Cristóbal R. SANTA MARÍA y Claudia S. BUCCINO

ELEMENTOS DE PROBABILIDAD Y ESTADÍSTICA

UNIVERSIDAD NACIONAL DE MORENO

Elementos de probabilidad y estadística

Cristóbal R. Santa María
Claudia Soraya Buccino

Elementos de probabilidad y estadística

Cristóbal R. Santa María
Claudia Soraya Buccino

UNIVERSIDAD
NACIONAL
DE MORENO

Santa María, Cristóbal R
Elementos de probabilidad y estadística / Cristóbal Santa María ; Soraya Buccino. - 2a ed . - Moreno : UNM Editora, 2019.
Libro digital, PDF - (Biblioteca universitaria)
Archivo Digital: descarga y online
ISBN 978-987-782-010-2

1. Estadísticas. 2. Probabilidades. I. Buccino, Soraya. II. Título.
CDD 519.2

Departamento de Ciencias Aplicadas y Tecnología
Director General -Decano:
Jorge L. ETCHARRÁN

Colección: Biblioteca de Economía
Directores: Pablo A. TAVILLA y Alejandro L. ROBBA.

2a. edición: abril 2019

Av. Bartolomé Mitre 1891, Moreno
(B1744OHC), prov. de Buenos Aires, Argentina
(+54 237) 425-1619/1786, (+54 237) 460-1309,
(+54 237) 462-8629,
(+54 237) 466-1529/4530/7186, (+54 237) 488-3147/3151/3473
Interno: 154
unmeditora@unm.edu.ar
http://www.unm.edu.ar/editora
http://www.facebook.com/unmeditora/

ISBN (Edición digital): 978-987-782-010-2

La edición en formato digital de esta publicación se encuentra disponible en: http://www.unm.edu.ar/index.php/unm-virtual/biblioteca-digital http://www.unmeditora.unm.edu.ar/index.php/colecciones/biblioteca-universitaria

Libro de edición argentina
Queda hecho el depósito que marca la Ley 11.723

Libro Universitario Argentino

Presentación

Esta presentación reviste un carácter especial por ser Elementos de probabilidad y estadística la primera de las publicaciones del Departamento de Ciencias Aplicadas y Tecnología, especialmente dirigida a los estudiantes de la carrera de Ingeniería en Electrónica de la Universidad. El libro es el resultado del trabajo realizado por los docentes Cristóbal Santa María y Claudia Soraya Buccino de la asignatura Probabilidad y Estadística. Es necesario destacar que un equipo docente, además de su tiempo frente a curso, tiene la responsabilidad de la producción académica. Es decir, su tarea también es continuar con la formación en el aula y ampliar los conocimientos de los estudiantes a través de la investigación y la elaboración, que se deben volcar en producciones como la presente. Para la Universidad Nacional de Moreno adquiere una relevancia incuestionable ya que dentro de sus objetivos está el de contar con docentes que produzcan textos que sean el resultado del estudio, investigación, reformulación de hipótesis y reelaboraciones a partir de la reflexión acerca de su propia práctica. En tal sentido, el libro reúne los conocimientos básicos imprescindibles para entender los conceptos fundamentales de Probabilidad y Estadística que utilizarán los técnicos e ingenieros en Electrónica que egresen de nuestra Universidad. Asimismo, el texto tiene la riqueza y el acervo científico pertinente para que pueda ser utilizado en las otras tecnicaturas y carreras del Departamento de Ciencias Aplicadas y Tecnología, lo cual otorga mayor importancia a la obra.

Jorge L. ETCHARRÁN

Director Decano del

Departamento de Ciencias Aplicadas y Tecnología

Índice

Prólogo

La teoría de las probabilidades y la estadística ocupa hoy un lugar central entre el conjunto de conocimientos matemáticos que permiten abordar el modelado de los sistemas tecnológicos. Nuestra época se caracteriza por el reconocimiento de los factores aleatorios que gobiernan las relaciones entre causas y efectos en cualquier dominio del conocimiento, expresados a través de la recolección de una enorme cantidad de datos que requieren procesamiento para ser útiles como información. En este sentido, los conceptos básicos que aquí exponemos son las herramientas que posibilitan interpretar esas relaciones y construir esos procesos. Proporcionan en definitiva un lenguaje que intenta estar libre de inconsistencias lógicas y a su vez hacer más simple la exposición y la comprensión de ideas así como también facilitar el cálculo. Ese es, precisamente, el papel de la matemática en la Ingeniería: ser el lenguaje con el que se entiende, se diseña y se calcula, potenciando así —valga la redundancia— el "ingenio" del ingeniero.

Este libro surgió naturalmente como resultado de la preparación de las clases que los autores hemos dictado en distintos centros de enseñanza universitaria dentro de los planes de estudio de carreras de ingeniería. A lo largo de ese camino, hemos tenido que adaptar tópicos y metodologías de enseñanza y aprendizaje a distintas modificaciones, propuestas para ajustar el esfuerzo del estudiante y la duración de los planes de carrera a la amplia exigencia de contenidos que plantea el desarrollo tecnológico actual. Creemos que tales modificaciones no siempre tuvieron en cuenta la importancia relativa de la materia respecto del total de matemática que debe aprender un ingeniero. Así, ha operado una paulatina reducción de los tiempos de clases asignados que, entre otras causas, ha sido responsable de las dificultades que enfrentan los estudiantes para obtener madurez conceptual en los distintos puntos programáticos. De alguna manera, este libro trata de aportar un principio de solución a tales dificultades al restringir el temario y acotar su profundidad a los tiempos y posibilidades reales.

Al escribir cada capítulo, se ha procurado tener en cuenta no solo las

consideraciones realizadas en el párrafo anterior sino también el perfil sociológico y cultural de los jóvenes estudiantes actuales, inmersos de hecho en el mundo de las tecnologías de la información y la comunicación. Hemos observado que en ese mundo prima la comprensión de conceptos a través de imágenes y, ocasionalmente, de sonidos por sobre el discurso y el texto. Esto, que quizás resulte una ventaja para la captación inmediata de ideas en la relación social e interpersonal, puede a su vez constituirse en un obstáculo para aprehender conceptos cuyo carácter más abstracto requiere de la palabra y aun de la fórmula como forma de representación. Es relativamente sencillo comprender qué cosa es un lápiz viendo una imagen del mismo y a una persona empuñándolo para escribir; pero puede ser mucho más complicado entender a través de imágenes qué significan administración, burocracia o entropía. Creemos entonces que, para no limitar el universo de la comprensión conceptual, hay que poner énfasis en la palabra oral y sobre todo en la escrita. En tal sentido hemos tratado de presentar cada tema extremando el desarrollo del texto antes de pasar a la abstracción mayor que implica la fórmula y utilizando solo en forma ocasional imágenes, allí donde las consideramos imprescindibles para completar la comprensión. También hemos procurado evitar seudo razonamientos mnemotécnicos que intentan proporcionar "reglas" y hemos atenuado el uso de recuadros y resaltados de forma de inducir a la búsqueda del contenido en el interior del texto. La intención ha sido contraponer al hábito de la imagen el hábito del discurso y del texto, con la posterior abstracción matemática, para intentar ampliar las posibilidades de comprender. Nos parece oportuno señalar que tal marco metodológico de enseñanza surge, antes que por cualquier otra razón, por la naturaleza del contenido abstracto que posee el tema. En suma, el proceso de enseñanza y aprendizaje que se intenta no solo involucra la comprensión de los contenidos sino también la de los métodos para lograrlo.

Lo expuesto hasta aquí no debe hacer pensar que nos negamos al uso de las tecnologías de la información y la comunicación. Muy por el contrario, las consideramos una parte imprescindible en el proceso de enseñanza y aprendizaje tanto en lo que atañe a la inmediatez de la comunicación por vía de una plataforma virtual como en la simplicidad del cálculo para la resolución de problemas por medio de software. La utilidad en ambos casos está fuera de toda discusión, pero hay que señalar que el beneficio se produce una vez que se ha realizado un primer proceso de comprensión que, para cada tema, abarca al menos la lectura inicial del libro, la participación en clase y la realización de ejercicios conceptuales entendidos como "pruebas de escritorio". Luego de esto, la interacción con los docentes por vía virtual y el uso de software para resolución de problemas cuyo cálculo podría ser

más tedioso, complejo o simplemente largo, suele servir para afirmar y ampliar la comprensión y además prepara adecuadamente para el desempeño profesional. De este modo la eficiencia en el conocimiento y en la aplicación del software pasa a ser un contenido más de los que pueblan la materia.

Hemos puesto especial empeño en desarrollar cada tema tratando que de la exposición y la ejercitación surja la estructuración mental de los conceptos fundamentales de la disciplina. Para esto hicimos a un lado, en tanto fue posible, detalles, complejidades y aun formalidades que pueden ser de indudable interés para el matemático profesional, pero que raramente un ingeniero necesita tener en cuenta. Al establecer esta jerarquización temática procuramos aportar a la adquisición sólida de conocimientos básicos mediante un proceso que, entendemos, comprende diferentes etapas. Se trata primero de entender, luego de ejercitar y resolver problemas para ampliar y consolidar la comprensión, y finalmente, de memorizar, aspecto este último que no suele contar con buena prensa pero que juzgamos también imprescindible. No creemos además que la matemática, y en particular nuestro tópico, pueda enseñarse ni aprenderse solo buscando el ingenio de la resolución de problemas para construir internamente el saber descubriéndolo. Tal camino, útil quizá para despertar gustos y vocaciones, resulta insuficiente en relación con la necesidad de saberes profesionales y con los tiempos en que estos deben ser adquiridos de acuerdo con nuestra capacidad. En consecuencia, en las etapas de aprendizaje señaladas, proponemos una dosificación parcial de esta idea al apuntar simultáneamente a una conceptualización teórica básica y a un razonable desarrollo de la capacidad de modelado y cálculo, compatibles no solo con la curiosidad vocacional sino también con el requerimiento del ejercicio profesional.

Realizamos este libro con la idea de que la ciencia y su aplicación técnica bien orientada puede aportar grandes y decisivos beneficios al desarrollo social. Si bien sabemos que en el pasado y en la actualidad muchos ingenios tecnológicos han servido y sirven para iniciar y consolidar desigualdades e injusticias sociales o para agredir nuestro medio ambiente natural, resulta impensable recorrer un camino de ampliación de derechos, inclusión y desarrollo económico sin la columna vertebral que constituye la formación en ciencia y tecnología. Tal formación, que garantiza un grado necesario de independencia, debe constituir una preocupación permanente de toda la comunidad y no solo de quienes dedicamos a ello nuestro trabajo. En ese sentido, intentamos aportar a los planes de formación de ingenieros para los próximos años un libro acerca de la Probabilidad y la Estadística escrito en la lengua de nuestra sociedad y de acuerdo con las formas de discurso que ella emplea. Procuramos además que exprese los contenidos y también

la metodología didáctica que, entendemos, colabora con el cumplimiento de aquellos planes.

Finalmente, queremos agradecer, en primer término, a la Universidad Nacional de Moreno, en cuyo ámbito concebimos la idea de publicar nuestros apuntes de clase en forma de libro, que aportó además horas-cátedra necesarias para desarrollar el trabajo al inaugurar así una práctica que redunda en beneficio de los estudiantes y que potencia la actividad docente. También queremos agradecer al Ingeniero Aldo Sacerdoti, a la Magister Maria Eugenia Angel por sus sugerencias y comentarios sobre el enfoque de los temas y, muy especialmente, a todos los estudiantes que fueron usando el material en las clases proponiendo modificaciones y ampliaciones que mejoraran la didáctica.

Capítulo 1

Probabilidad elemental

1.1. Introducción

La teoría de las probabilidades y la estadística cobra cada día mayor interés y aplicación en una gran variedad de campos de la actividad humana. Esto es así porque sirven para modelar muy diferentes tipos de sistemas, explicar sus comportamientos —aún aquellos azarosos—, clasificar y predecir con cierta exactitud, controlar el desenvolvimiento de tareas y tomar decisiones.

La teoría de las probabilidades es de carácter matemático y su objetivo principal es medir la incertidumbre provocada por el azar. La Estadística se basa en la recolección de datos que procesa, a efecto de transformarlos en información importante para tomar decisiones. Uno de los pilares de este proceso estadístico es precisamente la teoría de las probabilidades, que permite cuantificar la confianza que podemos tener en la información obtenida.

En ingeniería electrónica, las técnicas probabilísticas y estadísticas se utilizan para estudiar la fiabilidad de los componentes, diseñar redes de comunicaciones o filtrar y procesar señales digitales, entre otras aplicaciones. En computación, las teorías de la información o la de las colas se fundamentan en la probabilidad y la estadística. La biología molecular y la bioinformática utilizan esas técnicas para descifrar y comparar patrones genéticos establecidos por secuencias de ADN. En marketing, la explotación de grandes bases de datos de clientes se realiza con algoritmos que "aprenden" a clasificar y predecir comportamientos y hábitos de consumidores, y que se basan en análisis de probabilidades y frecuencias estadísticas. En sociología, son estas mismas ideas matemáticas las que permiten modelar el proceso de nacimiento y muerte y determinar la esperanza de vida. En economía, para

citar un ejemplo, la proporción de consumos de una sociedad se establece a través de modelos estadísticos. Ramas enteras de la física se explican a partir del lenguaje probabilístico. En medicina se testea la efectividad de tratamientos, vacunas o se realizan estudios epidemiológicos por medio de ambas disciplinas. Cualquier búsqueda realizada en Internet es posible gracias a modelos probabilísticos sobre frecuencias de letras, palabras y frases. La lista sigue y la enumeración, lejos de ser exhaustiva, alcanza a dar una idea sobre la profusa aplicación y utilidad de ambas teorías.

Aunque no hay registros históricos del concepto de probabilidad en civilizaciones antiguas, sí hay indicios de algunas ideas que se habrían utilizado, con la intención de predecir el futuro o de ganar en juegos de azar. También en el Talmud, libro que recoge las discusiones de los rabinos sobre la ley judía, se utilizan nociones de probabilidad para justificar decisiones. En la Europa medieval se conocían ciertos conceptos elementales y hacia el siglo XVI algunos planteos sobre juegos de azar con iguales chances fueron analizados por el matemático Cardano. Sin embargo, se considera que la teoría comenzó a desarrollarse un poco más tarde, durante 1654, en las cartas intercambiadas por los matemáticos franceses Pascal y Fermat a propósito de dos problemas que le planteara a Pascal el Caballero de Méré. De Méré, un hábil jugador, quería saber qué cantidad de veces debía arrojar dos dados para tener la misma probabilidad de sacar al menos un doble seis que de no sacarlo. También buscaba saber cuál debía ser la división equitativa de las apuestas en un juego interrumpido antes de su conclusión. Pascal y Fermat supieron desarrollar formas para contar casos que permitieron luego evaluar y resolver estos y muchos otros problemas probabilísticos.

La actividad estadística humana comenzó en la prehistoria, y adquirió formas más elaboradas mucho más adelante. Se cree, por ejemplo, que el primer censo de población y tierras lo realizaron los egipcios unos cinco mil años antes de construir pirámides. Por esa época hubo también estadísticas en Babilonia e incluso hay registros griegos de censos hacia el siglo VI antes de Cristo. Mas aquí, alrededor del 1200 después de Cristo, los Incas, en lo que hoy es América, realizaban recuentos de habitantes y ganado. Los propósitos de todas estas actividades eran usualmente la contabilización de bienes y de quiénes debían pagar impuestos al estado. Precisamente el nombre "Estadística" surgió más modernamente, en 1749, en Alemania, con referencia al análisis de datos del estado. Los historiadores están de acuerdo en que la primera persona que se ocupó del problema de la inferencia –que consiste en establecer los comportamientos de una población a partir de una pequeña muestra conocida de individuosâ", fue el inglés Thomas Bayes, durante la primera mitad del siglo XVIII. Para nuestra materia este es un hecho

importante, pues registra el momento en que la probabilidad comienza a auxiliar al conocimiento estadístico. El marqués Pierre Simon de Laplace, ministro de Napoleón y uno de los grandes matemáticos de la historia, es quien explica adecuadamente la inferencia y expone los conocimientos existentes hasta entonces en su *Théorie analytique des probabilités* publicada en 1812. Allí, entre otras cosas, se explicita la definición de probabilidad con la que usualmente habían trabajado los matemáticos, por lo menos desde Pascal y Fermat.

1.2. Las definiciones de probabilidad

La llamada definición "clásica" o "de Laplace" establece que la probabilidad de un suceso se calcula como el cociente entre el número de casos en los cuales ese suceso pueda presentarse y la cantidad de casos posibles. Esto se resume en:

$$p = \frac{numero\ casos\ favorables}{numero\ casos\ posibles}$$

El cálculo no requiere la observación de experiencia alguna; basta con establecer previamente cuántos son los casos favorables y cuántos los posibles. Es un cálculo "a priori" de cualquier experiencia, basado en un conocimiento suficientemente objetivo. Ése cálculo capta la variabilidad intrínseca que puede ofrecer cada resultado entre todos los posibles. en otras palabras, cuantifica la incertidumbre.

Así, por ejemplo, la probabilidad de que una moneda al ser arrojada al aire caiga "cara", surge al dividir el número de casos en que puede caer cara, que es 1, sobre el número de resultados posibles, que son 2, cara y ceca.

$$p = \frac{1}{2}$$

Aparecen aquí algunos problemas lógicos. El primero de ellos es que hay una presuposición de que es igualmente posible que caiga cara o que caiga ceca. Si, por ejemplo, la moneda estuviera cargada, ceca y cara no tendrían igual probabilidad de salir. Es decir, serían casos no "igualmente posibles" y entonces el cálculo de la probabilidad p sería erróneo. Para que la definición valga, hay que exigir que los casos posibles sean "igualmente posibles", o lo que es lo mismo, "igualmente probables". Pero entonces, para definir

la probabilidad tendría que usarse un concepto previo de probabilidad sin definir, lo que revelaría una inconsistencia lógica. Además, hay un segundo problema. Supongamos que nos piden elegir un número natural cualquiera: ¿Cuál es la probabilidad de que elijamos uno par?. Más allá de cualquier intuición, si queremos aplicar la definición de Laplace para calcularla, tendríamos que dividir la cantidad total de números pares sobre la cantidad total de números posibles. La cuenta $p = \frac{\infty}{\infty}$ conduciría a un resultado indeterminado. Sin embargo, nos es intuitivamente cierto que la probabilidad buscada es $p = \frac{1}{2}$, habida cuenta de que la experiencia nos muestra que uno de cada dos números naturales es par. En un caso así, la definición clásica no sirve y entonces hay que buscar otra forma de calcular la probabilidad.

Es posible observar hechos que se repiten en condiciones uniformes o similares. Por ejemplo, en una habitación arrojamos sobre una mesa de juego, utilizando un cubilete, un dado cuyas caras miden y pesan lo mismo, un dado "legal". En esas condiciones, repetimos la experiencia tantas veces como queramos y vamos anotando el número de la cara que cae hacia arriba. Así, por caso, luego de tirar 24 veces observamos que la cara 6 ha caído hacia arriba 5 veces. Decimos entonces que la frecuencia relativa con la que ha caído 6, es decir las veces que salió 6 en relación con el total de veces que se arrojó el dado, es $f_r = \frac{5}{24} \approx 0{,}20833333$. Continuamos arrojando el dado y cuando lo hemos hecho, por ejemplo, 48 veces, observamos que el 6 ha salido 7 veces. La frecuencia relativa ahora es $f_r = \frac{7}{48} \approx 0{,}145833333$. Al tirar el dado 144 veces vemos que 6 ha caído hacia arriba 23 veces y entonces $f_r = \frac{23}{144} \approx 0{,}159722222$. Observamos entonces que el fenómeno exhibe *regularidad estadística*, es decir su frecuencia relativa va adquiriendo, en la medida que el número de ensayos crece, un carácter estable. En efecto, para el caso, se observa que la diferencia entre las frecuencias tirando 24 y 48 veces el dado respectivamente, es aproximadamente 0.0625, mientras que la diferencia luego de tirar 48 y 144 veces se aproxima a 0.0138888889, un número sensiblemente menor. Las frecuencias relativas están entonces cada vez más cerca en torno a un valor que, aproximado suficientemente, se adoptará como probabilidad del suceso "caer hacia arriba la cara 6 del dado". En este caso, ese valor deberá aproximarse a $f_r = \frac{1}{6} \approx 0{,}166666667$. Este enfoque empírico para definir la probabilidad "experimental" o "a posteriori" de los hechos, es más general pues considera no sólo experiencias como el tirar una moneda o un dado, cuyos resultados posibles sabemos son finitos, sino que permite calcular la probabilidad en casos como el del número par, donde bastaría enunciar al azar números e ir calculando las frecuencias relativas correspondientes, del suceso número par. Eso es lo que implícitamente hacemos cuando decimos la mitad de los números son pares.

El cálculo de la probabilidad de sucesos que observen regularidad estadística es realizado a partir de los hechos, "a posteriori", y tiene toda la *objetividad* que pueda aportarle nuestra percepción de los mismos.

Llegado a este punto, se puede notar que las definiciones de probabilidad ensayadas se basan en experiencias realizadas o que pueden realizarse, cuyos resultados son excluyentes uno de otro. Si la moneda se arroja una vez, cae cara o cae ceca excluyentemente, no puede caer simultáneamente cara y ceca. Uno solo de ambos resultados debe registrarse. Lo mismo ocurre si se arroja una vez un dado: si cae 6 no cae ninguna otra cara. De igual forma, un número natural o es par o es impar. Además, si se consideran todos los resultados posibles y excluyentes de una experiencia, se observa que la probabilidad de cada uno es un número que está entre 0 y 1. También se observa que su suma es 1. Por ejemplo, si se arroja un dado, la suma de la probabilidad de que caiga 1, más la de que caiga respectivamente 2, 3, 4, 5 o 6 es: $\frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = 1$. Esto es así, sea que cada $\frac{1}{6} \approx 0{,}166666667$ se haya calculado por la definición de Laplace o en forma empírica, por la frecuencia relativa, arrojando el dado un número suficientemente grande de veces.

Ahora parece que estamos en condiciones de calcular la probabilidad de cualquier suceso. Entonces, por ejemplo, podríamos intentar calcular la probabilidad de que la selección de fútbol de Holanda gane el próximo campeonato mundial. Con nuestra definición clásica tendríamos que analizar el número de casos favorables, que es 1 solo y corresponde al suceso Holanda campeón; y el número de casos posibles, que son 32, porque esa es la cantidad de equipos que participan. Así, la probabilidad de que Holanda sea el campeón es $\frac{1}{32} = 0{,}03125$. El cálculo plantea un interrogante, pues además de Holanda, tanto Brasil como cualquier país ignoto del mundo futbolístico que participara entre los 32, tendrían también la misma probabilidad. Y esta difícilmente sea una evaluación realista de las chances de cada uno.

Tratamos entonces de aplicar la definición empírica, basándonos en el número de campeonatos mundiales que se han jugado y en las veces que lo ha ganado cada equipo. Como se han jugado hasta ahora 19 torneos y Holanda nunca ha salido campeón, la probabilidad dada por la frecuencia relativa es $f_{Holanda} = \frac{0}{19} = 0$, otra evaluación que seguramente es poco realista. Basta pensar que de acuerdo con un cálculo similar, España tenía probabilidad 0 de ganar el último mundial, que efectivamente fue el único que ganó. Hasta aquí, los cálculos de probabilidad realizados con toda la objetividad de que somos capaces, no parecen conducirnos a una evaluación adecuada de las posibilidades de Holanda. Lo que se pone de manifiesto aquí es que no cualquier forma de calcular la probabilidad mide de manera

realista las posibilidades de Holanda de obtener la copa del mundo. ¿Hay alguna otra forma de asignar una medida a la posibilidad de ese hecho? La respuesta es sí. Veamos lo que hacen los que reciben apuestas. Ellos dicen, por ejemplo, por cada peso que un jugador apueste yo le entregaré dos si gana Brasil. Como quien recibe las apuestas no pierde dinero, eso quiere decir que la mitad de las personas cree que Brasil va a ganar el campeonato. Por lo tanto, si 100 apostadores han puesto un peso, a 50 se les entregarán dos pesos, en caso de ganar Brasil, que es el resultado de dividir el monto total acumulado de 100 pesos entre esos 50 jugadores que acertaron. A cada apostador que crea que Argentina va a ganar el mundial se le entregarán, si eso ocurre, 10 pesos por cada peso apostado. Eso quiere decir que la décima parte de los 100 apostadores cree que Argentina va a ganar. Así en el caso de Holanda: supongamos que hay 5 de los 100 apostadores que creen que ganará el campeonato del mundo. El cociente $\frac{5}{100} = \frac{1}{20}$ representa entonces la proporción de apostadores a favor de Holanda, así como $\frac{10}{100} = \frac{1}{10}$ es la de Argentina y $\frac{50}{100} = \frac{1}{2}$ la de Brasil. Para fijar ideas: supongamos que las apuestas de los 100 apostadores, de un peso cada uno, se distribuyeran como en la Tabla 1:

Tabla 1

Selección	**Apostadores a favor**	**Proporción p**
Brasil	50	$50/100 = 1/2$
España	15	$15/100 = 3/20$
Argentina	10	$10/100 = 1/10$
Alemania	10	$10/100 = 1/10$
Italia	10	$10/100 = 1/10$
Holanda	5	$5/100 = 1/20$

Es claro que si gana Argentina no puede ganar Brasil. Que gane uno u otro son sucesos excluyentes. Además por ser proporciones, los p son números que están entre 0 y 1, es decir $0 \leq p \leq 1$. Finalmente, la suma de todas las proporciones da 1, incluso sumando las de los $32 - 6 = 26$ selecciones que no han recibido apuestas entre estos 100 apostadores y que por lo tanto valen 0. En efecto:

$$\frac{50}{100} + \frac{15}{100} + \frac{10}{100} + \frac{10}{100} + \frac{10}{100} + \frac{5}{100} + \frac{0}{100} + \cdots + \frac{0}{100}$$

Desde el punto de vista matemático, las proporciones calculadas cumplen con las mismas propiedades que las probabilidades halladas para las caras de una moneda o de un dado. Y además, parecen ser un poco más realistas respecto de las posibilidades de ganar de cada equipo. Claro que si aceptamos que estas proporciones son probabilidades, tendremos también que aceptar que su cálculo no ha sido tan *objetivo* como los que realizamos para las caras de una moneda "a priori" o para las caras de un dado "a posteriori". Esta *subjetividad*, que depende de los 100 apostadores, hay que suponer, se haría mas objetiva cuantos más apostadores hubiera. En el fondo, nada nos aparece como puramente objetivo, pues todo juicio que hagamos depende de nuestra *subjetividad*; también es cierto que estos apostadores conocen de fútbol como para haber hecho sus apuestas de forma no totalmente subjetiva. Es decir, la reserva sobre la validez de esta probabilidad llamada "subjetiva" es más filosófica que matemática y se relaciona con nuestra concepción sobre cómo y por qué ocurren los hechos. Desde el punto de vista de las propiedades matemáticas, las probabilidades *subjetivas* obtenidas subjetivamente de acuerdo con algún procedimiento, son tan válidas como las calculadas en forma clásica o empírica. Por último, si pensamos que la distribución de proporciones así calculada se ajusta razonablemente a la posibilidad de los hechos, no existen, en principio, razones para no utilizarla como "modelo" probabilístico de los mismos. Esto es precisamente lo que a veces se hace en dominios como la economía o el marketing, cuando se trabaja sobre la base de opiniones de personas.

1.3. Definición axiomática de probabilidad

Más allá de las formas en que lleguen a calcularse, lo que distingue a las probabilidades es, en cualquier caso, el cumplimiento de ciertas propiedades.

Todos los resultados posibles de una experiencia constituyen el llamado *espacio muestral* de la misma. Por ejemplo:

- si el experimento es arrojar una moneda y observar la cara que cae hacia arriba, el espacio muestral será $S = \{cara, ceca\}$
- si el experimento es arrojar un dado y ver que cara cae hacia arriba resulta el espacio muestral $S = \{1, 2, 3, 4, 5, 6\}$
- si el experimento es elegir un número natural se tiene $S = \{1, 2, 3, \cdots\}$

En estos ejemplos hemos considerado solo eventos simples, formados por un solo suceso. Pero también podemos estar interesados en eventos compuestos que se presenten al ocurrir cualesquiera de varios sucesos. Por ejemplo:

- Arrojar un dado y ver si cae hacia arriba un número primo. Este evento es $P = \{1, 2, 3, 4, 5\}$ y es un subconjunto del espacio muestral $S = \{1, 2, 3, 4, 5, 6\}$
- Elegir un número natural y ver si es impar. El evento es $P = \{1, 3, 5 \cdots\}$ y está incluido en $P = \{1, 2, 3, 4, 5, 6 \cdots\}$

Lo importante, por ahora, es comprender que los eventos simples o compuestos son una parte del espacio muestral y como tales son elementos del *conjunto de partes* del mismo. En ese conjunto de partes están todos los eventos. Por ejemplo, si la experiencia es tirar una moneda, el espacio muestral es $S = \{cara, ceca\}$, pero su conjunto de partes es $\Omega = Partes\,(S) = \{\emptyset, \{cara\}, \{ceca\}, S\}$. Aquí como los sucesos son excluyentes, el conjunto vacío representa el evento imposible en el cual simultáneamente caen cara y ceca: $\{cara\} \cap \{ceca\} = \emptyset$. Además caer cara o ceca es el evento $\{cara\} \cup \{ceca\} = S$. Hablando en general, es a los elementos del conjunto Ω a los que les asignamos una medida, que constituirá una probabilidad si se cumplen las siguientes condiciones:

I. Para cualquier evento A tal que $A \in \Omega$ (o lo que es lo mismo $A \subset S$) se cumple que: $P(A) \geq 0$

II. $P(S) = 1$

III. Si A y B son mutuamente excluyentes (es decir $A \cap B = \emptyset$) entonces: $P(A \cup B) = P(A) + P(B)$

Toda medida de incertidumbre que cumpla con estas condiciones, no importa por cuál método o fórmula haya sido calculada, en términos matemáticos es una probabilidad. Debe observarse que, como se ha comentado en la Sección 2, cualquiera de los ejemplos allí introducidos, satisface las condiciones expresadas.

Ejemplo 1: *Describir cuál es el espacio muestral asociado a cada uno de los siguientes experimentos aleatorios. En el caso de ser finito, indicar cuál es su cardinal.*

a) Se lanza al aire una moneda tres veces.

Si c simboliza cara y x representa ceca, se pueden presentar todos los resultados que indican las ternas del conjunto S:

$$S = \{(c,c,c)\,;(x,x,x)\,;(c,c,x)\,;(x,x,c)\,;(c,x,x)\,;(x,c,c)\,;(c,x,c)\,;(x,c,x)\}$$

El cardinal se puede calcular directamente contándolas:

$$\#S = 8$$

o bien a través de las variaciones con repetición de dos elementos (ver pág. 37):

$$\#S = VR_{2,3} = 2^3 = 8$$

b) Se arrojan dos dados simultáneamente.

Si se supusiera para mayor claridad que un dado es rojo y otro azul, los distintos pares posibles serían los listados en S. Así, por ejemplo, $(1,4)$ querría decir que el dado rojo cayó 1 y el azul cayó 4. Por supuesto, aunque los dados fueran de igual color, el conjunto S sería el mismo pues siempre se trataría de dos dados diferentes.

$$S = \left\{ \begin{array}{c} (1,1)\,;(1,2)\,;(1,3)\,;(1,4)\,;(1,5)\,;(1,6)\,;(2,1)\,;(2,2)\,;(2,3)\,;(2,4)\,; \\ (2,5)\,;(2,6)\,;(3,1)\,;(3,2)\,;(3,3)\,;(3,4)\,;(3,5)\,;(3,6)\,;(4,1)\,;(4,2)\,; \\ (4,3)\,;(4,4)\,;(4,5)\,;(4,6)\,;(5,1)\,;(5,2)\,;(5,3)\,;(5,4)\,;(5,5)\,;(5,6)\,; \\ (6,1)\,;(6,2)\,;(6,3)\,;(6,4)\,;(6,5)\,;(6,6) \end{array} \right\}$$

$$\#S = 36$$

También se puede calcular el cardinal a través de las variaciones de 6 elementos con repetición (ver pág. 37):

$$\#S = VR_{6,2} = 6^2 = 36$$

c) Se extraen dos fichas sucesivamente y sin reposición de una caja que contiene seis fichas numeradas del 1 al 6.

Supongamos que, por ejemplo, se extrae una ficha y esta resulta 2. Para la elección de la siguiente ficha, como la ficha 2 no se puede volver a poner en la caja, quedan solo cinco fichas: la 1, la 3, la 4, la 5 y la 6. Es decir, por cada elección de la primera ficha hay cinco posibles para la segunda. Ahora bien; hay seis formas distintas de elegir la primera ficha y está claro que la ficha que salió en primer término no puede repetirse en el segundo por lo que el número de pares que pueden resultar según la regla de elección es $6 \times 5 = 30$. Se tiene entonces:

$$S = \left\{ \begin{array}{c} (1,2)\,;(1,3)\,;(1,4)\,;(1,5)\,;(1,6)\,;(2,1)\,;(2,3)\,;(2,4)\,;(2,5)\,;(2,6)\,; \\ (3,1)\,;(3,2)\,;(3,4)\,;(3,5)\,;(3,6)\,;(4,1)\,;(4,2)\,;(4,3)\,;(4,5)\,;(4,6)\,; \\ (5,1)\,;(5,2)\,;(5,3)\,;(5,4)\,;(5,6)\,;(6,1)\,;(6,2)\,;(6,3)\,;(6,4)\,;(6,5) \end{array} \right\}$$

$$\#S = 30$$

También se puede calcular el cardinal a través de las variaciones de 6 elementos tomados de a 2 (ver pág. 37):

$$\#S = V_{6,2} = \frac{6!}{(6-2)!} = 30$$

d) Se lanza al aire una moneda hasta que sale cara por primera vez

Puede ocurrir que salga cara al primer lanzamiento. El resultado es entonces C. Si se tira la moneda y en el primer lanzamiento sale ceca, hay que seguir tirando. Si sale cara el segundo ahí se para de lanzar. El resultado es XC y se tiró dos veces la moneda. Así XXC estaría indicando que se tiró tres veces la moneda y que recién en la tercera salió cara. Evidentemente existe la posibilidad de que se tire la moneda muchas veces hasta que salga una cara. En términos matemáticos se diría que el número de tiros hasta que salga una cara puede crecer todo cuanto sea necesario. Es decir, es potencialmente infinito. Se anota entonces que el número de lanzamientos hasta que salga una cara puede ser uno del conjunto:

$$S = \{1, 2, 3, 4, \cdots\}$$

Hemos presentado aquí una definición axiomática de probabilidad. Los axiomas *i*) a *iii*) son las condiciones requeridas que nos permiten independizarnos de la forma en que la probabilidad se construye. Tal definición tiene

algunas implicancias importantes:

Propiedad 1 ***Ley del Complemento:*** Dado un evento A cualquiera, la probabilidad de su complemento es:

$$P\left(\bar{A}\right) = 1 - P(A)$$

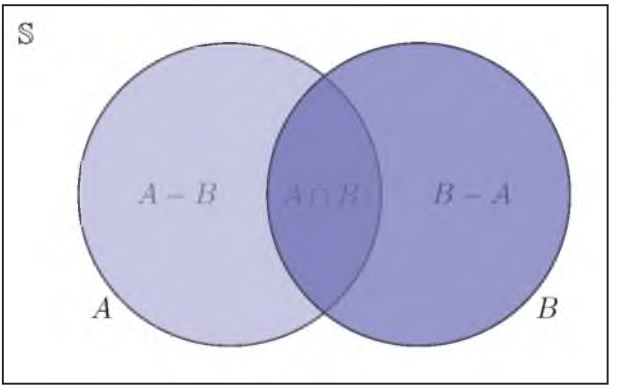
En efecto, por ser $\bar{A}$ el complemento de A, se tiene que son eventos mutuamente excluyentes pues $A \cap \bar{A} = \emptyset$. De allí que, como $S = A \cup \bar{A}$ se tenga: $1 = P(S) = P\left(A \cup \bar{A}\right) = P(A) + P\left(\bar{A}\right)$ de dónde $P\left(\bar{A}\right) = 1 - P(A)$ como se quería demostrar.

Propiedad 2 ***Ley de la Suma:*** Si A y B son dos eventos cualesquiera la probabilidad de que ocurra A o B o ambos simultáneamente es:

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

En efecto, por la teoría elemental de conjuntos sabemos que $A = (A - B) \cup (A \cap B)$ según se ve en la figura 1:

Figura 1.



Además $(A - B) \cap (A \cap B) = \emptyset$, Es decir, considerados como sucesos, $(A - B)$ y $(A \cap B)$ son mutuamente excluyentes. Algo enteramente análogo ocurre si se toma $B = (B - A) \cup (A \cap B)$ De tal modo:

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

Como se trata de sumas de números reales, podemos escribir entonces:

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

Si se observa con atención se ve que en el segundo miembro de esta igualdad se tiene sumada la probabilidad de tres conjuntos disjuntos. Como, por resultado de la teoría elemental de conjuntos, se sabe que $A \cup B = (A - B) \cup (B - A) \cup (A \cap B)$ aplicando el axioma *iii*) se tiene:

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

que es lo que quería probarse.

Nótese que en el caso particular en que $A \cap B = \emptyset$, la mutua exclusión de los sucesos implica la imposibilidad de ocurrencia simultánea de los mismos por lo que la probabilidad es $P(A \cap B) = 0$ y queda entonces $P(A \cup B) = P(A) + P(B)$.

Ejemplo 2: *En una pequeña ciudad se publican dos diarios: Nuevos Aires y El Eco. El* 40 % *de los habitantes lee Nuevos Aires, el* 27 % *lee El Eco, y el* 9 % *lee ambos diarios.*

a) ¿Qué porcentaje de habitantes lee un solo diario?

Si se emplea el modelo de la probabilidad elemental se puede pensar que la probabilidad de que un habitante lea el diario *Nuevos Aires* es:

$$P(NA) = 0{,}4$$

De la misma forma para el diario *El Eco* se tiene:

$$P(EE) = 0{,}27$$

También la probabilidad de que un habitante lea ambos diarios es:

$$P(NA \cap EE) = 0{,}09$$

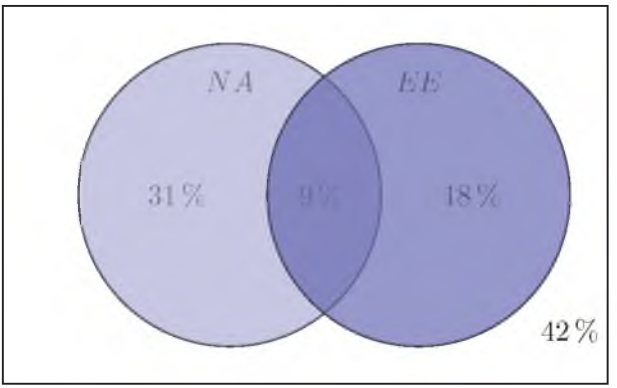
Así la probabilidad de que un habitante lea uno o los dos diarios es:

$$P(NA \cup EE) = P(NA) + P(EE) - P(NA \cap EE) = 0{,}4 + 0{,}27 - 0{,}09 = 0{,}58$$

Siempre según el modelo probabilístico esto implica que de cada 100 habitantes 58 leen al menos uno de los dos diarios. Está claro que de esos 58, 40 leen *NA*. A su vez de entre éstos, hay 9 que también leen *EE*. Por lo tanto, hablando en porcentajes del total:

$$40\,\% - 9\,\% = 31\,\% \ \textit{lee sólo} \ NA$$
$$27\,\% - 9\,\% = 18\,\% \ \textit{lee sólo} \ EE$$

Es posible organizar la información en el siguiente esquema:



Como el 31 % lee solo *Nuevos Aires* y el 18 % lee solo *El Eco*, sumando se tiene que el 49 % lee un solo diario.

b) ¿Qué porcentaje de habitantes no lee ningún diario?

El 31 % lee solo *Nuevos Aires*, el 18 % lee sólo *El Eco* y el 9 % lee los dos diarios. Entonces: 100 %−31 %−18 %−9 % = 42 % *restante no lee ninguno.*

1.4. Principios de conteo

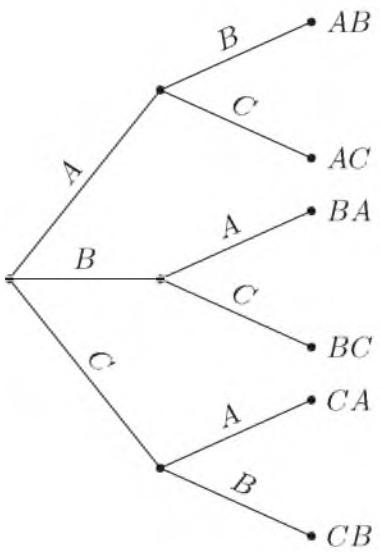
Cuando puede utilizarse la construcción clásica de la probabilidad surge la necesidad de contar los casos posibles y los favorables al suceso, cuya incertidumbre quiere medirse. Esto requiere la aplicación de ciertos principios, que ahora van a explicitarse.

Principio de multiplicación: Supongamos que una acción (1) puede desarrollarse de n_1 formas distintas y que a continuación de ella, puede realizarse otra acción (2) de n_2 maneras diferentes. Entonces la cantidad de formas en que puede realizarse la cadena de acciones (1) − (2) es $n_1 \times n_2$.

Por ejemplo, supongamos tener un conjunto de tres letras ABC. Si la acción (1) es elegir una letra dejando dos para la acción (2) que es elegir una segunda letra, se tiene que la acción (1) puede concretarse de $n_1 = 3$ formas distintas, mientras que la acción (2) puede realizarse de $n_2 = 2$. Entonces la

cantidad de maneras distintas en que puede hacerse la cadena de elecciones $(1)-(2)$ es $n_1 \times n_2 = 3 \times 2 = 6$. La figura 2 ayuda a entender este principio:

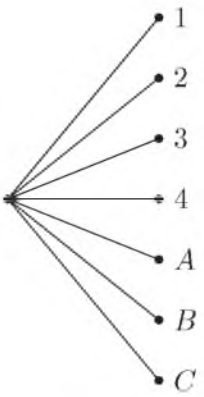
Figura 2.



Principio de adición: Supongamos que una acción (1) puede desarrollarse de n_1 formas diferentes y que otra acción (2) tiene n_2 maneras distintas de efectuarse, siendo que, además, no es posible realizar ambas acciones simultáneamente entonces hay $n_1 + n_2$ formas distintas de realizar una cualquiera de ellas dos.

Por ejemplo, sea la acción (1) consistente en seleccionar una letra del conjunto ABC y sea la acción (2) tal que hay que elegir un número del conjunto $1, 2, 3, 4$. Claramente hay $n_1 = 3$ maneras de seleccionar la letra y $n_2 = 4$ formas de elegir el número. La cantidad de maneras distintas de seleccionar una letra o un número es entonces $n_1 + n_2 = 3 + 4 = 7$. En la figura 3 se esquematiza esta cuenta:

Figura 3.



Función factorial: Esta es una función muy útil a efecto de contar distintos tipos de agrupamientos de elementos. Se define para todos los números naturales y el cero de la siguiente manera:

$$0! = 1$$

$$1! = 1$$

$$\vdots$$

$$n! = (n-1)! \times n$$

La notación que coloca el número n y a continuación el signo de admiración ! se lee como "factorial de n ". Así por ejemplo el "factorial de 3" se escribe 3! y se calcula:

$$3! = 2! \times 3 = 1! \times 2 \times 3 = 1 \times 2 \times 3 = 6$$

Se cumple en general que:

$$n! = n \times (n-1) \times (n-2) \times \cdots \times 3 \times 2 \times 1$$

De esta forma cada vez que en un cálculo aparezca un producto similar, éste podrá anotarse como el factorial de un número.

Permutaciones: Supongamos ahora que se tienen las tres letras ABC y se desea averiguar en cuantos órdenes distintos pueden escribirse sin repetirlas. Es claro que la primera acción será elegir una letra de entre las tres, a

continuación habrá que elegir la segunda letra entre las dos que quedaron disponibles y luego habrá una tercera acción que consistirá en agregar la letra sobrante en el último lugar de la palabra formada. El *principio de multiplicación* visto nos permite calcular el número de palabras distintas que pueden formarse con esas tres letras sin repetirlas: $n_1 \times n_2 \times n_3 = 3 \times 2 \times 1 = 3! = 6$

En general los distintos órdenes en los que puedan elegirse n elementos distintos se denominan *permutaciones* de n elementos, que se escriben y calculan:

$$P_n = n!$$

Variaciones: La mismas letras ABC pueden servir para formar pares. Es claro que la palabra AB y la palabra BA si tuvieran algún significado, podrían no tener el mismo. Es decir, en ese caso no es igual el par formado por AB que el formado por BA. Todos los pares que pueden formarse de esta manera, sin repetir ninguna letra son: AB, AC, BA, BC, CA y CB. El número total de pares que puedan formarse corresponde a las *variaciones* de 3 elementos tomados de a 2. En general si se tienen m elementos distintos y quieren conocerse todos los órdenes posibles en que pueden seleccionarse n de ellos sin repetirlos, hay que calcular las *variaciones* de m elementos tomados de a n:

$$V_n^m = V_{m,n} = \frac{m!}{(m-n)!}$$

Combinaciones: Si con ABC quieren formarse ahora pares para los que se tenga en cuenta sólo cuáles letras lo conforman, se tendrán las duplas AB, CB y AC. En efecto, bajo la condición de que no se consideren diferentes los pares cuando las mismas letras se citen en distintos órdenes, el par AB es el mismo que el BA, el CB el mismo que el BC y el AC es el mismo que CA. Cada dupla puede citarse de la forma en que se quiera, pero debe contarse solo una vez. En este caso la cantidad de duplas de tal tipo se calcula como las combinaciones de 3 elementos tomados de a 2. En forma genérica cuando se tienen m elementos diferentes y quiere hallarse el número de selecciones de n de estos elementos, sin que ninguno de ellos se repita ni se consideren distintas las selecciones que incluyen los mismos elementos citados en distinto orden, se deben calcular las *combinaciones* de m tomados de a n:

$$C_n^m = C_{m,n} = \binom{m}{n} = \frac{m!}{n!(m-n)!}$$

Obsérvese como dato interesante que:

$$\binom{m}{0} = \frac{m!}{0!(m-0)!} = \frac{m!}{1m!} = 1$$

cualquiera sea m.Además:

$$\binom{m}{1} = \frac{m!}{1!(m-1)!} = \frac{m(m-1)!}{(m-1)!} = m$$

y

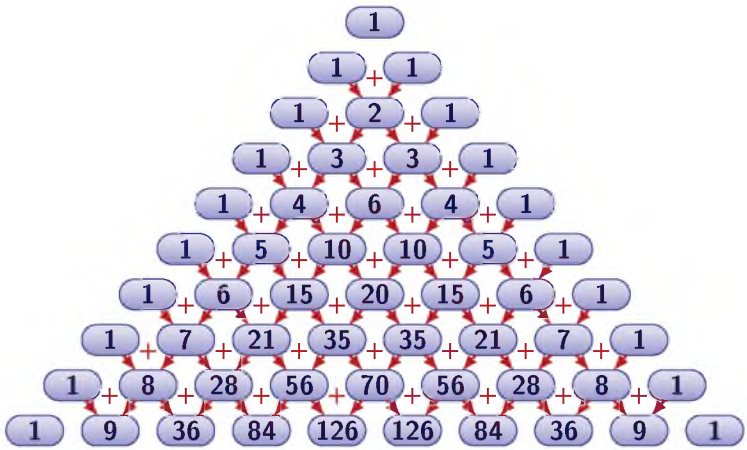
$$\binom{m}{m} = \frac{m!}{m!(m-m)!} = \frac{m!}{m!0!} = 1$$

Estos números, que pueden parecer no muy familiares, en realidad lo son bastante. Consideremos las sucesivas potencias de un binomio:

$(a+b)^0 = 1$
$(a+b)^1 = a+b = \binom{1}{0}a^{1-0}b^0$
$(a+b)^2 = a^2+2ab+b^2 = \binom{2}{0}a^{2-0}b^0 + \binom{2}{1}a^{2-1}b^1 + \binom{2}{2}a^{2-2}b^2$
$(a+b)^3 = a^3+3a^2b+3ab^2+b^3 = \binom{3}{0}a^{3-0}b^0 + \binom{3}{1}a^{3-1}b^1 + \binom{3}{2}a^{3-2}b^2 +$
$+\binom{3}{3}a^{3-3}b^3$
...
y así. La fórmula general para la *m-ésima* potencia la proporciona el llamado "binomio de Newton":

$$(a+b)^m = \sum_{n=0}^{m} \binom{m}{n} a^{m-n}b^n$$

Cada coeficiente de este desarrollo, cada número combinatorio, puede calcularse mnemotécnicamente por medio del "triángulo de Pascal", aquél matemático que como dijimos en la *sección 1* desarrolló, junto con Fermat, fórmulas para conteo de casos. El triángulo se forma con dos lados de unos y cada renglón se completa sumando los dos coeficientes ubicados en el renglón anterior a ambos lados de la posición calculada:



Por ejemplo, el número de combinaciones de 5 elementos tomados de a 3 es $\binom{5}{3} = 10$

Conteo con repetición: Si queremos calcular el número de palabras de tres letras que pueden formarse con las letras ABC cuando A debe repetirse 2 veces y B debe aparecer solo 1, construimos las ternas: AAB, ABA y BAA. Hay entonces 3 permutaciones de 3 elementos con repetición, de un elemento 2 veces y de otro 1 vez. Si por ejemplo se tuvieran cuatro letras $ABCD$, las permutaciones de estas 4 letras que contienen 2 veces a A, 1 vez a B y 1 a C serían:

$AABC$,$ABAC$,$ABCA$,$AACB$,$ACAB$,$ACBA$,$BAAC$,$BACA$,$BCAA$, $CAAB$,$CABA$ y $CAAB$ que son 12.

