tener o no un trabajo; y otras que, como el caso del peso corporal, poseen todas las propiedades de los números.

En tal sentido, Stevens (1946), establece cuatro niveles de medición en orden creciente de isomorfismo entre los números y la realidad. Los niveles de medición se denominan nominal, ordinal, intervalar y racional, tema que desarrollaremos en el Capítulo que sigue a continuación.

La importancia de la estadística en la Educación Física

¿Por qué estudiar estadística?

La asignatura Estadística Aplicada a la Educación Física está ubicada en el segundo año del Plan de Estudios 2000, de las carreras de Licenciatura y Profesorado en Educación Física⁶. Es de carácter obligatorio y corresponde al Trayecto de la Formación en Investigación, junto con la asignatura Metodología de Investigación en Educación Física, de 4to año. Tiene como materias correlativas a Educación Física I y Teoría de la Educación Física I, al tiempo que es correlativa de Metodología de Investigación en Educación Física y articula con contenidos de otras asignaturas.

El desarrollo de los contenidos de la asignatura no está pensado de forma aislada, sino en consonancia con las materias del mismo año, como así también en relación con materias de años posteriores. Asignaturas como: Capacitación en informática, Fisiología Aplicada a la Educación Física, Metodología de la investigación en Educación Física, Observaciones y Prácticas de la enseñanza en Educación Física, entre otras, utilizan como saber previo, los conocimientos que ofrece la materia.

⁶ Profesorado y Licenciatura en Educación Física – Plan de estudios 2000, U.N.L.P. Facultad de Humanidades y Ciencias de la Educación, disponible en <https://www.memoria.fahce.unlp.edu.ar/planes/pl.2/pl.2.pdf>

En el trayecto de formación del tronco común y el trayecto de formación pedagógica, la estadística se convierte en una herramienta útil en el marco de las prácticas sociales y educativas, como el juego, la gimnasia, los deportes, la vida en la naturaleza y al aire libre, y la natación. Su uso facilita el análisis de la relación entre estas prácticas y los procesos de enseñanza y aprendizaje, considerando la diversidad de marcos institucionales y político-contextuales en los que se desarrollan, ya sean escuelas, clubes, colonias de vacaciones, etc.

En el trayecto de formación biológica, su aplicación resulta clave para el análisis biomecánico de los movimientos corporales, el estudio de variables fisiológicas en reposo y en ejercicio, entre otros. Estos estudios permiten optimizar el rendimiento, prevenir lesiones y diseñar programas de entrenamiento más efectivos; todo se realiza e interpreta con técnicas estadísticas.

En lo que respecta al trayecto de formación general la estadística puede ser útil en el análisis de procesos sociales, educativos y culturales, muchos de los cuales pueden medirse y estudiarse con datos cualitativos como cuantitativos.

Si bien, históricamente, la estadística se la ha vinculado únicamente a las áreas de investigación, reduciéndose sólo a ello, el saber que aportan los contenidos de la materia son mucho más amplios e indispensables para el buen desempeño de las acciones más elementales de nuestra práctica profesional cotidiana.

Al realizar el registro de nuestras clases y balances, las comparaciones entre escuelas, entre cursos, el promedio de notas, el seguimiento particular de cada estudiante y la comparación con su grupo, las evaluaciones periódicas, el seguimiento de los jugadores en un club, de aquello que están bajo nuestra instrucción, en un gimnasio, plaza o colonia; para leer y comprender documentos, estudios, datos publicados.

Ya sea como docentes o instructores en el ámbito deportivo, o recreativo, o de salud, como profesionales en instituciones de gestión pública o privada, o cualquier otro espacio del ejercicio profesional, todos ellos requieren de este tramo de nuestra formación profesional: el del aprendizaje de las técnicas estadísticas básicas.

En el campo de la Educación Física nos encontramos con una gran variedad de estudios de investigación que nos aportan conocimientos y robustecen nuestro ejercicio profesional. La diversidad de temas de nuestra profesión es muy rica y aún hay mucho por descubrir. Aunque la mayoría de que los/as ingresantes eligen en su mayoría la carrera de profesorado (es un dato obtenido con técnicas estadísticas)⁷ debe señalarse que muchos de ellos optan también por la licenciatura, rama cuyo aporte a la investigación en educación física, es sumamente importante y urgente.

La inclusión de la materia en el actual Plan de Estudios refiere a la formación necesaria en ese campo, en el que aparecen conjugados aspectos formativos e instrumentales. Es formativa en tanto desarrolla una forma especial de pensamiento y es instrumental ya que proporciona herramientas de trabajo al futuro profesional de la Educación Física: de alguna u otra, todo Profesor o Licenciado en Educación Física necesita del uso de las técnicas estadísticas.

⁷ Información proporcionada por el Departamento de Educación Física, FaHCE. UNLP.

Debe señalarse también que la estadística es una metodología utilizada por distintas profesiones; luego, se trata de un saber que permite el vínculo interdisciplinario, el trabajo y la interacción con otras miradas y perspectivas, permitiendo imprimir la mirada de la Educación Física y la riqueza de su comprensión de la realidad, a las otras áreas de trabajo (Vilchez Guizado y Ramón Ortiz, 2018).

Los/as estudiantes de la Carrera reconocen que las estadísticas están más cerca de nuestra profesión de lo que creemos, por lo tanto, nos referimos a un tipo de conocimiento que se considera fundamental para el perfil del egresado/a. Su uso permite comprender dinámicas sociales, adaptar estrategias pedagógicas para optimizar los procesos de enseñanza y aprendizaje, evaluar el impacto de programas educativos, recreativos y deportivos, en distintos contextos (Fornessi, 2022).

En tal sentido, hay un aspecto realmente clave que queremos remarcar: el carácter político, social, cultural e histórico que tiene atraviesa la enseñanza de la estadística para nosotros/as, al tiempo que complejiza su relación con la Educación Física interpelando a los/as estudiantes en su cotidianidad y sus modos de estar en el mundo sobre el cual les toca trabajar.

La aplicación de herramientas estadísticas resulta de utilidad tanto en los procesos de investigación científica como siendo soporte técnico de la propia práctica docente y profesional. Casi todas las investigaciones aplicadas requieren algún tipo de análisis estadístico para que sea posible evaluar sus resultados. En algunos casos, para resolver un problema de carácter empírico, es preciso llevar a cabo análisis de características complejas; otras veces, basta con efectuar un análisis muy simple y directo.

El conocimiento estadístico alienta el espíritu crítico frente a los datos publicados y problematiza sobre su construcción e interpretación. Permite, en cualquier instancia del proceso educativo o de la práctica profesional, darle un sentido a los datos disponibles al poder ordenarlos, sistematizarlos, resumirlos, analizarlos, compartirlos y repensarlos.

Tal como señala Botanero (2002) al hablar de *cultura estadística*, ésta no sólo sirve como herramienta clave en la formación profesional del Profesor y Licenciado en Educación Física, sino que busca desarrollar la capacidad para, reflexionar y comunicar una posición crítica respecto de los resultados y conclusiones a los que se arriben basados en el uso de la estadística.

Somos sujetos sociales, atravesados por nuestra formación. El estudio de la estadística debe despertar en nosotros la necesidad ética de ser rigurosos en aquello que afirmamos, críticos de los conocimientos disponibles y creativos en los aportes al campo propio de la educación física.

Si bien está basada en la matemática, la estadística es teoría y aplicación; si bien trabaja con razonamientos complejos, esta materia nos hace incorporar un tipo de razonamiento de índole lógica deductiva/inductiva. Cabe destacar, que el propósito es que los/as estudiantes se apropien de nociones que les permita pensar que la estadística más allá de cálculos y operaciones algebraicas, sino que contribuye a la formación de profesionales críticos, éticos, cuyo modo de entender la realidad que nos circunda cuenta con útiles herramientas que

aportan fundamento a diversas áreas dentro de la disciplina que nos ocupa. Por ello, nuestra perspectiva se propone superar la mirada de un modelo biologicista o uno basado exclusivamente en hechos y resultados; buscamos enfatizar que la enseñanza-aprendizaje de análisis e investigación con métodos estadísticos se encuentra dentro del paradigma de las ciencias sociales.

Por todo lo dicho, el desarrollo de los contenidos de la materia busca proporcionar a los/as estudiantes conocimientos acerca de los procedimientos que permiten resumir e interpretar información con técnicas estadísticas descriptivas e introducirlos en el uso de técnicas estadísticas inferenciales que permitan obtener conclusiones a partir de bases rigurosas y que ayuden a tomar decisiones bajo incertidumbre.

Ello permitirá desarrollar las competencias básicas para la lectura comprensiva y el análisis crítico de trabajos de investigación que utilicen técnicas estadísticas, vincular y aplicar los saberes a la práctica en Educación Física, aplicar los principios éticos del ejercicio profesional a la construcción e interpretación de los datos, posicionándose como profesionales anclados históricamente, en el contexto socio-político y cultural actual.

Referencias

- Batanero, C. (2002). "Los retos de la cultura estadística", Conferencia Inaugural de las Jornadas Interamericanas de Enseñanza de la Estadística. Buenos Aires, Argentina. Disponible en: <https://www.ugr.es/~batanero/pages/ARTICULOS/CULTURA.pdf>
- Bologna, E. y Morales, M. (2011) "Presentación: ¿Estadística en psicología y educación?", en *Estadística para psicología y educación*. Editorial Brujas.
- Castro Nogueira, L.; Castro Nogueira M.; Morales Navarro, J. (2005). "Concepto de medición", en *Metodología de las Ciencias sociales, una introducción crítica*. Editorial Tecnos.
- Fornesi, R. (2022). *Caracterización y análisis de la experiencia de estudiantes de Educación Física en Estadística Aplicada a la Educación Física (2017-2020)*. Revista Educación Física y Ciencia, 24 (4), e237, <https://doi.org/10.24215/23142561e237>. Disponible en: http://www.scielo.org.ar/scielo.php?script=sci_arttext&pid=S2314-25612022000400237&lng=es&nrm=iso
- Johnson, R. y Kuby P. (1999) Estadística elemental. Cap. 3 México: ITP. Disponible en: [file:///C:/Users/Usuario/Downloads/Estadistica%20Elemental%20-%20Johnson%20Kuby%2011ed%20\(1\)-1.pdf](file:///C:/Users/Usuario/Downloads/Estadistica%20Elemental%20-%20Johnson%20Kuby%2011ed%20(1)-1.pdf)
- Sosa Escudero, W. (2020). "Mentiras verdaderas. A modo de introducción, en *¿Qué es (y qué no es) la estadística? Usos y abusos de una disciplina clave en la vida de los países y las personas*. Editorial Siglo XXI.
- Vilchez Guizado, J.; Ramón Ortiz, J. (2018). Elementos de estadística y probabilidades. Editorial Académica Española.

Stevens, S. S. (1946). Sobre la Teoría de las Escalas de Medición. USA: Asociación Estadounidense para el Avance de la Ciencia. Science. Nueva Serie, Vol. 103, No. 2684, p. 677-680. Traducción de la Facultad de Psicología, de la Universidad Nacional Autónoma de México. Disponible en https://berniececl.weebly.com/uploads/7/2/5/3/72531/02_s.s._stevens_-_sobre_teora_escalas_de_medicin_esp.pdf

CAPÍTULO 2

Sobre las unidades de análisis y sus características

Lucía Girardi y Rocío Belén Villafañe

Para desarrollar los temas que nos competen en este capítulo, resulta necesario partir del concepto de **dato**, por lo que comenzaremos con una breve definición de este concepto para luego explicar cada uno de sus elementos.

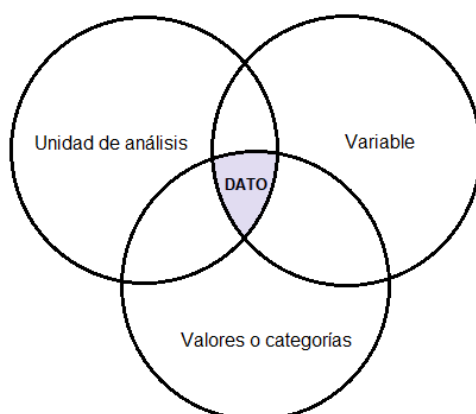
Dato

Si bien se pueden encontrar distintas definiciones de lo que es un dato tomaremos aquí como punto de partida la definición de Galtung (1966). Según este autor, el dato posee una estructura tripartita, dado que está compuesto por tres elementos: **unidad de análisis, variable y valor o categoría de la variable**. Lo que constituye el dato son estos tres elementos considerados conjuntamente.

Entonces un dato consiste en una unidad de análisis que, en una variable específica, presentará un determinado valor.

Figura 2.1

Elementos del dato



Nota. Fuente: elaboración propia

Unidades de análisis

Las unidades de análisis hacen referencia a aquello que se quiere estudiar, son el objeto de la investigación, identifican a qué o a quién se investiga, son los elementos entre los que se compara alguna cualidad. Tienen características constantes y características variables. Nos centramos en las características constantes para definir cada uno de los individuos que son objeto de nuestro estudio. De esta manera, podemos establecer los límites entre quienes integran la población en estudio y quienes no la integran.

Características de las unidades de análisis

Las unidades de análisis pueden o no ser diferentes de las unidades de observación (aquellas de las que proviene la información). Tienen la propiedad de poseer cualidades o características constantes y, de ser posible, se encuentran acotadas espacio-temporalmente.

A su vez pueden ser individuales o colectivas, dependiendo de si se centran en características de un individuo o de un grupo.

A continuación, veremos algunos ejemplos de unidades de análisis para poder profundizar en lo explicado anteriormente:

- Personas: alumnas/os de 1° año de la Escuela Media N°1 “Manuel Belgrano” de la ciudad de La Plata del año 2025; clientes de 50 años en adelante del gimnasio “X” (año 2026); profesores de Educación Física trabajando en gimnasios de Ensenada en el período 2024-2026.
- Equipos: equipos de primera división de fútbol femenino argentino en el año 2024; equipos de vóley participantes de los JJOO de Tokio 2020, equipos de la Liga Nacional de Vóley Sentado del año 2025.
- Instituciones: escuelas públicas de la ciudad de La Plata cuyas clases de Educación Física se dictan en el Centro de Educación Física N°2 en el ciclo lectivo 2025; clubes de barrio de la Ciudad Autónoma de Buenos Aires registrados en el Registro Único de Instituciones Deportivas en el año pasado.

Dijimos que, de ser posible, las unidades de análisis deben estar acotadas espacio-temporalmente, ahora bien, ¿qué significa esto?, si volvemos al ejemplo de los equipos de primera división de fútbol femenino argentino del año 2026 podemos observar que la delimitación espacial es Argentina y la delimitación temporal el año 2026.

También habíamos dicho que las unidades de análisis pueden ser individuales o colectivas, si se observan los ejemplos antes citados se puede ver que se encuentran agrupados en “personas”, “equipos” e “instituciones”, los que se encuentran en el primer grupo corresponden a unidades de análisis individuales, debido a que se centran en las características de un individuo (alumna/o, cliente, profesores de Educación Física). Los que se encuentran dentro de

los dos siguientes grupos son ejemplos de unidades de análisis colectivas debido a que se centran en las características de un grupo (equipo deportivo, escuela, club).

Variables

Las **variables** representan aquellas propiedades de la unidad de análisis que se caracterizan por no ser constantes, es decir son aquellas propiedades que varían. Es aquello que se quiere estudiar de la unidad de análisis y se puede someter a comparación, es una característica observable que varía entre las diferentes unidades que conforman una población y es susceptible de adoptar diferentes valores.

En la siguiente tabla podemos observar algunos ejemplos de variables según la unidad de análisis seleccionada:

Tabla 2.1

Ejemplos de unidades de análisis y variables

Unidades de análisis	Variables
Personas: alumnos/as, clientes, profesores de educación física	Edad; estatura; peso; fuerza; resistencia; nacionalidad; género
Equipos: equipo de fútbol, equipo de vóley, equipo de vóley sentado	Cantidad de partidos ganados; cantidad de partidos jugados; cantidad de torneos ganados
Instituciones: escuelas, clubes de barrio	Capacidades condicionales ⁸ incluidas en el programa anual de EF; deportes ofrecidos

Nota. Fuente: elaboración propia

Valor o categoría

Un valor o categoría es cada uno de los distintos resultados que se pueden obtener en un estudio estadístico en relación con una variable. El conjunto de los valores que una variable puede asumir conforma su sistema de categorías.

Tabla 2.2

Ejemplos de valores o categorías según las unidades de análisis y variables

Unidad de análisis	Variable	Valores o Categorías
Alumna/o de 1º año de la Escuela Media N°1 “Manuel Belgrano”	estatura	155 cm, 157 cm, 163 cm, 164 cm ...
Equipo de primera división de fútbol femenino argentino	cantidad de torneos ganados	0, 1, 2, 3, 4 ...

⁸ Las capacidades condicionales básicas en EF son: fuerza, resistencia, velocidad, flexibilidad y coordinación.

año 2024		
Escuela de Educación Primaria N°5 (La Plata)	deportes de equipo ofrecidos	vóley, fútbol, handball ...

Nota. Fuente: elaboración propia

Las categorías de las variables deben ser exhaustivas y mutuamente excluyentes. Que sean exhaustivas hace referencia a que se deben considerar todas las posibles alternativas de respuesta. De no incluirse todas las posibles opciones se corre el riesgo de una errónea clasificación u omisión del dato. Que sean mutuamente excluyentes significa que un dato no debiera identificarse con más de una categoría.

Al respecto, Bologna afirma:

Hay dos propiedades que debemos asegurar que cumplan las categorías que construyamos. La primera se llama exclusión mutua, es decir que cada categoría excluya a todas las demás. Dicho de otra manera, si a un individuo le corresponde una categoría, entonces sabemos que no le corresponde ninguna otra. (Bologna, 2010, p.22)

“El segundo requisito que solicitaremos a las categorías de una variable es que agoten todas las posibilidades de variación, es decir, que todos los valores posibles estén contemplados. Esta cualidad se llama exhaustividad.” (Bologna, 2010, p.23)

A continuación, abordaremos un ejemplo donde podremos identificar los conceptos desarrollados hasta aquí: unidad de análisis, variable y categorías o valores de la variable.

Ejemplo: Se quiere realizar una investigación acerca de los deportes de equipos más difundidos en los clubes de barrio de la Ciudad de Buenos Aires en el período que va desde el año 2000 hasta el año 2023. Para identificar las unidades de análisis cabe hacerse la pregunta ¿sobre quiénes se investiga?, la respuesta a esta pregunta serán los distintos clubes de barrio de la Ciudad de Buenos Aires (año 2000 a 2022), estos serán las unidades de análisis. Ahora bien, si nos preguntamos qué es lo que se quiere investigar sobre esas unidades obtendremos las variables. Para el caso de este ejemplo la variable es: deportes de equipo ofrecidos, pero podrían haber sido otras características de los clubes como la cantidad de socios o los barrios donde se ubican.

Como ya hemos visto, estas variables pueden adquirir distintos valores y variar de un club a otro. Para el caso de la variable “deportes de equipo ofrecidos” las categorías o valores que puede tomar esta variable serían: handball, vóley, básquet, fútbol, hockey.

Tipos de variables y niveles de medición

Tipos de variables

Las variables son muy diversas, por lo que debemos clasificarlas dado que no todos los cálculos estadísticos se pueden hacer con todas las variables.

Las mismas pueden ser clasificadas según distintos criterios, por ejemplo: según el tipo de atributo que pretenden medir; el papel que cumplen en el análisis; el grado de complejidad; la unidad de análisis a la que se aplican; el nivel de medición.

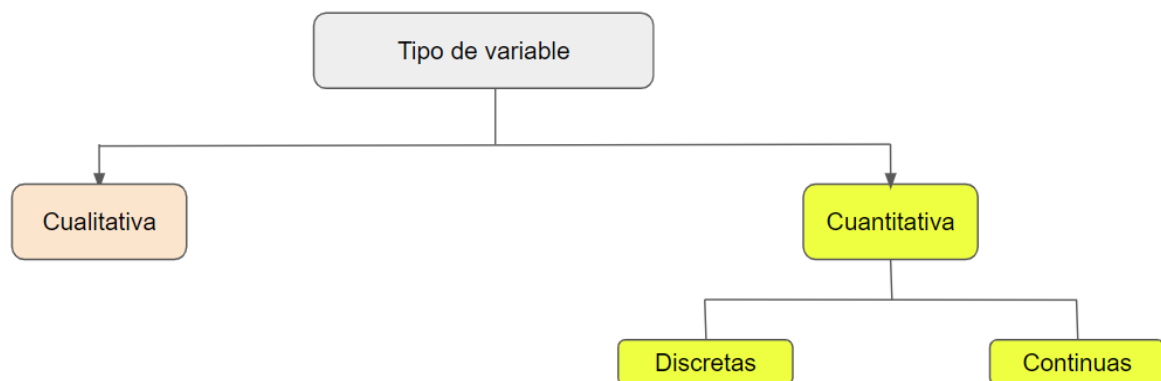
Nos ocuparemos aquí del desarrollo de las variables según el tipo de atributo que pretenden medir y según el nivel de medición.

Según el tipo de atributo que pretenden medir las variables se clasifican en:

- **variables cualitativas o categóricas**
- **variables cuantitativas o numéricas.** Dentro de las cuantitativas o numéricas, podemos encontrar las variables: **discretas y continuas**

Figura 2.2

Clasificación de las variables según el atributo que pretenden medir



Nota. Fuente: elaboración propia

Variables cualitativas o categóricas

Una variable es cualitativa cuando los valores que asume la misma son atributos que no pueden ser cuantificados. Este tipo de variables tiene como características que pueden ser divididas en clases separadas y mutuamente excluyentes.

Siguiendo los ejemplos de variables citados anteriormente algunos de los que se corresponden con este tipo son: nacionalidad, capacidades condicionales, deportes ofrecidos.

Tabla 2.3*Ejemplos de variables cualitativas y algunos de sus posibles valores o categorías*

Variable	Ejemplos de valores o categorías de las variables
Nacionalidad	- Argentino - Brasileño - Venezolano
Capacidades condicionales	- Fuerza - Resistencia - Flexibilidad
Deportes ofrecidos	- Handball - Vóley - Fútbol

Nota. Fuente: elaboración propia

Estas variables son muy frecuentes en las ciencias sociales y del comportamiento y son las más difíciles de analizar desde el punto de vista estadístico.

Variables cuantitativas

Las variables cuantitativas son aquellas cuyos valores o categorías pueden ser cuantificados, es decir, pueden ser expresados mediante números (siempre que expresen cantidades), son susceptibles de ser medidas o contadas.

Estas variables a su vez se clasifican según los valores que pueden asumir, en variables cuantitativas discretas y variables cuantitativas continuas.

- **Discretas:** cuando los valores que asume la variable pueden contarse y expresarse en números enteros. Ejemplo: número de socios de un club, cantidad de partidos jugados, número de torneos ganados, etc.
- **Continuas:** estas variables no se cuentan, sino que se miden, admiten valores intermedios y pueden tomar cualquier valor en un ámbito infinito de valores, por lo que hay continuidad entre un valor y otro. Por ejemplo, entre 20 m y 21 m hay una infinidad de valores posibles (20,51 m; 20,52 m; 20,57 m, etc).

Veamos un ejemplo más concreto: si una profesora está cronometrando el tiempo que tardan sus alumnos en recorrer trotando a un ritmo moderado 400 m, podrá elegir entre tomar el tiempo solo en minutos (min) o tomar también los segundos (s). Observemos en la siguiente tabla los tiempos que podría obtener según el criterio que adopte:

Tabla 2.4*Tiempos en recorrer 400 m por alumno*

Trote 400 m		
Alumno	Tiempo (min)	Tiempo (min y s)
Mía	3 min	3 min 7 s
Lucas	3 min	3 min 22 s
Agustín	4 min	4 min 46 s

Nota. Fuente: elaboración propia

Observamos que entre los 3 min y los 4 min hay otros valores posibles y que el límite entre un tiempo y otro estará dado por el criterio que se adopte para tomarlo.

En síntesis, las variables cualitativas son esencialmente variables para la clasificación de las entidades; las variables cuantitativas discretas son variables en donde las entidades son susceptibles de ser *contadas* y las variables continuas son aquellas variables que pueden ser medidas.

Niveles de medición

Tomaremos la teoría de Stevens (1951) para definir a la medición como el proceso mediante el cual se representan propiedades mediante números. Medir, entonces, es asignar números a objetos o datos de acuerdo con ciertas reglas. Las variables pueden ser medidas con mayor o menor grado de precisión según la escala de medida utilizada para su observación.

Según Maureira Cid (2015, p.26) es fundamental determinar correctamente el nivel de medición de las variables, dado que esto determina que tipos de análisis se puede realizar con ellas. Cada nivel de medición requiere de la utilización de un conjunto de instrumentos estadísticos distintos.

Las variables bien definidas se clasifican en uno de los 4 niveles de medición: nominal, ordinal, intervalar, de cocientes o razones. Estos niveles tienen un orden, siendo el nivel de medición nominal el más bajo y el nivel de medición de cocientes o razones el más alto. A su vez, el nivel de medición ordinal es un nivel más alto que el nominal y más bajo que el intervalar. Este último es más bajo que el de cocientes y más alto que el nominal y el ordinal.

Nivel nominal

El nivel de medición nominal involucra el proceso de denominar o etiquetar, esto es, colocar los casos dentro de categorías y contar su frecuencia de ocurrencia, sustituyendo los objetos

reales por números o símbolos, indicando sólo la diferencia respecto a una cualidad dada. Es el nivel más bajo de medición y permite la clasificación. Las categorías deben ser exhaustivas y mutuamente excluyentes.

Los datos nominales se rotulan con el fin de agrupar los casos en categorías separadas para indicar semejanza o diferencia respecto de una cualidad o característica dada. El nivel nominal clasifica objetos, personas o características.

No pueden teóricamente realizarse operaciones matemáticas con las variables del nivel nominal.

Ejemplos de variables de nivel nominal: nacionalidad, deportes ofrecidos, capacidades condicionales.

La variable nacionalidad pertenece al nivel de medición nominal ya que nos permite clasificar a las personas por el país en el que nacieron. Así, los valores de la variable nacionalidad, separan en categorías a los argentinos, brasileños, venezolanos, uruguayos, chilenos, bolivianos, paraguayos, etc. En esta clasificación no hay un orden intrínseco entre los valores o categorías de la variable nacionalidad; el orden que asumen es indistinto. Lo mismo sucede con la variable “deportes ofrecidos”, si lo que se busca saber es cuál es la oferta de deportes de los clubes barriales de alguna ciudad, algunos de los posibles valores que podría asumir la misma serían: handball, fútbol, vóley, vemos en estas categorías que no existe un orden implícito entre las mismas.

Veamos un último ejemplo, se realiza un estudio para saber cuáles son las capacidades condicionales más incluidas dentro de los programas de Educación Física de la provincia de Córdoba. Las categorías de la variable a analizar son: velocidad, resistencia, fuerza, flexibilidad, coordinación, todos valores distintos de los cuales ninguno es mejor o peor que otro.

Nivel ordinal

En este nivel se clasifican y ordenan los casos en términos del grado en que poseen una determinada característica.

Los valores que asume la variable, además de ser distintos entre sí, poseen un orden intrínseco, es decir, entre las distintas categorías se podrá establecer un orden jerárquico en el que una categoría implica más o menos que la otra, hay un orden que puede ser de menor a mayor o de mayor a menor.

Este nivel de medición permite clasificación y orden. Además de clasificar nos da información acerca de la organización de las categorías, pero no indica la magnitud de las diferencias entre las mismas.

Al igual que en el nivel nominal no se pueden realizar operaciones matemáticas en este nivel. Sin embargo, las variables pueden cuantificarse siempre y cuando se asignen números a las categorías manteniendo la relación “mayor que” y “menor que”. Es decir, se puede hacer cualquier transformación mientras se mantenga el orden jerárquico intrínseco.

Ejemplo de variable de nivel ordinal: el índice de masa corporal (IMC) cuando los valores que asume esta categoría son: bajo peso; peso normal; sobrepeso y obesidad.

Luego de calcular el índice de masa corporal (IMC) se puede clasificar el peso de adultos mayores de 20 años según los resultados obtenidos como se observa en la siguiente tabla.

Tabla 2.5

Interpretación del IMC para adultos mayores de 20 años

Nivel de peso	IMC
Bajo peso	menor a 18.5
Peso normal	entre 18.5 y 24.9
Sobrepeso	entre 25 y 29.9
Obesidad	30 o más

Nota. Adaptado de

https://www.cdc.gov/healthyweight/spanish/assessing/bmi/adult_bmi/index.html

Vemos que los valores que asume la variable IMC, además de clasificar a las unidades de análisis, nos permite ordenarlas de acuerdo al valor que asume el IMC medido. Es decir, el nivel de medición ordinal separa a las personas en categorías (bajo, normal, sobrepeso u obeso) pero al mismo tiempo, dichas categorías, nos muestran la relación que les asigna un orden. Quien obtenga un IMC en la categoría “normal” registrará un valor menor de IMC que aquella persona que tenga un IMC asociado con la categoría “obesidad”. Hay un orden jerárquico establecido entre una categoría y otra.

Nivel intervalar o de intervalos

Este nivel clasifica, ordena e indica distancia entre distintas categorías. Las distancias que existen entre un valor y otro se denominan intervalos, en el nivel de medición intervalar esos intervalos son distancias fijas. Por ejemplo, la distancia entre 12 y 14 es la misma que entre 24 y 26.

Lo que caracteriza a una escala de intervalo es la existencia de una unidad de medida común y constante que permite asignar un número real a todos los pares de objetos del conjunto ordenado.

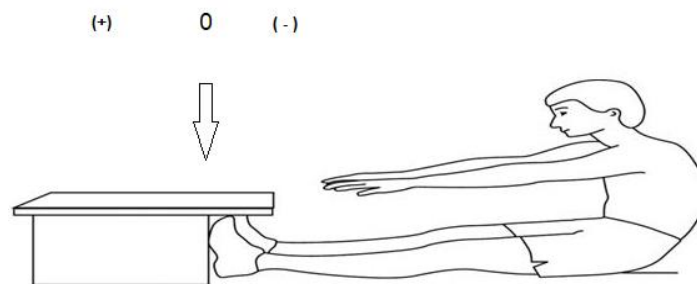
En una escala de intervalo el punto cero y la unidad de medición son arbitrarios. El punto cero está arbitrariamente determinado y no representa ausencia completa de la característica que se mida.

Por ejemplo, el test *sit and reach* (o test de Wells y Dillon) mide la flexibilidad de los isquiotibiales y la flexión del tronco. El mismo se realiza desde la posición de sentado y

consiste en que el evaluado apoye los pies en una tarima de madera sobre la cual hay una escala de graduación numérica y realice una flexión del tronco extendiendo los miembros superiores. Se mide el punto al que llegan los dedos de las manos pudiendo ser el resultado positivo o negativo. En esta escala numérica el cero coincide con el punto donde se apoyan los pies. Si los dedos pasan la punta de los pies el resultado será positivo, si pasa lo contrario será negativo.

Figura 2.3

Test “sit and reach”



Nota. Figura adaptada de <https://www.efdeportes.com/efd177/niveles-de-flexibilidad-del-grupo-adulto-mayor.htm>

Observamos entonces que el cero en este *test* no representa la ausencia de flexibilidad, sino que es un cero arbitrario.

En este nivel se puede hablar de cuantificación de variables y permite todo tipo de operaciones matemático-estadísticas. A su vez posee las características de los niveles anteriores y permite la utilización de las mismas técnicas estadísticas que los anteriores.

Ejemplos: coeficiente de inteligencia, temperatura en Centígrados y Fahrenheit, los puntajes de los test de inteligencia, de aptitudes o de conocimiento.

Nivel de razón o de cocientes

Es similar al nivel de medición intervalar puesto que permite clasificación, orden y distancia. La única diferencia es que el cero en este nivel es *absoluto*, representando la ausencia completa de la característica que mide. En el nivel de cocientes lo único arbitrario es la unidad de medida. Este nivel permite la clasificación, ordena, indica distancia y parte de un 0 (cero) absoluto⁹.

Al igual que el nivel de medición intervalar este nivel permite todo tipo de operaciones matemáticas.

⁹ Tiene un punto cero verdadero en su origen, el “cero” de la escala ha de coincidir con la *nada* o ausencia real de la variable a medir. Cero auténtico.

Ejemplo: peso en kilogramos, distancia recorrida en metros, altura en centímetros, tiempo en segundos, velocidad en metros por segundos.

Debemos hacer una aclaración con respecto a los niveles de medición. El nivel de medición de una variable siempre estará definido por el sistema de categorías y este último es una construcción que hacemos en función de nuestros objetivos, es por ello por lo que una variable puede definirse a partir de más de un sistema de categorías.

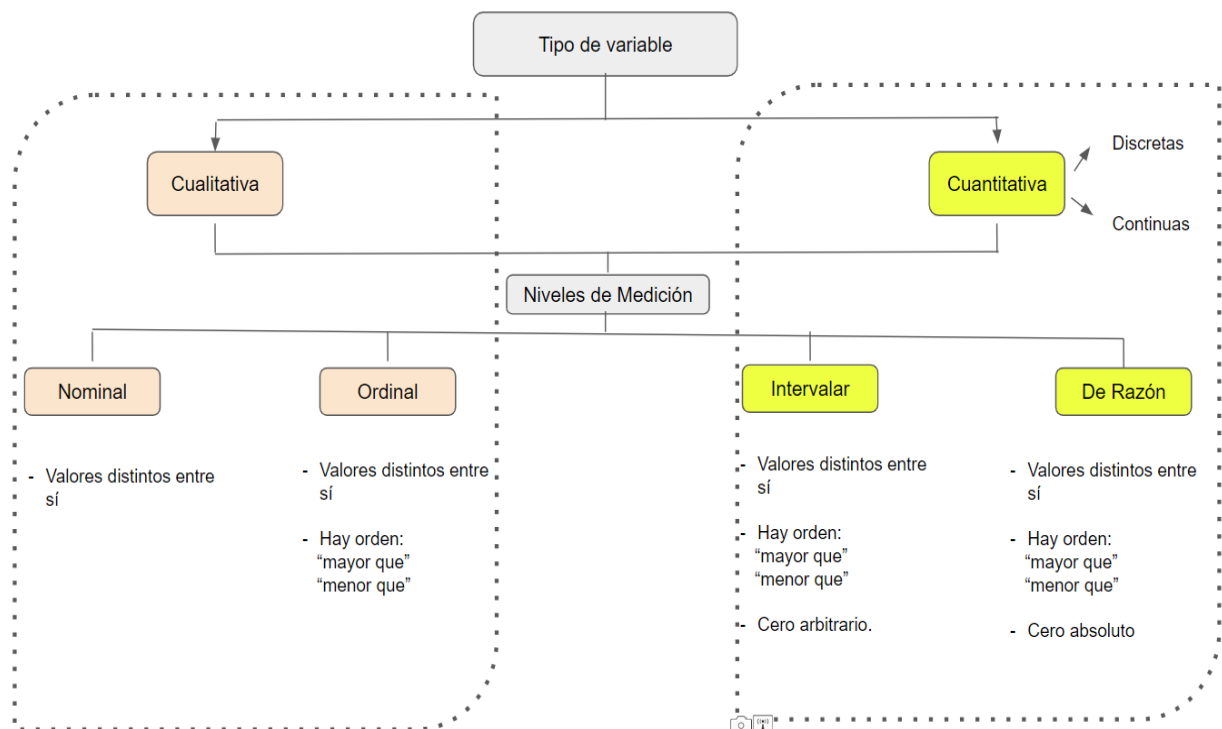
La definición de las variables a estudiar se asocia directamente con nuestros objetivos de conocimiento. Si estudiamos la variable peso medida en kilos, la variable pertenecerá al nivel de medición de cocientes o razones. Los valores que puede asumir la variable serán por ejemplo 64,5 kg; 77 kg; 80 kg. Pero, si al peso lo medimos a partir de las categorías bajo peso, normal y sobrepeso, estamos clasificando a las personas a partir de otro sistema de categorías, y la variable pertenecerá al nivel de medición ordinal.

En todos los niveles de medición debe regir la exhaustividad y las categorías deben ser mutuamente excluyentes. Cualquier procedimiento de medición debe ser exhaustivo, es decir, con categorías suficientes para que puedan clasificarse todos los datos dados. Las categorías deben ser mutuamente exclusivas y excluyentes.

Los niveles de medición nominal y ordinal son niveles que se utilizan para medir variables de tipo cualitativas mientras que el intervalar y de razón se utilizan para medir variables de tipo cuantitativas.

Figura 2.4

Gráfico resumen de los tipos de variables y los niveles de medición



Nota. Fuente: Elaboración propia

Guía para reconocer el Nivel de Medición de la variable

Una vez identificadas la unidad de análisis, el tipo de variable/s y su/s sistemas de categorías se debe identificar a qué Nivel de Medición corresponde cada variable. Como mencionamos anteriormente, se puede establecer un orden incremental entre los Niveles de Medición siendo el orden resultante el siguiente: nominal; ordinal; intervalar; de razón.

Es por ello que para reconocer el Nivel de Medición de cada variable es recomendable verificar si los valores que asume la hacen susceptible de ser identificada con el Nivel de Medición Nominal, de ser así se debe constatar la probabilidad de ubicarla en el Nivel de Medición inmediatamente siguiente, que en este caso sería el ordinal, y así sucesivamente. Si puede ser ubicada bajo el nivel inmediatamente siguiente se debe seguir probando hasta llegar al nivel de cociente o razón, de no poder ubicarla en el nivel siguiente entonces corresponde al nivel inmediatamente anterior.

Paso a paso

1- Verificar si es una variable de **nivel de medición nominal**.

Pregunta clave: *¿los valores son distintos entre sí?*

Si la respuesta es afirmativa se debe corroborar si pertenece o no al nivel de medición siguiente.

2- Verificar si es una variable **de nivel de medición ordinal**.

Pregunta clave: *¿puedo identificar un orden implícito?*

Si la respuesta es negativa entonces el nivel de medición es el nominal (el inmediatamente anterior).

Si la respuesta es afirmativa se debe corroborar si pertenece o no al nivel de medición siguiente.

3- Verificar si es una variable de **nivel de medición intervalar**.

Pregunta clave: *¿puedo calcular la distancia entre las categorías?*

Si la respuesta es negativa entonces el nivel de medición es el ordinal (el inmediatamente anterior).

Si la respuesta es afirmativa se debe corroborar si pertenece o no al nivel de medición siguiente.

4- Verificar si es una variable de **nivel de medición de razón**.

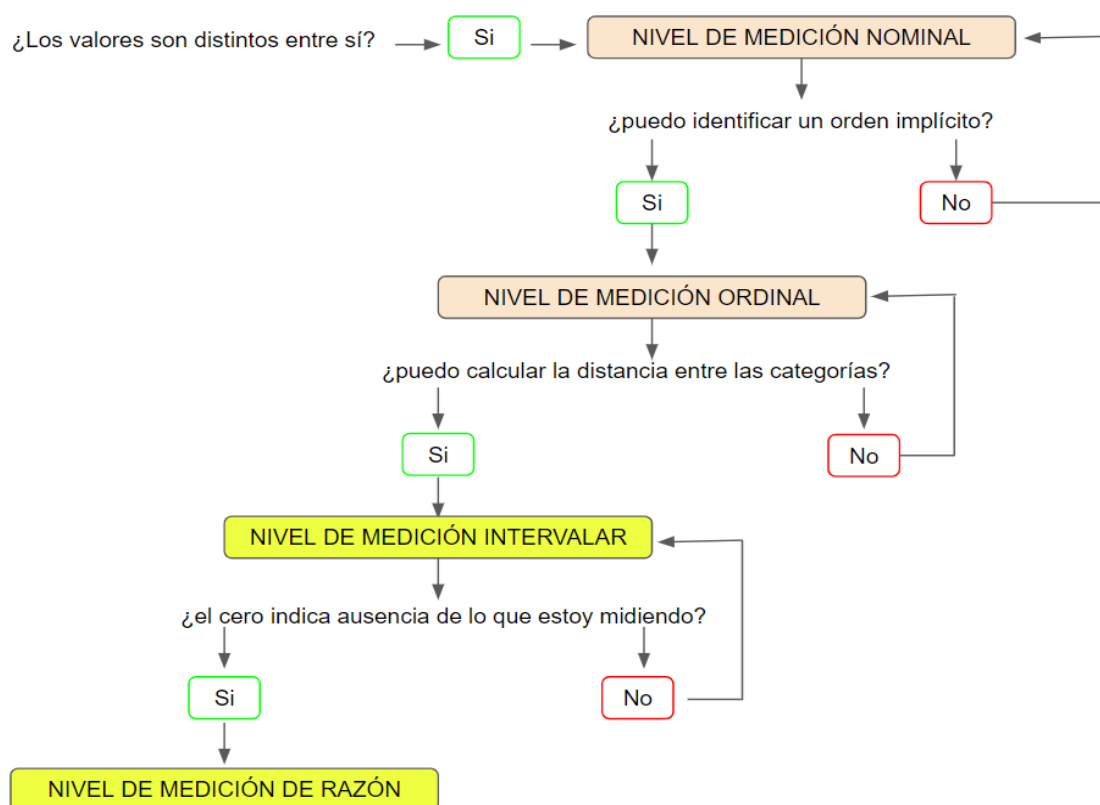
Pregunta clave: *¿el cero indica ausencia de lo que estoy midiendo?*

Si la respuesta es negativa entonces el nivel de medición es el intervalar (el inmediatamente anterior).

Si la respuesta es afirmativa entonces el nivel de medición de la variable es el de razón.

Figura 2.5

Gráfico guía para reconocer el nivel de medición de la variable



Nota. Fuente: elaboración propia

Referencias

Departamento de Salud y Servicios Humanos de EE.UU., Centros para el Control y la Prevención de Enfermedades. (2024, marzo 19). *Categorías de IMC en adultos*. CDC. <https://www.cdc.gov/bmi/adult-calculator/bmi-categories.html>

Valderrama, F. y Flores Rivera, C. (2013). Aplicación de un programa de actividad física integral y su incidencia en los niveles de flexibilidad del grupo adulto mayor D.I.P.R.E.C.A. de la comuna de Concepción, Octava Región, Chile. *EFDeportes.com, Revista Digital, Volumen 17* (Nº 177). <https://www.efdeportes.com/efd177/niveles-de-flexibilidad-del-grupo-adulto-mayor.htm>

Bibliografía

Galtung, Johan (1966). *Teoría y métodos de Investigación social*. Editorial Eudeba.

González de Menne, M. C.; Faas, A.; Bologna, E. (2010). "Las variables y su nivel de medición", en Bologna E., *Estadística para psicología y educación*, (pp.19-41). Editorial Brujas.

Maureira Cid, Fernando (2015) *Estadística básica para educación física*, Editorial Académica Española.

Cortada de Kohan, Nuria (1994) *Diseño Estadístico*. Editorial Eudeba.

CAPÍTULO 3

Organización, presentación y comunicación de los datos

Martina Díaz y Román Fornessi

Presentación del capítulo

Hasta acá describimos a grandes rasgos qué es la estadística, sus funciones descriptiva e inferencial, dijimos que trabaja con datos, abordamos su estructura tripartita, vimos el concepto de variable y cómo estas pueden clasificarse. En este capítulo, veremos cómo deberíamos proceder operativamente en el marco de una investigación o una exploración sobre algún aspecto de la realidad que nos interese conocer. Cuando decimos “operativamente” nos referimos a cuáles son las decisiones y los procedimientos concretos que vamos a llevar adelante para trabajar con datos. En este sentido, presentaremos distintas formas de organizar y presentarlos para facilitar su comunicación y apropiación atendiendo a características de las variables con las que estemos trabajando. Avanzaremos así en la presentación de la matriz de datos, las tablas de frecuencias y las representaciones gráficas.

Por otro lado, abordaremos crítica y propositivamente algunas herramientas y estrategias de comunicación de datos, no solamente pensándonos como sus productores/as sino también en nuestro rol de ciudadanos/as que consumimos permanentemente datos estadísticos. La apuesta aquí será allanar el camino para la elaboración de una lectura crítica de esas elaboraciones, a partir de la consolidación de lo que Batanero (2001) llama una “cultura estadística”. En esta línea, presentaremos lo que llamamos “medidas de resumen” para referirnos a tasas, proporciones y razones.

En suma, en este capítulo ofreceremos instrumentos, herramientas y estrategias para la elaboración de datos, pero también para su comprensión crítica.

Matriz de datos

Las variables pueden ser analizadas en dimensiones, es decir que puedo delimitar y definir con mayor precisión qué es lo que realmente me interesa medir en cada caso. Por ejemplo: podemos plantear como variable a la edad, pero sería mucho más preciso si se aclarara a qué

dimensión de esa variable quiero ceñirme: ¿me interesa la edad propiamente dicha o el grupo etario de mi unidad de análisis? Si pensamos el sistema de categorías para una u otra definición teórica de la variable podríamos trabajar de la siguiente manera:

- Variable Edad (cronológica, en años cumplidos). Valores posibles de la variable: 10 años, 11 años, 12 años, 13 años, etc. Si pensamos en el grupo etario, un posible sistema de categorías podría ser de 10 a 14 años; de 15 a 19 años, de 20 a 24 años, etcétera.
- Otra opción podría ser como etapa evolutiva: Niño/a, adolescente, joven, adulto/a, adulto/a mayor (en este caso, la variable deberíamos modificarla ligeramente, ya no hablamos de edad cronológica, sería más exacto hablar de “grupo poblacional en función de la edad”, por ejemplo).

Como se ve, en función de cómo construyamos la variable, vamos a tener distintas opciones para elaborar el sistema de categorías, en cada caso.

Una vez definidas las variables con las que deseamos trabajar y sus sistemas de categorías, procederemos con la recolección de los datos. Los instrumentos para esa recolección son variados: encuesta, test, entrevistas, medición directa, etcétera.

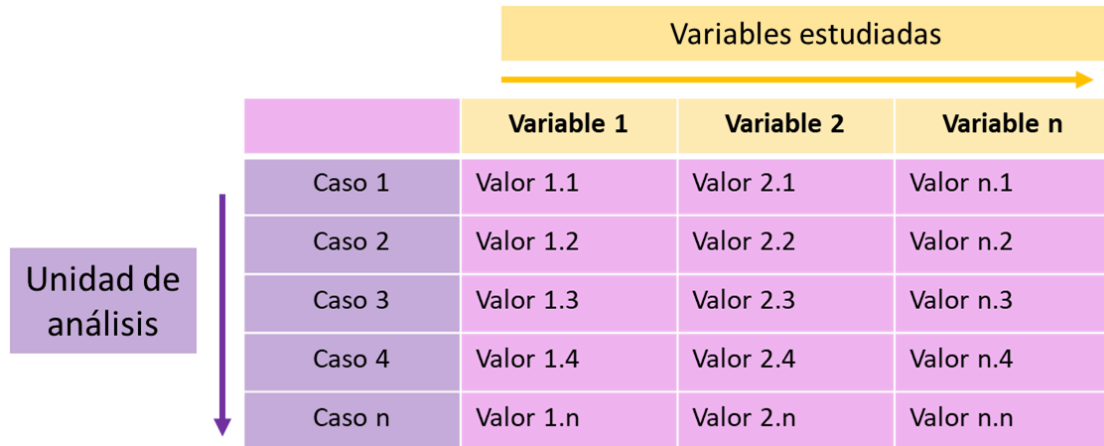
Al finalizar la recolección de esos datos, lo que debemos hacer es sistematizarlos, disponerlos de forma ordenada, para que podamos empezar a trabajar con ellos. La forma habitual de realizar esta sistematización es elaborando una matriz de datos, que es una tabla de doble entrada cuyas hileras se corresponden con los casos que representan a la unidad de análisis, y las columnas corresponden a las variables estudiadas. Esta forma de presentar los datos es sumamente práctica para realizar, luego, los análisis estadísticos, ya sea manualmente o a partir de programas computarizados.

La matriz es una forma de presentar los datos que recolectamos de manera ordenada y sistemática: cuánto más desordenado y trabajoso sería tener todas las encuestas, una por una, en una caja; mientras que la matriz de datos ofrece toda esa información en una sola planilla (como una hoja de cálculos de Excel, por ejemplo).

En la matriz se cristaliza la estructura tripartita del dato, que mencionamos en el capítulo anterior, podremos ver cómo cada observación se corresponde con un valor que, para cada variable, asume un caso de nuestra unidad de análisis. Si, en la matriz, nos posicionamos sobre un valor, vamos a ver este entrecruzamiento entre: el propio valor, el caso (de la unidad de análisis) y la variable. Se llama “estructura tripartita” porque está relacionando tres partes: el caso de la unidad de análisis, la variable, y el valor.

En la matriz, los datos se organizan de manera tal que en cada fila se ubican los casos de la unidad de análisis, y en las columnas cada una de las variables que estoy midiendo. El cruce entre cada variable y cada caso se corresponderá con un valor.

De manera gráfica, podemos observar a continuación lo mencionado:

Figura 3.1*Matriz de datos*


Variables estudiadas			
	Variable 1	Variable 2	Variable n
Caso 1	Valor 1.1	Valor 2.1	Valor n.1
Caso 2	Valor 1.2	Valor 2.2	Valor n.2
Caso 3	Valor 1.3	Valor 2.3	Valor n.3
Caso 4	Valor 1.4	Valor 2.4	Valor n.4
Caso n	Valor 1.n	Valor 2.n	Valor n.n

Nota. Fuente: elaboración propia

A continuación, podrán ver un ejemplo de una matriz de datos construida en Excel:

Figura 3.2*Matriz de datos en Excel*

Archivo		Inicio		Insertar		Disposición de página		Fórmulas		Datos		Revisar		Vista		Ayuda							
 Pegar		Calibri 11		 N		 <i>K</i>		 <u>S</u>		 		 		 		  				 Ajustar texto		General	
Portapapeles		Fuente		Alineación		Número																	
J22																							
A		B		C		D		E		F		G		H		I							
UNIDADES		VARIABLES																					
DE ANALISIS		V1	V2	V3	V4	V5	V6	VX															
UA 1		1	50	3	2	5	6																
UA 2		2	43	4	2	3	4																
UA 3		2	47	2	3	5	4																
UA 4		1	51	2	2	4	2																
UA 5		2	47	3	2	3	5																
UA 6		2	39	1	3	6	3																
UA 7		1	41	3	2	2	1																
UA 8		1	38	4	3	5	1																
UA 9		1	57	1	1	7	2																
UA 10		2	36	1	1	6	6																
UA 11		2	40	1	3	5	4																
UA 12		1	55	2	2	4	3																
UA 13		1	59	1	2	5	2																
UA 14		1	40	4	1	2	4																
UA 15		2	49	3	2	7	2																
UA x																							
V1: Género		V4: ¿Vive en La Plata, Berisso o Ensenada?		V6: Situación ocupacional																			
1. Masculino		1. LP		1. Trabaja en relación de dependencia																			
2. Femenino		2. Berisso		2. Desocupado/a																			
V2: Edad cronológica en años cumplidos		3. Ensenada		3. Jubilado/a																			
V3: Estado civil		V5. Máximo nivel de estudios alcanzado		4. Rentista																			
1. Soltero/a		1. Sin estudios		5. Trabaja por cuenta propia																			
2. Casado/Concubinato		2. Primario		6. Otro																			
3. Separado/a		3. Secundario																					
4. Viudo		4. Terciario																					
5. Ns/Nc		5. Universitario																					

Nota. Fuente: elaboración propia

La matriz de datos habilita dos lecturas:

- Centrándose en la variable, las columnas se analizan de forma separada en cuanto a la información que brindan sobre la variable correspondiente ("lectura vertical"),
- Centrándose en un caso de la unidad de análisis, se observan las filas separadamente, en este caso vemos cómo se comporta cada unidad de análisis en cada variable.