En general si se tienen m elementos distintos y se quieren formar sus permutaciones considerando que uno de ellos en particular debe aparecer n_1 veces, otro en particular n_2 veces y así hasta otro, que debe aparecer n_k veces, de modo tal que $n_1 + n_2 + \cdots + n_k = m$, hay que calcular las permutaciones con repetición de m elementos repetidos $n_1, n_2, \cdots, n_k$ veces según:

$$P_m^{n_1,n_2,\cdots,n_k} = \frac{m!}{n_1!n_2!\cdots n_k!}$$

Ahora corresponde analizar las variaciones que podrían darse con las letras ABC si cualquiera de ellas puede repetirse y se quieren seleccionar, por ejemplo 2 letras. Supongamos que elegimos la primera letra entre ABC.

Esto podemos hacerlo de 3 maneras posibles. Si ahora elegimos la segunda de entre las 3, tenemos también 3 formas de hacerlo. Los pares posibles resultan AA, AB, BA, BB, AC, CA, CC, CB y BC. El número de estas variaciones formadas con 3 letras tomadas de a 2, si las letras pueden *repetirse*, es $3 \times 3 = 3^2$. En general las *variaciones con repetición* de m elementos tomados de a n se calculan como:

$$V_n'^m = V'_{m,n} = VR_{m,n} = m^n$$

Para finalizar consideremos el caso en que sobre las letras ABC se establece una selección de dos de ellas con repetición admitida, pero ahora considerando que los pares se toman en conjunto, de modo que, por ejemplo, es lo mismo listar AB que BA. En tal caso las combinaciones con repetición posibles serán AA, AB, BB, AC, CC, CB cuya cantidad surge de la cuenta $\frac{(3+2-1)!}{2!(3+2-1-2)!} = \frac{4!}{2!2!} = 6$. En términos generales las *combinaciones con repetición* de m tomados de a n son:

$$C_n'^m = C'_{m,n} = CR_{m,n} = \binom{m+n-1}{n} = C_n^{m+n-1}$$

Los principios de conteo aquí explicados se suelen aplicar aisladamente o combinando algunos de ellos cuando es necesario calcular casos, tanto posibles como favorables, al aplicar la definición clásica de probabilidad. Como se ve, y a pesar de lo que previamente pudiera suponerse, contar casos puede ser una actividad no sencilla para la cual aguzar el ingenio y poseer entrenamiento resulta fundamental.

Ejemplo 3: *De cuantas formas distintas pueden entregarse dos premios entre 10 personas si:*

a) no se pueden conceder ambos premios a la misma persona.

- Si los premios son iguales, no importa cual de ellos corresponda a cada ganador. Por lo tanto se calcula el número de formas posibles de elegir dos ganadores entre diez personas sin que importe el orden en que se lo haga. Es decir se calculan las combinaciones de 10 elementos tomados de a 2:

$$C_{10,2} = \frac{10!}{2!(10-2)!} = \frac{10!}{2!8!} = \frac{10 \cdot 9}{2} = \mathbf{45}$$

- Si los premios son distintos, no da lo mismo el orden en que se seleccionen los ganadores. Si se supone que al primer ganador le corresponde el premio A y al segundo el B, si se invirtiera el orden de los ganadores el resultado sería distinto. Por esa razón cada par de ganadores debe considerarse en sus dos órdenes posibles. Se calculan entonces las variaciones de 10 elementos tomados de a 2:

$$V_{10,2} = \frac{10!}{(10-2)!} = \frac{10!}{8!} = 10 \cdot 9 = \mathbf{90}$$

b) pueden otorgarse ambos premios a la misma persona.

- Si los premios son iguales será una combinación donde además se puede repetir el ganador. Se calculan las combinaciones con repetición.

$$CR_{10,2} = \frac{(10+2-1)!}{2!(10-1)!} = \frac{11!}{2!9!} = \frac{11 \cdot 10}{2} = \mathbf{55}$$

También se puede resolver:

$$C_{10,2} + 10 = \frac{10!}{2!(10-2)!} + 10 = \frac{10!}{2!8!} + 10 = \frac{10 \cdot 9}{2} + 10 = 45 + 10 = \mathbf{55}$$

- Si los premios son distintos y además se puede repetir el ganador, se calculan las variaciones con repetición.

$$VR_{10,2} = 10^2 = \mathbf{100}$$

1.5. Probabilidad condicional

Hay sucesos que dependen parcial o totalmente de que ocurran otros. Por ejemplo, para que la suma de las caras de dos dados que se han arrojado resulte mayor o igual que 11 es necesario que uno de los dados haya caído 6. Si ninguno de los dos dados cayó seis, la suma nunca podrá dar 11 o más. En este caso el suceso "sumar 11 o más" depende totalmente de que se de el suceso "caer 6" en un dado. En cambió si el suceso a analizar fuera sumar 10 o más, esto podría ocurrir a causa de que un dado caiga 6, por ejemplo si el otro cayese 4 o 5 o 6, pero también ocurriría si ambos dados

resultasen 5. Aquí el suceso "sumar diez o más' depende parcialmente de la ocurrencia de una "cara 6". En cualquier caso se puede querer averiguar cuál es la probabilidad de que ocurra un suceso cuando ha tenido o tiene lugar otro. Es decir, calcular la probabilidad de que se aprecie un *efecto* cuando sucede una causa. Si continuamos analizando nuestro ejemplo, la suma 10 o mayor se da en los casos en que los pares de resultados, un número por dado, son $\{(4,6),(5,5),(5,6),(6,4),(6,5),(6,6)\}$. Ahora bien, caer 6 al menos uno de los dos dados es algo que ocurre con los pares del conjunto: $\{(6,1),(6,2),(6,3),(6,4),(6,5),(6,6),(5,6),(4,6),(3,6),(2,6),(1,6)\}$

Si definimos los eventos A y B acordando que A es "sumar 10 o más" y B "caer 6 al menos uno de los dos dados" se tiene:

$$A = \{(4,6),(5,5),(5,6),(6,4),(6,5),(6,6)\}$$

$$B = \{(6,1),(6,2),(6,3),(6,4),(6,5),(6,6),(5,6),(4,6),(3,6),(2,6),(1,6)\}$$

Se observa que de las 6 formas en las que puede ocurrir A, en 5 ocurre simultáneamente B. Así resulta $A \cap B = \{(4,6),(5,6),(6,4),(6,5),(6,6)\}$ el conjunto en el que se cumplen simultáneamente ambas situaciones: sumar 10 o más y caer 6 al menos uno de los dados. Es claro que hay en total 36 formas distintas en que pueden caer las caras de dos dados, es decir 36 pares distintos posibles, de tal modo que puede calcularse la probabilidad de que ocurra que "la suma sea 10 o más" cuando "cae 6 al menos una de las caras de los dados" de la forma:

$$P(A/B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{5}{36}}{\frac{11}{36}} = \frac{5}{11} \approx 0{,}45454545$$

Con $P(A/B)$ queremos decir: probabilidad de que ocurra A cuando ocurre o ha ocurrido B. Es decir, medimos la probabilidad de que sucedido B, suceda A. Si se observa con atención, la ocurrencia de B implica una restricción sobre los 36 casos distintos que pudieran darse al sumar las caras resultantes de tirar dos dados. En este ejemplo, si se contabilizan sólo estos casos como los posibles, y se ve en cuales de ellos se cumple también la condición que determina al evento A, es decir se cuentan como casos favorables los pares de caras que están en $A \cap B$, resulta naturalmente: $p = \frac{\#A \cap B}{\#B} = \frac{5}{11}$. La condición que expresa a B restringe los casos posibles contenidos en el espacio muestral y requiere de los favorables a la propiedad que caracteriza a A su simultáneo cumplimiento. Si se considera, en el ejemplo dado, el caso del evento "sumar 11 o más" la probabilidad de que esto ocurra, cuando ha

salido 6 al menos un dado, es:

$$P(A/B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{3}{36}}{\frac{11}{36}} = \frac{3}{11} \approx 0{,}27272727$$

siendo

$$A = \{(5,6),(6,5),(6,6)\}$$

y

$$B = \{(6,1),(6,2),(6,3),(6,4),(6,5),(6,6),(5,6),(4,6),(3,6),(2,6),(1,6)\}$$

En general se dirá entonces que la probabilidad condicional se calcula según la fórmula $P(A/B) = \frac{P(A \cap B)}{P(B)}$ que expresa precisamente la probabilidad de que ocurra A dado que ocurre B.

Cabría ahora preguntarse por la probabilidad de que suceda B cuando haya sucedido A. En nuestro ejemplo eso equivaldría a calcular la probabilidad de que caiga 6 al menos una cara, cuando la suma da 10 o más. Es decir, queremos calcular: $P(B/A) = \frac{P(B \cap A)}{P(A)}$ es claro que los conjuntos A,B y $A \cap B$ son los mismos que antes, pero aquí resulta:

$$P(B/A) = \frac{P(B \cap A)}{P(A)} = \frac{\frac{5}{36}}{\frac{6}{36}} = \frac{5}{6} \approx 0{,}83333333$$

La restricción que plantea A sobre los 36 pares distintos que constituyen el espacio muestral es obviamente diferente a la que establecía B. De tal modo aquí esos 36 pares se ven reducidos a sólo 6 que son aquellos para los cuales la suma es 10 o más. De ellos, 5 son tales que simultáneamente contienen al menos un dado 6 y entonces la probabilidad podría calcularse directamente como $\frac{5}{6}$. Como se ve ahora las probabilidades condicionales $P(A/B)$ y $P(B/A)$ no tienen porque ser iguales. Sin embargo está claro que $P(A \cap B) = P(A/B)P(B) = P(B/A)P(A)$.

1.6. Independencia

No siempre los sucesos dependen unos de otros. A veces son *independientes*. Si al continuar sobre el ejemplo de los dados, consideramos ahora el evento "sumar 2 o mas" es claro que el conjunto A, que lo representa, está formado por los 36 pares posibles que constituyen todo el espacio muestral.

Esto es así porque la suma de las caras que resulta menor es 2, que corresponde al par $(1,1)$. Todos los demás pares suman más de 2. Por otra parte el suceso "caer al menos un 6" al arrojar los dos dados, está representado por el mismo conjunto que llamamos B en la *sección* anterior. Así las cosas, tenemos:

$$A = \left\{ \begin{array}{c} (1,1);(1,2);(1,3);(1,4);(1,5);(1,6);(2,1);(2,2);(2,3);(2,4); \\ (2,5);(2,6);(3,1);(3,2);(3,3);(3,4);(3,5);(3,6);(4,1);(4,2); \\ (4,3);(4,4);(4,5);(4,6);(5,1);(5,2);(5,3);(5,4);(5,5);(5,6) \\ (6,1);(6,2);(6,3);(6,4);(6,5);(6,6) \end{array} \right\}$$

$$B = \{(6,1);(6,2);(6,3);(6,4);(6,5);(6,6);(5,6);(4,6);(3,6);(2,6);(1,6)\}$$

Entonces la probabilidad de que sumar 2 o más cuando sale al menos un 6 es:

$$P(A/B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{11}{36}}{\frac{11}{36}} = \frac{11}{11} = 1$$

Si se observa bien, éste es el valor de la probabilidad del evento A, "sumar 2 o más", calculada con *independencia* del evento B, "caer 6 al menos una cara". En efecto: $P(A) = \frac{36}{36} = 1$ así resulta $P(A/B) = P(A)$. Finalmente, en razón de la igualdad $P(A \cap B) = P(A/B)P(B))$ resulta, bajo esta condición de independencia, $P(A \cap B) = P(A)P(B)$.

Precisamente ésta es la forma general que adopta la *definición de independencia*. Dos sucesos A y B son independientes si y solo si:

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

Debe notarse que al calcular la probabilidad de que "caiga al menos un 6" cuando "la suma es 2 o más" se tiene:

$$P(B/A) = \frac{P(B \cap A)}{P(A)} = \frac{\frac{11}{36}}{\frac{36}{36}} = \frac{11}{36} \approx 0{,}30555556$$

que resulta la misma cantidad que la calculada para la probabilidad de que "caiga al menos un 6" sin tener en cuenta cual es la suma de las dos caras $P(B) = \frac{11}{36} \approx 0{,}30555556$.

La igualdad $P(B/A) = P(B)$ revela nuevamente el hecho, ya conocido, de la independencia pues, como era de esperar, resulta una vez más:

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

Es decir, la independencia es una relación recíproca entre dos sucesos.

Ejemplo 4: *En una clase hay 12 niños y 4 niñas. Si se seleccionan tres estudiantes al azar; ¿Cuál es la probabilidad de que sean todas niñas?*

- Hay en total 16 estudiantes. El primer estudiante seleccionado tiene una probabilidad de ser mujer de:

$$P(M_1) = \frac{4}{16}$$

Para elegir el segundo estudiante quedan en total 15 de las cuales sólo 3 son mujeres. La probabilidad de que ese segundo estudiante sea mujer cuando el primero lo ha sido es:

$$P(M_2/M_1) = \frac{3}{15}$$

Ahora quedan 14 estudiantes y solamente 2 niñas. La probabilidad de que la tercera elección recaiga sobre una mujer cuando las primeras dos también lo han sido se calcula:

$$P(M_3/M_1M_2) = \frac{2}{14}$$

Entonces la probabilidad de que la primera elección resulte mujer, la segunda también y la tercera también es:

$$P(M_1M_2M_3) = P\left(M_1 \cap (M_2/M_1) \cap (M_3)/M_1M_2\right) =$$

$$= P(M_1) \cdot P(M_2/M_1) \cdot P(M_3/M_1M_2)$$

En definitiva:

$$P(M_1M_2M_3) = \frac{4}{16} \cdot \frac{3}{15} \cdot \frac{2}{14} = \frac{1}{140} \approx 0{,}007$$

- Este cálculo pudo simplificarse calculando directamente la cantidad de ternas posibles de estudiantes elegidos. Casos posibles:

$$C_{16,3} = \frac{16!}{3!(16-3)!} = \frac{16 \cdot 15 \cdot 14 \cdot 13!}{3!13!} = \frac{16 \cdot 15 \cdot 14}{3!} = 560$$

Casos favorables: La cantidad de ternas favorables integradas por tres de las cuatro niñas,

$$C_{4,3} = \frac{4!}{3!(4-3)!} = \frac{4 \cdot 3!}{3!1!} = 4$$

Entonces la probabilidad buscada es:

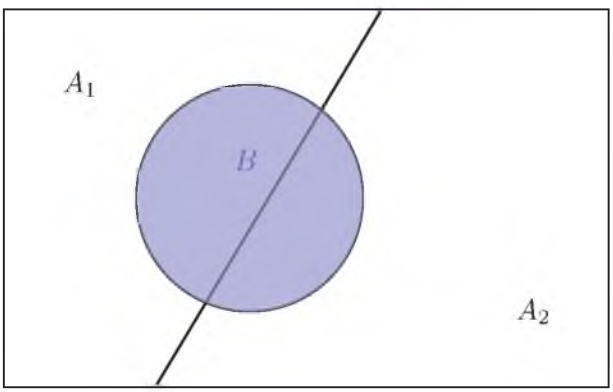
$$P(M_1M_2M_3) = \frac{4}{560} \approx 0{,}007$$

1.7. Probabilidad de las causas

A veces un evento puede producirse porque se da otro, de entre varios que podrían causarlo. En tal caso puede desearse obtener la probabilidad de que una determinada de esas causas sea la que lo haya producido.

Para comprender esto tomemos el siguiente ejemplo. Supongamos que una empresa fabrica lámparas de larga duración utilizando 2 máquinas. La máquina A_1 produce el 60 % del total de lámparas fabricadas y la máquina A_2 fabrica el otro 40 %. El 5 % de las lámparas producidas por la máquina A_1 es defectuoso mientras que ocurre lo mismo para el 3 % de las que produce la máquina A_2. Luego de fabricadas las lámparas se juntan en un depósito, mezcladas sin un orden fijo, y se analiza su calidad. Supongamos que se selecciona con tal propósito una lámpara y se comprueba que es defectuosa, ¿cuál es la probabilidad de que haya sido fabricada por la máquina A_1? Comencemos por graficar el total de las lámparas fabricadas haciendo corresponder una parte de ellas a las provenientes de la máquina A_1 y otra a las que vienen de la máquina A_2. También dibujemos el conjunto de las lámparas defectuosas que podemos llamar B. El gráfico se ve en la figura 4:

Figura 4.



Sabemos que la probabilidad de que una pieza, extraída del total representado en el cuadrado de la figura 4, provenga de cada máquina es, respectivamente:

$$P(A_1) = 0{,}6$$

$$P(A_2) = 0{,}4$$

También conocemos la probabilidad de que, proviniendo la pieza de la máquina A_1, resulte defectuosa. Y lo mismo si proviene de la máquina A_2. Esto se expresa mediante las probabilidades condicionales:

$$P(B/A_1) = 0{,}05$$

$$P(B/A_2) = 0{,}03$$

Ahora bien, la probabilidad de que siendo defectuosa una pieza elegida provenga de la máquina A_1 será:

$$P(A_1/B) = \frac{P(A_1 \cap B)}{P(B)} \tag{1.7.1}$$

Por un lado, sabemos que de la probabilidad condicional $P(B/A_1) = \frac{P(A_1 \cap B)}{P(A_1)}$ podemos despejar:

$$P(A_1 \cap B) = P(B/A_1)P(A_1) \tag{1.7.2}$$

Por otro, de la teoría elemental de conjuntos, obtenemos que B es la unión de sus intersecciones con los conjuntos A_1 y A_2 respectivamente, $B = (A_1 \cap B) \cup (A_2 \cap B)$. Como es claro, si una pieza es fabricada por una máquina no puede ser fabricada por la otra, estos sucesos son mutuamente excluyentes y se tiene que $A_1 \cap A_2 = \emptyset$. Entonces también son excluyentes los eventos representados por las dos intersecciones cuya unión da B. Es decir: $(A_1 \cap B) \cap (A_2 \cap B) = \emptyset$. De aquí que la probabilidad de que la pieza sea defectuosa pueda escribirse como suma de las probabilidades de esas intersecciones, al aplicar el axioma *iii*) de la definición axiomática de probabilidad expuesta en la *sección 3*: $P(B) = P(A_1 \cap B) + P(A_2 \cap B)$.

Y como en general vale para $i = 1, 2$ que $P(A_i \cap B) = P(B/A_i \cap B)P(A_i)$, se obtiene al reemplazar:

$$P(B) = P(B/A_1)P(A_1) + P(B/A_2)P(A_2) \tag{1.7.3}$$

Entonces al sustituir en la fórmula (7.1) el numerador y el denominador del cociente, por las fórmulas (7.2) y (7.3) respectivamente se obtiene la forma

de calcular la probabilidad de que sea defectuosa la pieza seleccionada, a *causa* de provenir de la máquina A_1:

$$P(A_1/B) = \frac{P(B/A_1)P(A_1)}{P(B/A_1)P(A_1) + P(B/A_2)P(A_2)}$$

Haciendo las cuentas:

$$P(A_1/B) = \frac{0{,}05 \times 0{,}3}{0{,}05 \times 0{,}3 + 0{,}03 \times 0{,}4} \approx 0{,}7143$$

Para calcular la probabilidad de que la pieza defectuosa seleccionada haya sido fabricada por la máquina A_2 basta con hacer:

$$P(A_2/B) = \frac{P(B/A_2)P(A_2)}{P(B/A_1)P(A_1) + P(B/A_2)P(A_2)} = \frac{0{,}03 \times 0{,}4}{0{,}05 \times 0{,}3 + 0{,}03 \times 0{,}4} \approx 0{,}2857$$

Se observa que la probabilidad de que, siendo defectuosa, provenga de alguna de las dos máquinas es precisamente 1 ya que no hay otras máquinas que la puedan haber fabricado. Esto se evidencia en la cuenta:

$$P((A_1 \cup A_2)/B) = P(A_1/B) + P(A_2/B) \approx 0{,}7143 + 0{,}2857 = 1$$

Si en vez de existir dos posibles causas de un suceso, existiesen muchas, digamos n, la probabilidad de la ocurrencia de ese suceso a causa de una cualquiera de ellas podría hallarse por la expresión general:

$$P(A_1/B) = \frac{P(B/A_i)P(A_i)}{\sum_{j=1}^{n} P(B/A_j)P(A_j)} \; con \; j = 1, 2, \cdots, n$$

Esta es la llamada *fórmula de Bayes*, aquel estadístico citado en el apartado *I*, también conocida como *fórmula de la probabilidad de las causas.*

Ejemplo 5: *Dos máquinas automáticas producen piezas idénticas que son colocadas en un transportador común. La primera máquina tiene el doble de rendimiento que la segunda y produce un promedio de 60 % de piezas de calidad excelente. El 84 % de la producción de la segunda máquina es de excelente calidad. Para control se toma al azar una pieza de la cinta transportadora y resulta ser de calidad excelente; ¿cual es la probabilidad de que haya sido fabricada por la primera máquina?*

Si la segunda máquina produce una cantidad de piezas C, entonces la cantidad de piezas que produce la primera es $2C$. De tal modo la producción total es lo que produce la primera máquina más lo que produce la segunda:

$$Totalpiezas = 2C + C = 3C$$

La probabilidad de que una pieza seleccionada al azar provenga de la primera máquina es:

$$P(M_1) = \frac{2C}{3C} = \frac{2}{3} \approx 0{,}67$$

De igual forma puede obtenerse:

$$P(M_2) = \frac{C}{3C} = \frac{1}{3} \approx 0{,}33$$

Además la probabilidad de que una pieza de la primera máquina sea excelente es:

$$P(E/M_1) = 0{,}6$$

Mientras que para la segunda máquina resulta:

$$P(E/M_2) = 0{,}84$$

La probabilidad de que una pieza excelente provenga de la primera máquina se expresa:$P(M_1/E)$ y puede calcularse por la fórmula de Bayes:

$$P(M_1/E) = \frac{P(E/M_1)P(M_1)}{P(E/M_1)P(M_1) + P(E/M_2)P(M_2)} \approx$$
$$\frac{0{,}67 \times 0{,}6}{0{,}67 \times 0{,}6 + 0{,}33 \times 0{,}84} \approx 0{,}592$$

Se puede interpretar que hay una probabilidad 0.592 de que la causa de la excelencia de la pieza es que haya sido producida por la primera máquina.

1.8. Ejercicios

Ejercicio N°1: Calcular $V_{20,2}$, $V_{8,5}$, P_7, $C_{8,5}$, $C_{8,8}$.

Ejercicio N°2: Un estudiante tiene que elegir un idioma y una asignatura entre 5 idiomas y cuatro asignaturas. ¿De cuántas formas distintas puede hacerlo?

Ejercicio N°3*: ¿De cuántas formas distintas puede construirse una fila con 5 personas?

Ejercicio N°4: ¿De cuántas formas distintas puede armarse una mano de póker para 4 personas?

Ejercicio N°5*:

a) ¿De cuántas maneras distintas 3 niños y 2 niñas pueden sentarse en una fila?

b) ¿De cuantas maneras pueden sentarse si las niñas deben hacerlo juntas y los niños también?

c) ¿De cuantas maneras si se pide solo que las niñas estén siempre juntas?

Ejercicio N°6: ¿De cuántas maneras puede escogerse un comité compuesto por 3 hombres y 2 mujeres, de un grupo de 7 hombres y 5 mujeres?

Ejercicio N°7*: Un estudiante tiene que contestar 8 de 10 preguntas en un examen.

a) ¿Cuántas maneras de escoger tiene?

b) ¿Cuántas si las tres primeras preguntas son obligatorias?

c) ¿Cuántas si tiene que contestar 4 de las 5 primeras?

Ejercicio N°8:

a) ¿De cuántas maneras distintas pueden escogerse 2 vocales para un comité de entre 7 personas?

b) ¿De cuantas formas si deben elegirse Presidente y Vice del comité?

Ejercicio N°9*: Sea el experimento lanzar una moneda y un dado simultáneamente.

a) Determinar el Espacio Muestral S.

b) Determinar que elementos de S forman los eventos siguientes:

1- $A = \{aparece\ cara\ y\ un\ numero\ par\}$

2- $B = \{aparece\ un\ numero\ primo\}$

3- $C = \{aparece\ ceca\ y\ un\ numero\ impar\}$

c) Detallar los eventos:

1- Sucede A o B.

2- Sucede A y B.

3- Sucede solamente B.

d) ¿Cuáles de los eventos A, B, y C son mutuamente excluyentes?

Ejercicio N°10: Determinar la probabilidad de cada evento:

a) Que salga un número par al lanzar un dado.

b) Que resulta rey al sacar una carta de un mazo de cartas francesas.

c) Que aparezca una ceca al lanzar tres monedas.

d) Que aparezca una bola blanca al sacar una sola bola de una urna que contiene 4 bolas blancas, 3 rojas y 5 azules.

Ejercicio N°11*: Se escogen al azar 3 lámparas de entre 15 de las cuales 5 son defectuosas. Hallar la probabilidad de que:

a) Ninguna de las tres lámparas seleccionadas sea defectuosa.

b) Solo una lo sea.

c) Al menos una sea defectuosa.

Ejercicio N°12: Se lanza un dado 100 veces. La tabla adjunta detalla los 6 casos posibles y la frecuencia con la que ha aparecido cada número:

Número:	1	2	3	4	5	6
Frecuencia:	14	17	20	18	15	16

Hallar las frecuencias de los siguientes eventos:

a) Aparezca un 3.

b) Aparezca un 5.

c) Aparezca un número par.

d) Aparezca un número primo.

Ejercicio N°13*: Se lanzan dos dados. Hallar la probabilidad de que la suma de sus números sea 10 o mayor si:

a) No hay condiciones.

b) Apareció un 5 en el primer dado.

c) Aparece un 5 en uno de los dados por lo menos.

Ejercicio N°14: Se reparten a una persona 4 tréboles de una baraja corriente de 52 cartas. Si ahora se le dan tres cartas adicionales hallar la probabilidad de que por lo menos una de ellas sea trébol.

Ejercicio N°15*: Una urna contiene 7 bolas rojas y 3 blancas. Se extraen tres bolas una tras otra. Hallar la probabilidad de que las dos primeras sean rojas y la tercera blanca.

Ejercicio N°16: De una caja con 5 bolillas rojas numeradas del 1 al 5 y 3 bolillas blancas numeradas del 6 al 8 se extrae una bolilla al azar. Juzgar la validez de los siguientes enunciados.

a) Sólo hay dos resultados posibles: rojo o blanco. Por lo tanto, la probabilidad de cada uno de ellos es $\frac{1}{2}$.

b) Cualquier bolilla tiene igual probabilidad de salir: $\frac{1}{8}$.

c) Las bolillas rojas tienen mayor probabilidad de salir que las blancas. 3 es roja y 7 es blanca, por lo tanto, es más probable que salga el 3 que el 7.

d) El experimento ya se realizó una vez y salió el 5. Si se vuelve a realizar, sería mucha casualidad que volviese a salir el 5. Por lo tanto, en la siguiente repetición del experimento es más probable que salga el 6, por ejemplo, que el 5.

Ejercicio N°17*: Se arrojó una moneda perfectamente balanceada 4 veces y se obtuvieron cuatro caras. ¿Cuál es la probabilidad de que si se arroja por quinta vez vuelva a obtenerse cara? Justificar brevemente el resultado obtenido.

Ejercicio N°18*: Sean A y B dos eventos con sus respectivas probabilidades: $P(A) = \frac{1}{2}$, $P(B) = \frac{1}{3}$. Se sabe además que $P(A \cup B) = \frac{7}{12}$.

a) Calcular la probabilidad de ocurrencia simultanea de A y B.

b) ¿A y B son independientes?

Ejercicio N°19*: Un análisis para detectar cierta enfermedad de los caballos ofrece un 92 % de seguridad en los enfermos y 98 % en los sanos. Se sabe además que el 9 % de la población caballar del país padece dicha enfermedad. En un laboratorio se hizo un análisis que arrojó resultado positivo. ¿Cuál es la probabilidad de que el caballo analizado esté efectivamente enfermo?

Ejercicio N°20: En una dada población el 0,1 % de los individuos padece una enfermedad. Las pruebas para verificar la presencia del virus que la produce son imperfectas y solo el 95 % de los infectados es detectado positivamente por el análisis respectivo. También ocurre que el 2 % de aquellos que no padecen en realidad la enfermedad dan positivo al hacerse el estudio. ¿Cuál es entonces la probabilidad de que al dar una prueba positiva el paciente esté realmente infectado?

Capítulo 2

Variables aleatorias discretas

2.1. Variables aleatorias

Hasta aquí hemos trabajado con la llamada probabilidad elemental que asigna probabilidades a los sucesos en sí mismos considerados como conjuntos y simbolizados por letras como A, B, etc. Pero la teoría de las probabilidades aumenta significativamente su poder de modelado sobre los hechos del mundo real, cuando incorpora la representación de los posibles resultados de una experiencia por medio de conjuntos numéricos. En particular, si esos números son los reales, que se corresponden geométricamente con los puntos de una recta, es posible aplicar, para el desarrollo y justificación de la teoría, herramientas del análisis matemático y funcional.

El primer concepto que es necesario analizar y construir es el de *variable aleatoria*. Comencemos diciendo que consideraremos, por ahora, sucesos que pueden ser representados por una sola cantidad numérica. Por ejemplo, está muy claro que los posibles resultados de arrojar un dado y registrar la cara que cae hacia arriba son directamente los números que representan esas caras, es decir, 1, 2, 3, 4, 5 y 6. Pero también el tiempo que transcurre hasta que somos atendidos en la caja del supermercado es un suceso que puede ser representado por números. En este caso, llamando a al instante en que llegamos a la cola de la caja y b al momento en que somos atendidos, transcurre un tiempo $x = b - a$ hasta que efectivamente nos atienden, El suceso "tiempo

transcurrido desde la llegada en el instante a hasta ser atendido" es el intervalo $[a, b]$ que puede representarse por su duración o medida X. En particular obsérvese que si se considerase que el tiempo inicial es 0, el intervalo sería el $[0, X]$. Como el tiempo que efectivamente permanece un cliente en la cola se supone varía en forma aleatoria, es decir que en principio le podemos asignar una probabilidad, decimos que X es una *variable aleatoria.*

Si consideramos los ejemplos del dado y del tiempo en la cola de la caja, lo que tienen en común es que a cada suceso que integra el espacio muestral se le asigna un número. Generalizamos entonces el concepto de variable aleatoria como sigue.

Variable aleatoria: Dado el espacio muestral S asociado a una experiencia, una función $X : S \to \Re$, que a cada elemento de S le asigna un número real, se denomina variable aleatoria[1].

Como es claro, solo una parte de los números reales puede ser necesaria para representar los sucesos. Ése es el caso del dado en que para representar las 6 caras bastan precisamente los naturales del 1 al 6. En cambio, para representar el tiempo posible en la cola del súper tendremos que usar todos los números reales en un intervalo entre 0 y el tiempo que falte, por ejemplo, hasta que el supermercado cierre.

2.2. Distribuciones de probabilidad

Una vez que los sucesos del espacio muestral están representados por números, es posible medir la incertidumbre de distintos eventos que ellos pueden conformar. Por ejemplo, se puede querer evaluar la probabilidad de que el dado caiga par ó la de que tengamos que esperar en la cola desde las 18.30 hasta las 19 a lo sumo o desde las 20 hasta las 20.20 como máximo. Es decir; los elementos a los cuales queremos asignar probabilidad no son solo los simples del espacio muestral sino

[1] Hay que observar que la terminología, que es universalmente aceptada, es sin embargo desafortunada. En realidad, como se ve, una variable aleatoria ¡es una...función!

conjuntos formados por ellos. Para el dado, se busca la probabilidad del conjunto $\{2, 4, 6\}$, mientras que para la espera en la cola se trata de medir la probabilidad del intervalo $[18:30, 19] \cup [20, 20:20]$.

En general, entonces, hablamos de eventos que pueden ser simples o compuestos y que son subconjuntos del espacio muestral S. Esos subconjuntos integran, como elementos, el conjunto de partes de S, $Partes(S) = \Omega$.

Si se denota la probabilidad del suceso, correspondiente a que la variable X tome el valor real a, por medio de $P(X = a)$, se puede escribir $P(a \leq X \leq b)$ para representar la probabilidad de que la variable aleatoria X tome un valor en el intervalo $a \leq X \leq b$. De tal forma, si conocemos la probabilidad para todos los valores posibles de a y b, sabemos en realidad cómo se *distribuye* la probabilidad. Consideremos un número real x y la probabilidad de que la variable aleatoria X tome un valor menor o igual que él. Es decir $P(X \leq x)$. El valor de esta probabilidad es función del valor que adopte x. De esta forma podemos anotar: $F(x) = P(X \leq x)$. Tomemos ahora los subconjuntos de la recta real que cumplen las propiedades $X \leq a$ y $a < X \leq b$. En la figura 1 se muestran los intervalos $(-\infty, a]$ y $(a, b]$ que ellas caracterizan:

Figura 1.

Como se aprecia, se trata de conjuntos disjuntos que, por ende, representarán eventos mutuamente excluyentes.

Se tiene:

$$P(X \leq b) = P\left((-\infty, a] \cup (a, b]\right)$$

De acuerdo con nuestra definición de $F(x) = P(X \leq x)$ podemos escribir:

$$F(b) = F(a) + P(a < x \leq b)$$

y entonces:

$$P(a < x \leq b) = F(b) - F(a) \tag{2.2.1}$$

Es decir, conocida la función $F(x)$ para todos los valores que pueda tomar la variable aleatoria, la probabilidad para cualquier evento como el $(a, b]$, queda determinada por la fórmula (2.1). Como la probabilidad es un número no negativo, la forma de construcción garantiza que $F(x) \geq 0$ y entonces, de acuerdo a (2.1) se tiene $F(a) \leq F(b)$. De aquí resulta también que $F(+\infty) = \lim_{x \to +\infty} F(x) = 1$ y $F(-\infty) = \lim_{x \to -\infty} F(x) = 0$. En particular el suceso representado por un solo punto $x = a$, es un evento de probabilidad 0. En efecto: $P(a < x \leq a) = F(a) - F(a) = 0$

Función de distribución: La función $F(x) = P(X \leq x)$ que cumple con las propiedades antedichas se denomina *función de distribución* (o de repartición) de la variable aleatoria X.

2.3. Variables y distribuciones discretas

Hasta aquí venimos aceptando en forma más o menos implícita que la variable aleatoria, tal como se la ha definido, es una variable real. En efecto, así queríamos pensarlo aunque no todos los eventos que puedan representarse tengan el carácter continuo de tal conjunto numérico. Por ejemplo, para el caso de querer determinar la probabilidad de que un dado caiga par nos bastará considerar, como ya hemos mencionado, el conjunto finito y discreto $\{2, 4, 6\}$. En cambio, el intervalo [18.30,

19] de espera en la cola de la caja requerirá pensar en los infinitos no numerables puntos del segmento de tiempo correspondiente. En suma, pensar en la continuidad de ese lapso. Esto quiere decir que la variable aleatoria podría considerarse discreta o continua según lo fueran los eventos que aspirara a representar[2].

En lo que sigue, adoptando el punto de vista de los manuales clásicos de teoría de la probabilidad, utilizaremos una analogía física. Supongamos que el eje real que simboliza a la variable aleatoria X es una barra infinitamente delgada sobre la cual se halla distribuida una masa cuyo valor total es 1. La densidad de esa masa puede entonces variar en distintas zonas de la barra de forma que en algunas habrá más concentración de masa que en otras. Podrá darse, siguiendo este esquema, alguna zona o punto donde la masa sea nula y otros puntos o lugares donde la masa esté muy concentrada. Es decir, hay una forma de distribución de la masa sobre la barra que, en la terminología de probabilidades, se corresponde con la función de distribución de la que hemos hablado.

Supongamos que tenemos, siguiendo nuestra analogía física, el caso en que la masa total se distribuye de manera concentrada en puntos aislados. Hay entonces una sucesión finita o infinita de puntos $x_1, x_2, \ldots$ que representan los únicos valores que puede tomar la variable aleatoria X. De este modo, resulta que $P(X = x_i) = p_i$. Claramente, debe ocurrir $\sum_i p_i = 1$ y la función de distribución tendrá la forma:

$$F(x) = P(X \leq x) = \sum_{x_i \leq x} P(X = x_i) = \sum_{x_i \leq x} p_i$$

La función de distribución acumula probabilidad, sumando las probabilidades de cada uno de los x_i que resultan menores o iguales que x.

Veamos un ejemplo. Sea la probabilidad P de una variable aleatoria X según se presenta en la Tabla 1:

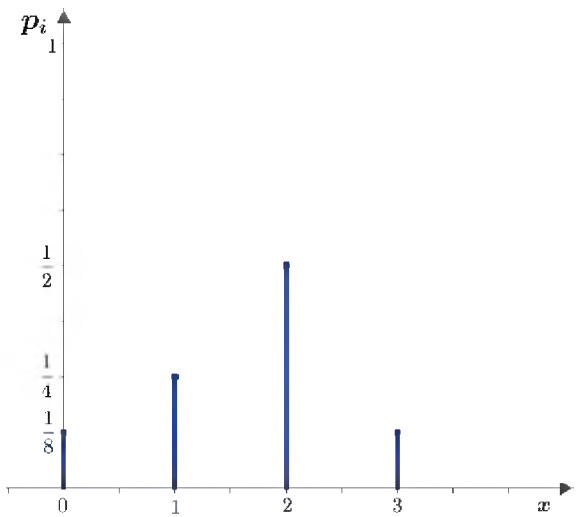
[2]Para la abrumadora mayoría de las aplicaciones no resulta necesario considerar variables que puedan exhibir el doble cariz de continuas y discretas. Tener en cuenta este caso excede los alcances del libro.

Tabla 1.

x_i	$P(X = x_i) = p_i$
0	1/8
1	2/8
2	4/8
3	1/8

La probabilidad concentrada en los puntos x_i se muestra en la figura $2a$):

Figura $2a$.



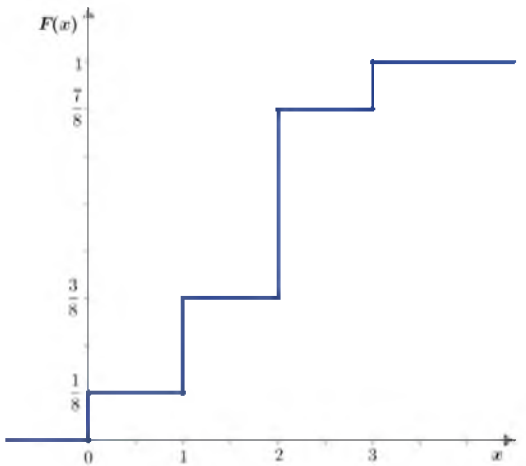
De acuerdo con esto, la función de distribución en cada uno de los puntos considerados acumula probabilidad según la tabla 2:

Tabla 2.

x_i	$F(x_i)$
0	1/8
1	3/8
2	7/8
3	1

Y su gráfica correspondiente, teniendo en cuenta todos los x reales, se ve en la figura 2*b*):

Figura 2*b*.



Al usar nuestra analogía física, si la masa está concentrada en puntos aislados diremos que la variable aleatoria y su distribución pertenecen al tipo discreto, como es el caso que ejemplificamos arriba.

Ejemplo 1: *Se selecciona al azar una muestra sin remplazo de 3 artículos de un total de 10, de los cuales 2 son defectuosos. Si X es la variable aleatoria: número de artículos defectuosos en la muestra, obtener la distribución de probabilidades de la variable aleatoria.*

Sea la variable aleatoria X : *número de artículos defectuosos.*

Como en la muestra hay 2 artículos defectuosos, la variable aleatoria X, que expresa la cantidad de artículos defectuosos, solo podrá tomar los valores 0,1 ó 2. Calculemos para empezar la probabilidad de que X adopte el valor 0, lo que significa que ninguno de los tres artículos seleccionados resultará defectuoso.

Si el primer artículo seleccionado tiene una probabilidad de ser defectuoso de:

$$P(D_1) = \frac{2}{10}$$

entonces, la probabilidad de que no lo sea, al haber 8 artículos que no son defectuosos, es de:

$$P(\bar{D}_1) = \frac{8}{10}$$

Como cada extracción se realiza sin reposición, para elegir el segundo artículo quedan en total 9, de las cuales quedan 7 que no son defectuosos. La probabilidad de que el segundo artículo no sea defectuoso, cuando el primero tampoco lo ha sido, es entonces:

$$P(\bar{D}_2/\bar{D}_1) = \frac{7}{9}$$

Ahora quedan 8 artículos y 6 de ellos no son defectuosos. La probabilidad de que el tercer artículo no sea defectuoso es:

$$P(\bar{D}_3/\bar{D}_2\bar{D}_1) = \frac{6}{8}$$

De acuerdo con este análisis, la probabilidad de que ninguno de los tres artículos seleccionados sea defectuoso es:

$$P(\bar{D}_1\bar{D}_2\bar{D}_3) = P(\bar{D}_1 \cap (\bar{D}_2/\bar{D}_1) \cap (\bar{D}_3/\bar{D}_1\bar{D}_2)) =$$
$$= P(\bar{D}_1)P(\bar{D}_2/\bar{D}_1)P(\bar{D}_3/\bar{D}_1\bar{D}_2)$$

Y entonces la probabilidad de que la variable aleatoria X tome el valor 0 artículos defectuosos es:

$$P(X = 0) = P(\bar{D}_1\bar{D}_2\bar{D}_3) = \frac{8}{10} \cdot \frac{7}{9} \cdot \frac{6}{8} = \frac{7}{15}$$

Para analizar la probabilidad de que la variable aleatoria X tome el valor 1, es necesario considerar tres posibilidades: que la primera de las tres extracciones resulte defectuosa, que la segunda de las tres extracciones resulte defectuosa o que la última de las tres extracciones resulte defectuosa (cada una de estas posibilidades son sucesos mutuamente excluyentes):

$$P\left[(D_1\bar{D}_2\bar{D}_3)\cup(\bar{D}_1D_2\bar{D}_3)\cup(D_1\bar{D}_2D_3)\right]=$$
$$=P(D_1\bar{D}_2\bar{D}_3)+P(\bar{D}_1D_2\bar{D}_3)+P(D_1\bar{D}_2D_3)=$$
$$=\frac{2}{10}\cdot\frac{8}{9}\cdot\frac{7}{8}+\frac{8}{10}\cdot\frac{2}{9}\cdot\frac{7}{8}+\frac{8}{10}\cdot\frac{7}{9}\cdot\frac{2}{8}=\frac{7}{15}$$

Para analizar la probabilidad de que la variable aleatoria X tome el valor 2, es necesario considerar también tres posibilidades:

$$P\left[(D_1D_2\bar{D}_3)\cup(\bar{D}_1D_2D_3)\cup(D_1\bar{D}_2D_3)\right]=$$
$$=P(D_1D_2\bar{D}_3)+P(\bar{D}_1D_2D_3)+P(D_1\bar{D}_2D_3)=$$
$$=\frac{2}{10}\cdot\frac{1}{9}\cdot\frac{8}{8}+\frac{8}{10}\cdot\frac{2}{9}\cdot\frac{1}{8}+\frac{2}{10}\cdot\frac{8}{9}\cdot\frac{1}{8}=\frac{1}{15}$$

Obteniéndose la distribución de probabilidades de la variable X:

x_i	0	1	2
$P(X=x_i)$	7/15	7/15	1/15

Los cálculos pueden simplificarse considerablemente al advertir que se trata de una distribución hipergeométrica, porque la extracción de artículos se realiza sin remplazo lo que quiere decir que una vez seleccionado el artículo no puede ser elegido nuevamente. Esta situación es equivalente a elegir en un solo acto los 3 artículos y si $X=0$ significará que de los 2 defectuosos existentes, en el total de artículos, no se ha elegido ninguno. Tal cosa puede darse de $C_{2,0}$ maneras distintas. A su vez hay $C_{8,3}$ maneras distintas de elegir 3 artículos buenos de entre 8, de modo tal que el número de casos favorables a que el valor de la variable aleatoria X sea 0 es $C_{2,0}\times C_{8,3}$. Como el número de ternas posibles que pueden formarse con 10 artículos es $C_{10,3}$ se tiene que:

$$P(X=0)=\frac{\binom{2}{0}\times\binom{8}{3}}{\binom{10}{3}}=\frac{1\times 56}{120}=\frac{7}{15}$$

En forma análoga se calculan las probabilidades para $X = 1$ y $X = 2$:

$$P(X = 1) = \frac{\binom{2}{1} \times \binom{8}{2}}{\binom{10}{3}} = \frac{2 \times 28}{120} = \frac{7}{15}$$

$$P(X = 2) = \frac{\binom{2}{2} \times \binom{8}{1}}{\binom{10}{3}} = \frac{1 \times 8}{120} = \frac{1}{15}$$

2.4. Distribuciones conjuntas

A veces los sucesos que se presentan como resultado de una experiencia pueden representarse con más de una variable aleatoria. Por ejemplo, si en un control médico se registra el peso y la altura de las personas, el peso se representa por la variable aleatoria X y la altura por la variable aleatoria Y. Así puede determinarse la *probabilidad conjunta* de que una persona pese, por ejemplo, 70 kilogramos y mida 1.75 metros. Es decir $P(X = 70, Y = 1{,}75)$. De acuerdo con ello, y en forma análoga al caso de una sola variable, se define la *función de distribución conjunta* $F(X, Y) = P(X \leq x, Y \leq y)$ que es *bidimensional* y tiene similares propiedades a las analizadas para aquel caso. En general esta idea puede ampliarse a conjuntos de n variables aleatorias y distribuciones n-dimensionales. Si cada una de las variables aleatorias es de tipo *discreto*, la distribución correspondiente será *discreta*.

Para fijar ideas supongamos el caso de una variable aleatoria X que puede tomar los valores 1,2 y 3 y otra variable aleatoria Y cuyos valores pueden ser 10 o 20. El espacio muestral de los resultados posibles está formado por pares del tipo (X, Y).

$$S = \{(1, 10); (1, 20); (2, 10); (2, 20); (3, 10); (3, 20)\}$$

Es práctico representar los valores de probabilidad de estos sucesos en una así llamada *tabla de contingencia*, como puede verse en la tabla 3:

Tabla 3.

X/Y	$Y=10$	$Y=20$	P_X Dist. Marginal de X
$X=1$	0,10	0,20	0,30
$X=2$	0,25	0,15	0,40
$X=3$	0,15	0,15	0,30
P_Y Dist. Marginal de Y	0,50	0,50	1

En general lo que se tiene dentro de la tabla es la probabilidad de que se obtenga el caso i de la variable X *conjuntamente* con el caso j de la variable Y. Es decir $P(X=x_i, Y=y_j)=p_{ij}$. Así, por ejemplo, $P(X=2, Y=10)=0{,}25=p_{21}$. En los márgenes están las llamadas, precisamente, distribuciones *marginales* de X e Y. Estas representan la probabilidad de ocurrencia de cada caso de una de las variables como suma de las probabilidades de ocurrencia de los casos de la otra. Por ejemplo:

$$P_X(X=1)=P(X=1,Y=10)+P(X=1,Y=20)=$$

$$=0{,}10+0{,}20=0{,}30=p_{1*}$$

ó

$$P_Y(Y=20)=P(X=1,Y=20)+P(X=2,Y=20)+$$

$$+P(X=3,Y=20)=0{,}20+0{,}15+0{,}15=0{,}50=p_{*2}$$

Obsérvese que si p_{ij} representa la probabilidad conjunta, las cantidades p_{i*} y p_{*j} simbolizan las respectivas probabilidades marginales.

Recordemos ahora brevemente nuestra definición de sucesos independientes. La independencia de dos sucesos A y B tiene lugar si y solo si $P(A\cap B)=P(A)P(B)$. En el caso de las variables aleatorias, números que representan a sucesos, esta definición se traduce en $P(X=x_i, Y=y_j)=P_X(X=x_i)P_Y(Y=y_j)$ para todo i y para todo j o, lo que resulta equivalente, $p_{ij}=p_{i*}p_{*j}$. Es decir, si en todos los casos posibles de ambas variables aleatorias, la probabilidad conjunta es igual al producto de las marginales entonces las variables son independientes. Recíprocamente se cumple también que si las variables son

independientes, la probabilidad conjunta será el producto de las marginales. Además una propiedad similar vale para la función de distribución conjunta y las funciones de distribución marginales: las variables aleatorias X e Y son independientes si y solo si $F(XY) = F_X(x)F_Y(y)$. En la distribución conjunta de la Tabla 3 se ve claramente que las variables aleatorias no son independientes pues, por ejemplo, la igualdad no se cumple para el caso $0{,}25 = p_{21} \neq p_{2*}p_{*1} = 0{,}40 \times 0{,}50 = 0{,}20$.

2.5. Esperanza y varianza

Cuando los sucesos están representados por variables aleatorias podemos preguntarnos, frente a la realización de una experiencia, qué resultado esperar. Por supuesto habrá distintos resultados posibles pero habrá uno, "en promedio", tal que si se repitiese la experiencia muchas veces sería el que cabría esperar. Por ejemplo, al arrojar un dado muchas veces cierta proporción de las veces caerá 1, ciertas 2 y así. Para fijar ideas supongamos que el dado se ha tirado 100 veces y que el número de veces que ha salido cada cara y la proporción de estas sobre el total son las de la tabla 4:

Tabla 4.

cara	n°de caras	proporción
1	14	14/100
2	18	18/100
3	18	18/100
4	15	15/100
5	19	19/100
6	16	16/100

En "promedio" ha salido el número:

$$\frac{1 \times 14 + 2 \times 18 + 3 \times 18 + 4 \times 15 + 5 \times 19 + 6 \times 16}{100} =$$

$$1 \times \frac{14}{100} + 2 \times \frac{18}{100} + 3 \times \frac{18}{100} + 4 \times \frac{15}{100} + 5 \times \frac{19}{100} + 6 \times \frac{16}{100} = 3{,}51$$

Obsérvese que en realidad 3,51 no es exactamente ninguna cara del dado pero es la que "en promedio" se ha obtenido y la que, si nos basamos exclusivamente en esta experiencia, cabría entonces esperar, al menos en teoría. También hay que notar que, en el segundo miembro de la igualdad, los factores que multiplican a cada valor de la cara del dado son precisamente las proporciones en las que esas caras han caído. Esto nos permite pensar ahora que si la experiencia se repitiese más veces, un número suficiente de veces digamos, esas proporciones coincidirían con la probabilidad 1/6 de salir de cada cara, según hemos visto cuando se analizó la definición experimental de la probabilidad. Entonces un cálculo más ajustado del valor que debiera esperarse "en promedio" surgiría, sin necesidad de hacer ninguna experiencia, de la cuenta que suma los productos de cada cara por su probabilidad:

$$1 \times \frac{1}{6} + 2 \times \frac{1}{6} + 3 \times \frac{1}{6} + 4 \times \frac{1}{6} + 5 \times \frac{1}{6} + 6 \times \frac{1}{6} = 3{,}5$$

Si llamamos X a la variable aleatoria, podemos escribir la fórmula de cálculo del *valor esperado* como $E(X) = \sum_{i=1}^{6} x_i p_i$. Al generalizar para cualquier variable aleatoria discreta se tiene el concepto de *esperanza matemática* que ahora se explicita.

Esperanza matemática: Es el valor esperado de una variable aleatoria discreta X que se calcula como:

$$E(X) = \sum_i x_i p_i$$

donde p_i indica la probabilidad de cada valor x_i.

Es interesante analizar brevemente algunas propiedades de la esperanza matemática.

Propiedad 1: Si X es una variable aleatoria y tanto a como b son constantes se cumple:

$$E(aX + b) = aE(X) + b$$

En efecto:

$$E(aX+b) = \sum_i (ax_i + b)p_i = \sum_i (ax_i p_i + bp_i) =$$

$$= \sum_i ax_i p_i + \sum_i bp_i = a\sum_i x_i p_i + b\sum_i p_i$$

En el último miembro de la cadena de igualdades se tienen las expresiones $\sum_i (x_i p_i) = E(X)$ y $\sum_i (p_i) = 1$. Reemplazando queda $E(aX+b) = aE(X) + b$ que es lo que quería demostrarse.

Propiedad 2: Si X,Y y $X+Y$ son variables aleatorias se cumple que:

$$E(X+Y) = E(X) + E(Y)$$

Admitiremos aquí que $p_{ij} = P(X+Y = x_i + y_i)$ y, sin entrar en demasiado detalle[3], utilizaremos las probabilidades marginales para escribir:

$$E(X+Y) = \sum_{ij} p_{ij} x_i y_j = \sum_i p_{i*} x_i + \sum_j p_{*j} y_j = E(X) + E(Y)$$

que es lo que queríamos demostrar. Es necesario resaltar que esta propiedad es válida para variables aleatorias sin tener en cuenta que una dependa de la otra o que no lo haga.

Propiedad 3: Si las variables aleatorias X e Y son independientes entonces:

$$E(XY) = E(X)E(Y)$$

Resulta:

$$E(XY) = \sum_{ij} p_{ij} x_i y_j = \sum_{ij} p_{i*} p_{*j} x_i y_j =$$

$$= \sum_i p_{i*} x_i \sum_j p_{*j} y_j = E(X)E(Y)$$

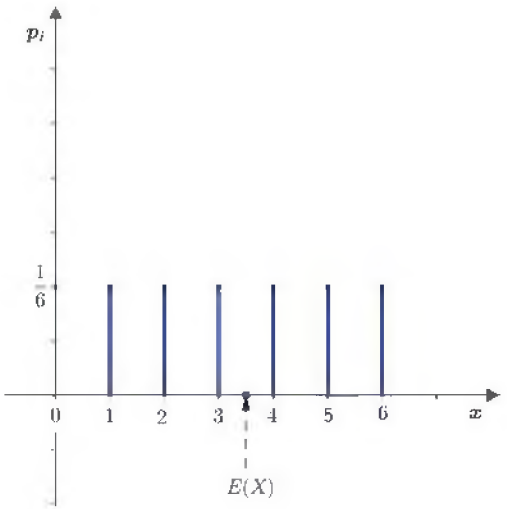
[3]Los detalles necesarios para una prueba rigurosa pueden encontrarse, por ejemplo, en Harald Cramér: *Teoría de probabilidades y aplicaciones*, Aguilar, pp. 73-74.

que es lo que se pretende probar.

Obsérvese que la hipótesis de independencia de las variables X e Y se utilizó en el tercer miembro de la cadena de igualdades al hacer $p_{ij} = p_{i*}p_{*j}$.

Al estudiar la distribución de probabilidad de una variable aleatoria se tiene en cuenta, además de su esperanza matemática, un valor que representa la *variabilidad* de la variable. Al tirar un dado los valores que pueda tomar la cara que cae hacia arriba, simbolizada por la variable aleatoria X, pueden variar en un rango de cantidades enteras desde 1 hasta 6 y parecería que con saber esto, y simultáneamente conocer que la esperanza matemática de X es $E(X) = 3{,}5$, basta para caracterizar adecuadamente la distribución de la probabilidad. El gráfico de la función de probabilidad resulta el que se exhibe en la figura 3:

Figura 3.



Se observa que las posibilidades de variación de la variable a izquierda y derecha del valor esperado son las mismas. A derecha de $E(X)$, X puede tomar 3 valores 4, 5 y 6, todos con probabilidad 1/6 de ocurrir individualmente y a izquierda otros 3 que son el 1, el 2 y el 3, que también tienen, cada uno, probabilidad 1/6 de salir. Pero si

el dado estuviera cargado y fuera más probable que salieran 3 y 4 que 2 y 5 y a su vez estos fueran más probables que 1 y 6, la distribución resultante sería, por ejemplo, la de la tabla 5:

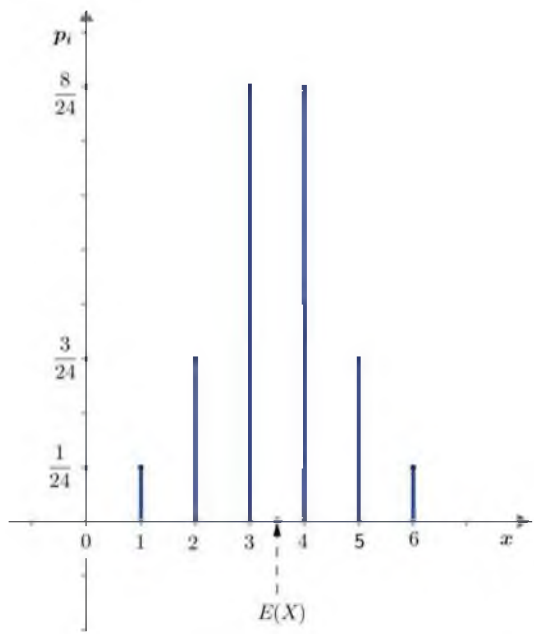
Tabla 5.

x_i	p_i
1	1/24
2	3/24
3	8/24
4	8/24
5	3/24
6	1/24

En este caso la esperanza matemática resultaría la misma. En efecto:

$$E(X) = \frac{1}{24} + 2 \times \frac{3}{24} + 3 \times \frac{8}{24} + 4 \times \frac{8}{24} + 5 \times \frac{3}{24} + 6 \times \frac{1}{24} = 3{,}5$$

Figura 4.



Sin embargo, ahora si bien hay la misma cantidad de valores posibles para la variable a derecha y a izquierda de su esperanza, no todos son igualmente probables. Como se ve en la figura 4 la probabilidad está más concentrada alrededor de los valores 3 y 4 de la variable aleatoria.

Es fácil pensar que si un dado con las caras cargadas de esta forma se arrojara un número grande de veces debiera ocurrir que fueran mucho más frecuentes los resultados 3 y 4 que los 2 y 5, y menos frecuentes fueran todavía los resultados 1 y 6. En suma, habría menor variabilidad de la cara que caería hacia arriba, al arrojar el dado. El asunto es entonces poder medir este efecto.

Se puede considerar cada desvío del valor esperado $(x_i - E(X))$. Como x_i tiene probabilidad p_i es natural asignar a cada desvío esta probabilidad de ocurrir. Entonces el valor esperado de los desvíos se calcula como:

$$E(X - E(X)) = \sum_i (x_i - E(X))p_i = \sum_i x_i p_i - \sum_i E(X)p_i =$$

$$= E(x) - E(X)\sum_i p_i = E(X) - E(X) = 0$$

En el cuarto miembro de esta cadena de igualdades hemos tenido en cuenta que la esperanza es un número que se calcula como $E(X) = \sum_i x_i p_i$ y que, precisamente por ser un número, puede ser extraído como factor común de los sumandos p_i. Además en el quinto miembro se tuvo en cuenta que $\sum_i p_i = 1$. Así las cosas, sabemos ahora que, bajo cualquier distribución de probabilidad, el valor esperado para los desvíos de la variable aleatoria es 0. Dado que siempre es así, esto no nos proporciona información suficiente sobre la variabilidad de la variable.

Una propuesta de cálculo de la variabilidad podría consistir en considerar los valores absolutos de los desvíos para eliminar así la influencia del signo de los mismos. Ocurre que cuando $x_i < E(X)$ resulta $(x_i - E(X)) < 0$, es decir desvío negativo, y que cuando $x_i > E(X)$ se tiene $(x_i - E(X)) > 0$, desvío positivo. Al sumarse multiplicadas por las respectivas probabilidades, las cantidades se compensan y se obtiene, como vimos, $E(X - E(X)) = 0$. Al tomar los valores absolutos de los desvío esto no ocurriría pues todos los sumandos serían positivos y no habría compensación posible. Es decir $\sum_i |(x_i - E(X)|\, p_i \geq 0$ y en

particular solo sería 0 en el caso en que todos los xi resultaran iguales a la $E(X)$ y por ende la variabilidad no existiese. Sin embargo medir la variabilidad de esta forma traería ciertos problemas operativos, pues al aplicar los resultados del análisis matemático a la teoría de las probabilidades, muchas veces resulta necesario derivar o integrar expresiones. En ese caso, operar sobre fórmulas donde intervenga la función módulo, puede agregar innecesarias complicaciones. Por ello resulta universalmente aceptada otra forma de medir la variabilidad[4].

Para lograr una suma que tenga signo positivo sin necesidad de emplear la función módulo se recurre al cuadrado de cada desvío. Así se forma la cantidad $E\left[(X - E(X))^2\right] = \sum_i (x_i - E(X))^2 p_i \geq 0$ que es el valor esperado del cuadrado de los desvíos. Esta cantidad es 0 solo cuando todos los valores de la variable aleatoria son iguales y si no, resulta estrictamente mayor que 0 pues suma cantidades positivas.

Varianza: Se define como *varianza* (o variancia) de una variable aleatoria a la cantidad

$$Var(X) = \sum_i (x_i - E(X))^2 p_i$$

Si se considera ahora el ejemplo anterior del dado legal, con probabilidades constantes 1/6 se tiene:

$$Var(X) = (1 - 3{,}5)^2 \times \frac{1}{6} + (2 - 3{,}5)^2 \times \frac{1}{6} + (3 - 3{,}5)^2 \times \frac{1}{6} +$$
$$+(4 - 3{,}5)^2 \times \frac{1}{6} + (5 - 3{,}5)^2 \times \frac{1}{6} + (6 - 3{,}5)^2 \times \frac{1}{6} \approx 2{,}917$$

mientras que para el dado cargado, cuya distribución de probabilidad se dio en la Tabla 5, la varianza es:

$$Var(X) = (1 - 3{,}5)^2 \times \frac{1}{24} + (2 - 3{,}5)^2 \times \frac{3}{24} + (3 - 3{,}5)^2 \times \frac{8}{24} +$$

[4]En realidad hay un enfoque alternativo, más riguroso matemáticamente, para explicar y deducir las fórmulas de la centralidad y variabilidad de una variable aleatoria. Esto se realiza por medio de la función generadora de momentos y sus fundamentos pueden verse, por ejemplo, en Paul Meyer: *Probabilidad y aplicaciones estadísticas*, Ed. Addison:Wesley Iberoamericana.

$$+(4-3{,}5)^2 \times \frac{8}{24} + (5-3{,}5)^2 \times \frac{3}{24} + (6-3{,}5)^2 \times \frac{1}{24} = 1{,}25$$

Se observa como la concentración de las probabilidades alrededor de los valores 3 y 4 de la variable aleatoria ha resultado en una reducción de la variabilidad de la misma. En efecto, la representación de esa variabilidad por la varianza, para el caso de probabilidad constante del dado legal, es mayor, como cabía esperar, que en el caso del dado cargado. En este sentido podemos decir que la varianza mide adecuadamente la variabilidad.

Sin embargo hay aquí un problema. Cuando se utiliza una variable aleatoria para modelar distintas magnitudes como la cara que caiga un dado o el número de personas que descienda de un colectivo en una parada, a la cantidad en sí resulta siempre adosada la unidad correspondiente. Decimos, por ejemplo, "en la parada A bajaron 3 personas" y no simplemente 3, pues así dicho podrían haber sido 3 perros o 3 marcianos. La unidad "personas" va siempre adosada a la cantidad para que la variable represente a personas y no a otras cosas. Por lo tanto, también la esperanza de la variable estará en "personas" y lo mismo ocurrirá con los desvíos $(x_i - E(X))$. Pero entonces al elevar cada desvío al cuadrado resulta que la cantidad obtenida $(x_i - E(X))^2$ está en "personas al cuadrado" lo que desde un punto de vista fáctico carece de todo sentido. También en "personas al cuadrado" estará entonces la varianza al ser adimensional el factor p_i de cada sumando. Si bien la varianza está captando adecuadamente la variabilidad, la debemos poder tratar en las mismas unidades en que esté la variable aleatoria X. Para ello se introduce la idea de aplicar a la varianza la raíz cuadrada positiva y obtener así una cantidad expresada en unidades adecuadas.

Desviación estándar: La cantidad obtenida al aplicar la raíz cuadrada positiva a la varianza de una variable aleatoria se denomina *desvío estándar* se denota por la letra griega σ(sigma). Se calcula:

$$\sigma = \sqrt{\sum_i (x_i - E(X))^2 p_i}$$

Por tal motivo es común denotar a la varianza por $\sigma^2 = Var(X)$

Analizamos ahora una interesante propiedad de la varianza muy útil para el cálculo de la misma.

Propiedad 4: $Var(X) = E(X^2) - E^2(X)$

En efecto:

$$Var(X) = \sum_i (x_i - E(X))^2 p_i = \sum_i (x_i^2 - 2x_i E(X) + E^2(X)) p_i =$$

$$= \sum_i x_i^2 p_i - 2E(X) \sum_i x_i p_i + E^2(X) \sum_i p_i$$

En el último miembro de la cadena de igualdades se tiene que $E(X) = \sum_i x_i p_i$ y también $\sum_i p_i = 1$ por lo que resulta:

$$Var(X) = E(X^2) - 2E(X)E(X) + E^2(X) =$$
$$= E(X^2) - 2E^2(X) + E^2(X) = E(X^2) - E^2(X)$$

que es lo que se quería demostrar.

Para concluir este apartado consideramos la varianza de dos variables aleatorias independientes.

Propiedad 5: Si X e Y son variables aleatorias independientes entonces

$$Var(X + Y) = Var(X) + Var(Y)$$

Por la Propiedad 4 y teniendo en cuenta además que si las variables aleatorias son independientes, resulta $E(XY) = E(X)E(Y)$, se tiene entonces:

$$Var(X + Y) = E((X + Y)^2) - E^2(X + Y) = E(X^2 + 2XY + Y^2) -$$
$$-E(X + Y)E(X + Y) = E(X^2) + 2E(XY) + E(Y^2) - E^2(X) -$$
$$-E^2(Y) - 2E(X)E(Y) = E(X^2) + 2E(X)E(Y) + E(Y^2) -$$
$$-E^2(X) - E^2(Y) - 2E(X)E(Y) = E(X^2) - E^2(X) + E(Y^2) -$$
$$-E^2(Y) = Var(X) + Var(Y)$$

2.6. Distribuciones discretas

Las distribuciones de probabilidad de las variables aleatorias se obtienen por distintos métodos que involucran aquellas maneras clásica, experimental o subjetiva que hemos considerado en el capítulo 1 y pueden expresarse por medio de tablas como las vistas y también de acuerdo a fórmulas de cálculo que las tipifican. En este caso, existen ciertas formas de distribución de la probabilidad que se presentan frecuentemente al modelar una gran variedad de sistemas y que, por ello, requieren un análisis pormenorizado.

Consideremos un juego consistente en arrojar 3 veces una moneda y contar el número de veces que ha caído cara. Cada vez que se arroja la moneda hay un *binomio* de resultados posibles y mutuamente excluyentes: cara o ceca. Si consideramos un éxito el hecho de que la moneda caiga cara, la probabilidad de éxito será $p = \frac{1}{2}$ y la de fracaso será $q = 1 - p = \frac{1}{2}$. Además en cada tiro, realizado en similares condiciones que los demás, el resultado no depende de lo que haya ocurrido en tiros anteriores y tampoco influirá en los resultados que aparezcan cuando se tire la moneda nuevamente. Las probabilidades de cara y ceca serán siempre las mismas y los ensayos consistentes en arrojar al aire la moneda serán independientes. Resumiendo hay un binomio de resultados posibles, éxito=cara y fracaso=ceca, y $n = 3$ ensayos independientes.

En este contexto uno podría preguntarse, por ejemplo, ¿cuál es la probabilidad de que en los 3 tiros la moneda caiga cara exactamente 2 veces? Podría ocurrir que la sucesión de las tres experiencias repetidas arrojara los resultados CCX, dónde C representa "cayó cara" y X representa "cayo ceca". Como los ensayos son independientes, la probabilidad de que esto ocurra es $p(CCX) = p(C)p(X)p(C) = ppq = p^2q$. Está claro que si la sucesión de 3 ensayos arrojara por ejemplo los resultados CXC la probabilidad $P(CXC) = P(C)P(X)P(C) = pqp = p^2q$ resultaría la misma. Y la misma también resultaría para todas las formas en que de una sucesión de tres tiros de la moneda, exactamente 2 veces cayese cara. Éstas son las combinaciones de 3 elementos tomados de a 2, en este caso $\binom{3}{2} = \frac{3!}{2!(3-2)!} = 3$. En otras palabras, hay 3 formas distintas en que puedan caer exactamente 2 caras cuando se arroja 3 veces la moneda. Todas tienen como vimos igual probabilidad y, como si caen en una cierta sucesión no pueden caer en otra, cada formato de terna es excluyente de cada otro. Por lo tanto, en este caso,

considerando la variable aleatoria k "numero de veces que cae cara la moneda", se tiene que la probabilidad:

$$P(k=2) = P(CCX) + P(CXC) + P(XCC) = \binom{3}{2} p^2 q^{3-2} = 3\left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^1$$

En general entonces, si se realiza n veces un ensayo, cada vez en forma independiente y cada vez con igual binomio de resultados posibles, éxito y fracaso, la probabilidad de que ocurran exactamente k éxitos puede calcularse mediante:

$$P(k) = \binom{n}{k} p^k q^{n-k}$$

Que esta es efectivamente una distribución de probabilidad se comprueba al ver que satisface las condiciones de la definición axiomática dada. En particular se cumple que:

I. $P(k) = \binom{n}{k} p^k q^{n-k} \geq 0$ pues tanto el combinatorio como las probabilidades p y q son números positivos.

II. $\sum_{k=0}^{n} P(k) = \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} = (p+q)^n = (p+1-p)^n = 1^n = 1$
En el tercer miembro de la cadena de igualdades se ha tenido en cuenta que la sumatoria que aparece en el miembro anterior es el desarrollo de la potencia n-ésima de un binomio, como ya se comentó al analizar técnicas de conteo. Además la suma de p y q es 1 pues son probabilidades complementarias.

Distribución binomial: Sean n ensayos independientes del mismo experimento cuyos únicos dos resultados excluyentes pueden ser éxito o fracaso, de probabilidades p y $q = 1-p$ respectivamente. La variable aleatoria k que representa el número de éxitos que puedan ocurrir en los n ensayos, se distribuye en forma *binomial* según:

$$b(k,n,p) = P(k) = \binom{n}{k} p^k q^{n-k}$$

La variable aleatoria discreta k, que como es claro toma valores enteros entre 0 y n, cumple las dos propiedades siguientes.

Propiedad 6:

$$E(k) = \sum_{k=0}^{n} kb(k, n, p) = np$$

Propiedad 7:

$$Var(k) = \sum_{k=0}^{n} (k - E(k)^2)b(k, n, p) = npq$$

Se sugiere al lector intentar demostrarlas o buscar la demostración en la bibliografía. Como corolario de la Propiedad 6 surge que el desvío estándar de la variable aleatoria es:

$$\sigma = \sqrt{npq}$$

Ejemplo 2: *La probabilidad de que un tirador pegue en el blanco es* $1/4$.

a) Si dispara 7 veces ¿cuál es la probabilidad de que dos veces por lo menos de en el blanco?

La variable aleatoria X: *número de disparos que acierta en el blanco*, como se realizan 7 disparos, sólo podrá tomar los valores 0,1,2,3,4,5,6 ó 7. Por lo menos dos veces acertar en el blanco, es lo mismo que decir acertar dos veces o más, significa 2,3,4,5,6 ó 7 veces. Para hacerlo más sencillo y realizar menos cálculos, se puede obtener la probabilidad del complemento de este conjunto, es decir, 0 ó 1 vez y luego restarla de 1.

Además el experimento tiene las siguientes características:

- Consiste en una secuencia de n ensayos, donde n se fija antes del experimento.

- Los ensayos son similares y cada uno puede resultar en uno de los posibles resultados; "éxito" o "fracaso".
- Los ensayos son independientes, por lo que el resultado de cualquier intento particular no influye sobre el resultado de cualquier otro ensayo. Hay que observar que en la práctica se está suponiendo que el tirador es experimentado y ya no mejora su probabilidad de hacer blanco con la reiteración de nuevos ensayos.
- Por eso la probabilidad de éxito y fracaso es constante de un ensayo a otro.

Estamos, entonces, en presencia de un experimento binomial:

$$P\left[(X=0)\cup(X=1)\right] = \binom{7}{0}\left(\frac{1}{4}\right)^0\left(\frac{3}{4}\right)^7 + \binom{7}{1}\left(\frac{1}{4}\right)^1\left(\frac{3}{4}\right)^6 \approx 0{,}445$$

La probabilidad obtenida se resta a la probabilidad del espacio muestral, es decir se le resta a 1, para obtener la probabilidad pedida:

$$P\left[(X=2)\cup(X=3)\cup(X=4)\cup(X=5)\cup(X=6)\cup(X=7)\right] =$$

$$= P(X\geq 2) = 1 - P\left[(X=0)\cup(X=1)\right] \approx 1 - 0{,}445 \approx 0{,}555$$

b) ¿Cuántas veces tiene que disparar para que la probabilidad de pegar por lo menos una vez en el blanco sea mayor que 2/3?

Por lo menos una vez, es una vez o más:

$$P(X\geq 1) \geq \frac{2}{3}$$

Nuevamente, se puede obtener la probabilidad del complemento de este conjunto, es decir:

$$P(X\geq 1) = 1 - P(X=0)$$

$$1 - P(X = 0) \geq \frac{2}{3}$$

Despejando:

$$P(X = 0) \leq \frac{1}{3}$$

$$P(X = 0) = \binom{n}{0} \left(\frac{1}{4}\right)^0 \left(\frac{3}{4}\right)^n \leq \frac{1}{3}$$

$$\left(\frac{3}{4}\right)^n \leq \frac{1}{3}$$

$$n \cdot ln\left(\frac{3}{4}\right) \leq ln\left(\frac{1}{3}\right)$$

$$n \geq \frac{ln\left(\frac{1}{3}\right)}{ln\left(\frac{3}{4}\right)}$$

resulta:

$$n \geq 4$$

Otra distribución de probabilidad discreta muy utilizada es la de Poisson. Si consideramos el número de clientes que llega a un cajero automático en una hora comprendemos que este varía entre 0 y una cierta cantidad entera que desconocemos pero que en principio podríamos suponer tan grande como fuera necesario. Esto último quiere decir matemáticamente que esa cantidad puede crecer indefinidamente. Ahora bien, ¿cuál es la probabilidad de que lleguen al cajero en la siguiente hora exactamente k clientes? Si aceptamos el modelo de Poisson para esta actividad de los clientes[5], podemos calcular la probabilidad requerida por medio de $P(k) = \frac{e^{-\lambda}\lambda^k}{k!}$ donde $\lambda > 0$ es un parámetro cuyo significado veremos enseguida. Pero antes debemos comprobar si efectivamente la fórmula ofrecida corresponde a una probabilidad.

[5]Porqué una variable aleatoria distribuida Poisson representa bien al número de clientes que llegan a un cajero en una unidad de tiempo dada o a otras cantidades similares, puede consultarse en Kai Lai Chung: *Teoría elemental de los procesos estocásticos*, Ed. Reverté, pp. 225-245.

I.

$$P(k) = \frac{e^{-\lambda}\lambda^k}{k!} \geq 0$$

pues $e \approx 2{,}718282 > 0$, k es un entero positivo o 0 y $\lambda > 0$, y por lo tanto los respectivos producto, potencia, factorial y cociente también lo serán.

II.

$$\sum_{k=0}^{\infty} \frac{e^{-\lambda}\lambda^k}{k!} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{-\lambda}e^{\lambda} = e^0 = 1$$

En el tercer miembro de la cadena de igualdades se ha utilizado el desarrollo en *serie de Maclaurin* $e^{\lambda} = 1 + \frac{\lambda}{1} + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \cdots = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!}$. Se trata entonces de una función de probabilidad pues cumple con los requisitos para ello.

Distribución de Poisson: La variable aleatoria discreta k, que toma valores $k = 0, 1, 2 \ldots$ se distribuye Poisson cuando:

$$P(k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

con $\lambda > 0$.

De aquí se desprenden un par de propiedades interesantes.

Propiedad 8: La esperanza de una variable aleatoria distribuída Poisson es:

$$E(k) = \lambda$$

En efecto,

$$E(k) = \sum_{k=0}^{\infty} k\frac{e^{-\lambda}\lambda^k}{k!} = \sum_{k=1}^{\infty} k\frac{e^{-\lambda}\lambda\lambda^{k-1}}{k(k-1)\cdots 1} = \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!}$$

Hasta aquí en el tercer miembro de la cadena de igualdades se ha quitado el sumando para $k = 0$ pues se hace precisamente 0. Ahora

puede realizarse el cambio de variable $u = k - 1$ con lo cual si $k = 1$ resulta $u = 1 - 1 = 0$, y si $k \to \infty$ u también lo hace. Se tiene entonces:

$$E(k) = \lambda e^{-\lambda} \sum_{u=0}^{\infty} \frac{\lambda^u}{u!} = \lambda e^{-\lambda} e^{\lambda} = \lambda e^0 = \lambda$$

como quería probarse. Ahora sí se tiene claridad sobre que es el parámetro λ utilizado en la definición de la probabilidad.

Propiedad 9: La varianza de una variable aleatoria distribuida Poisson es:

$$Var(k) = \lambda$$

Veamos:

$$Var(k) = E(k^2) - E^2(k) = \left(\sum_{k=0}^{\infty} k^2 \frac{e^{-\lambda} \lambda^k}{k!} \right) - \lambda^2 =$$

$$= \left(\sum_{k=1}^{\infty} k \frac{e^{-\lambda} \lambda \lambda^{k-1}}{(k-1)!} \right) - \lambda^2 = \left(\lambda e^{-\lambda} \sum_{k=1}^{\infty} k \frac{\lambda^{k-1}}{(k-1)!} \right) - \lambda^2$$

Si se realiza el cambio de variable $u = k - 1$ se tiene $k = u + 1$ y los valores a sumar de u, como en la demostración de la Propiedad 8, van desde 0 a ∞. Entonces:

$$Var(k) = \left[\lambda e^{-\lambda} \sum_{k=0}^{\infty} (u+1) \frac{\lambda^u}{u!} \right] - \lambda^2 =$$

$$= \left[\lambda e^{-\lambda} \sum_{k=0}^{\infty} u \frac{\lambda^u}{u!} + \lambda e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^u}{u!} \right] - \lambda^2 =$$

$$= \left[\lambda \sum_{k=0}^{\infty} u \frac{e^{-\lambda} \lambda^u}{u!} + \lambda e^{-\lambda} e^{\lambda} \right] - \lambda^2$$

y se tiene finalmente

$$Var(k) = (\lambda\lambda + \lambda e^0) - \lambda^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$$

que es lo que se pretendía probar.

La esperanza y la varianza de una variable aleatoria distribuida según Poisson es entonces el mismo parámetro λ que figura en la fórmula de la función de probabilidad.

Ejemplo 3: *Se sabe que durante ciertas horas del día las llamadas telefónicas a una central están distribuidas al azar según un proceso de Poisson con un promedio de 4 llamadas por minuto. Calcular la probabilidad de que:*

a) Transcurran dos minutos sin llamadas.

El valor esperado de llamadas por minuto es 4, pero aquí la variable aleatoria X deberá ser número de llamadas en dos minutos. Como se trata de un proceso de Poisson, ésta variable también se distribuye Poisson y entonces su esperanza se calcula:

$$E(X) = 2 \times 4 = 8 = \lambda$$

Entonces para encontrar la probabilidad de que transcurran dos minutos sin llamadas, se deberá buscar la probabilidad de que la variable X tome el valor 0:

$$P(X = 0) = \frac{8^0 \cdot e^{-8}}{0!} = e^{-8} \approx 0{,}0003$$

b) En un minuto haya por lo menos dos llamadas.

Por lo menos dos llamadas es lo mismo que decir dos llamadas o más. La variable aleatoria X, número de llamadas por minuto, puede tomar los valores 2,3,4,5,6,7,8,9,10,... El inconveniente es que no tenemos un límite superior. Por eso debemos obtener la probabilidad del complemento de este conjunto, es decir, 0 ó 1 y restarla a la probabilidad del espacio muestral (restarla a 1).

El valor esperado de llamadas por minuto es $\lambda = 4$.

$$P(X \geq 2) = 1 - P\left[(X = 0) \cup (X = 1)\right]$$

$$P(X \geq 2) = 1 - \left(\frac{4^0 \cdot e^{-4}}{0!} + \frac{4^1 \cdot e^{-4}}{1!}\right) \approx 0{,}9084$$

c) En tres minutos se produzcan exactamente 10 llamadas.

La variable X aquí debe representar número de llamadas en tres minutos y siendo el valor esperado de llamadas por minuto 4, su esperanza será:

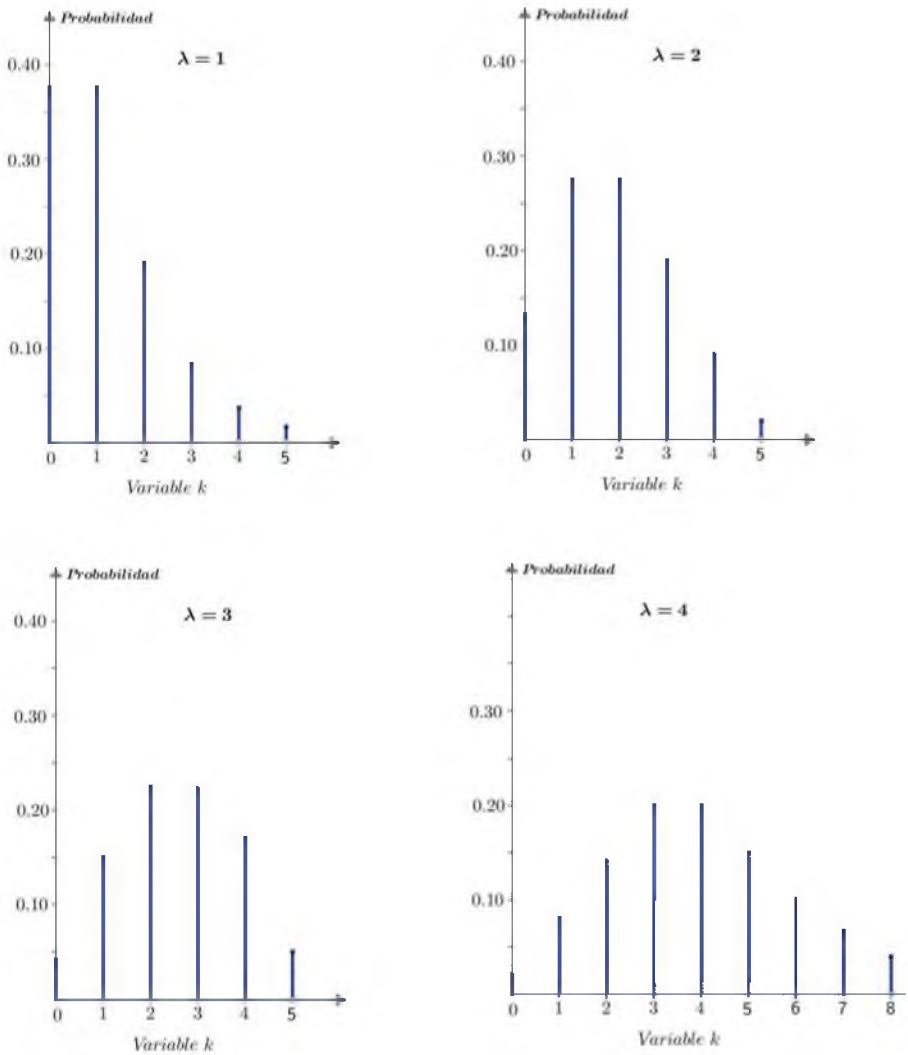
$$E(X) = 3 \times 4 = 12 = \lambda$$

Si se busca encontrar la probabilidad de que se produzcan 10 llamadas en ese lapso. Se debe calcular la:

$$P(X = 10) = \frac{12^{10} \cdot e^{-12}}{10!} \approx 0{,}105$$

Es interesante observar el efecto que la variación de este parámetro tiene sobre la gráfica de la probabilidad. Vemos en la figura 5 las gráficas correspondientes a la distribución de probabilidad conforme va creciendo el valor de λ.

Figura 5.



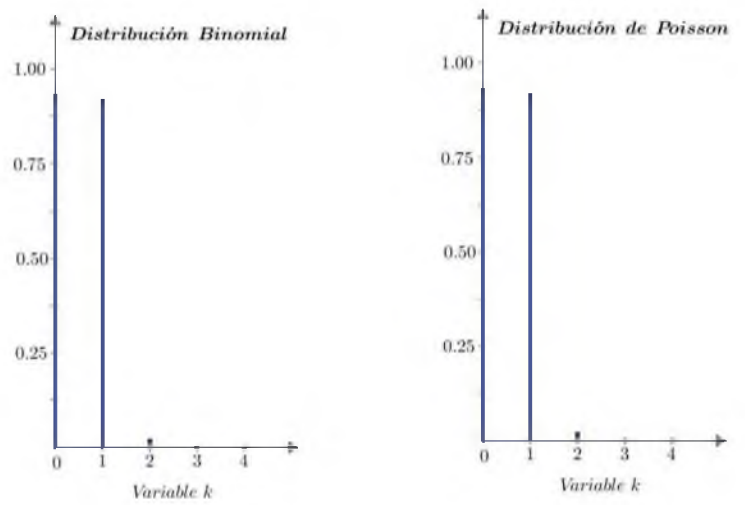
Cuando λ crece la gráfica de la distribución se va haciendo más simétrica. Para valores de $\lambda < 1$ se puede observar un notorio parecido de la gráfica con las correspondientes a ciertas distribuciones binomiales. Si, por ejemplo, se realizan 10 ensayos independientes, cada uno de los cuales tiene solo dos resultados posibles, éxito o fracaso, con una probabilidad de éxito $p = 0{,}01$, la probabilidad de ocurran k éxitos se distribuye en forma binomial según:

$$b(k, 19, 0{,}01) = \binom{10}{k}(0{,}01)^k(0{,}99)^{10-k}$$

La figura 6 ofrece, para comparar, las gráficas de esta distribución binomial y una de Poisson con valor esperado:

Figura 6.

$$\lambda = 0{,}1$$



Se observa un notable parecido entre ambas distribuciones lo que además puede confirmarse inspeccionando los valores de probabilidad respectivos dados en la tabla 6:

Tabla 6.

V.A.	Distr.Binomial	Distr.Poisson
$k = 0$	0,9044	0,9048
$k = 1$	0,9014	0,9005
$k = 2$	0,0042	0,0045
$k = 3$	0,0001	0,0002
$k = 4$	0,0000	0,0000

Hay que notar que al calcular el valor esperado de la variable aleatoria binomial este coincide con el valor de la esperanza de la variable aleatoria[6] distribuída Poisson:

$$E(X) = np = 10 \times 0{,}01 = 0{,}1 = \lambda$$

En general si el valor λ es *pequeño*, en adelante aceptaremos que la distribución de probabilidad binomial puede aproximarse por una distribución de Poisson con igual valor esperado para la variable aleatoria. Esto proporciona una buena forma de evaluar probabilidades binomiales en casos donde el número de ensayos es muy *grande*, pues se elude así la dificultad de tener que calcular un número combinatorio con factoriales de n.

2.7. Ejercicios

Ejercicio N°1*: Considerar el experimento de lanzar un par de dados. Sean las variables aleatorias X: suma de los puntos obtenidos e Y: mínimo puntaje de los dos obtenidos. Construir las tablas de función de probabilidad de estas variables aleatorias discretas.

Ejercicio N°2: Hallar el valor esperado $E(x)$, la varianza σ^2 y el desvío estándar σ para cada una de las variables aleatorias distribuidas como sigue:

x_i	2	3	11
$f(x_i)$	1/3	1/2	1/6

x_i	1	3	4	5
$f(x_i)$	1/4	1/8	1/2	1/8

x_i	-5	-4	1	2
$f(x_i)$	0,4	0,1	0,2	0,3

[6]Una justificación matemática rigurosa puede verse en Paul Meyer ob. cit., pp. 165 a 170.

Ejercicio N°3*: Se lanza un dado. Designamos a X como el doble del número que aparezca y denotamos Y como 1 o 3 según el número sea impar o par respectivamente. Hallar entonces la distribución de probabilidades, la esperanza y la varianza de las variables aleatorias:

a) X *b)* Y *c)* $X+Y$ *d)* XY

Ejercicio N°4*: Se lanza una moneda hasta que salga cara o hasta que salgan 5 cruces. Hallar el valor esperado del número de lanzamientos de la moneda.

Ejercicio N°5*: Se selecciona al azar una muestra con reemplazo de 3 artículos de un total de 10, de los cuales 2 son defectuosos. Si X es la variable aleatoria: número de artículos defectuosos en la muestra, obtener la distribución de probabilidades de la variable aleatoria.

Ejercicio N°6*: Una moneda se lanza tres veces. Hallar la probabilidad de que salgan:

a) 3 caras exactamente.

b) 2 caras exactamente.

c) 1 cara exactamente.

d) 1 cara al menos.

e) Ninguna cara.

Ejercicio N°7: Graficar la función de distribución de probabilidad binomial para: $n = 7$; $p = 0{,}1$; 0.5 ; 0.9

a) Observar la simetría para $p = 0{,}5$ y la dirección del sesgo en los otros casos.

b) Calcular $E(X)$ y $Var(X)$ en cada caso.

Ejercicio N°8*: El equipo A tiene 2/3 de probabilidad de ganar cada vez que juega. Si juega 4 partidas hallar la probabilidad de que A gane:

a) dos partidos exactamente.

b) un partido por lo menos.

c) mas de la mitad de los partidos.

Ejercicio N°9*: Determinar el número esperado de niños de una familia con ocho hijos suponiendo que la distribución del sexo es igualmente probable. ¿Cuál es la probabilidad de que el número esperado de niños suceda?

Ejercicio N°10*: La probabilidad de que un artículo producido por una fábrica sea defectuoso es 0,02. Un cargamento de 10000 artículos se envía a sus almacenes. Hallar el número esperado de artículos defectuosos y la desviación estándar.

Ejercicio N°11: Sea $k = 0, 1, 2, \ldots$ y $\lambda > 0$ probar que $p(k) = \frac{e-\lambda^k}{k!}$ es una función de probabilidad. Graficar para $\lambda = 1$,2,5 y 10.

Ejercicio N°12*: Un artillero dispara a un blanco y sabe que la probabilidad de acertar es $p = 0{,}01$. ¿Cuántos disparos tendrá que realizar para tener probabilidad mayor que 90 % de dar en el blanco por lo menos una vez? Resolver usando las distribuciones binomial y de Poisson.

Ejercicio N°13*: El 2 % de los artículos que produce una fábrica es defectuoso. Hallar la probabilidad de que haya 3 artículos defectuosos en una muestra de 100 artículos fabricados.

Ejercicio N°14: Un comité de 5 personas se selecciona al azar entre 3 abogados y 5 contadores. Hallar la probabilidad de que el número de abogados en el comité sea dos.

Ejercicio N°15: Unos lotes de 40 componentes cada uno son aceptados si no contienen más de tres componentes defectuosos. Para muestrear el lote se eligen al azar 5 componentes y se rechaza el lote si hay uno o más defectuosos. ¿Cuál es la probabilidad de encontrar un componente defectuoso en la muestra si hay 3 en el lote?

Ejercicio N°16*: Un fabricante de neumáticos para automóvil dice que en un embarque de 5000 neumáticos, enviados a una distribuidora local, 1000 están ligeramente dañados. Si se le compran 10 neumáticos al azar ¿cuál es la probabilidad de que exactamente 3 estén dañados?

Capítulo 3

Variables aleatorias continuas

3.1. Función de densidad

Hemos citado ya la analogía entre la recta real y una barra infinitamente larga y delgada cuya masa total es 1. Si en vez de concentrarse en ciertos puntos aislados la masa se distribuye en forma continua a lo largo de la barra, la variable aleatoria y su distribución de probabilidad serán continuas. La *densidad* que tiene la masa en cada punto x es función del mismo y la denotamos $f(x)$. Se expresa así una magnitud positiva ó a lo sumo 0, cuando no hay masa en un punto, por lo que resulta $f(x) \geq 0$. La cantidad de masa contenida en un intervalo infinitesimal $[x, x+dx)$ es $f(x)dx$. Esta cantidad, en virtud de nuestra analogía, no es otra cosa que la probabilidad correspondiente al intervalo infinitesimal. Se trata entonces de un diferencial de la función de distribución tal que $dF = f(x)dx$ por lo que claramente $F'(x) = \frac{dF}{dx} = f(x)$ ya que definimos $F(x) = P(X \leq x)$ que acumula la probabilidad del intervalo $(-\infty, x]$.

Función de densidad de probabilidad: Así se denomina a la derivada de la función de distribución $F'(x) = f(x)$.

Surgen de inmediato dos propiedades características de toda función de densidad de probabilidad.

Propiedad 1: $f(x) \geq 0$ para todo x real.