Para realizar una matriz de datos ofrecemos, a continuación, un ejemplo:

Recordemos que en primer lugar vamos a definir las variables que queremos estudiar, luego, a partir del instrumento que elijamos vamos a recoger la información, es decir, los valores de nuestras variables presentados para todos y cada uno de los casos que conforman nuestra unidad de análisis, luego con esa información vamos a realizar la matriz de datos.

Supongamos que una docente de Educación Física está trabajando, en la escuela, el deporte Handball, y quiere realizar una evaluación al final del bimestre. Como la docente no acuerda con los modos de evaluación convencionales decidió que las/os estudiantes se evalúen entre sí. Les facilitó una planilla (que sería un instrumento de medición directa) cada dos estudiantes, confeccionada de la siguiente manera:

Figura 3.3

Modelo de planilla para recolección de datos

Recolección de datos

Nombre del estudiante:		
Deporte evaluado : Handball (Segundo año de secundaria)		
Género:		
Escuela:		
Edad:		
Indicadores o variables	Si	No
Se desplaza al espacio vacío de la cancha para recibir la pelota		
Pasa la pelota al compañero/a desmarcado/a		
Lanza a pie firme		
Lanza con ritmo de tres tiempos		
Se desmarca para recibir la pelota		
Utiliza códigos comunicacionales convenidos		
Participa de las situaciones tácticas acordadas		

Nota. Fuente: elaboración propia

En la parte superior presenta información relacionada al/a la estudiante, el contenido a ser evaluado, el año al que pertenece, y la escuela. En la parte inferior están las variables propias de la evaluación. Es decir:

- Se desplaza al espacio vacío de la cancha para recibir la pelota.
- Pasa la pelota al compañero/a desmarcado/a.
- Lanza a pie firme.
- Lanza con ritmo de tres tiempos.
- Se desmarca para recibir la pelota.
- Usa códigos comunicacionales convenidos.
- Participa de las situaciones tácticas acordadas.

Estas variables son cualitativas, porque están midiendo atributos o cualidades, y dicotómicas, porque van a indicar solamente la ausencia o presencia de eso que se está midiendo, que en la planilla se traduce en las opciones (categorías) “sí” y “no”.

Recordemos que la variable edad es cuantitativa y la variable género es cualitativa.

Veamos el procedimiento para llevar a cabo la recolección de datos: la docente tiene 30 estudiantes. En principio, armará dos grupos de 15 personas, un grupo evalúa y el otro juega

un partido de handball y luego rotan, quienes evaluaron en la primera etapa, luego jugarán y serán evaluados/as.

Una vez que todo el grupo fue evaluado, la docente se llevará toda esa información a su casa para organizar los datos obtenidos. A modo de ejemplo, la evaluación que arrojó el desempeño de un estudiante fue la siguiente:

Figura 3.4

Ejemplo de planilla para recolección de datos

Recolección de datos

Nombre del estudiante: Juan Martín		
Deporte evaluado : Handball (Segundo año de secundaria)		
Género: Varón		
Escuela: Escuela Secundaria Juana Azurduy de La Plata		
Edad: 13 años		
Indicadores o variables	Si	No
Se desplaza al espacio vacío de la cancha para recibir la pelota	X	
Pasa la pelota al compañero/a desmarcado/a	X	
Lanza a pie firme		X
Lanza con ritmo de tres tiempos	X	
Se desmarca para recibir la pelota	X	
Utiliza códigos comunicacionales convenidos		X
Participa de las situaciones tácticas acordadas	X	

Nota. Fuente: elaboración propia

Con los datos relevados, la docente construirá la matriz de datos en una tabla de Excel para sistematizar y organizar las evaluaciones de sus estudiantes:

Figura 3.5*Ejemplo de matriz de datos*

Unidad de análisis:
5° año de una
escuela secundaria
de la ciudad de La
Plata

Los casos de mi unidad de
análisis

	Se desplaza al espacio vacío de la cancha para recibir la pelota	Pasa la pelota al compañero/a designado/a	Lanza a pie firme	Lanza con ritmo de tres tiempos	Se detiene para recibir la pelota	Utiliza códigos comunicacionales en conversión	Participa de las situaciones lúdicas acordadas	Nivel de desempeño	Valoración numérica	Género	Edad
Estudiante 1	si	si	si	no	si	no	si	C	7	mujer	13
Estudiante 2	no	si	si	si	no	si	si	C	7	varon	13
Estudiante 3	no	no	si	no	no	si	no	D	3	varon	13
Estudiante 4	no	si	si	si	no	si	no	C	5	varon	12
Estudiante 5	si	si	si	si	si	si	si	A	10	no binarie	12
Estudiante 6	si	si	si	no	si	si	si	B	8	mujer	13
Estudiante 7	no	no	si	no	no	si	no	D	3	mujer	12
Estudiante 8	si	si	si	no	si	no	si	C	7	mujer	13
Estudiante 9	si	si	si	si	si	si	si	A	10	varon	13
Estudiante 10	no	no	si	no	no	no	no	D	3	no binarie	12
Estudiante 11	no	si	si	si	no	si	si	C	7	mujer	13
Estudiante 12	si	si	si	no	si	si	si	B	8	mujer	13
Estudiante 13	si	si	si	si	si	si	si	A	10	mujer	12
Estudiante 14	no	si	si	si	no	si	si	C	7	varon	13
Estudiante 15	si	si	si	si	si	si	si	A	10	no binarie	13
Estudiante 16	no	si	si	si	no	si	si	C	7	mujer	14
Estudiante 17	no	si	si	si	no	si	no	C	5	mujer	14
Estudiante 18	no	si	si	si	no	si	si	C	7	varon	13
Estudiante 19	no	si	si	si	no	si	si	C	7	varon	13
Estudiante 20	si	si	si	no	si	no	si	C	7	varon	13
Estudiante 21	si	si	si	no	si	no	si	C	7	mujer	13
Estudiante 22	no	no	si	no	no	si	no	D	3	varon	13
Estudiante 23	no	si	si	si	no	si	no	C	5	varon	12
Estudiante 24	si	si	si	si	si	si	si	A	10	no binarie	12
Estudiante 25	no	si	si	si	no	si	si	C	7	mujer	14
Estudiante 26	no	si	si	si	no	si	no	C	5	mujer	14
Estudiante 27	no	si	si	si	no	si	si	C	7	varon	13
Estudiante 28	no	si	si	si	si	si	si	B	8	varon	13
Estudiante 29	si	si	si	no	si	no	si	C	7	no binarie	13
Estudiante 30	si	si	si	si	si	si	si	A	10	no binarie	13

Variables cualitativas dicotómicas, nivel de medición nominal

Variable cualitativa, nivel de medición ordinal

Variable cuantitativa, nivel de medición intervalar

Variable cualitativa, nivel de medición nominal

Variable cuantitativa, nivel de medición de cocientes o razón

Esto es un dato.
Estructura tripartita: el estudiante número 9 asume el valor "sí" para la variable "lanza con ritmo de tres tiempos"

Nota. Fuente: elaboración propia

Como vemos, en esta tabla podemos identificar los elementos con los que venimos trabajando en estos capítulos: las variables, sus niveles de medición, la unidad de análisis, los valores que asumirán los distintos casos en cada una de las variables. En síntesis, podemos identificar, en cada una de las celdas de esa tabla, **datos**.

Recategorización de una variable

Esta matriz que construyó la docente de Educación Física nos muestra un ejemplo de recategorización. Recategorizar es reagrupar los datos que obtuvimos en la instancia de observación en nuevas categorías. Este tipo de operación se utiliza cuando queremos resumir la cantidad de datos con la que contamos, por ejemplo, cuando tenemos variables cuyos sistemas de categorías son extensos. Para ilustrar esto, pensemos en la variable "Edad en años cumplidos", el sistema de categorías de esta variable podríamos establecerlo entre 0 y 99 años. Una vez que yo observo las edades de todos los casos que componen mi muestra o mi

población, puedo recategorizarlos en “Grupos de edad”. Así, pondría a todas las personas que tienen 0, 1, 2, 3, 4 y 5 años en la categoría “Entre 0 y 5 años”. Aquellas personas que cuenten con 6, 7, 8, 9 y 10 años, van a pasar a formar parte del grupo “de 6 a 10 años”, y así sucesivamente. En resumen, estoy tomando la categoría original (supongamos 3 años para una persona con esa edad) y la reubico en una categoría nueva: “Entre 0 y 5 años”

Entonces si por ejemplo, en la escuela donde trabaja esta docente de Educación Física le solicitaran notas numéricas, con una recategorización de las variables estudiadas, podrá fácilmente calificar con nota numérica a sus estudiantes. La profesora decidió categorizar esas variables de la siguiente manera:

Figura 3.6

Recategorización de variables

RECATEGORIZACIÓN		
Nivel de desempeño	Valoración de criterios	valoración numérica
A (Alcanza objetivos)	Siete criterios demostrados	10
B (Satisfactorio)	Seis criterios demostrados	8
C (Suficiente)	Cinco criterios demostrados	7
	Cuatro criterios demostrados	5
D (No alcanza objetivos)	Tres o menos criterios demostrados	3

Nota. Fuente: elaboración propia

Una aclaración: la expresión “criterios demostrados” hace referencia a las variables que procuró medir la profesora en la planilla inicial: “se desplaza al espacio vacío de la cancha para recibir la pelota”, “pasa la pelota al/a la compañero/a desmarcado/a”, “lanza a pié firme”, “lanza con ritmo de tres tiempos”, “se desmarca para recibir la pelota”, “utiliza códigos comunicacionales convenidos”, “participa de las situaciones tácticas acordadas”. Siguiendo con el ejemplo de Juan Martín, vemos que alcanza un “sí” en 5 criterios, lo que le valdría una calificación de 7 puntos.

Así, en función de la cantidad de “sí” que tenga un/a estudiante, la docente podrá contabilizar y asignar una nota numérica a cada uno/a de ellos/as.

Una vez que tengamos construida la matriz de datos con toda la información recolectada vamos a poder analizar y describir las distribuciones de datos a través de la utilización de algunas herramientas que nos permitirán ordenar, resumir y representar los conjuntos con los que trabajamos. Estas herramientas son: tablas de frecuencias, representaciones gráficas y/o cálculos de resumen (razón, tasa, etc.)

Tablas de frecuencias

Las tablas de frecuencias son un modo sencillo de resumir los datos con los que estemos trabajando. Nos va a indicar con qué frecuencia se presenta cada una de las categorías de la variable, en el conjunto de las unidades de análisis en estudio. Facilita la apreciación de resultados, la realización de cálculos y la confección de gráficos.

Si continuamos con el ejemplo que venimos trabajando, la docente de Educación Física podrá visualizar y extraer conclusiones de manera sencilla y rápida. No sólo podrá saber quiénes aprobaron o no, sino que podrá detectar, por ejemplo, qué aspectos tendrá que seguir trabajando, qué contenidos tendrá que reforzar o qué temas nuevos podrá incorporar.

Por ejemplo, si tomamos la variable “Se desplaza al espacio vacío de la cancha para recibir la pelota”, podemos contabilizar y ver que sólo 13 estudiantes de los 30 realizaron esa acción, por lo tanto, si lo pensamos en términos de porcentaje, podríamos decir que un 43% de ellos/as aún no se desplazan a espacios vacíos para recibir un pase. Por lo tanto, la docente tendría que seguir trabajando ese aspecto porque casi la mitad del grupo no ejecuta esa acción. Distinto es con la variable **Lanza a pie firme** ya que, si observamos la matriz, el total de estudiantes lo hacen.

Todas estas conclusiones y deducciones las podemos observar fácilmente si construimos tablas de frecuencias. Dicha tabla es, también, un cuadro de doble entrada donde podemos observar cómo se comporta una variable determinada.

Continuando con nuestro ejemplo, vamos a observar cómo se comporta la variable “Valoración numérica” (recordemos que es una variable cuantitativa discreta y que la construimos en base a la cantidad de criterios demostrados por cada estudiante):

Tabla 3.1

Tabla de frecuencias

Valoración numérica (calificación) del/de la estudiante	Frecuencia absoluta	Frecuencia absoluta acumulada	Frecuencia relativa en proporción	Frecuencia relativa en proporción acumulada	Frecuencia relativa en porcentaje	Frecuencia relativa en porcentaje acumulada
3	4	4	0,1333	0,1333	13,33%	13,33%
5	4	8	0,1333	0,2666	13,33%	26,66%
7	13	21	0,4333	0,6999	43,33%	69,99%
8	3	24	0,1	0,7999	10%	79,99%
10	6	30	0,2	1	20%	100%
Total (n)	30	-	1	-	100%	-

Nota. Fuente: elaboración propia

En la primera columna, encontramos las calificaciones obtenidas por los/as estudiantes, inmediatamente después, la columna de las “Frecuencias absolutas” nos indica cuántos/as estudiantes obtuvieron cada una de esas calificaciones: 4 estudiantes alcanzaron un 3, 4 estudiantes obtuvieron un 5, 6 estudiantes fueron calificados con 10, etcétera. Obviamente, la cantidad total de estudiantes (que se simboliza con la letra n , al final de la tabla de frecuencias) suma 30, ya que era el número de personas que componían este grupo.

Las “Frecuencias relativas en proporción” nos indica cuánto representa cada categoría de la variable (cada calificación de la valoración numérica) en el total, por ejemplo: una proporción de 0,4333 del total del grupo obtuvo 7 puntos. La proporción es un dato “convertido” a la escala de la unidad (1): la mitad será la proporción equivalente a 0,5, un cuarto será 0,25, un tercio será 0,33, etcétera.

Las “Frecuencias relativas en porcentaje” expresan lo mismo que las anteriores, solo que en la escala que va del 0% al 100%. Siguiendo con el ejemplo anterior, el 43,33% de los/as estudiantes de este grupo obtuvo 7 puntos como calificación.

La potencia que tiene esta forma de organizar los datos es que nos facilita varias lecturas en una misma tabla:

- en términos absolutos: 13 estudiantes fueron calificados/as con 7 puntos;
- en términos relativos en proporción: 0,4333 es la proporción de estudiantes fueron calificados/as con 7 puntos;
- en términos relativos, en porcentajes: el 43,33% de los/as estudiantes fueron calificados/as con 7 puntos.

Otra ventaja de esta forma de expresar los datos se advierte al observar las columnas de “Frecuencias acumuladas”, esta expresión nos indica cuántos casos se encuentran en esa categoría o por debajo de ella.

En nuestro ejemplo, podríamos afirmar que “21 estudiantes obtuvieron hasta 7 puntos” como calificación. O que “21 estudiantes obtuvieron 7 o menos”. O que “9 estudiantes obtuvieron una nota superior a 7”, etcétera.

Entonces, elaborada esta tabla de frecuencias, esta docente de Educación Física podrá detectar por ejemplo que:

- si se aprueba con una nota de 6, entonces 8 estudiantes desaprobaron, es decir, un 26,6% del total.
- si se aprueba con una nota de 6, entonces 22 estudiantes aprobaron, es decir, un 73,4% del total.
- de los/as 22 estudiantes que aprobaron, 6 obtuvieron como nota un 10 y representan un 20% del total.

Estas y otras lecturas se construyen con facilidad a partir de la construcción de esa tabla de frecuencias.

Representaciones gráficas

Además de las tablas de frecuencias, existe otra forma de representar los datos y los comportamientos de las variables: los gráficos.

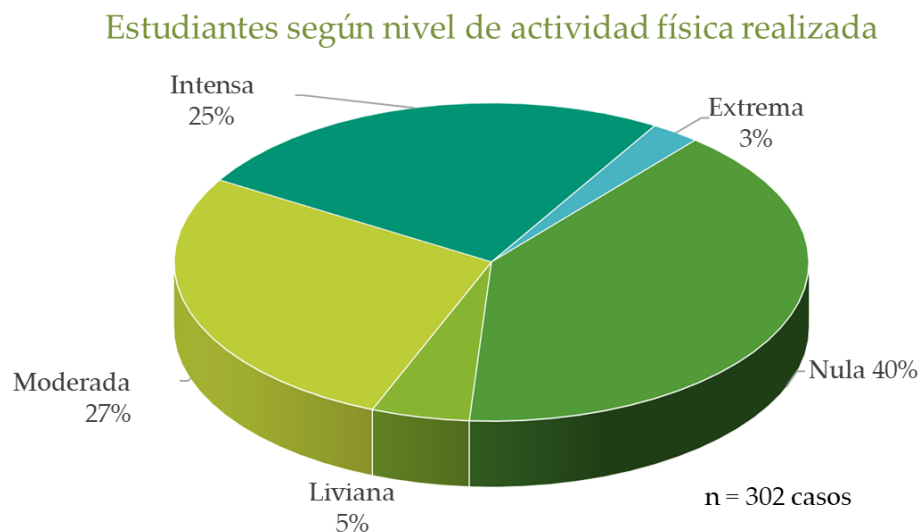
¿Qué me muestra un gráfico? La unidad de análisis, la/s variable/s, también, al ser una representación gráfica, podremos identificar tendencias, lecturas comparativas de las distintas magnitudes, la forma de la distribución, entre otros elementos.

Hay cierto consenso en la bibliografía (Maureira Cid, 2017, p. 42-53; Grima Cintas, 2011, p. 28-35; Bologna, 2010, p. 59-65) en torno a qué gráfico utilizar según el tipo de variable y el nivel de medición con los que estemos trabajando.

Por ejemplo, el gráfico sectorial (también conocido como gráfico de torta) es recomendable para variables cualitativas (nominales u ordinales), sobre todo cuando cuentan con un número pequeño de categorías (las variables dicotómicas son ideales para ser ilustradas en este gráfico).

Figura 3.7

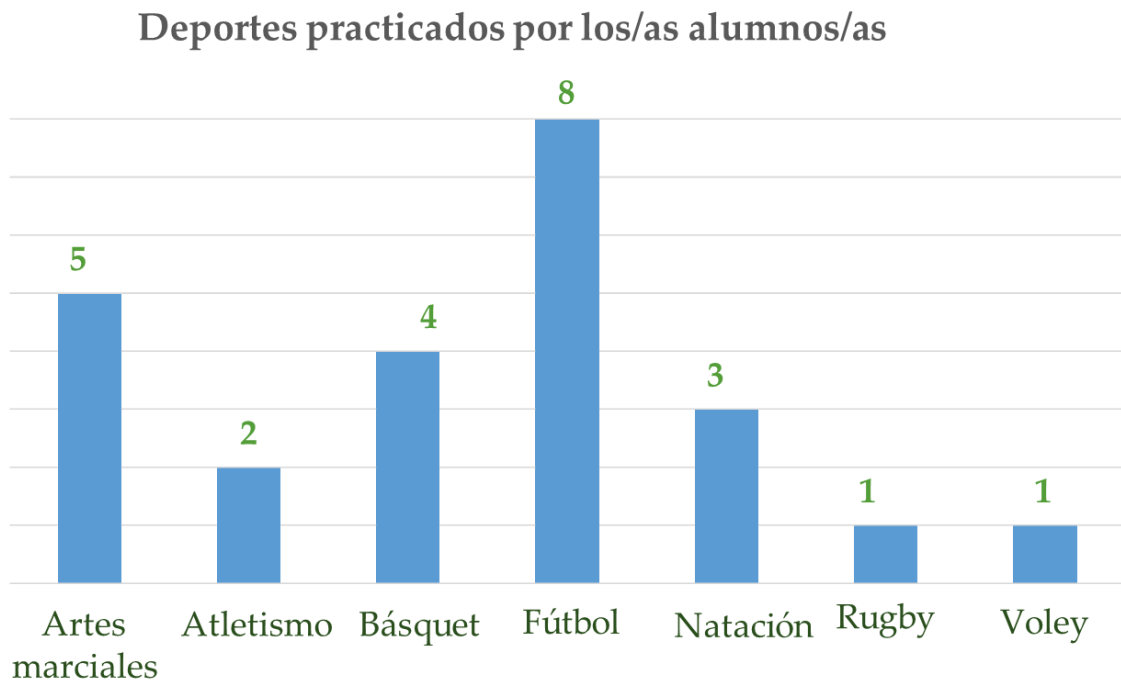
Gráfico sectorial



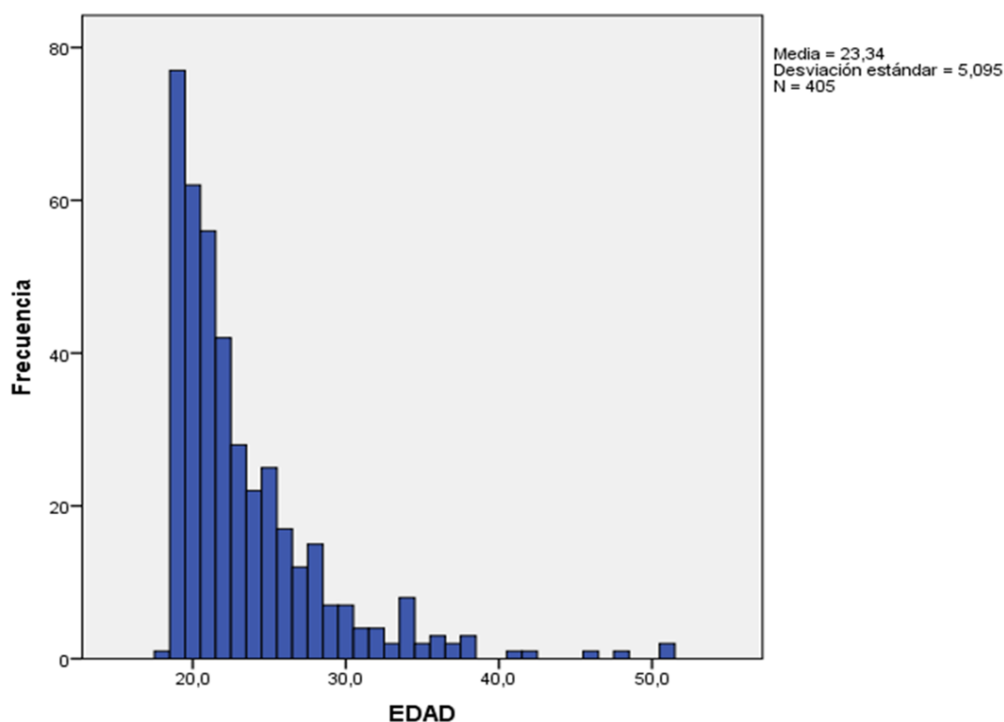
Nota. Fuente: elaboración propia

Este gráfico no es recomendable cuando la variable cuenta con una cantidad muy grande de categorías, porque su lectura deja de resultar intuitiva y rápida.

El diagrama de barras es recomendable cuando trabajamos con variables cualitativas (nominales y ordinales) y cuantitativas discretas, su potencia reside en habilitar una lectura rápida de las magnitudes, de las frecuencias de cada categoría. En el ejemplo que sigue, el gráfico nos permite leer, rápidamente, que en este grupo practican fútbol el doble de estudiantes que los/as que juegan al básquet.

Figura 3.8*Gráfico de barras**Nota.* Fuente: elaboración propia

Para representar el comportamiento de variables cuantitativas continuas, el histograma es una muy buena opción:

Figura 3.9*Histograma*

Nota. Cantidad de alumnas/os de Estadística Aplicada a la Educación Física por edad, año 2020.

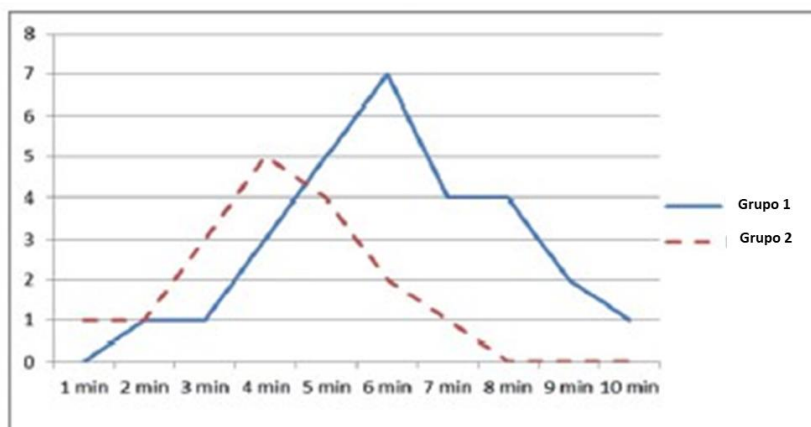
Fuente: elaboración propia

Nos muestra, rápidamente, donde se “concentran” los datos, pero, además, gráficamente, ayuda a demostrar que estamos trabajando con una variable continua, porque al estar las barras en contacto entre sí, da la sensación de continuidad en la escala de la variable.

Para este tipo de variable, otra representación gráfica recomendable es el polígono de frecuencias: es un gráfico similar al histograma, pero entrega una imagen de la curva que genera la distribución de la variable. Se construye teniendo en cuenta los puntos medios de cada intervalo de clase de la variable y nos permite comparar el desempeño de más de un caso en el mismo gráfico.

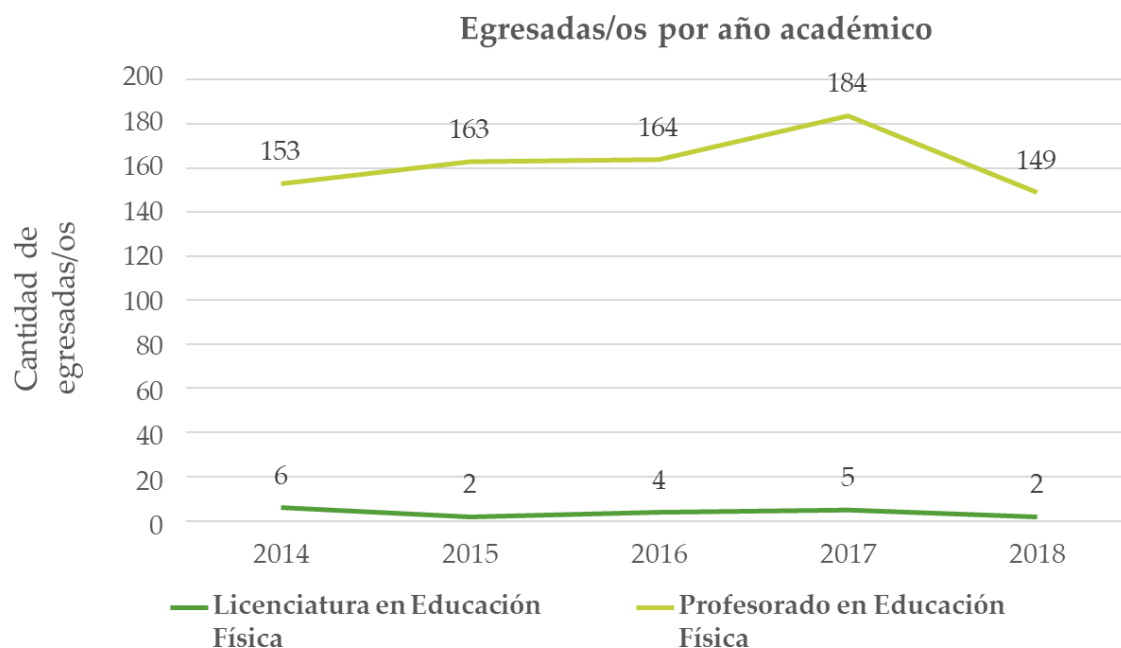
Figura 3.10*Polígono de frecuencias*

Puntuaciones en el test de Naveta según grupo

*Nota.* Fuente: elaboración propia

Como se desprende de este ejemplo, el grupo 1 presenta una mayor cantidad de casos que realizó el test en 6 minutos, mientras que el grupo 2 presenta una mayor cantidad de casos que realizó el test en 4 minutos.

Si nos interesa advertir cómo evoluciona una variable en el tiempo, es decir, qué resultados o valores va asumiendo en distintos momentos, el gráfico de líneas temporales es muy útil. En el siguiente ejemplo podemos observar cómo evolucionaron, entre el 2014 y el 2018, las variables “Cantidad de egresados/as del Profesorado en Educación Física” y “Cantidad de egresados/as de la Licenciatura en Educación Física” (UNLP):

Figura 3.11*Gráfico de líneas temporales*

Fuente: Registros en SIU-Guaraní

Nota. Fuente: elaboración propia en base a registros en SIU-Guaraní, FaHCE: UNLP.

Como vemos, las representaciones gráficas son instrumentos muy útiles y potentes para comunicar los datos con los que disponemos, no solamente porque cada gráfico concentra una cantidad de información que, si tuviéramos que expresarla de forma escrita nos insumiría mucho más tiempo, extensión y seguramente resultaría en una lectura menos fluida, sino porque, además, cada tipo de representación nos ayuda a comunicar mejor las lecturas de los datos en las que queremos enfocarnos.

Medidas “de resumen”

Llamamos medidas de resumen a datos que surgen a partir de la elaboración y el cálculo del/de la investigador/a, que tienen como finalidad comunicar una lectura determinada sobre la iniciativa o investigación que esté llevando adelante.

El abordaje de este tipo de medidas es fundamental en nuestros tiempos, ya que solemos encontrarnos con una enorme cantidad de afirmaciones en medios de comunicación, redes sociales, espacios de formación, etcétera, que, si bien comunican datos que pueden ser fidedignos, también se prestan para sustentar operaciones de lecturas parciales y sesgadas sobre la realidad.

Avancemos en una definición de cada una de estas medidas:

Una proporción se puede pensar, intuitivamente, como una porción, una parte de un total mayor, que la contiene. Por ejemplo: una porción de pizza representa una determinada proporción de toda la pizza. Si la cortamos en 8 porciones, podríamos decir que cada una de ellas representa una octava parte del todo. Técnicamente, la proporción se expresa como una parte de la unidad (la unidad sería 1).

Entonces, las proporciones siempre asumirán un valor entre 0 y 1:

Ejemplos:

0,25 = proporción que representa una cuarta parte del total.

0,50 (también se expresa 0,5) = proporción que representa la mitad del total.

0,75 = proporción que representa tres cuartas partes del total.

Atención: en la bibliografía y otras fuentes de información, pueden encontrar esta forma de expresar las proporciones junto con otra, que también es muy común y que es quitando el “0,” de la expresión. Ejemplo: en vez de indicar **0,5** (“cero coma cinco”) para una proporción que expresa la mitad de la unidad, pueden encontrar **.5** (“punto cinco”).

Sepan que es lo mismo:

0,5 = .5

0,15 = .15

0,99 = .99

Etcétera.

¿Cómo identificar la proporción que representa una porción en relación con el todo mayor, que la contiene? Sigamos con el ejemplo de la pizza: si la cortamos en 8 porciones, la proporción que represente cada porción surgirá de dividir 1 sobre 8:

$\frac{1}{8} = 0,125$ (o, lo que es lo mismo: .125)

Entonces, si me como dos porciones y alguien me pregunta “¿Qué proporción de la pizza te comiste?”, yo debería responder “0,25” (o, lo que es lo mismo: .25), ya que cada porción de la pizza representa, proporcionalmente, el 0,125 de la pizza. Entonces, $0,125 + 0,125$ (porque me comí dos porciones) = 0,25

Notemos cómo, si sumamos proporciones, vamos a ir representando una parte cada vez mayor del total (de la pizza): dos porciones serán una proporción de 0,25 de la pizza (la cuarta parte), cuatro porciones serán el 0,5 de la pizza (la mitad), etcétera.

Traslademos esto a un ejemplo propio de Educación Física. A continuación, mostramos una rutina de gimnasio:

Figura 3.12*Ejemplo rutina de gimnasio*

Rutina semana 1			
Nombre: Aike			
Profe: Amil			
	Series	Repeticiones	Carga
Estocadas con mancuernas	4	10	7,5 k
Sentadilla con despegue/salto	4	10	sin carga
Puente de gluteo con barra	4	6	40 k
Remo con TRX	4	8	sin carga
Press de hombro con barra	4	6	15 k
Vuelos frontales	4	6	4 k
Flexiones de pecho	3	8	sin carga
Biceps con mancuerna	3	8	4 k
Pelota pared	4	8	4 k
Elíptico	15 minutos		

Nota. Fuente: elaboración propia

Supongamos que quiero conocer cuál es la proporción que representa, en esta rutina, cada grupo muscular a ser trabajado. Lo primero que voy a hacer es identificar los ejercicios que responden a cada uno de ellos: los primeros tres podemos agruparlos dentro de “tren inferior”; remo, flexiones, *press* de hombro, vuelos laterales y bíceps, podría agruparlos dentro de “tren superior”. Para los ejercicios de pelota pared y elíptico, podría incorporarlos en un grupo llamado “otros”:

Tabla 3.2*Tabla de frecuencias con recategorización de variable*

		Cálculo para la proporción	Proporción
Tren inferior	3 ejercicios	$3/10=$	0,3
Tren superior	5 ejercicios	$5/10=$	0,5
Otros	2 ejercicios	$2/10=$	0,2
Total	10 ejercicios	$10/10=$	1

Nota. Fuente: elaboración propia

Así, identificamos que, en esta rutina, la proporción de ejercicios destinados a trabajar tren superior es de 0,5. Mientras que la proporción de ejercicios en los que se trabaja tren inferior es de 0,3, y otro tipo de ejercicios representan una proporción de 0,2 en la rutina.

Otra de las medidas resumen es el porcentaje. Este puede pensarse de la misma manera que una proporción: como una fracción o una parte de un todo mayor, que la contiene. La diferencia está en la forma en la que se van a expresar esas porciones: si las vamos a expresar

como porcentajes, entonces nuestra escala irá desde el 0% hasta el 100%. Lo que en la proporción era 1, en los porcentajes es el 100%: la totalidad.

Es importante que, cuando trabajamos con porcentajes, no olvidemos indicarlo en la expresión con el símbolo “%”: no es lo mismo decir 50, que 50%.

Siguiendo con el ejemplo de la rutina del gimnasio, podríamos decir que el 50% de los ejercicios trabajan los grupos musculares del tren superior, o que el 30% son ejercicios para tren inferior, y que el 20% son trabajos de otro tipo. Para dar con el porcentaje, como verán, solamente debo multiplicar la proporción por 100, y agregar el símbolo “%”.

La tercera medida resumen que vamos a presentar es la razón o relación. Técnicamente, podemos definirla como el cociente entre dos categorías de una misma variable. Se expresa generalmente en términos de “tantos de una categoría por cada tantos de otra”. Se utiliza comúnmente al trabajar con variables cualitativas. Veamos un ejemplo: de un total de 240 socias/os de un Club, se observa que 160 de ellas/os abonan la cuota por débito automático y 80 lo hacen manualmente. Si quisiera conocer cuál es la relación entre ambos grupos debo proceder dividiendo (para calcular el cociente) 160 por 80: $160/80 = 2$. Entonces, puedo afirmar que por cada 2 socias/os que abonan la cuota por débito automático, 1 lo hace de forma manual. O, también: de cada 3 socias/os del club, 2 abonan a través de débito automático y 1 de manera manual. Esta sería la relación o razón entre quienes abonan de una forma y otra.

La última medida resumen con la que trabajamos es la tasa: es un índice que mide la magnitud de cambio. Es un tipo especial de razón o de proporción. Establece una relación entre un subgrupo y la población total, durante un período de tiempo, por un número alto de personas (expresado en múltiplos de 10). Encontramos algunos ejemplos en datos sociodemográficos, que se expresan en relación con grandes cantidades, como la tasa de mortalidad al nacer, tasa de empleo y desempleo, o en las tasas que indican la prevalencia de las enfermedades en la población.

Balance del capítulo

Como dijimos en capítulos anteriores: el dato es una construcción. Al introducirnos en las distintas herramientas y estrategias que nos sirven para organizar y presentar esos datos, se vuelve evidente que la forma en que optemos por llevar adelante esta tarea será clave en dicho proceso de construcción. Ahora bien, ¿esto indica forzar los datos? De ninguna manera. Como vimos, según la bibliografía con la que trabajamos, hay cierto consenso en torno a qué tipo de gráfico es conveniente usar según el tipo de variable y su nivel de medición. No obstante, es importante que tengamos esto presente no solamente en tanto gestoras/es de datos, es decir, no solo como personas que trabajamos con ellos y los producimos, sino también como sujetos sociales, como ciudadanas/os expuestos permanentemente a la circulación de estadísticas en medios de comunicación, redes sociales, documentos, etcétera. De manera que, nuestra recomendación, consiste en que cada vez que esté frente a un dato se pregunten ¿qué hay

detrás de él? ¿cómo eligen comunicarlo quienes lo difunden? ¿por qué optan por ese gráfico?
¿Utilizan una proporción, una tasa, una razón?

Por otro lado, en este capítulo vimos cómo llegamos a condensar en una simple expresión un dato que en sí mismo está cargado de mucho significado. Por ejemplo: para poder decir que, durante el mes de enero de 2025, en Argentina se produjo un femicidio cada 28 horas, es necesario llevar adelante un procedimiento de relevamiento y de cálculos que permiten resumir, condensar, toda esa información es esa breve premisa u oración.

Justamente, los distintos elementos que vimos en este capítulo habilitan una forma resumida, concisa, y gráfica de representar esos datos y, en ocasiones, ponerlos en relación entre sí. También, habrán advertido que hay distintas formas de decir lo mismo: una proporción de 0,5 es equivalente al 50%, al igual que decir “la mitad”. El objetivo es que se armen con estas herramientas y las utilicen diferencialmente para realizar sus propias expresiones y elaborar sus propios elementos de trabajo en sus clases, entrenamientos, rendimientos, performances deportivas, estudios vinculados a la salud, etcétera.

Bibliografía

- Batanero, C. (2002). “Los retos de la cultura estadística”, Conferencia Inaugural de las Jornadas Interamericanas de Enseñanza de la Estadística. Buenos Aires, Argentina. Disponible en: <https://www.ugr.es/~batanero/pages/ARTICULOS/CULTURA.pdf>
- Bologna, E. y Morales, M. (2011) “Presentación: ¿Estadística en psicología y educación?”, en Estadística para psicología y educación. Editorial Brujas.
- Cortada de Kohan, Nuria (1994) Diseño Estadístico. Editorial Eudeba
- Grima Cintas, P. (2011) *La certeza absoluta y otras ficciones los secretos de la estadística*. Editorial RBA.
- Maureira Cid, Fernando (2015) Estadística básica para educación física, Editorial Académica Española.

CAPÍTULO 4

Medidas estadísticas descriptivas de posición, centrales y no centrales

Lucía Girardi y Rocío Belén Villafañe

Una vez que se han reunido y ordenado los datos en una tabla de frecuencias la cual, como se mencionó en el capítulo anterior, facilita la lectura de los datos recolectados, se puede profundizar aún más el análisis estadístico. Para poder comparar distintas tablas de frecuencias es importante conocer las propiedades fundamentales que las caracterizan. Al reducir el conjunto de observaciones a un pequeño grupo de características básicas, no sólo se facilita la comparación de las distribuciones, sino que da lugar a interpretaciones que sirven para la comprensión de los fenómenos a los que se ha aplicado la estadística.

Como ya sabemos, dentro de las tablas de frecuencias se pueden observar los distintos valores o categorías que definen a una variable. Se puede caracterizar una distribución de una variable según sea el criterio que se utiliza para agrupar los valores o categorías de las mismas. Esto da lugar a la existencia de distintos tipos de medidas:

- **Medidas de tendencia central:** caracterizan la distribución usando un valor que tiende a ubicarse en el centro de datos.
- **Medidas de posición no central / medidas de orden:** caracterizan la distribución utilizando valores que la dividen en un determinado número de partes iguales.
- **Medidas de variabilidad:** caracterizan la distribución según el nivel de heterogeneidad o dispersión de los datos.

En este capítulo nos abocaremos al desarrollo de las medidas de tendencia central y las medidas de posición no central siendo las medidas de variabilidad objeto del siguiente capítulo.

Medidas de tendencia central

Las medidas de tendencia central se refieren a los valores o categorías de la variable que suelen estar en el centro de la distribución. Las más utilizadas (no son las únicas) son la **moda o modo (Mo)**, la **mediana (Md)**¹⁰ y la **media aritmética ($\bar{x}$)**.

¹⁰ Según la bibliografía también puede ser encontrada con las siguientes simbologías: Ma; Mdn; Me.

Moda o Modo (Mo)

Es el valor o categoría de una variable que registra mayor frecuencia en un conjunto de observaciones, es decir, la moda es el valor más común de la distribución. En la vida cotidiana hacemos referencia a la moda en el mismo sentido en que se expresa esta medida estadística, es decir, si pensamos en las zapatillas más vendidas o en el deporte más elegido en un momento dado, estaremos haciendo referencia a la moda (las zapatillas “de moda” o el deporte “de moda”).

La moda puede ser determinada en una tabla de frecuencias identificando, a través de la columna de frecuencias absolutas, la frecuencia máxima y observando a qué categoría o valor de la variable corresponde esa frecuencia. Dicha categoría o valor de la variable es la moda. La moda o modo se refiere siempre a categoría/s de la variable, no se refiere a la frecuencia, esta última solo se utiliza para poder identificarla.

La moda puede no existir; incluso es posible que no sea única, es decir que haya más de una moda en una distribución. Las distribuciones que contienen una sola moda se llaman *unimodales*, las que contienen dos modas se denominan *bimodales* y las que tienen más de dos modas se denominan *multimodales*.

La moda tiene la virtud de ser fácilmente reconocible por simple observación, por lo que se utiliza como el índice más rápido y directo para determinar la posición de una distribución. Además, la moda, no se ve afectada por los valores extremos de una distribución. Sin embargo, exige un gran número de observaciones, y tiene el inconveniente de no ser necesariamente única o de no ser calculable si todos los valores numéricos son diferentes.

Esta medida estadística puede ser aplicada para el análisis de variables de *cualquier nivel de medición*.

¡IMPORTANTE!

La moda o modo se refiere siempre a categoría/s de la variable, no se refiere a la frecuencia, esta última solo se utiliza para poder identificarla.

Veamos la determinación de la moda en un ejemplo:

Se realiza una investigación para saber cuál o cuáles son las capacidades condicionales más incluidas en las planificaciones anuales de Educación Física (EF) de las escuelas de la ciudad de La Plata durante el año 2024. Con la información obtenida se pudo confeccionar la siguiente tabla de frecuencias:

Tabla 4.1*Capacidades condicionales incluidas en las planificaciones anuales de Educación Física*

Capacidades condicionales incluidas en las planificaciones anuales de EF	Frecuencia absoluta	Frecuencia relativa en porcentaje (%)
Fuerza	25	34%
Resistencia	22	30%
Velocidad	10	13%
Flexibilidad	6	8%
Coordinación	11	15%
Total	74	100%

Nota. Fuente: elaboración propia

En la tabla de frecuencias que antecede podemos identificar la variable (capacidades condicionales incluidas en las planificaciones anuales de Educación Física) con sus categorías (fuerza, resistencia, velocidad, flexibilidad, coordinación), las frecuencias relativas y las absolutas, que nos indican, como ya hemos visto, la cantidad de veces que se repite cada categoría de la variable. La mayor frecuencia en esta tabla se corresponde con el número 25, ahora bien, 25 no es la moda, la categoría modal en este caso es “fuerza”; es el valor de la variable al cual corresponde esa frecuencia.

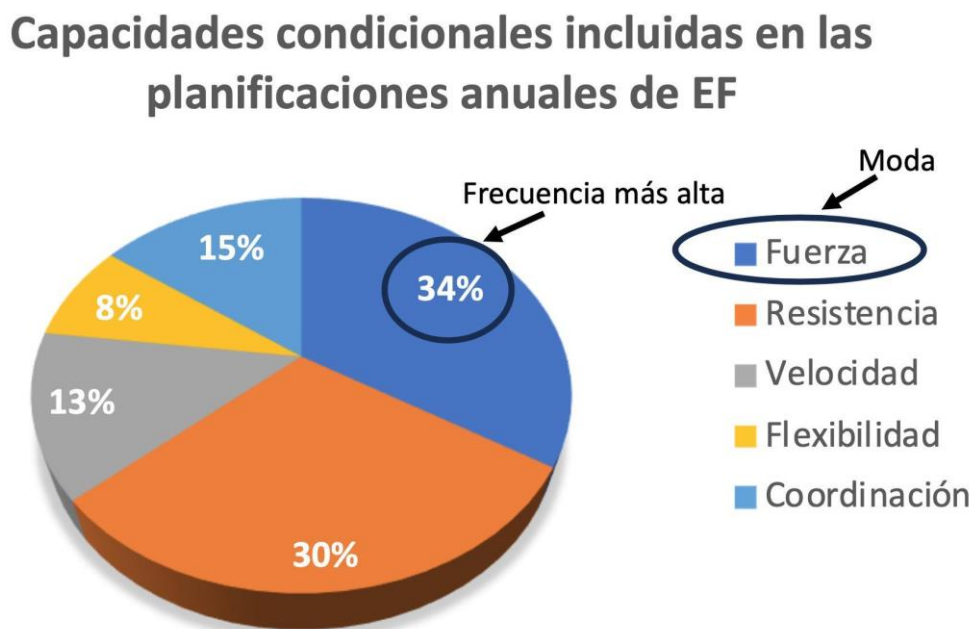
Observamos en esta tabla que la cantidad de planificaciones que se analizaron para esta distribución fue de 74 y que, de ese total de 74 planificaciones, la fuerza aparece en 25 de ellas como una de las capacidades a trabajar en las clases de Educación Física, siendo ésta la categoría modal (la moda) de la variable por ser la capacidad condicional que más se repite entre las planificaciones. La resistencia fue contabilizada en 22 de ellas, la coordinación en 11, mientras que la velocidad y la flexibilidad son las categorías de la variable que aparecieron con menor frecuencia, la primera apareció en 10 planificaciones y la flexibilidad solo en 6.

En el caso de este ejemplo, se trata de una distribución unimodal dado que hay una sola una moda, un solo valor cuya frecuencia es la más alta.

Si representamos gráficamente los datos de la tabla de frecuencias podremos identificar, a través de la observación de las frecuencias relativas porcentuales, la moda de la distribución. En el siguiente gráfico de sectores se representan las capacidades condicionales registradas en las 74 planificaciones relevadas. Cada una de las categorías abarca un sector o porción definido a partir del valor de su frecuencia porcentual. Se observa así que es la “fuerza” la capacidad condicional que constituye la moda de este conjunto de datos, ya que cuenta con la frecuencia más alta, un 34%.

Figura 4.1

Gráfico capacidades condicionales incluidas en las planificaciones anuales de Educación Física



Nota. Fuente: Elaboración propia

Mediana (Md)

La mediana se define como aquel valor de la variable por encima y por debajo del cual hay la misma cantidad de observaciones. Es un punto que divide en dos partes iguales a una distribución, es el valor del medio en un conjunto ordenado de datos.

Si por ejemplo tenemos estas 5 edades: 12, 15, 17, 22, 30.

Observamos que la edad 17 es la que se ubica en el medio de la distribución, dejando 2 edades por encima y 2 edades por debajo, por lo tanto, la mediana en esta distribución será: 17.

La mediana es un valor de posición que no presupone conocimiento de la distancia. No la afectan los valores extremos porque le interesa la cantidad de valores de la distribución y no cuánto valen estos. La mediana puede utilizarse con variables de los niveles de medición ordinal, intervalar y de cocientes.

El cálculo de la mediana depende de cómo estén organizados los datos, es decir, si se encuentran sin agrupar, en una tabla de frecuencias simple, o agrupados en intervalos.

En cualquier caso, el cálculo de la mediana se realiza en dos pasos: primero se establece el lugar o la ubicación que ocupa la mediana y luego se establece el valor de la misma, que será la categoría que se encuentre en esa posición del medio.

Cálculo de la mediana (Md) para datos sin agrupar:

1. Calcular la ubicación de la Md : una vez que los datos han sido ordenados, se calcula la posición que ocupa el valor central con la siguiente fórmula:

$$Md = \frac{n+1}{2} \text{ en donde: } n \text{ es la cantidad total de casos}$$

Usaremos esta fórmula siempre que $n < 30$ casos.

Cuando $n \geq 30$ casos, optamos por la siguiente fórmula: $Md = \frac{n}{2}$

2. Determinar el valor de la Md : una vez calculada la posición donde se ubica la mediana se pueden presentar dos situaciones dependiendo de si la cantidad de datos en una distribución es par o impar.

- Para una cantidad total de datos impares: se toma el valor central de un conjunto ordenado de datos. Por ejemplo, para los datos 12, 15, 17, 22, 30 la mediana es 17. Aplicando la fórmula: $Md(\text{posición}) = \frac{5+1}{2} = 3$

Este primer paso me indica que la mediana será el valor que ocupa la 3º posición. En nuestro ejemplo, la tercera posición la ocupa el 17, por lo que el valor de la Md es 17.

- Para una cantidad total de datos pares: se toma el par central de los valores y se calcula el promedio. Por ejemplo: para los datos 12, 15, 17, 18, 22, 30, se toman los datos centrales 17 y 18 y se promedian, siendo la Mediana 17,5.

$$\text{Aplicando la fórmula: } Md(\text{posición}) = \frac{6+1}{2} = 3,5$$

Este primer paso me indica que la mediana será el valor que se encuentra entre la 3º y la 4º posición. En nuestro ejemplo, la tercera posición la ocupa el 17 y la cuarta posición la ocupa el 18, al promediar ambos valores obtendremos el valor de la mediana. $Md = \frac{17+18}{2} = 17,5$

Si los datos están organizados en una tabla de frecuencias simple el procedimiento para calcular la mediana también sigue los dos pasos mencionados arriba. El primer paso consiste en calcular la ubicación de la Md y el segundo paso consiste en determinar el valor de la Md .

1. Calcular la ubicación de la Md : una vez que los datos han sido ordenados en una tabla de frecuencias, se calcula la posición que ocupa el valor central con la siguiente fórmula: $Md = \frac{n}{2}$
2. Determinar el valor de la Md : una vez calculada la posición donde se ubica la mediana nos dirigimos a la columna de la frecuencia acumulada para identificar cuál es el valor de la variable que contiene el caso que ocupa la ubicación calculada.

Veamos el cálculo de la mediana en un ejemplo:

La siguiente tabla muestra la cantidad de arqueras en la categoría menores que tienen los equipos de Handball de la ciudad de La Plata 2024.

Tabla 4.2

Cantidad de arqueras

Cantidad de arqueras (x)	Frecuencia absoluta (f)	Frecuencia absoluta acumulada (fa)	Frecuencia relativa en porcentaje (%)	Frecuencia relativa en porcentaje acumulada (%a)
1	4	4	25%	25%
2	7	11	44%	69%
3	4	15	25%	94%
4	1	16	6%	100%
Total	16		100%	

Nota. Fuente: elaboración propia

1. Calculamos la ubicación de la Md: $Md(\text{posición}) = \frac{16}{2} = 8$. Indica que el valor central se encuentra en el octavo lugar de la serie de datos.
2. Determinamos el valor de la Md: observamos en la columna de la frecuencia acumulada que el octavo lugar se encuentra en la frecuencia acumulada 11. Por lo que el valor de la Md es 2. en nuestro ejemplo: Md = 2 arqueras.

Esto significa que el 50% de los equipos de Handball de la ciudad de La Plata, el año pasado, en la categoría menores, tienen 2 arqueras o menos. Debido a que en nuestro relevamiento por debajo del valor 2 arqueras solo existe la categoría 1 arquera, el 50% de los equipos que menos arqueras tienen, cuentan con 1 o 2 arqueras. Del 50% restante, los equipos por encima de la mediana, tienen 2 o más arqueras, en nuestro ejemplo 3 o 4 arqueras.

Esta medida también puede ser identificada en la tabla de frecuencias a través de la columna de los porcentajes acumulados. Dado que la mediana, es el valor que divide a la distribución en dos partes iguales, tenemos que identificar entre los porcentajes acumulados dónde se ubica el 50% de los casos. En el ejemplo, observamos que el 50% está incluido en el 69% que está en la segunda fila de la tabla siendo entonces Md= 2 arqueras.

Media aritmética ($\bar{X}$)

La media aritmética es la más usual de las medidas de tendencia central y se puede utilizar con variables de tipo *cuantitativas*, variables de *niveles de medición intervalar y de cocientes*.

La media aritmética es la suma de un conjunto de valores dividido por la cantidad total de las observaciones y representa el **promedio** de los valores de una distribución. A diferencia de la mediana, la media aritmética, es muy sensible y se ve afectada por los valores extremos de las distribuciones. Esto representa una desventaja porque la media es corrida del centro de la distribución por el valor extremo generando un sesgo. La media también se simboliza **M**.

Para calcular la media aritmética para datos sin agrupar utilizamos la siguiente fórmula:

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

en donde:

$\sum$ es la sumatoria

x es el valor de cada una de las observaciones registradas para esa variable y

n es el total de los casos

Retomando los casos de nuestro ejemplo anterior, si tenemos 5 edades y queremos calcular la media aritmética:

12, 15, 17, 22, 30 ($n=5$)

$$\bar{X} = \frac{12 + 15 + 17 + 22 + 30}{5} = \frac{96}{5} = 19,2 \text{ años}$$

Cuando los datos están organizados en una tabla de frecuencias simple, el cálculo de la media debe hacerse a través de la fórmula:

$$\bar{X} = \frac{\sum (x \cdot f)}{n}$$

en donde:

$x \cdot f$ resulta de multiplicar cada valor de x con la frecuencia absoluta de ese valor

Veamos el cálculo de la media aritmética en un ejemplo:

Se registró el número de goles convertidos por las jugadoras de hockey de los Juegos Universitarios Argentinos (JUAR) en la edición 2023. Los resultados se muestran en la tabla de frecuencias:

Tabla 4.3*Cantidad de goles*

Cantidad de goles x	Frecuencia absoluta f	$x * f$
0	17	0
1	20	20
2	11	22
3	8	24
4	6	24
5	5	25
Total	67	132

Nota. Fuente: elaboración propia

En esta tabla de frecuencias:

x representa a la variable medida: cantidad de goles

f representa a la frecuencia absoluta: la cantidad de jugadoras que convirtió x número de goles

Para calcular la media de goles convertidos incorporamos una columna a la tabla $x * f$, es decir, multiplicamos la cantidad de goles por la cantidad de jugadoras que hicieron ese número de goles.

Aplicamos la fórmula: $\bar{X} = \frac{\sum (x*f)}{n}$

$$\bar{X} = \frac{132}{67} = 1.9701$$

A partir del cálculo realizado concluimos que el promedio de goles convertidos por las jugadoras de hockey en los JUAR 2023 es de 2. Media = 2 goles.

Una de las propiedades que tiene la media aritmética es que es susceptible de operaciones algebraicas, es decir, los promedios pueden sumarse, restarse, etc. Por lo tanto, si conocemos la media de varios grupos de observaciones, podemos calcular la media del conjunto total. Incluso si los grupos tienen diferente tamaño, es decir, si no tienen la misma cantidad de casos.

La media total (MT) se utiliza para ponderar el valor de la media, de dos o más grupos y se calcula de la siguiente manera:

$$Media\ Total = \frac{(\bar{X}_{grupo\ A} * n_{grupo\ A}) + (\bar{X}_{grupo\ B} * n_{grupo\ B})}{n_{grupo\ A} + n_{grupo\ B}}$$

en donde:

$\bar{X}$ es la media de cada grupo en estudio

n es la cantidad de casos de cada grupo en estudio

Tabla 4.4

Medidas de Tendencia Central

Medidas de Tendencia Central	Simbolización ¹¹	Interpretación	Nivel de Medición
Moda o modo	Mo	Es el valor de la variable que aparece con mayor frecuencia	- Todos los niveles de medición
Mediana	Md	Es el valor de la variable que divide en dos partes iguales a un grupo de observaciones	- Nivel de medición ordinal - Nivel de medición intervalar - Nivel de medición de razón
Media	$\bar{X}$	Representa el promedio de las observaciones	- Nivel de medición intervalar - Nivel de medición de razón
Media total	MT	Representa el promedio de dos o más grupos de observaciones	- Nivel de medición intervalar - Nivel de medición de razón

Nota. Fuente: elaboración propia

Medidas de posición no centrales: cuartiles, deciles y percentiles. Diagrama de cajas

Para caracterizar a las distribuciones contamos también con otras medidas de posición, estas son los cuartiles, deciles y percentiles. Son medidas de posición no centrales que se utilizan para localizar en una sucesión ordenada de datos, los puntos que dividen las distribuciones de frecuencias en grupos de igual número de datos cada uno.

¹¹ Se exponen en el cuadro las simbolizaciones utilizadas por la cátedra que no son las únicas, sino que varían según el autor (ver nota de pie¹)

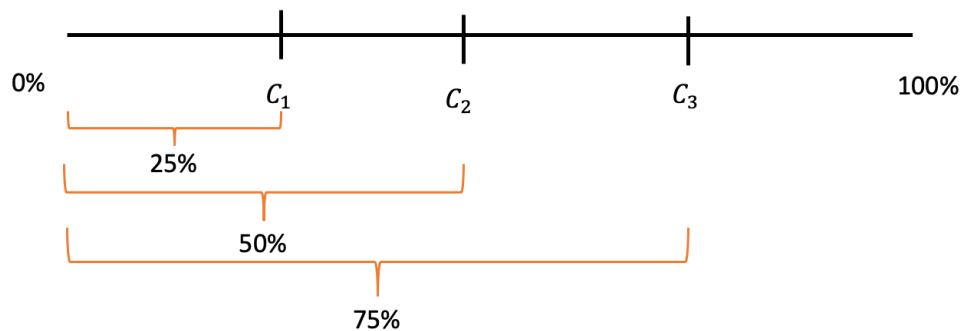
Cuartiles (C o Q)

Los cuartiles son tres y dividen a la distribución en cuatro partes iguales, representando cada una de estas partes el 25% del total de casos.

Si un conjunto de datos se ordena por orden de magnitud, el valor del medio que divide el conjunto en dos partes iguales sabemos que es la mediana. Ampliando esta idea, podemos dividir el conjunto en cuatro partes iguales y tendremos los tres cuartiles C_1 , C_2 , C_3 . El primer cuartil (C_1) es el valor antes del cual quedará una cuarta parte (25%) de todas las observaciones; el segundo cuartil (C_2) es idéntico a la mediana (50%) y el tercer cuartil (C_3) es el valor antes del cual quedarán tres cuartas partes (75%) de las observaciones.

Figura 4.2

Cuartiles

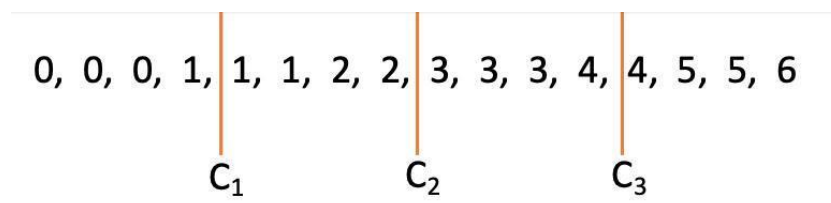


Nota. Fuente: elaboración propia

Los cuartiles dividen en 4 partes iguales a un conjunto de datos. En el siguiente grupo de datos se encuentran señalados los cuartiles:

Figura 4.3

Cuartiles



Nota. Fuente: elaboración propia

Deciles (D)

Los deciles son nueve medidas y dividen en 10 partes iguales un conjunto de datos, es decir, dividen al conjunto en porciones que representan un 10% de los casos cada uno. El decil

1 (D_1) deja por debajo de su valor el 10% de los casos y por encima de su valor el 90% restante. El decil 2 (D_2) deja por debajo de su valor el 20% de los casos y por encima de su valor el 80% restante. Así sucesivamente hasta el decil 9 (D_9) que deja por debajo de su valor el 90% de los casos y por encima el 10% restante.

Percentiles (P)

En una distribución hay 99 percentiles que dividen al conjunto de datos (ordenados de menor a mayor) en cien partes iguales. El valor por debajo del cual queda un porcentaje determinado de casos es un percentil, teniendo un valor que oscila entre 1 y 100. Si se divide un conjunto de observaciones en 100 partes iguales se obtendrán 99 percentiles representados por $P_1, P_2, P_3, P_4, P_5 \dots P_{99}$. Así, cuando se habla del percentil 15 se quiere expresar que es el valor de la variable que deja el 15% de los datos a su izquierda y el 85 % de los mismos a su derecha o lo que es lo mismo decir que es el valor de la variable que deja al 15 % de los datos por debajo de él y el 85% por encima. Entonces, por ejemplo, el percentil 80 es el valor de la variable que es igual o deja por debajo de sí, al 80% del total de los datos.

Los tres cuartiles (C_1, C_2 y C_3) coinciden con los percentiles 25, 50 y 75. Lo mismo ocurre con los deciles siendo por ejemplo $D_1=P_{10}$. De lo anteriormente dicho se desprende que: $M_d = Q_2 = D_5 = P_{50}$

Para el cálculo de cualquier cuartil, decil o percentil son válidas las consideraciones que se hicieron para el cálculo de la mediana. En cualquier caso, el cálculo de las medidas de orden se realiza en dos pasos: primero se establece el lugar o la ubicación que ocupa el cuartil, decil o percentil y luego, se establece el valor de la medida. Los valores que asumen los cuartiles, deciles o percentiles siempre son un valor de la variable.

El primer paso que haremos consiste en calcular la posición de los cuartiles, deciles o percentiles para lo que tendremos en cuenta, a) en cuántas partes queremos dividir el grupo de datos 4, 10 o 100 y, b) el número de cuartil, decil o percentil.

Para calcular la posición de los cuartiles $C_h = \frac{h * n}{4}$	donde h puede valer: 1, 2 o 3
Para calcular la posición de los deciles $D_h = \frac{h * n}{10}$	donde h puede valer: 1, 2, 3, 4, 5, 6, 7, 8 o 9
Para calcular la posición de los percentiles $P_h = \frac{h * n}{100}$	donde h puede valer: 1, 2, 3, ... 22, 23, 24 ... o 99

Una vez identificada la ubicación del cuartil, decil o percentil determinaremos su valor.

Si bien las medidas de posición no centrales no suelen calcularse para grupos pequeños de datos, solo con el objetivo de simplificar la explicación, mostraremos en un grupo reducido de datos cómo identificar estas medidas de orden.

Buscaremos calcular los cuartiles 1 y 3 con el siguiente grupo de observaciones:

0, 0, 0, 1, 1, 1, 2, 2, 2, 3, 3, 3, 4, 4, 5, 5, 6

donde $n=17$.

Para el cálculo del cuartil 1 tendremos que hallar la posición del cuartil 1:

$$C_{1(\text{posición})} = \frac{1 \cdot 17}{4} = 4,25$$

Este cálculo indica que el cuartil 1 se ubica entre el cuarto y quinto lugar, por lo que habrá que promediar los valores que ocupan dichas posiciones. En nuestro ejemplo, en el cuarto lugar está el valor 1 y en el quinto lugar se encuentra otro 1, por lo que el valor del cuartil 1 será igual a 1. $C_1 = 1$

Emplearemos el mismo procedimiento para el cálculo del cuartil 3: como primer paso buscaremos la posición en la que se ubica el cuartil 3: $C_{3(\text{posición})} = \frac{3 \cdot 17}{4} = 12,75$

Este cálculo muestra que el cuartil 3 se ubica entre la posición 12 y 13. En nuestro ejemplo, en la ubicación 12 hallamos un 4 y en la posición 13 también un 4. El promedio de los valores que ocupen las posiciones 12 y 13 será el valor del cuartil 3, por lo que el cuartil 3 es igual a 4. $C_3 = 4$

Retomemos nuestro ejemplo acerca de los goles convertidos por las jugadoras de hockey en los JUAR 2023. La tabla de frecuencias que organiza los datos relevados nos muestra las frecuencias absolutas y las frecuencias relativas porcentuales. Si observamos la columna de las frecuencias porcentuales acumuladas podemos visualizar el porcentaje de casos que se acumulan para identificar el 25%, el 50% y el 75%. Estos porcentajes se asocian con los casos que dejan por debajo de su valor el cuartil 1 (25%), el cuartil 2 (50%) y el cuartil 3 (75%) respectivamente.

Figura 4.4*Cuartiles en tabla de frecuencias*

	Cantidad de goles (x)	Frecuencia absoluta (f)	Frecuencia relativa en porcentaje (%)	Frecuencia relativa en porcentaje acumulada (%a)	
C ₁ →	0	20	25%	25%	← Contiene el 25% de los datos
C ₂ →	1	23	28%	53%	← Contiene el 50% de los datos
	2	14	17%	70%	
C ₃ →	3	10	12%	83%	← Contiene el 75% de los datos
	4	8	10%	93%	
	5	6	7%	100%	
		81	100%		

Nota. Fuente: elaboración propia

La tabla de frecuencias nos permite identificar cualquier medida de posición no central siguiendo el mismo procedimiento que se utilizó para calcular los cuartiles.

Si analizamos nuevamente la tabla de frecuencias, observaremos que la columna de las frecuencias relativas en porcentaje acumulado nos permite identificar cómo se distribuyen las porciones de casos.

Tabla 4.5*Tabla de frecuencias de cantidad de goles*

Cantidad de goles (x)	Frecuencia absoluta (f)	Frecuencia relativa en porcentaje (%)	Frecuencia relativa en porcentaje acumulada (%a)
0	20	25%	25%
1	23	28%	53%
2	14	17%	70%
3	10	12%	83%
4	8	10%	93%
5	6	7%	100%
Total	81	100%	

Nota. Fuente: elaboración propia

Por ejemplo, sabemos que el decil 2 deja por debajo de su valor el 20% de los casos y que el decil 7 deja por debajo de su valor al 70% de los casos. Al observar la columna de los porcentajes acumulados podemos encontrar que el 25% de las jugadoras registró un valor 0 en la medición de nuestra variable, es decir, hizo 0 número de goles o no hizo goles. El 20% de las jugadoras está incluido en ese primer porcentaje de casos que no registró goles. Por lo tanto, el valor del decil 2 es igual a 0 ($D_2 = 0$). El decil 7 puede ser identificado de la misma forma, al observar la columna de los porcentajes acumulados, podemos encontrar que el 70% de las jugadoras marcó hasta 2 goles durante los partidos de los JUAR 2023. Decimos que las jugadoras marcaron “hasta 2 goles” porque estamos analizando los porcentajes acumulados, es decir, el 70% de las jugadoras incluye a las que no marcaron goles ($x = 0$), a las que marcaron 1 gol ($x = 1$) y a las que marcaron 2 goles ($x = 2$). Por lo tanto, el valor del decil 7 es igual a 2 ($D_7 = 2$).

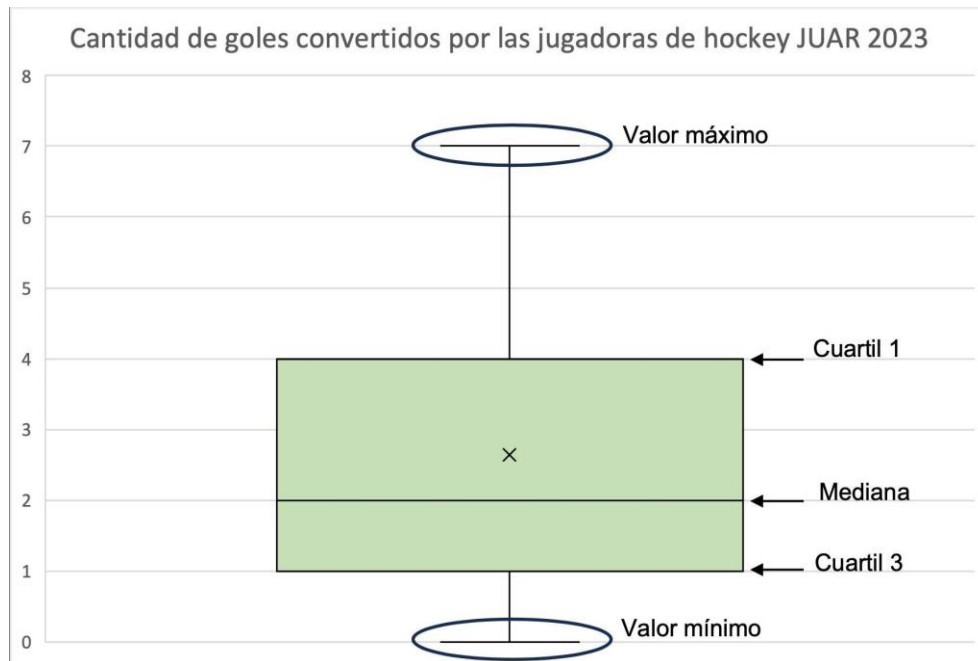
Ahora buscaremos identificar el percentil 80 (cuyo valor deja por debajo el 80% de los casos) y el percentil 93 (cuyo valor deja por debajo el 93% de los casos). Observamos nuevamente la columna de las frecuencias porcentuales acumuladas para localizar los percentiles. La tabla nos muestra que el 83% de las jugadoras marcaron hasta 3 goles o, en otras palabras, marcaron 3 goles o menos. El 83% de las jugadoras incluye a las que convirtieron 0, 1, 2 o 3 goles. El percentil 80, es el valor de la variable (el número de goles) convertido por el 80% de las jugadoras, por lo que el valor del percentil 80 es igual a 3 ($P_{80} = 3$). El percentil 93 puede ser identificado en la tabla de frecuencias a través del valor de la frecuencia porcentual acumulada 93%. Este valor porcentual nos indica que el 93% de las jugadoras marcaron hasta 4 goles durante el torneo, por lo que el valor del percentil 93 es igual a 4 ($P_{93} = 4$).

Diagrama de Cajas

Un diagrama de caja es una representación gráfica que nos permite analizar la manera en que se distribuye un conjunto de datos. El diagrama muestra los cinco números resumen: el valor mínimo y el valor máximo registrado, la mediana y el 1º y 3º cuartil.

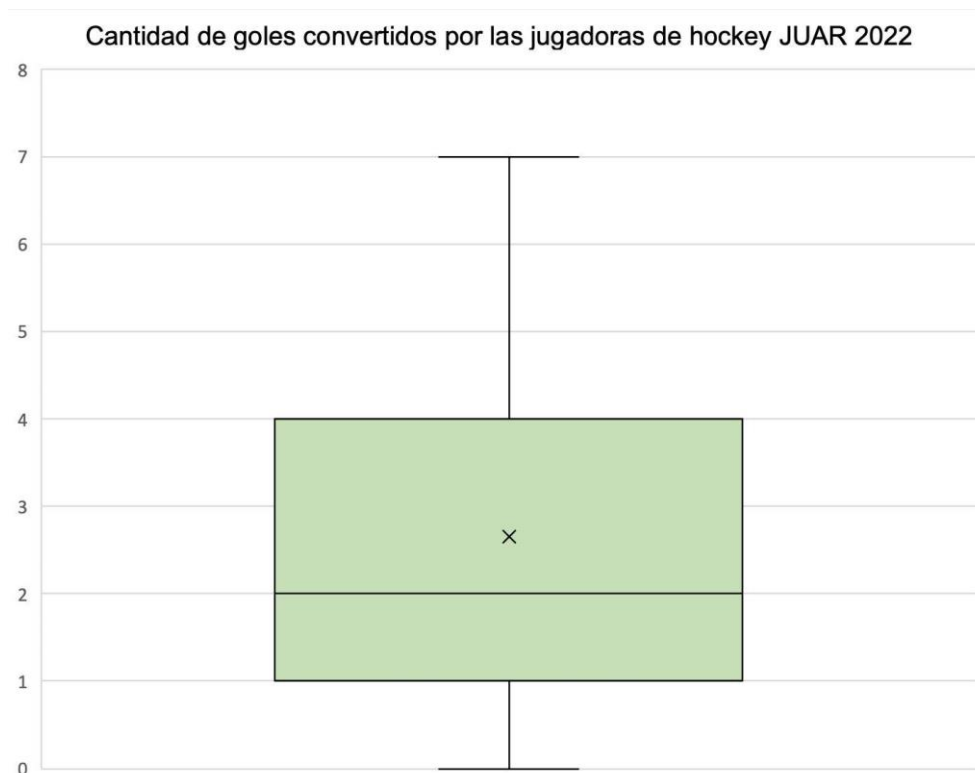
El diagrama se construye sobre un eje vertical partiendo del valor mínimo hasta el valor máximo relevado. La caja representa al 50% central de los datos debido a que sus límites se corresponden con el cuartil 1 y el cuartil 3. Las líneas que se extienden hacia abajo y hacia arriba de la caja se denominan “bigotes”. La caja se divide por una línea que refleja el valor de la mediana, esta línea deja por debajo el 50% de los casos y por encima el 50% restante. En el diagrama de caja suele representarse la media con un punto o una cruz.

En la siguiente imagen se observan los elementos mencionados:

Figura 4.5*Gráfico de cajas para cantidad de goles**Nota.* Fuente: Elaboración propia

El diagrama de cajas es una representación gráfica que se utiliza sólo para variables del nivel de medición intervalar y de razón. Aunque cuenta con menos detalles que otras representaciones gráficas, tiene mucha utilidad para establecer comparaciones entre dos o más distribuciones, incluso en el mismo gráfico.

El siguiente diagrama de cajas nos muestra la distribución de goles convertidos por las jugadoras de hockey en la edición de los JUAR 2022. El valor mínimo de goles registrados es 0 y el valor máximo registrado es 7 goles. La caja señala que el 50% central de las jugadoras marcaron entre 1 y 4 goles, ya que los extremos inferior y superior de la caja, se establecen a partir del valor de los cuartiles 1 y 3 respectivamente. La línea que atraviesa la caja nos muestra a la mediana con un valor de 2 goles. Los “bigotes” del diagrama contienen un 25% de casos cada uno, pero al observarlos en este ejemplo, vemos que los segmentos tienen distinta longitud. El segmento inferior que contiene el primer 25% de los casos, es decir, las jugadoras que menos goles hicieron, tienen una longitud que abarca de 0 a 1 gol. El segmento que contiene el 25% superior de los casos, tiene una longitud mayor que el primero debido a que están comprendidas las jugadoras que hicieron entre 4 y 7 goles. Por su parte, la media aritmética se encuentra señalada con una cruz por encima del valor de la mediana, pero sin alcanzar el valor de 3 goles.

Figura 4.6*Gráfico de cajas para cantidad de goles jugadoras de hockey**Nota.* Fuente: Elaboración propia

En una distribución simétrica los cuartiles 1 y 3 se encuentran, cada uno de ellos, aproximadamente a la misma distancia de la mediana. Cuando la distancia entre el cuartil 1 y la mediana y el cuartil 3 y la mediana no es similar o idéntico, la distribución refleja mayor o menor grado de asimetría.

En una distribución simétrica los “bigotes” del diagrama tienen aproximadamente la misma longitud. Cuando uno de los bigotes es más largo que el otro, el diagrama está evidenciando la asimetría de la distribución y un mayor grado de dispersión de las observaciones. Simetría y dispersión son conceptos que se desarrollarán más adelante en este libro.

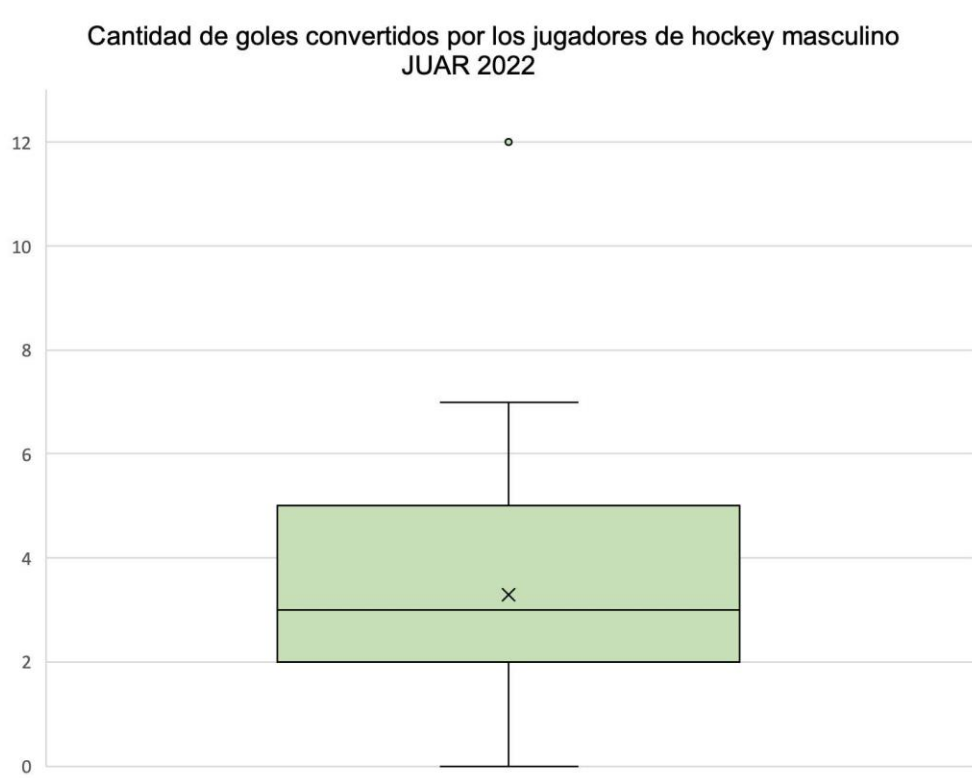
Observemos ahora el diagrama de cajas que representa los goles convertidos por los jugadores de los equipos de hockey masculino que participaron de la edición de los JUAR 2022. Podemos advertir algunas diferencias con el diagrama anterior. En primer lugar, las cajas que contienen el 50% central de los casos tienen características distintas. En el caso de los jugadores de equipos masculinos, el lado inferior de la caja refleja un valor para el cuartil 1 equivalente a 2 goles, y el lado superior muestra al cuartil 3 con un valor de 5 goles. La mediana, al igual que en la anterior distribución, no se encuentra en el centro de la caja. Si bien sabemos que a ambos lados de la mediana se encuentra un 25% de casos (dentro de la caja), cada uno de esos lados presenta mayor o menor dispersión de goles. En la distribución de los

equipos masculinos la media se encuentra en un valor más cercano a la mediana que en el diagrama anterior analizado.

Por último, observamos un punto a la altura de los 12 goles, que se encuentra por fuera de la caja y los segmentos denominados “bigotes”. Este punto representa un valor atípico en la distribución, un valor muy alejado de los extremos de la caja, que no se asemeja a la mayoría de los datos.

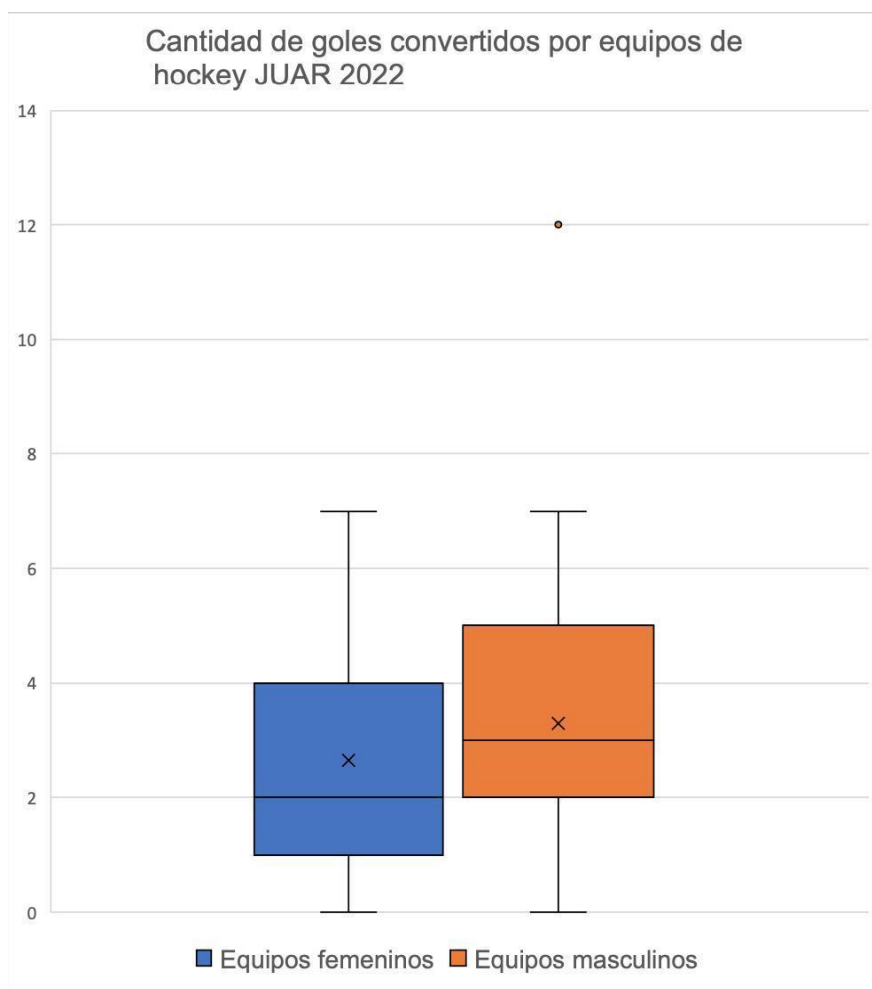
Figura 4.7

Gráfico de cajas para cantidad de goles jugadores de hockey



Nota. Fuente: Elaboración propia

Una forma sencilla de comparar estos diagramas de cajas consiste en realizar un mismo gráfico para dos o más distribuciones. A continuación, se muestran los datos analizados anteriormente sobre la cantidad de goles marcados por las jugadoras y los jugadores de los equipos de hockey en los JUAR 2022.

Figura 4.8*Gráfico de cajas para cantidad de goles jugadoras y jugadores de hockey**Nota.* Fuente: Elaboración propia

Bibliografía

- Aguirre, C.; M. F. Niño y E. Simonetti (2005) Estadística aplicada en las ciencias sociales y humanas, Universidad Nacional de Misiones-Editorial Universitaria.
- Maureira Cid, Fernando (2015) Estadística básica para educación física, Editorial Académica Española.
- Cortada de Kohan, Nuria (1994) Diseño Estadístico, Editorial Eudeba.

CAPÍTULO 5

Medidas de variabilidad y de forma

Román Fornessi y María Agustina Gandini

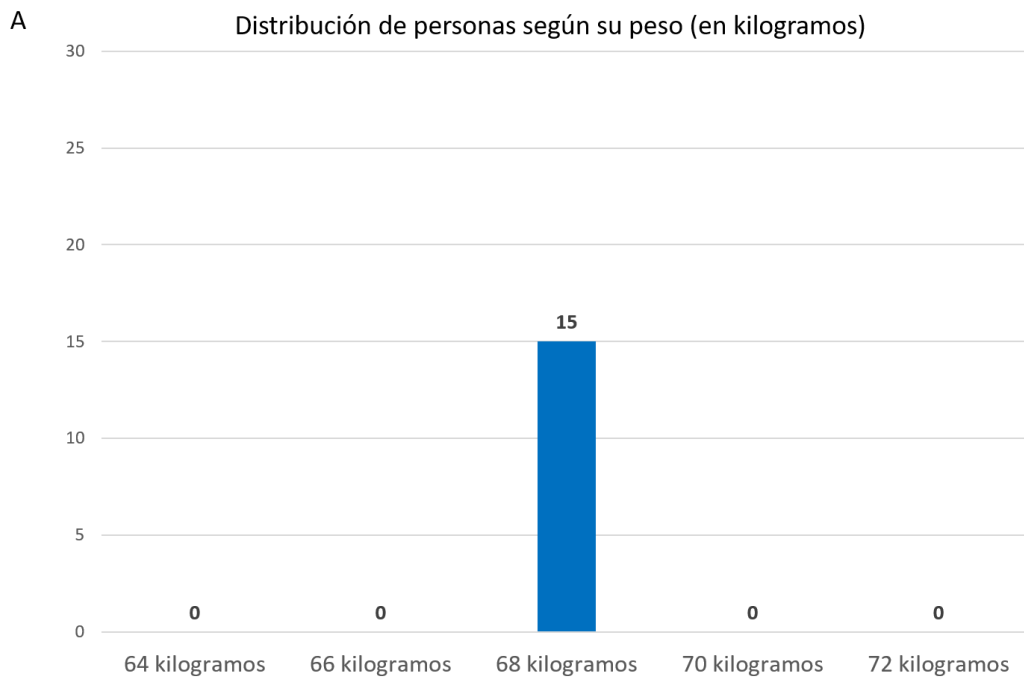
Introducción a las medidas de variabilidad

Hasta aquí hemos recorrido algunos conceptos introductorios en estadística, que nos permiten identificar a las entidades que estamos estudiando (la unidad de análisis), los atributos o características que estamos midiendo (las variables), las formas de presentación de la información (matriz de datos, tablas de frecuencias, representaciones gráficas). Luego, avanzamos hacia una batería de herramientas que nos permitieron comenzar a describir conjuntos de datos: las medidas de tendencia central y las medidas de tendencia no central.

En este capítulo, avanzaremos con otro conjunto de técnicas de análisis que, complementariamente con las recién mencionadas, permitirán complejizar y aportar más información sobre las distribuciones de datos: las medidas de variabilidad, por un lado, y herramientas que nos permitirán describir la asimetría y la curtosis de una distribución, por otro.

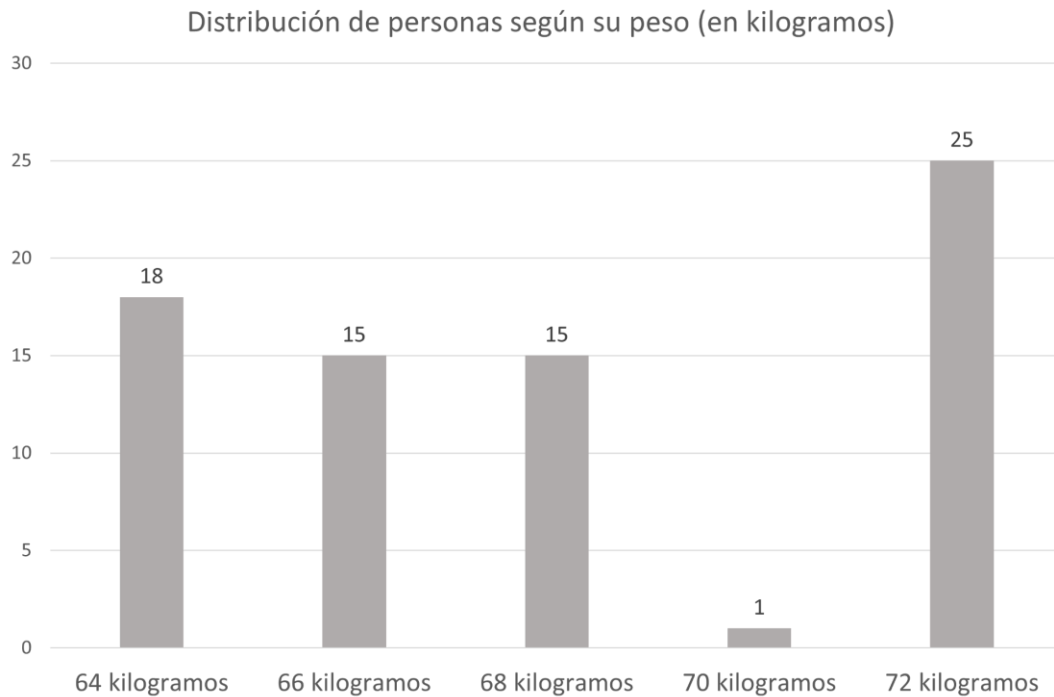
Es importante destacar que todas las medidas que estamos abordando se emplean de manera complementaria, es decir, tenemos que abordar los conjuntos de datos con todas ellas (siempre dependiendo del nivel de medición de cada variable), de manera tal que podamos enriquecer la descripción que estamos construyendo. Para ello, es necesario identificar qué tipo de información nos aporta cada grupo de medidas.

En este capítulo veremos “cómo se dispersan las puntuaciones de una variable de intervalo/razón de menor a mayor y la forma de la distribución de estas” (Ritchey, 2006, p. 136). En esta definición ya encontramos un primer elemento fundamental: las medidas que veremos en este capítulo se aplicarán a variables cuyo nivel de medición sea de intervalos o de razón. Las medidas de variabilidad nos informan el nivel de dispersión de los datos. Les proponemos un primer acercamiento para comprender el concepto de dispersión, a partir del siguiente gráfico:

Figura 5.1*Gráfico de barras de personas según peso en kg**Nota.* Fuente: elaboración propia

Supongamos que estamos frente a una distribución de la variable “peso, en kilogramos”, como nos muestra el gráfico, todas las personas que participan de esta distribución (que son 15 sujetos) muestran el mismo valor (68 kilogramos), por eso es que vemos solo una barra para el valor 68 kg. Como podemos calcular rápidamente, el promedio (la media aritmética) de esta distribución es 68 kilogramos.

Ahora veamos qué sucede con otra distribución de la misma variable y con el mismo promedio (media aritmética):

Figura 5.2*Gráfico de barras de personas según peso en kg*

Nota. El peso está definido aquí como una variable cuantitativa discreta (cantidad de kilos completos alcanzados). Fuente: elaboración propia

Como podemos observar, en esta distribución se incorporaron más personas, con distintos pesos, es decir sujetos que aportan valores distintos para la misma variable (por eso vemos que, en el gráfico, hay barras para otros valores, aparte del 68 kg).

A partir de una primera lectura rápida sobre estas dos distribuciones de datos, podemos responder la pregunta ¿en cuál de las dos hay mayor dispersión?: hay mayor dispersión en la distribución B, porque en la A los casos se concentran en un mismo valor, en cambio en el segundo gráfico, los casos tienden a asumir distintos valores.

Ahora bien, fíjense que siempre estamos hablando de una distribución más dispersa o menos dispersa en relación con otra; esto debemos tenerlo siempre muy claro: la descripción, en términos de variabilidad, de un conjunto de datos siempre se realiza en función de otro/s. No podemos decir “esta distribución es dispersa” o “este conjunto de datos presenta una alta variabilidad”, la utilidad de este análisis siempre será en términos comparativos: la distribución A presenta una variabilidad menor que la distribución B.

Vean que, si nos quedáramos con la descripción que nos permiten las medidas de tendencia central, advertiríamos que el promedio en ambos conjuntos de datos es el mismo (68 kilogramos), sin embargo, una simple lectura al gráfico nos muestra que estamos ante dos distribuciones distintas (aunque el promedio sea el mismo), entonces necesitamos aplicar otro conjunto de medidas que nos den información sobre eso en lo que se diferencian, que es,

justamente, la variabilidad o la dispersión. Por eso decíamos, al inicio del capítulo, que estas medidas hay que utilizarlas de forma complementaria entre sí.

Para reforzar la comprensión sobre el concepto de variabilidad, aportamos otro ejemplo sencillo, antes de avanzar: supongamos que la Profesora de Educación Física de tercer grado de una escuela primaria de La Plata advierte que el curso de la mañana y el de la tarde tienen desempeños distintos en la clase. Pero ella identifica esto de forma general, “le parece” que ese desempeño es distinto. Entonces, para poder construir un diagnóstico más elaborado y basado en datos, decide registrar los desempeños de cada grupo en algunos ejercicios de la clase. A continuación, se presentan los resultados para la prueba de salto al banco en 1 minuto¹²:

En esta prueba, registró la cantidad de saltos que realizaron 12 alumnos/as del turno mañana y 12 alumnos/as del turno tarde:

Turno mañana (cantidad de saltos logrados en un minuto por cada estudiante):

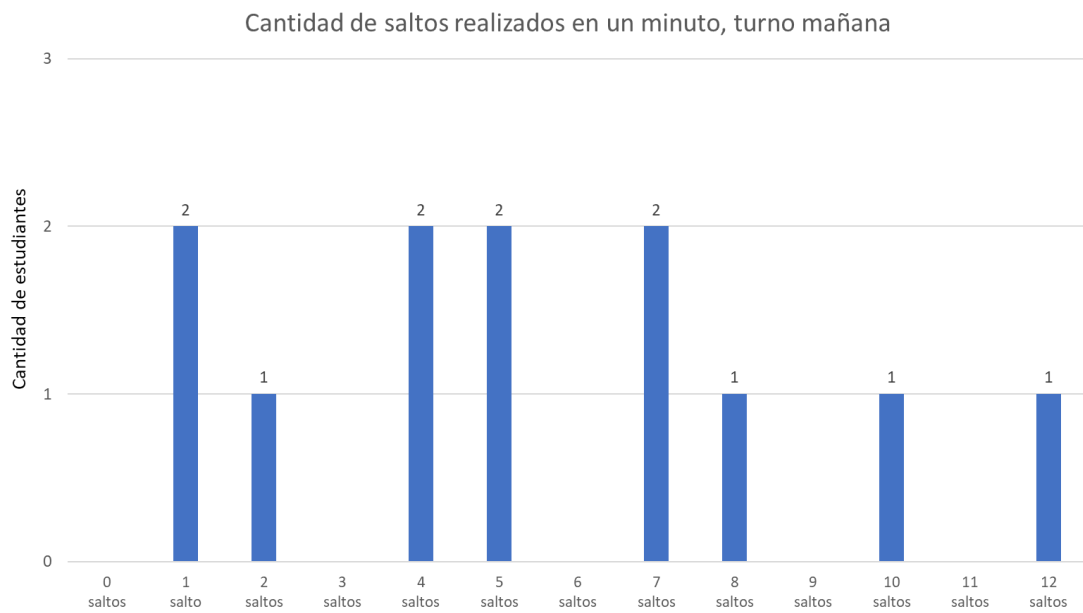
1 – 1 – 2 – 4 – 4 – 5 – 5 – 7 – 7 – 8 – 10 – 12

Turno tarde (cantidad de saltos logrados en un minuto por cada estudiante)

0 – 4 – 4 – 5 – 5 – 6 – 6 – 6 – 7 – 7 – 8 – 8

Para poder tener una primera lectura de estas distribuciones, las representamos gráficamente:

¹² Esta prueba consiste en hacer que una persona salte desde el piso hasta un banco que se encuentra a 30 cm de altura, realizando la mayor cantidad de saltos posibles, en un minuto.

Figura 5.3*Gráfico de barras de estudiantes del turno mañana según saltos realizados en un minuto**Nota.* Fuente: elaboración propia**Figura 5.4***Gráfico de barras de estudiantes del turno tarde según saltos realizados en un minuto**Nota.* Fuente: elaboración propia

En estas distribuciones sucede algo curioso: al calcular la media aritmética advertimos que el promedio de saltos realizados por cada persona, en cada turno, es el mismo: 5,5 saltos; también la cantidad total de saltos realizados por cada grupo es la misma: 66.

Sin embargo ¿cómo se comportaron ambas distribuciones en términos de dispersión o de variabilidad?

A simple vista, podemos aventurar que el grupo de la mañana muestra una variabilidad o una dispersión mayor que el grupo de la tarde, en cuanto al desempeño en esta prueba.

¿Esto quiere decir que hay un desempeño mejor o peor de uno u otro grupo? De ninguna manera. No hablamos en términos de mejor o peor (porque esto sería valorativo) sino que nos referimos a desempeños más parejos o desparejos, más homogéneos o más heterogéneos, más dispersos o menos dispersos, etcétera. Y esto lo podremos afirmar en términos comparativos, es decir que vamos a poder aseverar en base a datos estadísticos que, por ejemplo, *el grupo A presenta un desempeño más homogéneo que el grupo B*, o que *de los grupos analizados A, B, C y D, el que muestra una dispersión mayor en cuanto a la variable X es el grupo D*, etcétera.

Ahora bien, volviendo a nuestro ejemplo, ya identificamos que un grupo mostró un desempeño **más disperso** que otro, pero ¿cuán variable o heterogéneo son los desempeños de ambos grupos?

La estadística nos brinda medidas que nos van a permitir, valga la redundancia, **medir** esa dispersión: **la amplitud total, el recorrido intercuartílico, el desvío estándar, el coeficiente de variación**

Hay que tener en cuenta que las medidas de variabilidad o de dispersión se pueden calcular en variables cuyo nivel de medición es de intervalos o de razón, es decir, variables cuantitativas (ya sean continuas o discretas), porque vamos a cuantificar ese nivel de dispersión o de variabilidad.

Amplitud total

Como su nombre lo indica, la amplitud total nos brinda una primera aproximación a la dispersión de un conjunto de datos a partir de la magnitud del recorrido total de la variable, es decir, nos indica la distancia existente entre el caso que asume el menor valor y el que asume el mayor valor de la distribución. Por ejemplo, si contamos con la siguiente distribución de la variable edad: 23 – 21 – 18 – 31 – 29 – 30 – 26 – 32, la amplitud total de este conjunto de datos sería de 14 años, porque es la distancia que existe entre el menor valor (18 años) y el mayor valor (32 años) del conjunto de datos. Entonces, diremos que el recorrido total de la variable edad, en esta distribución es de 14 años. Si comparamos este dato con otra distribución cuya amplitud total sea de, por ejemplo, 7 años, tendremos un primer indicador del nivel de dispersión de ambas distribuciones: una de ellas cuenta con una amplitud total de 14 años, y la otra de 7 años.

Retomando el ejemplo anterior, que mostraba el registro de la cantidad de saltos que realizaron alumnos/as de dos cursos distintos, calculemos la amplitud total para cada grupo.

Lo primero que debemos hacer es identificar la mayor y la menor cantidad de saltos realizada en cada grupo por algún/a alumno/a. Así, advertimos que para el turno mañana, el valor más alto es de 12 saltos y el más bajo es de 1 salto. Mientras que en el turno tarde, el valor más alto es 9 saltos, y el más bajo 0 saltos.

La fórmula para el cálculo de la amplitud (A_t) es la siguiente:

$$A_t = V - v$$

Donde:

V = valor mayor de la distribución

v = valor menor de la distribución

Apliquemos la fórmula a nuestros conjuntos de datos:

Turno mañana

$$A_t = V - v = 12 - 1 = \mathbf{11 \text{ saltos}}$$

Turno tarde

$$A_t = V - v = 9 - 0 = \mathbf{9 \text{ saltos}}$$

Como vemos, la amplitud total, el recorrido total de la variable, es mayor en la primera distribución. Incluso podemos interpretar que la amplitud total de la variable, para el grupo del turno tarde, es la mitad que para el grupo del turno mañana.

Sin embargo, para un correcto análisis y una pertinente descripción de la variabilidad de una o más distribuciones, es necesario que las medidas de dispersión se apliquen en conjunto, lo que aportará una lectura integral.

Recorrido intercuartílico

El recorrido intercuartílico representa la distancia entre dos cuartiles, específicamente el cuartil 1 y el cuartil 3. A diferencia de la Amplitud Total, el recorrido intercuartílico no considera la totalidad de los datos, sino que permite identificar el 50% central de la distribución, delimitado por el Cuartil 1 y Cuartil 3. Señalamos que es una medida más estable que la Amplitud Total porque no se ve afectada por valores extremos de la variable. Se utiliza, principalmente, cuando estamos ante una distribución asimétrica.

Para calcular el recorrido intercuartílico se procede restando los valores de los cuartiles 3 y 1. Cabe aclarar que, para este cálculo, debemos tener en cuenta que los datos se encuentren ordenados de menor a mayor y, posteriormente, saber cuál es la posición en la que se ubican estos cuartiles.

Retomando el ejemplo en el apartado anterior, teníamos la siguiente distribución de la variable edad: 23 – 21 – 18 – 31 – 29 – 30 – 26 – 32.

En primer lugar, ordenamos los datos:

18 - 21 - 23 - 26 - 29 - 30 - 31 - 32

Luego, avanzamos identificando las posiciones en las que se encuentran los cuartiles 1 y 3. Para ello, aplicamos la fórmula de cuartil 1 y cuartil 3 de orden (recuerden que las trabajamos en detalle en el capítulo 4, dentro de medidas de orden o de posición).

Cuartil 3 de orden = $(3 * n) / 4$

$(3 * 8) / 4 = 24 / 4 = 6$ (el cuartil 3 es el valor que se ubica en la sexta posición, estando los valores ordenados de menor a mayor) → El valor que ocupa la sexta posición es 30 años.

Cuartil 1 de orden = $(1 * n) / 4$

$(1 * 8) / 4 = 8 / 4 = 2$ (el cuartil 1 es el valor que se ubica en la segunda posición, estando los valores ordenados de menor a mayor) → El valor que ocupa la segunda posición es 21 años.