En efecto, como ya se ha comentado, la función de distribución acumulada, definida para todos los puntos de la recta real, es monótona creciente por lo que su derivada debe resultar mayor o igual a 0 en todos los puntos de definición. Es decir, $\frac{dF}{dx} = f(x) \geq 0$ como quería probarse.

Propiedad 2: $\int_{-\infty}^{\infty} f(x)dx = 1$

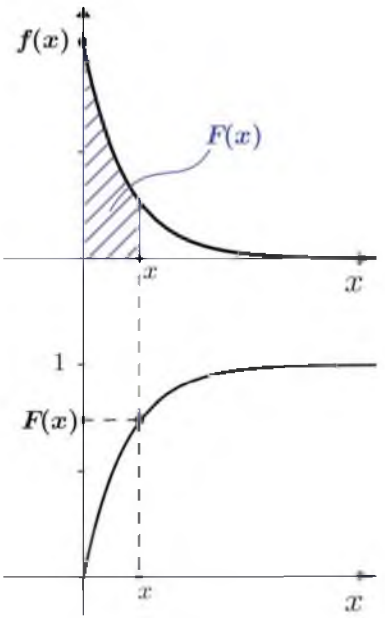
Si $F'(x) = f(x)$ entonces $P(a \leq x \leq b) = F(b) - F(a) = \int_a^b f(x)dx$. Claramente si $a \to -\infty$ resulta $F(a) = 0$ y si $b \to \infty$ se tiene que $F(b) = 1$. De acuerdo a esto $\int_{-\infty}^{\infty} f(x)dx = 1 - 0 = 1$ como se buscaba demostrar.

Hay que notar que si $f(x)$ satisface estas dos propiedades, la expresión:

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

es la de una función de distribución de probabilidad. La x colocada en el extremo de integración indica el punto hasta el cual se acumula la probabilidad ya que, como se sabe, $F(x) = P(X \leq x)$. Así, t es la forma de llamar a la variable de integración que toma valores en el intervalo $(-\infty, x]$. La relación entre la función de distribución y la de densidad puede verse en la figura 1. El área sombreada bajo la curva de $f(x)$, función de densidad de probabilidad, tiene por medida el valor $F(x)$ de la función de distribución:

Figura 1.



Ejemplo 1: *Dada* $f(x) = \frac{1}{4}(x+1)^3$ *si* $-1 \leq x \leq 1$ *y 0 en otro caso.*

a) Probar que es una función de densidad de probabilidad.

La función de densidad dada para la variable aleatoria continua x, se puede reescribir:

$$f(x) = \begin{cases} \frac{1}{4}(x+1)^3 & \text{si } -1 \leq x \leq 1 \\ 0 & \text{para otro caso} \end{cases}$$

Tal función debe cumplir las siguientes dos condiciones:

i. $f(x) \geq 0$

En palabras, la función deberá ser siempre positiva o 0. En este caso, como $(x+1) \geq 0$ para todo $x \in [-1;1]$, $f(x)$ es mayor o igual que 0 en ese intervalo.

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

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

$$= \int_{-\infty}^{-1} 0 + \int_{-1}^{1} \frac{1}{4}(x+1)^3 dx + \int_{1}^{+\infty} 0 = \frac{1}{4}\int_{-1}^{1} (x+1)^3 dx$$

Sustituyendo $t = x + 1$ resulta $\frac{dt}{dx} = 1$ y por lo tanto $dt = dx$. Además si $x = -1$, $t = -1 + 1 = 0$ y si $x = 1$, $t = 1 + 1 = 2$. Se tiene entonces:

$$\frac{1}{4}\int_{-1}^{1} (x+1)^3 dx = \frac{1}{4}\int_{0}^{2} t^3 dt = \frac{1}{4}\frac{t^4}{4}\Big|_0^2 = \frac{1}{4}(\frac{2^4}{4} - 0) = 1$$

Como la función es 0 fuera del intervalo $[-1; 1]$, la operación realizada calcula el área baja la curva $f(x) = \frac{1}{4}(x + 1)$ definida en ese intervalo.

b) Construir la distribución acumulada.

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

Para cada x, $f(x)$ representa, geométricamente hablando, el área bajo la curva de densidad a la izquierda de x y es la acumulación de las probabilidades desde $-\infty$ hasta el valor de x.

$$\int_{-\infty}^{x} \frac{(u+1)^3}{4} du = \frac{1}{4}\int_{-\infty}^{-1} 0\, du + \frac{1}{4}\int_{-1}^{x} \frac{(u+1)^3}{4} du =$$

$$= \frac{1}{4}\int_{0}^{x+1} t^3\, dt = \frac{1}{4}\frac{t^4}{4}\Big|_0^{x+1} = \frac{1}{16}\left[(x+1)^4 - 0\right] = \frac{1}{16}(x+1)^4$$

Para calcular esta integral, primero se tuvo en cuenta que la variable de la función F es la x, supremo de la integración y se nombró por u a

la variable operatoria de integración. Además se realizó la sustitución $t = u + 1$ con lo cual $dt = du$ y se obtuvieron los extremos $t = 0$ para $u = -1$ y $t = x + 1$ para $u = x$. Finalmente queda:

$$F(x) = \begin{cases} \frac{(x+1)^4}{16} & \text{si } -1 \leq x \leq 1 \\ 0 & \text{en otra} \end{cases}$$

Se observa en particular que $F(-1) = 0$ y $F(1) = 1$ con lo cual se comprueba que toda la probabilidad se acumula desde $x = -1$ hasta $x = 1$.

c) Hallar $F(x > 0{,}2)$

En la práctica $F(x)$ permite hallar la probabilidad acumulada desde -1 hasta un cierto valor de x, en términos geométricos el área bajo la curva de densidad que se ubica a la izquierda de x. En este caso la probabilidad pedida coincidirá con el área a la derecha de 0.2 y se hará necesario calcularla a través del complemento:

$$F(x > 0{,}2) = 1 - F(x \geq 0{,}2) = 1 - F(0{,}2) =$$

$$= 1 - \frac{(0{,}2 + 1)^4}{16} = 0{,}8704$$

Conviene explicar el uso libre que se hace del término distribución. Hemos definido *función de distribución o de repartición* de la probabilidad por un lado y por el otro, *función de densidad*. Sin embargo la palabra distribución suele ser empleada, en el contexto de la teoría, sobreentendiendo el conjunto formado por ambas, en un sentido amplio y como sinónimo del término función. De tal forma a veces se utiliza la palabra distribución, seguida de un nombre específico, haciendo referencia a la función de densidad. Por ejemplo se cita, como veremos, a la *distribución uniforme*, la *distribución normal* o la *distribución exponencial*. Esta es una práctica largamente extendida en la bibliografía y es por ese motivo que realizamos la aclaración.

Una extensión de los razonamientos efectuados para las variables aleatorias discretas culmina naturalmente en las siguientes definiciones.

Esperanza matemática: Para una variable aleatoria continua X se define la esperanza matemática como:

$$E(X) = \int_{-\infty}^{+\infty} x \cdot f(x)\, dx$$

Varianza: Si X es una variable aleatoria continua la varianza es:

$$Var(X) = \int_{-\infty}^{+\infty} (x - E(x))^2 f(x)\, dx$$

En forma análoga se define el *desvío estándar* como:

$$\sigma = \sqrt{Var(X)}$$

Además se extienden también las propiedades vistas para estas cantidades cuando las variables aleatorias son discretas.

Propiedad 3: Si se trata de variables aleatorias continuas resultan válidas:

(a) $Var(X) = E(X^2) - E^2(X)$

(b) $E(aX + b) = aE(X) + b$

(c) $E(X + Y) = E(X) + E(Y)$

(d) $E(XY) = E(X)E(Y)$ si y solo si X e Y son variables aleatorias independientes.

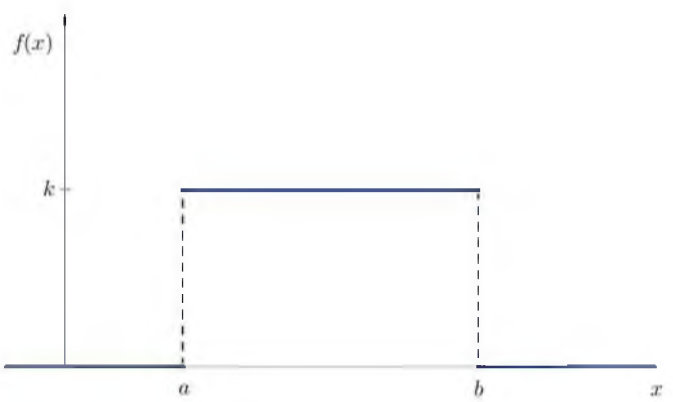
(e) Si X e Y son variables aleatorias continuas independientes entonces $Var(X + Y) = Var(X) + Var(Y)$

3.2. Distribuciones continuas

En el caso de las variables aleatorias continuas son bien conocidas ciertas funciones de densidad y sus correspondientes funciones de distribución que son utilizadas en una amplia gama de situaciones. Analizaremos en primer término el caso de la llamada distribución *uniforme*.

Si X es una variable aleatoria con función de densidad de probabilidad dada por una constante diremos que la distribución de la probabilidad es uniforme. Supongamos el caso de $f(x) = k$ en un intervalo $[a, b]$ y 0 fuera de él. La figura 2 muestra la gráfica de tal función:

Figura 2.



Para que una función de este tipo sea de densidad de probabilidad deberán verificarse las propiedades 1 y 2 del apartado anterior. Para el primer caso bastaría con que se cumpla $f(x) = k \geq 0$, es decir, que k fuese positivo o 0. Sin embargo la segunda propiedad determina unívocamente el valor de k. En efecto, debe cumplirse que $\int_{-\infty}^{+\infty} f(x)\,dx = 1$, entonces $\int_{-\infty}^{+\infty} f(x)\,dx = \int_{-\infty}^{a} 0\,dx + \int_{a}^{b} k\,dx + \int_{b}^{+\infty} 0\,dx = \int_{a}^{b} k\,dx = 1$ y resolviendo la integral se tiene $\int_{a}^{b} k\,dx = kx|_{a}^{b} = kb - ka = k(b-a) = 1$ de donde resulta finalmente el valor de $k = \frac{1}{b-a}$ que como $a < b$ resulta siempre positivo. En resumen, para $a \leq x \leq b$ la función $f(x) = \frac{1}{b-a}$, y en otra parte 0, es de densidad de probabilidad:

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{si } a \leq x \leq b \\ 0 & \text{en otra parte} \end{cases}$$

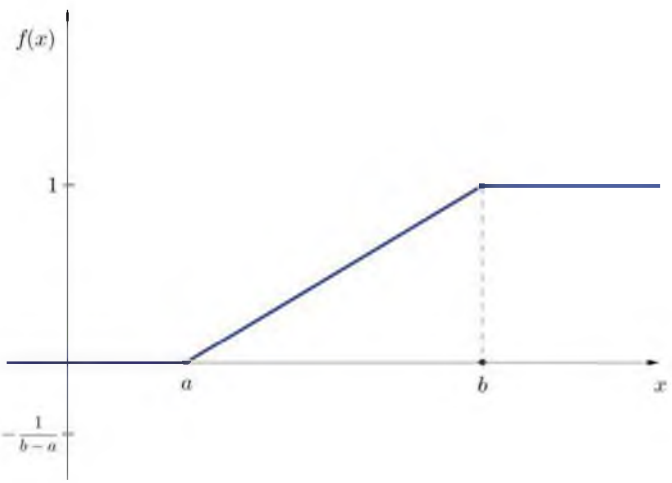
La *distribución acumulada* o *función de distribución* se calcula mediante:

$$F(x) = \int_{-\infty}^{x} f(t)\,dt = \int_{-\infty}^{a} 0\,dt + \int_{a}^{x} \frac{1}{b-a}\,dt = \frac{1}{b-a}t\Big|_{a}^{x} =$$

$$= \frac{1}{b-a}(x-a) = \frac{1}{b-a}x - \frac{a}{b-a}$$

La figura 3 muestra la función de distribución que entre a y b es un segmento de la recta de pendiente $\frac{1}{b-a}$ y ordenada al origen $-\frac{a}{b-a}$. En el conjunto $(-\infty, a)$ la función vale 0 y 1 es su valor en $(b, +\infty)$:

Figura 3.

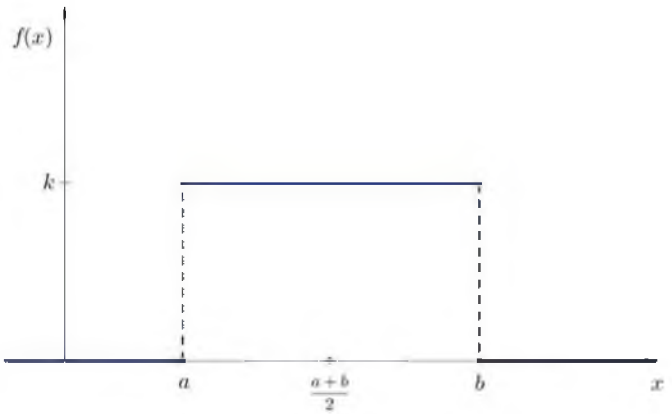


La esperanza matemática de una variable distribuida uniformemente se calcula como:

$$E(X) = \int_{-\infty}^{+\infty} x(x)\,dx = \int_{a}^{b} x\frac{1}{b-a}\,dx = \frac{1}{b-a}\int_{a}^{b} x\,dx =$$

$$= \frac{1}{b-a}\frac{x^2}{2}\Big|_{a}^{b} = \frac{1}{2(b-a)}(b^2-a^2) = \frac{1}{2(b-a)}(b-a)(b+a) = \frac{b+a}{2}$$

En la figura 4 se puede observar como la esperanza de la variable se encuentra en el punto medio del intervalo $[a, b]$:

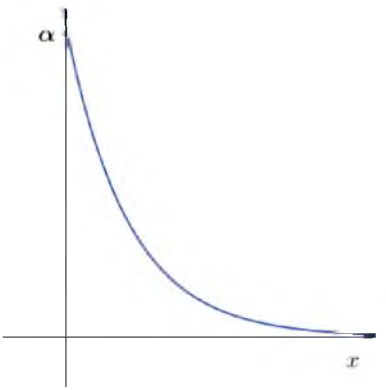


El tiempo que una persona permanece dentro de un cajero automático puede modelarse con la llamada distribución exponencial. La fórmula correspondiente a la función de densidad de probabilidad es en este caso:

$$f(x) = \begin{cases} \alpha e^{-\alpha x} & \text{si} x \geq 0 \\ 0 & \text{si} x < 0 \end{cases}$$

donde $\alpha > 0$ es una constante. Es relativamente sencillo probar que esta función, cuya gráfica se ve en la figura 5, cumple con las propiedades 1 y 2.

Figura 5.



Para empezar debe considerarse que tanto α como $e^{-\alpha x} = \frac{1}{e^{\alpha x}}$ son cantidades mayores que cero y por lo tanto resulta para todo x que $f(x) \geq 0$. Además se tiene que:

$$\int_{-\infty}^{+\infty} f(x)\,dx = \int_{0}^{+\infty} \alpha e^{-\alpha x}\,dx = \alpha \lim_{t\to\infty} \int_{0}^{t} e^{-\alpha x}\,dx =$$

$$= \alpha \lim_{t\to\infty} \frac{e^{-\alpha x}}{-\alpha}\bigg|_0^t = -(\lim_{t\to\infty} e^{-\alpha t} - 1) = 1$$

La función de distribución se obtiene como sigue:

$$F(x) = \int_{-\infty}^{x} \alpha e^{-\alpha t}\,dt = \alpha \int_{0}^{x} e^{-\alpha t}\,dt = \alpha \frac{e^{-\alpha t}}{-\alpha}\bigg|_0^x =$$

$$-(e^{-\alpha x} - 1) = 1 - e^{-\alpha x}$$

Al continuar con los cálculos se puede ver que $E(X) = \frac{1}{\alpha}$ y resulta interesante comentar aquí que el valor esperado de una variable exponencial resulta el recíproco de la esperanza de una variable que se distribuye Poisson. Por ejemplo si el tiempo que permanece una persona en el cajero automático es una variable aleatoria distribuida exponencialmente con un valor esperado de 2 minutos se tiene $E(x) = 2 \textit{minutos/persona}$. El recíproco es $1/E(x) = 1/2$ *personas/minuto*. Si ahora se multiplica por los 60 minutos que tiene la hora resulta $\lambda = 30$ *personas/hora*. Se puede demostrar que este valor λ es el esperado para una variable aleatoria distribuida Poisson que modela en este caso el número de clientes que son atendidos por el cajero automático cada hora[1].

3.3. La distribución normal

Es frecuente modelar sistemas usando la distribución normal, llamada así precisamente por su presencia habitual en la explicación de un gran número de hechos. Son variables aleatorias distribuidas normalmente la nota que pueda obtener un estudiante en un examen, el peso que registre una persona, la altura etc. A pesar de esto, su

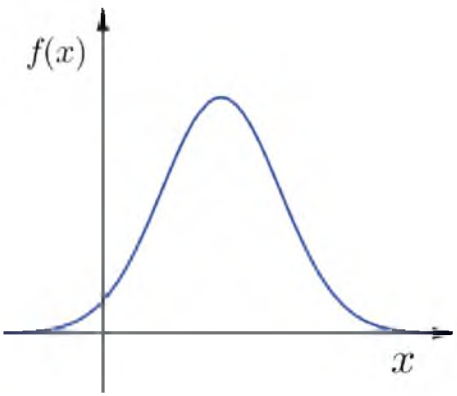
[1]La relación teórica entre las distribuciones de Poisson y exponencial puede verse en George C. Canavos: *Probabilidad y estadística. Aplicaciones y métodos*, Editorial McGraw-Hill.

tratamiento matemático no es tan sencillo como el de las distribuciones continuas vistas hasta aquí. Comencemos por dar la *función de densidad* correspondiente:

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

con $\sigma > 0$ y μ parámetros cuya significación luego veremos. La variable x toma todos los valores reales del intervalo $(-\infty, \infty)$. La gráfica de esta función puede verse en la figura 6:

Figura 6.



Esta forma acampanada y simétrica es la más popular en la probabilidad y la estadística. La fórmula de la distribución fue establecida por Abraham De Moivre[2] aunque generalmente se atribuye al alemán Karl Friederich Gauss reconocido como el matemático más grande de la historia moderna. Así el nombre habitual de esta figura es *campana de Gauss*.

Los problemas comienzan cuando se trata de probar que la función así definida es de densidad de probabilidad intentando demostrar para ello que se cumplen las propiedades 1 y 2 de la *sección 1*. Resulta simple ver que $f(x) \geq 0$ pues al ser $\sigma > 0$ resulta $\frac{1}{\sqrt{2\pi}\sigma} > 0$ y además como $\frac{1}{e^{\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}} > 0$ el producto de ambos factores será positivo. Esto explica, por otra parte, que las "colas" de la campana, véase figura 6,

[2]Véase Pablo Jacovskis y Roberto Perazzo: *Azar, ciencia y sociedad*, EUDEBA, p.52

resulten asintóticas con el eje x siendo siempre los valores positivos. Pero al tratar de probar que $\int_{-\infty}^{\infty} f(x)dx = 1$ se tropieza con el hecho de que la primitiva de la función es un desarrollo en serie y por lo tanto no hay una fórmula cerrada, es decir una fórmula que posea un número limitado de expresiones simples, que la represente. En principio podría entonces hacerse un cálculo aproximado truncando el desarrollo en serie a partir de cierto término o como suele ser de práctica utilizar un artificio matemático que considera una integral doble a efecto de probar la igualdad[3]. Contaremos entonces, aunque aquí no hayamos presentado la demostración, con que la expresión:

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

según las precisiones realizadas, es una función de densidad de probabilidad.

Pero la cuestión sobre la forma de la primitiva de la función de densidad subsiste cuando hay que determinar la función de distribución:

$$F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2} dt$$

de manera que, en principio, se requeriría calcular los valores aproximados de la integral expresados por el desarrollo en serie de la misma. Esto traería aparejado la necesidad de un alto nivel de conocimientos matemáticos por parte de los habituales usuarios de la teoría de las probabilidades y de la estadística, tales como ingenieros, médicos, economistas, sociólogos, etc, lo que resultaría impracticable. Afortunadamente existe un procedimiento alternativo para calcular la probabilidad $F(x) = P(X \leq x)$ cuando la variable aleatoria se distribuye normalmente. Pero antes de explicarlo detallaremos brevemente el significado de los parámetros μ y σ que aparecen en la expresión de la función de densidad.

Propiedad 4: El valor esperado de una variable aleatoria normal es

$$E(X) = \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = \mu$$

[3]La prueba puede verse en Paul Meyer ob. cit., p.188

En efecto, si se realiza el cambio de variables $z = \frac{x-\mu}{\sigma}$ resulta $x = \mu + \sigma z$ y se tiene que $\frac{dz}{dx} = \frac{1}{\sigma}$ de donde $\sigma \cdot dz = dx$. Además cuando x tiende a $+\infty$ y $-\infty$ ocurre que z también lo hace respectivamente. De esta forma $E(X) = \int_{-\infty}^{+\infty} x \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} (\sigma z + \mu) e^{-\frac{z^2}{2}} dz = \frac{1}{\sqrt{2\pi}} \sigma \int_{-\infty}^{+\infty} z e^{-\frac{z^2}{2}} dz + \mu \frac{1}{\sqrt{2\pi}} \sigma \int_{-\infty}^{+\infty} e^{-\frac{z^2}{2}} dz$. Al analizar las dos integrales del último miembro de la cadena de igualdades, se observa que en la primera la función integrando es impar, es decir que se verifica para ella la igualdad $g(z) = -g(z)$ con lo cual la integral se hará 0. La segunda integral es la de una función de densidad normal de parámetros $\mu = 0$ y $\sigma = 1$, por lo tanto su valor será 1. Así resulta entonces $E(X) = \mu$ como se quería demostrar.

Propiedad 5: La varianza de una variable aleatoria normal es

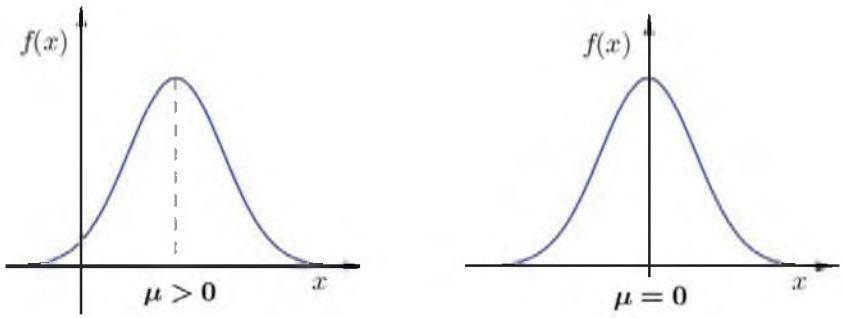
$$Var(X) = \int_{-\infty}^{+\infty} (x - E(X))^2 \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx = \sigma^2$$

Como $Var(X) = E(X^2) - E^2(X)$ calculamos[4] $E(X^2) = \sigma^2 + \mu^2$ así queda $Var(X) = \sigma^2 + \mu^2 - \mu^2 = \sigma^2$ que es lo que quería probarse. Claramente, el desvío estándar es entonces σ.

Ahora analizamos el papel que juegan los parámetros σ y μ en la gráfica de la distribución. Para ello observamos primero que cuando $x = \mu$ se tiene $f(\mu) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{\mu-\mu}{\sigma}\right)^2} = \frac{1}{\sqrt{2\pi}\sigma}$ que es el valor máximo de la función de densidad de probabilidad ya que para cualquier otro valor de la variable aleatoria x resultará $e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2 < 1}$ y por lo tanto $f(x) < f(\mu)$. De esta forma la recta $x = \mu$ resulta el eje de simetría de la campana que pasa por el "pico" de la misma, según se ve a la izquierda en la figura 7:

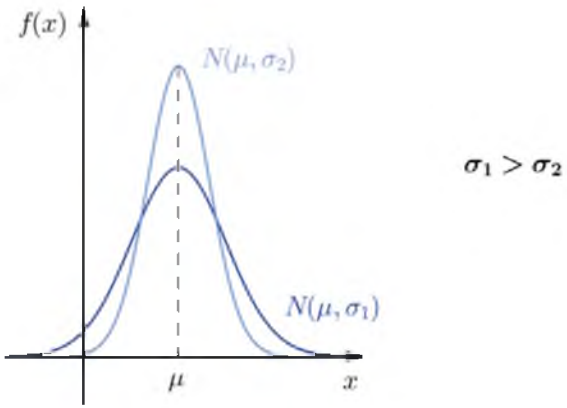
[4]Los detalles de este cálculo pueden verse, por ejemplo, en Paul Meyer, ob. cit., p.190

Figura 7.



También se ve en la misma figura cómo un cambio en el valor de μ se traduce en un desplazamiento de la campana sobre el eje de abscisas. En cuanto al desvío estándar σ ocurre que cuanto más pequeño es, más concentrada está la gráfica alrededor del eje de simetría. Si consideramos dos distribuciones normales con distintos desvíos estándar e igual esperanza se obtienen gráficas como las que se muestran en la figura 8:

Figura 8.



La notación $N(\mu, \sigma)$ se utiliza para referirse a la distribución normal de una variable aleatoria con esperanza μ y desvío estándar σ. Ahora estamos en condiciones de enunciar otro resultado importante[5].

[5]La prueba de la Propiedad 6 puede verse en Paul Meyer, ob. cit. p. 90 y p. 191.

Propiedad 6: Si X es una variable aleatoria distribuída normalmente con esperanza μ y varianza σ^2, la variable aleatoria $Z = aX + b$, donde a y b son constantes reales, se distribuye normalmente con esperanza $a\mu + b$ y varianza $a^2\sigma^2$.

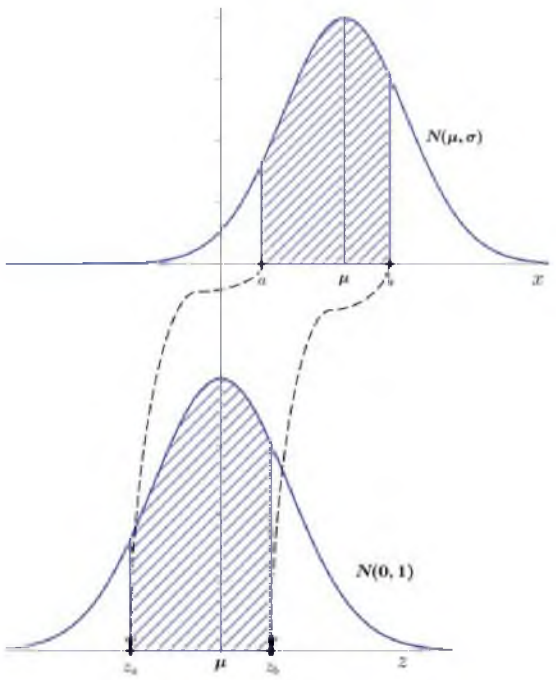
Una consecuencia relevante de esta propiedad surge al hacer $z = \frac{x-\mu}{\sigma}$. Aquí la variable X se transforma multiplicándola por $a = \frac{1}{\sigma}$ y restándole a ese producto la cantidad $b = \frac{-\mu}{\sigma}$. Entonces, según la propiedad, la esperanza resulta $\frac{1}{\sigma}\mu - \frac{\mu}{\sigma} = 0$ y la varianza $\left(\frac{1}{\sigma}\right)^2 \sigma^2 = 1$ por lo que también $\sigma = 1$. Utilizando la notación vista podemos poner entonces que $z = \frac{x-\mu}{\sigma}$ se distribuye $N(0, 1)$. Como veremos enseguida, esto se usará para el cálculo de probabilidades de cualquier variable aleatoria normal.

En efecto, ya hemos referido que la imposibilidad de contar con una expresión "cerrada" para la primitiva de la función de densidad normal, complica el cálculo de la función de distribución en cualquier punto. No tenemos entonces una expresión sencilla para $P(X \leq x)$ cualquiera sea la esperanza y la varianza de x, y se tendrían que hacer los cálculos aproximados en cada oportunidad, con el desarrollo en serie truncado de la primitiva. En vez de esto, que complicaría en gran medida el uso de una herramienta fundamental de modelado, se tabulan los distintos valores de probabilidad de una sola variable normal que llamaremos por ello *estándar* y todos los cálculos de probabilidad que deban realizarse con otras variables normales se remiten a la tabla de esa variable estándar. En general si x es una variable aleatoria normal de esperanza μ y desvío estándar σ, es decir $N(\mu, \sigma)$, se realiza la transformación $z = \frac{x-\mu}{\sigma}$, denominada en adelante *variable normal estandarizada*. Como los valores correspondientes a la función de distribución acumulada $F(z)$ se han calculado por el procedimiento de aproximación apuntado más arriba, estos se utilizan para hallar los correspondientes valores de la función de distribución $F(x)$. Se tiene $F(x) = P(X \leq x) = F(z) = P(Z \leq z)$. Es decir; con una sola tabla cuyos valores se han establecido por un trabajoso procedimiento matemático previo, se puede calcular la probabilidad de cualquier variable aleatoria normal x haciendo $z = \frac{x-\mu}{\sigma}$. Así

$$P(a \leq x \leq b) = P\left(\frac{a-\mu}{\sigma} \leq z \leq \frac{b-\mu}{\sigma}\right)$$

La figura 9 explica el procedimiento. Los puntos a y b se transforman en $z_a = \frac{a-\mu}{\sigma}$; $z_b = \frac{b-\mu}{\sigma}$ y en vez de calcular la probabilidad para la variable aleatoria x, representada por el área sombreada bajo la curva superior, se obtiene, por medio de la tabla, el área sombreada bajo la curva inferior que corresponde a la variable normal estandarizada:

Figura 9.



Los detalles de la tabla y su uso se presentan en el siguiente ejemplo.

Ejemplo 2: *Sabiendo que X es una variable aleatoria con distribución normal con media 30 y varianza 16, calcular usando la tabla:*

a) $P(X < 25)$

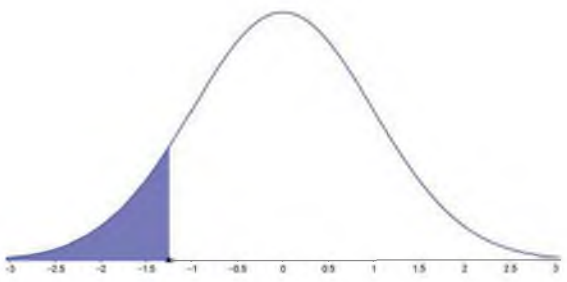
Si una variable aleatoria es normal, la imagen de su función de densidad es una campana en torno a un eje de simetría que pasa precisamente por el valor esperado de la variable.

$\mu = 30$ es en este caso el valor de abscisas perteneciente el eje de simetría de la campana de Gauss.

$\sigma^2 = 16$ varianza de la distribución, mide la variabilidad de la variable. Esto se hace más visible cuando se considera su raíz cuadrada positiva denominada desvío estándar. Cuando el desvío estándar es pequeño la campana se encuentra más concentrada alrededor del eje de simetría, mientras que si el desvío estándar crece, y por ende también lo hace la varianza, la distribución gaussiana se extiende más, en términos geométricos, porque hay mayor dispersión. Como se ha visto la forma analítica de una distribución acumulada, para una variable aleatoria normal, no es una expresión funcional cerrada sino un desarrollo en serie de complicado cálculo. Por tal motivo los cálculos de probabilidad acumulada se realizan utilizando una tabla. Es claro que no puede disponerse de una tabla para cada normal caracterizada por su esperanza y su desvío estándar, pues ambos valores varían en un rango infinito. Por esta razón se transforma linealmente la variable original x utilizando el parámetro μ para cambiarla de origen, centrándola en 0, y el parámetro σ para cambiarla de escala. Así se tiene la variable "estándar" $z = \frac{x-\mu}{\sigma}$ que tiene esperanza 0, desvío estándar 1 y se distribuye normalmente. z se denomina "estándar" porque a ella hay que remitir todas las variables normales para calcular probabilidades, usando una única tabla de la distribución normal.

$$P(X < 25) = P(z < \frac{25 - 30}{40}) = P(z < -1{,}25) = F(-1{,}25)$$

Se está buscando el área sombreada, representada a la izquierda de la gráfica:

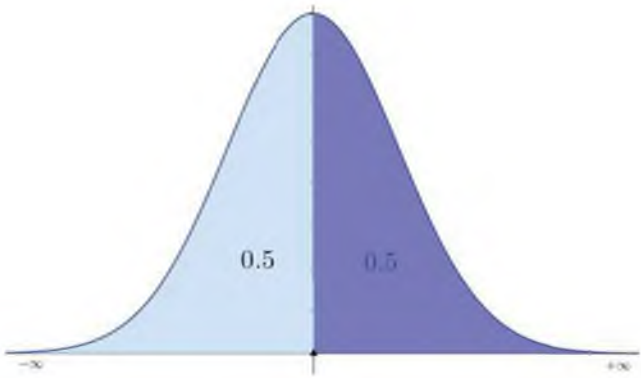


Ahora bien; la tabla está construida acumulando la probabilidad desde 0 hasta un valor de z positivo como lo muestra la zona sombreada en la tabla del Anexo A:

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147

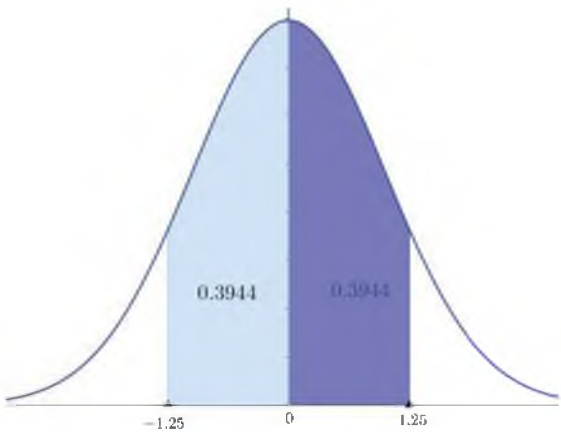
Surge de la tabla, que el valor acumulado de probabilidad entre 0 y 1,25 es entonces $P(0 \leq z < 1{,}25) = F(1{,}25) - F(0) = 0{,}3944$

Pero la curva es simétrica y la probabilidad acumulada desde $-\infty$ hasta 0 es la misma que la acumulada desde 0 hasta $+\infty$. Por lo tanto ambas deben valer 0,5 como se ve en la siguiente figura:



Además, por la misma simetría debe ocurrir que la cola a izquierda, entre $-\infty$ y $-1{,}25$, contenga en total la misma probabilidad que contiene la cola a derecha entre 1,25 y $+\infty$.

Situación que se puede observar claramente en el siguiente gráfico:



Entonces se tiene:

$$P(z < -1{,}25) = P(z > 1{,}25) = 0{,}5 - P(0 \leq z < 1{,}25) =$$

$$= 0{,}5 - 0{,}3944 = 0{,}1066$$

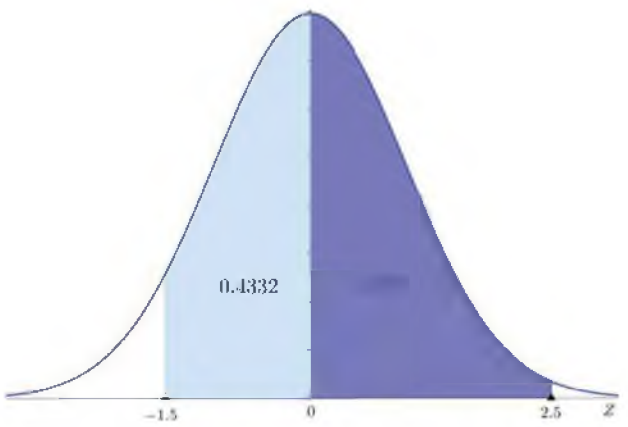
b) $P(24 < x < 40)$

La cadena de igualdades para remitir este cálculo a uno que pueda realizarse por la tabla es:

$$P(24 < x < 40) = P(\frac{24 - 30}{4} < z < \frac{40 - 30}{4}) =$$

$$= P(-1{,}25 < z < 2{,}5)$$

Situación que podemos representar mediante el siguiente gráfico:



Por simetría entonces y a partir de la tabla, se puede encontrar la probabilidad buscada, representada por el área sombreada más suave, de acuerdo a:

$$P(-1{,}25 < z < 2{,}5) = P(-1{,}25 < z \leq 0) + P(0 \leq z < 2{,}5) =$$

$$= P(0 \leq z < 1{,}25) + P(0 \leq z < 2{,}5)$$

Es decir:

$$P(24 < x < 40) = 0{,}4332 + 0{,}4838 = 0{,}927$$

c) $P(x > 32)$

$$P(x > 32) = P(z > \frac{32 - 30}{4}) = P(z > 0{,}5) =$$
$$= 0{,}5 - P(0 \leq z < 0{,}5) = 0{,}5 - 0{,}1915 = 0{,}3085$$

3.4. Ejercicios

Ejercicio N°1*: Dada $f(x) = 2x$ si $0 \leq x \leq 1$ y 0 en otra parte.

a) Probar que es una función de densidad de probabilidad.

b) Construir la distribución acumulada.

c) Hallar $P(x < 0{,}5)$.

Ejercicio N°2: Sea x una variable aleatoria continua cuya función de densidad de probabilidad es $f(x) = \frac{1}{6}x + k$ en $0 \leq x \leq 3$ en otra parte.

a) Calcular k

b) Hallar $P(1 \leq x \leq 2)$

Ejercicio N°3*: Sea x una variable aleatoria continua cuya distribución es uniforme en un dado intervalo $[10, 20]$ y 0 en otra parte:

a) Determinar k.

b) Hallar $E(X)$.

c) Hallar la función de probabilidad acumulada $F(x)$.

Ejercicio N°4: Sea la variable aleatoria x tal que $P(x = 0) = \frac{1}{2}$ y que si $x \neq 0$ se distribuye uniformemente entre -5 y 15

a) Hallar su función de densidad de probabilidad.

b) Calcular la probabilidad acumulada hasta $x = 10$.

Ejercicio N°5*: El tiempo que una persona pasa dentro de un cajero automático se distribuye en forma exponencial y se sabe que su valor esperado es de 4 minutos. Calcular:

a) La probabilidad de que una persona permanezca en el cajero solo 1 minuto.

b) La probabilidad de que tarde más de 5 minutos.

Ejercicio N°6*: La esperanza y la desviación estándar de las notas de un examen calificado al décimo son 7,4 y 1,2 respectivamente.

a) Hallar los resultados en unidades estándar de los estudiantes que recibieron notas: *i)* 6,5 *ii)* 7,4 *iii)* 8,6 *iv)* 9,2

b) Hallar las notas que corresponden a los resultados estándar: *i)* -1 *ii)* 1,25 *iii)* 0,5 *iv)* 1,75

Ejercicio N°7*: Sea z una variable aleatoria con distribución normal estándar. Utilizar la tabla conveniente para calcular:

a) $P(0 \leq x \leq 1{,}42)$

b) $P(-0{,}73 \leq x \leq 0)$

c) $P(-1{,}37 \leq x \leq 2{,}01)$

d) $P(0{,}65 \leq x \leq 1{,}26)$

e) $P(-1{,}79 \leq x \leq -0{,}54)$

f) $P(x \geq 1{,}13)$

g) $P(|x| \leq 0{,}5)$

Ejercicio N°8*: Sabiendo que z es una variable aleatoria normal estándar, calcular el valor de la constante a, usando la tabla:

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

b) $P(z > a) = 0{,}8$

c) $P(z < a) = 0{,}1$

d) $P(z > a) = 0{,}25$

e) $P(|z| < a) = 0{,}25$

f) $P(|z| > a) = 0{,}01$

Ejercicio N°9: Se ha estudiado que el tiempo de espacio dedicado a publicidad en TV sigue una distribución normal con promedio de 40 segundos y desvío de 5 segundos. ¿Cuál es la probabilidad de que una publicidad elegida al azar dure:

a) más de 43 segundos?

b) menos de 35 segundos?

c) entre 37 y 42 segundos?

d) más de un minuto?

Ejercicio N°10*: Supóngase que las estaturas de 800 estudiantes están normalmente distribuidas con una media de 66 pulgadas y un desvío estándar de 5 pulgadas. Hallar entonces el número de estudiantes con estaturas:

a) entre 65 y 70 pulgadas.

b) mayor o igual a 72 pulgadas.

Capítulo 4

Complementos teóricos

4.1. Desigualdad de Chebyshev

Siempre que se conoce la distribución de probabilidades de una variable aleatoria es posible calcular el valor esperado y la varianza de la misma. Sin embargo, esto no vale al revés. Es decir; conocer esperanza y desvío estándar no basta para poder reconstruir la distribución de probabilidad. En este caso puede surgir el problema de calcular $P(|X - E(X)| \leq C)$. Tal cálculo intenta medir la probabilidad de que los valores de la variable aleatoria se encuentren a una distancia del valor esperado menor o igual que C y para realizarlo se requiere contar con la distribución de probabilidad. Al desconocerse la misma, la desigualdad de Chebyshev al menos proporciona una cota a la probabilidad buscada. En particular suele ser útil, en muchas aplicaciones, tomar un múltiplo del desvío estándar haciendo $C = k\sigma$.

Desigualdad de Chebyshev[1]: $P(|X - E(X)| \geq k\sigma) \leq \frac{1}{k^2}$

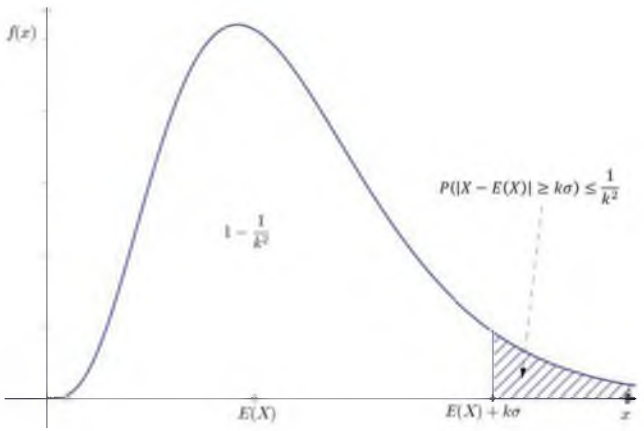
Esta versión de la desigualdad, que puede expresarse en realidad de distintas maneras, pone una cota a la probabilidad que los valores de la variable aleatoria se encuentren más allá de k desvíos estándar del valor esperado. En la figura 1 se aprecia el intervalo considerado y el área bajo la curva que representa la probabilidad.

De forma complementaria resulta $P(|X - E(X)| \leq k\sigma) \geq 1 - \frac{1}{k^2}$ que está representada por el área sin sombrear bajo la curva. En cual-

[1]Una demostración de la desigualdad puede verse en Paul Meyer, ob. cit., pp. 146-147.

quier caso lo importante es que, sin conocer exactamente cual es la distribución de la probabilidad, puede establecerse una cota para las cantidades apuntadas si se tienen esperanza y desvío estándar. El número k razonablemente tiene que cumplir $k > 1$ para que la desigualdad tenga sentido.

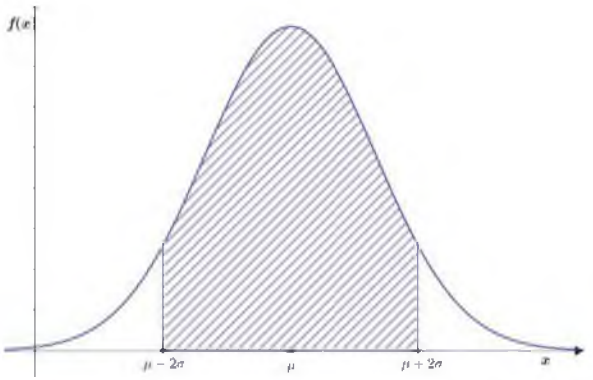
Figura 1.



Es interesante analizar el caso de la distribución normal cuando se toma, por ejemplo, $k = 2$. Dados la esperanza μ y el desvío estándar σ puede querer averiguarse la probabilidad del intervalo $[\mu - 2\sigma, \mu + 2\sigma]$. Este intervalo está constituido por todos los puntos x que satisfacen $\mu - 2\sigma \leq x \leq \mu + 2\sigma$ lo que puede escribirse también $|x - \mu| \leq 2\sigma$. Por la desigualdad de Chebyshev se debe cumplir que $P(|x - \mu| \leq 2\sigma) \geq 1 - \frac{1}{2^2} = \frac{3}{4} = 0{,}75$. Sin embargo, como aquí sabemos que la variable aleatoria se distribuye normalmente, podemos calcular la probabilidad utilizando la tabla. En efecto, al estandarizar se aplica la transformación $z = \frac{x-\mu}{\sigma}$ y se tiene $-2 \leq z \leq 2$ y con la tabla calculamos entonces:

$$P(|x - \mu| \leq 2\sigma) = P(|z| \leq 2) = 2 \times 0{,}4772 = 0{,}9544$$

La probabilidad obtenida corresponde al área sombreada en la figura 2.



Obsérvese que la desigualdad de Chebyshev proporciona, en este caso, una cota inferior para la probabilidad acumulada sin que sea necesaria la expresión de la distribución para calcularla. Cuando ésta se tiene en cuenta, la probabilidad para el intervalo considerado resulta mayor. En otras palabras, Chebyshev permite conocer el "piso" del valor de probabilidad acumulada teniendo sólo la esperanza y la varianza.

4.2. Ley de los grandes números

Hemos visto al comienzo que una de las formas de llegar a establecer la probabilidad de un resultado, de entre los que puede arrojar una experiencia, consiste en realizarla un gran número de veces hasta que la proporción de veces que aparece el suceso buscado se estabiliza y entonces asignar esa proporción como su probabilidad. Este camino empírico da un valor "a posteriori" de las experiencias pero, como también hemos analizado, la probabilidad definida matemáticamente, es decir en forma axiomática, no requiere experiencia previa y solo basta con que se cumplan tres propiedades para definirla. Supongamos que ese fuera el caso, es decir que se cumplen las propiedades matemáticas exigidas sin importar de que manera nos han sido sugeridas, si en una forma clásica, experimental o subjetiva. ¿Si se realiza la experiencia un número suficiente de veces, la frecuencia de veces que aparece el resultado buscado convergerá a su probabilidad definida axiomáticamente?

La respuesta es sí, como estaba fuertemente sugerido, y tal propiedad se expresa como *Ley de los grandes números*[2].

Ley de los grandes números (*o de Bernoulli*[3]): Sea un experimento y un suceso A uno de sus resultados posibles. Sean n repeticiones independientes del experimento, n_A el número de veces en que el resultado es A y la razón o frecuencia relativa de A, $f_A = \frac{n_A}{n}$. Sea p la probabilidad que suceda A. Para cualquier número $\epsilon > 0$ se tiene que:

$$P\left(|f_A - p| \geq \epsilon\right) \leq \frac{p(1-p)}{n\epsilon^2}$$

Obsérvese que si n es suficientemente grande, es decir si $n \to \infty$, el segundo miembro de la desigualdad tiende a 0 y puede escribirse $\lim_{n\to\infty} P\left(|f_A - p| \geq \epsilon\right) = 0$. Complementariamente resulta entonces,

$$\lim_{n\to\infty} P\left(|f_A - p| < \epsilon\right) = 1$$

Esto explicita que la probabilidad de que la frecuencia relativa observada del suceso converja a su probabilidad es 1. Es lo que se denomina *convergencia en probabilidad*. No debe pasar inadvertido, además, que el suceso A se presenta con probabilidad p o en forma complementaria no se presenta, con probabilidad $1-p$. Hay un binomio de posibilidades y como cada repetición del experimento es independiente de cualquier otra se está en la situación estudiada para la probabilidad binomial.