Habiendo identificado los valores que se corresponden con el cuartil 3 y con el cuartil 1, procedemos a calcular la distancia que hay entre ellos. Esto lo realizamos a partir de la resta entre cuartil 3 y cuartil 1:

$$\text{Recorrido intercuartílico (RI)} = C3 - C1 = 30 - 21 = 9 \text{ años}$$

Diremos, entonces, que el recorrido intercuartílico de la variable edad es de 9 años. Es decir, 9 años es el rango de variación del 50% central de la distribución.

Asimismo, podemos interpretar que el 50% central de los casos (que son aquellos comprendidos entre el cuartil 1 y el cuartil 3) se concentran a lo largo de 9 años. También podemos afirmar que la mitad de los casos tiene entre 21 y 30 años.

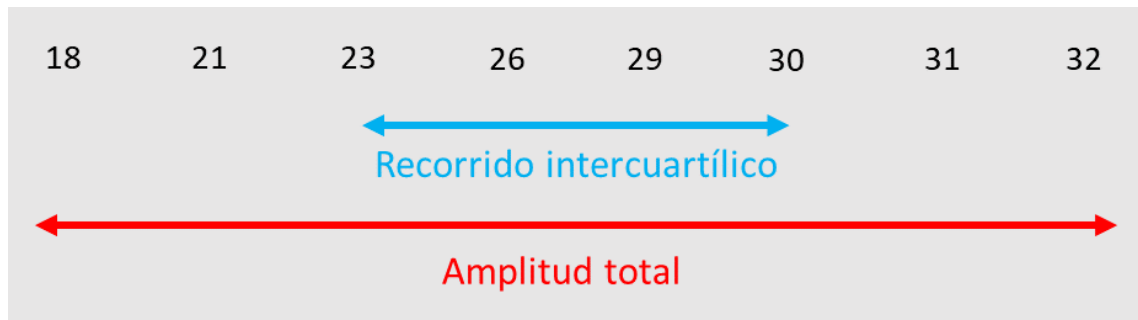
Fíjense que el cuartil 1 y el cuartil 3 guardan una relación de identidad con los percentiles 25 y 75, respectivamente.

Veamos, gráficamente, qué nos muestra cada una de estas dos medidas que vimos recientemente. A continuación, encontraremos la distribución del ejemplo con el que trabajamos:

18 21 23 26 29 30 31 32

La amplitud total nos indicará cuál es el rango, el recorrido total que realiza la variable, en esta distribución, va a considerar el 100% de los casos.

En cambio, el recorrido intercuartílico nos brinda dimensión del recorrido de la variable, pero concentrándose solo en el 50% central:

Figura 5.5*Amplitud total y recorrido intercuartílico para datos en lista**Nota.* Fuente: elaboración propia

¿Cómo procederemos en caso de contar con los datos agrupados, en una tabla de frecuencias?

Para el caso de la amplitud total, no debemos avanzar por el mismo camino, es decir, tomando el valor más alto de la variable y restándole el valor más bajo.

En el caso del recorrido intercuartílico, recuerden que, como dijimos anteriormente, el cuartil 1 y el cuartil 3 se corresponden con los percentiles 25 y 75, respectivamente. Entonces, debemos ubicar el valor que le corresponde al caso que deja por debajo al 25% del conjunto, y aquel que deja por debajo al 75%. Veamos un ejemplo:

Supongamos que estoy trabajando con la variable “cantidad de saltos al cajón realizados correctamente en 30 segundos”, cuya distribución es:

Tabla 5.1*Tabla de frecuencias de cantidad de saltos realizados al cajón en 30 segundos*

Cantidad de saltos al cajón realizados correctamente en 30 segundos	Frecuencias absolutas	Frecuencias absolutas acumuladas	Frecuencias relativas (%)	Frecuencias relativas (%) acumuladas
0 saltos	6	6	4,28%	4,28%
1 salto	21	27	15%	19,29%
2 saltos	35	62	25%	44,29%
3 saltos	26	88	18,57%	62,86%
4 saltos	12	100	8,57%	71,43%
5 saltos	8	108	5,71%	77,14%
6 saltos	14	122	10%	87,13%

7 saltos	18	140	12,85%	100%
n=	140	-	100%	-

Nota. Fuente: elaboración propia

Así, vemos que, por ejemplo, 35 personas realizaron 2 saltos; o que 18 personas realizaron 7 saltos, etcétera.

Si decimos que el cuartil 1 es el valor que deja por debajo al 25% de los casos, entonces podemos proceder calculando el cuartil 1 de orden $\rightarrow (1 * n) / 4 = (1 * 140) / 4 = 140 / 4 = 35$

El cuartil 1 será el valor que se encuentre en la posición 35. Entonces, puedo tomar como referencia la columna de frecuencias acumuladas, y pensar estas frecuencias acumuladas como si fueran posiciones: así, tendré las primeras 6 posiciones ocupadas por las personas que realizaron 0 saltos; luego, tendré 21 posiciones (es decir desde la 7° hasta la posición 27°) ocupadas por las personas que realizaron 1 salto; luego, 35 posiciones (desde la 28° hasta la 62°) ocupadas por quienes hayan realizado 2 saltos. Y así sucesivamente, hasta llegar a la posición 140°. Debería identificar cuál es el valor, es decir, qué cantidad de saltos realizó la persona que está en la posición 35°. Así, vemos que esa persona realizó **2 saltos**.

Si procedo de la misma forma para averiguar el cuartil 3 de orden $\rightarrow (3 * n) / 4 = (3 * 140) / 4 = 420 / 4 = 105$ (lugar).

El cuartil 3 será el valor que se ubique en la posición 105° y ese valor es 5 saltos.

Procedemos a calcular el valor del recorrido intercuartílico, para esta distribución:

C3 - C1 = 5 saltos - 2 saltos = 3 saltos

Podemos afirmar que el 50% central de los casos, en esta distribución, realizó entre 2 y 5 saltos.

Otra forma de identificar los valores de los cuartiles 1 y 3 es concentrarnos en la columna de frecuencias relativas acumuladas. Como dijimos en más de una oportunidad, el cuartil 1 se identifica con el valor que supera al 25% de los casos, mientras que el cuartil 3 lo hace con aquel que supera al 75%. Entonces, basta con encontrar, en la columna de frecuencias relativas acumuladas, la menor frecuencia que contenga al 25% y al 75%.

Figura 5.6*Cuartiles en tabla de frecuencia*

	Cantidad de saltos al cajón realizados correctamente en 30 segundos	Frecuencias absolutas	Frecuencias absolutas acumuladas	Frecuencias relativas (%)	Frecuencias relativas (%) acumuladas
	0 saltos	6	6	4,28%	4,28%
	1 salto	21	27	15%	19,29%
Cuartil 1 →	2 saltos	35	62	25%	44,29%
	3 saltos	26	88	18,57%	62,86%
	4 saltos	12	100	8,57%	71,43%
Cuartil 3 →	5 saltos	8	108	5,71%	77,14%
	6 saltos	14	122	10%	87,13%
	7 saltos	18	140	12,85%	100%
	n	140	-	100%	-

← Contiene al 25%

← Contiene al 75%

Nota. Fuente: elaboración propia

Desvío estándar

Es la medida de variabilidad más utilizada dado que, a diferencia de la Amplitud Total y el Recorrido Intercuartílico, toma en cuenta a todo el conjunto de datos. Su interpretación se construye a partir de la diferencia entre el valor observado en cada individuo y la media del conjunto. Contempla en su cálculo todas las desviaciones individuales entre cada valor y la media, por eso diremos que la desviación estándar asumirá un valor para un conjunto de datos o una distribución, y no para cada caso individual. Es pertinente usar el desvío estándar cuando estamos frente a distribuciones simétricas, esto ocurre cuando la mayoría de los datos tienden a concentrarse en los valores centrales de la variable, “por ejemplo, en un test de fuerza muy pocas personas tendrán valores muy bajos o muy altos, la mayoría obtendrán puntuaciones medias” (Maureira Cid 2017, p. 53).

¿Cómo debemos proceder para calcular el desvío estándar?

Para responder a esta pregunta, podemos apelar a la fórmula:

Figura 5.7*Fórmula del desvío estándar*

The diagram shows the formula for standard deviation, $S = \sqrt{\frac{\sum (X - \bar{x})^2}{N - 1}}$, with several annotations and arrows pointing to specific parts:

- An arrow points to the S symbol with the text: "Símbolo de 'desvío estándar'".
- An arrow points to the summation symbol $\sum$ with the text: "Símbolo que me indica 'sumatoria' (es decir, me está dando la orden de sumar)".
- An arrow points to the variable X with the text: "Cada valor de la variable".
- An arrow points to the mean $\bar{x}$ with the text: "Promedio o media aritmética".
- An arrow points to the denominator $N - 1$ with the text: "Cantidad total de casos".

Nota. Fuente: elaboración propia

La fórmula nos dice que el desvío estándar es igual a la raíz cuadrada de la sumatoria de los desvíos de cada valor de la variable con respecto a la media (el promedio) al cuadrado, dividido la cantidad total de casos menos uno.

Para proceder con su cálculo, nos conviene ir resolviendo esa fórmula paso a paso, partiendo de la resolución de lo que está entre paréntesis y, desde allí, ir avanzando.

La fórmula de la varianza es exactamente igual, solo que no se calcula la raíz cuadrada.

Asimismo, debemos tener en cuenta que el desvío estándar se utiliza cuando hemos hallado la media aritmética como medida de tendencia central.

Para calcular desvío estándar se procede mediante una serie de pasos que explicaremos retomando el ejemplo del apartado anterior. Teníamos la siguiente distribución de la variable edad: 18 - 21 - 23 - 26 - 29 - 30 - 31 - 32

1° paso: Calculamos la media aritmética

$$(18 + 21 + 23 + 26 + 29 + 30 + 31 + 32) / 8 = 26,25 \text{ años}$$

2° paso: Calculamos los desvíos (la distancia) de cada puntaje con respecto a la media.

x	x-M
18	$(18 - 26,25) = - 8,25$
21	$(21 - 26,25) = - 5,25$
23	$(23 - 26,25) = - 3,25$
26	$(26 - 26,25) = - 0,25$
29	$(29 - 26,25) = 2,75$
30	$(30 - 26,25) = 3,75$
31	$(31 - 26,25) = 4,75$
32	$(32 - 26,25) = 5,75$

3° paso. Elevamos al cuadrado cada desvío a la media.

x	x-M	(x-M)²
18	-8,25	$-8,25^2 = 68,06$
21	-5,25	$-5,25^2 = 27,56$
23	-3,25	$-3,25^2 = 10,56$
26	-0,25	$-0,25^2 = 0,06$
29	2,75	$2,75^2 = 7,56$
30	3,75	$3,75^2 = 14,06$
31	4,75	$4,75^2 = 22,56$
32	5,75	$5,75^2 = 33,06$

4° paso. Calculamos la sumatoria de los cuadrados de los desvíos a la media.

X	X-M	(X-M) ²
18	-8,25	68,06
21	-5,25	27,56
23	-3,25	10,56
26	-0,25	0,06
29	2,75	7,56
30	3,75	14,06
31	4,75	22,56
32	5,75	33,06

Sumatoria de las
distancias de cada
valor de x con
respecto a la media =
183,48

5° paso. Dividimos la sumatoria por la cantidad de casos $n - 1$

$$\frac{183,48}{8-1} = 26,21$$

6° paso. Calculamos la raíz cuadrada de esa división.

$$\sqrt{26,21} = 5,11$$

Pero... ¿cómo interpretamos este resultado?

Una de las potencias que tiene conocer el desvío estándar de una distribución en comparación con otra. Es decir: si la distribución A tiene un desvío estándar de, supongamos, 3,41 y la distribución B un desvío estándar de 8,26, entonces podremos decir que la distribución más dispersa o la más heterogénea es la B. También podemos decir lo contrario: que la distribución menos dispersa o más homogénea, es la A. Cualquiera de estas interpretaciones es acertado.

¿Y por qué se llama desvío “estándar”? Ritchey (2006, p. 148) lo explica así: “la desviación estándar recibe su nombre por el hecho de que proporciona una unidad de medida común (un estándar) para comparar variables con unidades observadas de medida muy diferentes”. Esto resultará más claro más adelante, cuando avancemos sobre la distribución probabilística normal y veamos los *puntajes z*.

Ahora bien, ¿cómo podemos interpretar la variabilidad o dispersión de dos o más distribuciones con unidades de medida distintas?

Pongamos un ejemplo:

El desvío estándar de la altura de un grupo de 120 personas es de 35 centímetros. El promedio es de 171 centímetros.

El desvío estándar del **peso** de ese mismo grupo de personas es de 3,75 kilogramos. El promedio es de 81,4 kilogramos.

¿En cuál de las dos variables este grupo de personas presenta menos variabilidad o dispersión?

Como se desprende del ejemplo, tenemos dos variables: peso y altura, por lo que no podemos comparar directamente sus desvíos estándar. Entonces, necesitamos llevar esas dispersiones a una unidad común: el coeficiente de variación. Este coeficiente se calcula a partir del cociente entre el desvío y el promedio, y su multiplicación por 100. El resultado lo expresamos en porcentaje.

$$CV = (s/M) * 100$$

Aplicaremos, entonces, el **coeficiente de variación (CV)** para cada uno de nuestros ejemplos:

Altura

Media aritmética: 171 centímetros

Desvío estándar: 35 centímetros

$$(35/171) * 100 = 20,46 \%$$

Peso

Media aritmética: 81,4 kilogramos

Desvío estándar: 3,75 kilogramos

$$(3,75/81,4) * 100 = 4,60 \%$$

Como vemos, el coeficiente de variación es menor al calcularlo sobre la variable **peso**. Por ende, este grupo de 120 personas presenta una dispersión menor en cuanto a su peso, en comparación con su altura.

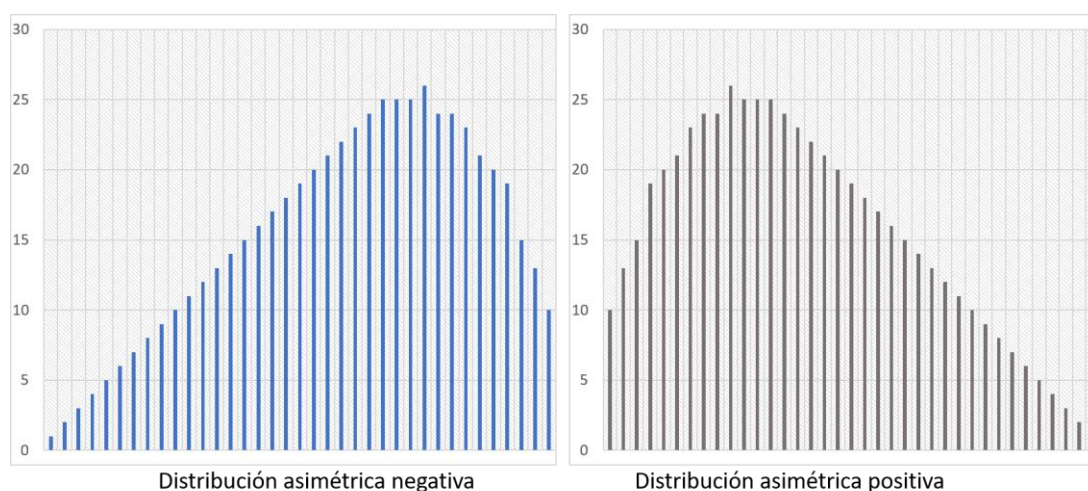
Como el coeficiente de variación se expresa en porcentajes, el menor valor posible será 0, mientras que el mayor será 100. A medida que se acerca a este último, diremos que la variabilidad de esa distribución será mayor.

Medidas de forma: asimetría y curtosis

Recién dijimos que el desvío estándar es útil para aplicar en distribuciones simétricas, y que estas se caracterizaban por concentrar gran parte de los casos en los valores centrales de la variable, y es por ello que la medida de tendencia central recomendada en estos casos es la media aritmética. ¿Pero qué sucede cuando las distribuciones son asimétricas? En este caso, la medida de tendencia central más adecuada será la mediana, pero, además, podemos caracterizar esa asimetría según sea positiva o negativa. En el primer caso, veremos que los casos tienden a concentrarse en los valores más bajos de la variable; mientras que, en el segundo, lo harán hacia los valores más altos. Gráficamente quedaría ilustrado así:

Figura 5.8

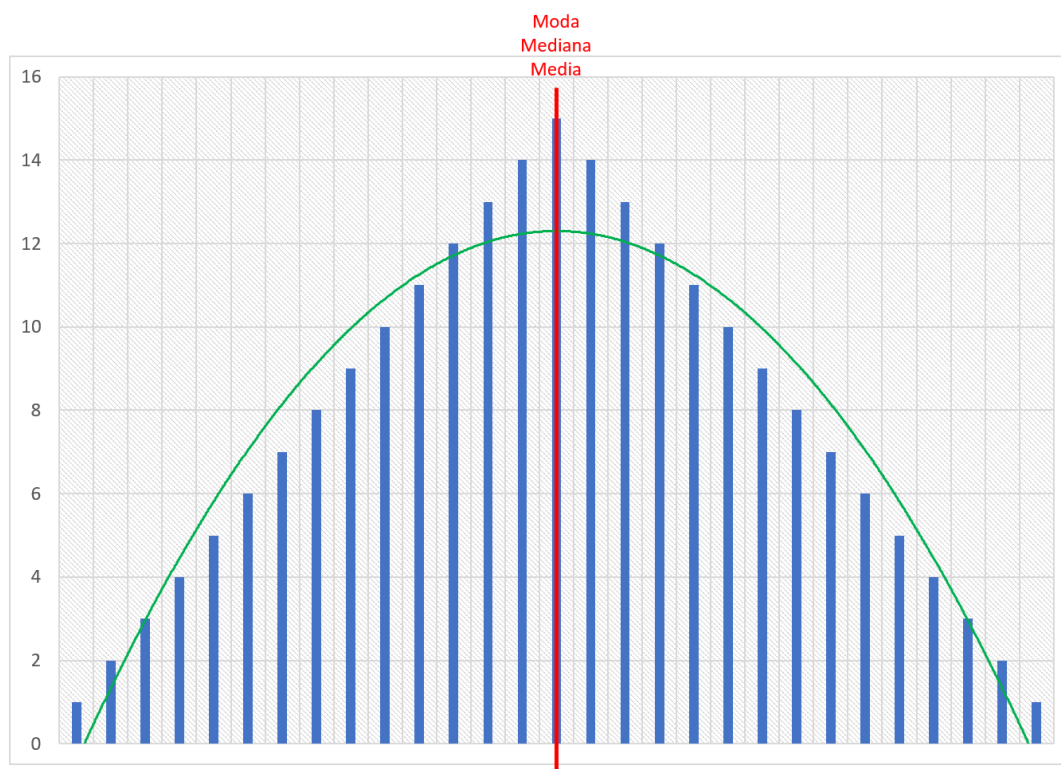
Distribuciones asimétricas



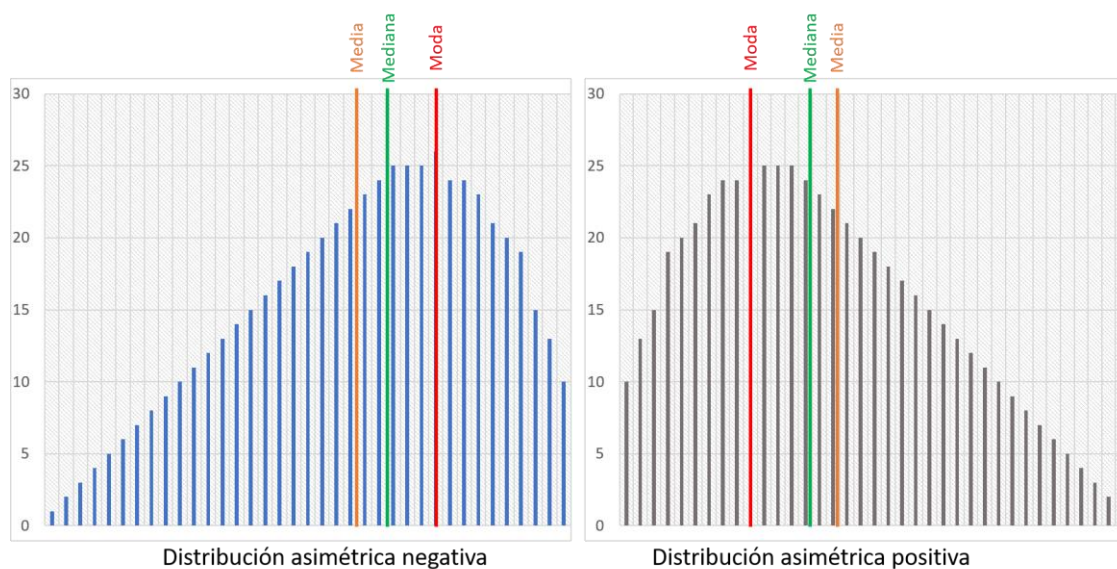
Nota. Fuente: elaboración propia

Las cajas que están debajo de las curvas delimitan el 50% central de los casos (el recorrido intercuartílico, es decir la distancia entre el cuartil 1 y el cuartil 3). Como se aprecia en las ilustraciones en la gráfica asimétrica negativa, esa caja tiende a posicionarse hacia la derecha (viendo el gráfico de frente), que es donde se ubican los valores más altos de la variable (que está en el eje x). Por otro lado, en la distribución asimétrica positiva, la caja que concentra al 50% central de los casos, se ubica con un sesgo hacia la izquierda, donde se encuentran los valores más bajos de la variable.

En una distribución simétrica, esa caja se encontraría en el centro, porque la mayoría de los casos tenderían a concentrarse en torno a los valores centrales de la variable:

Figura 5.9*Distribución simétrica**Nota.* Fuente: elaboración propia

También es importante tener en cuenta cómo se ubicarán, en cada uno de estos casos, las medidas de tendencia central (modo, mediana y media aritmética):

Figura 5.10*Medidas de tendencia central en distribuciones asimétrica positiva y asimétrica negativa**Nota.* Fuente: elaboración propia

En las distribuciones asimétricas positivas, la media aritmética queda por arriba de la mediana, mientras que, en las distribuciones asimétricas negativas, la mediana se ubicará por encima del promedio. Esto sucede porque, como vimos anteriormente, la media aritmética es susceptible a valores extremos no compensados, entonces esos valores extremos van a “arrastrar” al promedio, ya sea hacia valores más altos o más bajos de la variable.

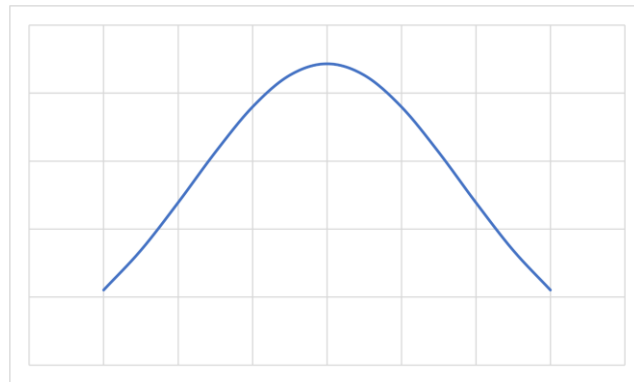
Técnicamente este nivel de asimetría se puede calcular, pero lo importante es que logremos identificar en torno a qué valores de la variable se concentran los casos en cada situación.

Otro elemento que nos permite describir la forma de las distribuciones es la curtosis. Para expresarlo coloquialmente, podemos pensar a la curtosis como una medida que nos indicará cuán “en punta” se percibe la gráfica de un conjunto de datos, como se muestra a continuación.

Distribución platicúrtica:

Figura 5.11

Distribución platicúrtica

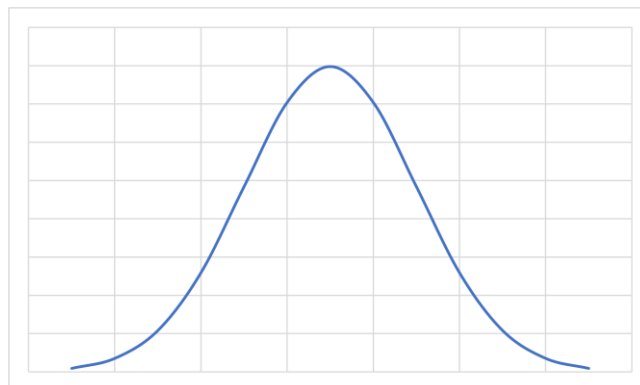


Nota. Fuente: elaboración propia

Distribución mesocúrtica:

Figura 5.12

Distribución mesocúrtica

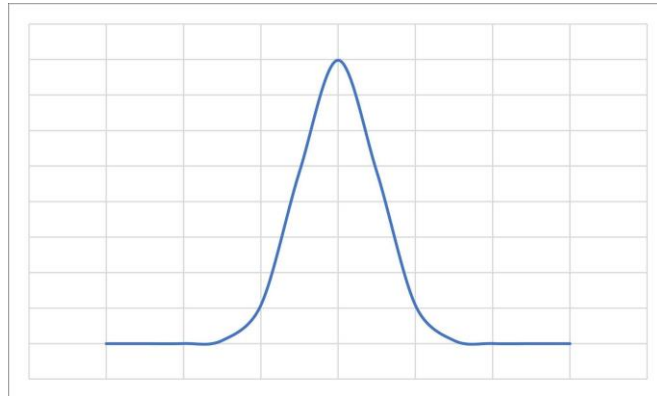


Nota. Fuente: elaboración propia

Distribución leptocúrtica:

Figura 5.13

Distribución leptocúrtica



Nota. Fuente: elaboración propia

La curtosis nos permite identificar el nivel de dispersión de los datos que conforman las distribuciones. En este caso, vemos que la distribución leptocúrtica es la menos dispersa, seguida de la mesocúrtica. Luego, encontramos la distribución platicúrtica que, de las tres, es la que muestra un mayor grado de dispersión.

Ahora bien, más allá de la identificación de la curtosis a partir de los gráficos, esta medida se puede calcular a partir de la siguiente fórmula:

$$\text{curtosis} = Q / (\text{percentil } 90 - \text{percentil } 10)$$

Donde **Q** es igual a la mitad del recorrido intercuartílico, es decir:

$$Q = (\text{Cuartil } 3 - \text{Cuartil } 1) / 2$$

Según el resultado de esta operatoria, consideraremos que:

la distribución es mesocúrtica si el resultado es igual a 0,263

la distribución es leptocúrtica si el resultado es mayor a 0,263

la distribución es platicúrtica si el resultado es inferior a 0,263.

Debemos tener en cuenta que, para poder describir densamente la variabilidad de una distribución, será necesario aplicar la batería de medidas con las que trabajamos a lo largo de este capítulo, de manera tal que podamos decir varias cosas sobre la heterogeneidad u homogeneidad de los datos. Asimismo, en estadística es recomendable la aplicación combinada de las medidas de variabilidad junto con las de tendencia central y de posición. Por eso, si bien en este libro de cátedra presentamos los conjuntos de medidas en distintos capítulos, este agrupamiento no persigue otro fin que no sea el didáctico para la exposición de los contenidos. En estadística se trabaja con el conocimiento simultáneo de todos estos conjuntos de medidas.

Balance del capítulo

Con este capítulo concluimos una parte de los temas de la materia en la que abordamos distintos grupos de herramientas que nos brindan información muy útil para describir conjuntos de datos: las medidas de tendencia central, las medidas de posición, las medidas de variabilidad. También, desarrollamos herramientas que nos ayudan a organizar y presentar los datos con los que trabajamos: las representaciones gráficas, las tablas de frecuencias, las medidas resumen.

Es imperioso advertir que todas estas herramientas, en general, deben aplicarse a la vez, porque todas ellas ofrecen una lectura integral, de conjunto sobre las distribuciones con las que trabajamos. Esto nos permitirá elaborar descripciones densas y ganar complejidad en la interpretación de los datos.

Concentrándonos en este capítulo, la variabilidad nos indica cuán dispersos se encuentran los datos entre sí, cuán “pareja” demuestra ser la distribución. No nos brinda elementos para determinar si tal o cual variable tuvo un “mejor” o un “peor” desempeño, sino, simplemente, si ese desempeño fue más o menos estable. Así, el desvío estándar es un coeficiente muy útil para medir la variabilidad, pero exhibe una limitación cuando nos interesa comparar dos distribuciones con distinta unidad de medida, allí es cuando echamos mano al coeficiente de variabilidad. La amplitud total y el recorrido intercuartílico nos permitirán una primera aproximación al nivel de variabilidad de la distribución. Por otro lado, la forma que adquiera la distribución, cuando lo grafiquemos, nos permitirá identificar intuitivamente el nivel de curtosis y asimetría (aunque, como vimos, hay medidas técnicas que permiten calcularlas exactamente).

A continuación, nos enfocaremos en un tipo de distribución probabilística que es simétrica y que está desarrollada sobre un modelo matemático que nos permitirá identificar algunas características del conjunto de datos, siempre que se aproxime a este modelo: la curva normal o “campana de Gauss”.

Referencias

- Bologna E., Morales M., González M. Cristina de Menne, Faas A., Medrano L., Romero W., Reyna C., Urrutia A. (2013). *Estadística para psicología y educación*. Editorial Brujas 3° Ed., (p. 69; pp. 91-109). Córdoba: Argentina.
- Maureira Cid, F. (2017). *Estadística básica para Educación Física*. Estadística descriptiva e inferencial univariada. 2da. Ed., Capítulo 3 (pp. 51-54) y capítulo 4 (pp. 61-66). Editorial Académica Española.
- Ritchey, F. (2006) *Estadística para las Ciencias Sociales*. Capítulo 5. Medición de la dispersión o distribución de una medición de puntuaciones (pp. 136-167). Editorial Mc Graw-Hill Interamericana.

Sandoval Dávila, J. (2012) *Fundamentos de estadística aplicada a la investigación social*.
Capítulos 6 y 7 (pp. 70-82).

CAPÍTULO 6

Probabilidad y distribuciones probabilísticas. La distribución normal

Stella Maris Doná y Martina Díaz

Introducción

La Estadística proporciona herramientas para abordar el estudio de fenómenos en los cuales no se conocen las leyes que los rigen; a estos eventos los llamamos fenómenos aleatorios. La comprensión de su naturaleza supone la aplicación de los conocimientos proporcionados por la Teoría de la Probabilidad.

Su estudio es imprescindible para prever la ocurrencia de fenómenos de nuestro interés, realizar generalizaciones inferenciales al ajustar de la manera más exacta posible los imponderables debidos al azar que inciden en el desarrollo de todo hecho.

Un fenómeno aleatorio es aquel que ejecutado n cantidad de veces bajo las mismas condiciones, no siempre resulta igual, puede tener varios resultados diferentes. En cada ocurrencia del fenómeno es imposible predecir, anticipar, el resultado, atribuyéndose al azar.

Un experimento es un fenómeno aleatorio cuando se cumplen las siguientes condiciones: se puede repetir indefinidamente, siempre en las mismas condiciones previa realizarlo, no se puede predecir el resultado que se va a obtener con total certeza luego de realizarlo, el resultado que se obtuvo pertenece a un conjunto de resultados posibles y conocidos previamente.

Ejemplo: si un tenista va a ir a jugar un partido, puede ser que gane o pierda. Antes de jugar el partido, no puede asegurar que vaya a ganar ni que vaya a perder. Ganar y perder integran el conjunto de resultados posibles y conocidos previamente del evento *jugar el partido de tenis*. Una vez que terminó el partido, su resultado fue o bien ganó, o bien perdió. La posibilidad de *empate* no pertenece al conjunto de resultados posibles.

Denominaremos espacio muestral al conjunto formado por cada uno de los resultados posibles de ocurrencia, previo al suceso aleatorio. Suele denotarse con la letra omega mayúscula: Ω

A los distintos resultados posibles de un fenómeno aleatorio lo denominamos evento aleatorio.

En nuestro ejemplo, el espacio muestral lo conforman dos eventos aleatorios posibles de ocurrir antes de que se juegue el partido: triunfo y victoria. Empate podría ser un resultado del

espacio muestral para el fenómeno aleatorio “jugar un partido de fútbol” pero no es un evento posible de encontrar en el espacio muestral de nuestro fenómeno aleatorio “jugar un partido de tenis”.

Concepto de probabilidad

La probabilidad, denotada con la letra *p*, es una medida de la posibilidad de ocurrencia de un evento entre otros de la misma índole.

La medida de la probabilidad se establece matemáticamente como una razón entre la frecuencia con que aparece el resultado deseado y el número total de resultados posibles que existen.

El resultado de tal cociente es siempre un número que se encuentra dentro del rango de 0 a 1.

Si el resultado de dicha división es 0, se interpreta como la imposibilidad de que ocurra el resultado deseado. Cualquier valor bajo o próximo a 0 estará indicando poca, o tendiente a nula, la probabilidad de que se verifique un resultado esperado.

Por el contrario, si el resultado numérico del cálculo de ese cociente es igual a 1, significa la seguridad de la ocurrencia del evento deseado. Y cuanto más alto sea el valor calculado de una probabilidad, cuanto más se aproxime a 1, se interpretará como altamente probable la posibilidad de que ocurra el evento deseado (Levin y Levin, 2004).

Entonces:

Figura 6.1

Fórmula probabilidad

$$p = \frac{\text{frecuencia con que ocurre el evento deseado}}{\text{cantidad total de resultados posibles}} \quad \text{donde } 0 \leq p \leq 1$$

Nota. Fuente: elaboración propia

Ejemplo:

¿Cuál es la probabilidad de tirar un dado y que salga la cara con el número 4?

Un dado tiene seis caras; en cada cara hay un número. El 4 es uno de esos seis números del dado. Por lo tanto, suponemos que el evento deseado es que salga el número 4. La frecuencia con que aparece el 4 en un dado es sólo una (un dado tiene sólo un número 4 entre sus seis lados). La cantidad total de números que pueden salir al tirar el dado son seis (puede que tire el dado y salga el 1, o salga el 2, o salga el 3, o salga el 4 –que es lo que quiero que

ocurra-, o salga el 5, o salga el 6). De esas seis posibilidades, sólo una es la que deseo: que salga el 4.

A priori, la probabilidad de que vaya a tirar un dado y de que salga el número 4 es sólo una de seis:

$$p = \frac{1}{6}$$

El resultado de ese cociente es $p = 0,17$. El valor numérico calculado de la probabilidad de arrojar un dado y que salga el lado con el número 4 es $p = 0,17$; un valor bajo, indicando que tal evento es poco probable que ocurra.

El fenómeno aleatorio “tirar un dado” admite un espacio muestral con seis elementos, los resultados posibles antes de tirar el dado son [1, 2, 3, 4, 5, 6]. El evento que se busca estudiar es el de obtener el número 4 al tirar el dado.

Es importante entender el hecho de que a priori un evento que tenga poca probabilidad de presentarse no significa que a posteriori de realizado el experimento, no se presente. Solamente será imposible de que ocurra cuando la probabilidad calculada sea igual a 0.

De igual manera, sólo estaremos seguros de que un evento deseado ocurra cuando la probabilidad calculada sea igual a 1.

Se vuelve necesario diferenciar entre la probabilidad a priori, aquella que se calcula antes de que se produzca el experimento (por ejemplo, tirar un dado y que salga el 4), y la probabilidad *a posteriori* de haberse producido el experimento. Hay un tercer tipo de probabilidad llamada subjetiva pues se asigna un valor de probabilidad a los sucesos basado en el grado de creencia o de confianza de que el suceso pueda ocurrir (Elorza Pérez Tejada, 2008).

La probabilidad a posteriori es aquella calculada cuando ya ocurrió el hecho (por ejemplo, ya se tiró el dado y se observa qué número salió) la probabilidad también se calcula matemáticamente como un cociente, pero esta vez entre la cantidad de veces en que efectivamente ocurrió lo que deseaba, dividido el número total de veces en que se realizó el experimento esperando que ocurra lo que se deseaba (pues no siempre ocurre lo deseado).

Siguiendo con el mismo ejemplo, ya calculamos que el valor de la probabilidad $p=0,17$ es a priori, lo que se espera que vaya a ocurrir.

A posteriori de realizado el experimento, cuando se tira el dado, efectivamente se observa si salió el 4, o no salió.

En el caso en que el dado se tire no una sola vez, sino muchas veces, se puede ir contando la cantidad de veces que se tira el dado y, de cada una de esas veces, la cantidad de veces en que sale el número 4.

Vamos a suponer que se tira el dado 30 veces. Y vamos a suponer que en siete oportunidades sale el número 4.

El cálculo de la probabilidad a posteriori sería:

$$p = \frac{\text{frecuencia con que ocurrió el evento deseado}}{\text{cantidad total de experimento realizados}}$$

Donde la probabilidad p siempre es un valor numérico entre 0 y 1. Esto se escribe así:
 $0 \leq p \leq 1$.

A posteriori de haber tirado el dado, la probabilidad de que haya salido el número 4 fue:

$$p = \frac{7}{30}$$

El resultado de ese cociente es $p = 0,23$. El valor numérico calculado de la probabilidad de que, al arrojar 30 veces un dado, y de que salga el lado con el número cuatro en 7 oportunidades, es $p = 0,23$. Este valor de probabilidad indica que tal evento (tirar el dado 30 veces y que 7 veces sea un cuatro el lado que sale) tiene probabilidad de ocurrencia baja.

Nótese que este valor calculado de la p -valor a posteriori no es más que la conocida frecuencia relativa; o lo que es lo mismo: la proporción de casos; o, como podemos también llamarlo, la probabilidad con que se constató la ocurrencia de un evento en relación con el total de veces efectivamente realizado. Tres maneras equivalentes de describir una misma realidad a posteriori de que el hecho suceda.

La teoría probabilística es la bisagra que permite pasar de la descripción a la inferencia estadística. Debemos tener en cuenta que:

- Las bases para el desarrollo del concepto y cálculo de la probabilidad a priori se deben a Pierre-Simon de Laplace¹³.
- El enfoque de la probabilidad a posteriori es aportado por Andrén Nikoláyevich Kolmogórov¹⁴.

Las bases desarrolladas por Laplace se expresan en tres leyes:

1ra. Ley. El valor numérico de la probabilidad se encuentra siempre entre 0 y 1:

$$0 \leq p \leq 1$$

2da. Ley. Cuando se desea calcular la probabilidad de que ocurra el evento A o el evento B, siendo ambos eventos mutuamente excluyentes –que no pueden ocurrir en el mismo experimento–, entonces, la probabilidad de que ocurran A ó B, se calcula sumando las probabilidades de A y la de B, cada una por separado, y luego sumándolas:

$$p(A \text{ o } B) = p(A) + p(B)$$

¹³ Pierre-Simon de Laplace (Francia, 1749-1827). Astrónomo y matemático, sentó las bases sobre la Teoría Analítica de la probabilidad.

¹⁴ Andrén Nikoláyevich Kolmogórov (Rusia, 1903-1987). Matemático; realizó grandes aportes sobre la Teoría de la Probabilidad y uso de la Teoría de Conjuntos.

Ejemplo: tiramos un dado. El evento “A” podría ser la probabilidad de que salga el número 4. El evento “B” podría ser la probabilidad de que salga el número 5. Entonces, la probabilidad de que al tirar el dado salga un 4 o salga un 5 es:

$$p(4 \text{ ó } 5) = p(4) + p(5)$$

$$p(4 \text{ ó } 5) = 1/6 + 1/6 \quad p(4 \text{ ó } 5) = 2/6$$

$$p(4 \text{ ó } 5) = 0,33$$

El p-valor calculado indica que, a priori, en 1 de cada 3 tiradas del dado deberían salir los números 4 o 5.

A esta ley se la conoce como **de la suma o de la probabilidad simple**.

3ra. Ley. Cuando se desea calcular la probabilidad de que ocurra un evento A y un evento B, siendo ambos **eventos independientes entre sí**, pero la ocurrencia de uno de los eventos condiciona la aparición del otro evento-, entonces, la probabilidad de que ocurran A y B, se calcula multiplicando las probabilidades de A y la de B:

$$p(A \text{ y } B) = p(A) * p(B)$$

Ejemplo: Tenemos un mazo de cartas francesas (52 cartas: 26 de color rojo y otras 26 de color negro: cuatro palos -pica, corazón, trébol, diamante- de 13 cartas por cada palo- 1 al 10 y J, Q, K).

El evento “A” podría ser la probabilidad de que salga una carta roja. El evento “B” podría ser la probabilidad de que salga una carta J.

Entonces, la probabilidad de que salga una carta roja y que sea una J:

$$p(\text{roja y J}) = p(\text{roja}) * p(J)$$

$$p(\text{roja y J}) = 26/52 * 4/52$$

$$p(\text{roja y J}) = 0,50 * 0,08$$

$$p(\text{roja y J}) = 0,04$$

El valor de probabilidad calculado de nuestro ejemplo indica que, a priori, es muy baja la probabilidad de sacar una carta del mazo y que esa carta sea justo, una J-roja. A esta ley se la conoce como de la multiplicación o de la probabilidad compuesta.

La probabilidad a priori es lo que se espera que ocurra, lo que teóricamente debería suceder.

Esta probabilidad se conoce por haberse estudiado cómo se podría presentar un hecho, en una cantidad infinita de repeticiones de la misma.

Se repite un mismo evento una vez, y otra vez, y otra vez, y así hasta un número infinito de veces; y en cada una de las repeticiones se va registrando el resultado que aparece.

La probabilidad a posteriori es lo empíricamente observado, lo que efectivamente ocurrió en una cantidad finita de repeticiones de un suceso. Es la frecuencia relativa de cada valor de la variable: la relación entre la cantidad de casos observados para un valor de la variable, y el total de casos de la muestra: la proporción de casos.

Entonces, frente a cada suceso tendremos lo que en teoría debería esperarse que se dé, la frecuencia con que teóricamente debe esperarse que ocurra un evento; y, por otro lado, está lo que realmente se observa luego de realizado el experimento.

Ejemplo: si administró un test de Aptitud Física a un grupo de sujetos.

A priori puede predecirse qué puntajes deberían sacarse esos sujetos pues ya se lo ha estudiado en un sinnúmero de casos.

Luego que se administra al Test, se obtienen los valores que efectivamente sacaron los sujetos que conforman la muestra.

El registro preexistente de lo que debería esperarse a priori que ocurra (lo que se conoce como baremo) y el registro de los puntajes observados a posteriori, se comparan. Entre lo esperado teóricamente y lo observado empíricamente, siempre hay diferencias. La clave es determinar si la discrepancia entre la teoría y la práctica es o no es significativa. La cantidad de veces en las que, en teoría debería esperarse que un determinado suceso se dé, es una frecuencia de probabilidad teórica a priori de que se realice un experimento.

La cantidad de veces en que ocurre efectivamente un suceso es la frecuencia absoluta con que se dio ese suceso, a posteriori de realizado.

Las frecuencias con que los eventos de una misma índole –en teoría- deberían presentarse, constituyen un modelo teórico de referencia con el cual se puede comparar y referenciarse lo que empíricamente observamos sobre una muestra.

Las distribuciones de frecuencias de probabilidades teóricas, a priori de un evento, ofrecen un patrón o modelo de lo que debería esperarse que fuese a ocurrir en la práctica concreta, con una distribución empírica de una variable que se investiga.

Concepto de distribución probabilística

Una distribución probabilística es un modelo teórico-matemático basado en distribuciones empíricas. Es un modelo de distribución de frecuencias relativas teóricas, no están definidas por la observación, sino que son postuladas a priori, pero atendiendo a cómo se comportan los fenómenos aleatorios en la experiencia.

Una distribución probabilística es un patrón del modo de distribución de las frecuencias relativas teóricas de una variable aleatoria; muestra la forma en que se esperan que varíen los datos empíricos observados, esos datos de una característica en estudio, específica, tomada sobre una muestra concreta.

Una distribución probabilística muestra las infinitas probabilidades (frecuencias relativas), con que un valor pueda ser observado. Como los fenómenos son de muy variados, hay diferentes Distribuciones Probabilísticas, cada una desarrollada para poder anticipar cuán probable es que ocurra un evento deseado.

Esos datos empíricos que recogemos, organizamos y analizamos (recuerde la definición de “Estadística”) surgen de una determinada variable que se desea describir. Y como ya sabemos, no hay un único tipo de variable.

Luego, es lógico pensar que a una variable cualitativa le corresponda un modelo probabilístico de lo que se esperaría observar, diferente al modelo probabilístico de referencia de una distribución de datos cuantitativos.

Según sea el tipo de variable en estudio, la cantidad de casos que conforman la muestra y la forma gráfica que adquiere su distribución de frecuencias, será el modelo de distribución teórico probabilístico de referencia para conocer, en teoría, qué debería esperarse de la distribución de frecuencias empíricas, concretamente observada sobre la variable en estudio.

Los modelos más usados en ciencias sociales son:

Tabla 6.1

Tipo de variable y Modelo de distribución correspondiente

Tipo de variable	Modelo de Distribución probabilística
Cualitativa	Chi cuadrado
Cuantitativa discreta	Binomial
Continua – pocos casos	t de Student
Continua – muchos casos	Normal

Nota. Fuente: elaboración propia

Distribución probabilística binomial

Es un modelo para una variable discreta donde sólo existen dos posibles resultados a encontrar: éxito o fracaso. Los valores paramétricos de la Distribución Binomial son: Media= $n \cdot p$; Varianza= $n \cdot p \cdot q$.

La Distribución se basa en el desarrollo de un binomio de Bernoulli para resolver problemas probabilísticos, siempre que se cumplan con las siguientes condiciones (Maureira Cid, 2017):

- **La variable debe ser dicotómica.** Sólo toma dos valores, situación dicotómica. (Ejemplo: cara o ceca una moneda, varón o mujer, ganar o perder un partido, aprobar o desaprobado un examen, etc). Los dos valores de una variable de este tipo se denominan “éxito” o “fracaso”, la probabilidad asignada al éxito se la denota con la letra **p** y la del fracaso con la letra **q**.
- **Condición de estabilidad.** La probabilidad de éxito **p** debe permanecer constante en cada uno de los ensayos (ejemplo: la probabilidad de obtener un as al arrojar un dado es $\frac{1}{6}$, si lo arrojó cinco veces esa probabilidad debe ser la misma en cada una de las tiradas que haga).
- **Condición de independencia.** La probabilidad de obtener éxito en un ensayo no aumenta ni disminuye si se conoce el resultado de otro ensayo. (Ejemplo: la probabilidad de obtener as en un segundo lanzamiento no debe aumentar ni disminuir

cuando se sabe el resultado anterior, cuando se conoce que en el lanzamiento anterior hubo as).

La probabilidad de eventos cuando tenemos una variable dicotómica se calcula desarrollando el siguiente coeficiente binomial: $(p + q)^n$

Donde **p** es la probabilidad de éxito; **q** es la probabilidad de fracaso, **n**= número de ensayos.

Ejemplo: ¿Cuál es la probabilidad de tirar 4 dados y sacar el número 3, dos veces?

p es la probabilidad a priori de éxito, de sacar 3 cuando se tira un dado, es sólo una posibilidad (sólo hay un número 3) entre los números posibles (el dado tiene seis números: 1, 2, 3, 4, 5, 6)

$$p = 1 / 6 = 0,17$$

q es la probabilidad *a priori* de fracaso, de no sacar el número 3 cuando se tira un dado, es la posibilidad de obtener cualquier de los otros cinco números del dado (1, 2, 4, 5, 6)

$$q = 5 / 6 = 0,83$$

n es el número de veces en que se tira un dado; **n**=4

De ello, resulta que:

$$(p + q)^4 = [p^4] + [4 \cdot p^3 \cdot q] + [6 \cdot p^2 \cdot q^2] + [4 \cdot p \cdot q^3] + q^4$$

$$(1/6 + 5/6)^4 = [(1/6)^4] + [4 \cdot (1/6)^3 \cdot (5/6)] + [6 \cdot (1/6)^2 \cdot (5/6)^2] + [4 \cdot (1/6) \cdot (5/6)^3] + (5/6)^4$$

$$(1/6 + 5/6)^4 = 0,12090$$

Interpretación: si tiro un dado 4 veces (o una sola vez, 4 dados juntos), la probabilidad de que salga el número 3, dos veces es 0,12. Expresado en porcentaje, se puede decir que hay un 12% de probabilidades de sacar un número 3, al tirar cuatro veces un dado.

Distribución probabilística t de Student

Es un modelo de distribución de probabilidad para variables continuas con muestra de pocos casos ($n < 30$). Su importancia también radica en que es utilizada para probar hipótesis estadísticas. Conforme aumenta la cantidad de casos de la muestra, la Distribución t se va pareciendo más a la distribución normal.

La Distribución t, describe la distancia de la media de una muestra, a la de la población, cuando la desviación estándar de la población no se conoce, y las observaciones vienen de una población con una distribución normal. La Distribución tiene la capacidad para trabajar con muestras pequeñas, donde la varianza poblacional es desconocida. En tales casos, la Distribución t proporciona una herramienta valiosa para realizar inferencias estadísticas con mayor precisión que la Distribución normal estándar.

Se caracteriza por ser simétrica respecto del valor de media y se abre en dos colas hacia cada lado. Su gráfica se construye atendiendo a los grados de libertad, directamente relacionados con la cantidad de casos **n** de la muestra.

Se aplica para comparar el promedio de los valores de una variable continua, obtenida en una muestra pequeña, con la media poblacional; también para comparar las medias de dos muestras independientes y determinar si existe una diferencia significativa entre ellas; o comparar medias de dos muestras dependientes o apareadas. Otro uso importante es para calcular intervalos de confianza para la media de una población, especialmente cuando el tamaño de la muestra es pequeño.

Su fórmula es:

$$t = \frac{M - \mu}{s / \sqrt{n}}$$

En el denominador se resta los valores de la media muestral $\bar{x}$, y la media poblacional μ ; s es el valor del desvío estándar de la muestra y n su cantidad de casos.

El procedimiento para determinar la significancia estadística de una Probar de hipótesis consiste en obtener dos valores de t, compararlo y tomar una decisión. Se aplica una fórmula y se alcanza un **valor t** llamado **calculado**. Por otro lado, se busca otro valor de t llamado crítico o teórico (se obtiene de una Tabla, teniendo en cuenta los grados de libertad y el nivel de significación que se establezca). Una vez que se tienen ambos valores de t, se los compara. Si el t calculado es mayor que el t crítico, se acepta que rechaza la hipótesis nula.