Como la Ley de los grandes números efectivamente se cumple, puede utilizársela para realizar una aproximación de la distribución binomial por medio de la normal[4]. Obviando la demostración matemática desarrollada por el ya citado De Moivre en 1733, basta tomar como valor esperado de la normal aproximante al de la variable binomial, es decir, $\mu = np$ y como desvío estándar a $\sigma = \sqrt{np(1-p)}$.

En la práctica ocurrirá que si el valor de p está próximo a 0.5 bastará que el número de ensayos n no sea muy pequeño para que la aproximación resulte buena. En la medida en que p sea más próximo a 0 o a 1, se requerirá un número n mayor de ensayos. En la figura 2

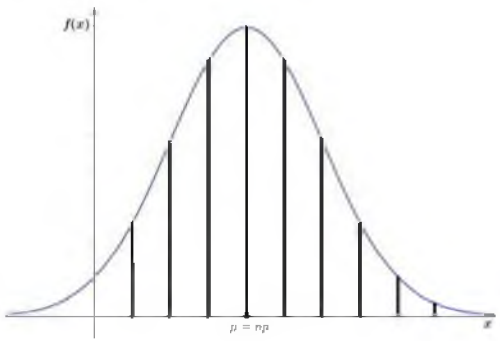
[2]Un comentario más extenso sobre las implicancias de este resultado se encuentra en Pablo Jacovskis y Roberto Perazzo, o. cit., pp.39 a 49.

[3]La prueba puede verse en Paul Meyer, ob.cit., p. 253.

[4]La fórmula de aproximación se encuentra desarrollada por ejemplo en Harald Cramér: *Teoría de probabilidades y aplicaciones*, Aguilar, pp. 102-108.

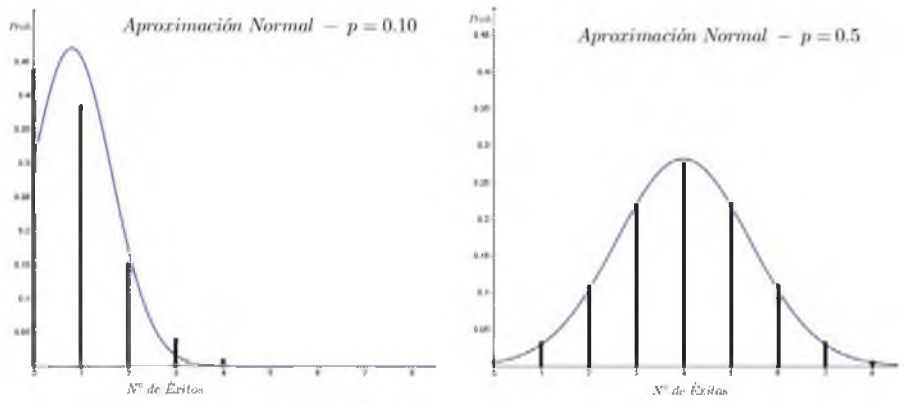
se muestra la distribución binomial representada por los bastones y la curva envolvente que representa a la normal aproximante.

Figura 2.



Si n no es muy grande puede ser necesario realizar una corrección de media unidad de forma tal que la probabilidad binomial acumulada entre dos valores enteros positivos, por ejemplo k y l sea aproximada por la probabilidad normal acumulada en el intervalo $[k - 0{,}5; l + 0{,}5]$. La figura 3 muestra como la aproximación resulta mejor cuanto más se acerca la probabilidad de éxito p a 0.5:

Figura 3.



A la izquierda se observa la distribución de probabilidad binomial para $n = 8$ y $p = 0{,}1$. La normal superpuesta tiene igual esperanza y

desvío estándar que la variable binomial: $\mu = np = 8 \times 0{,}1 = 0{,}8$ y $\sigma\sqrt{npq} = \sqrt{8 \times 0{,}1 \times 0{,}9} = 0{,}8485$. Como se puede apreciar, la aproximación de la probabilidad binomial por la normal, aún con la corrección por continuidad que aproxima la probabilidad en cada valor k de la variable discreta por la del intervalo $[k - 0{,}5, k + 05]$ correspondiente a la variable normal continua, no resulta adecuada. Por ejemplo:

$$P_{binomial}(k = 1) = \binom{8}{1} (0{,}1)^1 (0{,}9)^{8-1} = 0{,}3826$$

no resulta bien aproximada por:

$$P_{normal}\,[1 - 0{,}5 \leq x \leq 1 + 0{,}5] = P(0{,}5 \leq x \leq 1{,}5) =$$

$$= P(-0{,}35 \leq z \leq 0{,}82) = 0{,}4307$$

En cambio en la figura de la derecha la gráfica de la distribución binomial con $n = 8$ y $p = 0{,}5$ está superpuesta con la de la normal aproximante de esperanza $\mu = np = 8 \times 0{,}5 = 4$ y desvío estándar $\sigma\sqrt{npq} = \sqrt{8 \times 0{,}5 \times 0{,}5} = 1{,}4142$. En este caso se obtiene:

$$P_{binomial}(k = 1) = \binom{8}{1} = (0{,}5)^1 (0{,}5)^{8-1} = 0{,}0312$$

$$P_{normal}\,[1 - 0{,}5 \leq x \leq 1 + 0{,}5] = P(0{,}5 \leq x \leq 1{,}5) =$$
$$= P(-2{,}47 \leq z \leq -1{,}77) = 0{,}0316$$

Como se ve, aquí la aproximación es muy buena pues el error cometido se evidencia recién en el cuarto dígito decimal y es por ello menor que 0.001. En la medida que la probabilidad de éxito p esté próxima a 0.5 la normal permitirá calcular sin demasiado error la probabilidad binomial. Si el valor de p se aleja hacia 0 ó 1 la aproximación normal producirá un error mayor aunque, por supuesto, si el número n de ensayos se acrecienta p podrá acercarse más a 0 ó a 1 sin que el error de aproximación resulte muy grande.

4.3. Teorema del límite central

Si tenemos un binomio de posibilidades que denominamos éxito y fracaso, cada vez que se realiza una experiencia, si se presenta éxito podemos poner una variable en 1 y si se presenta fracaso colocarla en 0. Así la variable aleatoria X_i que representa al resultado de la experiencia realizada por *i-ésima* vez, será:

$X_i = 1$ si se produce un éxito en la *i-ésima* repetición de la experiencia.

$X_i = 0$ si se produce un fracaso en la *i-ésima* repetición de la experiencia.

El número X de éxitos que se produce cuando se repite n veces la experiencia es una variable aleatoria binomial y puede pensarse entonces como la suma de n variables aleatorias independientes $X = X_1 + X_2 + \ldots + X_n$. Como ya hemos visto una variable aleatoria de tales características puede aproximarse por la distribución normal. Al generalizar esta idea, se da lugar a un resultado importante.

Teorema del límite central[5]: Si $X_1, X_2, \ldots, X_n, \ldots$ es una sucesión de variables aleatorias independientes con valor esperado $E(X_i) = \mu_i$ y varianza $Var(X_i) = \sigma_i^2$, la distribución de la variable aleatoria $X = X_1 + X_2 + \ldots + X_n$ converge a una distribución normal de esperanza:

$$E(X) = \sum_{i=1}^{n} \mu_i = \mu$$

y desvío estándar:

$$\sigma = \sqrt{\sum_{i=1}^{n} \sigma_i^2}$$

cuando $n \to \infty$.

Es decir, si el número de variables sumadas es suficientemente grande, su suma tiene una distribución muy próxima a la normal con la esperanza y desvío estándar indicados. Obsérvese que, según el teorema, esto ocurre cualquiera sea la distribución de cada una de las

[5]Si bien la prueba general de este teorema no es sencilla, una versión con ciertas hipótesis simplificadas puede verse esbozada en Paul Meyer, ob.cit., p. 260

variables X_i y aunque todas las distribuciones de probabilidad de estos sumando fueran distintas entre si. Esta es otra de las causas de la importancia de la distribución normal en la teoría de las probabilidades pues en muchas aplicaciones la variable aleatoria involucrada puede representarse como una suma de variables aleatorias y por ende su distribución puede aproximarse por una normal. El teorema del límite central nos permitirá establecer las bases de la estadística inferencial que estudiaremos más adelante.

4.4. Ejercicios

Ejercicio N°1*: Dada una distribución de probabilidad cualquiera, ¿cuál es la máxima probabilidad de que la variable aleatoria tome valores a más de 1.8 desviaciones estándar del valor esperado?

Ejercicio N°2*: Utilizar la tabla de la distribución normal estándar para evaluar $P(-3 \leq z \leq 3)$. ¿Contradice el resultado obtenido la previsión teórica aportada por la desigualdad de Chebyshev ? Justificar la respuesta.

Ejercicio N°3*: Una moneda corriente se lanza 12 veces. Determinar la probabilidad de que el número de caras que salgan estén entre 4 y 7 inclusive por medio de:

a) la distribución binomial

b) la aproximación normal a la distribución binomial

Ejercicio N°4*: Hallar la probabilidad de que entre 10000 dígitos tomados al azar, el dígito 3 aparezca 950 veces a lo sumo.

Capítulo 5

Estadística descriptiva

5.1. Introducción

La estadística descriptiva se ocupa de establecer las tendencias principales de las poblaciones a partir de medir o calificar todos o solamente algunos de los atributos que las caracterizan. Para empezar hay que decir que el término *población* se refiere a un universo o conjunto total de individuos que satisfacen una propiedad. Si los individuos son personas una propiedad puede ser, por ejemplo, estudiar en la Universidad. Así tenemos la población de estudiantes de una Universidad. Pero no es necesario que se trate de personas para que tengamos una población en sentido estadístico. También podemos hablar de poblaciones de microbios, de tornillos, de billetes o de programas de computadora, por ejemplo, reunidas en cada caso por alguna propiedad común a todos sus integrantes.

Consideremos el caso de la población de estudiantes de una Universidad. Cada estudiante tiene atributos como apellido, edad, altura, peso, nota de probabilidad y estadística, color de ojos, bebida preferida etc. Un atributo es lo que se denomina también una *variable*, porque puede tomar valores diferentes o representar casos distintos. Por ejemplo; el estudiante se llama Gómez, que es un caso de apellido, mide 1.77 que es un valor posible de altura, pesa 75 kg, cantidad factible de peso y prefiere las gaseosas, categoría de bebida preferida. Si bien existen varias formas para clasificar los tipos de variables, se ve a simple vista que hay esencialmente dos: el tipo *cuantitativo* que mide o cuenta precisamente una cantidad, como el peso o la altura, y

el tipo *cualitativo* que expresa una cualidad o categoría como apellido o bebida preferida, es decir no mide ni cuenta nada sino que nombra o clasifica. Las variables numéricas pueden a su vez dividirse entre las que adoptan valores *discretos*, por lo general números enteros, y las que toman valores reales que se denominan *continuas*. Así la figura 1 resume esta clasificación.

Cuando se analiza una sola variable la estadística se denomina *univariada*, mientra que cuando se consideran dos o más se trata de estadística *multivariada*. Por ejemplo, los *censos* de habitantes, que se realizan cada diez años, son estadísticas multivariadas sobre toda la población. En nuestro caso estudiaremos estadística univariada, es decir que nos ceñiremos a los registros obtenidos para una sola variable, e intentaremos con ello *describir* el comportamiento de una determinada población respecto de esa variable. Es importante mencionar que el objetivo general de hacer una estadística suele ser el de tomar ulteriores decisiones. La estadística es una herramienta poderosa que ayuda a tomar decisiones adecuadas. Veremos también que es importante al respecto contar con una visualización de los resultados fidedigna pues, en ocasiones, puede malinterpretarse un resultado estadístico si no se muestra adecuadamente.

Figura 1.

5.2. Distribuciones de frecuencias

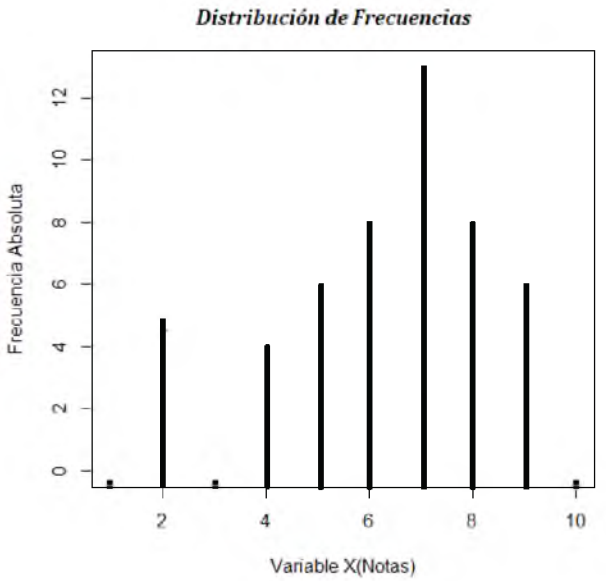
Los datos de una variable recogidos para todos los individuos de una población son eso, simplemente datos. Para que se conviertan en información útil, en conocimiento acerca de esa población, deben procesarse adecuadamente. El primer paso de éste proceso es agruparlos en las llamadas *distribuciones de frecuencias*. Por ejemplo: si las notas de un grupo de 50 estudiantes son 7 7 2 4 8 9 7 6 5 9 2 7 7 5 6 9 8 7 6 6 5 7 7 7 8 8 5 4 8 2 2 9 7 6 6 8 4 2 9 7 7 8 8 6 4 7 6 5 5 9, se puede construir la distribución de frecuencias que se muestra en la tabla 1:

Tabla 1.

Variable X	F	f	$\sum F$	$\sum f$
1	0	0/50	0	0/50
2	5	5/50	5	5/50
3	0	0/50	5	5/50
4	4	4/50	9	9/50
5	6	6/50	15	15/50
6	8	8/50	23	23/50
7	13	13/50	36	36/50
8	8	8/50	44	44/50
9	6	6/50	50	1
10	0	0/50	50	1
$\sum$	50	1		

X es la variable "nota", F es la *frecuencia absoluta* de cada nota, es decir la cantidad de veces que la nota se presentó y f su *frecuencia relativa* al total de estudiantes. $\sum F$ representa la *frecuencia absoluta acumulada* y $\sum f$ la *frecuencia relativa acumulada*. Así el i-ésimo caso de la variable X, denotado X_i, tiene frecuencia absoluta F_i y relativa f_i. La figura 2 muestra la *distribución de frecuencias absolutas*:

Figura 2.



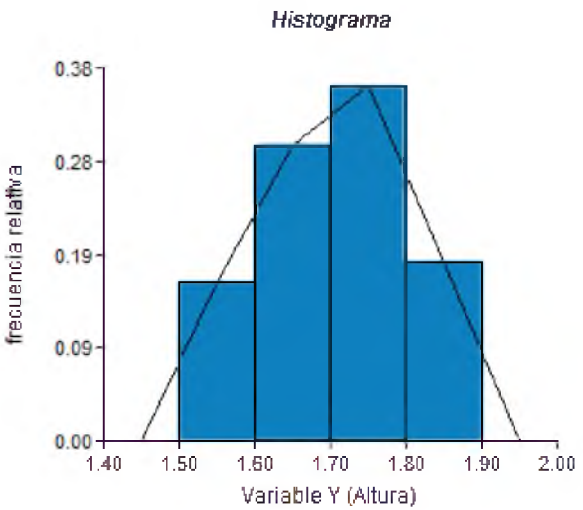
No siempre una variable numérica es discreta. Si la variable es continua se pueden utilizar *intervalos de clases* para establecer la distribución. Así por ejemplo, si se consideran las alturas de los 50 alumnos la distribución es la de la tabla 2:

Tabla 2.

Variable Y (Altura)	F	f	$\sum F$	$\sum f$
[1,50; 1,60)	8	8/50	8	8/50
[1,60; 1,70)	15	15/50	23	23/50
[1,70; 1,80)	18	18/50	41	41/50
[1,80; 1,90)	9	9/50	50	1

Aquí los valores de altura se han agrupado por intervalos. El paréntesis que cierra el extremo superior de los mismos indica que, dada la continuidad de la variable, en realidad el valor que corresponde al límite superior no pertenece a ese intervalo. Esto es así para que, si por ejemplo, un estudiante mide 1.60 se lo considere solamente en el intervalo [1,60; 1,70). Obsérvese que al agrupar en clases en cierto sentido se oculta o pierde información pues sabemos, por ejemplo, que hay 8 estudiantes con alturas dentro del intervalo [1,60; 1,70) pero no tenemos a la vista el dato de sus alturas individuales. Con una variable continua así agrupada puede realizarse un *histograma* como se ve en la figura 3:

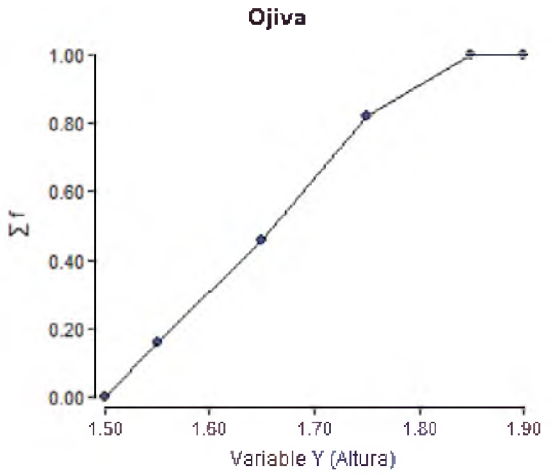
Figura 3.



Considerando que la variable es continua y estableciendo el punto medio de cada intervalo en la figura 3 se ha superpuesto el polígono

de frecuencias. También puede graficarse, con las frecuencias relativas acumuladas, la *ojiva* de la distribución según muestra la figura 4:

Figura 4.



Podemos considerar también variables cualitativas. Supongamos por ejemplo que se sirve una bebida a elección entre café, gaseosa o agua a los 30 pasajeros de un avión con la siguiente distribución de frecuencias dada por la tabla 3:

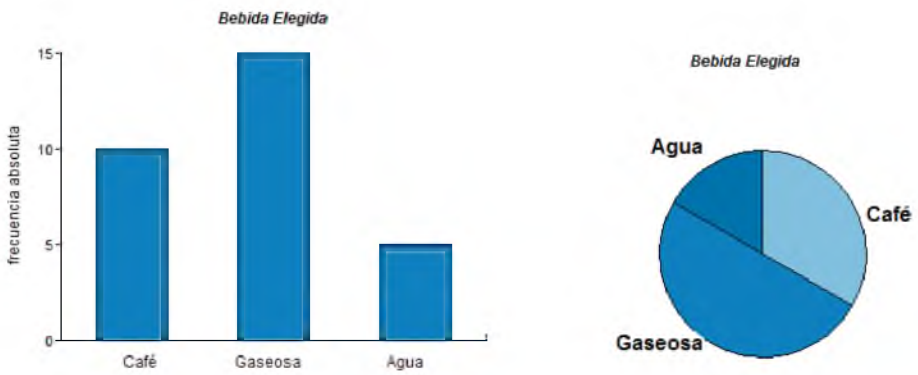
Tabla 3

Bebidas	F	f
Café	10	10/50
Gaseosa	15	15/50
Agua	5	5/50

Se puede realizar un *diagrama de barras* o un *gráfico de pastel* como se muestra en las figuras 5-a y 5-b respectivamente.

Figura 5.

a) b)



Cualquiera de los ejemplos vistos implica que se ha medido o se ha clasificado una característica representada por una variable. Esto es; se ha realizado una experiencia. Si con base en ella se quisiera evaluar la probabilidad de que un pasajero del avión prefiriera café o que un estudiante midiera entre 1.60 y 1.70, se podría apelar a las distribuciones de las frecuencias relativas respectivas para tratar de averiguarlo. Claro que la experiencia estuvo limitada a los 30 pasajeros de un solo avión o sólo a los 50 estudiantes considerados, que constituyeron en esos casos toda la población. Sin embargo, bajo ciertas condiciones que más adelante veremos, las distribuciones de frecuencias relativas así obtenidas pueden indicarnos mucho sobre la probabilidad de sucesos referidos a poblaciones mas generales.

Ejemplo 1: *La siguiente lista revela el número de interrupciones diarias que sufre un proceso de fabricación por distintas causas.*

Interrupciones diarias	Frecuencia
0	3
1	5
2	9
3	15
4	10
5	6
6	2

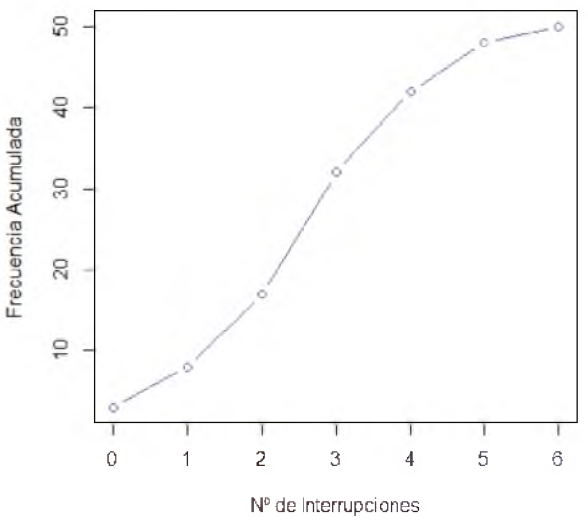
a) Calcular frecuencias relativas y acumuladas.

Las frecuencias relativas f son el cociente entre las frecuencias absolutas F y el total de datos. Mientras que las frecuencias acumuladas $\sum F$ se obtiene realizando las sumas parciales de las frecuencias:

Interrupciones diarias	F	f	$\sum F$	$\sum f$
0	3	3/50	3	3/50
1	5	5/50	8	8/50
2	9	9/50	17	17/50
3	15	15/50	32	32/50
4	10	10/50	42	42/50
5	6	6/50	48	48/50
6	2	2/50	50	1

b) Dibujar la ojiva de la distribución.

Es la representación de la frecuencia acumulada:



5.3. Medidas de tendencia central

Una vez que los datos están estructurados en una distribución de frecuencias, el paso siguiente consiste en estudiar las *tendencias* que evidencia la variable. Nos interesan fundamentalmente tres: la *tendencia central*, la *variabilidad* y la *asimetría*. Comenzaremos entonces por las medidas de la tendencia central.

La tendencia central es la tendencia del grueso de los datos. Expresa los valores o casos de la variable que se ubican centralmente en la distribución. Hay varias formas de medirla y puede ocurrir que todas esas maneras de cálculo no arrojen puntualmente iguales resultados. Estudiaremos sólo las tres medidas que son más utilizadas.

La primera de ellas es la llamada *media aritmética*, o simplemente *media*, y se calcula como el promedio de los datos. Como cada dato x_i según la distribución de frecuencias se presenta F_i veces, se tienen en total $n = \sum_i F_i$ datos y teniendo en cuenta que la frecuencia relativa es $f_i = \frac{F_i}{N}$, la media aritmética puede calcularse por cualquiera de las dos fórmulas que aparecen en la siguiente cadena de igualdades: $m = \frac{\sum_i x_i F_i}{N} = \sum_i x_i f_i$

Como se ve, la última fórmula es análoga a la utilizada para el

cálculo teórico de de la esperanza matemática de una variable aleatoria. En este caso la diferencia consiste en que las probabilidades son reemplazadas por las frecuencias relativas calculadas no a partir de un concepto teórico, como la probabilidad, sino utilizando los datos de la experiencia. Claro que la ley de los grandes números autoriza a pensar que la frecuencia relativa representa a la probabilidad y este hecho justifica que a veces se nombre a la esperanza matemática directamente como media.

Cabe además realizar una aclaración complementaria respecto del cálculo de la media. Cuando la variable es continua y se utilizan intervalos de clase en la distribución de frecuencias, se toman las marcas de clase como representantes de las mismas y con esos valores x_i se calcula la media según las fórmulas dadas.

La media aritmética tiene propiedades de linealidad similares a las de la esperanza matemática pero además resulta de sencillo cálculo. Sin embargo, como medida de tendencia central, tiene dos inconvenientes. El primero de ellos es que si la variable es cualitativa no puede aplicarse. En nuestro ejemplo de la bebida elegida por los pasajeros del avión ¿cómo se calcularía y que querría significar un promedio de bebidas sino una mezcla desagradable que no fuera ninguno de los casos posibles de la variable? Un segundo problema que se presenta al calcular la media aritmética es que resulta muy sensible a valores extremos. Si, por ejemplo una clase tiene 5 estudiantes y sus notas de examen son 10, 10, 10, 10 y 1 respectivamente, está claro que la tendencia central es la de un curso excelente con estudiantes de 10, donde además hay uno que no estudia nada. La media no refleja bien la tendencia pues en este caso $m = \frac{10\times4+1\times1}{4} = 8{,}2$, nota que corresponde a un curso sin dudas bueno pero no sobresaliente. La media se ha visto afectada aquí por la lejanía de la nota 1 respecto del grueso de los datos y no expresa con exactitud la situación.

Otra medida de tendencia central a menudo utilizada es la *mediana* que es el dato ubicado en la posición central una vez que todos los datos han sido ordenados de menor a mayor. Supongamos por ejemplo la serie ordenada de datos 1, 1, 3, 5, 7, 7, 9. El valor central $me = 5$ es la mediana. Si la cantidad de datos fuera par se toma como mediana el promedio de los dos centrales. Obsérvese que la mediana es de fácil cálculo siempre que los datos no estén agrupados en intervalos de clase. Si esto fuera así habría que usar una fórmula interpolatoria para calcularla, cuya presentación no se hará en este caso. La mediana

tiene sin embargo una ventaja frente a la media: no es *sensible* a los extremos. Si volvemos a considerar nuestro ejemplo del curso con 5 estudiantes de notas 10, 10, 10, 10 y 1, ordenando los datos de menor a mayor tenemos: 1 10 10 10 10. El dato ubicado en la posición tercera es la mediana que, en este caso, representa la tendencia central mejor que la media aritmética.

En una playa, un vendedor ha vendido 50 gorros rojos, 10 azules y 5 verdes. Están de moda los gorros rojos. Precisamente la *moda* o *modo* es la medida de tendencia central que se calcula como el caso o valor de la variable que tenga mayor frecuencia absoluta. Aquí entonces $mo = rojo$. Como se ve, el modo puede aplicarse para evaluar la tendencia central cuando la variable es cualitativa y si el vendedor hubiera vendido también 50 gorros amarillos la distribución tendría dos modos. Hay en general distribuciones unimodales, bimodales, trimodales, multimodales. Nada impide además aplicar el modo a variables cuantitativas. En nuestro ejemplo de las notas de los 5 estudiantes el modo sería $mo = 10$, que es la nota de mayor frecuencia y también estaría revelando la tendencia central en forma adecuada.

Ejemplo 2: *Utilizar los datos del Ejemplo 1 para evaluar media, mediana y modo*

Media aritmética:

$$m = \frac{\sum_i x_i F_i}{N} =$$

$$= \frac{0 \times 3 + 1 \times 5 + 2 \times 9 + 3 \times 15 + 4 \times 10 + 5 \times 6 + 6 \times 2}{50}$$

$$m = \frac{150}{50} = 3$$

Mediana: En un conjunto de n elementos ordenados (creciente o decrecientemente), la mediana es el valor que se encuentra en el centro. Es decir, divide al conjunto en dos partes iguales, de manera tal que el 50 % de los datos es mayor o igual que la mediana y el 50 % restante es menor o igual a esta. Si n es impar, entonces la mediana se encuentra

en la posición central, que es exactamente la posición que separa los datos en dos grupos de igual cantidad. Si n es par, entonces la mediana será el promedio entre los elementos que ocupan la posición $\frac{n}{2}$ y la posición $\frac{n}{2}+1$.

En este ejemplo podríamos extraer, de la tabla, todos los datos y ordenarlos crecientemente:

00011111222222222333333333333333444444444455555566

Son 50 datos (cantidad par). Por lo dicho anteriormente, se debe calcular el promedio de los datos que ocupan la posición 25 y 26. Es decir:

$$me = \frac{3+3}{2} = 3$$

Sin embargo, podría ahorrarse esta ordenación aprovechando la distribución de frecuencias de la tabla. En la última columna, se tienen las frecuencias acumuladas. Si necesitamos encontrar el elemento que ocupa la posición central, bastará con buscar el primer valor de x, para el cual la frecuencia acumulada, verifica:

$$\sum F \geq \frac{n}{2}$$

es decir,

$$\sum F \geq \frac{50}{2} = 25 \Rightarrow me = 3$$

Modo: La moda o modo se define como el valor que más se repite en un conjunto de datos, o sea, el valor que ocurre con más frecuencia. Puede suceder que no se presente un único valor modal; por ejemplo, podría ser bimodal, si tiene dos modas, multimodal, si tiene más de tres y podría no existir el modo si todos los valores se presentan sólo una vez. En este caso, hay sólo un valor que se repite 15 veces:

$$mo = 3$$

5.4. Medidas de variabilidad

La variabilidad es la segunda tendencia que estudiaremos. Hay en principio una variabilidad general establecida por el llamado *rango* o *amplitud* que es la diferencia entre el mayor valor observado de la variable y el menor. Por supuesto nos estamos refiriendo a variables cuantitativas pues, cuando se trata de atributos cualitativos, medir la variabilidad se hace más complicado y no consideraremos aquí ese caso.

Más allá de la variabilidad general citada, importa estudiar como varían los datos respecto de la tendencia central. Si se elige para medir esta última a la media aritmética, cada dato se desviará una cantidad $(x_i - m)$. Si la frecuencia de cada dato es F_i y $N = \sum_i F_i$ se podría intentar promediar la suma total de los desvíos de la forma $\frac{\sum_i (x_i - m)F_i}{N}$ o bien, en forma equivalente, calcular $\sum_i (x_i - m)_i$. Como ocurre en la teoría de probabilidades cuando se considera una fórmula similar, reemplazando la frecuencia relativa por la probabilidad, tal cuenta compensará cantidades positivas con negativas y se hará 0. Análogamente también puede definirse la *desviación media* haciendo $DM = \sum_i |x_i - m| \, f_i$ lo que derivará en las mismas dificultades operatorias que se presentan en aquella teoría. La solución a la mano, para positivizar los desvíos y hacer su sumatoria distinta de 0, es nuevamente elevarlos al cuadrado. Se tiene así una medida que se denomina directamente *varianza* o *variancia* pero que ahora está calculada "a posteriori" de la experiencia de recoger los datos. Su fórmula es entonces:

$$Var = \frac{\sum_i (x_i - m)^2 F_i}{N} = \sum_i (x_i - m)^2 f_i$$

Las propiedades ya vistas para la varianza de variables aleatorias son por supuesto válidas aquí. En particular nos interesa señalar la formula de cálculo análoga:

$$Var = \frac{\sum_i x_i^2 F_i}{N} - \left(\frac{\sum_i x_i F_i}{N}\right)^2 = \sum_i x_i^2 f_i - \left(\sum_i x_i f_i\right)^2$$

Como en el caso de las variables aleatorias, los datos se dan en unidades por lo cual la variancia calculada queda en esas unidades elevadas al

cuadrado. Para subsanar este inconveniente se define el *desvío estándar* como $\sigma = \sqrt{Var}$ y entonces puede anotarse $Var = \sigma^2$.

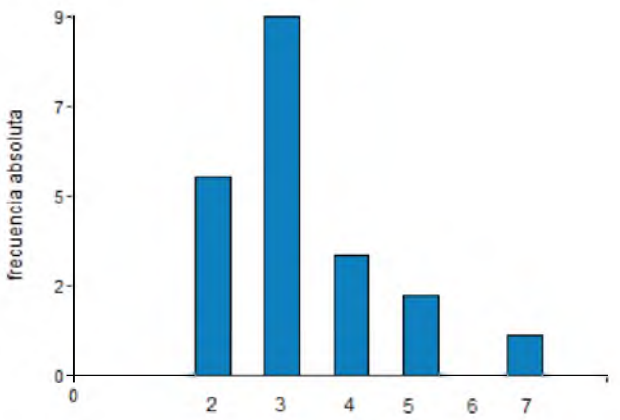
Hay otras medidas de variabilidad general como, por ejemplo, el *rango intercuartílico*. Si se consideran los datos ordenados de menor a mayor se pueden definir los *cuartiles*. El primer cuartil es el dato ubicado en la cuarta parte de los mismos, el segundo cuartil es la mediana y el tercer cuartil es el dato ubicado en las tres cuartas partes de los datos ordenados. El rango intercuartílico es entonces el valor absoluto de la resta entre el tercer y el primer cuartil. Como se ve, con los datos ordenados de menor a mayor y divididos en cuatro partes surgen tres cuartiles. En forma análoga al dividirlos en diez partes surgen los *deciles* y al dividirlos en cien, los *percentiles*, medidas todas que tienen algún uso en estadística aunque solo las mencionamos aquí.

Ejemplo 3: *Una compañía tiene veinte representantes de venta en todo el país. El número de unidades que el último mes vendió cada representante resultó: 2 3 2 3 3 4 2 4 3 2 3 4 5 3 3 3 3 5 2 7*

a) Graficar el comportamiento de esta población. Evaluar su media y su desvío estándar.

Es conveniente agrupar los datos en una tabla, detallando en la primera columna, cada uno de los datos y en la segunda columna su frecuencia, es decir, la cantidad de veces que aparece dicho valor en el conjunto. Esta es la distribución de frecuencias de la variable "número de unidades vendidas". Las frecuencias absolutas establecen la cantidad de representantes que venden la respectiva cantidad de unidades.

X_i (Unidades vendidas)	F_i
2	5
3	9
4	3
5	2
6	0
7	1



Media aritmética poblacional:

$$\mu = \frac{\sum_i x_i f_i}{N} = \frac{2 \times 5 + 3 \times 9 + 4 \times 3 + 5 \times 2 + 6 \times 0 + 7 \times 1}{20} = \frac{66}{20} = 3{,}3$$

Resulta importante observar que la media calculada no resulta una cantidad entera de unidades pero es un valor teórico que puede ser utilizado tal como aparece para análisis comparativos y, en general, para la toma de decisiones.

Desvío poblacional: Es la medida de la "variación" de los datos con respecto a la media. Es decir:

- El desvío es cero si todos los dato son iguales.
- El desvío es un número pequeño, si los datos son próximos entre sí.
- Puede aumentar significativamente si se introducen valores extremos.

$$\sigma = \sqrt{Var} = \sqrt{\frac{\sum_i (x_i - m)^2 F_i}{N}}$$

$$\sigma = \sqrt{\frac{5(2-3{,}3)^2 + 9(3-3{,}3)^2 + 3(4-3{,}3)^2 + 2(5-3{,}3)^2 + 1(7-3{,}3)^2}{20}} =$$

$$= \sqrt{\frac{8{,}45 + 0{,}81 + 1{,}47 + 5{,}78 + 13{,}69}{20}} = \sqrt{\frac{30{,}2}{20}} = \sqrt{1{,}51} \approx 1{,}23$$

b) Tomar 5 muestras aleatorias de 5 representantes de ventas. Evaluar el promedio de ventas de cada muestra y el promedio de estos promedios. Comparar con el promedio poblacional.

M_1 : 4 2 5 2 7

$$m_1 = \frac{4+2+5+2+7}{5} = \frac{20}{5} = 4$$

M_2 : 3 3 2 3 7

$$m_2 = \frac{3+3+2+3+7}{5} = \frac{18}{5} = 3{,}6$$

M_3 : 3 2 2 3 5

$$m_3 = \frac{3+2+2+3+5}{5} = \frac{15}{5} = 3$$

M_4 : 4 4 2 3 3

$$m_4 = \frac{4+4+2+3+3}{5} = \frac{16}{5} = 3{,}2$$

M_5 : 4 3 3 7 2

$$m_5 = \frac{4+3+3+7+2}{5} = \frac{19}{5} = 3{,}8$$

$$m = \frac{m_1 + m_2 + m_3 + m_4 + m_5}{5} = \frac{4 + 3{,}6 + 3 + 3{,}2 + 3{,}8}{5} = \frac{17{,}6}{5} = 3{,}52$$

Obsérvese que la media poblacional es $\mu = 3{,}3$ y que la media de estas 5 medias muestrales da 3.52, un valor próximo.

c) Tomar ahora diez muestras de 10 representantes de ventas cada una y repetir los cálculos y la comparación.

M_1 : 5 3 2 3 2 4 4 3 2 2

$$m_1 = \frac{5+3+2+3+2+4+4+3+2+2}{10} = \frac{30}{10} = 3$$

M_2 : 5 3 4 2 3 3 3 3 3 4

$$m_2 = \frac{5+3+4+2+3+3+3+3+3+4}{10} = \frac{33}{10} = 3{,}3$$

M_3 : 3 3 3 7 3 3 2 3 2 3

$$m_3 = \frac{3+3+3+7+3+3+2+3+2+3}{10} = \frac{32}{10} = 3{,}2$$

M_4 : 2 2 3 4 2 3 3 3 4 2

$$m_4 = \frac{2+2+3+4+2+3+3+3+4+2}{10} = \frac{28}{10} = 2{,}8$$

M_5 : 3 3 4 3 4 3 2 7 3 2

$$m_5 = \frac{3+3+4+3+4+3+2+7+3+2}{10} = \frac{34}{10} = 3{,}4$$

M_6 : 2 3 4 3 4 3 5 3 2 2

$$m_6 = \frac{2+3+4+3+4+3+5+3+2+2}{10} = \frac{31}{10} = 3{,}1$$

M_7 : 3 2 3 7 4 3 3 2 5 2

$$m_7 = \frac{3+2+3+7+4+3+3+2+5+2}{10} = \frac{34}{10} = 3{,}4$$

M_8 : 5 3 7 2 3 2 5 3 3 2

$$m_8 = \frac{5+3+7+2+3+2+5+3+3+2}{10} = \frac{35}{10} = 3{,}5$$

M_9 : 5 3 3 3 3 5 2 3 3 7

$$m_9 = \frac{5+3+3+3+3+5+2+3+3+7}{10} = \frac{37}{10} = 3{,}7$$

M_{10} : 2 3 3 2 3 3 4 5 4 4

$$m_{10} = \frac{2+3+3+2+3+3+4+5+4+4}{10} = \frac{33}{10} = 3{,}3$$

$$m = \frac{3+3{,}3+3{,}2+2{,}8+3{,}4+3{,}1+3{,}4+3{,}5+3{,}7+3{,}3}{10} = \frac{32{,}7}{10} = 3{,}27$$

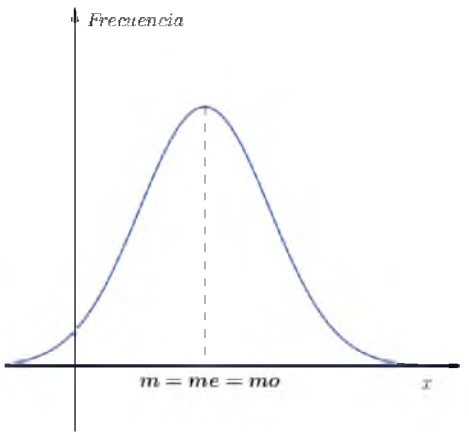
d) Extraer conclusiones

Por un lado es importante notar, a partir de las muestras de tamaño 5 tomadas en a), que la media es bastante sensible a valores extremos. Esto se hace visible en todas las muestras, en la que aparece 7 como valor más "atípico". Por otro lado, si bien la media muestral depende de los valores particulares incluidos en la muestra, pudiendo variar de una muestra a otra (más notorio en el item a)) el promedio de las mismas se acerca bastante a la media poblacional, en las muestras de mayor tamaño (ítem b)). En general cuanto mayor sea el tamaño muestral ocurrirá que el promedio de las medias muestrales más se acercará a la media poblacional.

5.5. Medidas de asimetría

Por *asimetría* se quiere entender, precisamente, la pérdida de la simetría en la distribución de los datos. Por ejemplo, si la variable fuese continua y su distribución fuese similar a una función de densidad de una variable aleatoria normal, cabría esperar un comportamiento simétrico y acampanado de las frecuencias. En tal ocasión debiera ocurrir que la media aritmética calculada promediando los datos, la mediana obtenida en la posición central de los mismos y el modo resultante del valor de la variable con mayor frecuencia, coincidieran aproximadamente. Es decir la simetría implica la cadena de igualdades $m = me = mo$ como se grafica en la figura 6:

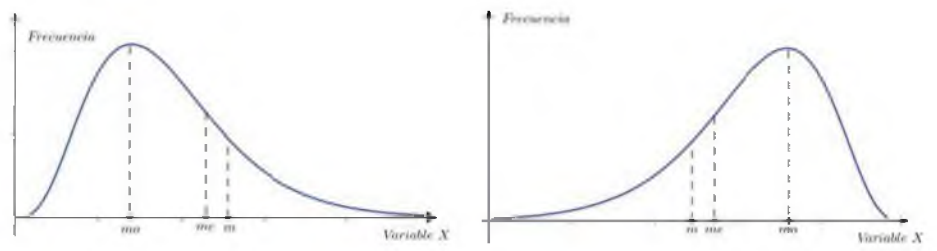
Figura 6.



Si por el contrario la distribución es aproximadamente acampanada pero no es del todo simétrica se pueden observar cualquiera de las dos alternativas que muestran las figuras 7-a y 7-b:

Figura 7.

a) b)



En la figura 7-a, con el grueso de los datos a la derecha del potencial eje de simetría marcado por el pico de la forma acampanada, se observa que $mo < me < m$ mientras que lo inverso ocurre en la figura 7-b, $m < me < mo$, en la cual la parte más grande de los datos queda a izquierda del posible eje de simetría.

Una relación empírica, es decir obtenida a través de la observación de muchas distribuciones de frecuencia de forma más o menos acampanada y no del todo simétrica, indica que la distancia entre el modo y la media suele ser aproximadamente el triple de la que hay entre la mediana y la media. En fórmulas esto puede ponerse $3(m-me) \approx (m-mo)$. Si se considera la figura 7-a es claro que $m - mo > 0$ y podemos hablar entonces de *signo de asimetría* o *sesgo* positivo. Mientras que en la figura 7-b el signo de asimetría es negativo ya que $m - mo < 0$.

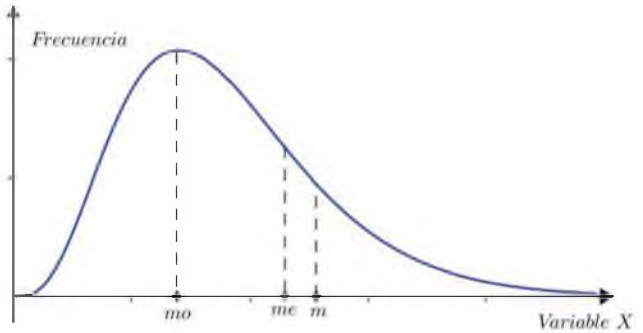
Con esto bastaría para caracterizar la asimetría si no fuera porque, como ya hemos visto, las variables pueden tener más de un modo, es decir ser multimodales. Esta complicación se zanja apelando a la relación empírica apuntada que permite utilizar la mediana en vez del modo. Así, dividiendo además por el desvío estándar para tener un coeficiente relativo a él y de carácter adimensional pues las unidades se eliminan, se define el *coeficiente de asimetría*:

$$CA = \frac{3(m - me)}{\sigma}$$

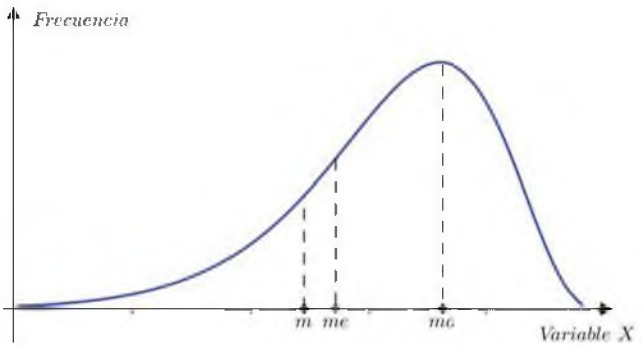
cuyo signo marca además el sesgo. Claramente si $CA = 0$ resulta que hay simetría pues se estaría en la situación descripta por la figura 6. En cambio en la medida en que crece en valor absoluto CA la asimetría es mayor.

Ejemplo 4: *Si el 66.67 % de los valores de una variable x están a la derecha de la media ¿Cuál es el signo de asimetría?*

El coeficiente de asimetría es un parámetro que permite establecer el grado de asimetría que presenta una distribución de frecuencias de datos. Una distribución es simétrica, si existe el mismo número de valores a la derecha que a la izquierda de la media. Tal lo que ocurre, por ejemplo, en una distribución exactamente normal de los datos. En general, para que una distribución sea simétrica, debe suceder que $m = me = mo$. Obsérvese que esto implica que $CA = \frac{3(m-me)}{\sigma} = 0$. Si $mo < me < m$, que es el caso de la figura de abajo, el signo de asimetría viene dado por $m - mo > 0$, es decir positivo:



Sin embargo, en el problema que estamos analizando el 66,67 % de los datos se encuentra a derecha de la media y por lo tanto uno de ellos deberá ser la mediana, que es el dato que deja exactamente 50 % para cada lado. Esto quiere decir que $m \leq me$. Si la distribución "se porta bien" debe ser suave, acampanada y cumplirse que $m < me < mo$. Por lo tanto el signo de asimetría que resulta de $m - mo$ es negativo. La figura de abajo muestra el caso:



5.6. Ejercicios

Ejercicio N°1*: El Director de Ingreso a una Universidad tiene los puntajes obtenidos por 20 alumnos en un examen: 98 65 70 62 85 71 56 72 90 51 59 79 66 80 94 55 79 62 63 73

a) Elaborar una distribución de frecuencias por clases.

b) Realice un histograma y un polígono de frecuencias con los datos agrupados en clases.

c) ¿Qué tipo de variable está considerando?

Ejercicio N°2*: Una azafata fue anotando las bebidas que sirvió a los pasajeros de un avión (una por pasajero): café, café, café, agua mineral, agua mineral, jugo, agua mineral, te, agua mineral, agua mineral, café, agua mineral, te, jugo, jugo, café, agua mineral, café, jugo, café, café, agua mineral, te, te, agua mineral, jugo, café, agua mineral, café, café.

a) ¿Qué tipo de variable está considerando?

b) Organice la distribución de frecuencias absolutas, relativas y porcentuales.

c) Realice un gráfico de barras y otro de pastel.

Ejercicio N°3: Dados los datos de los ejercicio 1 y ejemplo 1 evaluar en cada caso:

a) Media, mediana y modo.

b) Rango, desvío estándar y desviación media.

c) Cuartiles.

d) Coeficiente de variación $CV = \frac{\sigma}{m}$.

e) Coeficiente de asimetría.

Ejercicio N°4: Si se tienen dos distribuciones campaniformes y simétricas tal que la primera de las cuales tiene $modo = 16$ y $Var = 4$ y la segunda $mediana = 12$ y $desvío\ estándar = 2$ ¿Cuál de ellas tiene mayor dispersión relativa y cual mayor dispersión absoluta? Usar coeficiente de variación.

Ejercicio N°5: Si $\sum_1^7 x_i = 28$ y $\sum_1^7 {x_i}^2 = 140$ determinar la media y la desviación estándar de los valores de x.

Capítulo 6

Estimación de parámetros poblacionales

6.1. Introducción

Hasta aquí se han analizado las tendencias poblacionales y se han establecido medidas que permiten cuantificarlas. Las medidas de la tendencia central, la variabilidad o la asimetría se calculan conociendo el valor de la variable respectiva para todos los individuos o instancias que integran la población. Sin embargo esta situación, en la mayoría de los casos, dista de ser real pues muchas veces por razones de costos, otras muchas por el tamaño de la población o porque las mediciones implican la destrucción del objeto a medir, se hace difícil o aún imposible conocer el valor de la variable para cada individuo. Como de todas formas las tendencias deben establecerse se recurre a muestras, pequeños subconjuntos de la población, para *estimar* los valores de los parámetros poblacionales que las expresan. Este proceso, denominado en general *inferencia estadística*, requiere el cálculo de un *estadístico* o *estimador*, a partir de los datos de la muestra, para estimar el valor del *parámetro poblacional* con cierto margen de error o con cierta probabilidad de acierto. Por supuesto, el tamaño de la muestra juega un papel importante así como también lo hace la exactitud con que se conozcan los valores de otros parámetros de la misma población.

Por ejemplo, para estimar la altura media μ de los estudiantes de la Universidad se mide la altura de 25 estudiantes del curso de Probabilidad y Estadística. A continuación se calcula la media para esa muestra

que supongamos es $\bar{x} = 1{,}69$. $\bar{x}$ es un *estimador* de μ. La primera reflexión que hay que hacer es que, quizás, la altura media poblacional no sea exactamente $\mu = 1{,}69$ y que si se hubiera tomado la muestra integrada por los 30 alumnos del curso de Economía, el promedio muestral de altura podría resultar $\bar{x} = 1{,}71$ con lo cual tendríamos un valor estimado de μ distinto. Es decir, la estimación puede no dar un valor exacto y lo que suele hacerse es evaluar la probabilidad de que el valor obtenido para el estimador muestral represente adecuadamente al valor del parámetro poblacional. En general se tiene el esquema de la tabla 1:

Tabla 1.

Población/Parámetro	Muestra/Estimador
Media μ	$\bar{x}$
Varianza σ^2	s^2
Desvío Estándar σ	s
Proporción p	$\hat{p}$

Estos no son, por supuesto, todos los parámetros y estimadores que puedan interesar, pero son los más comunes e importantes. En verdad, cada tendencia poblacional puede estar expresada por distintos parámetros y a su vez cada uno de ellos puede tener diferentes formas de estimación. Conviene aclarar además que el valor que adopta un estimador suele estar relacionado no solo con el valor del parámetro respectivo sino también con los valores conocidos o desconocidos de otros parámetros poblacionales. Por ejemplo; si en la realidad hay gran variabilidad entre las alturas de los estudiantes de la Universidad, es decir varianza grande, dos muestras distintas podrían arrojar promedios $\bar{x}$ bastante distintos también. En cambio si la variabilidad poblacional fuera poca esos promedios deberían parecerse mucho más. Es claro entonces que conocer la variabilidad, por ejemplo el desvío estándar poblacional, puede ser útil para evaluar el error con que el promedio muestral $\bar{x}$ estima la media poblacional μ. Desde otro ángulo, el conocimiento de la variabilidad real poblacional puede ayudar a determinar el tamaño que debe tener la muestra cuando se desea cometer, en la estimación, un error pequeño prefijado.

Otra cuestión importante es la de las fórmulas de los estimadores, pues no siempre estos se calculan con iguales expresiones a las

que se usarían para calcular los parámetros. Además en el proceso de inferencia a veces se realizan suposiciones sobre el comportamiento poblacional que son útiles para enmarcar el problema, pero que simultáneamente introducen una cuota de error. En suma podríamos decir que la inferencia estadística es un arte, no una ciencia deductiva, que involucra la *forma de la inferencia* y una *medida de su bondad.*

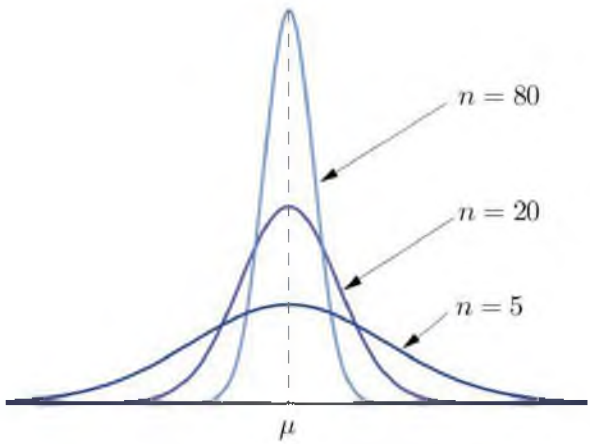
6.2. Los estimadores y sus propiedades

Si se considera la media muestral resulta claro que para cada muestra, aún del mismo tamaño, el valor de $\bar{x}$ depende de la elección azarosa de los individuos que la integran. De tal modo se comprende que cada valor posible de $\bar{x}$ tiene una cierta probabilidad de ocurrencia y que, por lo tanto, $\bar{x}$ es una variable aleatoria con una determinada distribución de probabilidad. Lo mismo ocurre para otros estadísticos que estiman otras tendencias poblacionales. Cabe entonces hacerse la pregunta general: ¿dado un estimador muestral de un parámetro poblacional determinado, cuál es su distribución de probabilidad?

En el caso del estimador $\bar{x}$ de la media poblacional μ, la respuesta puede obtenerse a partir del teorema del límite central presentado en el capítulo 4. Sin pretender una demostración exhaustiva, volvamos al ejemplo de las alturas de los estudiantes y consideremos que cada individuo que integra la muestra de tamaño n es elegido en forma independiente. De tal forma, el valor x que adopte su altura es una variable aleatoria. Es decir $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ es la suma de n variables aleatorias independientes dividida por n. Como vimos, el teorema asegura que para n suficientemente grande ($n \to \infty$) la distribución de tal suma es normal. Luego de algunas otras consideraciones se obtiene entonces como conclusión que, para n suficientemente grande, $\bar{x}$ se distribuye normalmente con esperanza $E(\bar{x}) = \mu$ y desvío estándar $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ donde μ y σ son la media y el desvío estándar poblacionales respectivamente. Claramente si crece el tamaño muestral n, disminuirá el desvío estándar en la distribución de la media muestral $\bar{x}$ según se ve en la figura 1. En la práctica, esto significará que una creciente cantidad de muestras arrojará valores de $\bar{x}$ cada vez más cercanos a μ. Tal efecto se constata para un mismo intervalo alrededor de la media

poblacional μ, pues conforme va creciendo n cada vez la probabilidad se concentra más sobre él.

Figura 1.



Hay que notar que el resultado obtenido acerca de la distribución del estimador $\bar{x}$ no depende de las características de la distribución poblacional. Es decir, la variable poblacional x puede tener cualquier distribución de media μ y desvío estándar σ, que no necesariamente sea normal, y de todas formas $\bar{x}$ tendrá el comportamiento apuntado, siempre que el tamaño muestral n sea el suficiente.

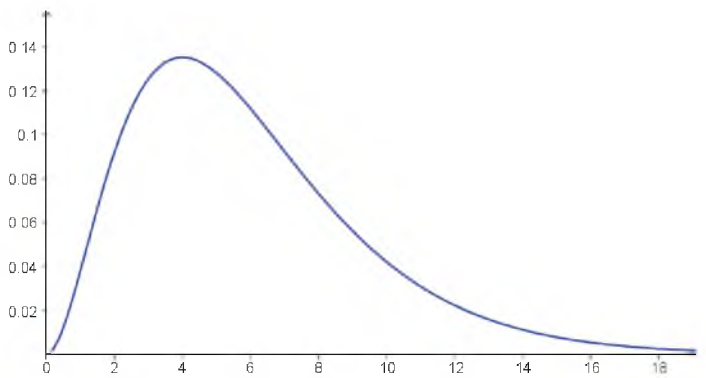
De acuerdo con lo expuesto $\bar{x}$ resulta un buen estimador para la media poblacional μ pues, además de resultar función *lineal* de los datos de la muestra, tiene la interesante propiedad de ser *insesgado*. Esta es la forma técnica de describir aquella situación por la cual, precisamente, el valor esperado del estimador coincide con el valor poblacional del parámetro que quiere estimarse. En este caso, como vimos, $E(\bar{x}) = \mu$. Un estimador para el que esto no ocurriera, es decir no pasara que su valor esperado fuera el del parámetro poblacional buscado, se denominaría *sesgado* y no sería tan buen estimador. En general se entiende que dado un parámetro θ a estimar, el mejor estimador $\hat{\theta}$ será aquel que sea lineal respecto de los datos muestrales, *insesgado* y de *varianza mínima*. Sin embargo, puede ocurrir que para ciertos parámetros no se hallen estimadores que cumplan simultáneamente con estas tres características. A veces también se tiene en cuenta una cuarta propiedad

deseable expresada como sigue: el estimador $\hat{\theta}$ *converge* "en probabilidad" al parámetro θ. Esto quiere decir que en la medida que crece el tamaño de la muestra, la probabilidad de que el valor del estimador se acerque tanto como se quiera al del parámetro, tiende a 1. En símbolos:

$$\lim_{n\to\infty} Prob\left(\left|\hat{\theta}-\theta\right| \leq \epsilon\right) = 1$$

La estimación de la varianza o del desvío estándar se apoya en resultados bien conocidos pero no tan simples como los que permiten construir el estimador de la media de una población. Hace falta, para comenzar, definir la variable aleatoria *chi-cuadrado* (o ji-cuadrado). Esta variable es la suma de los cuadrados de n variables aleatorias normales de esperanza 0 y desvío estándar 1 según: $\chi^2 = X_1^2 + X_2^2 + \ldots X_n^2$. Claramente resulta $\chi^2 \geq 0$. Las expresiones de la función de distribución y la correspondiente función de densidad no van a ser usadas en este libro introductorio[1] , aunque a título ilustrativo se muestra la forma de la función de densidad en la figura 2. Como se ve la distribución no es simétrica y es distinta de cero sólo cuando $\chi^2 \geq 0$:

Figura 2.



Por razones análogas a las apuntadas cuando se mencionó a la distribución normal, las probabilidades acumuladas para una variable aleatoria χ^2 se tabulan según se muestra en la tabla 1 que reproduce un fragmento de la tabla adjuntada como Anexo C.

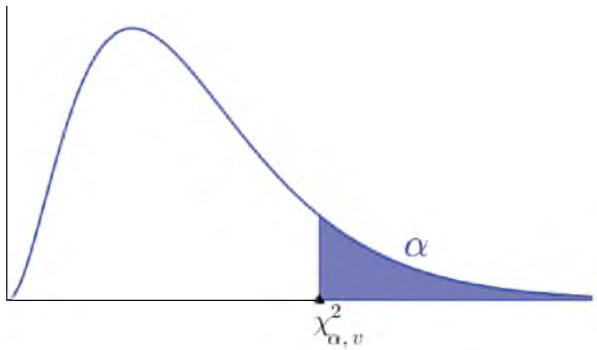
[1]Una deducción de esas fórmulas puede verse en Harald Cramér: *Teoría de probabilidades y aplicaciones*, Aguilar, pp. 131-134

Tabla 1.

ji-cuadrado	**Área de la cola,** α						
α/ν	**0.300**	**0.200**	**0.100**	**0.050**	**0.025**	**0.010**	**0.005**
1	1.07	1.64	2.71	3.84	5.02	6.63	7.88
2	2.41	3.22	4.61	5.99	7.38	9.21	10.60
3	3.66	4.64	6.25	7.81	9.35	11.34	12.84
4	4.88	5.99	7.78	9.49	11.14	13.28	14.86
5	6.06	7.29	9.24	11.07	12.83	15.09	16.75

En la primera columna se encuentra el número de variables normales sumadas que integran la variable aleatoria χ^2, al que por razones que veremos enseguida denominamos número de *grados de libertad.* En la primera fila se listan los distintos valores de áreas bajo la curva en la región sombreada de la figura 3:

Figura 3.



Los valores que se encuentran en el interior de la tabla corresponden a las abscisas a partir de las cuales se acumula el área α para cada uno de los grados de libertad ν.

La relación existente entre la variable aleatoria χ^2 y la estimación de la varianza poblacional σ^2 queda expresada por el siguiente resultado.

Propiedad 1: Sea una muestra aleatoria $X_1, X_2, \ldots, X_n$ de una variable aleatoria X, cuya esperanza es μ y su varianza σ^2. Sea la cantidad $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$. Se tiene que:

a) $E(s^2) = \sigma^2$

b) Si X se distribuye normalmente, se cumple que $\chi^2_{n-1} = \frac{n-1}{\sigma^2} s^2$ (El subíndice $n-1$ indica los grados de libertad de la variable χ^2 correspondiente)[2]

Como se intuye s^2 es el estimador de la varianza poblacional σ^2 que resulta insesgado por la parte a) de la propiedad. En cuanto al punto b), bajo la condición de comportamiento normal de la variable aleatoria X, el estimador s^2 multiplicado por la constante $\frac{n-1}{\sigma^2}$ tiene una distribución chi-cuadrado con $n-1$ grados de libertad. El término grados de libertad surge naturalmente cuando se considera que las cantidades $(x_i - \bar{x})$ no son todas independientes pues como ya hemos visto la suma de todos estos desvíos es 0 y entonces conocidos $n-1$ de ellos el n-ésimo surge de despejarlo en la ecuación:

$$(x_1 - \bar{x}) + (x_2 - \bar{x}) + \ldots + (x_n - \bar{x}) = x_1 + x_2 + \ldots + x_n - n\bar{x} = 0$$

De esta forma también resultarán independientes solo $n-1$ de los cuadrados $(x_i - \bar{x})^2$. En definitiva sólo hay $n-1$ términos independientes o libres y de ahí el nombre *grados de libertad* que se da a la cantidad de variables que forman la χ^2_{n-1} correspondiente a un s^2 que suma n desvíos observados en total.

Hasta aquí hemos construido estimadores para la media y la varianza poblacional. En el caso de la media, el estimador $\bar{x}$ se distribuye normalmente y puede estandarizarse haciendo $\bar{z} = \frac{(\bar{x}-\mu)}{\sigma_{\bar{x}}} = \frac{(\bar{x}-\mu)}{\frac{\sigma}{\sqrt{n}}}$ de tal forma que así se distribuya normalmente pero con media 0 y desvío estándar 1. Debe tenerse presente que μ y σ son la media y el desvío estándar poblacionales respectivamente. Cuando el desvío estándar poblacional no se conoce debe ser estimado por $s = \sqrt{s^2}$ ya que s^2 es el estimador de la varianza. En ese caso quedaría $= \frac{(\bar{x}-\mu)}{\frac{s}{\sqrt{n}}}$ y cabe preguntarse entonces si la distribución de esta variable sigue siendo normal. Ante esto, en primer lugar hay que señalar que la estimación s requiere como hipótesis que la variable aleatoria X sea normal, suposición que no hicimos cuando construimos el estimador $\bar{x}$ y dedujimos su distribución. La segunda cuestión importante es que si la muestra

[2]Un desarrollo más exhaustivo de esta propiedad puede verse en Paul Meyer, ob.cit., p. 282.

tiene pocos elementos esto redundará en una estimación pobre de la varianza y el desvío estándar poblacionales. En realidad solo para un tamaño muestral $n \geq 30$ suele considerarse que el estimador construido con el desvío estándar estimado s tiene un comportamiento similar al normal. Es decir, aquella expresión "*para n suficientemente grande*" que se citó cada vez que se utilizó el teorema del límite central, para justificar el comportamiento normal de una suma de variables independientes, adquiere aquí una forma práctica convencional. Con base empírica y observacional se considera que una muestra es grande si $n \geq 30$, y se la juzga pequeña si $n < 30$. Ahora bien; si la muestra es pequeña ¿como se distribuye entonces el estadístico $t = \frac{(\bar{x}-\mu)}{\frac{s}{\sqrt{n}}}$? La respuesta fue dada por William Gosset, estadístico británico y empleado de la cervecera Guinnes, quien en 1908 publicó su trabajo en referencia a este problema bajo el pseudónimo de Student. Desde entonces se dice que el estadístico t, útil cuando las muestras son pequeñas y no se conoce la varianza de la distribución, tiene una distribución *t de Student*. A nuestros fines no resultará importante conocer las fórmulas de la función de distribución o la de densidad[3] . En forma muy similar a lo realizado para la variable chi-cuadrado, la distribución de la t de Student se tabula teniendo en cuenta los grados de libertad involucrados en la estimación de s. La tabla 2 reproduce una parte de la tabla de la t de Student anexada en el Anexo B al final del libro:

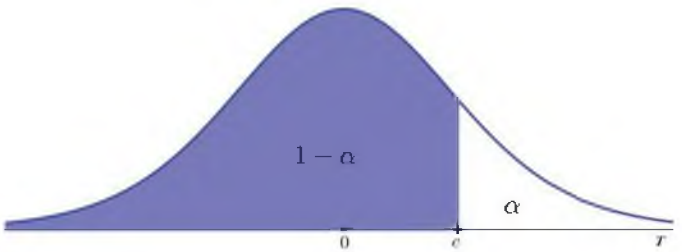
Tabla 2.

r	**0.75**	**0.80**	**0.85**	**0.90**	**0.95**	**0.975**	**0.99**	**0.995**
1	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657
2	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925
3	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032

En la primera columna se encuentra el número de grados de libertad con que se ha estimado s. En la primera fila se listan los distintos valores $1-\alpha$ de áreas bajo la curva en la región sombreada de la figura 4:

[3]La fórmula para la función de densidad de probabilidad de la variable t puede verse en Paul Meyer, ob.cit., p. 314.

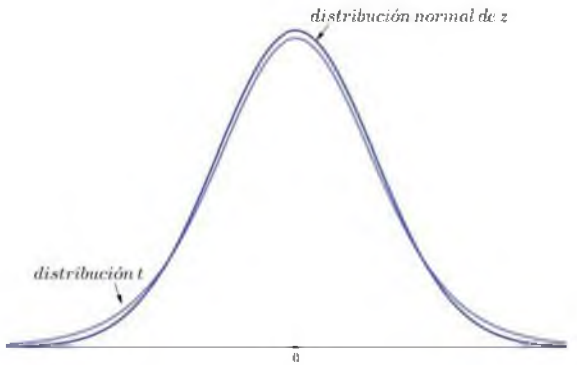
Figura 4.



Los valores que se encuentran en el interior de la tabla corresponden a las abscisas hasta las cuales se acumula el área $1 - \alpha$ para cada uno de los grados de libertad.

Finalmente es interesante observar, en la figura 5, que la distribución t de Student tiene colas un poco más gruesas que la normal por efecto de la mayor variabilidad que se presenta en las muestras chicas. De acuerdo a como va creciendo el tamaño muestral n, se va verificando un comportamiento de la t de Student cada vez más parecido al de la normal.

Figura 5.



A modo de resumen de esta sección digamos entonces que hemos visto, el estimador utilizado para la media poblacional con distribución normal, el estimador utilizado para la varianza y el desvío estándar

que multiplicado por una constante se distribuye chi- cuadrado y el efecto que se produce sobre la distribución del estimador de la media poblacional, al desconocer la varianza, lo que conduce a la variable t de Student. Vimos también propiedades deseables para un estimador y en particular la importancia de que no tenga sesgo. Con las ideas expuestas estamos ahora en condiciones de realizar la estimación de parámetros.

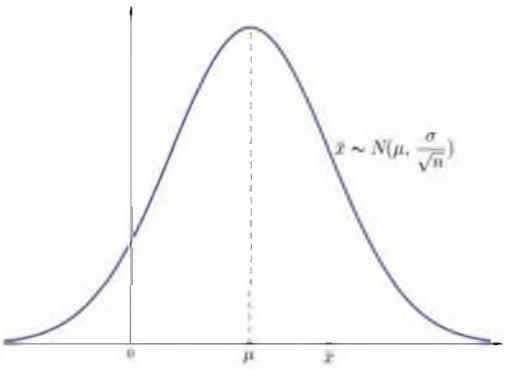
6.3. Estimación de parámetros

En el proceso de inferencia estadística se relacionan implícitamente tres distribuciones. La primera es la distribución de la variable poblacional cuya forma general suele desconocerse y a veces es necesario suponer. El objetivo de la inferencia es precisamente establecer valores probables para sus parámetros desconocidos. La segunda distribución involucrada es la que corresponde internamente a la muestra y que por ende es enteramente empírica. A partir de ella pueden calcularse los estimadores necesarios para evaluar los parámetros poblacionales. La tercera distribución es teórica, de probabilidad, y corresponde al estimador que se utilice. En ella, como resultado teórico verdadero, se apoya la inferencia. Es decir; el proceso de inferencia vincula la distribución muestral, empíricamente conocida, con la poblacional a través de la distribución teórica del estimador.

Una vez obtenida una muestra de tamaño n, la forma más directa y elemental de estimar un parámetro poblacional es calcular directamente el valor de su estimador. Este procedimiento se denomina *estimación puntual* y, en principio, tiene el inconveniente de no aportar simultáneamente una medida de la bondad de la estimación que realiza. Para adquirir una idea del error que puede estarse cometiendo se calculan cotas de error sobre la base de consideraciones empíricas. Por ejemplo; para estimar una media poblacional μ se utiliza el estadístico $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ que, como ya hemos visto, tiene una distribución normal con $E(\bar{x}) = \mu$ y desvío estándar $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$. Ocurre que el valor obtenido de $\bar{x}$ es uno de los posibles alrededor de μ ó, en el mejor de los casos, el propio μ. Calculado el promedio muestral se está en la

situación que ejemplifica la figura 6:

Figura 6.

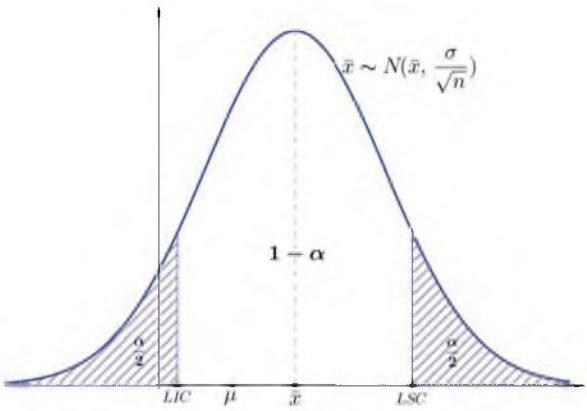


Sabemos, por la regla empírica acerca de la probabilidad normal, que para el intervalo formado por dos desvíos estándar desde la media, a izquierda y derecha, la probabilidad correspondiente es mayor que 0.95. Entonces la probabilidad de que un valor promedio observado caiga dentro de ese intervalo es mayor que 0.95. Hay una muy alta probabilidad de que el promedio muestral se encuentre dentro del intervalo $[\mu - 2\sigma_{\bar{x}}, \mu + 2\sigma_{\bar{x}}]$. Es decir; con probabilidad mayor que 0.95 debe ocurrir que $|\mu - \bar{x}| \leq 2\sigma_{\bar{x}}$. Por lo tanto es poco probable que el error $\mu - \bar{x}$ que se cometa al estimar sea mayor que $2\sigma_{\bar{x}}$ y entonces se adopta el valor $2\sigma_{\bar{x}}$ como cota del error de estimación. En resumen la estimación puntual es: $\mu \approx \bar{x} \pm \frac{2\sigma}{\sqrt{n}}$ Obsérvese la importancia que ha tenido en este proceso el hecho de que el estimador fuese insesgado pues se trabaja alrededor de su valor esperado.

Otro enfoque de la estimación de parámetros poblacionales es el de los *intervalos de confianza* cuyas estimaciones se denominan por ello *interválicas*. El procedimiento general consiste en suponer primero que el valor obtenido por el estimador calculado a partir de la muestra es, efectivamente el correspondiente al parámetro. A partir de esto y utilizando el valor del estimador obtenido como centro, se construye un intervalo con dos extremos denominados LIC, límite inferior de confianza, y LSC, límite superior de confianza. Entonces para tal intervalo se acumula una cantidad de probabilidad $1 - \alpha$, quedando $\frac{\alpha}{2}$ en cada

cola de la distribución. Nuevamente ejemplifiquemos con la estimación de una media poblacional. Obtenida la muestra de tamaño n, calculamos el estimador. Si suponemos que este es efectivamente μ, deberá ser el valor central de la campana de Gauss que representa la distribución del estimador como puede verse en la figura 7:

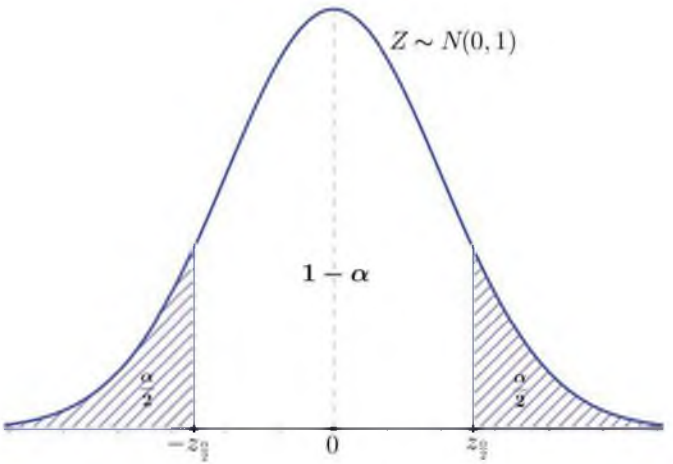
Figura 7.



Ahora bien, el verdadero valor del parámetro poblacional μ deberá ser uno de los valores del eje de abscisas de la gráfica y por lo tanto la probabilidad de que el intervalo $[LIC, LSC]$ contenga al mismo, será $1-\alpha$. En símbolos $P\left(LIC \leq \mu \leq LSC\right) = 1-\alpha$

Los valores hasta aquí desconocidos de los límites de confianza pueden conocerse al estandarizar la variable pues $1-\alpha$, ó $\frac{(1-\alpha)}{2}$ si se piensa en la tabla de la distribución normal tal como está adjuntada el Apéndice, es una probabilidad o *nivel de confianza* que se fija de antemano y que usualmente se expresa en forma porcentual. En la figura 8 se aprecia como los valores de LIC y LSC se transforman al estandarizarlos en cantidades indicadas como $-z_{\frac{\alpha}{2}}$ y $z_{\frac{\alpha}{2}}$ que dejan en ambas colas precisamente cantidades $\frac{\alpha}{2}$ de probabilidad.

Figura 8.



En fórmulas:

$$-z_{\frac{\alpha}{2}} = \frac{LIC - \bar{x}}{\frac{\sigma}{\sqrt{n}}} \quad y \quad z_{\frac{\alpha}{2}} = \frac{LSC - \bar{x}}{\frac{\sigma}{\sqrt{n}}}$$

despejando se obtiene:

$$LIC = \bar{x} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \quad y \quad LSC = \bar{x} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

Por lo tanto, finalmente tenemos:

$$P\left(\bar{x} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

Resulta entonces que el intervalo $\left[\bar{x} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right]$ contiene a la media poblacional μ con una probabilidad $1 - \alpha$.

Consideremos el siguiente ejemplo. Se tiene una muestra de tamaño $n = 36$ de una población cuya distribución y media son desconocidas. El promedio muestral es $\bar{x} = 6$ y se sabe que la varianza poblacional es $\sigma^2 = 9$. Si se desea establecer un intervalo de confianza del 95 % para la media poblacional, se debe proceder como sigue. Como 95 % de confianza implica una probabilidad $1-\alpha = 0{,}95$ se tiene que $\frac{\alpha}{2} = 0{,}025$. En la tabla de la distribución normal estándar se busca entonces la

abscisa cuya probabilidad acumulada desde 0 hasta ella es 0.4750. Esta abscisa es $z_{\frac{\alpha}{2}} = 1{,}96$ y por simetría $-z_{\frac{\alpha}{2}} = -1{,}96$. Ahora se calcula entonces el intervalo:

$$LIC = 6 - 1{,}96 \times \frac{3}{\sqrt{36}} = 5{,}02 \;\; y \;\; LSC = 6 + 1{,}96 \times \frac{3}{\sqrt{36}} = 6{,}98$$

De acuerdo a esto escribimos $P\,(5{,}02 \leq \mu \leq 6{,}98) = 0{,}95$ lo que significa que con una confianza del 95 % el intervalo [5,02 ; 6,98] contiene a la media poblacional μ.

Al estimar la media poblacional si la muestra es pequeña, $n \leq 30$, y no se conoce la varianza poblacional, el intervalo de confianza puede construirse utilizando la distribución t de Student con $n-1$ grados de libertad. Queda entonces:

$$P\left(\bar{x} - t_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}\right) = 1 - \alpha$$

Sin embargo, hay que señalar que tanto mejor será la estimación interválica cuanto más se acerque a la normal el comportamiento de la variable aleatoria pues, como vimos, la distribución t del estadístico media muestral se produce bajo el supuesto de que la población se distribuye normalmente.

Para establecer, a partir de una muestra de tamaño n, una estimación interválica de la varianza poblacional σ^2 hay que tener en cuenta la relación que la liga al estadístico s^2 a través de la variable *chi-cuadrado* con $n-1$ grados de libertad. Como $\chi^2 = \frac{(n-1)}{\sigma^2}s^2$ se tiene que $\sigma^2 = \frac{(n-1)}{\chi^2}s^2$ y por lo tanto para una confianza del $(1-\alpha)\,\%$ resulta:

$$LIC = \frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}}} \quad y \quad LSC = \frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2}}}$$

Como en ambas fracciones el numerador es el mismo, está claro que el denominador en la fórmula del límite inferior debe ser mayor que ese número en la expresión del límite superior. Por eso el valor de $\chi^2_{\frac{\alpha}{2}}$ es la abscisa que deja a derecha una probabilidad $\frac{\alpha}{2}$ o, lo que es lo mismo, que corresponde a una probabilidad $1-\frac{\alpha}{2}$ a izquierda, y similarmente $\chi^2_{1-\frac{\alpha}{2}}$ deja a derecha la probabilidad $1-\frac{\alpha}{2}$.

Ejemplo 1: *La longitud de los pernos fabricados por una máquina se distribuye normalmente. Se toma una muestra aleatoria, se los mide, resultando los siguientes valores en cm: 2.84 2.92 2.80 2.79 2.90 2.90 Determine:*

a) Un intervalo de confianza del 90 % para la longitud media.

Se pide estimar un parámetro poblacional, en este caso la media, a través de una muestra de tamaño $n = 6$. Es posible, a partir de ésta, obtener la media muestral:

$$\bar{x} = \frac{2{,}84 + 2{,}92 + 2{,}80 + 2{,}79 + 2{,}90 + 2{,}90}{6} = \frac{17{,}15}{6} \approx 2{,}86$$

y el desvío muestral:

$$\frac{\sum (x_i - \bar{x})^2}{n-1} \approx \frac{(2{,}84 - 2{,}86)^2 + (2{,}92 - 2{,}86)^2 + (2{,}80 - 2{,}86)^2}{5} +$$

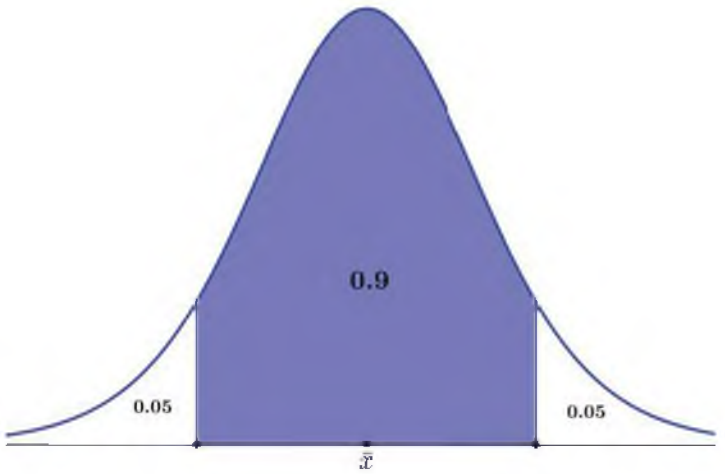
$$+ \frac{(2{,}79 - 2{,}86)^2 + (2{,}90 - 2{,}86)^2 + (2{,}90 - 2{,}86)^2}{5} \approx 0{,}00314$$

$$s = \sqrt{s^2} \approx \sqrt{0{,}00314} \approx 0{,}056$$

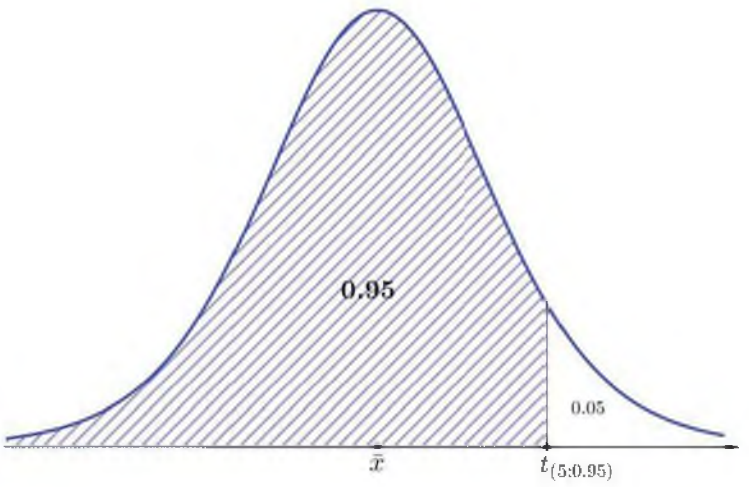
Como se trata de una variable aleatoria normal, pero no se conoce el valor del desvío poblacional, es entonces necesario observar el tamaño de la muestra. Al ser una muestra chica ($n \leq 30$) se debe utilizar la distribución *t-student* para obtener el intervalo de confianza (si la muestra fuera grande se podría utilizar la distribución normal):

$$\left(\bar{x} - t_{(n-1,1-\frac{\alpha}{2})} \frac{s}{\sqrt{n}} \ , \ \bar{x} + t_{(n-1,1-\frac{\alpha}{2})} \frac{s}{\sqrt{n}} \right)$$

Se pide un intervalo de confianza del 90 %, entonces $1 - \alpha = 0{,}9 \Rightarrow \alpha = 0{,}1$ gráficamente:



entonces $1 - \frac{\alpha}{2} = 0{,}95$ que también se puede representar gráficamente:



Utilizando la tabla de distribución *t-student* buscamos $t_{5;0,95}$, es decir, el valor de t para el cual el área a la izquierda es de 0,95, con 5 grados de libertad:

r	**0.75**	**0.80**	**0.85**	**0.90**	**0.95**	**0.975**	**0.99**	**0.995**
1	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657
2	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925
3	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604
5	0.727	0.920	1.156	1.476	**2.015**	2.571	3.365	4.032

Obteniéndose el intervalo de confianza:

$$\left(2{,}86 - 2{,}015 \times \frac{0{,}056}{\sqrt{6}}; 2{,}86 + 2{,}015 \times \frac{0{,}056}{\sqrt{6}}\right) \approx (2{,}81; 2{,}91)$$

Este intervalo contiene la longitud media, con una confianza del 90 %.

b) Un intervalo de confianza del 95 % para el desvío poblacional.

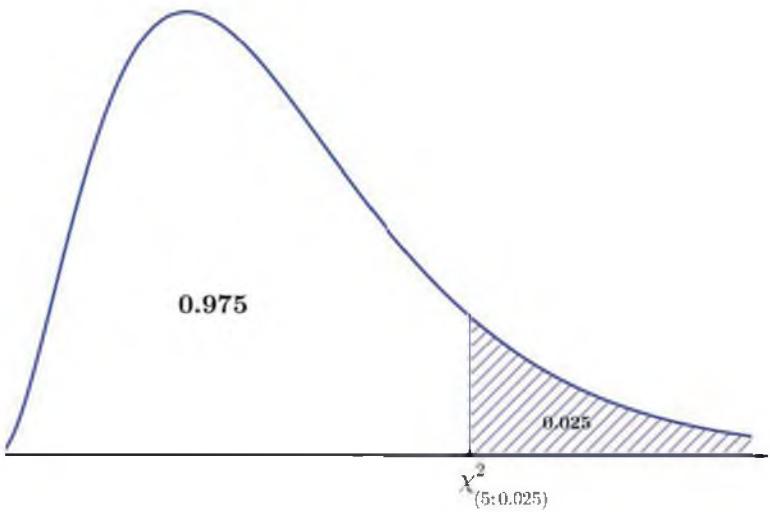
Como se ha dicho anteriormente, se trata de una variable aleatoria normal, pero no se conoce el valor del desvío poblacional. Para hallar un intervalo de confianza para la varianza se debe utilizar la distribución *ji-cuadrado*:

$$\left(\frac{(n-1)\cdot s^2}{\chi^2_{(n-1;\frac{\alpha}{2})}}; \frac{(n-1)\cdot s^2}{\chi^2_{(n-1;1-\frac{\alpha}{2})}}\right)$$

Por lo calculado en el ítem anterior, con $s^2 \approx 0{,}00314$ y $n = 6$ se obtiene el intervalo de confianza para la varianza:

$$\left(\frac{5 \times 0{,}00314}{\chi^2_{(5;0{,}025)}}; \frac{5 \times 0{,}00314}{\chi^2_{(5;0{,}975)}}\right)$$

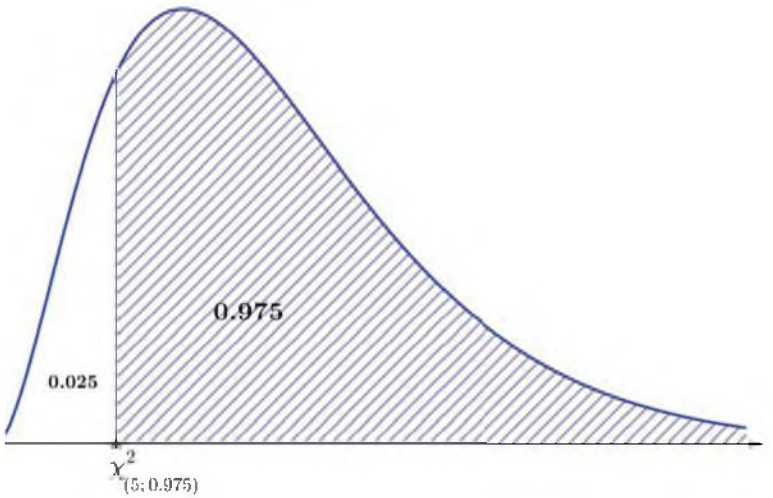
Utilizando la tabla de distribución *ji-cuadrado* se busca $\chi^2_{(5;0{,}025)}$, es decir, el valor de χ^2 para el cual el área a la izquierda es de 0,975, y a la derecha es de 0,025 con 5 grados de libertad:



Observación: Buscamos el valor 0,025 en la tabla ya que la tabla con la que estamos trabajando utiliza el área determinada a la derecha y no a la izquierda.

ji-cuadrado	**Área de la cola,** α						
α/ν	**0.300**	**0.200**	**0.100**	**0.050**	**0.025**	**0.010**	**0.005**
1	1.07	1.64	2.71	3.84	5.02	6.63	7.88
2	2.41	3.22	4.61	5.99	7.38	9.21	10.60
3	3.66	4.64	6.25	7.81	9.35	11.34	12.84
4	4.88	5.99	7.78	9.49	11.14	13.28	14.86
5	6.06	7.29	9.24	11.07	**12.83**	15.09	16.75

Debido a que la curva no es simétrica, para encontrar el otro extremo del intervalo, se busca de forma análoga , es decir, el valor de para el cual el área a la izquierda es de 0,025, y a la derecha es de 0,975 con 5 grados de libertad:



ji-cuadrado	**Área de la cola,** α						
α/ν	**0.995**	**0.990**	**0.975**	**0.950**	**0.900**	**0.800**	**0.700**
1	0.00	0.00	0.00	0.00	0.02	0.06	0.15
2	0.01	0.02	0.05	0.10	0.21	0.45	0.71
3	0.07	0.11	0.22	0.35	0.58	1.01	1.42
4	0.21	0.30	0.48	0.71	1.06	1.65	2.19
5	0.41	0.55	0.83	1.15	1.61	2.34	3.00

$$\left(\frac{0{,}0157}{12{,}83}; \frac{0{,}0157}{0{,}83}\right) \approx (0{,}00122; 0{,}01892)$$

El intervalo obtenido es para la varianza, pero se pidió el intervalo de confianza para el desvío. Para ello se debe, entonces, obtener la raíz cuadrada de los extremos del intervalo, recordar que el desvío es la raíz cuadrada de la varianza:

$$\left(\sqrt{0{,}00122}; \sqrt{0{,}01892}\right) \approx (0{,}035; 0{,}138)$$

Este intervalo contiene el desvío poblacional, con una confianza del 95 %

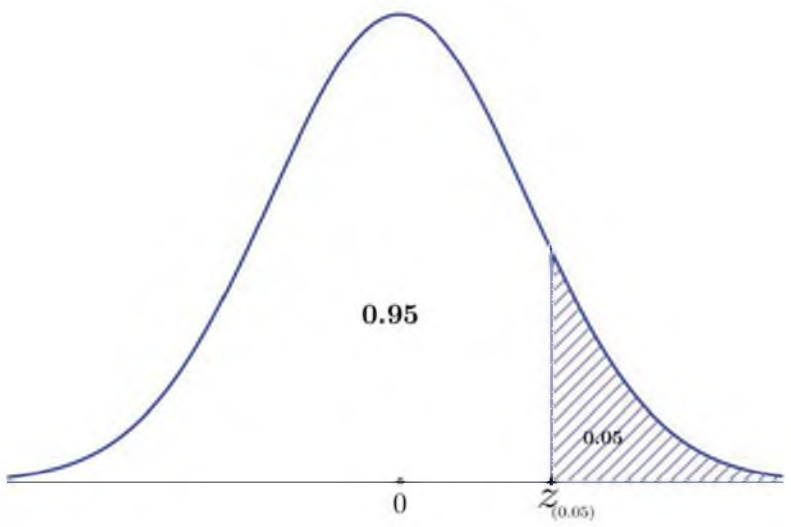
Otros parámetros pueden ser estimados con similar técnica. En el ejemplo siguiente se construye un intervalo de confianza para una proporción poblacional.

Ejemplo 2: *Se quiere estimar la proporción de artículos defectuosos en un proceso de manufactura. Se extraen 400 artículos al azar, encontrándose 10 defectuosos. Halle un intervalo de confianza para el porcentaje de defectuosos, con* $\alpha = 0{,}1$

$\hat{p} = \frac{10}{400} = 0{,}025$ es la proporción de éxitos de la muestra y el intervalo de confianza para el parámetro binomial p está dado por:

$$\left(\hat{p} - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}}; \hat{p} + z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}}\right)$$

Donde $z_{\frac{\alpha}{2}}$ es el valor de z que deja un área de $\frac{\alpha}{2}$ a la derecha. Si $\alpha = 0{,}1$ entonces $\frac{\alpha}{2} = 0{,}05$



Recordando que la tabla de distribución normal estándar del Anexo A da el área bajo la curva desde 0 hasta z, se busca el valor de z tal que dicha área sea de $0{,}5 - 0{,}05 = 0{,}45$. El valor de z buscado, resulta:

Z	**0.00**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616

Como hay dos valores que están igualmente próximos a la probabilidad buscada, y son 1.64 y 1.65, tomamos entonces el valor de z que se encuentra entre los dos, es decir: $z = 1{,}6450$

$$\left(0{,}025 - 1{,}645\sqrt{\frac{0{,}025 \times 0{,}975}{400}}; 0{,}025 + 1{,}645\sqrt{\frac{0{,}025 \times 0{,}975}{400}}\right) \approx (0{,}012; 0{,}038)$$

Este intervalo contiene el porcentaje de artículos defectuosos, con una confianza del 90 %

También puede realizarse una estimación sobre la diferencia existente entre medias de distintas poblaciones. Para ello es necesario obtener una muestra de cada población, calcular el estimador de la media muestral de ambas y armar la diferencia.

Ejemplo 3: *Se prueban dos fórmulas diferentes de un combustible oxigenado para motor en cuanto a octanaje. La dispersión para la Fórmula 1 es* $\sigma_1 = 1{,}5$ *y para la Fórmula 2 es* $\sigma_1 = 1{,}6$. *Se prueban dos muestras aleatorias de tamaño* $n_1 = n_2 = 40$. *Los octanajes promedios observados son* $\bar{x}_1 = 89{,}6$ *y* $\bar{x}_2 = 92{,}5$. *Construir un intervalo de confianza del* 95 % *para la diferencia en el octanaje promedio*

El intervalo de confianza para la diferencia entre medias está dado por:

$$\left(\bar{x}_1 - \bar{x}_2 - z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_1^2}{n_2}}; \bar{x}_1 - \bar{x}_2 + z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_1^2}{n_2}}\right)$$

También se puede escribir:

$$\left(\bar{x}_1 - \bar{x}_2 \mp z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_1^2}{n_2}}\right)$$

Se pide un intervalo de confianza de $1 - \alpha = 0{,}95 \Rightarrow \alpha = 0{,}05 \Rightarrow \frac{\alpha}{2} = 0{,}025$ Donde $z_{\frac{\alpha}{2}}$ es el valor de z que deja un área de $\frac{\alpha}{2}$ a la derecha. Se busca en la tabla de distribución normal estándar, el valor de z que deja un área de $0{,}5 - 0{,}05 = 0{,}475$ desde el centro hasta z. El valor buscado es 1,96.

Z	**0.00**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850

Entonces:

$$\left(89{,}6 - 92{,}5 \mp 1{,}96\sqrt{\frac{1{,}5^2}{40} + \frac{1{,}6^2}{40}}\right) \approx (2{,}9 \mp 0{,}347) \approx (2{,}55; 3{,}25)$$

Este intervalo contiene la diferencia de octanaje promedio, con una confianza del 95 %

6.4. Estimación por máxima verosimilitud

Hasta aquí hemos trabajado con los estimadores de parámetros poblacionales haciéndolos surgir por analogía con las expresiones de cálculo de las medidas de tendencia central y variabilidad analizadas en el capítulo 5. En realidad las fórmulas para estos estadísticos se obtienen teóricamente de distintas maneras. El método de momentos de K. Pearson[4] se suele emplear por la sencillez de los cálculos que involucra. Sin embargo los estimadores que se obtienen a partir del método de *máxima verosimilitud* desarrollado por R. A. Fischer suelen ser más eficientes. A continuación veremos en que consiste tal método.

Sea $x_1, x_2, \ldots, x_n$ una muestra de n valores de una variable aleatoria X. Como claramente cada valor x_i es uno de los posibles de la variable X con una cierta probabilidad de ocurrencia, puede pensarse que la muestra se compone con cada uno de los valores que han tomado n variables aleatorias $X_1, X_2, \ldots, X_n$. Se supone además que cada elección de un valor muestral x_i se ha realizado en forma independiente de todas las otras, de forma tal que las variables $X_1, X_2, \ldots, X_n$ son independientes.

Supongamos ahora que la distribución de probabilidades de cada X_i, que en definitiva es la de X en los n casos, depende de un cierto parámetro θ

[4] Véase por ejemplo Paul Meyer, ob.cit., p. 217.

que interviene en su fórmula como interviene, por ejemplo, la media μ en la normal, la proporción p en la binomial o el parámetro α en la exponencial. Entonces podemos armar la distribución de probabilidad conjunta de las n variables que dependen todas del parámetro θ por medio de la expresión:

$$L(X_1, X_2, \ldots, X_n, \theta) = f(X_1, \theta) f(X_2, \theta) \ldots f(X_n, \theta)$$

Aquí cada $f(X_i, \theta)$ es la función de densidad de cada variable aleatoria X_i y sus expresiones se multiplican pues estas variables se suponen independientes.

En particular la probabilidad de la muestra extraída $x_1, x_2, \ldots x_n$ es $L(x_1, x_2, \ldots, x_n, \theta)$ donde cada x_i es un valor concreto de la variable X_i. Si se contara con dos estimaciones numéricas distintas del parámetro θ, llamadas θ_1 y θ_2 cabría esperar que un valor de probabilidad fuese mayor que otro. Es decir, por ejemplo, $L(x_1, x_2, \ldots, x_n, \theta_1) < L(x_1, x_2, \ldots, x_n, \theta_2)$.