Distribución probabilística normal

Aplicaciones de la distribución probabilística curva normal

- Comparar el rendimiento individual de un sujeto con el rendimiento del grupo al cual pertenece.
- Comparar el rendimiento de un mismo sujeto en varias pruebas diferentes.
- Averiguar la probabilidad de ocurrencia de un determinado evento o el porcentaje de casos esperables de ser observados, por debajo de un puntaje o entre dos puntajes determinados.
- Averiguar hasta qué puntuación o entre qué puntuaciones, se encuentra con cierta probabilidad de ocurrencia de un evento o se encuentra cierto porcentaje de casos.
- Clasificar o dividir un grupo, en subgrupos a los que se puede asignar categorías diagnósticas.
- Se aplica en el campo de la estadística inferencial para la estimación de parámetros y en la prueba de hipótesis.

Características de la curva normal

Es el modelo teórico matemático de referencia para variables cuantitativas continuas, cuya muestra está integrada por un número alto de unidades de análisis. Es una distribución paramétrica que se aplica a distribuciones muestrales.

La distribución probabilística Curva Normal es la expresión de una fórmula matemática desarrollada por el francés Abraham De Moivre (1667-1774) cuya función se representa gráficamente con una curva continua que se define por su media (simbolizada con la letra griega minúscula “mú” μ) y su desviación estándar (simbolizada con la letra griega minúscula “sigma” σ).

No hay una única curva normal sino infinitas pues no hay un único par de valores de μ y σ sino infinitos.

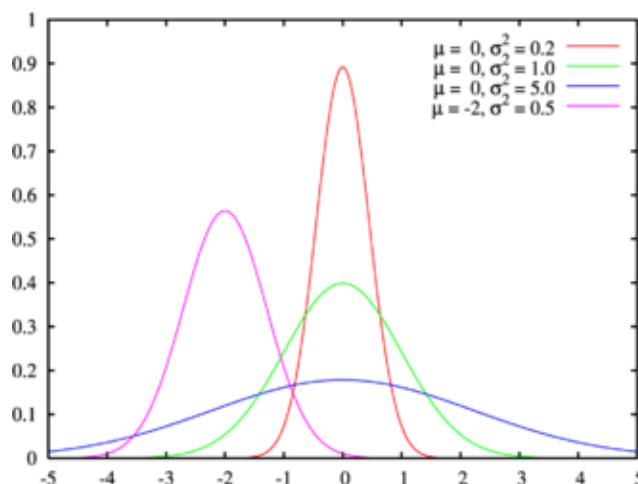
En todos los casos, la gráfica presenta forma de campana, apreciándose una concentración de casos en el centro de la distribución, disminuyendo simétricamente hacia ambas colas de la misma, de manera asintótica (no corta el eje de la abscisa), pues la cantidad de casos que se representan en el área bajo la curva, son infinitos.

A esta distribución probabilística también se la conoce con el nombre de Campana de Gauss.

Carl Friedrich Gauss¹⁵, un matemático posterior a De Moivre, continuó con los desarrollos estadísticos de probabilidad teórica, como modelos esperables del comportamiento de variables continuas aleatorias. La siguiente gráfica¹⁶ muestra distintas curvas normales según diferentes valores de media poblacional y de varianza.

¹⁵ Carl Friedrich Gauss (Alemania, 1777-1855). Matemático y astrónomo; fue niño prodigio. En la teoría de la probabilidad, desarrolló el método de los cuadrados mínimos y las leyes fundamentales de la distribución de la probabilidad. La curva normal de la probabilidad se sigue llamando curva de Gauss.

¹⁶ Imagen disponible en https://es.wikipedia.org/wiki/Distribuci%C3%B3n_normal

Figura 6.2*Curvas normales con diferentes medias poblacionales y varianza*

Nota. Fuente: adaptado de https://es.wikipedia.org/wiki/Distribuci%C3%B3n_normal

La distribución de probabilidad de una variable continua aleatoria x , es una función que asigna a cada valor posible de dicha variable, un valor de probabilidad p de ocurrencia.

Es habitual que los datos que un investigador recoja para describir un aspecto de la realidad de su interés sean valores numéricos reales, continuos de la variable (años, tiempos, puntajes, etc) y sean además aleatorios en tanto que cada resultado del espacio muestral se le asocia un número real, de manera azarosa.

Muchos de los fenómenos de las ciencias humanas se distribuyen como la curva normal dado que se trata de variables aleatorias continuas que, además, al graficarlas se observa que se distribuyen tendiendo a la forma de campana gausseana.

Luego, el modelo teórico de la distribución probabilística paramétrica de curva normal se usa como patrón para el análisis de datos obtenidos en dichas distribuciones empíricas muestrales de observaciones concretas.

El modelo, brinda una aproximación teórica a la probabilidad de que, en la práctica, se presente un cierto valor de la variable en estudio. Proporciona –al transformar el valor de la probabilidad y multiplicarlo por 100- el porcentaje de casos esperables hasta un cierto valor de la variable x .

Permite encontrar el porcentaje de casos bajo la Curva entre dos valores puntuales de la variable. Brinda una manera de comparar los rendimientos individuales con relación al rendimiento promedio de su grupo.

A la Curva Normal que se caracteriza por una $\mu=0$ (Media=0) y un $\sigma=1$ (Desvío), se la denomina Estándar. Es mesocúrtica y simétrica respecto de su punto medio.

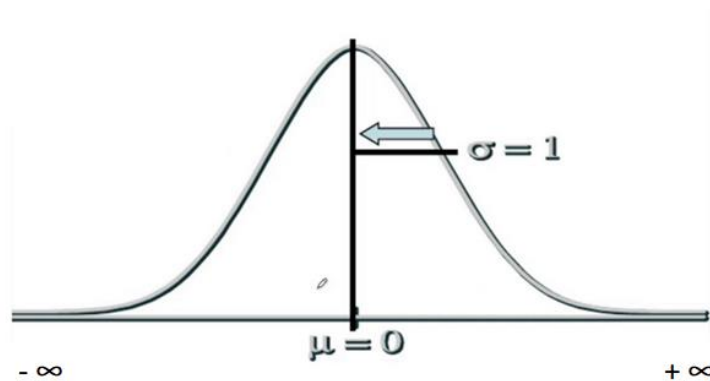
Ese punto medio es el que corresponde al valor de la Media que, por ser una Curva simétrica, coincide con el valor de la Mediana y del Modo.

Por debajo de la curva está determinada el área de probabilidad de ocurrencia de todos los eventos de una misma índole. Entonces, la probabilidad total bajo la curva normal es igual a 1, y la mitad de esa probabilidad es decir 0,50, se presenta desde $-\infty$ (menos infinito) hasta el valor la Media $\mu = 0$. Lo mismo desde el punto central $\mu = 0$, hacia $+\infty$ (más infinito).

Es característico de la Distribución Normal que, entre dos puntos de la abscisa, el área bajo la curva es constante. Es equivalente decir que, por debajo de la curva normal, están previstos el 100,00% de los casos de la misma naturaleza; el 50,00% de los casos hasta el punto central y otro 50,00% desde ese mismo punto medio. En ambas colas, hasta el ∞ (infinito).

Figura 6.3

Curva normal

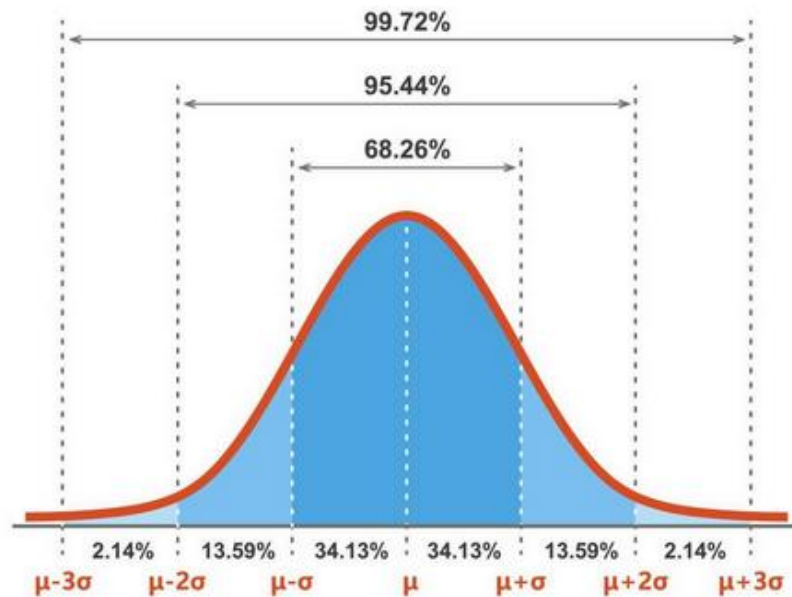


Nota. Fuente: elaboración propia

Otra propiedad de la Curva Normal es tener porcentajes iguales de áreas entre puntos equidistantes a su Media μ (μ), como lo muestra la siguiente figura:

Figura 6.4

Curva normal y porcentaje de casos según cantidad de unidades de desvío



Nota. Fuente: adaptado de <https://es.vecteezy.com/arte-vectorial/7695520-distribucion-de-gauss-estandar-normal-distribucion-gaussiana-campana-grafico-curva-negocios-y-marketing-concepto-matematicas-probabilidad-teoria-editable-trazo-vector-ilustracion-aislado-sobre-fondo-blanco>

La curva normal tiene otra característica y es que, el 99,74 % central de los casos, de ese 100,00% de los casos, se encuentra entre dos valores bien delimitados: los valores que resultan de restar y de sumar a la media $\mu=0$, tres veces el valor del desvío estándar $\sigma=1$.

Entonces, alrededor de la media y de manera equidistante a ella, el 99,74% central de los casos centrales se encuentran delimitados por los valores que resultan de restar y sumar 3 veces el desvío estándar ($\sigma=1$), al valor de la media ($\mu=0$).

Ejemplo: vamos a suponer que se ha administrado un test para evaluar la condición física a 120 sujetos, a través la administración de una Prueba y obteniendo un puntaje final donde, a mayor puntuación, mejor condición física. Suponemos también que la distribución de puntajes de la Prueba se distribuye siguiendo el modelo normal y mostró una Media $M=10$ puntos y un desvío estándar $s=3$ puntos.

La aplicación de la teoría de la curva normal estándar a este ejemplo de una distribución empírica concreta indicaría que es de esperarse que la mitad de los casos (60 casos) sean sujetos cuyas puntuaciones son menores o iguales a 10 puntos. Y los otros 60 casos, la otra mitad de los casos, es de esperarse que correspondan a sujetos que hayan obtenido puntuaciones iguales a 10 puntos o más altas.

Si la teoría dice que se debe esperar que casi todos los datos (el 99,74% central de los casos) corresponden a sujetos con puntuaciones entre los valores que resultan de restar y de

sumar al valor de la media M , tres veces el del desvío estándar s , observaremos que casi todos los sujetos, 119 de los 120 sujetos, obtendrían puntuaciones entre 1 y 19 puntos:

$$M - 3*s = 10 - 3*(3) = 10 - 9 = 1 \text{ punto}$$

$$M + 3*s = 10 + 3*(3) = 10 + 9 = 19 \text{ puntos}$$

El 99,74% central de 100,00% (120 casos) = 119 casos (regla de tres simple):

120 casos ----- 100,00%

¿? casos ----- 99,74%

$$\text{¿? casos} = 99,74 * 120 / 100$$

$$\text{¿? casos} = 11968,8 / 100$$

¿? casos = 119,69 es decir; 119 pues se trata de cantidad de sujetos, un número discreto.

Debajo de la curva normal, el área que queda delimitada alrededor de la media $\mu=0$ si le restamos y sumamos dos veces el valor del desvío estándar $\sigma=1$, es del 95,44% central de los casos. Entonces, alrededor de la media M y de manera equidistante a ella, el 95,44% de los casos centrales se encuentran delimitados por los valores que resultan de restar y sumar 2 veces el valor del desvío estándar s , al valor de la media M .

En teoría, como la curva normal es simétrica, es de esperarse que la mitad del 95,44% central de casos, es decir un 47,72% de los casos corresponda a sujetos cuyas puntuaciones registradas sean valores inmediatamente más bajos que la media y el otro 47,72% de los casos tenga valores inmediatamente por encima de la media.

Concepto de normalidad estadística

Se dice que un sujeto tiene un rendimiento estadísticamente normal si su puntuación cae dentro del rango de valores de la variable analizada, que queda delimitado entre la media y más/menos una vez el valor del desvío estándar.

El porcentaje de casos entre -1 desvío estándar a la media, y +1 desvío estándar a la media es del 68,26 % central de los casos; 34,13 % de los casos a ambos lados del valor central de la media.

Aplicando esta teoría de la Distribución probabilística Normal, a nuestro ejemplo, podemos decir que 81 sujetos son estadísticamente normales, aquellos que obtuvieron puntuaciones entre 7 y 13 puntos:

$$M - 1*s = 10 - 1*(3) = 10 - 3 = 7 \text{ puntos}$$

$$M + 1*s = 10 + 1*(3) = 10 + 3 = 13 \text{ puntos}$$

El 68,34% central de 100,00% (120 casos) =81 casos (regla de tres simple). En términos de Percentiles, la normalidad estadística se encuentra en el rango de valores de la variable comprendidos entre los Percentiles 16 y Percentil 84.

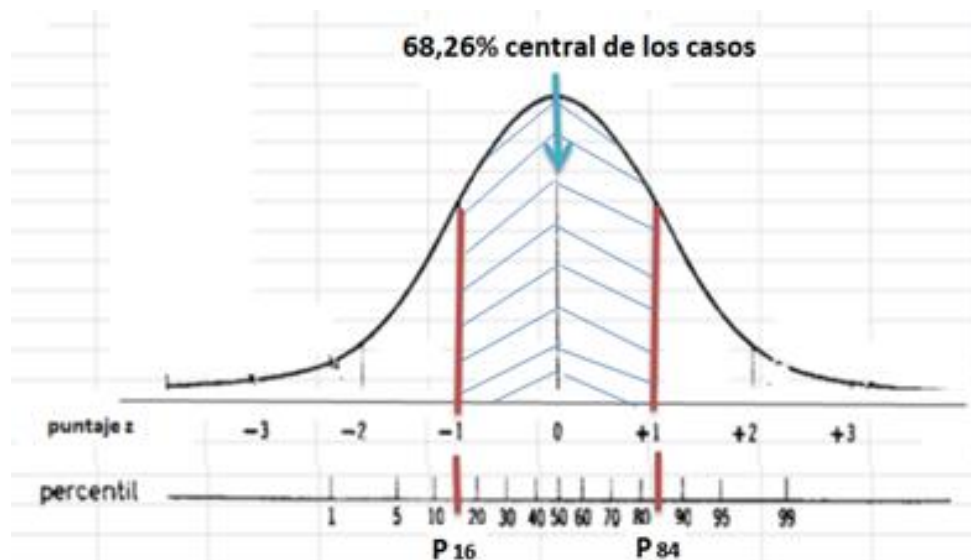
La probabilidad de encontrar casos desde $-\infty$ y hasta $z = -1$, es $p \leq 0,1587$. Equivale a decir que el 15,87% de los casos tienen puntuaciones iguales o inferiores a $z = -1$. Redondeando el valor del porcentaje y aplicando la definición de Percentil, un sujeto que se ubica en el Percentil 16, obtuvo una puntuación tal que se ubica, aproximadamente a puntaje $z = -1$.

De igual manera, la probabilidad de encontrar casos hasta $z = +1$ es $p \leq 0,8413$. Equivale a decir que el 84,13% de los casos tienen puntuaciones iguales a o inferiores a $z = +1$. Redondeando el valor del porcentaje y aplicando la definición de Percentil, un sujeto que se ubica en el Percentil 84 obtuvo una puntuación tal que se ubica, aproximadamente a $+1$ puntaje z .

La siguiente figura muestra lo expuesto:

Figura 6.4

Curva normal, zona de normalidad estadística



Nota. Fuente: elaboración propia

Es importante reflexionar sobre dos aspectos de este concepto.

Los sujetos estadísticamente normales representan el 68,26% de la distribución teórica. Por lo tanto, es de esperarse que los sujetos estadísticamente normales de una distribución empírica sean más de las dos terceras partes de los casos. Algo más de los $2/3$ de los casos (68,26%) es mucho más que la mitad (50,00%) de los casos.

El concepto de normalidad estadística no debe entenderse como opuesto al de patología.

La normalidad estadística no significa salud y, por lo tanto, no puede interpretarse en relación con ese concepto. “Normal” es un concepto asociado a la frecuencia con que aparecen los eventos: la gran mayoría de los casos, más de los dos tercios de los casos, se concentra

alrededor del valor central de la Media, alejándose de ella en la cantidad de un desvío estándar por abajo y por arriba del valor promedio de la Media.

Transformaciones de puntajes

El puntaje directo es la cantidad bruta, absoluta, de unidades en las que se mide una variable (Ejemplo: años, puntos, milisegundos, etc.) y presentan como valores de variable de cada unidad de análisis. El puntaje directo puede transformarse linealmente a puntuación z o no linealmente a percentil.

El puntaje z expresa el rendimiento de un sujeto respecto de los desvíos a la media aritmética, de la distribución de frecuencias, de una variable en estudio.

El percentil expresa el rendimiento de un sujeto según el porcentaje de casos que iguala o supera un sujeto, respecto de su grupo.

Ambas transformaciones se pueden realizar tanto manualmente como utilizando Excel o algún software estadístico. Cuando se realiza manualmente, para realizar el pasaje puntaje bruto a puntaje z , se utiliza la fórmula de puntaje z ; cuando se transforma puntaje bruto a Percentil, se utiliza una Tabla de probabilidades de la Curva Normal.

Puntuación z

Si la variable es Edad, la unidad de medida de la distribución de edades es “años”. Si la variable es tiempo de reacción, la unidad de medida de la distribución de tiempos de reacción puede ser “milisegundos”. Si la variable es Puntuaciones en un Test, la unidad de medida de la distribución de es “puntos”.

Dada la Curva Normal, ¿cuál es la unidad de medida de esa distribución de frecuencias teóricas de una variable aleatoria?: el puntaje z . Su fórmula es:

$$\text{pje. } z = \frac{x - M}{s}$$

Dada una distribución de frecuencias empíricas, x es el valor del puntaje bruto; M es el valor de la media aritmética; s : es el valor de la desviación estándar. Cada uno de los valores de una variable empírica puede transformarse en puntaje z . Para calcular el puntaje z , que le corresponde al rendimiento de un sujeto, se requieren conocer los valores de la media M y del desvío estándar s , de la distribución de valores de la variable x en estudio.

Esta transformación lineal del valor bruto de la variable, a puntaje z , permite describir cómo se ubican cada una de las observaciones en relación con el valor promedio registrado por el grupo. Al transformar el puntaje directo que obtuvo un sujeto, a puntaje z , estamos

transformando su rendimiento, a un sistema de puntuaciones equivalente que brinda la información de cómo se posiciona ese sujeto, con ese determinado rendimiento “x” específico, respecto del promedio M registrado por el grupo al cual pertenece.

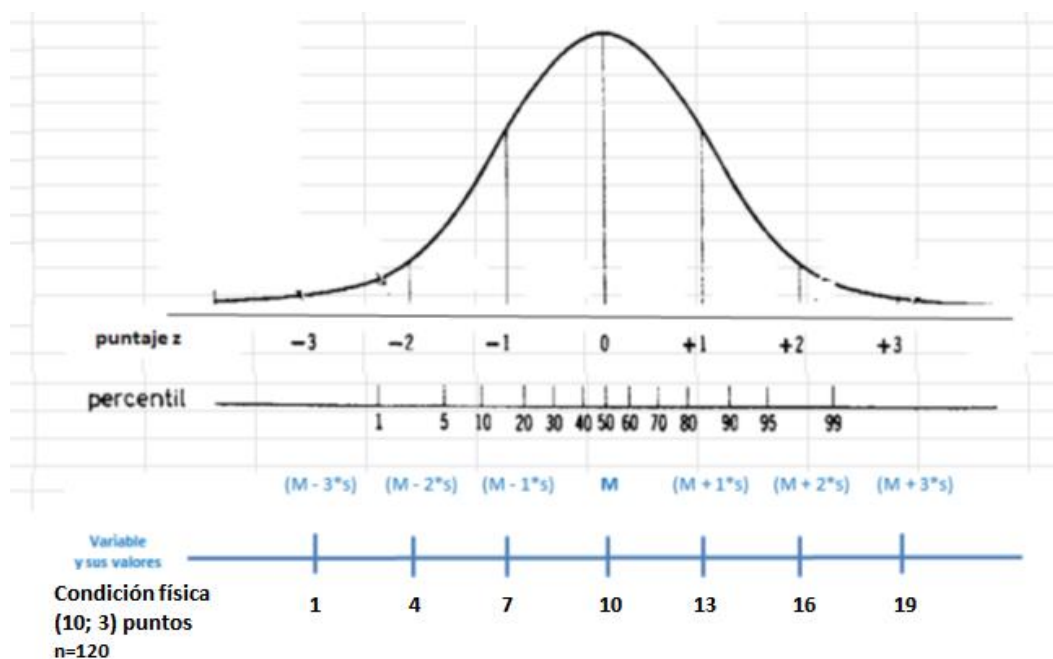
Un puntaje z es una puntuación que resulta de transformar linealmente, un valor observado de manera directa sobre un sujeto que conforma la muestra, a un sistema que traduce ese rendimiento individual al de posicionamientos respecto del valor promedio constatado en el grupo al que pertenece ese sujeto.

El puntaje z mide cuánto se aleja o se acerca el rendimiento puntual de un sujeto, al valor central promedio de una distribución de frecuencias de la muestra a la cual pertenece ese sujeto. Esa distancia se mide en términos de unidades de desvíos a la media.

Luego, cada puntuación z es equivalente a una distancia respecto del valor de la media aritmética, donde las distancias se expresan como partes de su desviación estándar (Lewin y Lewin, 2006, p 80).

Figura 6.5

Curva normal y transformación de puntajes



Nota. Fuente: elaboración propia

Adviértase que el puntaje z, es un valor de referencia respecto de la media aritmética.

Dado que el valor de la media de la Curva Normal Estándar es 0, cualquier valor de referencia que corresponda a un sujeto con rendimiento menor al promedio del grupo, se transforma a un valor de puntaje z menor a 0, es decir valor de puntaje z negativo.

De igual manera, si el rendimiento de un sujeto supera al rendimiento promedio del grupo al que pertenece, ello se verá reflejado en una transformación del puntaje directo obtenido, a

puntaje z que, al tener que reflejar un rendimiento mayor al promedio resultará un valor de puntaje z mayor a 0: un valor de pje. positivo.

Siguiendo con el ejemplo planteado, vamos a suponer que Juan y María son dos de los 120 sujetos testados. Que Juan obtuvo 9 puntos y María 12 puntos.

¿Cómo interpretar estos dos puntajes directos en relación con el rendimiento grupal?

Tabla 6.2

Ejemplo

Sujeto	Puntaje directo en el Test	Puntaje z
Juan	9 puntos	- 0,33
María	12 puntos	+ 0,67

Nota. Fuente: elaboración propia

Juan:

$$z = \frac{9 - 10}{3} = \frac{-1}{3} = -0,33$$

María:

$$z = \frac{12 - 10}{3} = \frac{2}{3} = +0,67$$

Tanto Juan como María tuvieron rendimientos normales en el Test de competencias físicas; sus puntuaciones se ubican dentro del rango de un desvío estándar por debajo y por arriba de la Media.

En la tabla que antecede se transformaron los puntajes obtenidos por cada uno de los sujetos, a puntaje z.

Se constata que Juan obtuvo un puntaje que lo ubica por debajo del promedio de su grupo, y María, por encima.

Juan: su puntuación z es -0,33. Ese valor de -0,33 tiene signo “-” negativo y significa que el rendimiento de Juan es más bajo que el promedio de su grupo pues el promedio del grupo es equivalente a un puntaje z=0.

El valor absoluto $|0,33|$ significa que es unas 0,33 partes del s, entendido este Desvío Estándar como un todo, como un entero, como la probabilidad total, como un 1. El valor

absoluto $|0.33|$ partes del entero 1 (que es el valor del s) es lo mismo que decir que es un tercio $1/3$ del s Desvío Estándar.

Aplicado a nuestro ejemplo donde el $s=3$ puntos, resulta que $1/3$ de 3 puntos, es 1 punto. Luego, aplicando la teoría de la curva normal, debemos tomar la Media (10 puntos) y restarle 0,33 (pues $z=-0,33$) partes del Desvío estándar ($s=3$ puntos y $1/3$ de $s=1$ punto). A partir de la puntuación z , la cantidad de puntos de Juan se obtendría de la siguiente manera: $10 (M) - 1 (1/3 \text{ del } s=3)= 9$ puntos.

Esa cantidad de puntos, 1 punto, es en este caso, la distancia que le “falta” a Juan para alcanzar el rendimiento típico promedio de su grupo. Y decimos “le falta”, pues el valor del pje. z es negativo, indicando que a Juan le faltan puntos para llegar al promedio de su grupo.

María: su puntuación z es 0,67. Ese valor de 0,67 tiene signo “+” y significa que el rendimiento de María es más alto que promedio. El valor absoluto $|0.67|$ significa –aplicando la teoría a la práctica- que María dista de la media en una 0,67 parte del s , entendido éste –en teoría- como un todo, como 1. $|0.67|$ partes de $1= s$, es $2/3$ del s . Si en nuestro ejemplo, el $s=3$ puntos, entonces $2/3$ de 3 puntos es: 2 puntos. Esa cantidad de puntos, 2 puntos, son los puntos que María tiene más que el promedio de su grupo; es la distancia de puntos en que María ha superado al rendimiento típico grupal.

Estandarización de una variable

Se llama normalización, estandarización o tipificación de una variable x , a su transformación de puntaje bruto, a puntaje z ; al hacerlo, estamos permitiendo la comparación de rendimientos de un sujeto, respecto del promedio de desempeños observados en su grupo o grupos de referencia.

Estandarizar o tipificar la variable significa pasar de puntaje bruto a puntaje z , cada uno de los valores de la variable x . Luego, al tipificar un variable se crea un sistema equivalente que permite la comparación de los rendimientos de varios sujetos entre sí respecto de lo esperado para su grupo, tanto como comparar los rendimientos expresados en distintas unidades de medida.

Al hacerlo, pasamos de medir cuántos puntos obtuvo un sujeto, a medir cuántas partes de lo que vale el desvío estándar de la distribución empírica, se aleja el rendimiento particular de un sujeto, del promedio de su grupo.

Nótese que puede alejarse en más o en menos, alejarse del promedio superándolo o, por el contrario, no alcanzándolo aún. Ello permite no sólo saber cómo se ubica en relación con el promedio, y comparar cuánto se alejan del promedio el rendimiento de varios sujetos, sino que permite comparar el rendimiento de un sujeto o varios, en diferentes Pruebas.

Al estandarizar las puntuaciones directas obtenidas en Pruebas con unidades de medidas diferentes (ejemplo: la variable tiempo de reacción medida en milisegundos, la variable capacidad aeróbica medida en volumen de oxígeno en sangre), se produce una transformación

lineal a un sistema de puntuaciones, puntuaciones z , que permite la comparación de rendimientos diversos, de diversos sujetos.

El puntaje z es el puntaje típico o estándar de referencia. “Típico” es aquello que representa la característica básica de algo. El valor típico de la distribución de una variable x es su Media aritmética.

El puntaje $z=0$ es el valor central equivalente al promedio de una distribución de frecuencias de una variable x . Todos los puntajes z expresan distancias a ese centro. Por ello también se dice centrar la variable al procedimiento de estandarización de la misma. Centrar la variable es expresar cada puntaje bruto en términos de cuántas partes del desvío estándar se “descentra” o aleja del centro de la distribución de frecuencias.

El concepto de estandarización de una variable se relaciona con el concepto de *baremos* de una variable.

Un baremo es una tabla que expresa de manera ordenada, la escala típica de los rendimientos esperados de un grupo respecto de una variable de interés, para cada región, según grupo de edad, y en una determinada época socio- histórica.

Percentiles

Recordamos que los percentiles son valores de la variable de orden, que se definen por el porcentaje de casos que deja por debajo de ese valor.

Constituyen una transformación no lineal del puntaje bruto a otro valor de la variable que está determinado por el porcentaje de casos que supera, el rendimiento particular de un sujeto respecto de su grupo.

Un Percentil es una medida de orden que expresa, en términos de porcentaje, el valor de la probabilidad p (p -valor) de encontrar sujetos que obtuvieron puntuaciones menores o iguales a la observada en un determinado valor de la variable. Se trata del porcentaje acumulado hasta ese valor de la variable expresado en el Percentil.

Ejemplo: Percentil 40 de la variable edad, indica la cantidad años tal que –quien la posea– tendrá una edad igual o mayor a la edad del 40% de los casos que –como ese mismos caso– integran la muestra.

Para transformar un puntaje directo a Percentil, utilizamos una Tabla (ver Anexo). Pero, en esta Tabla sólo se encuentran puntuaciones z y p -valores. Si los datos que tenemos son sólo los valores de la media (M) y del desvío estándar (s), debemos dar un paso previo al uso de la Tabla: transformar el puntaje directo a pje. z . Sólo entonces, con ese valor de pje. z , se puede obtener de la Tabla un determinado valor de probabilidad acumulada de encontrar en la muestra, puntajes iguales o menores al indicado. Si queremos expresar ese p -valor hallado en términos de porcentajes de casos, se multiplica el valor de probabilidad encontrados por 100%, y así se tiene el porcentaje de casos que se observan hasta ese determinado valor de la variable.

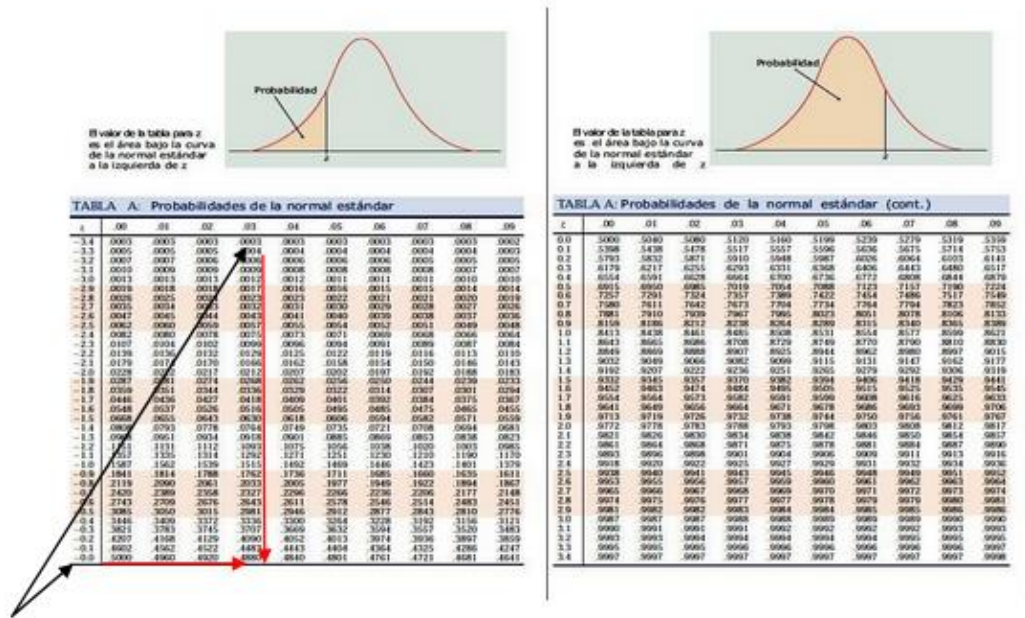
Por último, al redondear ese porcentaje de casos, tenemos el valor buscado del Percentil. La manera de redondear la decide el investigador en función de la variable en estudio.

El camino inverso se realiza si el dato que tenemos es de un determinado Percentil y debemos encontrar el puntaje asociado a él:

$z = -0,33 \rightarrow p\text{-valor} \leq 0,3707 \rightarrow 37,07\% \rightarrow \text{Percentil } 37$. Juan supera al 37% de su grupo.

Figura 6.6

Tabla de distribución de probabilidad normal

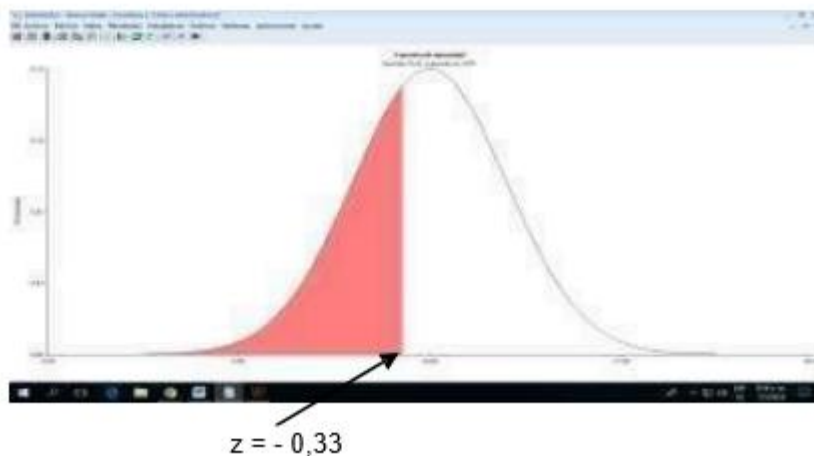


$z = -0,33$

Nota. Fuente: adaptada de <https://4geeks.com/es/lesson/tabla-normal-estandar>

Figura 6.7

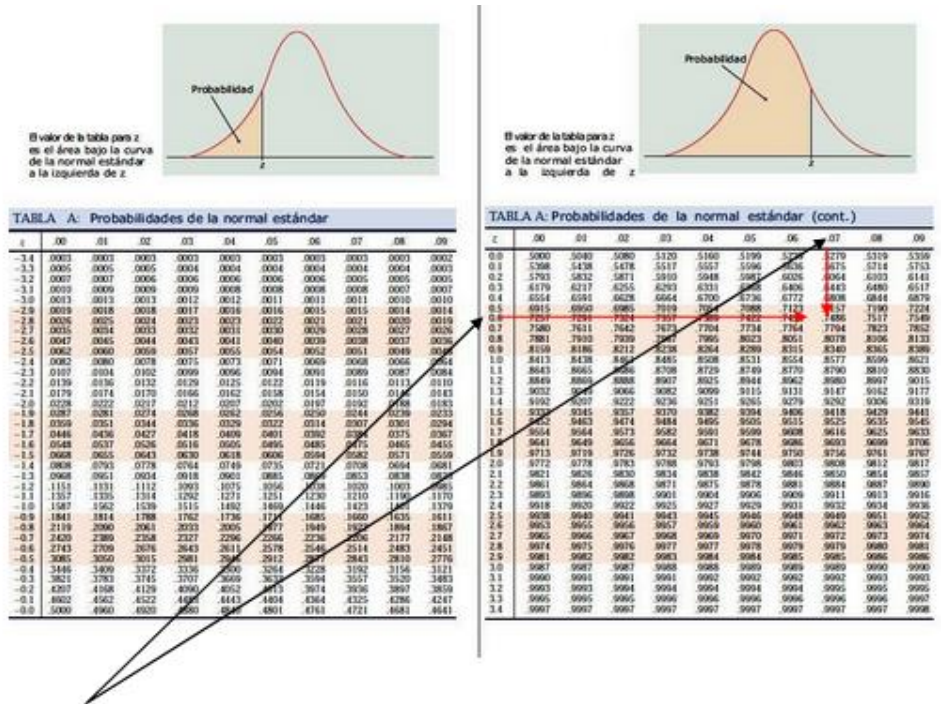
Área de la curva normal según probabilidad



Nota. Fuente: elaboración propia

Figura 6.8

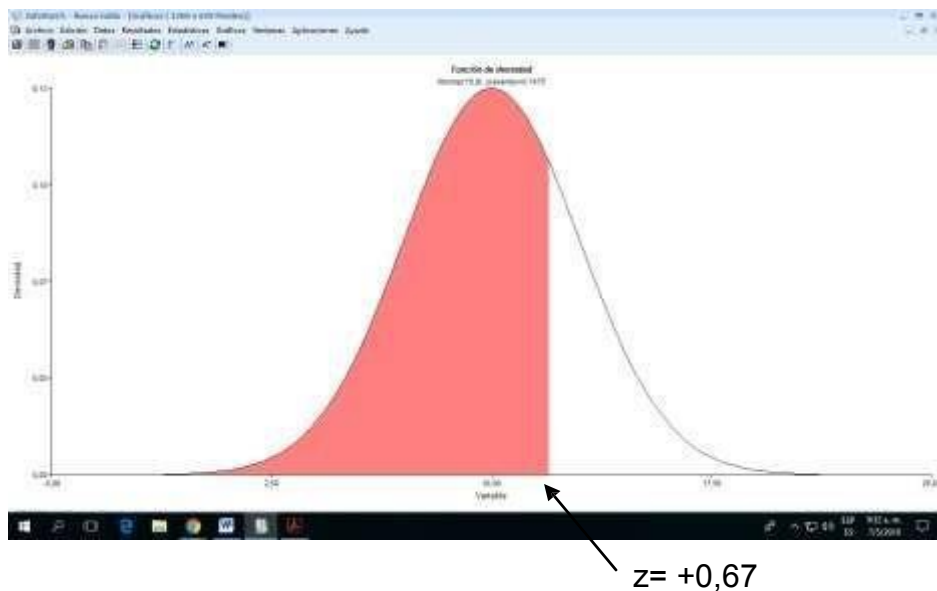
Tabla de distribución de probabilidad normal


 $z=0,67 \rightarrow p\text{-valor} \leq 0,7486 \rightarrow 74,86\% \text{ de los casos} \rightarrow \text{Percentil 74}$

Nota. Fuente: adaptada de <https://4geeks.com/es/lesson/tabla-normal-estandar>

Figura 6.9

Ejemplo de área de distribución de probabilidad normal



Nota. Fuente: elaboración propia

María se ubica en el P74. Significa que, con su puntuación, supera o es igual en rendimiento que el 74% de su grupo.

Por lo tanto, el rendimiento de un sujeto puede expresarse de distintas maneras, todas equivalentes. El puntaje bruto puede transformarse a **puntaje z** y a Percentil. La elección de una u otra forma depende de cómo se quiera comparar su rendimiento:

Puntaje directo: expresa de manera bruta, qué valor de la variable registró un sujeto (Ejemplo: cuántos años tiene un sujeto, cuántos puntos sacó en una Prueba, cuántos milisegundos tardó en responder a un ítem, etc.).

Puntaje z: transforma linealmente ese puntaje directo, a una escala que indica a cuántas partes del desvío estándar, un sujeto se aleja o acerca del promedio del grupo.

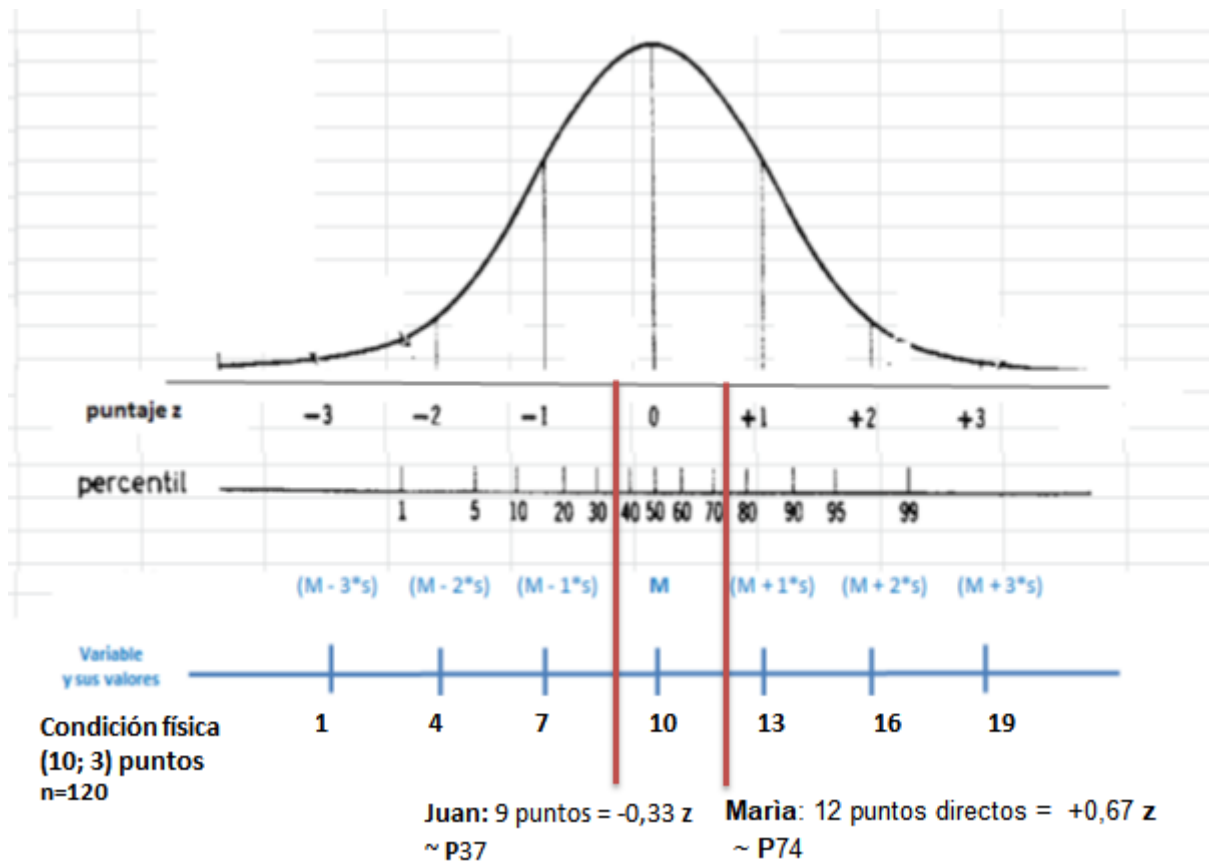
El **p-valor:** indica el valor de probabilidad p, acumulada hasta un determinado puntaje bruto. Señala la probabilidad de encontrar puntuaciones menores o iguales a un determinado puntaje bruto.

Percentiles: indica el porcentaje de casos que supera un sujeto al haber obtenido un determinado puntaje directo.

Resumiendo, el ejemplo de Juan y María:

Figura 6.9

Curva normal y transformación de puntajes



Nota. Fuente: elaboración propia.

Para expresar una probabilidad p , en términos de porcentaje y Percentil, hay que recordar las siguientes equivalencias:

Tabla 6.3

Transformación de puntajes

probabilidad p	porcentaje de casos	Percentil h
0 a 1	de 0,00% a 100,00%	h : de 1 a 99
Ejemplo:		
$p \leq 0,1963$	19,63 % de los casos	entre los P9 y P10

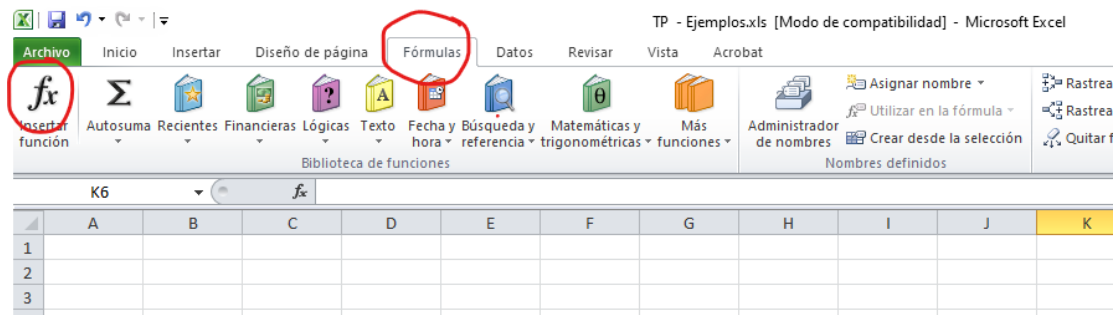
Nota. Fuente: elaboración propia

Uso de Excel

Al abrir un libro de Excel, las Fórmulas se insertan abriendo la solapa “Fórmulas”. En esa solapa, hay que cliquer sobre el ícono de “Insertar función”. Ahí se pueden buscar las fórmulas que necesitamos (recordar que al “Seleccionar categoría” esté puesta la opción, “Todo”, si no, sólo aparecen las últimas fórmulas usadas).

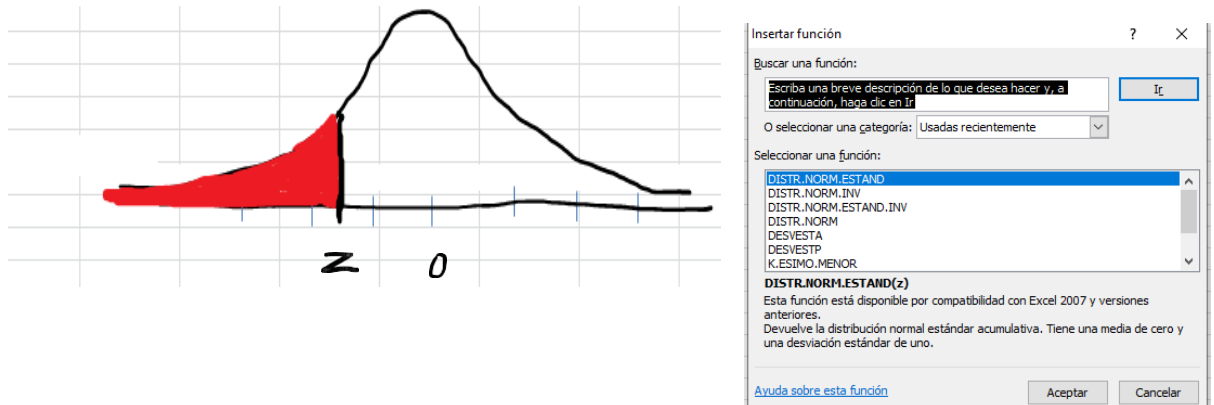
Figura 6.10

Procedimiento para cálculo de probabilidad curva normal en Excel



Nota. Fuente: elaboración propia.

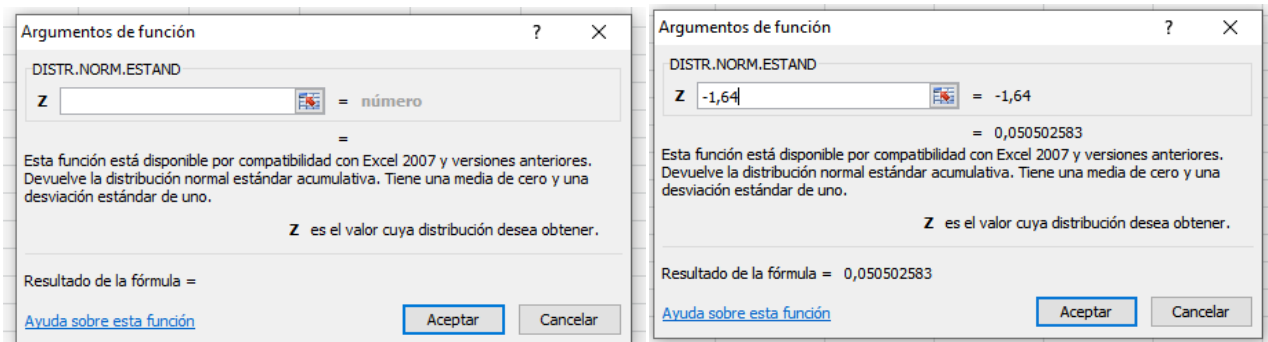
Con el $p_{je.z}$, obtengo la probabilidad acumulada hasta ese valor de puntaje z (el área pintada de rojo), con la fórmula en Excel: =DISTR.NORM.ESTAND(z)

Figura 6.11*Procedimiento para cálculo de probabilidad curva normal en Excel*

Nota. Fuente: elaboración propia.

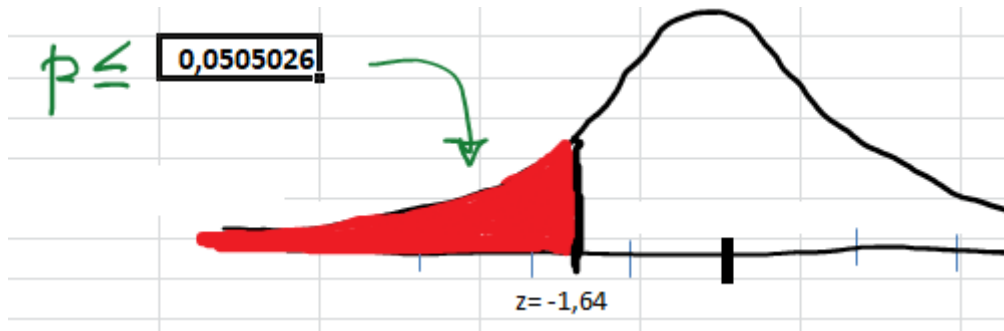
Figura 6.12*Funciones en Excel: puntaje z*

Ejemplo: si $z=-1,64$, el valor de probabilidad acumulada hasta $z=-1,64$ es...



Nota. Fuente: elaboración propia

Al dar “Aceptar”, aparece en una celda, el valor de probabilidad acumulada, hasta ese determinado valor de puntaje z . En este caso aparecerá el valor de $p=0,0505026$

Figura 6.13*Área bajo la curva normal*

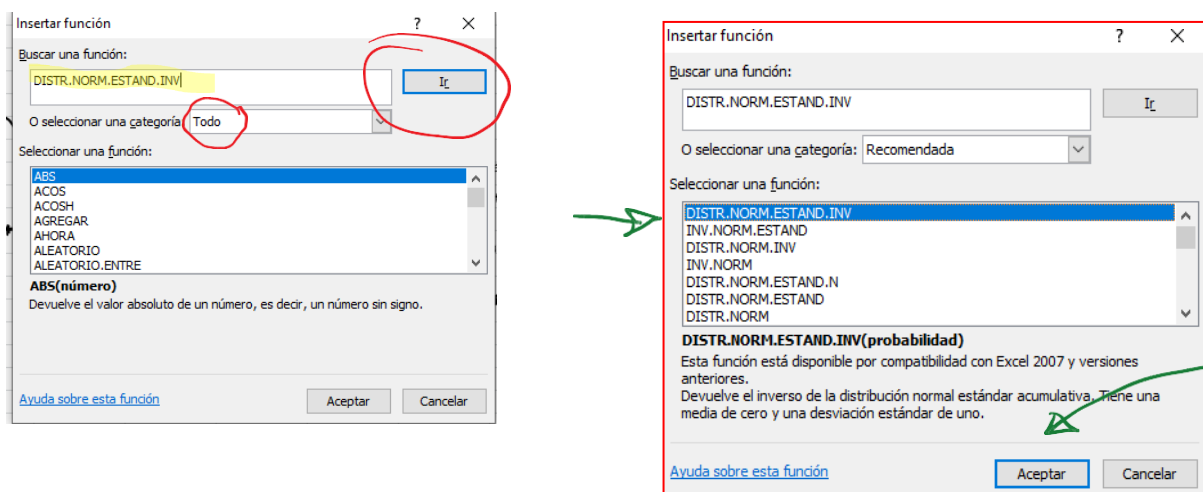
Nota. Fuente: elaboración propia.

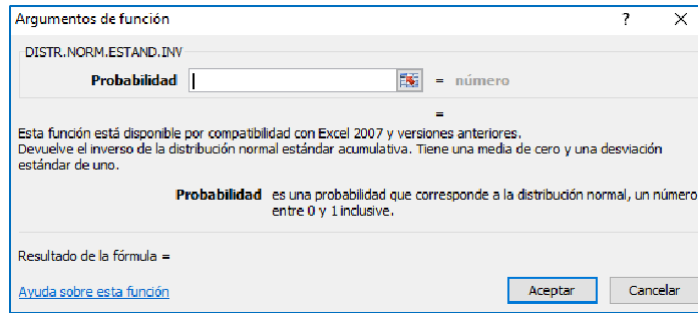
Para un $z = -1,64$, la probabilidad acumulada es de 0,0505 (con cuatro decimales alcanza para las probabilidades; dos decimales para los porcentajes).

Interpretación: significa que ese resultado obtenido está indicando que la probabilidad de encontrar puntuaciones que se ubican por debajo del valor promedio muestral, a 1,64 partes de desviaciones estándar, es baja, es una probabilidad menor o, a lo sumo igual que 0,0505.

Al revés si tengo la probabilidad p , puedo obtener el puntaje z con la siguiente fórmula:
`=DISTR.NORM.ESTAND.INV(p)`

En “Buscar una función”, se escribe la función. En “seleccionar categoría” debe estar la opción “Todo”. Luego, se cliquea sobre “Ir”, y aparece el cuadro con la función. Dar “Aceptar” y aparece la fórmula.

Figura 6.14*Cuadros de diálogo para cálculo de probabilidad curva normal en Excel*

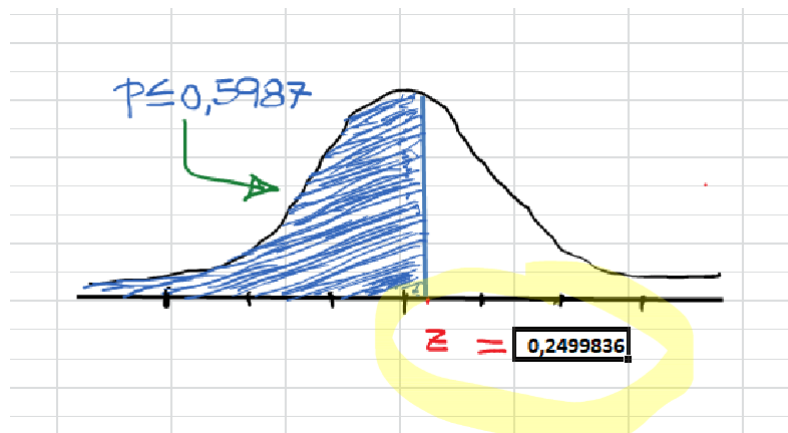


Nota. Fuente: elaboración propia.

Ejemplo: si tenemos una probabilidad acumulada de 0,5987 ¿a qué pje.z le corresponde? Escribimos en el sector “Probabilidad” de la ventana que se abrió. Al dar “Aceptar”, aparece una celda con el valor del pje. z que estábamos trabajando (podemos utilizar dos decimales).

Figura 6.15

Curva normal en Excel



Nota. Fuente: elaboración propia.

Interpretación: una probabilidad menor o igual a 0,5987, se corresponde con un valor máximo de puntaje $z=+0,25$

Es exactamente igual: en vez de usar Excel, usar la Tabla de áreas acumuladas (las interpretaciones son las mismas, pues se describe lo mismo).

Ejemplo: si $z=-1,64$, el valor de probabilidad acumulada hasta $z=-1,64$ es...

Para encontrarlo, el puntaje z, se desarticula en:

signo negativo: tomó los valores del lado izquierdo

el valor $z = -1,6$: lo encuentro en la primera columna

el valor $0,04$: lo encuentro en la primera fila, la fila superior

En el cruce de los valores $-1,6$ y 04 , se encuentra la probabilidad p buscada: 0,0505. Recordamos que este valor de probabilidad es equivalente a determinar el 5,05% de los casos en estudio; y la ubicación de ese rendimiento es en el Percentil 5.

Figura 6.16

Curva normal en la Tabla de probabilidades

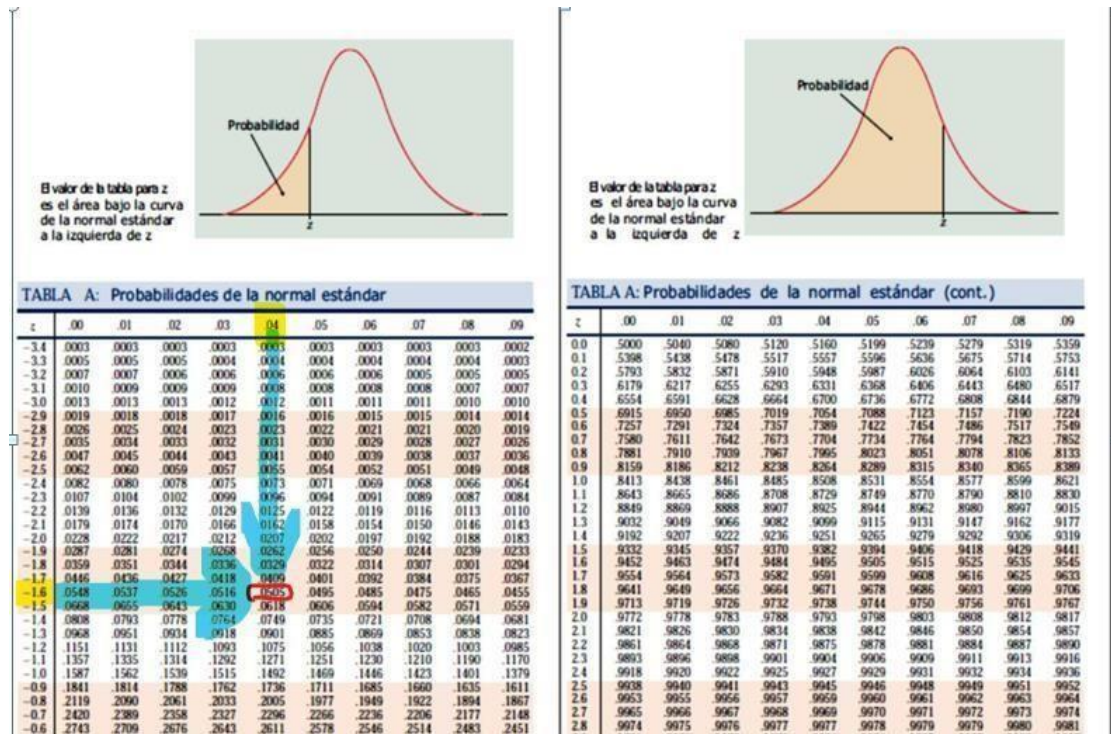
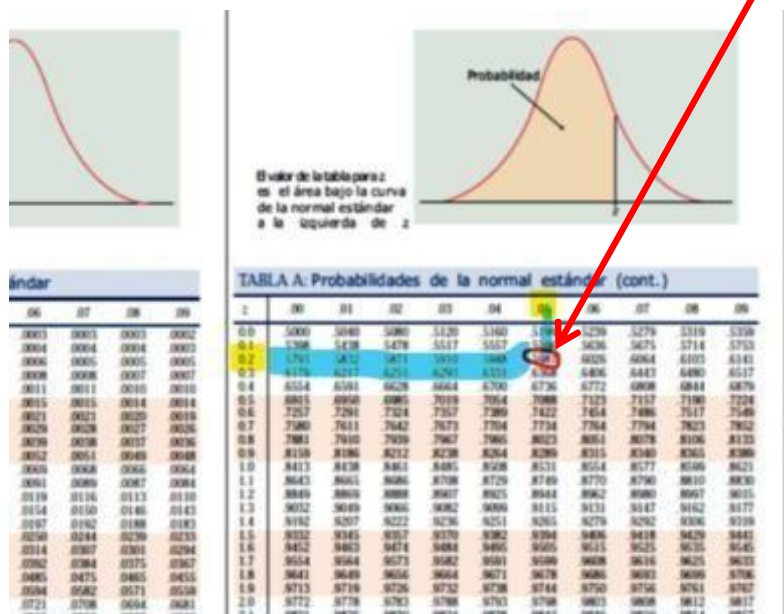
Nota. Fuente: adaptada de <https://4geeks.com/es/lesson/tabla-normal-estandar>

Figura 6.17

Curva normal en la Tabla de probabilidades



Ejemplo: si tenemos una probabilidad acumulada de 0,5987 ¿a qué puntaje z le corresponde?

Al ser un valor de probabilidad que supera el 0,50 indica que supera la probabilidad esperada hasta el valor promedio, es decir que el puntaje z que le corresponderá será positivo, mayor a puntaje $z=0$. Luego, hay que buscar dentro de la tabla, del lado derecho, el valor más parecido a 0,5987. Una vez encontrado, el valor del puntaje z que le corresponde, se obtiene componiendo el de la columna con la fila. Para este caso, el valor del puntaje z asociado a una probabilidad de 0,5987 es de: 0,2 (en la primera columna) agregado un ,05 de la primera fila; es decir: $z= 0,25$.

Referencias

- Aron, A., & Aron, E. (2003). Estadística para Psicología. (Cap. 2 y 5) Brasil: Prentice Hall.
- Cortada de Kohan, N. (1994). Diseño estadístico. Buenos Aires: EUDEBA
- Elorza Pérez Tejada, H. (2008). Estadística para las ciencias sociales, del comportamiento y la salud. Cap. 6 (pp. 211-231). México DF: Cengage Learning.
- Garrett, H. (1990). Estadística en Psicología y Educación. Cap. V. Buenos Aires: Paidós.
- Levin, J., & Levin, W. (2004). Fundamentos de Estadística en la Investigación Social. Cap. 6. México. Alfaomega. Oxford.
- Maureira Cid, F. (2017). Estadística básica para Educación Física. Estadística descriptiva e inferencial univariada. 2da. Ed., Cap. 6: pp.87-100. Editorial Académica Española.

Bibliografía de software estadístico

- Carrascal Arranz, U. (2007). Estadística descriptiva con Microsoft® Excel 2007. RA-MA, Madrid.
- InfoStat (2008). InfoStat, versión 2020. Manual del Usuario. Grupo InfoStat, FCA, Universidad Nacional de Córdoba. Primera Edición, Editorial Brujas Argentina.

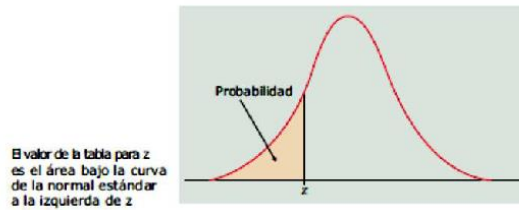
Anexo. Tabla de Distribución probabilística Curva Normal¹⁷

TABLA A: Probabilidades de la normal estándar

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
-3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
-3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
-3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
-3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
-2.9	.0019	.0018	.0018	.0017	.0016	.0015	.0015	.0014	.0014	.0014
-2.8	.0026	.0025	.0024	.0023	.0022	.0021	.0021	.0020	.0019	.0019
-2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
-2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
-2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
-2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
-2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
-2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
-2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
-2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
-1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
-1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
-1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
-1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
-1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
-1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
-1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
-1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
-1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
-1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
-0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
-0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
-0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
-0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
-0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
-0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
-0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
-0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
-0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
-0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

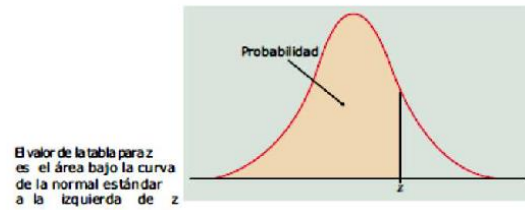


TABLA A: Probabilidades de la normal estándar (cont.)

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

¹⁷ Tabla disponible en: <https://www.studocu.com/es-ar/document/universidad-siglo-21/estadistica-ii/tabla-normal-acumulada/7382714>

CAPÍTULO 7

Asociación entre variables cualitativas

Martina Díaz y Mariela Cotignola

Introducción al análisis no paramétrico

En este capítulo abordamos las técnicas de análisis estadístico no paramétrico, centrándonos en las relaciones entre variables cualitativas medidas en escalas nominales u ordinales.

Cómo hemos visto, las herramientas de análisis estadístico se diferencian en dos grandes tipos: las paramétricas y no paramétricas.

Las pruebas estadísticas paramétricas requieren conocimiento previo sobre la distribución de los datos y que estos cumplan los supuestos de normalidad y homogeneidad de varianzas.

El análisis estadístico no paramétrico es un conjunto de técnicas estadísticas que no requiere asumir que los datos sigan una distribución específica, como la normal. Es especialmente útil cuando los datos que observamos no cumplen con los supuestos necesarios para aplicar métodos estadísticos paramétricos convencionales. Estas técnicas son particularmente relevantes en situaciones donde se observa que los datos se desvían significativamente de la distribución normal, siendo apropiados para aplicar cuando analizamos variables categóricas, ordinales y datos de muestras pequeñas.

Una de las principales fortalezas del análisis no paramétrico es su capacidad para abordar datos categóricos u ordinales, permitiendo evaluar relaciones y asociaciones sin la necesidad de medir diferencias precisas entre categorías. Asimismo, estos métodos son robustos frente a la presencia de valores extremos o atípicos, siendo menos sensibles a tales anomalías que los métodos paramétricos. La inferencia en análisis no paramétricos se basa frecuentemente en técnicas de ordenamiento, rangos o conteos de frecuencia, lo que simplifica su interpretación y facilita una comprensión directa y práctica de los resultados.

Además, el análisis no paramétrico ofrece versatilidad y adaptabilidad, siendo aplicable a diversas disciplinas como ciencias sociales, ciencias de la salud, educación física, deporte y actividades recreativas, donde es común encontrar datos que no se ajustan estrictamente a supuestos paramétricos. A pesar de estas ventajas, los métodos no paramétricos suelen tener una potencia estadística ligeramente menor que sus equivalentes paramétricos cuando se cumplen los supuestos necesarios para estos últimos, requiriendo a veces tamaños de muestra más grandes para alcanzar niveles comparables de significancia.

Es importante destacar que en una misma investigación pueden llevarse a cabo análisis paramétricos para algunas hipótesis y variables, y análisis no paramétricos para otras.

Análisis bivariado: asociación entre variables cualitativas

Muchos de los problemas en la investigación social suponen analizar la relación entre dos o más variables:

(...) podemos hablar de relación entre variables, cuando en un mismo conjunto de individuos (de unidades de análisis) se observa un comportamiento sincrónico o coordinado en el comportamiento de las mismas (al cambiar los valores de una variable cambian al mismo tiempo y de manera determinada, los valores de la otra). (Aguirre, Niño y Simonetti, 2006, p. 106)

Al iniciar el análisis bivariado debemos considerar en primer término dos aspectos:

- La naturaleza¹⁸ de la relación entre las variables.
- El tipo de variables que se están analizando.

En primer lugar, al establecer qué tipo de relación se desea estudiar entre variables, es necesario precisar su naturaleza, pudiendo tratarse de diferentes casos según Barbancho (1978, citado por Aguirre, Niño y Simonetti, 2005):

- Dependencia causal unilateral, donde una variable influye sobre otra sin que la segunda variable influya sobre la primera. Ejemplos de esto son la influencia del nivel educativo sobre las preferencias de lectura, o el impacto del nivel de ingresos en la elección del lugar de alojamiento.
- Interdependencia, caracterizada por la existencia de una influencia recíproca o causalidad bilateral. Ejemplos claros son la relación entre el precio de un producto y la cantidad demandada, la relación entre la postura sobre el aborto y la afiliación política, o la elección del destino vacacional en relación con el medio de transporte elegido.
- Dependencia indirecta, que ocurre cuando dos variables están relacionadas a través de una tercera variable que afecta simultáneamente a ambas. Por ejemplo, la tasa de natalidad y el consumo de proteínas animales están vinculados indirectamente por el nivel de vida; o la cantidad de accidentes de tránsito y la cantidad de semáforos, cuya relación indirecta se explica mediante la concentración urbana.
- Covariación casual, que corresponde a dos variables cuyo comportamiento coincide temporalmente, pero cuya asociación es meramente casual o accidental, ya que no se identifican factores directos ni indirectos que expliquen dicha relación observada.

¹⁸ Al decir “la naturaleza” nos referimos al tipo de variable y su nivel de medición.

En segundo lugar, el tipo de variables que se están analizando determinará las herramientas estadísticas que deben ser utilizadas.

Así tenemos que:

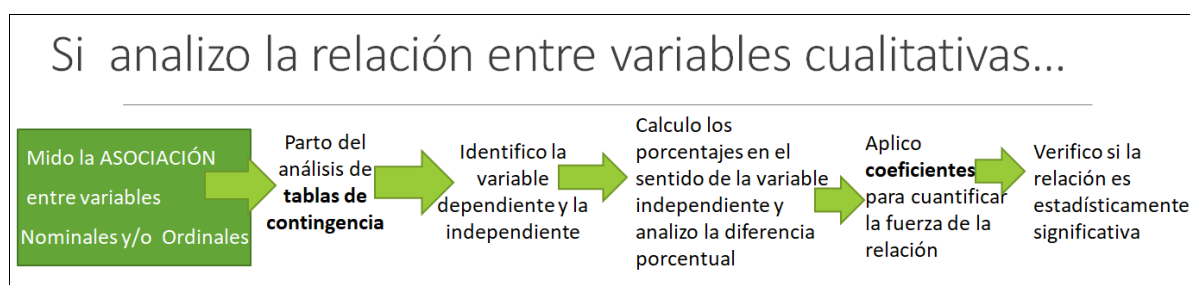
Si se trata de...	Recurrirnos a...
Dos variables categóricas	Tablas de contingencia
Una variable numérica y una categórica.....	Comparación de medias entre grupos
Dos variables numéricas.....	Análisis de Correlación

En términos metodológicos, se designa con el nombre de asociación a la relación que se da entre dos variables de naturaleza cualitativa nominal u ordinal, y con el nombre de correlación la relación que se da entre dos variables de naturaleza intervalar o racional.

La siguiente figura resume la secuencia a tener en cuenta al momento de describir la relación entre atributos:

Figura 7.1

Asociación entre variables



Nota. Fuente: elaboración propia

Análisis de contingencia

¿Cómo se presentan los datos para describir simultáneamente a las unidades de análisis a partir de variables categóricas? Los datos se organizan en una Tabla de Contingencia.

Se define qué es una tabla de contingencia y se señalan sus partes.

La tabla de contingencia consiste en un cierto número de celdas en las que, como resultado de un proceso de tabulación, realizado de forma manual, mecánica o electrónica, se han volcado las frecuencias (número de casos) correspondientes a cada combinación de valores de al menos dos variables.

La tabla de contingencia consiste en un cierto número de celdas en las que, como resultado de un proceso de tabulación, realizado de forma manual o con algún programa estadístico, se

volcaron las frecuencias (la cantidad de casos) correspondientes a cada combinación de valores de, por lo menos, dos variables.

La tabla de contingencias es el resultado del cruce (tabulación simultánea) de dos o más variables. Se aplica al cruce de variables de nivel de medición nominal y ordinal.

Su ejemplo más sencillo es la tabla bivariada, también llamada de clasificación cruzada o tabulación cruzada.

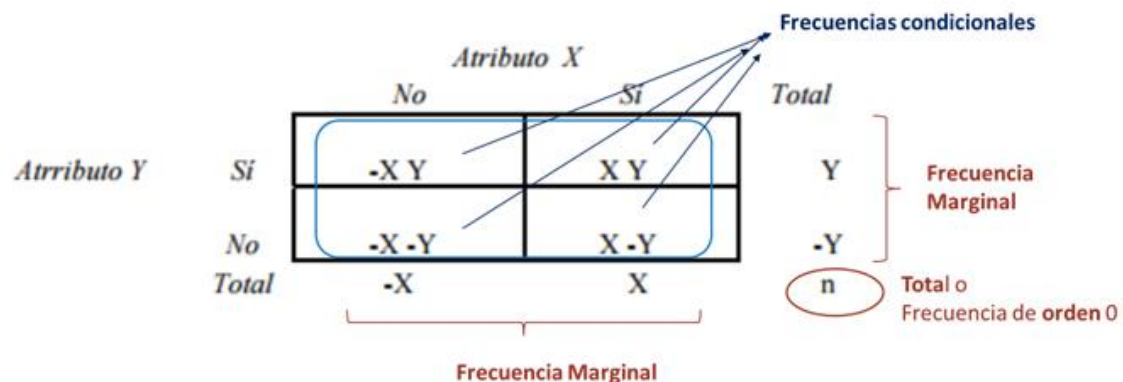
En el cuerpo de la tabla se presenta la distribución conjunta que da cuenta del número de unidades de análisis que presenta cada una de las combinaciones posibles de las categorías de las variables.

Los marginales de la tabla corresponden a la última fila y la última columna de la tabla que, encabezada por la palabra “total”, presentan la distribución univariada según cada variable.

Se presenta de forma gráfica¹⁹ una tabla bivariada, sus características y componentes:

Figura 7.2

Tabla de contingencia



Nota. Fuente: elaboración propia

A continuación, se presenta un ejemplo de análisis bivariado. La tabla 7.1 muestra la cantidad de nacimientos distribuidos según el peso al nacer según condición de fumadora de la madre.

Tabla 7.1

Bajo peso al nacer	Condición de fumadora		Total
	Fumadora	No Fumadora	
Si	51	128	179

¹⁹ Figura básica tomada de Baranger (1999, p.44) y enriquecida a los fines de los contenidos del actual capítulo.

No	269	2.052	2.321
Total	320	2.180	2.500

Nota. Fuente: elaboración propia

El cuadro muestra los valores absolutos, es decir el número de casos que cumplen con las distintas características. A modo de ejemplo se ensayan lecturas de las frecuencias absolutas condicionales y marginales del cuadro:

- Son 51 los nacimientos de niños con bajo peso cuyas madres son fumadoras.
- 320 son los nacimientos de niños, de madres fumadoras.
- 179 niños nacieron con bajo peso y 2.321 con peso normal.
- Se analizaron 2.500 nacimientos.

Hasta aquí se describieron los valores absolutos. Pero la necesidad de comparar y sacar conclusiones nos lleva al cálculo de los porcentajes.

¿Para qué sirven los porcentajes?

- Verificar diferencias importantes entre categorías de las variables/entre subpoblaciones.
- Para considerar el peso relativo de cada grupo respecto del total.
- Determinar si existe alguna relación.
- Confirmar o rechazar hipótesis.

Pero ¿cómo calcular los porcentajes? ¿sobre qué total los calculamos?

Existen tres posibles totales de referencia: Total de fila, total de columna y total general o frecuencia de orden 0.

Según el tipo de lectura que quiera realizar sobre los datos será la forma de cálculo de los porcentajes que corresponda. Se pueden realizar dos tipos de análisis sobre la tabla de contingencia:

- Análisis simétrico
- Análisis asimétrico

Análisis simétricos: características

- En el análisis simétrico no se presupone que una variable sea causa de la otra. No es apto para examinar la existencia de una relación de dependencia entre las dos variables. No estandariza valores
- Los porcentajes de las frecuencias marginales y condicionales se computan sobre la misma base, el total de casos.

Se opta por este tipo de análisis cuando no hay interés en indagar acerca del presunto “efecto” de una variable sobre la otra.

- Se utiliza a veces con el propósito de estimar o describir la proporción de personas/unidades de análisis de cada categoría de la población. Necesita que la muestra sea representativa.

Tabla 7.2

Lectura simétrica. Porcentajes calculados sobre el total de casos.

Bajo peso al nacer	Condición de fumadora		Total
	Fumadora	No Fumadora	
Si	2%	5%	7%
No	11%	82%	93%
Total	13%	87%	100%

Nota. Fuente: elaboración propia

Ejemplos de lectura de los porcentajes.

- El 7% de los niños analizados nacieron con bajo peso.
- El 82% de los niños nacieron con peso normal y sus madres no son fumadoras.
- El 2% de los niños nacieron con bajo peso y sus madres son fumadoras.
- El 87% de las madres de los niños analizados no son fumadoras.

Análisis asimétrico

Se plantea siempre que se decide considerar analíticamente que una variable -la independiente- incide sobre la distribución de la otra-la variable dependiente. Requiere identificar la variable independiente y la dependiente y enunciar una hipótesis sobre la relación entre ellas.

Ejemplo: ¿El peso al nacer, está determinado por la condición de que la madre del/la bebé sea fumadora? ¿Tiene relevancia la condición de ser una madre fumadora en el peso de los/as bebés que nacen?

Hipótesis: Los hijos de madres fumadoras tienen mayor probabilidad de nacer con bajo peso.

¿Cómo deben calcularse los porcentajes para medir la existencia de relación entre las variables?

Para responder a la pregunta se hace referencia a la primera regla de Zeisel, la regla de causa-efecto (Baranger, 1999): los porcentajes se calculan en el sentido de la variable independiente y se comparan en el sentido de la variable dependiente.

Otra consideración o advertencia en el modo de calcular los porcentajes se trabaja con la segunda regla de Zeisel (Baranger, 1999): cuando un conjunto de marginales no es

representativo de la población, los porcentajes deben computarse en la dirección en que la muestra no es representativa.

Cuando se observa mediante el análisis asimétrico de una tabla que una de las variables aparece determinando o afectando a la otra, podemos decir que ambas variables están asociadas.

La determinación o el condicionamiento indica la probabilidad diferencial de una unidad de análisis de pertenecer a un grupo u otro.

¿En qué medida se pone en evidencia la asociación entre las variables?

La medida de asociación más básica y frecuentemente utilizada es la diferencia porcentual; cuando se trata de información surgida de muestras se plantea el problema adicional de determinar la significación estadística que se le puede prestar a una asociación entre variables.

Como se señaló anteriormente para este tipo de análisis los porcentajes se computan en dirección a la variable independiente y se comparan en dirección inversa: la diferencia porcentual se calcula en la dirección en que se realiza la comparación.

En el ejemplo seguido en la explicación se observa que la diferencia porcentual es de 10 puntos porcentuales.

Tabla 7.3

Distribución de las frecuencias porcentuales de las variables Bajo peso al nacer y Condición de fumadora de la madre

Bajo peso al nacer	Condición de fumadora		Diferencia en %
	Fumadora	No Fumadora	
Si	16%	6%	10%
No	84%	94%	-10%
Total	100%	100%	

Nota. Fuente: elaboración propia

Ahora bien, ¿esta diferencia es alta, media o baja? ¿Pone en evidencia que hay relación entre las variables?

Para responder en primera instancia nos quedamos con que si la muestra es suficientemente grande a partir de 5 puntos porcentuales podemos decir que las variables están relacionadas (levemente).

Responder a estas preguntas implica varias consideraciones:

Se debe valorar el tamaño de la muestra o de la población en estudio.

Considerar el coeficiente de asociación que mejor se ajuste a la naturaleza de los datos relevados para poder medir la fuerza de la relación.

Aplicar un test de hipótesis que permita confirmar si la relación evidenciada es estadísticamente significativa.

Conceptos básicos de la estadística cualitativa o de atributos

Independencia y asociación

Cuando analizamos variables cualitativas hablamos de asociación entre atributos categoriales. La independencia implica que la ocurrencia de un evento no influye en otro. Por ejemplo, la preferencia por un deporte no influye en la edad del participante. Por el contrario, la asociación indica que existe una relación entre eventos o variables, como la relación entre nivel de actividad física y salud cardiovascular.

Intuitivamente, la independencia entre eventos implica que la probabilidad de que uno ocurra es igual independientemente de la ocurrencia del otro evento. Cuando decimos que la variable x no tiene efectos sobre la variable y , indicamos que y sucede tanto si x está presente como si no lo está. La independencia de dos variables equivale a que no haya asociación entre ellas.

Análisis de relaciones entre variables

Para evaluar la relación entre variables cualitativas nominales u ordinales se utiliza la tabla de contingencia, una herramienta esencial para examinar y cuantificar asociaciones entre variables. Inicialmente, se elabora una tabla de doble entrada (por ejemplo, 2×2) para variables dicotómicas.

Coeficiente de Asociación Q de Kendall (Q de Yule)

Utilizado para variables dicotómicas (dos categorías). Varía entre -1 y 1. Se interpreta por su valor absoluto, siendo valores cercanos a 1 una asociación fuerte y cercanos a 0 una asociación débil. Ejemplo: medir la asociación entre interés en practicar fútbol y género.

Donde A, B, C, D son frecuencias de la tabla de contingencia.

El signo (+ o -) del coeficiente Q no se interpreta, debido a que depende de cómo se hayan ordenado las categorías en la tabla.

El coeficiente Q de Kendall es apropiado para variables dicotómicas (dos categorías). Su fórmula es:

$$Q \text{ de Kendall} = \frac{A*D - B*C}{A*D + B*C}$$

Donde A, B, C, D corresponden a las frecuencias observadas en una tabla de contingencia 2x2.

Variable x	Variable y	
	valor de variable y ₁	valor de variable y ₂
Valor de variable x ₁	A	B
Valor de variable x ₂	C	D

Por ejemplo, supongamos que queremos evaluar la asociación entre el género (hombres/mujeres) y la preferencia por deportes colectivos (sí/no):

Deportes colectivos	Género	
	Hombres	Mujeres
Sí	40	20
No	10	30

Aplicando la fórmula:

$$Q \text{ de Kendall} = \frac{A*D - B*C}{A*D + B*C}$$

$$Q \text{ de Kendall} = \frac{(40 * 30) - (20 * 10)}{(40 * 30) + (20 * 10)}$$

$$Q \text{ de Kendall} = \frac{(1200) - (200)}{(1200) + (200)}$$

$$Q \text{ de Kendall} = \frac{1000}{1400} = 0,71$$

El valor resultante del coeficiente Q de Kendall es 0.71 indica una asociación fuerte entre ser hombre y la preferencia por deportes colectivos.

Coeficiente por Rangos rho de Spearman (ρ)

El coeficiente por rangos de Spearman (ρ) mide la relación entre variables ordinales. Se calcula asignando rangos a las observaciones y resolviendo empates mediante promedios. Su fórmula es:

$$\rho = \frac{6 * \sum d^2}{(N+1) * N * (N-1)}$$

Donde d es la diferencia entre los rangos asignados a cada par de observaciones y N el número total de casos. Utilizado para medir la relación entre variables ordinales. Se asigna un rango a cada observación, resolviendo empates mediante el promedio de posiciones.

Donde:

6 es un valor constante

d es la diferencia entre rangos

N es el número total de observaciones

Ejemplo: Evaluar la asociación entre niveles subjetivos de esfuerzo percibido y clasificación final en una carrera deportiva, obteniendo un coeficiente ρ que indique el grado de asociación.

El Coeficiente por Rango rho de Spearman, es una medida no paramétrica que mide la fuerza y dirección de la asociación entre dos variables ordinales o cuantitativas que no cumplen con los supuestos para usar el coeficiente de correlación de Pearson.

Su valor oscila entre **-1 y +1**, indicando:

+1: correlación positiva perfecta (cuando una aumenta, la otra también).

-1: correlación negativa perfecta (cuando una aumenta, la otra disminuye).

0: ausencia de correlación.

Es particularmente útil cuando:

- Las variables están medidas en escala ordinal (posiciones, clasificaciones)
- Las variables no cumplen los requisitos de distribución normal
- Existen observaciones extremas (valores atípicos)

Por ejemplo, se puede aplicar en situaciones como:

- Relación entre la posición en un ranking competitivo y la frecuencia semanal de entrenamiento.
- Relación entre posición en una carrera y tiempos de reacción ante estímulos

Procedimiento paso a paso para aplicar el Coeficiente de Asociación rho de Spearman

El procedimiento para la aplicación del coeficiente Rho de Spearman comienza con la definición de las variables involucradas y la formulación de la pregunta de investigación. A partir de los datos recolectados se asignan rangos ordenados a cada variable, asignando promedios en caso de empates. Posteriormente, se calculan las diferencias entre los rangos asignados a cada par de observaciones, elevándose estas diferencias al cuadrado para obtener una suma total. Finalmente, se aplica la fórmula del coeficiente Rho de Spearman, interpretando el resultado según su cercanía a +1 (asociación positiva fuerte), -1 (asociación negativa fuerte) o 0 (ausencia de asociación).

Paso 1: Plantear la pregunta o hipótesis de investigación

Por ejemplo: ¿Existe relación entre la posición en una carrera de natación y la cantidad semanal de horas entrenadas por deportista?

Paso 2: Recolectar datos (variables ordinales o numéricas)

Imaginemos un ejemplo con 8 nadadores/as:

Deportista	Posición en carrera	Horas de entrenamiento semanal
A	1	12
B	2	10
C	3	8
D	4	6
E	5	7
F	6	5
G	7	3
H	8	4

Paso 3: Asignar rangos a cada variable

Para calcular Spearman, convertimos los datos en rangos (orden). Rangos de posición (ya están en orden ascendente):

Deportista	Posición carrera	Rango posición	Horas entrenamiento	Rango horas entrenamiento
A	1	1	12	1
B	2	2	10	2
C	3	3	8	3
D	4	4	6	5
E	5	5	7	4
F	6	6	5	6
G	7	7	3	8
H	8	8	4	7

Aclaración importante: en caso de empates, asignar el promedio de rangos.

Paso 4: Calcular diferencias entre rangos (d) y diferencias al cuadrado (d^2)

Deportista	Rango posición	Rango horas entrenamiento	d (diferencia entre rangos)	d^2 (cuadrado de la diferencia entre rangos)
A	1	1	0	0
B	2	2	0	0
C	3	3	0	0

D	4	5	-1	1
E	5	4	1	1
F	6	6	0	0
G	7	8	-1	1
H	8	7	1	1
				Suma $\sum d^2 = 4$

Paso 5: Aplicar fórmula de Spearman

Figura 7.4

Aplicación del Coeficiente rho de Spearman

La fórmula de Rho de Spearman es:

$$\rho = 1 - \frac{6 \sum d^2}{n(n^2 - 1)}$$

Siendo:

- $\sum d^2$: Suma de diferencias al cuadrado.
- n : Número de casos (8 deportistas en este ejemplo).

Reemplazamos valores:

$$\rho = 1 - \frac{6 \times 4}{8(8^2 - 1)} = 1 - \frac{24}{8(64 - 1)} = 1 - \frac{24}{8 \times 63} = 1 - \frac{24}{504} = 1 - 0.0476 = 0.9524$$

Nota. Fuente: elaboración propia

Paso 6: Interpretar el resultado

$\rho=0,9524 \rightarrow$ **Correlación positiva muy fuerte.**

Significa que, a mayor cantidad de horas de entrenamiento semanal, mejor posición en carrera.

Otros ejemplos posibles en actividad física y deporte:

Ejemplo 1: Asociación entre percepción subjetiva del esfuerzo (escala ordinal 1-10) y frecuencia cardíaca registrada en deportistas durante un entrenamiento.

Ejemplo 2: Relación entre el ranking nacional en tenis (posición ordinal) y cantidad de partidos ganados durante el último año (variable numérica).

Prueba Estadística Chi-Cuadrado (χ^2)

Permite determinar si existe independencia o asociación entre variables cualitativas nominales u ordinales.

Hipótesis nula (H_0): Las variables x e y, son independencia entre sí.

Hipótesis alternativa (H_1): Las variables x e y, no son independencia entre sí.

Se compara la frecuencia observada (f_o) con la frecuencia esperada (f_e):

El valor χ^2 se calcula mediante:

Grados de libertad (gl) se calculan mediante:

Ejemplo: Analizar si existe asociación entre la pertenencia a un estrato social y la preferencia deportiva (fútbol, atletismo, tenis). Si el valor χ^2 calculado supera al crítico (obtenido de tablas estadísticas), se rechaza la hipótesis nula de independencia.

Una de las técnicas más comunes del análisis no paramétrico es la prueba de Chi-cuadrado (χ^2), que permite determinar si existe una relación significativa entre dos variables categóricas. Para aplicar el test de Chi-cuadrado, es necesario seguir un procedimiento claro y estructurado.

En primer lugar, se deben establecer las hipótesis del estudio:

Hipótesis nula (H_0): establece que no existe asociación entre las variables.

Hipótesis alternativa (H_1): establece que sí existe una asociación entre las variables.

Por ejemplo, si deseamos investigar si existe relación entre el nivel de actividad física (alto, medio, bajo) y el género (hombres, mujeres), la hipótesis nula sería que el nivel de actividad física es independiente del género, mientras que la hipótesis alternativa sería que existe alguna asociación entre estas dos variables.

En segundo lugar, se construye una tabla de contingencia donde se ubican las frecuencias observadas (f_o) de las combinaciones de categorías de ambas variables.

Por ejemplo:

Nivel de Actividad Física	Hombres	Mujeres	Total
Alto	30	40	70
Medio	50	60	110
Bajo	20	30	50
Total	100	130	230

Luego, se calculan las frecuencias esperadas (f_e) para cada celda mediante la fórmula:

$$f_e = (\text{total fila} \times \text{total columna}) / \text{total general}$$

Por ejemplo, para la celda de "Nivel Alto - Hombres": $f_e = (70 \times 100) / 230 \approx 30.43$

Este cálculo se realiza para todas las celdas.

El siguiente paso consiste en calcular el valor Chi-cuadrado (χ^2), utilizando la siguiente fórmula:

$$\chi^2 = \sum ((f_o - f_e)^2 / f_e)$$

Cada celda aporta un valor a esta sumatoria.

Luego se determinan los grados de libertad (gl) mediante la siguiente fórmula:

$$gl = (\text{número de filas} - 1) \times (\text{número de columnas} - 1)$$

En nuestro ejemplo, $gl = (3-1) \times (2-1) = 2$.

Finalmente, el valor obtenido se compara con el valor crítico de la tabla de Chi-cuadrado según los grados de libertad calculados y un nivel de significancia previamente definido (comúnmente 0,05). Si el valor calculado de Chi-cuadrado es mayor que el valor crítico de la tabla, se rechaza la hipótesis nula, concluyendo que existe una asociación estadísticamente significativa entre las variables. Por el contrario, si el valor calculado es menor, no se rechaza la hipótesis nula.

Este procedimiento permite un análisis riguroso de la relación entre variables categóricas, facilitando interpretaciones significativas sobre asociaciones relevantes en contextos prácticos como el educativo y el deportivo.

Errores en la Decisión Estadística

Al realizar inferencias estadísticas, existe la posibilidad de cometer errores:

Error Tipo I (Alfa): Rechazar una hipótesis nula siendo ésta verdadera.

Error Tipo II (Beta): Aceptar una hipótesis nula y no rechazarla cuando en realidad, es falsa.

El nivel de significación alfa es la probabilidad de cometer un error Tipo I. Los valores más utilizados son $\alpha=0.05$ o $\alpha= 0.01$. La probabilidad de cometer un error Tipo II es Beta es siempre $1-\alpha$.

Limitaciones de la prueba Chi-Cuadrado:

- Prueba no paramétrica
- Las muestras deben ser aleatorias e independientes.
- Tamaño de muestra recomendado superior a 40 casos.
- Las categorías deben ser excluyentes y exhaustivas.
- Las variables son cualitativas y al menos una, nominal.

Referencias

- Aguirre, C.; Niño, F.; Simonetti, E. (2005). Estadística aplicada en las ciencias sociales y humanas. Coordinado por Rodolfo Nicolás Capaccio - 1a ed. 1a reimp. - Posadas: Univ. Nacional de Misiones-Editorial Universitaria.
- Baranger, D. (1999). *Construcción y análisis de datos. Una introducción al uso de técnicas cuantitativas en la investigación social*. Misiones. Ed. Universitaria.
- Batanero, C. (2001). *Didáctica de la Estadística*. Departamento de Didáctica de la Matemática Universidad de Granada.

- Maceira, V (2015) “Un abordaje teórico-metodológico para la investigación de la estructura, la movilidad social y las condiciones de vida: la propuesta ENES-PISAC”, En RELMECS, diciembre 2015, vol. 5, nº 2, ISSN 1853-7863 <http://www.relmecs.fahce.unlp.edu.ar/>
- Moore, D. (1995) *Estadística aplicada básica*. Barcelona, Antoni Bosch.

CAPÍTULO 8

Correlación bivariada y regresión entre variables numéricas

*Stella Maris Doná, Mariela Cotignola y Gisela Marcos
Castillo*

Introducción

En el capítulo anterior, hemos desarrollado las técnicas estadísticas que utilizamos al momento de describir la relación entre variables, en particular, cuando estas son medidas en los niveles nominal y ordinal.

En el presente Capítulo abordaremos una técnica específica para establecer la relación entre variables medidas en los niveles intervalar y proporcional.

Cuando estudiamos un tema de nuestro interés, se describen cada una de las variables intervinientes y es muy habitual que, a partir de lo observado, surjan preguntas referidas a la relación entre esas características de la muestra.

Por ejemplo, un Profesor de Educación Física que trabaja como docente en una escuela de enseñanza media debe desarrollar un bloque de contenidos sobre condición física, expresión corporal y salud. Uno de los aspectos a tener en cuenta, es el referido al conocimiento del propio grupo de alumnos con los que trabaja. Para ello, recoge datos tales como edad, medida en años; talla, expresada en centímetros; y peso, expresada en kilogramos. Se trata de variables numéricas continuas, medidas en el nivel de medición intervalar la primera, y racional las dos últimas. Una vez que se describe cada una de estas variables sobreviene la pregunta por la relación entre variables: ¿Se puede afirmar que a medida que aumenta la edad también lo hace la talla? ¿Hay relación entre el peso y la talla?

Sabemos que, en estadística, todas las preguntas de investigación se responden a partir de la aplicación de una medida que es elegida teniendo en cuenta el tipo de variable, su nivel de medición, la cantidad de casos y la forma que asume la distribución de sus frecuencias. Cuando el interés es describir la relación entre dos variables cuantitativas del nivel intervalar o racional, el Coeficiente de Correlación r de Pearson es el índice a aplicar (Bologna, 2011).

El término *correlación* lo introduce Francis Galton²⁰ hacia fines del siglo XIX para referirse a la influencia que tiene una variable sobre otra. Luego, Karl Pearson,²¹ continuó realizando desarrollos sobre los estudios de Galton de correlación y regresión entre variables centrándose en conocer la variabilidad observada en el estudio de las relaciones entre variables (Maureira Cid, 2017).

El estudio de las correlaciones entre pares de variables abrió el camino para responder al surgimiento de una nueva pregunta de investigación: si se ha constatado la existencia de una correlación bivariada, ¿es posible predecir el comportamiento de esos aspectos que caracterizan a un grupo?

Ejemplo: si se ha podido encontrar una correlación estadísticamente significativa entre el tiempo de entrenamiento y el porcentaje de lesiones en ciclistas; conociendo el tiempo de entrenamiento de un ciclista, ¿podría estimar cuál sería el porcentaje de lesiones que se esperaría que tenga? La respuesta es sí.

Las técnicas estadísticas de regresión lineal entre variables son las medidas que permiten establecer esa predicción probable.

En este capítulo abordaremos los temas referida a la correlación y regresión bivariada a fin de conocer cuáles son las preguntas de investigación que se responden con esos coeficientes, cuáles deben ser las condiciones para poder aplicarlas y cómo se interpretan los resultados alcanzados.

¿Qué es la correlación bivariada?

La correlación es una medida que indica si existe relación entre variables, la magnitud de su fuerza y el sentido de la misma (Johnson y Kuby, 1999).

Dadas dos variables numéricas, denominada genéricamente x e y , llamamos correlación a la medición del grado y sentido en que varía de manera concomitante los pares de valores de las variables. Ejemplo: dado el siguiente par de variables,

variable x : talla (medida en centímetros)

variable y : envergadura²² (medida en centímetros)

La correlación entre ambas variables es la medida en que, para cada una de las unidades de análisis que conforman la muestra que estudiamos, los cambios en la talla se dan junto con un cambio en su envergadura. Podría ocurrir que exista o no un cambio concomitante de los

²⁰ Francis Galton (1822-1911): científico inglés. Era primo de Charles Darwin. Desarrolló métodos cuantitativos para el estudio de la herencia humana, la psicometría, la correlación y regresión entre variables. Aunque basado en ideologías conservadoras, buscaba responder preguntas sobre la evolución y las diferencias individuales entre los hombres, con demostraciones matemáticas.

²¹ Karl Pearson (1857-1936): científico inglés. Sentó las bases de la estadística del siglo XX. Analizó la correlación bivariada y regresión lineal, para gran cantidad de datos. Al igual que F. Galton, la ideología eugenésica impulsó el desarrollo de coeficientes que, con independencia de aquella doctrina, son de gran utilidad en el ámbito de las ciencias sociales, como su Prueba de Independencia de atributos Chi cuadrado.

²² Envergadura: es la medida de la distancia de los brazos humanos completamente extendidos en cruz. Su unidad de medida es centímetros.

valores de ambas variables; y si lo hay, podría ser que la fuerza de esa relación sea fuerte o débil; además, podría darse que a medida que aumenta la talla también aumenta la envergadura o no. El valor numérico que resulte de la aplicación de la técnica de correlación, al conjunto de datos de las propiedades estudiadas en un grupo, proporciona la información sobre todos esos aspectos.

El Coeficiente de Correlación lineal r de Pearson es una técnica estadística descriptiva que sirve para analizar, para un mismo conjunto de individuos, la relación existente entre las variables. La manera de establecer si hay correlación es analizando la covariación de las variables en estudio; es decir, cómo cambian al mismo tiempo, el valor de la variable x e y , en cada uno de los sujetos de la muestra (Aguirre, Niño y Simonetti, 2006).

Ejemplo: dada una muestra de n casos, para poder establecer si existe correlación entre las variables talla y envergadura, se analiza -para cada uno de los n sujetos- cómo es la relación entre su talla y su envergadura.

El Coeficiente r de Pearson es la medida numérica de la fuerza de la relación lineal entre dos variables; refleja el efecto que tienen los cambios de los valores de una variable, sobre la otra. Su cálculo ayuda a responder la pregunta: ¿existe una correlación lineal entre las dos variables bajo estudio? (Johnson y Kubly, 1999).

La fórmula para el cálculo de una correlación r de Pearson es la siguiente (Bologna, 2011). Su nombre es producto-momento de Pearson²³:

$$r = \frac{\sum (x - Mx) * (y - My)}{n * Sx * Sy}$$

donde,

Mx ; My : son los valores de la media aritmética de cada una de las variables a correlacionar.

sx ; sy : son los valores calculados de la desviación estándar en cada una de las variables a correlacionar.

n : cantidad de casos que integran la muestra

x ; y : son el par de valores de las variables obtenidos para cada uno de los sujetos de la muestra

Σ : es el signo de sumatoria. Calcula para cada sujeto un valor de r ; luego se suman los n valores calculados, uno de cada uno de los sujetos que conforman la muestra.

Condiciones para aplicar el Coeficiente r de Pearson

Como todos los índices estadísticos, para poder aplicar r de Pearson de manera adecuada, el conjunto de datos disponible debe cumplir ciertas condiciones:

²³ El denominador de la fórmula del Coef. r de Pearson varía según los textos. Ello se debe a cómo se realiza el cálculo del desvío estándar: $n-1$ (para la muestra), o n , (para la población).

- Las dos variables a correlacionar deben ser cuantitativas.
- De esto se deduce que las variables se miden en los niveles intervalar o racional.
- Las dos variables a correlacionar deben tener una distribución que tienda a ser normal, Luego, cada una de las variables fue descrita estadísticamente con la media (M) como Medida de Tendencia Central, y a la Desviación Estándar (s) como Medida de Variabilidad. No deben observarse valores extremos descompensados.
- La población de la cual es extraída la muestra también debe ser normal.
- La cantidad de casos que integran la muestra debe ser grande.
En Estadística, hablar de muchos casos es referirse siempre a un número que supera los 30 casos en cada muestra ($n > 30$). Cuanto más grande sea la muestra, mayor será la validez de la correlación entre las variables analizadas.
- La correlación es lineal; mide la tendencia lineal entre dos variables cuantitativas relacionadas linealmente. Hay correlaciones entre variables cuyo modelo no es lineal y, por ende, no puede aplicarse el Coeficiente r de Pearson. Se debe corroborar que la tendencia de la relación entre las variables que se describen x e y, sea lineal.

Si se desea describir la correlación entre dos variables que cumplen las condiciones indicadas, el Coeficiente r de Pearson es la medida adecuada. Por el contrario, si las variables son categóricas, el Coeficiente de Asociación Q de Kendall o la Prueba de Independencia Chi cuadrado -según el caso- son las medidas adecuadas. Si una variable es ordinal y otra numérica, el Coeficiente de Correlación rho de Spearman es el adecuado. Este último Coeficiente se utiliza como la opción alternativa cuando no se puede aplicar r de Pearson.

Interpretación del resultado del coeficiente de correlación lineal r de Pearson

Una vez que se aplica la fórmula sobre el conjunto de datos de cada sujeto en ambas variables, se obtiene un valor numérico que requiere ser interpretado.

El valor calculado de r de Pearson es independiente de la unidad de medida de las variables intervinientes y es un número que está siempre dentro de los siguientes límites:

$$-1 \leq r \leq +1$$

Tres aspectos de aprecian del resultado calculado:

a) Si existe o no correlación entre las variables estudiadas.

Si $r=0$, no hay correlación bivariada; se ha hecho un hallazgo. La indagación no continúa.

Si $r \neq 0$, existe correlación entre las variables en estudio sobre la muestra; se debe continuar para determinar otros dos aspectos:

b) Dirección de la correlación.

Es el sentido de la covariación de los valores de las variables. Se obtiene analizando el signo del valor numérico del r calculado.

Cuando el signo es negativo, la correlación lineal es negativa o inversa. Significa que a medida que los valores de la variable x aumentan, los de la variable disminuyen. Ejemplo: cuanta más velocidad se imprime a una carrera, menor tiempo empleado en recorrer esa distancia propuesta.

Cuando el signo es positivo, la correlación lineal es positiva. Significa que a medida que los valores de la variable x aumentan, los de la variable y también aumentan. Ejemplo: cuanta más largas son las extremidades inferiores, mayor será el número de centímetros saltados de manera vertical.

c) Grado de la correlación.

Es la intensidad de correlación lineal encontrada. Se obtiene al analizar el valor absoluto del resultado numérico del cálculo de r .

Tabla 8.1

Interpretación del valor de la correlación

Valor de r (en valores absolutos)	Intensidad de la correlación
0	Nula
< 0,50	Baja
0,50 a 0,70	Media
> 0,70	Alta
1	Perfecta

Nota. Fuente: una adaptación del texto de Maureira Cid (2017, p.186).