Recordemos que L es un valor de probabilidad de una muestra y por lo tanto, si vale la desigualdad de arriba, es más probable obtener la muestra $x_1, x_2, \ldots, x_n$ cuando el parámetro es θ_2 que cuando es θ_1. Dado que la muestra $x_1, x_2, \ldots, x_n$ efectivamente se ha obtenido, resulta θ_2 más *verosímil* que θ_1 pues es más probable. El método de *máxima verosimilitud* consiste entonces en hallar el valor de θ que, dada la muestra obtenida, hace máxima la probabilidad L.

Nuestro objetivo es ahora hallar la expresión de θ que maximiza L para lo cual comenzamos por aplicar logaritmo natural a efecto de transformar los productos en sumas y luego, según las técnicas del cálculo infinitesimal, resolver la llamada *ecuación de verosimilitud.*

$$\frac{\partial}{\partial \theta} ln\left[L\left(x_1, x_2, \ldots, x_n, \theta\right)\right] = 0$$

para hallar los puntos críticos candidatos a maximizar el $ln\,[L]$. Obsérvese que siendo el logaritmo natural una función estrictamente creciente, el valor de θ crítico que lo maximice también maximizará a la probabilidad L.

Ejemplo 4: *Obtener la expresión de un estimador de la proporción de artículos defectuosos para el total de artículos producidos por una fábrica.*

La proporción p poblacional es desconocida y se la desea estimar a partir de los datos de una muestra de n artículos. La muestra esta integrada por n observaciones de una variable aleatoria X que vale 1 si el artículo es defectuoso y 0 si no lo es. De tal forma ocurre que cada variable X_i vale 1

o 0 en los respectivos casos. La probabilidad con que adopte estos valores puede escribirse:

$$P(X_i = 0) = p^0(1-p)^1 = 1-p$$

$$P(X_i = 1) = p^1(1-p)^0 = p$$

Dados los n valores de la muestra $x_1, x_2, \ldots, x_n$, habrá una cantidad k de ellos que valen 1 porque corresponden a artículos defectuosos y una cantidad $n-k$ que valen 0 ya que representan a artículos no defectuosos. Dados estos valores, la cantidad de artículos defectuosos en la muestra es $k = \sum_{i=1}^{n} x_i$ y la probabilidad de que aparezca ese número de defectuosos resulta $L(x_1, x_2, \ldots, x_n, p) = p^k(1-p)^{n-k}$. Obsérvese que los valores $x_1, x_2, \ldots, x_n$ son de la muestra y ya se cuenta con ellos sean 0 ó 1 en cada caso, por lo que no corresponde calcular todas las formas posibles en que puedan aparecer listados una cantidad k de unos y una cantidad $n-k$ de ceros, sino solamente considerar el orden en que fueron elegidos. Aplicando ahora el logaritmo natural se tiene

$$ln\left[L(x_1, x_2, \ldots, x_n, p)\right] = ln\left[p^k(1-p)^{n-k}\right] = k\,lnp + (n-k)ln(1-p)$$

Al derivar

$$\frac{\partial}{\partial p} ln\left[L(x_1, x_2, \ldots, x_n, p)\right] = \frac{k}{p} - \frac{n-k}{1-p} = 0$$

Operando se obtiene

$$\frac{k}{p} = \frac{n-k}{1-p}$$

$$(1-p)k = (n-k)p$$

$$k - pk = np - pk$$

y resulta $\hat{p} = \frac{k}{n}$ que maximiza la *función de verosimilitud* $L(x_1, x_2, \ldots, x_n, p) = p^k(1-p)^{n-k}$ pues expresa el único punto crítico de la función logarítmica. Téngase en cuenta que ésta es estrictamente creciente y además tiende a $-\infty$ cuando la probabilidad L tiende a 0. Es decir que para el caso de la muestra obtenida el *valor máximo verosímil* de la proporción poblacional p viene dado por la expresión del estimador $\hat{p} = \frac{k}{n}$ que se calcula a partir de ella.

Este cálculo del estimador, cuya fórmula podría sugerirse en este caso por analogía con la manera en que se obtendría la proporción poblacional

si se conociera su tamaño y la cantidad de artículos defectuosos, resulta así justificado por el razonamiento matemático. En general, los estimadores de máxima verosimilitud son convergentes en probabilidad a los parámetros que estiman lo quiere decir que, en la medida en que aumenta el tamaño muestral, la probabilidad de que el valor del estimador resulte el del parámetro tiende a 1. Por otra parte, si bien pueden resultar sesgados, en muchos casos el sesgo se corrige en forma simple multiplicando por una constante[5].

6.5. Ejercicios

Ejercicio N°1*: La duración de una pieza de un equipo es una variable aleatoria normal con una dispersión de 4 horas y una media que se desea estimar. Una muestra aleatoria de 100 piezas que fueron probadas produjo una media muestral de 501,2 hs. Obtenga un intervalo de confianza para la media con un nivel de confianza de:

a) 95 %

b) 99 %

Ejercicio N°2*: La densidad de un producto químico tiene una distribución normal con una dispersión de 0.005 g/cm^3. ¿Cuál debe ser el tamaño de la muestra, como mínimo, para que al estimar la densidad media con un intervalo de confianza del 95 % el error resulte menor que 0.002 g/cm^3?

Ejercicio N°3*: Se quiere estimar el peso medio de una producción de tubos de hormigón. Mediciones previas permitieron saber que la dispersión es de 2.4 kg. Se toma una muestra de tamaño 100 ¿con qué nivel de confianza se puede asegurar que el peso medio muestral no difiere del poblacional en mas de 0.8 kg?

Ejercicio N°4*: La resistencia eléctrica de ciertos cables tiene una distribución normal. Para estimar la resistencia media se efectúan 16 mediciones obteniendo un promedio de 10.48 Ω con una desviación estándar de 1.36 Ω. Halle un intervalo de confianza para la media poblacional ($\alpha = 0{,}05$).

[5]Para un análisis más extenso de las propiedades de los estimadores de máxima verosimilitud puede consultarse George Canavos, ob.cit., pp. 264-268.

Ejercicio N°5*: La dispersión muestral de una muestra de 30 lámparas es de 100 horas. Halle un intervalo de confianza para la dispersión poblacional con: (Suponer distribución normal)

a) $\alpha = 0{,}05$

b) $\alpha = 0{,}01$

Ejercicio N°6: En una muestra aleatoria de 1500 teléfonos residenciales tomada en una ciudad se encontró que 387 números no aparecían en la guía telefónica. Encuentre un intervalo de confianza del 90 % para el porcentaje de números que no aparecen en la guía.

Ejercicio N°7: Se sabe que la cantidad de días de lluvia que se producen en una zona desértica sigue una distribución de Poisson. Este año se han tenido 2 días de lluvia. Se desea utilizar esta información para estimar el parámetro λ correspondiente a la distribución de la cantidad de días de lluvia.

Ejercicio N°8: En un experimento binomial se observan 3 éxitos en 6 ensayos. Obtener el estimador de máxima verosimilitud de la probabilidad de éxito en cada ensayo.

Capítulo 7

Test de hipótesis

7.1. Introducción

En el marco de la toma de decisiones puede resultar necesario evaluar hipótesis acerca de parámetros poblacionales. Por ejemplo; de acuerdo a las especificaciones del fabricante cada soporte para ciertos equipos resiste un promedio de 500 kg de peso y se quiere comprobar si esto realmente es así a efecto de ordenar la compra de un gran lote de los mismos. Con tal finalidad se puede tomar una muestra de una pequeña cantidad de soportes y someterlos a prueba de resistencia para obtener una resistencia promedio y compararla con la declarada. Si el promedio de resistencia de los soportes de la muestra resultara suficientemente próximo al especificado por el fabricante para todos los soportes, se decidirá la compra del lote completo. En caso contrario ésta no se realizará. Es claro que tal prueba estadística requiere, para ser sólida, una medida adecuada de proximidad entre ambos promedios y apreciar además la probabilidad de cometer un error en nuestra decisión para darnos la posibilidad de minimizarlo en cuanto sea posible.

El procedimiento que realiza este análisis se conoce con el nombre de *Test de hipótesis* pues, precisamente, coloca el valor de un parámetro poblacional, en el ejemplo la especificación del fabricante, como una hipótesis cuya veracidad se probará a partir de un ensayo con datos de una muestra. Con ayuda de la distribución del estadístico correspondiente se podrá además establecer un criterio adecuado de aceptación o rechazo de la hipótesis y las probabilidades de equivocación.

En ocasiones puede ser necesario establecer si una distribución empírica de datos se parece a una determinada distribución teórica. Por ejemplo, medidos los pesos del conjunto de pacientes de un hospital, se desea saber

si siguen una distribución normal. Como veremos, también en estos casos un test de hipótesis permite evaluar si el conjunto de datos obtenidos sigue aproximadamente la fórmula de la distribución de frecuencias teórica supuesta, es decir si los datos según sus frecuencias se *ajustan* a ella.

Se presenta el método a continuación.

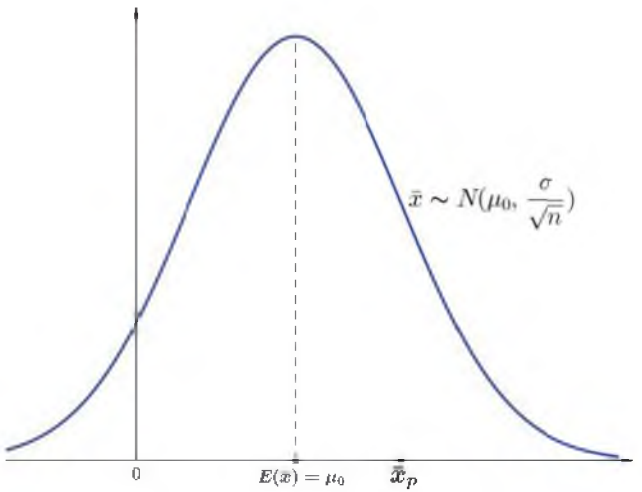
7.2. Test de hipótesis

La realización de la prueba involucra una suposición sobre el valor o el conjunto de valores posibles de una cantidad. Esta presunción se denomina, por razones que más adelante quedarán en claro, *hipótesis nula* y se denotará con H_0. La existencia de la hipótesis nula ofrece la alternativa de otra hipótesis que constituye su negación complementaria llamada *hipótesis alterna* y nombrada como H_a. Así, por ejemplo, si el test se efectuara sobre una media poblacional μ que se supone igual a cierto valor μ_0 las hipótesis nula y alterna quedarían:

1. $H_0 : \mu = \mu_0$
2. $H_a : \mu \neq \mu_0$

Obsérvese que estas hipótesis se realizan sobre el parámetro poblacional y que, si se considera el promedio $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ obtenido a partir de cualquier muestra de tamaño n tomada para testearlo, el valor esperado del mismo debiera ser μ_0 siempre que sea válida la hipótesis nula. Según se vio en la sección 2 del capítulo 6, $\bar{x}$ tiene una distribución que tiende a ser normal para n suficientemente grande, en este caso con $E(\bar{x}) = \mu_0$ y $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$. Si $\bar{x}_p$ es el valor calculado del estadístico sobre la muestra efectivamente seleccionada, la figura 1 ilustra la diferencia que éste puede tener con respecto a la cantidad μ_0 tomada como hipótesis nula. La proximidad nos indicaría que la hipótesis realizada es plausible mientras que si $\bar{x}_p$ resultara en el gráfico muy lejano a μ_0, sería razonable dudar sobre la certeza de la hipótesis adoptada acerca de la media poblacional μ. Como por tales razonamientos se evalúa en definitiva lo acertado de la suposición $\mu = \mu_0$, $\bar{x}_p$ se denomina *estadístico de prueba* de la hipótesis.

Figura 1.



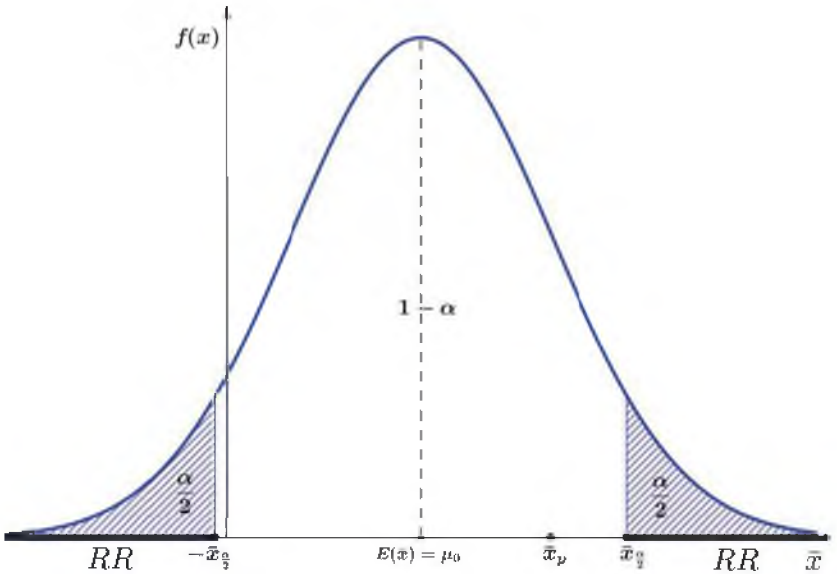
Ya se ha señalado antes que la concentración de los promedios muestrales alrededor de la media μ_0 está en relación con el valor del desvío estándar $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$. Como se ve, éste a su vez, no solo depende del tamaño de las muestras sino también del desvío estándar poblacional que supondremos por ahora conocido. Es claro que cualquier variación del valor de σ traería aparejado un cambio en la gráfica concentrándola más o menos alrededor del eje de simetría según decreciera o creciera esa cantidad.

El tercer elemento constitutivo del test es entonces el conjunto formado por la muestra, con sus respectivos tamaño y estadístico de prueba, y el desvío estándar, en principio conocido, de la población.

3. *Muestra: tamaño n y estadístico de prueba $\bar{x}$. Desvío estándar poblacional σ*

Para aceptar o rechazar la hipótesis nula realizada hace falta además fijar un criterio que establezca cuando el valor de prueba $\bar{x}_p$ está suficientemente próximo a μ_0 y cuando resulta demasiado lejano. Este criterio viene dado por el cuarto elemento que posibilita la realización del test que es el llamado *nivel de significancia*. Se trata en realidad de un valor de probabilidad α que se fija de antemano suficientemente pequeño como para que se corresponda solo con la suma de la probabilidad acumulada en las colas de la distribución. Cada cola se corresponderá con una probabilidad acumulada desde un valor $\pm\bar{x}_{\frac{\alpha}{2}}$ hacia $\pm\infty$ respectivamente como se muestra en la figura 2. Así, α es la cuarta componente a tener en cuenta para el test.

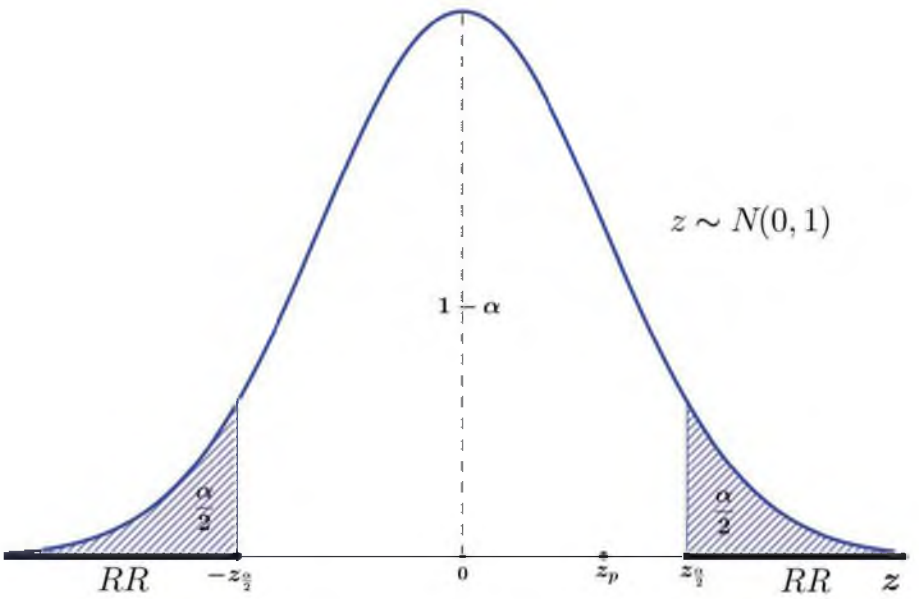
4. *Nivel de significación* α
Figura 2.



Los valores $\pm\bar{x}_{\frac{\alpha}{2}}$ se denominan *críticos*. Si el estadístico de prueba, obtenido de la muestra, cae dentro del intervalo $\left(-\bar{x}_{\frac{\alpha}{2}}, \bar{x}_{\frac{\alpha}{2}}\right)$ se considerará que es lo suficientemente próximo a μ_0 como para aceptar que ése es el valor de la media poblacional. Tal intervalo se denomina *región de aceptación* (RA) de la hipótesis nula. Si por el contrario $\bar{x}_p$ cae fuera de esta región, dentro de la llamada *región de rechazo* (RR) resaltada con color negro en la figura 2, la hipótesis nula deberá rechazarse por ser el valor obtenido del promedio muestral demasiado lejano al valor supuesto de la media poblacional. En esencia, este es el mecanismo del test de hipótesis. Sin embargo todavía hay que agregar, para el caso particular de la media poblacional que estamos considerando, que una vez fijada la probabilidad α, la determinación de los valores críticos $\pm\bar{x}_{\frac{\alpha}{2}}$ debe realizarse utilizando la tabla de la normal estándar de acuerdo a las probabilidades $\frac{\alpha}{2}$ de las colas. De tal forma se pueden establecer los valores críticos estandarizados $z_{\frac{\alpha}{2}}$ y resulta entonces más cómodo efectuar el test directamente sobre esta normal de media 0 y desvío estándar 1. Así el estadístico de prueba se calcula como $z_p = \frac{\bar{x}_p - \mu_0}{\frac{\sigma}{\sqrt{n}}}$. También, $\pm z_{\frac{\alpha}{2}}$ son los correspondientes transformados de $\pm\bar{x}_{\frac{\alpha}{2}}$ que señalan

el comienzo de cada una de las colas y establecen las regiones de aceptación y rechazo para la hipótesis nula original. La figura 3 exhibe esta situación.

Figura 3.



En forma general y resumida podemos apuntar entonces que el test de hipótesis requiere constituir los siguientes elementos:

1. *Hipótesis nula* H_0
2. *Hipótesis alterna* H_a
3. *Muestra de tamaño* n *y estadístico de prueba.*
4. *Nivel de significación* α *y regiones de aceptación y rechazo de la hipótesis nula.*

Una vez que se han establecido estos aspectos hay que analizar en que región cae el estadístico de prueba para aceptar o rechazar la hipótesis nula.

Ejemplo 1: *Se analiza una partida de combustible sólido. Una de las características importantes de este producto es la rapidez de combustión medida en centímetros por segundo. Para que el combustible sea apto se requiere*

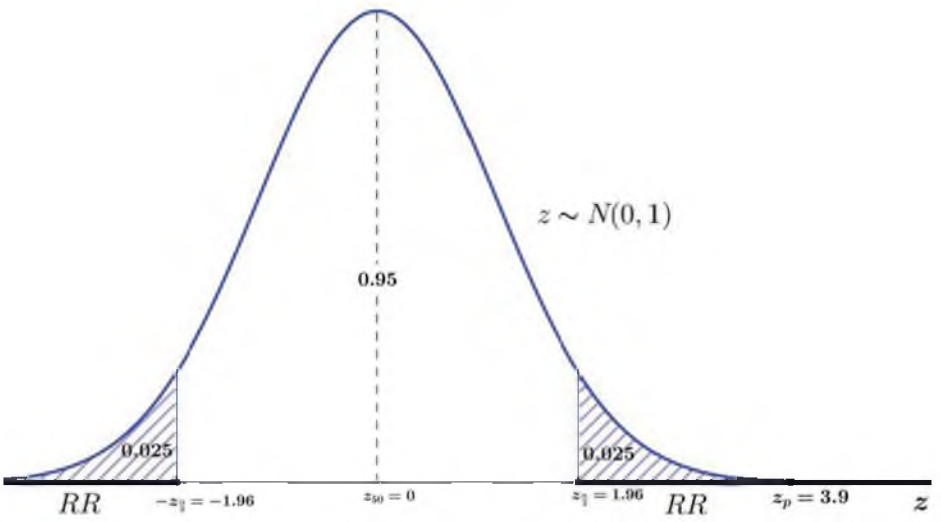
que la rapidez promedio de combustión sea de 50cm/s. Si fuera muy superior se consumiría innecesariamente demasiado combustible; en tanto que si fuera inferior, podría disminuir sensiblemente la eficiencia de los sistemas que provee. En cualquier caso, se sabe que la desviación estándar de esta rapidez es de 2cm/s. Se desea realizar una prueba de hipótesis para establecer si la partida de combustible sólido cumple con la especificación de rapidez de combustión requerida. A tal fin se selecciona una muestra aleatoria de 36 unidades de combustible y se obtiene una rapidez promedio muestral de combustión de 51,3cm/s. Se fija también un nivel de significancia del 5 %.

La media poblacional a testear es, de acuerdo al enunciado,$\mu_0 = 50cm/s$. El desvío estándar poblacional resulta conocido: $\sigma = 2cm/s$ mientras que el valor de prueba del estadístico es $\bar{x}_p = 51{,}3cm/s$. El nivel de significancia es el 5 % de probabilidad, o sea $\alpha = 0{,}05$ y dado que la rapidez promedio no puede ser ni mucho mayor ni mucho menor que la de la hipótesis nula, habrá que considerar el rechazo si $\bar{x}_p$ se ubica lejos de su valor, en cualquiera de las dos colas de probabilidad. Se tiene:

1. *Hipótesis nula* $H_0 : \mu = 50$
2. *Hipótesis alterna* $H_a : \mu \neq 50$
3. *Muestra de tamaño* $n = 36$, $\sigma = 6$ *y estadístico de prueba* $z_p = \frac{51{,}3-50}{\frac{2}{\sqrt{36}}} = 3{,}9$
4. *Nivel de significación* $\alpha = 0{,}05$. *La región de rechazo de la hipótesis nula está ubicada en las colas de la distribución que contienen 0.025 de probabilidad cada una y establecida a partir de los valores críticos* $\frac{\alpha}{2} = \pm 1{,}96$ *que surgen de la tabla de la normal estándar al considerar* $P\left(Z \leq -z_{\frac{\alpha}{2}}\right) = 0{,}025$ *y* $P\left(Z \leq z_{\frac{\alpha}{2}}\right) = 0{,}975$ *respectivamente.*

En la figura 4 está volcada toda esta información.

Figura 4.



Como se observa en la figura 4, el estadístico de prueba cae en la región de rechazo por lo que corresponde rechazar la hipótesis nula y concluir que la partida de combustible sólido analizada no cumple con la característica de rapidez de combustión requerida.

En el problema planteado en el ejemplo 1 se ha visto la necesidad de rechazar la hipótesis nula tanto cuando el valor del promedio muestral fuese mucho mayor como cuando resultara mucho menor que el supuesto. Pero esta alternativa no siempre se da. A veces, si el estadístico de prueba resulta sensiblemente mayor que la hipótesis nula, ésta debe aceptarse. En otras ocasiones al ser muy menor que la hipótesis nula no se la rechaza. En realidad el criterio de rechazo a adoptar depende de las características del problema que influyen sobre la forma de las regiones de rechazo y aceptación. Otra cuestión que puede presentarse es que no se conozca de antemano el desvío estándar poblacional en cuyo caso solo queda estimarlo mediante el estadístico s calculado a partir de la muestra. En tal situación, como vimos en el capítulo 6, el estimador de la media $\bar{x}$ se distribuirá según *t-student* con $n-1$ grados de libertad si la población tiene una distribución aproximadamente normal. Lo propio ocurrirá, aun cuando se conozca el desvío estándar poblacional, si la muestra fuese "pequeña", es decir si $n < 30$, siempre bajo el supuesto de distribución normal de la población. Consideraremos estas cuestiones en el siguiente ejemplo.

Ejemplo 2: *Una compañía que procesa fibras naturales afirma que sus fibras tienen una resistencia a la ruptura de 20 kg. Un posible comprador sospecha que esta resistencia es menor y decide realizar un test de hipótesis al respecto con un nivel de significancia del 1 %. Selecciona una muestra aleatoria de 25 fibras y obtiene para ellas un promedio de 19 kg con una desviación estándar muestral de 3 kg.*

El primer punto que hay que destacar es que si la media muestral resultara mayor que 20 kg, y aún mucho mayor que esto, no habría razones para rechazar la hipótesis de la compañía pues sería razonablemente seguro que la resistencia de las fibras de toda la población superaría a esa cantidad declarada y que, de así ocurrir, esto más que un perjuicio constituiría una ventaja. El problema se presenta entonces cuando, como en este caso, la media muestral resulta menor o a lo sumo igual que la cantidad 20. ¿Hasta que punto puede reducirse la resistencia muestral sin dudar de la especificación dada por la compañía? Tal punto quedará establecido por el nivel de significancia que ahora determinará, según el razonamiento apuntado, la región de rechazo solo en la cola correspondiente al lado izquierdo de la normal estándar. Dicha región corresponderá a los valores de la media muestral de resistencia que, a ése nivel de significación, son suficientemente menores al valor de 20 kg, asegurado para la población de fibras, como para rechazarlo. De acuerdo a esto la hipótesis nula deberá ser $\mu \geq 20$.

Otro aspecto importante a tener en cuenta es que no se tiene el desvío estándar poblacional que debe aproximarse por $s = 3$ calculado a partir de la muestra. Por otra parte esta cuenta con sólo 25 mediciones. Ambas cuestiones llevan a tener que suponer un comportamiento aproximadamente normal de la resistencia de todas las fibras para poder utilizar entonces la distribución *t-student* del estadístico $\bar{x}$.

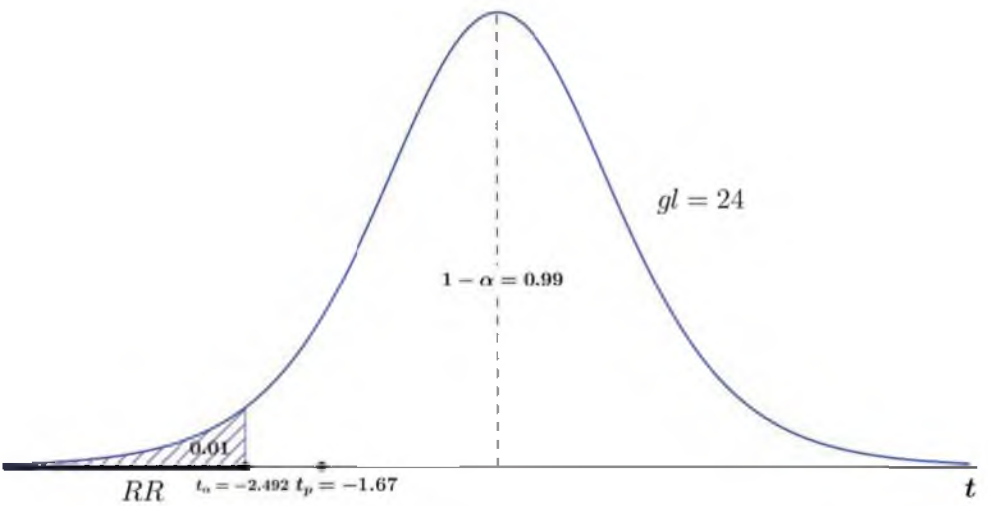
Estos análisis llevan a plantear el test de hipótesis como sigue:

1. *Hipótesis nula* $H_0 : \mu \geq 20$
2. *Hipótesis alterna* $H_a : \mu < 20$
3. *Muestra de tamaño* $n = 25$, $s = 3$ *y estadístico de prueba* $t_p = \frac{19-20}{\frac{3}{\sqrt{25}}} = -1{,}67$
4. *Nivel de significación* $\alpha = 0{,}01$. *Se supone distribución normal de la población a efecto de utilizar la distribución t-student con* $gl = 25-1 = 24$ *grados de libertad del estimador de la media poblacional. La región*

de rechazo de la hipótesis nula está ubicada en la cola izquierda de la distribución que contiene 0.01 de probabilidad, establecida a partir del valor crítico $t_\alpha = -2{,}492$ *Este surge de la tabla de la t-student al considerar* $P(t \leq t_\alpha) = 1 - P(t \leq t_{1-\alpha}) = 1 - 0{,}99 = 0{,}01$.

Los valores se expresan en la figura 5:

Figura 5.



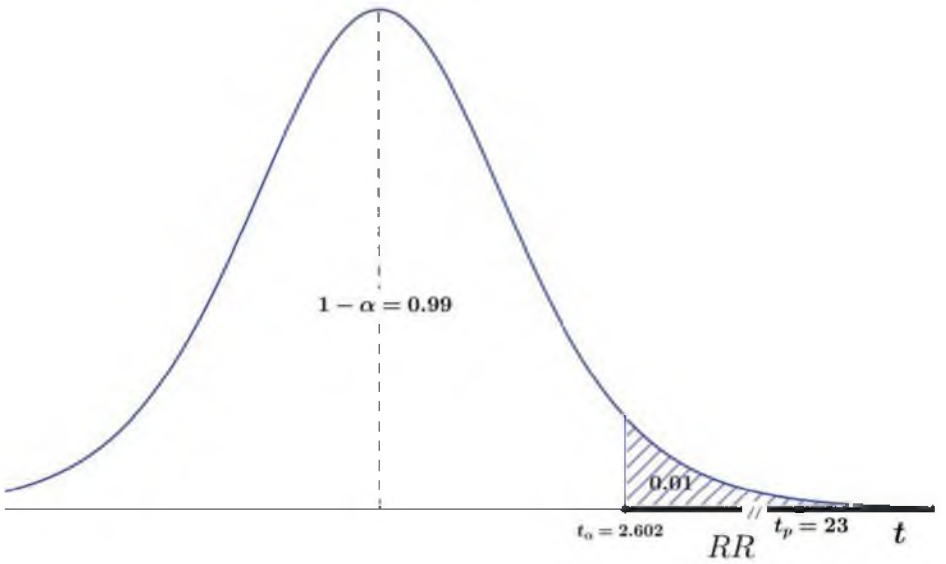
Como se ve, el estadístico de prueba cae dentro de la región de aceptación, lo que quiere decir que éste no es lo suficientemente menor que la hipótesis nula como para que debamos rechazarla. Corresponde, entonces no rechazar la hipótesis nula, es decir no descartar la afirmación de la compañía respecto de la resistencia media de sus fibras.

7.3. La probabilidad de error

Si se lee con atención se observa que en el Ejercicio 1 hemos rechazado sin dudas la hipótesis nula mientras que en el Ejercicio 2 no la hemos rechazado pero tampoco hemos afirmado tan enfáticamente que debíamos aceptarla, por lo menos no con el mismo énfasis que expresamos el rechazo en el otro caso. ¿A qué se debe esta sutil diferencia? ¿Tiene el mismo significado estadístico no rechazar la hipótesis nula que aceptarla con seguridad? Estas preguntas nos llevan, en primer término, a la consideración de los errores que pueden cometerse al realizar un test de hipótesis.

Cuando una hipótesis se rechaza o se acepta puede muy bien ocurrir que nos estemos equivocando. Que la rechacemos no quiere decir que sea falsa sin ninguna duda pues subsiste, aunque seguramente con baja probabilidad, la posibilidad que hayamos tomado una muestra muy particular que nos lleve a error. Por ejemplo, si se quisiera probar la hipótesis de que la altura promedio de los hombres adultos que viven en la ciudad de Córdoba es menor o igual que 1.71 m y se supusiera que tales alturas tienen una distribución normal con un desvío estándar conocido de 0.04 m, podría tomarse una muestra de 16 hombres del lugar para testearla realizando la prueba sobre una distribución t-student con 15 grados de libertad. Supongamos que en un hecho poco habitual, pero que sin embargo tiene una cierta probabilidad de ocurrencia, sucede que los 16 hombres seleccionados al azar para integrar la muestra resultan, casualmente, los 16 jugadores, titulares y suplentes, del primer equipo del campeón de básquet Atenas de Córdoba, cuyo promedio de altura es 1.94 m. Según este promedio muestral debemos rechazar la hipótesis de que los hombres cordobeses miden en promedio, a lo sumo, 1.71m. Esto es así porque claramente el 1.94 caerá dentro de la región de rechazo. En efecto, si $\bar{x}_p$ y $\sigma_{\bar{x}} = \frac{0{,}04}{\sqrt{16}} = 0{,}01$ al elegir, por ejemplo, un nivel de significancia del 1 % el test a realizar sobre la t-student con 15 grados de libertad tiene la forma exhibida en la figura 6 siendo el valor de prueba $t_p = \frac{1{,}94-1{,}71}{0{,}01} = 23$.

Figura 6.

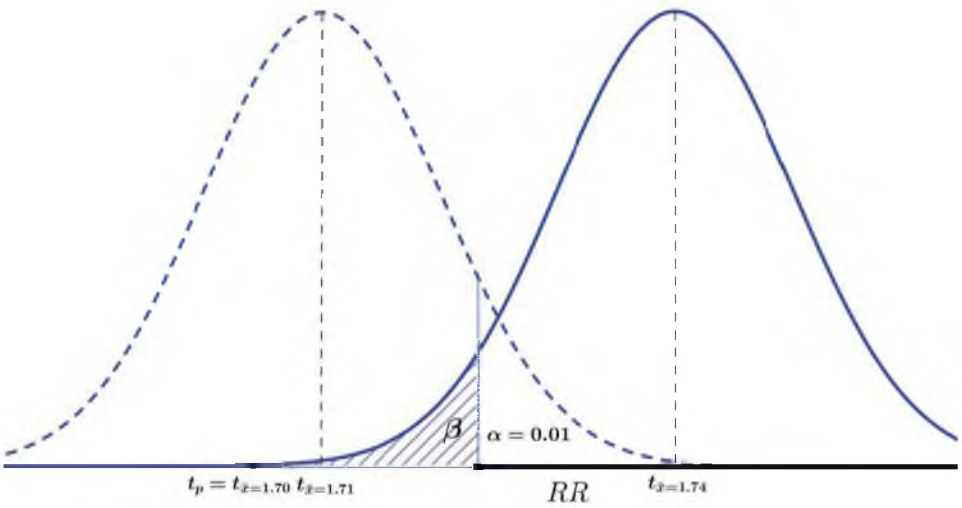


Obsérvese de paso que no solo el valor del promedio muestral $\bar{x}_p = 1{,}94$ obliga al rechazo de la hipótesis nula sino también cualquier otro valor de $\bar{x}$ que arroje un estadístico de prueba t_p mayor que $t_\alpha = 2{,}602$. Esto significa que habría una probabilidad $\alpha = 0{,}01$ de equivocarse cuando la hipótesis del promedio poblacional con $\mu_0 = 1{,}71$ fuera verdadera. Se ve entonces que la hipótesis nula puede ser verdadera aunque la rechacemos porque siempre tenemos una probabilidad, pequeña pero existente al fin, de equivocarnos. Cuando sucede tal cosa se comete un *error de tipo I* y la pequeña probabilidad de que esto ocurra es el nivel de significancia elegido para el test.

De manera parecida podría ocurrir que al aceptar una hipótesis nula en realidad nos estuviéramos equivocando porque una muestra poco probable nos indujera a error. En el mismo ejemplo de las alturas de los hombres cordobeses supongamos que la hipótesis nula $\mu \leq 1{,}71$ fuera en realidad falsa. Supongamos también que la verdadera altura promedio, para nosotros desconocida, fuera $\mu \leq 1{,}74$. Como consecuencia habría una distribución real del estimador $\bar{x}$ cuya media es 1.74 y una errónea, proveniente de la hipótesis nula con media 1.71. Sobre esta última se realizaría el test tomando, por ejemplo, un nivel de significancia $\alpha = 0{,}01$. Supongamos ahora que la muestra para testear nuestra hipótesis $\mu \leq 1{,}71$ tiene un promedio de

alturas $\bar{x}_p = 1{,}70$ siendo el ya conocido desvío estándar poblacional 0.04. En estas circunstancias debiéramos aceptar la hipótesis nula pues el estadístico de prueba no cae dentro de la región de rechazo del test. Ahora bien; no cometeríamos error sólo si rechazáramos la hipótesis nula y esto ocurriría cuando los valores de prueba t_p resultaran mayores que el valor crítico t_α. La probabilidad de cometer un error hay que medirla entonces sobre la distribución real del estadístico y no sobre la determinada por la hipótesis nula falsa. En la figura 7 la distribución real del estadístico se representa con línea llena y la errónea con línea punteada. El área β corresponde a la región de aceptación que erróneamente fue establecida y por lo tanto representa la probabilidad, medida sobre la verdadera distribución de los promedios muestrales, de aceptar la hipótesis nula cuando ésta es falsa.

Figura 7.



Se dice que al aceptar la hipótesis nula siendo en realidad falsa se comete un *error de tipo II*. La probabilidad de que tal cosa ocurra es, precisamente, β. De la figura 7 surge que tal probabilidad habrá de crecer cuando disminuya la probabilidad α de cometer un error de tipo I, que debe calcularse sobre la distribución erróneamente supuesta del estadístico. Sin embargo la forma funcional que adopta la relación entre α y β no resulta tan sencilla de establecer pues depende a su vez del valor real desconocido de μ[1]. A efectos

[1]Un análisis mas detallado exige definir una función de operación característica tal

prácticos daremos por establecido aquí que la alternativa para "achicar" la probabilidad β de un error de tipo II es agrandar el nivel de significancia α. Es claro además que se conoce tal valor, pues se lo establece para realizar el test, mientras que apreciar β requeriría análisis más profundos. Esto tiene a su vez una implicancia importante: es preferible plantear el test de modo de que sea la hipótesis nula aquella que se rechace pues se conoce directamente la probabilidad α de cometer un error en el caso de que fuera verdadera. En cambio, si se acepta la hipótesis nula, la probabilidad del error que puede cometerse debe ser estudiada con más detalle y por eso termina siendo preferible decir que no se la rechaza a aseverar con seguridad que se la acepta cuando ese detalle, en realidad, no se analiza. El nombre *hipótesis nula* con el que hemos venido trabajando hasta aquí surgió en el análisis comparativo de medias aritméticas del enunciado $\mu_1 - \mu_2 = 0$. Sin embargo, revela también la idea de que, en tanto sea posible y razonable, identifica a la hipótesis que habrá de ser rechazada.

En suma los errores posibles al realizar un test y sus respectivas probabilidades se esquematizan en la tabla 1:

Tabla 1.

Hipótesis nula H_0	Rechazo	Aceptación
Verdadera	Error Tipo I Probabilidad α	No se comete error
Falsa	No se comete error	Error Tipo II Probabilidad β

Ejemplo 3: *En una zona de la Patagonia se piensa instalar un parque de generación de energía eólica. Para que la zona sea apta para colocar los costosos molinos generadores, el viento promedio debe ser de no menos de 50 km/h. Para analizar si éste es el comportamiento del viento en la zona se realizan 36 mediciones en distintos momentos de la semana que arrojan un promedio de 47 km/h y se realiza un test de hipótesis sabiendo que por la condición patagónica de la zona el desvío estándar del viento es aproximadamente $\sigma = 12$. ¿Es la zona adecuada para instalar el parque?*

1. *Hipótesis nula* $H_0 : \mu \geq 50$
2. *Hipótesis alterna* $H_a : \mu < 50$

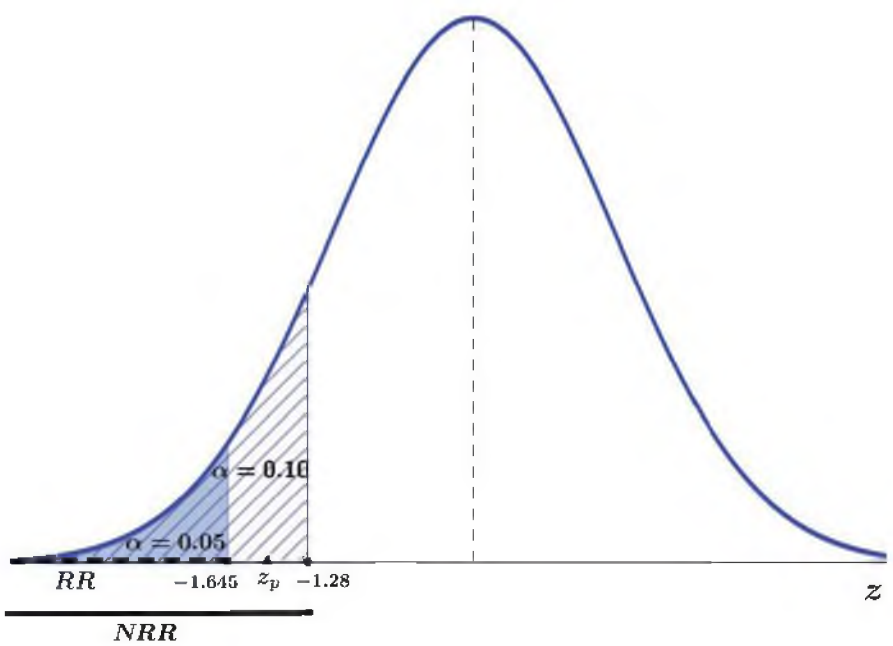
como se expone en Paul Meyer, ob.cit., pp. 326-333

3. *Muestra de tamaño* $n = 36$, $\sigma = 12$ *y estadístico de prueba* $z_p = \frac{47-50}{\frac{12}{\sqrt{36}}} = -1{,}5$

4. *Si se elige un nivel de significancia* $\alpha = 0{,}05$ *el valor crítico es* $z_\alpha = -1{,}645$ *que determina como región de aceptación el intervalo* $RA = [-1{,}645, \infty)$ *y como región de rechazo* $RR = (-\infty, -1{,}645)$. *De acuerdo a esto, no debiera rechazarse la hipótesis nula. Sin embargo, como resulta mucho más costoso instalar los molinos, y que luego no funcionen, que radicar el parque en otra zona, debe minimizarse la probabilidad de que no haya viento suficiente. Es decir, conviene hacer razonablemente pequeña la probabilidad de cometer un error de tipo II. Para ello se sube el nivel de significancia a* $\alpha = 0{,}10$ *con lo cual la nueva región de aceptación es* $NRA = [-1{,}28, \infty)$*y la nueva región de rechazo resulta* $NRR = (-\infty, -1{,}28)$. *En este caso el estadístico de prueba cae dentro de la región de rechazo por lo que se rechaza la hipótesis nula con una probabilidad 0.1 de equivocarse y que la región efectivamente fuera apta para la instalación de los molinos.*

En la figura 8 se esquematiza el razonamiento efectuado.

Figura 8.

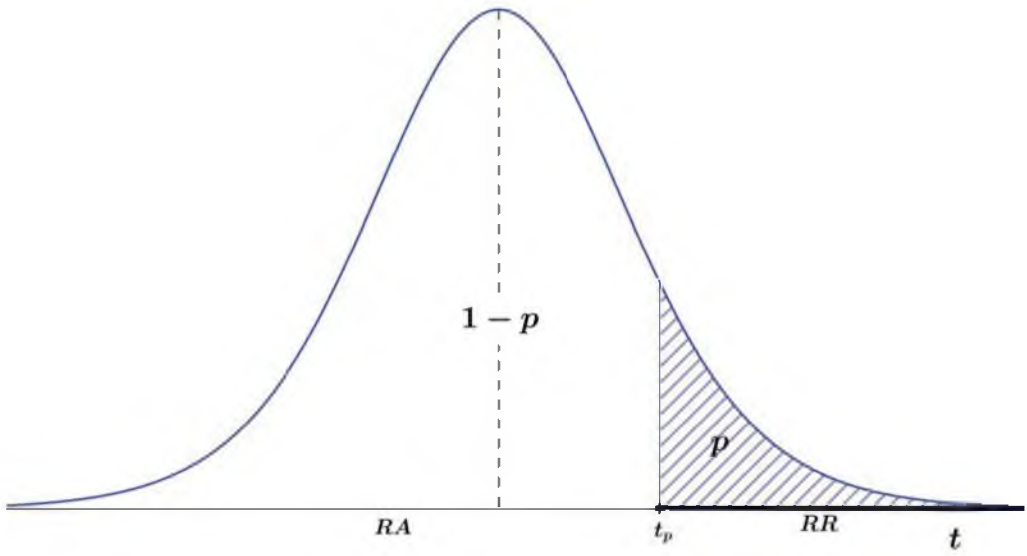


La línea gruesa punteada corresponde a la región de rechazo para el nivel de significancia 0.05. La línea continua es la de la nueva región de rechazo delimitada al considerar una probabilidad de error de tipo I de 0.10, mayor que la anterior. La posición del estadístico de prueba fuerza el rechazo de la hipótesis nula para tal nivel de significancia.

Un aspecto importante luego de todas las consideraciones efectuadas es el modo de proceder real empleado al hacer el test utilizando cualquier software de los disponibles. En primer lugar, cuando el tamaño muestral es pequeño, y más aún si se desconoce el desvío estándar poblacional, no queda más alternativa, para testear una hipótesis sobre una media poblacional, que utilizar la t-student suponiendo una distribución normal en la población. Pero dado que la distribución t-student se aproxima suficientemente a la normal cuando el tamaño muestral supera los 30 elementos y que para esta cantidad ya no hace falta suponer la normalidad en la distribución de la población ni conocer estrictamente la varianza, la mayoría de los paquetes tienen sólo en cuenta la t-student. Una segunda cuestión a observar es que a partir del valor calculado del estadístico de prueba t_p quedan determinadas dos regiones de probabilidades respectivas $1-p$ y p según se ve en la figura 9. De acuerdo a esto se puede pensar que el nivel de significancia α se fija directamente en el valor $\alpha = p$. Si en tal caso se rechaza la hipótesis nula, como si el estadístico de prueba hubiera caído dentro de la región de abscisas correspondiente al área p, la probabilidad de rechazar la hipótesis nula cuando es en realidad verdadera estaría dada por el *valor p*.

En suma, el software no requiere ni introducir ni calcular un valor del nivel de significancia, sino que dado el estadístico de prueba calcula el *valor p* respectivo y deja al usuario la decisión de aceptar o rechazar la hipótesis, caso este último en que, con probabilidad precisamente p, sabe que podrá cometer un error.

Figura 9.



7.4. El test para otros parámetros

Mediante el test de hipótesis puede también analizarse el comportamiento de otros parámetros. En cada caso habrá que tener en cuenta las características del estimador que se obtiene al partir de la muestra y su distribución.