Ejemplo: un Profesor de Educación Física está trabajando con un grupo de estudiantes y desea conocer si hay relación entre la velocidad y el tiempo empleado en recorrer cierta distancia. Supongamos que la correlación calculada entre las variables velocidad y tiempo empleado en recorrer cierta distancia fuere $r = -0,73$ estaría indicando que, hay correlación entre las variables ($r \neq 0$); el signo del valor calculado es negativo lo que indica que a medida que aumenta la velocidad en la carrera, disminuye el tiempo que se tarda en completar la distancia del recorrido propuesto. Que el valor numérico absoluto sea $|0,73|$ indica que, para este grupo, la fuerza de esa relación entre las variables velocidad y tiempo, es de grado alto.

Ejemplo: se busca establecer si existe relación entre el entrenamiento regular y una mejora en la flexibilidad²⁴, en grupo de personas jóvenes. Se recogen los datos relativos a ambas variables, se aplica r de Pearson y, supongamos, el valor calculado es de $r=0,85$. Al ser $r \neq 0$, podemos afirmar que sí existe relación entre las características estudiadas. Dado que el signo

²⁴ Flexibilidad: se define como: “la capacidad para desplazar una articulación o una serie de articulaciones a través de una amplitud de movimientos completos, sin restricciones ni dolor, influenciada por músculos, tendones, ligamentos, estructuras óseas, tejido graso, piel y tejido conectivo asociado”. Carrillo, E., Aguilar, V., y González, Y. (2020). El desarrollo de las capacidades físicas del estudiante de Mecánica desde la Educación Física. Mendive.

del resultado es positivo, está indicando que, a medida que aumenta la regularidad en el entrenamiento del grupo, también aumenta su flexibilidad.

Para la correcta interpretación de la correlación observada entre variables, es importante tener en cuenta que no se puede afirmar que la correlación supone una relación de causa/efecto entre las variables estudiadas; sólo determina un valor de la existencia de la covariación concomitante, su intensidad y sentido.

Diagrama de dispersión

La correlación r de Pearson entre dos variables puede anticiparse visualmente a través de la realización de una representación gráfica particular cuyo nombre es Diagrama de Dispersión. El análisis de la correlación comienza por su construcción e interpretación gráfica; sólo entonces se sabrá si tiene sentido continuar con el cálculo numérico y la descripción de la relación entre las variables estudiadas.

Es una gráfica que muestra una nube de puntos donde cada punto de esa nube representa, a la vez, el par de rendimientos de cada sujeto, en las variables que se correlacionan. El Diagrama de Dispersión tendrá tantos puntos como cantidad de unidades de análisis conformen la muestra y permite observar, justamente, cómo se dispersan los pares de valores de variables, para cada integrante del grupo estudiado.

Se dibujan los ejes cartesianos y sobre cada uno de ellos se ubican las variables y sus valores. Llamaremos variable x , a la variable que se comporta como independiente o de entrada y se la ubica sobre el eje horizontal, el eje de la abscisa. Llamaremos variable y , a la variable de salida que resulta ser dependiente de la variable x (Aguirre, Niño, y Simonetti; 2006). A la variable x , también se la llama variable explicativa; y a la variable y , se la llama variable respuesta (Kelmansky, 2009).

Debe entenderse que llamar al par de variables (x ; y) como independiente/dependiente o explicativa/respuesta hace referencia de la fortaleza de la asociación de las variables cuando se presentan juntas, a cómo los cambios en los valores de una variable, ejemplo Velocidad, producen al mismo tiempo, una modificación en otra variable, ejemplo Tiempo empleado en recorrer una cierta distancia. Para cada uno de los sujetos del grupo en estudio, los cambios en la Velocidad imprimen cambios en el Tiempo y los cambios en el Tiempo dependen de la Velocidad.

Ejemplo: sobre una muestra conformada por 17 jugadores profesionales de básquet de un total de 52 deportistas, la Tabla 8.1 es la matriz de los datos antropométricos de algunas de las variables que se analizan: edad, peso, talla, envergadura, tiempo de entrenamiento y lesiones en rodilla.

Tabla 8.2*Matriz de datos*

JUGADOR	EDAD (años)	PESO (kg)	TALLA (cm)	ENVERGADUR A (cm)	TIEMPO DE ENTRENAMIENT O (horas*)	LESIONES EN RODILLA (%**)
Santiago	23,74	79,1	181	182	9,7	14,2
Matías	31,36	97,7	176	180	13,1	9,4
Ián	24,18	69,7	180	182	13,6	8,9
Juan	39,93	88,5	191	193	10,2	12,4
Alexis	23,85	67,5	173	178	8,3	15,0
Pedro	20,42	57,7	173	189	6,1	13,7
Máximo	23,87	77,5	182	188	9,9	11,2
Alejo	20,95	89,0	184	191	13,6	5,9
Gabriel	29,50	99,9	192	192	12,8	7,8
Agustín	30,40	102,3	183	188	7,8	10,2
Enzo	24,72	72,8	175	178	10,6	9,6
Astor	25,48	109,2	192	193	8,7	12,5
Mariano	19,10	70,6	171	178	13,2	8,4
Nicolás	27,99	86,7	190	195	11,0	9,0
Esteban	25,5	65,4	179	184	9,8	11,8
Mateo	23,77	92,7	188	187	10,8	12,3
Miguel	19,92	64,5	184	188	12,3	9,8

Nota. Fuente: elaboración propia, Prof. en Educación Física Gisela Marcos Castilla, Adscripta graduada a la Cátedra de Estadística Aplicada a la Educación Física (período 2024-2026).

* Cantidad promedio de horas semanales de entrenamiento.

** Calculado por tiempo total de exposición, durante una temporada regular.

Tomaremos el siguiente par de variables para correlacionar:

variable x: Edad (medida en años)

variable y: Peso (medida en kilogramos)

La edad es la variable independiente y el Peso es la variable dependiente de ella. El siguiente gráfico 8.1 muestra el Diagrama de Dispersión cuya confección marca los pares ordenados de (Edad ; Peso) para cada uno de los 17 jugadores.

Las siguientes gráficas muestran los Diagramas de Dispersión para (x ; y): (Edad ; Talla), (Talla ; Envergadura), (Tiempo de Entrenamiento; Lesiones en Rodilla).

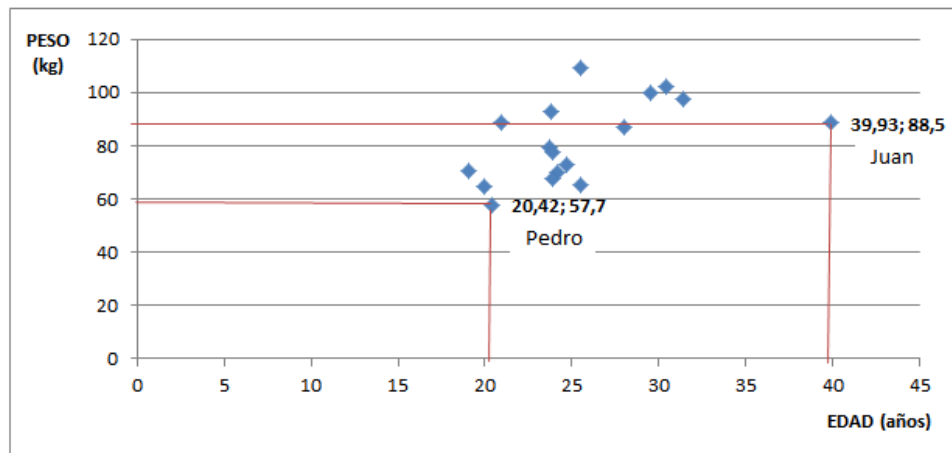
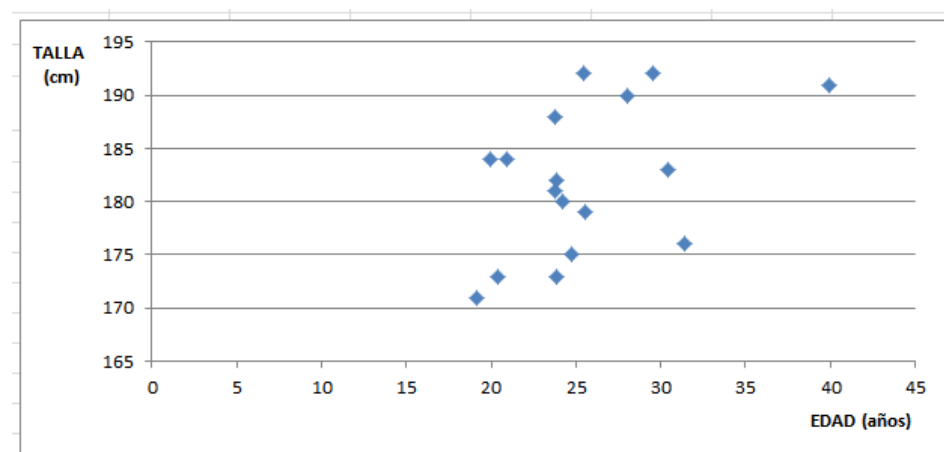
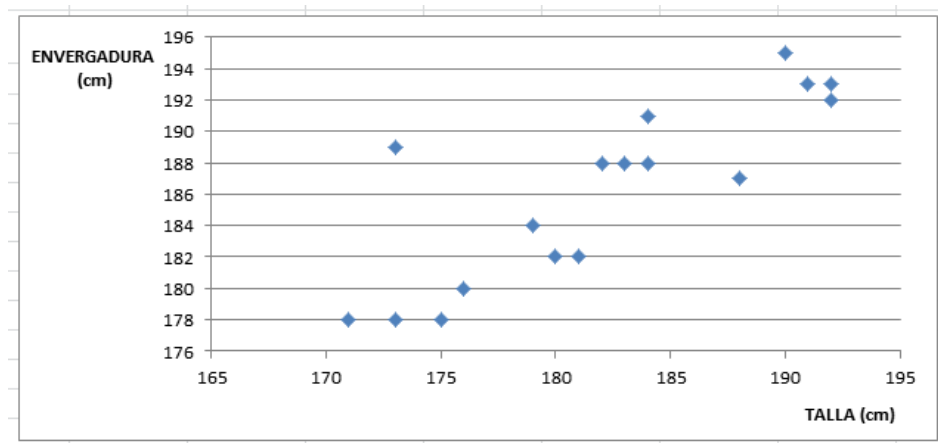
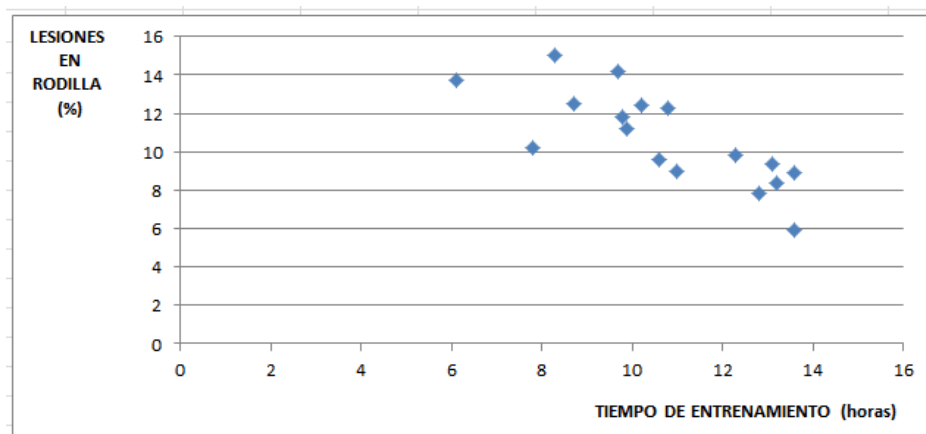
Figura 8.1*Diagrama de dispersión de las variables Edad y Peso**Nota.* Fuente: elaboración propia**Figura 8.2***Diagrama de dispersión de las variables Edad y Talla**Nota.* Fuente: elaboración propia

Figura 8.3*Diagrama de dispersión de las variables Talla y Envergadura**Nota.* Fuente: elaboración propia**Figura 8.4***Diagrama de dispersión de las variables Tiempo de Entrenamiento y Lesiones en rodilla**Nota.* Fuente: elaboración propia

Un diagrama de dispersión es una excelente aproximación visual para conocer la existencia o no de una relación lineal bivariada, su intensidad y su sentido. Una vez confeccionado, su correcta interpretación requiere atender a ciertas consideraciones.

Sobre el diagrama también se marca el par ordenado correspondiente a los valores promedios (Media aritmético) de cada una de las variables. A ese en particular se lo denomina centroide.

Ejemplo:

x: Edad (años); $M_x = 25,57$ años

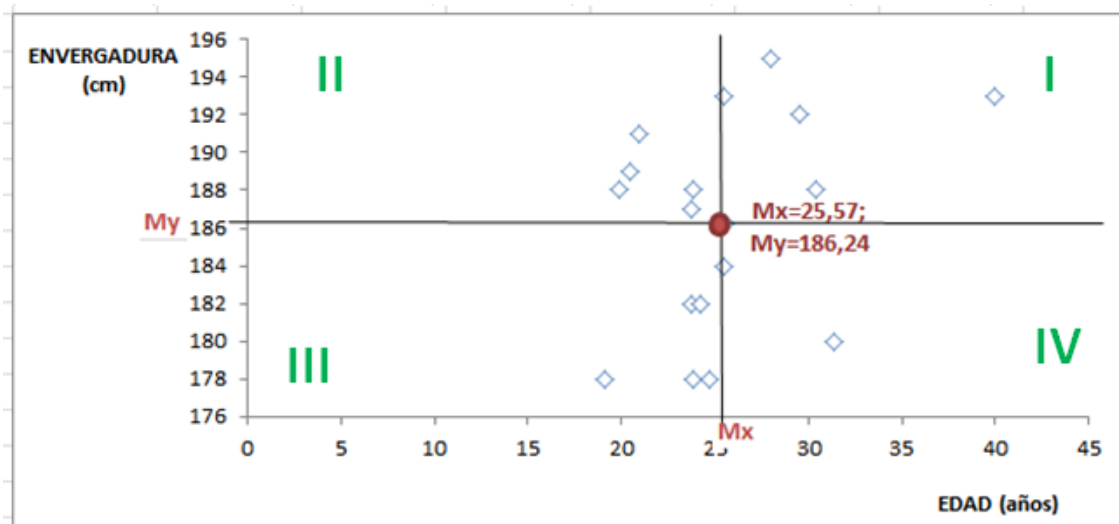
y: Envergadura (cm); 186,24 cm

En la gráfica 7.5 se observa marcado el centroide y dos líneas paralelas a cada eje, a partir de los valores de cada Media.

Quedan determinados cuatro cuadrantes que se designan con los números romanos del I al IV, en sentido contrario a las agujas del reloj (Maureira Cid, 2017).

Figura 8.5

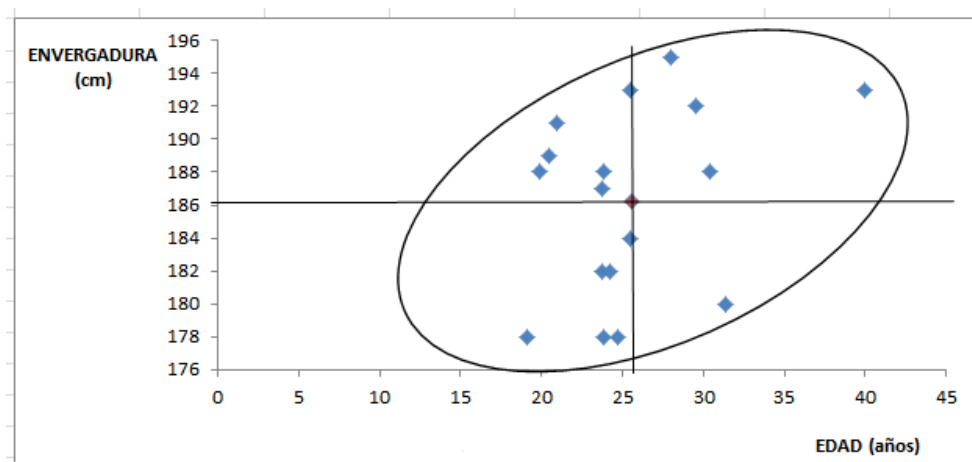
Diagrama de dispersión



Nota. Fuente: elaboración propia

Lo que resta, para poder interpretar el Diagrama de Dispersión, es observar la nube de puntos y poder determinar si existe un patrón de comportamiento de los puntos de la nube. Observemos la Figura 8.6:

Figura 8.6



Nota. Fuente: elaboración propia

Lo primero que hay que observar es que ese patrón que sigue la nube de punto tienda a ser lineal y no curvilínea. Si la forma que adquiere no es lineal, no se cumple una de las

condiciones para usar el coeficiente de Correlación r de Pearson (Aguirre, Niño, y Simonetti; 2006).

Interpretación: dado que se puede apreciar un comportamiento lineal de los puntos del diagrama, es posible usar el coeficiente r de Pearson para determinar la correlación bivariada.

Lo siguiente es observar qué forma adquieren los puntos del diagrama. ¿Los puntos tienden a seguir algún patrón, hay alguna tendencia más o menos clara en la manera de ubicarse? Si los puntos están tan dispersos que no siguen ningún patrón, decimos que no existe correlación entre las variables, sería equivalente obtener un valor calculado de $r=0$ o muy cercano a cero (Kelmansky, 2009).

Interpretación: la dispersión observada de la nube de puntos sigue un patrón lineal por lo que se puede decir que $r \neq 0$, hay correlación bivariada.

Habiendo establecido que hay un patrón de la correlación gráfica y es lineal, lo siguiente es determinar la dirección de la nube de puntos. Siguiendo con el ejemplo, el Gráfico 8.6 muestra que la nube de puntos tiende a alinearse en dirección de los cuadrantes I a III. Cuando esto ocurre, el sentido de la correlación es positivo, equivale a un valor calculado de $r > 0$.

Interpretación: La correlación es de sentido positiva lo que significa que, a medida que aumenta la Edad de cada uno de los 17 jugadores de Básquet que integran esta muestra, también tiene a aumentar su Envergadura.

Resta establecer la intensidad de la correlación. Para ello, se debe apreciar cuán próximos o separados están entre sí los puntos del Diagrama. Independientemente de la dirección de la nube de puntos, cuanto más alineados están los puntos, más alta es la correlación, cuanto más separados entre sí están los puntos, más baja será la correlación. Una correlación donde los puntos están ubicados formando una recta, equivale a un valor absoluto de $r = 1$ (Kelmansky, 2009).

Interpretación: la intensidad de la correlación es baja. Los puntos en la nube se encuentran muy dispersos entre sí.

Sabemos que se puede aplicar el Coeficiente r de Pearson (linealidad de la nube de puntos), existe correlación entre las variables x e y ($r \neq 0$; hay un patrón no caótico del comportamiento de los puntos representados en el Diagrama de Dispersión), esa correlación indica que a medida que la edad aumenta también lo hace la envergadura ($r > 0$, correlación positiva); pero esta relación es débil.

Cálculo del valor del coeficiente r de Pearson

Dado que el Diagrama de Dispersión permite evaluar la existencia de una correlación entre las variables, continuamos con el cálculo numérico que lo expresa. En la Tabla 8.3 se observa el procedimiento aplicado para la obtención de r de Pearson: $r = 0,32$.

Tabla 8.3*Procedimiento para obtener el valor de r de Pearson*

JUGADOR	EDAD (años)	ENVERGADURA (cm)	(x-Mx)	(y-My)	(x-Mx) * (y-My) =
Santiago	23,74	182	= 23,74 - 25,57 = -1,83	= 182 - 186,24 = -4,24	= (-1,83) * (-4,24) = 7,75
Matías	31,36	180	= 31,36 - 25,57 = 5,79	= 180 - 186,24 = -6,24	= (5,79) * (-6,24) = -36,11
Ián	24,18	182	= 24,18 - 25,57 = -1,39	= 182 - 186,24 = -4,24	= (-1,39) * (-4,24) = 5,88
Juan	39,93	193	= 39,93 - 25,57 = 14,36	= 193 - 186,24 = 6,76	= (14,36) * (6,76) = 97,15
Alexis	23,85	178	= 23,85 - 25,57 = -1,72	= 178 - 186,24 = -8,24	= (-1,72) * (-8,24) = 14,16
Pedro	20,42	189	= 20,42 - 25,57 = -5,15	= 189 - 186,24 = 2,76	= (-5,15) * (2,76) = -14,24
Máximo	23,87	188	= 23,87 - 25,57 = -1,70	= 188 - 186,24 = 1,76	= (-1,70) * (1,76) = -3,00
Alejo	20,95	191	= 20,95 - 25,57 = -4,62	= 191 - 186,24 = 4,76	= (-4,62) * (4,76) = -22,01
Gabriel	29,5	192	= 29,5 - 25,57 = 3,93	= 192 - 186,24 = 5,76	= (3,93) * (5,76) = 22,66
Agustín	30,4	188	= 30,4 - 25,57 = 4,83	= 188 - 186,24 = 1,76	= (4,83) * (1,76) = 8,52
Enzo	24,72	178	= 24,72 - 25,57 = -0,85	= 178 - 186,24 = -8,24	= (-0,85) * (-8,24) = 7,00
Astor	25,48	193	= 25,48 - 25,57 = -0,09	= 193 - 186,24 = 6,76	= (-0,09) * (6,76) = -0,60
Mariano	19,1	178	= 19,1 - 25,57 = -6,47	= 178 - 186,24 = -8,24	= (-6,47) * (-8,24) = 53,28
Nicolás	27,99	195	= 27,99 - 25,57 = 2,42	= 195 - 186,24 = 8,76	= (2,42) * (8,76) = 21,22
Esteban	25,5	184	= 25,5 - 25,57 = -0,07	= 184 - 186,24 = -2,24	= (-0,07) * (-2,24) = 0,16
Mateo	23,77	187	= 23,77 - 25,57 = -1,80	= 187 - 186,24 = 0,76	= (-1,80) * (0,76) = -1,38
Miguel	19,92	188	= 19,92 - 25,57 = -5,65	= 188 - 186,24 = 1,76	= (-5,65) * (1,76) = -9,97
M=	25,57	186,24			
s=	4,98	5,56			
n= 17 casos					

$$\Sigma(x - Mx) * (y - My) = 150,46$$

$$r = \frac{\Sigma(x - Mx) * (y - My)}{n * Sx * Sy} = \frac{150,46}{17 * 4,98 * 5,58} = \frac{150,46}{472,40} = 0,3185$$

r = 0,32

Nota. Fuente: elaboración propia

Si analizamos el resultado alcanzado, su interpretación debería coincidir con la anticipación visual brindada por el Diagrama de Dispersión. Efectivamente, hay correlación ($r \neq 0$) y sabemos que es lineal por dicho Diagrama. Es de sentido positivo ($r > 0$) y de grado bajo. Es decir, en este grupo de 17 jugadores de básquet, se observa una correlación entre la edad y la envergadura, a medida que aumenta la edad, también tiende a hacerlo la envergadura; sin embargo, es débil la intensidad de la relación.

La Tabla 8.4 es una matriz de correlaciones entre las variables de la Tabla 7.1.

Tabla 8.4*Matriz de correlaciones*

Matriz de Correlaciones	EDAD	PESO	TALLA	ENVERGADURA	ENTRENAMIENTO	LESIONES EN RODILLA
EDAD	1	0,54	0,47	0,32	-0,05	0,06
PESO		1	0,67	0,46	0,11	-0,23
TALLA			1	0,83	0,10	-0,15
ENVERGADURA				1	-0,13	-0,14
ENTRENAMIENTO					1	-0,78
LESIONES EN RODILLA						1

Nota. Fuente: elaboración propia

Análisis de la varianza

Es posible tomar una medida de la proporción de los cambios de los valores de la variable dependiente, explicados por los cambios de los valores de la variable independiente.

La expresión “proporción de los cambios” se refiere a la parte de la varianza observada, que es explicada por el modelo de la correlación. La variación concomitante de los valores las variables se deben, por una parte, a los cambios de la variable independiente pero los valores de la variable dependiente también cambian por otros factores, además del modelo de correlación lineal.

Ejemplo: dadas las variables x: Tiempo de entrenamiento (horas semanales) y Lesiones de la rodilla (porcentaje durante la temporada), se estableció que la correlación entre ambas es $r = -0,78$ (ver Tabla 7.3).

Sabemos que, a medida que cambia la cantidad de horas que cada jugador entrena durante la semana, también cambia el porcentaje de lesiones de rodilla que sufren durante la temporada regular de juego. La correlación $r = -0,78$ indica que la intensidad de la relación es alta, pero, claro, no es perfecta. Entonces, si bien al variar las horas de entrenamiento semanales también cambia el porcentaje de lesiones sufridas, éstas se reducen, pero no desaparecen con sólo entrenar. Una parte de la reducción del porcentaje de lesiones en la rodilla se explica por el aumento de las horas de entrenamiento semanal, pero hay otros factores que también influyen en la baja del porcentaje de lesiones de rodilla y no es porque los deportistas entrenan mucho. Una parte de la disminución de las Lesiones en rodilla se explica por el entrenamiento, pero otra parte de lo que ocurre (menor porcentaje de lesiones) no se debe al entrenamiento.

Podríamos preguntarnos cuál es la proporción de la disminución del porcentaje de Lesiones de la rodilla que pueden explicarse por el aumento del Tiempo de entrenamiento.

Llamaremos Coeficiente de Determinación al valor de r^2 , la parte de la varianza que se explica por el modelo lineal de correlación. A la parte de la varianza que no está explicada por el modelo, la llamaremos Coeficiente de No determinación y la designaremos con la letra k^2 .

Dada una correlación bivariada, el valor numérico de la varianza explicada por el modelo lineal (r^2), le sumamos el valor numérico de la varianza que el modelo no explica (k^2), entonces tendríamos la comprensión total de cómo se da la relación entre las variables (1).

Luego,

$$r^2 + k^2 = 1$$

Siguiendo con el ejemplo, si $r = -0,78$, entonces:

$$(-0,78)^2 + k^2 = 1$$

$$0,61 + k^2 = 1$$

$$k^2 = 1 - 0,61$$

$$k^2 = 0,39$$

La proporción de la varianza explicada por la correlación misma es de 0,61; un valor superior a la proporción de la varianza que no se explica por el modelo y se debe a otros factores que inciden en el suceso.

Es importante comprender el peso de cada componente de esta proporción: debería desearse que r^2 siempre sea mayor que k^2 . Se espera que, habiéndose constatado la existencia de una relación entre dos aspectos que caracterizan a un grupo (Ejemplo: Entrenamiento / Lesiones), lo deseable es que los cambios de la variable independiente (Horas de entrenamiento) expliquen la mayor parte de los cambios en la variable dependiente (Porcentaje de lesiones en rodilla).

Si realmente se desea reducir las lesiones de rodilla en este grupo de 17 basquetbolistas, se ha demostrado que el aumento de las horas semanales de entrenamiento, lo favorecen en gran medida.

En este grupo, de todo lo que influye en la reducción del porcentaje de lesiones en rodilla (1), la mayor parte se explica por el aumento del tiempo de entrenamiento semanal ($r^2=0,61 > k^2 = 0,39$) y no por otros factores que son desconocidos.

Cuando el Coeficiente de Determinación r^2 es mayor que el de No determinación k^2 , para un grupo, la correlación entre las variables es significativa y explica los cambios en la variable dependiente. Para que ello ocurra, el valor absoluto del Coeficiente de Correlación r de Pearson debe ser de 0,71. El valor calculado de r debería caer dentro de los siguientes rangos: entre -1 y -0,71 (si la correlación es negativa); o entre 0,71 y 1 (si la correlación es de sentido positivo).

Teniendo en cuenta ello, observando la Tabla 7.3, la correlación entre las variables x : Talla e y : Envergadura presentan una correlación $r=0,83$. Al analizar la varianza, resulta:

$$r^2 + k^2 = 1$$

$$(0,83)^2 + k^2 = 1$$

$$0,69 + k^2 = 1$$

$$k^2 = 1 - 0,69$$

$$k^2 = 0,31$$

Dado que $r^2 > k^2$, la proporción de la varianza de la Envergadura en este grupo se explica más por la varianza de en la Talla que por varios factores no conocidos. Podemos expresarlo en porcentaje y decir que el 69% de las diferencias que los deportistas tienen en la envergadura se explica por la cantidad de centímetros que miden, por su Talla; el 31% de las diferencias en la envergadura se deben a otros factores. El Coeficiente de Determinación r^2 se puede expresar como el porcentaje de relación entre las variables: la relación entre las variables Talla y Envergadura es del 69%; es el porcentaje de varianza compartida entre las dos variables (Maureira Cid, 2017).

Prueba de significación para r de Pearson

Hasta aquí hemos visto cómo el Diagrama de Dispersión permite determinar si el Coeficiente r de Pearson, puede ser usado para determinar la correlación entre dos variables cuantitativas, al tiempo que nos anticipa visualmente la fuerza y sentido de la correlación; cómo interpretar resultado del cálculo aplicando la fórmula de r de Pearson y cómo analizar la varianza explicada y no explicada.

Habiendo llegado hasta aquí, nos podríamos hacer la siguiente pregunta: si el número poblacional de basquetbolistas es de 52, ¿se hubiesen observado los mismos resultados que muestra la Tabla 7.3 si en vez de trabajar con una muestra de 17 jugadores hubiesen sido 23? ¿Y si hubiésemos trabajado con una muestra de 38 deportistas? ¿Cómo sabemos si los resultados alcanzados se obtuvieron por mera casualidad o si realmente son representativos del grupo?

Para poder responder a esa pregunta es que se realiza una prueba de significación.

Ejemplo: si sobre la misma población de 52 basquetbolistas, además de la muestra de los 17 jugadores que hemos analizado, se obtuvieran otras 99 muestras más, cada una de ellas con diferentes combinaciones de jugadores y, además, me tomara el trabajo de calcular los otros 99 resultados de aplicar la fórmula de la correlación entre las variables Talla y Envergadura en cada una de esas nuevas noventa y nueve muestras, ¿podríamos afirmar que la correlación que efectivamente encontré en la muestra de los 17 jugadores, la obtendría también en cualquiera de las otras 99 muestras?

Claramente, si tenemos 100 muestras diferentes de una misma población, y nos tomamos el trabajo de calcular los 100 valores de r de Pearson para medir la correlación entre las variables Talla y Envergadura, tendríamos 100 valores de r de Pearson distintos.

Pero la pregunta es ¿cómo saber si el resultado que efectivamente encontré al trabajar con mis 17 deportistas es un valor de correlación representativo de la población de la cual se extrajo la muestra, si es representativo del total de 52 basquetbolistas? Esa correlación hallada entre Talla/Envergadura de $r=0,83$, fuerte y positiva, ¿sólo se limita a esos 17 jugadores? ¿Fue por mero azar o se trata de un valor representativo de toda la población de deportistas?

Aplicar una prueba de significación resuelve esta duda. Cada coeficiente tiene su propio estadístico de prueba. Para r de Pearson, disponemos de dos maneras de saber si los resultados obtenidos en la muestra con la que trabajamos fueron obtenidos por mero azar o, por el contrario, son significativos y representativos de la población, de la cual se seleccionó la muestra que analizamos.

Hay una manera sencilla: el valor calculado, observado de r de Pearson, se contrasta con un valor de r de Pearson llamado crítico o teórico que está disponible en una Tabla disponible en el Anexo. Para poder obtener ese valor crítico de r de Pearson, se requieren dos elementos: los grados de libertad (gl) y el nivel de significación alfa (α).

Los grados libertad de calculan restando 2 al total de casos de la muestra: **gl = n-2**

El nivel de significación α se establece previamente; usualmente, $\alpha = 0,05$; o bien $\alpha = 0,01$. Cuando no se especifica un valor, se toma $\alpha = 0,05$.

Entonces, se tienen dos valores de del Coeficiente de Correlación r de Pearson:

1) r calculado: es el observado como resultado de aplicar la fórmula sobre el conjunto de datos

2) r crítico: es el obtenido de la Tabla de valores teóricos de r (ver Anexo), para $n=17$ casos y nivel de significación alfa ($\alpha = 0,05$)

Los valores de r de Pearson, calculado y crítico, se comparan entre sí. Luego, se aplica la regla de decisión: $|r \text{ calculado}| > r \text{ crítico}$, lo encontrado en la muestra no fue por causalidad, es representativo de la población de la cual se obtuvo la muestra en estudio. De lo contrario, los hallazgos son sólo válidos para la muestra.

Si r calculado (siempre en valor absoluto, sin el signo del sentido de la correlación) es mayor que el valor de r crítico que se obtiene de la Tabla (ver Anexo), significa que la correlación encontrada en la muestra no se obtuvo por mero azar; por el contrario, el valor obtenido es estadísticamente significativo para esa cantidad de gl y nivel de significación α (Maureira Cid, 2017).

Ejemplo:

Variables correlacionadas: x es Talla; y es Envergadura;

1) Valor calculado de $r = 0,83$, se obtiene de aplicar la fórmula (manual o a través de Excel o de algún software estadístico)

2) El valor crítico de r se obtiene de la tabla (ver Anexo):

$$n = 17$$

$$\alpha = 0,05 \text{ (valor predeterminado)}$$

$$r \text{ crítico} = 0,482 \text{ (surge de la tabla, ver Anexo)}$$

Ejemplo:

Variables correlacionadas: x “Tiempo de entrenamiento”; y “porcentaje de lesiones en rodilla”; Valor calculado de $r = - 0,78$, se obtiene de aplicar la fórmula

El valor crítico de r se obtiene de la Tabla (ver Anexo):

$$n=17$$

$$\alpha = 0,05 \text{ (valor predeterminado)}$$

$$r \text{ crítico} = 0,482 \text{ (surge de la Tabla, ver Anexo)}$$

La regla de decisión e interpretación nos indica que, como el valor absoluto de r calculado $= 0,78 > r \text{ crítico} = 0,482$ corresponde interpretar que el resultado de la correlación como en el caso anterior. La correlación muestral entre las variables Talla y Envergadura (alta y positiva) es un buen estimador de la correlación poblacional con un nivel de significación $\alpha = 0,05$ y n casos.

La otra prueba de significación para el Coeficiente r de Pearson es utilizando la Distribución t de Student. Se obtiene un valor de t calculado (resultante de aplicar una fórmula) y un valor de t crítico (surge de una Tabla que puede verse en el Anexo, con $gl = n - 2$, y un valor

predeterminado de alfa, generalmente $\alpha = 0,05$). Luego se comparan ambos valores y si se cumple la regla de decisión que dice que si el t calculado $>$ t crítico, se acepta el hecho de que el valor de la correlación muestral es un buen estimador de la correlación poblacional, que en el 95% de las muestras diferentes que se pudieran extraer de una misma población, se observaría lo que encontramos en la muestra pues no fue obtenido por casualidad (Maureira Cid, 2017).

La fórmula para obtener el valor t calculado (Bologna, 2013, p 423), es:

$$t = \frac{r}{\sqrt{\frac{1 - r^2}{n - 2}}}$$

Si reemplazamos los valores de la correlación entre las variables Talla/Envergadura, el valor t calculado = 5,93. Se obtiene de la aplicación de la fórmula a los datos disponibles:

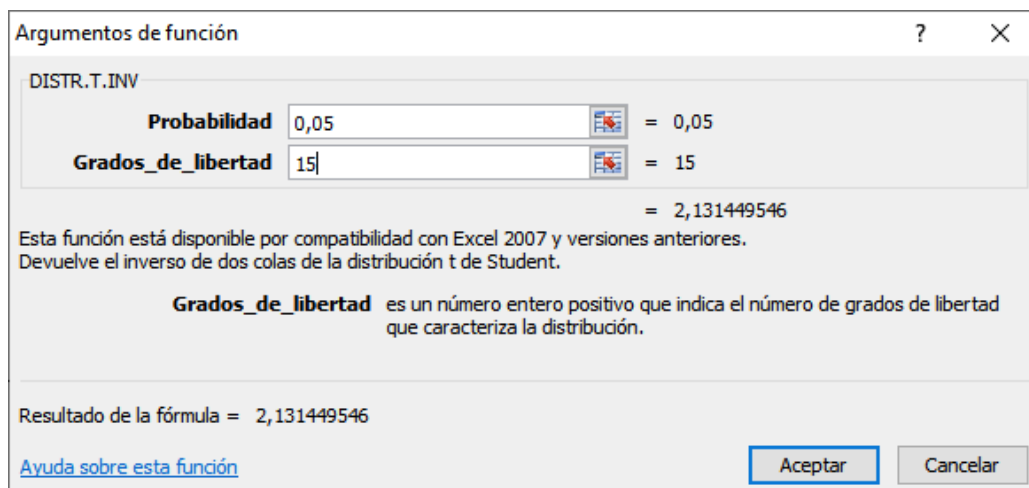
$$t = \frac{0,83}{\sqrt{\frac{1 - (0,83)^2}{17 - 2}}} = \frac{0,83}{\sqrt{\frac{0,31}{15}}} = \frac{0,83}{0,14} = 5,93$$

El valor del t crítico se obtiene de la Tabla o utilizando la función de Excel, DISTR.T.INV

El valor de t crítico se calculó para $gl = 17 - 2 = 15$; y $\alpha = 0,05$: **t crítico = 2,13**

Figura 8.7

Cálculo en Excel



Nota. Fuente: elaboración propia

El valor crítico de t se puede obtener de una Tabla de Distribución t , de dos colas, con un nivel de significación $\alpha=0,05$ (salvo que se especifique un valor diferente) y grados de libertad = cantidad de casos menos 2 ($gl=n - 2$)²⁵.

Regla de decisión: si el t calculado $>$ t crítico, se rechaza la idea que los valores obtenidos en la muestra hayan sido por mera casualidad; por el contrario, en el 95% de las diferentes muestras que pudieran seleccionarse de la misma población, se observarían los resultados encontrados en la muestra de los 17 jugadores que sí se analizó. Tal como comprobamos al realizar la prueba de significación del coeficiente r de Pearson usando la comparación con su valor crítico; al aplicar el método proporcionado por el valor de la Distribución t , se llega a la misma conclusión: el valor de la correlación muestral es un buen estimador de la correlación poblacional, considerando $\alpha = 0,05$ y $gl=15$: la talla correlaciona con la envergadura.

Sobre la realización de cálculos y gráficos

Cuando tenemos un conjunto grande de casos, ya sea por tiempo o precisión, los cálculos matemáticos y cualquier otra expresión que los describa se realizan utilizando Microsoft Excel y/o algún tipo de software para el procesamiento y análisis estadístico de datos, como lo son el SPSS (Maureira Cid, 2017), o InfoStat (Bologna, 2011), o el Programa Estadístico R²⁶, o similar que el vertiginoso avance tecnológico nos ofrezca.

Excel es un programa de cálculos que permite organizar la información disponible al modo de una Base de Datos, como hemos visto en el Capítulo 3. Cuenta con muchas funciones que, de manera simple e inmediata, arrojan los resultados de cálculos largos y complejos. La función COEF.DE.CORREL²⁷ calcula el valor numérico de la correlación bivariada. El Diagrama de Dispersión se puede insertar como un gráfico y además, es posible agregar la ecuación de la recta de regresión, y el valor del coeficiente de determinación r^2 .

Si utiliza una Tabla de valores de r de Pearson críticos, recuerde que pueden aparecer con n cantidad de casos de la muestra; o bien puede aparecer con grados de libertad $gl = n - 2$.

²⁵ Enlace a la Tabla de Distribución de valores críticos de t (utilice para dos colas) https://www.criticalvaluecalculator.com/t-table?_x

²⁶ R es un software de procesamiento y análisis de datos particularmente popular entre la comunidad científica y estadística, el cual es multiplataforma y está distribuido gratuitamente bajo licencia pública general.

²⁷ La actualización que se disponga de Microsoft Excel puede ofrecer otra denominación para el cálculo de la correlación.

Regresión bivariada

¿Qué es la regresión lineal simple?

La regresión lineal simple es una técnica estadística inferencial que, desarrollada a partir de la técnica descriptiva de correlación r de Pearson, permite predecir los valores que se obtendrían en una variable numérica, en base a otra variable cuantitativa que correlaciona con aquella (Maureira Cid, 2017).

Recordemos que el Coeficiente r de Pearson mide la existencia de la correlación lineal entre dos variables cuantitativas; es decir, permite determinar que una variable llamada dependiente o respuesta, puede expresarse en función del comportamiento de otra variable llamada independiente o predictora.

Si la correlación bivariada es estadísticamente significativa ($r^2 > k_2$), podemos asegurar que, dadas dos variables, las variaciones de los valores en una de ellas (variable explicativa), se acompaña de la variación en los valores de la otra variable (variable explicada).

Entonces, independientemente del sentido de la correlación, si el grado de la misma es alto, estaremos describiendo una relación fuerte entre esas dos variables analizadas, por ejemplo, los pares de variables velocidad y tiempo; edad y talla; talla y peso; kilómetros corridos y capacidad aeróbica; tiempo de precalentamientos y porcentaje de lesiones, etc, correlacionan siempre de manera fuerte.

A partir de esas correlaciones observadas, la pregunta que nos hacemos a continuación es esta: ¿es posible predecir el comportamiento de una de esas dos variables en función de la información que efectivamente se tiene de la otra? Dada una correlación alta entre dos variables, ¿Conocer el valor de la variable independiente de una correlación alta, permite estimar el valor de la variable dependiente?

Ejemplo: si se sabe que, en un grupo, a mayor edad, mayor talla; y que dicha correlación es significativa, ¿podría estimarse cuál sería la talla de un sujeto de ese mismo grupo, conociendo su edad?

Si se observó que existe una correlación fuerte entre el tiempo de entrenamiento de lanzamiento en baloncesto y el porcentaje de aciertos en los tiros libre durante los partidos jugados, ¿es posible inferir cuál será ese porcentaje sabiendo cuánto fue el tiempo de entrenamiento los lanzamientos? La respuesta es sí.

La regresión bivariada es la técnica estadística para poder realizar dicho análisis. Para su aplicación, deben cumplirse ciertas condiciones, a saber (Maureira Cid, 2017):

- las variables deben ser numéricas,
- las distribuciones de las variables deben tender a ser normales,
- los datos obtenidos para cada una de las variables bajo estudio, deben ser independientes entre sí,
- debe existir homogeneidad en las varianzas de ambas distribuciones de las variables,
- la correlación entre las variables debe ser lineal.

Las rectas de regresión lineal

Para poder predecir, la regresión se basa en la descripción de la forma que toma correlación bivariada. Cuando ésta es alta, en el Diagrama de Dispersión se observa que la nube de puntos tiende a tener una clara direccionalidad, una forma alineada, a agruparse formando una recta.

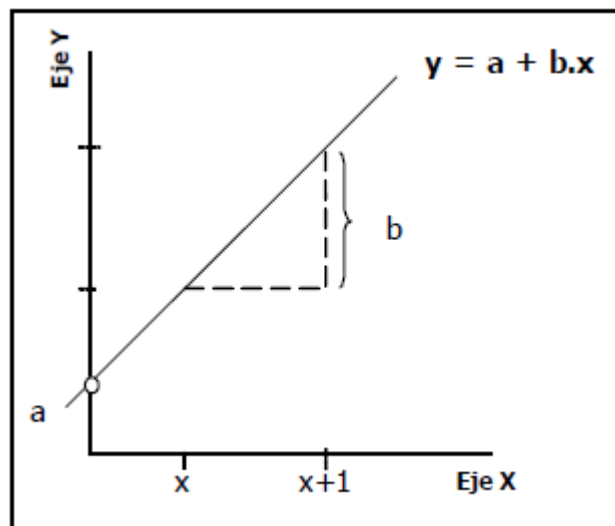
A partir de esta característica, la regresión determina cuál es la ecuación de esa recta que, idealmente, sería la línea sobre la que se ubicaría cada uno de los puntos del Diagrama de Dispersión, si la correlación fuera perfecta. La fórmula de una recta es la siguiente:

$$y = a + bx$$

donde a y b son valores constantes; x es el valor de la variable independiente, y es el resultado de la variable dependiente. Recordemos algunas particularidades más sobre la ecuación de una recta pues resultará de interés conocerlas a los fines de la mejor comprensión de la regresión, como se aprecia en el siguiente gráfico (Aguirre, Niño, Simonetti; 2006; p. 124).

Figura 8.8

Recta de regresión



Nota. Fuente: elaboración propia

Cuando el valor de $b=0$, la recta queda formulada como $y=a$ y es el punto en el cual la recta corta al eje y . El valor a , se denomina ordenada al origen.

El valor b , es la pendiente de la recta. Este valor de inclinación tiene la particularidad de ser positivo en correlaciones positivas, y ser negativo en correlaciones de sentido negativo.

La técnica de regresión buscará formular la ecuación de una recta que sea la expresión matemática de la línea más cercana posible de todos los puntos que se presentan en el Diagrama de Dispersión, y decimos la ecuación de la recta que más se ajusta a los puntos de la nube pues sabemos que las correlaciones sólo excepcionalmente son perfectas.

La recta de regresión sería el camino sobre el cual deberían ubicarse todos los pares de valores correlacionados si es que la correlación fuese $|1|$. Pero como las correlaciones son, en el mejor de los casos, muy altas, la regresión busca la fórmula de la ecuación que proporciona una recta que, atravesando la nube de puntos, es mejor expresión de su dirección y cercanía de los pares de valores.

Observemos el Gráfico 8.7. Está realizado a partir de la matriz de la Tabla 8.1 y el Gráfico 8.4. Se agregan tres elementos: el trazado de la recta de regresión, su ecuación y el valor del Coeficiente de Determinación r^2 .

El gráfico se realizó con Excel, insertando el gráfico de dispersión y agregando la línea de tendencia y el valor de r^2 . Resulta muy simple cargar la base de datos y luego, rápidamente obtener un Diagrama como este y la información detallada, de mucha importancia para el análisis estadístico. Los mismos valores los podemos calcular de manera manual.

En nuestro ejemplo, se ha descrito la correlación lineal bivariada entre:

x: Tiempo de entrenamiento (medido en horas /variable predictora)

y: Lesiones en rodilla (medida en porcentaje de lesiones durante la temporada regular / variable de resultado)

Se observó una correlación $r = -0,78$ (Correlación alta y negativa).

Coeficiente de determinación 0,61 (el 61% de lo que ocurre entre entrenar y las lesiones se explica por la misma relación entre ambas variables y sólo el 39% se debe a otros factores).

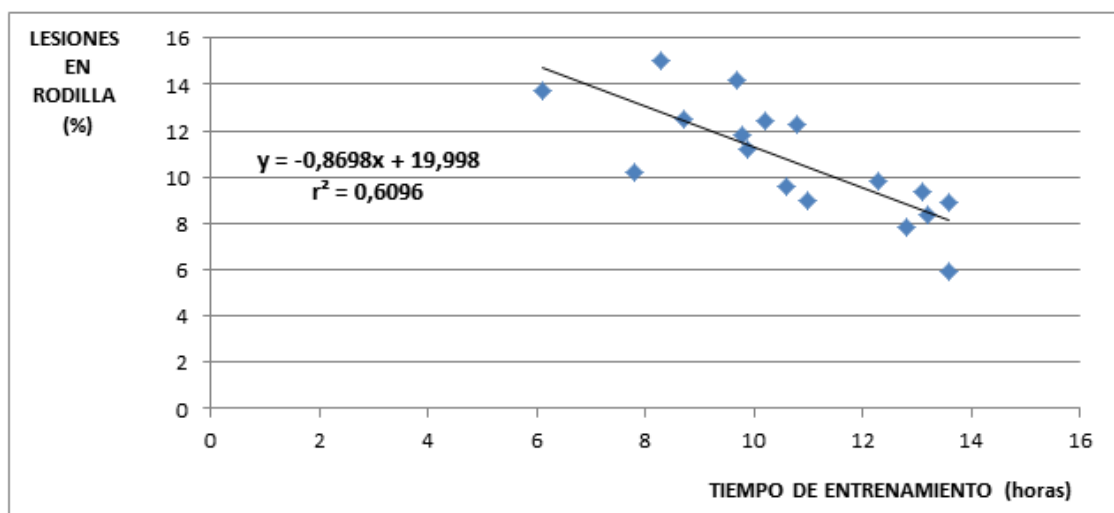
La ecuación de la recta de regresión de la variable x, para predecir los valores de y, quedó formulada así:

$$y = -0,8698x + 19,998$$

donde $a=19,998$ (si se prolonga la recta, verá cómo corta al eje vertical justo en el punto 19,998); $b=-0,8698$ (nótese que el valor de b , es negativo, como la correlación).

Gráfico 8.9

Diagrama de dispersión entre variables



Nota. Fuente: elaboración propia

Resulta importante citar aquí la siguiente afirmación:

Los valores de los coeficientes a y b se obtienen fácilmente a través de cualquier programa estadístico. Nuevamente aquí resulta importante comprender la lógica para determinar la recta que mejor ajusta la nube de puntos y la utilidad de contar con este modelo, más que los cálculos que requieren la determinación de estos coeficientes. (Aguirre, Niño y Simonetti, 2006, p 125).

La ecuación de la recta de regresión es la técnica que permite a la estadística inferir los porcentajes de lesiones en rodilla que podrían sufrir durante la temporada, los basquetbolistas de nuestro grupo de deportistas. Y es una herramienta potente pues sabemos que la correlación es fuerte. Sólo nos restaría conocer el tiempo de entrenamiento de un determinado jugador para poder estimar el porcentaje de lesiones en rodilla durante la temporada.

En realidad, las rectas de regresión lineal son dos, una ecuación de la recta cuando conocemos los valores de x e inferimos los de y; y otra cuando conocemos los valores de y e inferimos los valores de la variable x. También puede ocurrir que nos interese medir una variable explicada (ejemplo, peso) en dos grupos diferentes (ejemplo, los que hacen actividad física regular y los que no).

Ejemplo: Martín es un basquetbolista del grupo de deportistas estudiados. Martín, efectivamente realiza 7 horas de entrenamiento semanal. ¿Cuál es su porcentaje de lesiones en rodilla como deportista de básquet? Simplemente reemplazamos el valor de $x = 7$ horas de entrenamiento semanal.

Algo importante, nótese que, en lugar de consignar “y”, se ha escrito “ $\hat{y}$ ”, que se lee el valor estimado de y. Es estimado pues es un valor probable. No es el valor exacto. No conocemos exactamente cuál es su porcentaje de lesiones, sino que lo estamos infiriendo a partir de los datos que sí conocemos:

$$\begin{aligned} y &= bx + a \\ y &= -0,8698x + 19,998 \\ \hat{y} &= (-0,8698 * 7) + 19,998 \\ \hat{y} &= -6,0886 + 19,998 \\ \hat{y} &= 13,9094 \% \end{aligned}$$

Redondeando, el valor estimado del porcentaje de lesiones en rodilla para Martín es $\hat{y} = 13,91\%$

Interpretación: dado que efectivamente se observan las siguientes condiciones,

- una correlación alta, observada entre las variables x e y;
- Martín es un sujeto de la muestra que se correlacionó,
- se sabe que él entrena 7 horas semanales.

Entonces, se estima que su porcentaje de lesiones durante la temporada es de $\hat{y} = 13,91\%$.

La regresión lineal simple busca la recta que mejor se ajusta a la nube de puntos, buscando minimizar los errores que se producen al estimar un valor de una variable. A partir de esta condición se define el siguiente criterio para estimar la recta que mejor ajusta las observaciones: el criterio de cuadrados mínimos (Aguirre, Niño y Simonetti, 2006).

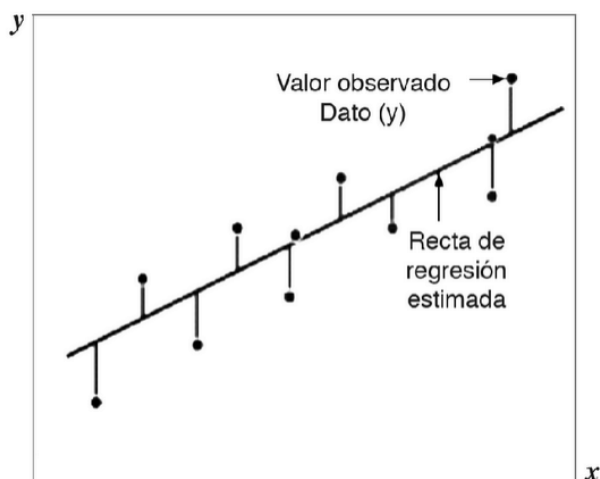
Los coeficientes de la recta de regresión a y b , proporcionan una ecuación que es la que reduce al mínimo la suma al cuadrado, de las distancias o errores o diferencias en la estimación predictiva de un valor de variable.

La ecuación de la recta es el patrón lineal que tienen a seguir los puntos de la nube. Mediante la técnica de los cuadrados mínimos, la regresión busca presentar la mejor ecuación que trace una recta que deje la menor distancia vertical entre todos los puntos y esa línea. (Kelmansky, 2009).

La siguiente figura²⁸ muestra cómo el método busca minimizar la suma de los cuadrados, de las distancias a la ecuación que será la de la recta de regresión y que es la que servirá para estimar el valor de y ($\hat{y}$), conociendo x .

Figura 8.10

Recta de los cuadrados mínimos



Nota, Fuente: elaboración propia

La distancia vertical de cada punto a la recta, se llama *residuo*. Los residuos son valores que pueden ser positivos (cuando sobrepasan la recta) o negativos (cuando quedan por debajo de la recta).

Para obtener la ecuación de la recta de regresión $y=a + bx$ podemos obtener los valores de a y b , de la siguiente manera:

$$a= My - b*Mx$$

²⁸ Ilustración disponible en: https://www.researchgate.net/figure/Figura-3-Representacion-de-los-minimos-cuadrados_fig1_311548169

donde M_y y M_x , son los valores de las Medias de cada una de las variables que se han correlacionado,

$$b = r * \frac{s_y}{s_x}$$

donde r es el valor calculado de la correlación, s_y y s_x son los valores de la desviación estándar de cada una de las variables correlacionadas.

De igual manera, si no se ha formulado la ecuación de la recta, puede estimarse el valor de y , con la siguiente fórmula que utiliza medidas descriptivas de ambas variables correlacionadas y que ya se disponen:

$$\hat{y} = r * \frac{s_y}{s_x} * (x - M_x) + M_y$$

Donde:

$\hat{y}$ es el valor estimado de la variable y

r es el valor del coeficiente de correlación de Pearson

s_x es la desviación estándar, de la variable x

s_y es la desviación estándar, de la variable y

X es el valor de la variable para el caso que se quiere estimar

M_x es la media de la variable x

M_y es la media de la variable y

En el ejemplo que seguimos, los estadísticos descriptivos de las variables, se presentan en la siguiente tabla:

Tabla 8.2

Ejemplo de correlación bivariada

$r = -0,78$	x: Tiempo de entrenamiento semanal (horas)	y: Lesiones en rodilla (%)
Media	10,68 h	10,71 %
Desvío estándar	2,21 h	2,46 %
x (Martín)	7 h	$\hat{y}$ %?

Nota. Fuente: elaboración propia

Si reemplazamos los valores, en la fórmula, obtendremos:

$$\hat{y} = r * \frac{sy}{sx} * (x - Mx) + My$$

$$\hat{y} = (-0,78) * \frac{2,46}{2,21} * (7 - 10,68) + 10,71$$

$$\hat{y} = (-0,78) * 1,11 * (-3,68) + 10,71$$

$$\hat{y} = 2,90 + 10,71$$

$$\hat{y} = 13,61 \%$$

Nótese cómo el valor de $\hat{y}$ es similar al calculado mediante la ecuación de la recta de regresión de la suma de los cuadrados mínimos.

Error en la estimación predictiva

La correlación entre el tiempo del entrenamiento y el porcentaje de lesiones en rodilla, si bien fue significativa ($r = -0,78$), no fue perfecta. Ello significa que hay otros factores que intervienen para disminuir el porcentaje de lesiones en rodilla. Esos factores podrían ser condiciones físicas preexistentes, largo de piernas, tiempo de juego como titular e, incluso, hasta la motivación deportiva. La consecuencia de ello es que, si la regresión es una técnica predictiva, dicha estimación nunca podría ser exacta.

Lo que se hace entonces es cuantificar esa distancia entre lo que se estima que ocurre y lo que realmente ocurre. A esta diferencia se la llama *error de predicción*, *error en la estimación predictiva*. El término error no debe entenderse como sinónimo de equivocación sino como la medida que permite conocer el grado de incertidumbre que se maneja al estimar un valor que describe un conjunto de datos.

Resulta evidente que los pares de valores de las dos variables correlacionadas no se alinean sobre la recta de regresión teórica, sino que ésta constituye el modelo, la forma en que se presentan los puntos de la nube en el Diagrama de Dispersión. Los puntos se ubican algunos por debajo de la recta y otros la sobrepasan.

Dicha variación es una desviación que se sabe que va a ocurrir pues, justamente, estimamos bajo algunas condiciones de incertidumbre. El error de predicción mide el rango o margen dentro del cual se espera que realmente aparezca el verdadero valor de la variable y ,

Error en la estimación predictiva se calcula sumando y restando al valor promedio, un valor numérico, el valor del error que se comete al hacer inferencias. La fórmula para realizar el cálculo suele designarse como σ^{29} o bien como ε^{30} (Kelmansky, 2009).

²⁹ σ : letra griega sigma.

³⁰ ε : letra griega épsilon.

$$\epsilon = S_y * \pm \sqrt{1 + r^2}$$

Es muy importante tener en cuenta que el margen de error calculado es válido para casos estadísticamente normales; es decir, para los sujetos cuyos rendimientos en las variables que se analizan, estén dentro del 68,26% central de los casos; alrededor de la media de su grupo, en ± 1 desvío a ese valor central.

Si calculamos el error en la estimación predictiva para nuestro ejemplo, obtenemos:

$$\epsilon = 2,46 * \pm \sqrt{1 + (-0,78)^2}$$

$$\epsilon = 2,46 * \pm \sqrt{1 + 0,61}$$

$$\epsilon = 2,46 * \pm \sqrt{0,39}$$

$$\epsilon = 2,46 * \pm 0,62$$

$$\epsilon = \pm 1,52 \%$$

Interpretación: el valor calculado del error debe sumarse y restarse (tal como indican los signos) al valor estimado de y ($\hat{y}$): $\hat{y} = 13,91\%$; $\epsilon = \pm 1,52 \%$

$\hat{y} - \epsilon$	$\hat{y}$	$\hat{y} + \epsilon$
$13,91\% - 1,52 \%=$	13,91%	$13,91\% + 1,52 \%=$
12,39%		15,43%

Del resultado de sumar y restar el valor de ϵ , al $\hat{y}$, queda determinado el rango (12,39%; 15,43%) y es la amplitud dentro de la cual se estima que esté el verdadero valor del porcentaje de lesiones en rodilla de Martín.

Dada una correlación significativa ($r=-0,78$) entre las variables tiempo de entrenamiento y lesiones en rodilla, se observó que Martín es un deportista de la muestra y que efectivamente se entrena $x=7$ horas. Se estima que su porcentaje de lesiones es de 13,91%, valor estimado con un margen de error de $\pm 1,52\%$; por lo tanto, se espera que el valor real del porcentaje de lesiones en rodilla sea entre el 12,39% de base y el 15,43% como máximo, siempre y cuando Martín sea un deportista estadísticamente normal en sus desempeños.

Referencias

- Aguirre, C.; Niño, M. F.; Simonetti, E. y (2006). “El estudio de la relación entre variables”, en *Estadística Aplicada a las Ciencias Sociales y Humanas*. (pp. 120-129). Universidad Internacional de Andalucía. Posgrado en desarrollo económico para América Latina, material de cátedra.
- Bologna, E (2011). *Estadística para Psicología y Educación*. Presentación y Cap. 5, 10 y 13. Córdoba, Argentina. Editorial Brujas.
- Johnson, R y Kuby P (1999) *Estadística elemental*. Cap. 3 México: ITP. Disponible en: [file:///C:/Users/Usuario/Downloads/Estadística%20Elemental%20-%20Johnson%20Kuby%2011ed%20\(1\)-1.pdf](file:///C:/Users/Usuario/Downloads/Estadística%20Elemental%20-%20Johnson%20Kuby%2011ed%20(1)-1.pdf)
- Kelmansky, D. (2009). “Relación entre variables”, en *Estadística para todos. Estrategias de pensamiento y herramientas para la solución de problemas*. Cap.22, apartados 1 al 4. Ministerio de Educación de la Nación. Instituto Nacional de Educación Tecnológica. Disponible en: <http://www.bnm.me.gov.ar/giga1/documentos/EL001858.pdf>
- Maureira Cid, F. (2017). “Correlación y Regresión”, en *Estadística Básica para Educación Física*. 2da. Ed., Cap. 13. Letonia: Editorial Académica Española.

Anexo. Tabla de valores críticos para r de Pearson

Tabla de valores críticos del coeficiente de correlación de Pearson para n es la cantidad de casos de la muestra. (Los valores críticos de la r de Pearson que se muestran son para una prueba unilateral según grados de libertad=n-2).

n	0,05	0,01	n	0,05	0,01
1	-	-	24	0,404	0,515
2	-	-	25	0,396	0,505
3	0,997	1	26	0,388	0,496
4	0,95	0,99	27	0,381	0,487
5	0,878	0,959	28	0,374	0,479
6	0,811	0,917	29	0,367	0,479

7	0,755	0,874	30	0,361	0,463
8	0,707	0,834	32	0,349	0,449
9	0,666	0,798	34	0,339	0,436
10	0,632	0,765	36	0,329	0,424
11	0,602	0,735	38	0,32	0,413
12	0,576	0,708	40	0,312	0,402
13	0,553	0,684	42	0,304	0,393
14	0,532	0,661	44	0,297	0,384
15	0,514	0,641	46	0,291	0,376
16	0,497	0,623	48	0,285	0,368
17	0,482	0,606	50	0,279	0,361
18	0,468	0,59	60	0,254	0,33
19	0,456	0,575	70	0,235	0,305
20	0,444	0,561	80	0,22	0,386
21	0,433	0,549	90	0,207	0,27
22	0,423	0,537	100	0,197	0,256
23	0,413	0,526			

Anexo. Tabla de distribución t de Student

t Tabla _ dos colas											
gl = n - 2	t _{.50}	t _{.75}	t _{.80}	t _{.85}	t _{.90}	t _{.95}	t _{.975}	t _{.99}	t _{.995}	t _{.999}	t _{.9995}
	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
Z	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										

CAPÍTULO 9

Muestreo y estimación paramétrica

Nociones de muestreo, estimación de parámetros y prueba de hipótesis

Stella Maris Doná y Mariela Cotignola

Nociones de muestreo

La estadística inferencial estudia un conjunto de procedimientos para extender las conclusiones que se obtienen en una muestra a la población general que es objeto de interés. Por razones de tiempo y costos operativos la mayor parte de las veces no es posible estudiar la población completa; por ello se opta por estudiar las variables en una muestra representativa de dicha población.

El muestreo entonces es un procedimiento mediante el cual se selecciona de un conjunto determinado llamado población, un subconjunto que recibe el nombre de muestra, con el objeto de llegar al conocimiento de determinadas características de los elementos de la población mediante la generalización de los correspondientes a los elementos de la muestra.

La teoría del muestreo tiene como objeto determinar los procedimientos mediante los cuales sea posible hacer generalizaciones sobre la población a partir de un subconjunto de la misma, con un mínimo de error. Es importante señalar que, no toda muestra se extrae con el propósito de proporcionar indicadores para realizar estimaciones a la población. Los valores muestrales son los estadísticos que describen las muestras, y con los cuales se busca inferir los parámetros poblacionales (Padua, 2018).

Para seleccionar una muestra, lo primero que se debe definir es la unidad de análisis (personas, organizaciones, objetos, etc.). El “quiénes van a ser medidos”, depende de precisar claramente el problema a investigar y los objetivos de la investigación. Cada uno de los objetos o sujetos que resultan de interés para nuestro análisis y son seleccionados, se llama unidad muestral (Kelmansky, 2006).

Una vez que se ha definido cuál será la unidad de análisis, se procede a delimitar la población que va a ser estudiada y sobre la cual se pretende generalizar los resultados. Una **población** es el conjunto de todos los casos que concuerdan con una serie de especificaciones. Objetos, sujetos o elementos con características determinadas, definido por

la explicitación de estas características. Cuando una población tiene más de 100.000 elementos, se la considera una **población infinita** (Baranquer, 2009). Para Padua (2018), una muestra es infinita cuando supera los 30.000 casos (p. 10).

La delimitación de las características de la población no sólo depende de los objetivos del estudio, sino de otras razones prácticas. Un estudio no será mejor por tener una población más grande, la calidad de un trabajo radica en delimitar claramente la población con base en los objetivos del estudio. La población debe situarse en torno a sus características de contenido lugar y en el tiempo. Por ejemplo: deportistas residentes en la ciudad de La Plata en el año 2020; egresados de carreras universitarias en los últimos cinco años en la República Argentina, etc.

Una **muestra** es un subconjunto de elementos que pertenecen a ese conjunto definido en sus características al que llamamos población, una parte seleccionada de la población, extraída con objeto de obtener información sobre la totalidad de ésta.

El requisito principal que debe reunir una muestra es la representatividad, o sea, que los resultados que se obtengan de la misma deben concordar con los que se hubiesen obtenido de haber sometido a observación toda la población. Cuando hablamos de representatividad de la muestra se debe especificar no sólo con respecto a qué población, sino también con respecto a cuál o cuáles variables o características.

La muestra debe tener las mismas características que posee la población, su estructura debe ser análoga a la misma. Representatividad y aleatoriedad son características fundamentales de toda muestra seleccionada (Kelmansky, 2009).

A cada uno de los elementos que componen la población se los denomina **unidad de muestreo** y el **marco de referencia** es el registro o listado de las unidades de referencia que sirven como base para determinar cuáles unidades integrarán la muestra.

El procedimiento que otorga mayores garantías de representatividad de la muestra es su conformación al azar, es decir siguiendo algún procedimiento según el cual se eligen aleatoriamente aquellos elementos de la población que se incluirán en la muestra.

Los valores estadísticos que se obtienen estudiando poblaciones completas se llaman **parámetros** como media aritmética, desvío estándar, varianza. Los valores estadísticos que se obtienen de las muestras reciben el nombre de **estadísticos**.

La estadística inferencial permite la estimación confiable de los parámetros gracias al conocimiento de los estadísticos siempre y cuando se haya realizado una correcta selección de la muestra.

La teoría sobre las muestras también se utiliza para poder determinar si las diferencias observadas entre dos muestras son realmente debidas al azar o si son significativas, lo que permite por medio de un proceso la toma de decisiones por medio de las pruebas de hipótesis que se pueden hacer.

El relevamiento total de una población se denomina **censo** y el trabajo sobre muestras **muestreo**. Las conclusiones obtenidas a partir de las muestras tienden a:

- **Probar hipótesis**, válidas para la población correspondiente conociendo la información de las muestras,
- **Estimar parámetros**, ciertas características de la población, o sea los parámetros a través del conocimiento de las mismas características en los valores de la muestra, que son los estadísticos.

Toda estadística de las muestras es una estadística inferencial, pues se infiere a partir de los valores estadísticos hallados en las muestras los valores paramétricos más probables para las poblaciones de las cuales hemos extraído las muestras.

Se diferencian dos tipos de muestreo: el muestreo **probabilístico** y el muestreo **no probabilístico**. Hay un tercer tipo de muestras y son aquellas que se lesiona para probar hipótesis sustantivas (Padua, 2018).

En el muestreo **probabilístico**, cada elemento de la población tiene una probabilidad conocida y no nula de ser seleccionado. Esta probabilidad está determinada por el diseño o tipo de muestreo.

Sólo con el muestreo probabilístico corresponde aplicar los procedimientos de estimación de parámetros y de prueba de hipótesis, y la ventaja principal reside en que puede medirse el tamaño del error en nuestras predicciones. Puede decirse que el principal objetivo en el diseño de una muestra probabilística es reducir al mínimo este error al que se lo llama **error estándar**.

En el muestreo **no probabilístico**, la elección de los elementos no depende de la probabilidad, ya que no se conoce el marco muestral, sino de causas relacionadas con las características del investigador o del que hace la muestra. Aquí el procedimiento no se basa en fórmulas de probabilidad, sino que depende del proceso de toma de decisiones de una persona o grupo de personas, y desde luego las muestras se seleccionan por decisiones subjetivas y tienden a estar sesgadas.

Muestreo probabilístico

Elegir entre una muestra probabilística o una no probabilística, depende de los objetivos del estudio, del esquema de investigación y la contribución que se piensa hacer con ella. Las técnicas de muestreo permiten seleccionar un conjunto de unidades de análisis de tal modo de obtener la máxima aproximación a los parámetros poblacionales compatible con restricciones de costo y tiempo (Baranguer, 2006)

Dentro del muestreo aleatorio o probabilístico, se encuentran:

- muestreo simple al azar
- muestreo sistemático
- muestreo estratificado (proporcional o no proporcional)
- muestreo por conglomerados

Muestreo aleatorio simple o simple al azar

$$p = \frac{1}{N}$$

Para realizar un muestreo aleatorio simple es necesario contar con el listado o registro de todos los elementos que componen la población, como marco de referencia. Las unidades de muestreo coinciden con los elementos de la población y todos deben tener la misma probabilidad de ser incluidos en la muestra. Del listado se extraen al azar algunos elementos de la población, tanto es como el número de casos que se desea para la muestra. Cada una de las unidades de análisis de los N casos de la población, debe tener la misma probabilidad (p) de poder ser elegida y pasar a ser una unidad de observación al integrar la muestra:

Se pueden usar una serie de medios para asegurar iguales probabilidades de ocurrencia para cada uno de los elementos de la lista. Se utilizan tablas de números aleatorios, que es un equivalente de un sorteo y consiste en una serie de números dispuestos al azar en hileras y columnas. También, programas informáticos como InfoStat, SPSS o similares, permiten obtener listas de números aleatorios.

La ventaja de este tipo de muestreo es que no se necesitan conocer las características previas de la población. No tienen sesgo las muestras; las estimaciones son simples de realizar pues los errores estándar siguen distribuciones de probabilidad conocidas, y cuando la cantidad de casos que integra la muestra, su representatividad se hace cada vez más grande.

La desventaja es que se requiere conocer el listado completo de los elementos de la población y ello, a veces resulta muy oneroso y, directamente, no se puede obtener.

Muestreo sistemático

Es una variante del muestreo por azar simple. El muestreo por azar sistemático también requiere disponer de un listado completo de todos los miembros de una población y nos asegura que para la selección se ha consultado y recorrido todo el listado completo. De igual manera, se debe saber cuántos elementos se desea que integren la muestra. Entonces, esta técnica consiste en determinar un intervalo de muestreo (k) que surge de dividir el total de la población por el número de la muestra (N/n) luego dentro de cada zona se selecciona una unidad que ocupa el mismo lugar en todas ellas. El primer elemento es seleccionado al azar dentro del primer intervalo k , el segundo se obtiene sumando k al primer elemento seleccionado y así sucesivamente hasta completar la muestra.

Muestreo estratificado

Consiste en dividir la población completa de unidades de muestreo en sus poblaciones que se denominan estratos. Luego dentro de cada estrato se selecciona una muestra aleatoria simple al azar o sistemática de las unidades de muestreo que conforman este estrato. Las unidades de muestreo que componen la población deberán figurar en uno y sólo uno de los estratos.

Se construyen estratos homogéneos dentro de sí y heterogéneos entre sí. La selección de la muestra en cada estrato puede realizarse proporcionalmente al número de casos de cada estrato. El porcentaje de casos representados en cada estrato debe respetarse en la muestra que se selecciona. Para tal fin podemos multiplicar el tamaño de grupo por el tamaño de la muestra y dividirlo por el total de la población.

Por el contrario, en el muestreo estratificado no proporcional se emplean fracciones de muestreo diferentes en distintos estratos. Se toman fracciones mayores cuando se sabe que la variabilidad de la variable en estudio dentro del estrato es mayor y fracciones menores cuando se sabe que es menor.

El muestreo estratificado requiere conocer la frecuencia relativa con que se dan las principales características de la población, en cada uno de los estratos, para que la muestra pueda estar libre de sesgos.

Muestreo por conglomerados

Este muestreo consiste en la extracción aleatoria de grupos de elementos o conglomerados. Para obtener una muestra por conglomerados se divide la población en grupos de elementos, pueden ser áreas o zonas geográficas, manzanas de una ciudad, clubes deportivos (para una muestra de deportistas) etc. luego se selecciona entre estos aleatoriamente y los elementos que componen los conglomerados elegidos forman todos parte de la muestra. Así, los conglomerados resultan heterogéneos dentro de sí y homogéneos entre sí.

Muestreo No probabilístico

Dentro del muestreo no probabilístico, se encuentran las siguientes técnicas:

- muestreo accidental o casual
- muestreo intencional
- muestreo por cuotas

Muestreo accidental o casual

La muestra se conforma con aquellos sujetos de la población a los que es más fácil acceder, en una situación dada, se encuentran accidentalmente a disposición. Es común que se tomen grupos de sujetos que se han reunido o conformado con anterioridad a la investigación y sin que nada haya tenido que ver en esto el investigador. Por ejemplo: alumnos de cierta materia, miembros de un club, encuestas en la vía pública etc.

Se utilizan generalmente en las fases iniciales de algunos estudios donde se necesita reunir información con el objeto de llegar a formular algunas hipótesis o mejorar el planteo de un problema.

Los resultados obtenidos mediante muestreo casual o accidental no son representativos de la opinión de toda la población, sino que el resultado es orientativo.

Muestreo intencional

Se basa en la selección de un subconjunto de elementos siguiendo un criterio que el investigador subjetivamente considera representativos de la población. Se seleccionan en base a sus conocimientos previos, pero esto no garantiza la representatividad. Es apropiado para investigaciones en su etapa exploratoria.

Muestreo por cuotas

El investigador en base a su criterio especifica cuotas en función de determinadas variables demográficas (edad, sexo, etc.) luego instruye a encuestadores para que completen dichas cuotas. Al completarlas éstos tienen cierta flexibilidad para encuestar a personas o familias (elementos) que se ajustan a las variables especificadas en la cuota.

Muestreo bola de nieve

La cantidad de casos que conforman la muestra va creciendo a medida que avanza la misma investigación. Se aplica para abordar temas complejos en lo que respecta a ser poco comunes o de difícil acceso a los datos; entonces, unos sujetos de la muestra brindan información sobre otros posibles que puedan tener las mismas características que buscamos encontrar en las unidades de análisis. Por ejemplo, si nuestro estudio es sobre sujetos cuya talla supera los 2 m, o edades superiores a los 100 años (Morales Vallejo, 2012).

Muestras para probar hipótesis sustantivas

Muchas veces, el interés de un estudio sobre una muestra no es para la estimación de parámetros sino, justamente, para probar hipótesis. El énfasis en la selección de las unidades de observación que integrarán la muestra debe atender a que contengan el o los elementos que interesa relacionar por lo que la muestra debe ser lo suficientemente heterogénea. El tamaño dependerá de cuántas variables se desean relacionar y de las condiciones que tengan los estadísticos de prueba que se apliquen.

Tamaño de la muestra

Se determina en función de los objetivos de la investigación y de los recursos disponibles. Debe entenderse que no hay ninguna regla para asegurar rápidamente el número de casos de una muestra, sino que depende de varios factores que deben ser considerados a la hora de determinar el tamaño de la muestra. Es sabido que cuanto mayor sea el requerimiento de confiabilidad de los resultados muestrales, mayor deberá ser el tamaño de la muestra (Baranguer, 2009).

El investigador debe responder a tres preguntas para decidir el tamaño de su muestra: ¿cuántas variables quiere se quieren analizar simultáneamente?; ¿cuál es el número máximo de valores que se utilizará por variable?, y ¿cuál será el valor mínimo de cada variable numérica? (Padua, 2018).

Veamos algunos casos:

Tamaño de Muestras para verificar hipótesis estadísticamente sustantivas

Para determinar la probabilidad de que un valor estadístico muestral no difiera significativamente del parámetro poblacional, hay que averiguar qué porcentaje de elemento de una población presenta el atributo x. Un plan de muestreo dependerá del nivel de confianza (ejemplo 95%) con que queremos que el parámetro no se aparte del valor muestral. Sin embargo, más allá del número de casos seleccionados, nada garantiza de que la muestra que hayamos tomado sea una de esas 5 entre 100, que nos conducirán a un posible error. El número no debería ser menor a 40 casos (Baranquer, 2009).

Tamaño de Muestras no probabilísticas

Por definición, son muestras que pueden ser sesgadas; no representativas de la población de la cual se extraen. El número de casos deberá cumplir las expectativas mínimas del objetivo diseñado en la investigación (exploratorio, descriptivo, etc).

Tamaño de Muestras probabilísticas

Se seccionan para extrapolar resultados muestrales a la población. Se considera que una muestra conformada por el 20% de la cantidad de casos que tiene la población de la que se extrae, garantiza un número lo suficientemente grande como para asegurar que la población tiene la misma característica muestral (Morales Vallejo, 2012).

La estadística inferencial proporciona las técnicas que permiten realizar afirmaciones sobre más elementos de los que ya se han medido, pero, justamente, como no se han medido todos los elementos sino sólo los de una parte de la población, los que integran la muestra. Esto conlleva a entender que siempre habrá discrepancias entre el valor descriptivo muestral y el parámetro poblacional.

Resulta necesario definir algunos conceptos al momento de seleccionar una muestra probabilística (Maureira Cid, 2017):

Nivel de confianza. Establece el grado de certeza de que ocurra un evento Recordemos que la probabilidad se mide entre los valores 0 y 1, siendo 0 la certeza absoluta de que un evento no se produzca, y 1 es la absoluta confianza que dicho evento ocurra. Los niveles de confianza más utilizados en ciencias sociales son 0,95 y 0,99.

Nivel de significación, o error de tipo I o alfa. Los niveles habituales son de 0,05 (5% de error) y 0,01 (1% de error). Es el error máximo de rechazar la hipótesis nula cuando ésta es verdadera. Se indica con la letra griega alfa, α .

Error muestral máximo permitido en la estimación. Es el valor de equivocación que aceptamos para los estadísticos de la muestra al ser extrapolados con los parámetros de la población. Los niveles habituales son de 0,03 (3% de error) y 0,02 (2% de error) y nunca superior al 5% (0,05). A medida que disminuye el error permitido, se debe aumentar el tamaño de la muestra. Cuando usamos una muestra, usamos una parte de la población para luego estimar una característica de la muestra a toda la población. Hay un error que es ajeno al muestreo que se comete al estimar el parámetro. En las fórmulas se representa con la letra e ; se expresa como probabilidad, ejemplo: $e=0,03$.

Diferenciar entre poblaciones finitas e infinitas. Por otra parte, debe tenerse en claro si se tiene conocimiento del número de casos de la población, o no. Ejemplo: un estudio sobre estudiantes de la Universidad es una población infinita: más de 100.000 casos pero tiene un número conocido de cantidad de estudiantes para Baranguer (2009) o 30.000 para Padua (2018).

Establecer si se conocen o no los valores paramétricos de media y desvío, y la forma de la distribución de la población de la cual se extrajo la muestra.

Ejemplo, en el informe elaborado en 2022 por el INTI³¹, se estableció un $e=0,01\%$; $\alpha=0,05$ y un nivel de confianza del 95%.

Como se podrá entender, calcular el tamaño de una muestra no es cosa rápida, hay muchas consideraciones a tener en cuenta y, por lo tanto, no hay una única fórmula para el cálculo de la cantidad de casos de una muestra. Una vez que se consideran todos estos aspectos, debe buscarse la fórmula adecuada para calcular el número mínimo de casos que debería tener nuestra muestra. Hay alternativas al cálculo (Morales Vallejo, 2012) y es usar algún enlace en la web que calcula el tamaño de la muestra a partir de los datos que aportamos³².

Fórmulas para determinar el tamaño de una muestra

Para el cálculo del tamaño de una muestra, hay que distinguir entre poblaciones infinitas (número de casos muy grande o una cantidad desconocida o incontables) y finitas (un número conocido de casos).

³¹ Informe antropométrico argentino realizado por el Instituto Nacional de Tecnología Industrial, INTI, disponible en https://www.argentina.gob.ar/sites/default/files/2023/10/estudio_antropometrico_ar_-_informe_tecnico.pdf

³² Enlaces web para el cálculo del tamaño de una muestra: http://www.macorr.com/ss_calculator.htm o <http://www.raosoft.com/samplesize.htm>

Tamaño de una muestra de poblaciones infinitas³³

$$N = \frac{z^2 pq}{e^2}$$

Establecido un nivel de confianza del 95% central (es decir: $\alpha=0,05$), los equivalentes valores de la puntuación z, de la Distribución Normal es $z = \pm 1,95$.

Los valores de p y q son dicotómicos; corresponden a la varianza poblacional, que, al ser valores desconocidos, consignamos para p, el mayor valor proporcional (que se indica entre 0 y 1) que se supone con que se presenta una característica en una categoría. El valor de $q=1-p$. Se suele tomar un valor de $p = q = 0,50$ (la mitad de los casos presentan la característica en estudio; y la otra mitad, no).

El error muestral se designa con la letra e. Es el margen máximo de error que establecemos al extrapolar los resultados muestrales a la población.

Ejemplo: ¿cuántos casos deben conformar una muestra cuando se establecen los siguientes valores?:

Nivel de confianza del 95% (es decir, $\alpha = 0,05$) y, por lo tanto, $z = \pm 1,95$.

Margen de error del 2% ($e = 0,02$)

Suponemos que la característica que estemos analizando se distribuye igualmente proporcional; luego, $p=q=0,05$

Al aplicar la fórmula queda de la siguiente manera:

$$N = \frac{(1,96)^2 * 0,5 * 0,5}{(0,02)^2}$$

$$N = \frac{3,8416 * 0,25}{0,0004}$$

$$N = \frac{0,9604}{0,0004}$$

$N=2401$ unidades de análisis deben integrar la muestra para ser observadas.

Si nos fijamos en la fórmula, siempre tendremos un aumento de la cantidad de casos seleccionados para conformar la muestra si aumenta el porcentaje del nivel de confianza y si disminuye el error e (Padua, 2018).

³³ Padua, J. (2018). *Técnicas de investigación aplicadas a las Ciencias Sociales*. México: Fondo de Cultura Económica. Cap. III, p. 5

Tamaño de una muestra de poblaciones finitas³⁴

$$n = \frac{N}{1 + \frac{e^2(N-1)}{z^2pq}}$$

Donde,

n: tamaño de la muestra

N: tamaño conocido de la población

e: margen de error en la estimación

z: valor del puntaje z correspondiente a un determinado porcentaje de nivel de confianza

p, q: proporción (de 0 a 1) con que se presenta en la población la característica dicotómica analizada.

Ejemplo: ¿Cuántos casos deben seleccionarse en para conformar una muestra?:

Tamaño conocido de la población de 600 estudiantes (N=600)

Nivel de confianza del 95% (es decir, $\alpha = 0,05$) y, por lo tanto, $z = \pm 1,95$.

Margen de error del 3% ($e = 0,03$)

Suponemos que la característica que estemos analizando se distribuye igualmente proporcional; luego, $p=q=0,05$

Al aplicar la fórmula queda de la siguiente manera³⁵:

$$n = \frac{600}{1 + \frac{.03^2(600-1)}{(1,96^2)(.25)}} = 384; \text{ necesitamos por lo tanto una muestra de 384 sujetos.}$$

La siguiente Tabla³⁶, ya tiene realizados los cálculos para determinar la cantidad mínima de casos de una muestra cuando se conoce la cantidad de casos de la población, para un $\alpha = 0,05$.

Recordamos que un profesional de la EF, de manera frecuente, trabaja con grupos pequeños, ya sea en el Aula, en un Gimnasio, en un Club.

³⁴ Padua, J. (2018). *Técnicas de investigación aplicadas a las Ciencias Sociales*. México: Fondo de Cultura Económica. Cap. III, p. 10.

³⁵ Padua (2018). ob. cit. Cap. III, p.10.

³⁶ Padua (2018). ob. cit. Cap. III, p.11.

Figura 9.1*Error máximo según el tamaño de la muestra*

<i>Tamaño de la población</i>	<i>error tolerado</i>		<i>Tamaño de la población</i>	<i>error tolerado</i>	
N	e = .05	e = .03	N	e = .05	e = .03
40	36	38	32	30	31
39	35	38	31	29	30
38	35	37	30	28	29
37	34	36	29	27	28
36	33	35	28	26	27
35	32	34	27	25	26
34	31	33	26	24	25
33	30	32	25	24	24

Nota. Fuente: cuadro adaptado de <https://rpubs.com/DiegoTrevino/981739>

Estimación de parámetros

El objeto de realizar estudios a través de muestras es conocer valores de una población, de donde ha sido extraída la muestra.

Los valores paramétricos pueden conocerse con precisión a través de estudios donde se observa toda la población (censo). Pero en general los estudios no se realizan con toda la población y para las infinitas variables que pueden ser de interés. No se conocen los correspondientes valores paramétricos.

Por lo tanto, se tratará de estimarlos a través de muestras. Si la muestra es aleatoria y ha sido bien seleccionada, el valor estadístico seguramente será semejante al valor paramétrico, estará próximo a él. Así, se emplea el valor estadístico, por ejemplo, la media muestral, para estimar el correspondiente parámetro μ , básicamente, de dos maneras: puntual o por intervalos.

Cómo se obtiene una muestra se relaciona con la representatividad de la misma, y el tamaño de una muestra debe tener un número tal que permita extrapolar los resultados con diferencias mínimas.

Estimación puntual

Mediante este procedimiento, se asigna al parámetro de la población un valor concreto. Los mejores estimadores puntuales de la media poblacional son la media y desviación estándar de una distribución muestral normal.

El valor calculado en la muestra se transforma en el **estimador puntual** obtenido en la muestra para representar a la población de la cual se obtuvo la muestra. El valor de los estadísticos muestrales de la media del desvío estándar es considerado como parámetros poblacionales.

Por ejemplo, en el informe del INTI, la estimación puntual del contorno de pecho en cuerpos de varones mayores de 18 años, de Argentina, es de 106,4 cm.

Una estimación se puede realizar si la muestra es probabilística y tiene un número grande de casos.

Estimación por intervalos

No siempre se considera verosímil que un estimador muestral sea el valor paramétrico exacto. Para este caso, los valores paramétricos se establecen no como un valor puntual sino como posible ser observado dentro de un determinado rango de valores. Suele calcularse sumando y restando al valor puntual, un cierto margen de error; luego, el parámetro quedará determinado dentro de ese rango alrededor del valor estimado como cierto, llamado **intervalo de confianza**.

Se trata de obtener dos valores equidistantes de un valor central, dentro de cuyos límites vamos a afirmar que se encuentra el verdadero parámetro, con cierta probabilidad en dicha estimación.

La amplitud del intervalo de confianza dependerá del nivel de confianza ($1-\alpha$), de la varianza poblacional (σ) y del tamaño de la muestra (n). Los límites del intervalo de confianza brindan un rango de probabilidad alta de que el parámetro se encontrará dentro de esos valores. Por lo tanto, la fórmula para la estimación del parámetro por intervalos no es única; hay varias y su elección depende de conocer o no la varianza poblacional y la cantidad de casos de la muestra.

Ejemplo, siguiendo con el informe del INTI, la estimación por intervalos del ancho de caderas en cuerpos de mujeres mayores de 18 años, de Argentina, se encuentra dentro del rango 102-105 cm.

Prueba de hipótesis

Es un procedimiento que busca contrastar enunciados que se hacen sobre la población. Es un tipo de razonamiento que permite evaluar cuán probable es que los resultados observados en una muestra puedan ser representativos de la población a la cual pertenece (Bologna, 2011).

Ejemplo, si la media muestral del contorno de busto en cuerpos de mujeres argentinas entre 20-49 años es efectivamente de 99,8 cm; ¿podría afirmarse que es un estimador del parámetro poblacional?

El procedimiento sigue siempre la misma lógica: se establece lo que denominamos **hipótesis nula**, un enunciado de que, a nivel poblacional, el contorno de busto es diferente a 99,8 cm. Si el estudio es correlacional, la hipótesis nula formulará que no hay relación entre las variables cuya asociación sí se encontró en la muestra, por ejemplo, entre el tiempo de entrenamiento físico y el porcentaje de lesiones sufridas durante una temporada regular de un deporte practicado.

Junto a la hipótesis nula, se formula la **hipótesis alternativa**, el enunciado que postula lo contrario, afirma que, a nivel poblacional, sí hay relación entre las variables estudiadas; o que el contorno de busto es de 99,8 cm.

Es decir, se trata de establecer si el resultado muestral no se aleja tanto de lo que esperaríamos si la hipótesis nula fuera cierta, por lo tanto, no podemos rechazarla. Se espera concluir que sí sea rechazada y la alternativa sea el nuevo conocimiento aceptado para la población de la cual se obtuvo la muestra.

Recordemos, llamaremos error de Tipo I o alfa (α), a la probabilidad que estamos dispuestos del rechazar la hipótesis nula siendo ésta verdadera. Es decir, concluimos que hay una relación entre dos variables o que un estadístico se acepte como parámetro cuando en realidad no debía concluirse eso sino lo contrario.

El valor de esta probabilidad máxima usada en ciencias sociales es de 0,05. Esto interpreta de la siguiente manera, de cada 100 muestras diferentes que pudieran extraerse de una población para realizar un mismo estudio, a lo sumo en 5 de ellas podríamos encontrar valores encontrados por mero azar, encontrados por casualidad en cómo nos quedó conformada la muestra y no serían representativos de la realidad poblacional.

Este error se debe diferenciar de aquel que se comete al estimar el valor poblacional pues esa variación se define como una diferencia esperable entre el estimador y el verdadero parámetro.

El error de Tipo II o beta (β) es la probabilidad de no haber rechazado la hipótesis nula cuando sí debíamos haberlo hecho, es creer que es verdadera cuando en realidad, no lo es.

El nivel de significación alfa (α) se utiliza también para poder determinar si la Hipótesis Nula se rechaza o no. Se compara el valor que se ha establecido como nivel de significación con un valor llamado **p-valor o p**.

El **p-valor** es una probabilidad, la de ser verosímil la Hipótesis Nula. El *p valor* lo generan de manera automática los programas estadísticos computacionales y aparecen asociados a cualquier valor calculado de un coeficiente estadístico descriptivo (chi cuadrado, r de Pearson, ro de Spearman, etc.).

Al ser una probabilidad, el p valor es una medida que se expresa entre 0 y 1. Cuanto más pequeño es el p valor, más chica es la posibilidad de que la Hipótesis Nula sea cierta. ¿Cómo saber si es un valor pequeño y poder entonces rechazar la Hipótesis Nula? Será un valor pequeño si no supera al nivel de significación fijado previamente.

Partimos del supuesto de que la Hipótesis Nula planteada en un procedimiento de Prueba de Hipótesis, es cierta. Para poder rechazar e imponer un conocimiento nuevo que se expresa en la Hipótesis Alternativa, nos centramos en el valor **p** obtenido por el software estadístico que se esté usando (su cálculo es complicado y supera el interés de este Curso): **si el valor p es menor que el nivel de significación alfa, se rechaza la Hipótesis Nula planteada** y se puede afirmar que, a un cierto nivel de significación (ejemplo: $\alpha=0,05$), es posible observar en la población de la cual obtuvimos la muestra en análisis, los mismos hallazgos que postula la Hipótesis Alternativa.

Si valor **p < 0,05** decimos que un **resultado es estadísticamente significativo** cuando permite rechazar la Hipótesis Nula y nos da información sobre la población (en general inaccesible) de la cual obtuvimos la muestra con la que estamos trabajando. Si $p < 0,05$ nos indica que es muy poco probable que los resultados que describen la muestra en estudio, hayan sido obtenidos por mero azar; es decir, si el valor **p < 0,05** se entiende que es muy poco probable cometer un Error de Tipo I o alfa (rechazar la Hipótesis Nula cuando es verdadera).

Un resultado estadísticamente significativo ($p < 0,05$) indica que te tiene la confianza de encontrar los mismos resultados que postula la Hipótesis Alternativa en al menos el 95% de las diferentes muestras que se extraigan de la misma población bajo estudio.

Ejemplo: A fin de mejorar la salud cardiovascular de un grupo de adultos en rehabilitación, se implementaron actividades de caminata, nado y bicicleta. Luego de un tiempo se evalúa el programa de cardio y se desea saber existe relación entre las actividades propuestas y la mejora o no de la salud cardiovascular.

Las variables son:

x: mejora de la salud cardiovascular (sí/no)

y: actividades del programa de cardio (caminata/nado/bicicleta)

Las hipótesis de investigación estadística quedan formuladas de la siguiente forma:

H_0 : La salud cardiovascular es independiente del programa de cardio que se realice.

H_1 : La salud cardiovascular no es independiente del programa de cardio que se realice.

Dado que las variables son cualitativas y del nivel nominal, se aplica la Prueba de Independencia Chi cuadrado. Para el análisis de los datos se utiliza el software estadístico InfoStat para un nivel de significación $\alpha=0,05$ (Bologna, 2011). Los resultados son los siguientes:

Estadístico	Valor	gl	p
Chi Cuadrado Pearson	7,00	2	0,0302

Esto significa:

- el valor calculado de Chi cuadrado es 7,00
- los grados de libertad son $gl=2$
- el valor de p asociado a ese Chi cuadrado calculado es 0,0302

Se rechaza la Hipótesis Nula para un valor $\chi^2=7$ (2;0,05) con una $p=0,0302$ asociada. Dado que el valor $p < 0,05$ podemos afirmar que los resultados son estadísticamente significativos; es decir, hemos encontrado una relación entre la mejora de la salud cardiovascular y el programa de actividades cardio que se realizan.

Referencias

- Baranger, D. (2009). "Muestra y muestreo", en *Construcción y análisis de datos. Introducción al uso de técnicas cuantitativas en la investigación social*. Cap. 3 (pp. 31-42). Misiones: Editorial Universitaria de Misiones. Disponible en: https://www.academia.edu/3793040/Construcci%C3%B3n_y_an%C3%A1lisis_de_datos_Introducci%C3%B3n_al_uso_de_t%C3%A9cnicas_cuantitativas_en_la_investigaci%C3%B3n_social
- Kelmansky, D. (2009). *Estadística para todos. Estrategias de pensamiento y herramientas para la solución de problemas*. Caps. 5 y 6 (pp. 25-40). Ministerio de Educación de la Nación. Instituto Nacional de Educación Tecnológica. Disponible en: <http://www.bnm.me.gov.ar/giga1/documentos/EL001858.pdf>
- Morales Vallejo, P. (2012). *Tamaños necesarios de la muestra ¿Cuántos sujetos necesitamos?* Madrid: Universidad Pontificia Comillas. Disponible en: https://www.academia.edu/5826247/Tama%C3%B1o_necesario_de_la_muestra_Cu%C3%A1ntos_sujetos_necesitamos_Tama%C3%B1o_necesario_de_la_muestra_Cu%C3%A1ntos_sujetos_necesitamos
- Padua, J. (2018). *Técnicas de investigación aplicadas a las Ciencias Sociales*. Cap. 3. México: Fondo de Cultura Económica. Disponible en: <https://ie42003cgalbarracin.edu.pe/biblioteca/LIBR-NIV312012023164926.pdf>

Quiénes escribimos este libro

Coordinadores

Doná, Stella Maris

Especialista en Metodología de la Investigación Científica, Especialista en Docencia Universitaria, Licenciada en Psicología y maestranda en Metodología de la Investigación Científica; Ayudante diplomada ordinaria en la cátedra de Estadística Aplicada a la Educación Física desde 2005, luego se desempeñó como Jefa de Trabajos Prácticos ordinaria y, actualmente, es la Profesora Adjunta interina a cargo. Asimismo, es Profesora Adjunta interina en la Cátedra de Estadística Aplicada a la Psicología (Facultad de Psicología, UNLP) y Docente a cargo del Seminario de Estadística Aplicada a la Educación Física en la Universidad Nacional de Rosario (UNR). Desde el año 2008 participa de Equipos de Investigación en la UNLP y continúa hasta la actualidad.

Fornessi, Román

Licenciado en Sociología (UNLP) y Doctor en Ciencias Sociales (UNLP). Desde el año 2017 es docente en la cátedra de Estadística Aplicada a la Educación Física, donde actualmente se desempeña como Jefe de Trabajos Prácticos. En posgrado, se ha desempeñado como docente del Seminario de Tesis de la Maestría en Ciencias Sociales (FaHCE-UNLP) y de la Maestría en Educación (FaHCE-UNLP). Participa del proyecto de extensión “Hacia clubes inclusivos. Jugar, gestionar, entrenar y dirigir con perspectiva de género para erradicar las violencias” (FaHCE-UNLP) y de los proyectos de investigación “Territorio y producción. Estudio de la economía mixta del Gran La Plata” (FaHCE-UNLP), y “Divulgación e intervención en debates públicos de y desde la sociología en Argentina (1983-2024)”.

Autores

Castillo, Gisela Marcos

Profesora en Educación Física (UNLP), Especialista en Entrenamiento Deportivo y Readaptación Físico Deportiva. Adscripta a la cátedra de Estadística Aplicada a la Educación Física desde el año 2024. Es Preparadora Física en Básquet y Coordinadora del Proyecto de Desarrollo de Jóvenes Atletas, en clubes amateurs de la Ciudad. Desde el año 2022, forma parte del Staff de Selecciones Nacionales de Básquet Formativo. Actualmente cursa la carrera de Ciencia de Datos en las Organizaciones y la Diplomatura en Ciencia de Datos (UNLP).

Cotignola, Mariela

Licenciada en Sociología, Magíster en Métodos y Técnicas de Investigación Social, Especialista en Gestión de la Educación Superior y Doctoranda en Educación. Formó parte, como Jefa de Trabajos Prácticos, de la cátedra de Estadística Aplicada a la Educación Física desde el año 2017. Actualmente, se desempeña como profesora adjunta en el Taller de Análisis Sociodemográfico y Construcción de Indicadores de la carrera de Sociología de la FaHCE y como Directora de Seguimiento de las Trayectorias Estudiantiles de la UNLP. Ha participado de proyectos de investigación de la FaHCE-UNLP desde el año 2005 vinculados al análisis del mercado de trabajo, los procesos educativos y las trayectorias estudiantiles.

Díaz, Martina

Profesora en Educación Física (UNLP), desde el año 2018 es docente en la cátedra de Estadística Aplicada a la Educación Física, en el cargo de Ayudante Diplomada Ordinaria. Actualmente, participa como coordinadora del área de Expresión y Movimiento del Programa de Educación Permanente para Adultos Mayores (PEPAM) de la Secretaría de Extensión de la FaHCE-UNLP. También es docente en nivel secundario, donde se desempeña en tareas de gestión institucional.

Gandini, María Agustina

Profesora en Educación Física (UNLP) actualmente se desempeña como ayudante diplomada interina en la cátedra de Estadística Aplicada a la Educación Física y en la cátedra de Fisiología Humana, de las carreras de Educación Física de la FaHCE-UNLP. Participa en el proyecto de extensión “Salugénesis: construyendo salud comunitaria” y en el proyecto de investigación “Gestión de espacios de clases y enseñanza virtual en la formación universitaria”.

Girardi, Lucía

Profesora de Sociología (UNLP). Desde el año 2022 se desempeña como docente, con el cargo de ayudante diplomada, en la cátedra de Estadística Aplicada a la Educación Física (FaHCE-UNLP). Es docente de Metodología de la Investigación y Estadística para carreras técnicas del nivel terciario desde 2010. Actualmente cursa la Diplomatura en Ciencia de Datos (UNLP) y la Especialización en Producción y Análisis de Información para Políticas Públicas (UNC).

Villafañe, Rocío Belén

Profesora en Educación Física (UNLP). Adscripta a la cátedra de Estadística Aplicada a la Educación Física desde el año 2022 hasta la actualidad. Participa del proyecto de extensión “Educación Física: igualdad, inclusión, derechos” y “Educación Física: gimnasia para mujeres” (FaHCE-UNLP).

Fornessi, Román

Análisis estadístico en educación física : medidas, usos e interpretaciones / Román

Fornessi; Stella Maris Doná ; Coordinación general de Román Fornessi ; Stella Maris Doná. -
1a ed. - Libro digital, PDF - (Libros de cátedra)

Archivo Digital: descarga

ISBN 978-950-34-2727-9

1. Educación Física. I. Doná, Stella Maris II. Fornessi, Román, coord. III. Doná, Stella Maris,
coord. IV. Título.

CDD 370

Diseño de tapa: Dirección de Comunicación Visual de la UNLP

Universidad Nacional de La Plata – Editorial de la Universidad de La Plata

48 N.º 551-599 / La Plata B1900AMX / Buenos Aires, Argentina

+54 221 644 7150

edulp.editorial@gmail.com

www.editorial.unlp.edu.ar

EduLP integra la Red de Editoriales Universitarias Nacionales (REUN)

Primera edición, 2026

ISBN 978-950-34-2727-9

© 2026 - EduLP

S
sociales


EDITORIAL DE LA UNLP



UNIVERSIDAD
NACIONAL
DE LA PLATA