En el caso de una proporción p como podría ser, por ejemplo, la proporción de mayores de 60 años en la población, habrá que considerar que el estimador $\hat{p}$, cuando n es suficientemente grande, tiene una distribución normal. Si la hipótesis nula fuera, por ejemplo: $H_0 : p_0$ se estaría suponiendo que la distribución normal del estimador tiene una media o valor esperado $E(\hat{p}) = p_0$ y que el desvío estándar está dado por la expresión $\sigma_{\hat{p}} = \sqrt{\frac{p_0(1-p_0)}{n}}$. De acuerdo a esto, una vez formulada complementariamente la hipótesis alterna, podría elegirse el nivel de significancia α y realizar el test con el valor de prueba obtenido a partir de la muestra de tamaño n

calculado como $z_p = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$

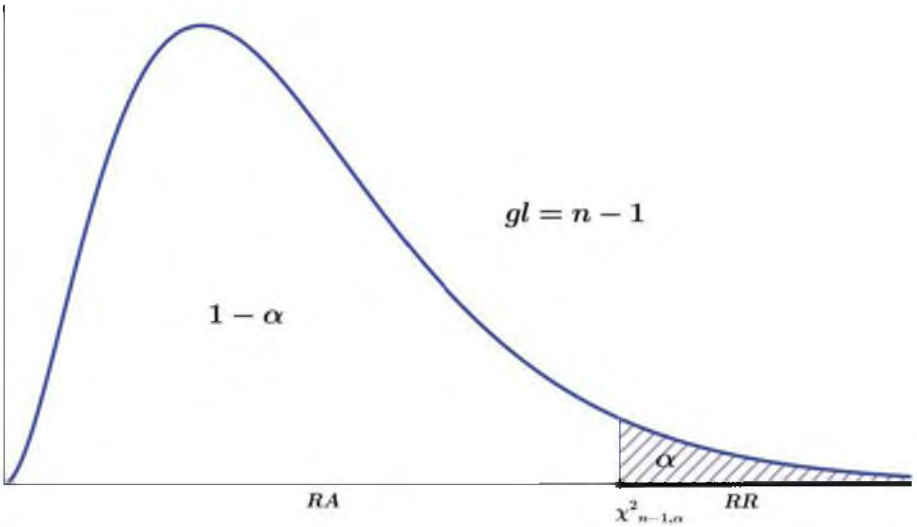
Si fuera necesario en cambio testear una hipótesis sobre el comportamiento de dos medias poblacionales, por ejemplo: $H_0 : \mu_1 \geq \mu_2$ deberá observarse que, en la medida que los tamaños n_1 y n_2 de las muestras tomadas sobre ambas poblaciones resulten suficientemente grandes, la distribución del estimador $\bar{x}_1 - \bar{x}_2$ será normal con valor esperado $E(\bar{x}_1 - \bar{x}_2) = \mu_1 - \mu_2$ y desvío estándar $\sigma_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$ siendo σ_1 y σ_2 los desvíos estándar de las respectivas poblaciones. De la hipótesis nula $\mu_1 \geq \mu_2$ surge en forma inmediata la equivalente $\mu_1 - \mu_2 \geq 0$ y entonces el test puede realizarse sobre la distribución del estadístico señalado. Para ello, una vez determinada la hipótesis alterna y fijado un nivel de significancia, se utiliza el valor de prueba:

$$\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Nótese que en el numerador del estadístico de prueba ha quedado sólo la resta $\bar{x}_1 - \bar{x}_2$ pues la cantidad que le fue restada para estandarizar es precisamente el 0 surgido de la hipótesis nula.

Cuando se trata de testear una hipótesis sobre la varianza poblacional la situación cambia un poco debido a que, bajo el supuesto de distribución normal de la variable poblacional, el estimador de la varianza $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ guarda la relación $\chi^2_{n-1} = \frac{(n-1)}{\sigma^2}s^2$ con la variable chi-cuadrado de $n-1$ grados de libertad, según quedó establecido por la Propiedad 1 del capítulo 6. De acuerdo a esto la hipótesis nula sobre la varianza poblacional, por ejemplo: $H_0 : \sigma^2 \leq \sigma_0^2$, debe testearse con el estadístico de prueba $\chi^2_{n-1,p} = \frac{(n-1)s^2}{\sigma_0^2}$ una vez fijado el nivel de significancia y establecida la hipótesis alterna. Para el caso ejemplificado, si el estadístico de prueba superara el valor crítico calculado de acuerdo al nivel de significancia α, debería rechazarse la hipótesis nula. Obsérvese si $\chi^2_{n-1,p}$ resultara mayor que el valor crítico $\chi^2_{n-1,\alpha}$ se tendría $\frac{(n-1)s^2}{\sigma_0^2} > \frac{(n-1)s^2}{\sigma^2}$ donde, dado que $n-1$ y s^2 provienen de la muestra y están fijos, debiera necesariamente ocurrir que $\sigma^2 > \sigma_0^2$ lo que contradeciría la hipótesis nula. De tal modo es claro que las regiones de aceptación y rechazo para la hipótesis ejemplificada serían las que se exhiben en la figura 10.

Figura 10.



Ejemplo 4: *Se desea comprobar que una balanza vial funciona correctamente de modo tal que el desvío estándar de sus mediciones de pesos no supere los 100 kg. A tal fin se toma una muestra de 25 pesos de camiones sin acoplado, vacíos, de la misma marca y modelo. La estimación del desvío estándar calculada a partir de la muestra resulta de 110 kg. Al nivel de significación del 5 % ¿puede afirmarse que la balanza funciona bien?*

Del enunciado surge en primer lugar que un desvío estándar requerido de 100 kg se corresponde con una varianza $\sigma^2 = 10000$. Análogamente el valor muestral $s = 110$ conduce a $s^2 = 12100$. Un test de hipótesis para probar la varianza deberá utilizar la distribución chi-cuadrado con $25 - 1 = 24$ *grados de libertad*, dado el tamaño de la muestra, y un nivel de significancia $\alpha = 0{,}05$ según se lo requiere.

1. *Hipótesis nula* $H_0 : \sigma \leq 100$
2. *Hipótesis alterna* $H_a : \sigma > 100$
3. *La muestra produce un estadístico de prueba* $\chi^2_{24,p} = \frac{(25-1)12100}{10000} = 29{,}04$
4. *Para el nivel de significancia* $\alpha = 0{,}05$ *el valor crítico en la distribución chi-cuadrado con 24 gl es* $\chi^2_{24,\alpha} = 36{,}42$ *con lo cual la región de rechazo*

es $[36{,}42, \infty)$*, Como el estadístico de prueba no cae dentro de ella corresponde no rechazar la hipótesis nula y concluir que la balanza funciona bien al menos al nivel de significación establecido.*

7.5. Bondad de ajuste

Hasta aquí hemos considerado el problema de decidir acerca de valores de parámetros poblacionales. En algunas ocasiones pudimos trabajar sin agregar hipótesis sobre la distribución de la población y, en otras, hemos tenido que postular el comportamiento normal de la misma. Pero puede plantearse también la necesidad de establecer, en términos de aproximación razonable, cual es la distribución que sigue la población. En tales casos es posible realizar un test de hipótesis para comprobar si el comportamiento poblacional se *ajusta* a una determinada distribución. Una muestra de valores empíricos obtenidos para la variable en cuestión permitirá decidir si, a un cierto nivel de significancia, dicho *ajuste* se produce. La hipótesis nula consistirá, precisamente, en suponer que la probabilidad de cada *celda* o resultado posible, corresponde a la probabilidad de una distribución teórica determinada. En símbolos:

1. *Hipótesis nula* $H_0 : p_i = P_{i0}$ para toda celda i
2. *Hipótesis alterna* $H_a : p_i \neq P_{i0}$ para alguna celda i

Aquí P_{i0} representa a la probabilidad teórica postulada para cada celda o resultado posible. Como en todo test de hipótesis, la muestra permite contrastar el comportamiento supuesto con los datos reales a través de la distribución de un estadístico que, en este caso, evalúa la *bondad del ajuste*. Siendo n el tamaño de la muestra, el estadístico tiene la forma $D^2 = \sum_{i=1}^{n} \frac{(F_i - np_{i0})^2}{np_{i0}}$ y, como F_i es la cantidad de veces que se cuenta el $i-$*ésimo* resultado posible, D^2 es una suma ponderada de los cuadrados de las diferencias entre cantidades F_i observadas de una celda y las respectivas cantidades teóricas esperadas np_{i0}. Aunque D^2 tiene una distribución con la que es difícil trabajar, se demuestra que esa distribución puede aproximarse, cuando el tamaño n de la muestra es suficientemente grande, por una chi-cuadrado con $n-1$ grados de libertad, siendo k el número de celdas consideradas[2]. Obsérvese

[2]La demostración puede verse en Harald Cramér: *Métodos matemáticos de estadística*, Aguilar, pp. 477-482

que cuanto mayor resulte, en valor absoluto, cada cantidad $F_i - P_{i0}$ mayor será la discrepancia entre la frecuencia absoluta con que aparece la $i-ésima$ celda en la muestra y la que debiera corresponder según la distribución asumida como hipótesis nula. El crecimiento de la cantidad D^2 debe asociarse con la discordancia entre las frecuencias muestrales y la distribución teórica supuesta de manera que, a partir de un cierto valor crítico de la variable aproximada χ^2, habrá que rechazar la hipótesis nula realizada[3]. Tal valor quedará establecido por el nivel de significancia que se adopte y marcará la separación entre la región de aceptación y la de rechazo.

Ejemplo 5: *Una terminal automotriz intenta establecer si la probabilidad de fabricar cierta pieza de motor defectuosa es la misma para las tres autopartistas que la producen. Esto significa que si una pieza es defectuosa y la automotriz está en lo cierto deberá tener probabilidad 1/3 de provenir de cada fabricante. La tabla 2 explicita esta distribución que en teoría se espera.*

Tabla 2.

Fabricante	Probabilidad del origen de una pieza defectuosa
A_1	$p_{10} = \frac{1}{3}$
A_2	$p_{20} = \frac{1}{3}$
A_3	$p_{30} = \frac{1}{3}$

Con la finalidad de comprobar que tal distribución teórica efectivamente se cumple se observan las 81 últimas piezas defectuosas halladas por los servicios de post-venta y se verifica cual de las tres empresas ha producido cada pieza. Los datos obtenidos junto con las cantidades esperadas se vuelcan en la tabla 3.

[3]Una explicación mas detallada sobre la forma en que crece D^2 cuando la distribución de las frecuencias en la muestra difiere en forma creciente de la propuesta en la hipótesis nula, se proporciona en Harald Cramér: *Teoría de probabilidades y aplicaciones*, Aguilar, pp. 246-247

Tabla 3.

	Autopartista 1	Autopartista 2	Autopartista 3
Cantidad observada de piezas defectuosas	$F_1 = 25$	$F_2 = 30$	$F_3 = 26$
Cantidad esperada	$np_1 = 27$	$np_2 = 27$	$np_3 = 27$

Al nivel de significancia del 5 %, ¿puede asegurase que todos los autopartistas tienen la misma probabilidad de producir una pieza defectuosa?

1. *Hipótesis nula* $H_0 : p_i = \frac{1}{3}$ para $i = 1, 2, 3$
2. *Hipótesis alterna* $H_a : p_i \neq \frac{1}{3}$ para alguna celda $i = 1, 2, 3$
3. *Tamaño muestral* $n = 81$ *Estadístico de prueba* $D^2 = \frac{(25-27)^2}{27} + \frac{(30-27)^2}{27} + \frac{(26-27)^2}{27} = 0{,}5185$ *La cantidad de celdas es* $k = 3$
4. *Para la distribución chi-cuadrado con* $gl = 3-1 = 2$ *que aproxima a la de* D^2 *se determina el valor crítico de acuerdo al nivel de significancia* $\alpha = 0{,}05$. *Resulta* $\chi^2_{2,\alpha} = 5{,}99$ *que determina la región de aceptación* $RA = [0; 5{,}99]$ *y la de rechazo* $RR = (5{,}99; \infty)$.

Como el estadístico de prueba cae dentro de la región de aceptación RA, no se rechaza la hipótesis nula y se concluye que una pieza defectuosa puede provenir con igual probabilidad de cada uno de los tres fabricantes.

Como vimos, el test se realiza sobre la base de que la distribución del estadístico D^2 tiene un comportamiento próximo a χ^2 con $n-1$ grados de libertad. En la práctica se requiere que las frecuencias esperadas cp_{i0} en cada celda resulten mayores o a lo sumo iguales a 10 y, para que esto ocurra, la cantidad n de observaciones que integren la muestra debe ser lo suficientemente grande. Este proceder tiene su incidencia también cuando se testea una distribución continua. En tal caso, cada celda habrá de representar un intervalo sobre el dominio de la variable aleatoria de forma que se cumpla con la cantidad mínima esperada. Analizamos esta situación en el siguiente ejemplo.

Ejemplo 6: *Para probar si los tiempos de permanencia de una persona dentro de un cajero automático son adecuadamente ajustados por una distribución exponencial cuya media es 2 minutos, se ha tomado una muestra de los tiempos que 120 personas han permanecido dentro del cajero. A efecto de realizar el test se han considerado las observaciones muestrales dentro de cuatro celdas distintas correspondientes a cuatro intervalos temporales según se muestra en la tabla 4*

Tabla 4.

	Celda 1= $[0,1]$	Celda 2= $(1,2]$	Celda 3= $(2,3]$	Celda 4= $(3,\infty]$
Cantidad de tiempos observados	38	32	28	22

Es necesario entonces evaluar la bondad del ajuste.

En primer lugar debemos tener en cuenta que la variable aleatoria continua que representa el tiempo tiene un valor esperado $E(t) = \frac{1}{\alpha}$ según se vio en el capítulo 3. De aquí es entonces posible determinar el parámetro α de la distribución haciendo $2 = \frac{1}{\alpha}$ y entonces $\alpha = \frac{1}{2} = 0{,}5$. En segundo término recordemos también que la función de distribución de la variable es $F(t) = 1 - e^{\alpha t}$. Al tener ambas cuestiones podemos formular las hipótesis.

1. *Hipótesis nula* $H_0 : F(t) = 1 - e^{-0{,}5t}$ *(Los tiempos poblacionales se distribuyen según una exponencial con media 2 minutos)*

2. *Hipótesis alterna* $H_a : F(t) \neq 1 - e^{-0{,}5t}$

3. *Los 120 valores muestrales agrupados en cuatro celdas determinan la necesidad de establecer que cantidad de observaciones debiera esperarse en cada una de ellas. Esto se logra calculando la probabilidad acumulada en cada intervalo si la hipótesis nula se cumpliera.*
   $p_{10} = P(t \leq 1) = F(t = 1) = 1 - e^{-0{,}5} = 0{,}39$
   $p_{20} = P(1 < t \leq 2) = F(t = 2) - F(t = 1) = 1 - e^{-0{,}5\times 2} - 0{,}39 = 0{,}24$
   $p_{30} = P(2 < t \leq 3) = F(t = 3) - F(t = 2) = 1 - e^{-0{,}5\times 3} - 0{,}63 = 0{,}15$
   $p_{40} = P(t > 3) = 1 - F(t = 3) = 1 - (1 - e^{-0{,}5\times 3}) = 0{,}22$

   Así las cosas, el número de celdas es $k = 4$ *y sus valores esperados son:* $np_{10} = 120 \times 0{,}39 = 46{,}8$
   $np_{20} = 120 \times 0{,}24 = 28{,}8$
   $np_{30} = 120 \times 0{,}15 = 18$

$np_{40} = 120 \times 0{,}22 = 26{,}4$

El estadístico de prueba es entonces

$D^2 = \frac{(38-46{,}8)^2}{46{,}8} + \frac{(32-28{,}8)^2}{28{,}8} + \frac{(28-18)^2}{18} + \frac{(22-26{,}4)^2}{26{,}4}$

4. *Para un nivel de significancia* $\alpha = 0{,}05$ *habitualmente usado, el valor crítico sobre la distribución* χ^2 *con* $gl = 4 - 1 = 3$ *resulta* $\chi^2_{3,\alpha} = 7{,}81$ *que establece las regiones de aceptación y rechazo respectivas* $RA = [0; 7{,}81]$ *y* $RR = (7{,}81; \infty)$

Corresponde rechazar la hipótesis nula, por lo cual se concluye que los tiempos de permanencia de las personas dentro de un cajero automático no se distribuyen exponencialmente con media 2 minutos.

En los casos ejemplificados la distribución de probabilidades testeada está completamente determinada. En particular es el parámetro que determina la particular distribución exponencial adoptada como hipótesis nula en el Ejemplo 6. Pero no siempre se cuenta con los valores paramétricos para especificar por completo la distribución. En ocasiones solo puede suponerse el tipo de distribución que siguen los datos y los parámetros que la determinan deben primero estimarse para recién luego, testear la hipótesis acerca de la distribución poblacional. En tales oportunidades la forma óptima de realizar la estimación paramétrica es la de máxima verosimilitud ya abordada en el capítulo 6.

7.6. Ejercicios

Ejercicio N°1*: Un auditor sostiene que el valor promedio de todas las cuentas por cobrar en una empresa determinada es de 2600$. Se sabe que el desvío estándar de todas las cuentas por cobrar es de 430$. Se desea poner a prueba la hipótesis del auditor a partir de una muestra aleatoria de 36 cuentas por cobrar de dicha empresa, con un nivel de significación del 5 %.

a) Plantear las hipótesis nula y alternativa adecuadas al problema.

b) Determinar la región de rechazo de la hipótesis nula y la de aceptación.

c) Si la muestra aleatoria de 36 cuentas dio un promedio de 2720$, ¿puede concluirse que el valor promedio de todas las cuentas por cobrar difiere de 2600$?

Ejercicio N°2*: Una planta manufacturera produce cierto tipo de fusibles. Se sabe que la duración de éstos se distribuye normalmente con una duración promedio de 1000 horas y un desvío estándar de 100 horas. Se instala un nuevo proceso de producción y se espera que los fusibles producidos con el nuevo proceso tengan una duración promedio mayor. Se selecciona una muestra aleatoria de 25 fusibles fabricados mediante el nuevo proceso y se obtiene una duración promedio de 1050 horas. Suponer que con el nuevo proceso el desvío estándar se mantiene en 100 horas.

a) Plantear las hipótesis nula y alternativa adecuadas al problema.

b) Con un nivel de significación del 5 %, ¿se deberá instalar el nuevo proceso de fabricación?

Ejercicio N°3: Una compañía de servicio público desea determinar si su nuevo horario de trabajo ha reducido de manera importante el tiempo de espera de los clientes para servicio. El tiempo promedio de espera en el pasado era de 30 minutos. con un desvío standard de 12 minutos. Se selecciona una muestra aleatoria de 144 observaciones y se obtiene una media de 28 minutos y un desvío estándar muestral de 12 min. Con un nivel de significación del 10 %, ¿apoya la evidencia muestral la hipótesis de que el tiempo de espera se ha reducido?

Ejercicio N°4: En una planta de armado se diseña una operación específica que toma un tiempo promedio de 5 minutos. El gerente de planta sospecha que para cualquier empleado el tiempo promedio es mayor. Toma una muestra de 11 tiempos obteniendo los siguientes datos en minutos: 4.9 5.6 5.3 5.0 4.8 5.2 5.4 4.9 5.1 5.2 5.6. Suponiendo que el tiempo de operación es una v.a. normal, ¿la evidencia muestral apoya la sospecha del gerente con $\alpha = 0{,}05$?

Ejercicio N°5*: Se prueban dos fórmulas diferentes de un combustible oxigenado para motor en cuanto al octanaje. La dispersión para la fórmula 1 es de $\sigma_1 = 1{,}5$ y para la fórmula 2 es $\sigma_2 = 1{,}6$. Se toman dos muestras aleatorias de tamaño $n_1 = n_2 = 40$ respectivamente. Los octanajes promedios observados son $\bar{x}_1 = 89{,}6$ y $\bar{x}_2 = 92{,}5$. Al nivel de significancia del 5 %, ¿puede concluirse que el octanaje de ambos tipos de nafta es el mismo?

Ejercicio N°6: Una empresa ha declarado que el 90 % de los artículos que produce no tienen fallas en el período de garantía. Se implementa un proceso de producción que se supone aumentará el porcentaje de artículos sin fallas. Una muestra de 100 artículos dio por resultado 6 fallados. ¿Apoya esta evidencia muestral que el nuevo proceso de producción es significativamente mejor con $\alpha = 0{,}1$? , ¿y con $\alpha = 0{,}05$?

Ejercicio N°7*: Un investigador está convencido de que su equipo de medición tiene una variabilidad que se traduce en una dispersión de $\sigma = 2$. Se efectúa una muestra aleatoria de n = 16 mediciones y da como resultado s = 2.47. ¿Están los datos en desacuerdo con su afirmación, con un nivel de significación $\alpha = 0{,}05$?

Ejercicio N°8: En un sistema de fabricación automático se inserta un remache en un agujero . Si el desvío estándar del diámetro del agujero es mayor que 0.01 mm, existe una probabilidad inaceptablemente grande de que el remache no entre en el agujero. Se toma una muestra aleatoria de $n = 15$ piezas, y se obtiene un desvío estándar muestral $s = 0{,}012$ ¿Es ésta suficiente evidencia para concluir que el desvío estándar del diámetro resultará predominantemente mayor que 0.01 mm? Utilizar $\alpha = 0{,}01$ y suponer que la longitud de los diámetros se distribuye normalmente.

Ejercicio N°9*: Se tiene la sospecha que cierto dado está cargado. A fin de comprobarlo se lo arroja 120 veces obteniéndose los siguientes resultados:

	Cara 1	Cara 2	Cara 3	Cara 4	Cara 5	Cara 6
N°de Veces	20	22	17	18	19	24

Con un 5 % de significancia, ¿puede concluirse que el dado no está cargado?

Capítulo 8

Análisis de la varianza

8.1. Introducción

Las técnicas estadísticas que veremos a continuación se conocen bajo la sigla ANOVA, acrónimo de Analysis of Variance, y constituyen un importante método de análisis experimental. Cuando el comportamiento de un determinado sistema está sujeto a incertidumbre se puede obtener conocimiento desarrollando experimentos sobre él. Para establecer la forma en que se realizarán tales ensayos se deberá tener en cuenta el objetivo del estudio a realizar. En primer lugar hay que identificar cual es la *variable respuesta* que nos informará sobre la característica estudiada. Ésta dependerá de una o varias variables independientes que denominaremos *factores*. En la jerga de ANOVA cada *factor* puede tener distintos niveles y cada combinación de niveles de los factores involucrados representa un *tratamiento*. En particular, si se analiza la influencia de un solo factor sobre la *respuesta*, cada *nivel* es un *tratamiento*. Es decir, tenemos que medir la *variable respuesta* para los distintos *niveles* o *tratamientos* correspondientes a un *factor*. Por ejemplo: supongamos que cada tres meses se ha realizado una campaña de medición de concentración de cromo (Cr) en el sedimento de un río. En cada oportunidad se han obtenido tres datos. La *variable respuesta* en este caso es la concentración de cromo que depende del *factor* número de campaña. Cada campaña en particular es un nivel o *tratamiento*. Si suponemos que las mediciones se han realizado, a través de las distintas campañas, en diferentes localidades del río, tales lugares constituyen lo que llamamos *unidades experimentales*.

El análisis de un experimento realizado para obtener conocimiento acerca de un sistema, cuyas propiedades son "a priori" inciertas, habitualmente

requiere la comparación de comportamientos en los distintos niveles o tratamientos. En nuestro ejemplo se podría querer establecer si el valor medio de concentración de cromo en el sedimento del río, durante todo el período en que se realizaron las distintas campañas, ha permanecido constante. Téngase en cuenta que las cantidades medidas en cada campaña constituyen una muestra y sus promedios son una estimación del valor medio poblacional de la concentración de cromo en el agua en ese momento.

Para que una investigación experimental tenga éxito hay que asegurarse que todos los *factores* que intervienen en la *respuesta* sean controlados por el investigador. Como esto en la realidad suele no ser posible, bien porque no se conozcan efectivamente todos los factores intervinientes, bien porque pueda no depender del investigador el control, se hace necesario en primer término realizar el muestreo de las unidades experimentales en forma aleatoria para atenuar cualquier efecto particular. Por ejemplo: si los lugares del río en los que se mide la concentración de cromo son siempre los mismos es posible que los promedios se vean afectados no solo por la época del año en que se realiza cada campaña, es decir por el *tratamiento*, sino también por las características contaminantes de esas localidades que si estuvieran alejadas de los centros urbanos ofrecerían una contaminación menor que la que debiera esperarse. Por supuesto que puede resultar válido a efecto de un objetivo particular realizar el estudio en tales condiciones, por ejemplo para establecer cual sería el nivel de cromo sin que intervenga la polución de las ciudades, pero esto debe ponerse en claro al efectuar el estudio para que luego no se extraigan falsas conclusiones sobre el nivel promedio de contaminación del río. La elección aleatoria de las unidades experimentales, es decir de los lugares del río donde se recogen muestras, nos pone a cubierto de cualquier sesgo que se produzca en la estimación como consecuencia de que se tomen en cuenta solo lugares escasamente contaminados o, por el contrario, muy contaminados. En segundo término puede medirse el *error experimental* estableciendo la variación aleatoria de la característica analizada al repetir varias veces el experimento. En nuestro ejemplo esto significaría que al realizar cada campaña se tomarían tres mediciones en tres lugares del río elegidos al azar, luego otras tres en otros tres lugares elegidos al azar y así. Todas estas mediciones realizadas en la misma campaña podrían ayudar a establecer el error aleatorio correspondiente a la variación de lugares. Este error sería tenido en cuenta al evaluar el promedio de concentración solo en función del momento del año en que se efectúa la campaña.

8.2. Análisis de la varianza de un factor

Comenzaremos por considerar el análisis de la varianza cuando se supone que la respuesta proviene de un solo *factor*, es decir cuando se ha desarrollado un experimento unifactorial. También supondremos que las unidades experimentales se asignan a cada tratamiento en forma completamente aleatoria. Esto último significa que la probabilidad de elegir cualquier unidad experimental es la misma y que cada elección resulta independiente de toda otra. Como ya se comentó, tal tipo de asignación se realiza para procurar que la respuesta sea observada en iguales condiciones durante todo el experimento. Así se podrán atribuir las variaciones que se observen a la diferencia entre tratamientos y no a otros factores que no estén controlados. Si se tienen r niveles o tratamientos y se intenta comparar las medias poblacionales de cada nivel se plantea la siguiente hipótesis nula:

Hipótesis nula $H_0 : \mu_1 = \mu_2 = \ldots = \mu_r$

En términos lógicos la igualdad de medias se niega cuando al menos una de las medias de los tratamientos es distinta a las otras por lo cual queda establecida la hipótesis alternativa,

Hipótesis alterna H_a : *Existe* μ_k *tal que es distinta de las demás medias*

Los datos obtenidos en cada uno de los r tratamiento pueden ordenarse según se muestra en la tabla 1.

Tabla 1.

Tratamientos

	1	2	$\cdots$	r
Observación 1	x_{11}	x_{12}	$\cdots$	x_{1r}
Observación 2	x_{21}	x_{22}	$\cdots$	x_{2r}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Observación nj	$x_{n_1 1}$	$x_{n_2 2}$	$\cdots$	$x_{n_r r}$

La cantidad n_j varía de acuerdo al tratamiento por lo que el número de unidades experimentales en cada nivel puede ser distinto. Por otra parte es claro que los rótulos colocados para nombrar a cada tratamiento indican cantidades o simplemente categorías distintas como por ejemplo A, B, etc., por lo que el factor puede ser una variable cuantitativa o una categórica.

Al llegar a este punto cabe preguntarse como hay que interpretar cada observación x_{ij}. Cada dato observado en cada tratamiento puede pensarse

como desviado del promedio general por dos causas concurrentes: el efecto de cada tratamiento y el error aleatorio producido por el muestreo en la experiencia. En fórmulas: $x_{ij} = \mu + \tau_j + \epsilon_{ij}$ donde μ es el promedio poblacional general, es decir considerados todos los niveles, τ_j el efecto del tratamiento j sobre la observación y ϵ_{ij} el error aleatorio experimental al tomar la observación x_{ij}. Este error se postula distribuido normalmente con media 0 y varianza σ_τ^2 lo que se anota: $\epsilon_{ij} \sim N(0, \sigma_\tau^2)$ Considerando fijo el efecto de cada tratamiento, es decir no sujeto a variación aleatoria, se supone adicionalmente que, si se cumple la hipótesis nula, la suma de los efectos sobre el total de observaciones del experimento es $\sum_{j=1}^{r} n_j \tau_j = 0$ y así resulta el llamado *modelo de efectos fijos* del ANOVA.

De acuerdo al modelo de efectos fijos las hipótesis nula y alterna formuladas son equivalentes a:

Hipótesis nula $H_0 : \tau_j = 0$ para todo $j = 1, 2, \cdots, r$

Hipótesis alterna $H_a : \tau_j \neq 0$ para algún j

Obsérvese que si para todo valor de j ocurre que $\tau_j = 0$ se puede considerar nulo el efecto de todos los tratamientos y por lo tanto, cada observación x_{ij} variará solo de acuerdo al error aleatorio que incorpore en el muestreo. Es decir: $x_{ij} = \mu + \epsilon_{ij}$ Como los errores distribuidos normalmente tienen una media 0, al promediar un número suficientemente grande de observaciones el valor del promedio muestral se acercará a μ. Éste valor resultará el mismo para todos los tratamientos, dado que el efecto τ_j de cada uno es nulo y por lo tanto esta hipótesis nula será equivalente a la que supuso las medias iguales.

Al avanzar en el análisis de nuestro modelo de efectos fijos definimos a $\tau_j = \mu_j - \mu$ como el efecto del tratamiento j, siendo μ_j la media de la población si se considera el nivel j y μ la media poblacional general, sin tener en cuenta el nivel particular que se establezca. De la misma forma, el error aleatorio de cada observación en cada tratamiento resulta $\epsilon ij = x_{ij} - \mu_j$ Estas cantidades se reemplazan en la expresión original $x_{ij} = \mu + \tau_j + \epsilon_{ij}$ para obtener entonces $x_{ij} = \mu + (\mu_j - \mu) + (x_{ij} - \mu_j)$ con lo cual queda $x_{ij} - \mu = (\mu_j - \mu) + (x_{ij} - \mu_j)$.

La fórmula revela explícitamente que la variación de cada observación respecto de la media global es la suma de la variación producida por el tratamiento, $(\mu_j - \mu)$ y la ocasionada por el error en el muestreo aleatorio $(x_{ij} - \mu_j)$. Así aceptar la hipótesis de la igualdad de las medias de los tratamientos equivale a establecer que estos no introducen variación en la observación y que la existente solo debe atribuirse al error aleatorio. Obsérvese que, en definitiva, analizar la igualdad de medias es evaluar variaciones de dos tipos. De ahí resulta el nombre de análisis de la varianza para estos

métodos.

La realización del test de hipótesis propuesto requiere de un estadístico y su distribución para determinar zonas de aceptación y rechazo o bien valores p. Se hace necesario entonces fijar algunas cuestiones de notación y desarrollar un trabajo algebraico que veremos solo someramente. Primero establecemos la siguiente notación:

$T_{*j} = \sum_{i=1}^{n_i} x_{ij}$ suma del total de las observaciones para cada tratamiento.

$\bar{X}_{*j} = \frac{T_{\cdot j}}{n_j}$ promedio de las observaciones del tratamiento j. Estimador de μ_j.

$T_{**} = \sum_{j=1}^{r} T_{\cdot j}$ suma de las observaciones de todos los tratamientos.

$N = \sum_{j=1}^{r} n_j$ total de observaciones del experimento.

$\bar{X}_{**} = \frac{T_{\cdot\cdot}}{N}$ promedio del total de observaciones. Estimador de la media global μ.

Recordemos que la fórmula que desglosa la variación es $x_{ij} - \mu = (\mu_j - \mu) + (xij - \mu_j)$ de forma que si reemplazamos las medias de los tratamientos y la global por sus estimadores queda $x_{ij} - \bar{X}_{**} = (\bar{X}_{*j} - \bar{X}_{**}) + (x_{ij} - \bar{X}_{*j})$

Si se elevan ambos miembros de la igualdad al cuadrado operando se obtiene:

$$\sum_{j=1}^{r}\sum_{i=1}^{n_j}(x_{ij} - \bar{x}_{**})^2 = \sum_{j=1}^{r}\sum_{i=1}^{n_j}(\bar{x}_{*j} - \bar{x}_{**})^2 + \sum_{j=1}^{r}\sum_{i=1}^{n_j}(x_{ij} - \bar{x}_{*j})^2$$

Es decir; la suma total de cuadrados STC resulta igual a la suma de los cuadrados de los tratamientos $SCTR$ más la suma de los cuadrados de los errores SCE. A efecto de simplificar algunos cálculos se puede probar que son equivalentes las fórmulas $STC = \sum_{j=1}^{r}\sum_{i=1}^{n_j} x_{ij}^2 - \frac{T_{**}^2}{N}$, $SCTR = \sum_{j=1}^{r} \frac{T_{*j}^2}{n_j} - \frac{T_{**}^2}{N}$ y $SCE = STC - SCTR$. En cualquier caso se tiene la fórmula: $STC = SCE + SCTR$. La ventaja de haber obtenido esta expresión queda clara cuando se recuerda que el cociente $\chi^2 = \frac{(n-1)s^2}{\sigma^2}$ es una variable aleatoria de distribución chi cuadrado con $gl = (n-1)$ grados de libertad. Así puede verse que bajo la hipótesis nula $H_0 : \tau_j = 0$, suponiendo que $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$, las variables $\frac{SCTR}{\sigma_{tr}^2}$ y $\frac{SCE}{\sigma_\epsilon^2}$ son independientes con $gl = r-1$ y $gl = N-r$ grados de libertad respectivamente. Se pueden hallar los llamados *cuadrados medios* de cada tratamiento y del error como $CMTR = \frac{SCTR}{r-1}$ y $CME = \frac{SCE}{N-r}$. Se construye entonces el estadístico $F = \frac{CMTR}{CME}$ que tiene una distribución de Fischer con $gl = r-1$ grados de libertad en el numerador y $gl = N-r$ grados de libertad en el denominador. Tal distribución se tabula

en el Apéndice D[1]. Se observa que valores pequeños de $SCTR$ originados en poca diferecia entre las medias de los distintos tratamientos conducen a valores pequeños de F. En la medida que F crece, la variación atribuible al efecto de los tratamientos habrá de hacerse mayor y menos incidirá en el estadístico la variación del error aleatorio. Veamos un ejemplo.

Ejemplo 1: *Se han recogido muestras de sedimento del lecho de un río en distintos momentos del año y se ha establecido la cantidad de cromo (Cr), medida en mg/kg, presente en cada observación. Los datos obtenidos se despliegan en la tabla 2. Se desea establecer si la presencia de Cr, en promedio, ha permanecido constante a lo largo del período observado.*

Tabla 2.

Campaña	Cr
1 Feb-2007	11.85
1 Feb-2007	22.88
1 Feb-2007	16.86
2 Ago 2007	74.00
2 Ago 2007	85.50
2 Ago 2007	62.70
3 Nov 2007	40.00
3 Nov 2007	13.70
3 Nov 2007	19.90

Comencemos por observar que hay $r = 3$ tratamientos señalados como Campañas 1, 2 y 3 respectivamente. En cada tratamiento se han registrado 3 observaciones. Es decir $n_1 = n_2 = n_3 = 3$. Resulta entonces que el número de grados de libertad del numerador en la distribución del estadístico es $gl = 3-1 = 2$. Se calcula $N = 3+3+3 = 9$ y resultan los grados de libertad del denominador $gl = 9-3 = 6$. Además se tienen:

$$T_{*1} = 11{,}85 + 22{,}85 + 16{,}86 = 51{,}59,\ T_{*2} = 222{,}20,\ T_{*3} = 73{,}60$$
$$\bar{x}_{*1} = \frac{51{,}59}{3} = 17{,}20,\ \bar{x}_{*2} = \frac{222{,}20}{3} = 74{,}06,\ \bar{x}_{*3} = \frac{73{,}60}{3} = 24{,}53,$$
$$T_{**} = 347{,}39 \text{ y } \bar{x}_{**} = \frac{347{,}39}{9} = 38{,}60.$$

[1]Un análisis mas pormenorizado de la teoría expuesta puede consultarse en George Canavos, ob.cit., pp. 240-243 y 407-411.

Con estas cantidades se calculan a su vez:

$STC = (11{,}85)^2 + (22{,}88)^2 + (16{,}86)^2 + (74{,}00)^2 + (85{,}50)^2 + (62{,}70)^2 + (40{,}00)^2 + (13{,}70)^2 + (19{,}90)^2 - \frac{(347{,}39)^2}{9} = 5741{,}57$

$SCE = STC - SCTR = 6440{,}55 - 5741{,}57 = 698{,}98$

$CMTR = \frac{SCTR}{r-1} = \frac{6440{,}55}{2} = 3220{,}28$ y $CME = \frac{SCE}{N-r} = \frac{698{,}98}{6} = 116{,}50.$

El estadístico de Fischer resulta:

$F = \frac{CMTR}{CME} = \frac{3220{,}28}{116{,}50} = 27{,}64$

Los datos se acomodan en una tabla ANOVA como se ve en la tabla 3:

Tabla 3.

	gl	SC	CM	Estadística F
Var.Tratamientos	2	5741.57	3220.28	27.64
Var.Error	6	698.98	116.50	
Total	8	6440.55		

La tabla de la distribución de Fischer, adjuntada en el Anexo D, acumula la probabilidad $1 - \alpha = 0{,}95$ hasta el valor de abscisa referenciado por los grados de libertad en el numerador y el denominador en primera fila y primera columna respectivamente. Es decir el valor de F crítico es en este caso $F_{critico} = 5{,}14$. Como el valor de prueba obtenido es mayor cae dentro de la región de rechazo para el nivel de significancia $\alpha = 0{,}05$. Como consecuencia se rechaza la hipótesis nula de igualdad de las tres medias poblacionales.

8.3. Análisis post hoc

La hipótesis nula de igualdad de medias utilizada en el ANOVA puede ser aceptada, y por lo tanto considerar que las medias poblacionales en los distintos niveles son iguales, ó puede ser rechazada, en cuyo caso habrá al menos dos medias poblacionales distintas. En verdad es factible que sean distintas dos medias, tres o todas las consideradas y el test realizado no nos sirve para aclarar esta situación. Con tal finalidad se suele realizar un

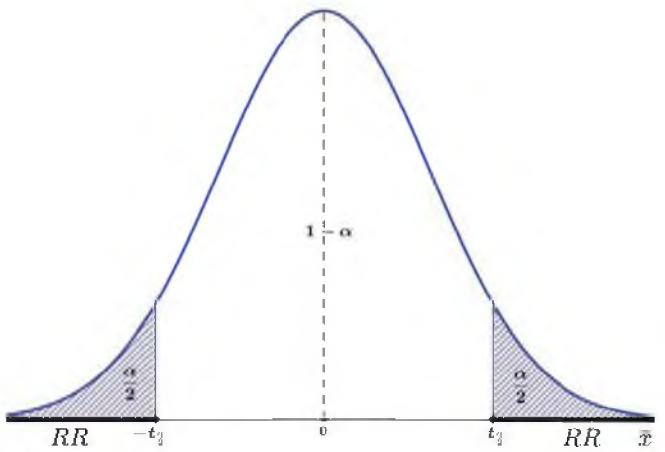
análisis "a posteriori" o "post hoc" que explora las diferencias entre medias que resulten estadísticamente discernibles. Tal análisis puede efectuarse de distintas formas y a continuación estudiaremos una de ellas. Consideramos en este caso que para todos los tratamientos se ha seleccionado la misma cantidad de unidades experimentales. Es decir; para los r tratamientos considerados se cumple que $n_1 = n_2 = \cdots = n_r$. Si el valor del estadístico de Fischer aconseja rechazar la hipótesis de igualdad de las medias o lo que es equivalente rechazar la hipótesis de efecto nulo en cada tratamiento, $\tau_j = 0$ para todo $j = 1, \cdots, r$, se define como diferencia mínima entre medias a la que existiere entre las medias de dos muestras significativamente diferentes. Si suponemos que la distribución en cada tratamiento es normal, siendo los tamaños muestrales pequeños, se puede construir el estadístico t-student,

$$t_p = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

para estimar la diferencia de medias y realizar con él la prueba de la hipótesis nula de igualdad de medias siendo los grados de libertad de su distribución $gl = n_1 + n_2 - 2$. Como además se supone en este caso la igualdad de todos los tamaños muestrales y también la igualdad de las variancias de los tratamientos, tal estadístico resulta $t_p = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{2\sigma^2}{n}}}$ y $gl = 2n - 2 = 2(n - 1)$ Obsérvese que aquí n es el tamaño de cada muestra obtenida en cada uno de los tratamientos y σ^2 la varianza atribuida al error experimental que se postula la misma en todos los tratamientos. En la distribución t-student, el test correspondiente a la igualdad de dos medias tendría una región de aceptación, correspondiente a la probabilidad $1 - \alpha$ acumulada, y otra de rechazo formada, en este caso, por las dos colas de la distribución que acumulan, cada una, la probabilidad $\frac{\alpha}{2}$ tal como se muestra en la figura 1. Los valores críticos del estadístico son entonces $t_{\frac{\alpha}{2}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{2\sigma^2}{n}}}$. Es decir que hay una *distancia mínima significativa*, medida por la diferencia de medias muestrales, a partir de la cual el estadístico de prueba caerá en la región de rechazo del test. Llamemos DMS a esta cantidad y pongamos $t_{\frac{\alpha}{2}} = \frac{DMS}{\sqrt{\frac{2\sigma^2}{n}}}$ con lo cual $DMS = t_{\frac{\alpha}{2}} \sqrt{\frac{2\sigma^2}{n}}$ Observando ahora que σ^2 es también la varianza de la muestra grande formada por todos los tratamientos, habida cuenta de la suposición de igualdad de varianza en todos ellos, se puede escribir $DMS = t_{\frac{\alpha}{2}} \sqrt{\frac{2CME}{n}}$. De tal modo, comparando de a dos las medias de los tratamientos, sus diferencias serán significativas si el valor absoluto de las

mismas resulta mayor que la cantidad DMS.

Figura 1.



Ejemplo 2: *Se trata de establecer para cuales tratamientos de los considerados en el Ejemplo 1 las medias resultan significativamente distintas como para concluir que son diferentes las medias poblacionales.*

Como el tamaño de las muestras en todos los tratamientos es $n = 3$, los grados de libertad del estadístico t utilizado para comparar las medias de a pares resultan $gl = 2(3-1) = 4$. Si seleccionamos un nivel de significación $= 0{,}05$ consultamos la tabla de la t-student y establecemos $t_{\frac{\alpha}{2}} = 2{,}776$. Por otra parte, de la tabla 3 extraemos el valor $CME = 116{,}50$. Calculamos entonces el valor de la distancia mínima significativa $DMS = 2{,}776\sqrt{\frac{2\times 116{,}50}{3}} = 24{,}46$. Ahora comparamos:

$|x_{*1} - x_{*2}| = |17{,}20 - 74{,}06| = 56{,}86 > DMS$ por lo tanto se debe concluir que las medias poblacionales de los tratamientos 1 y 2 son distintas.

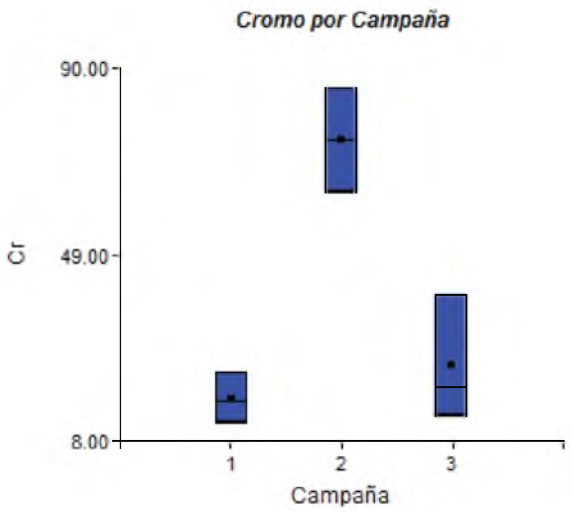
$|x_{*1} - x_{*3}| = |17{,}20 - 24{,}53| = 7{,}33 < DMS$ lo que indica que las medias poblacionales de los tratamientos 1 y 3 pueden suponerse iguales.

$|x_{*2} - x_{*3}| = |74{,}06 - 24{,}53| = 49{,}53 > DMS$ las medias poblacionales de los tratamientos 2 y 3 son distintas.

8.4. Los supuestos del análisis de la varianza

Los supuestos teóricos del análisis de la varianza comprenden la normalidad de las variables aleatorias que constituyen la respuesta en cada tratamiento, la igualdad de sus varianzas y la independencia de los errores aleatorios entre sí. El ANOVA es *robusto* en relación con el supuesto de normalidad de las variables. En efecto, en cada tratamiento bastan unas pocas observaciones con cierta simetría alrededor de la media para suponer que la hipótesis de normalidad se cumple En el caso del ejemplo 1 tal simetría puede observarse en la figura 2.

Figura 2.



Cada caja tiene los 3 valores de cromo observados en cada campaña y el punto negro representa la media de los tres. En este caso se supondrá entonces que se satisface la hipótesis de normalidad de las variables. La suposición de igualdad de varianzas para tales variables aleatorias debe satisfacerse al menos en forma aproximada. Ya hemos analizado que cada observación puede considerarse bajo el modelo $x_{ij} = \mu + \tau_j + \epsilon_{ij}$ y que, por lo tanto, el error aleatorio en cada tratamiento resulta $\epsilon_{ij} = x_{ij} - \mu_j$ Los residuos $e_{ij} = x_{ij} - \bar{x}_{*j}$, calculados a partir de los datos obtenidos, son entonces una estimación de esos errores aleatorios y hay que señalar que si, en cada tratamiento, tales diferencias conducen a gráficas de dispersión parecida, esto ocurrirá porque las varianzas de las variables son aproximadamente iguales. De todas formas antes de graficar conviene *estandarizar*

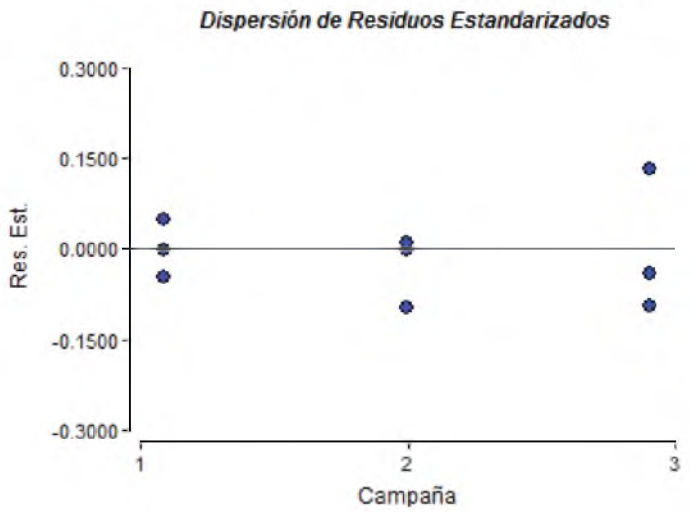
los residuos dividiéndolos por $\sqrt{CME}$ a efecto de compararlos en la misma escala. Si calculamos los *residuos estandarizados* a partir de los datos del ejemplo 1 se obtiene la tabla 4.

Tabla 4.

Res. Est. Campaña 1	Res. Est. Campaña 2	Res. Est. Campaña 3
-0.0459	-0.0005	0.1328
0.0488	0.0099	-0.0975
-0.0029	-0.0930	-0.0397

En la figura 3 se observan los residuos estandarizados de cada campaña:

Figura 3.



Como los errores aleatorios se suponen normalmente distribuidos, al estandarizar los residuos cabe esperar que tengan un comportamiento normal estándar y por lo tanto rara vez se deberían obtener residuos mayores que 3 o menores que -3. En la figura 3 las diferencias entre los residuos en cada una de las campañas son pequeñas y parecen no ser demasiado relevantes como para considerar distintas varianzas en los tratamientos. Así, razonablemente, puede suponerse que las varianzas de cada uno de los niveles son iguales. Tal suposición es plausible además porque el estadístico F utilizado para testear la hipótesis de igualdad de medias es también robusto respecto de la desigualdad de las varianzas cuando los tamaños muestrales de los tratamientos son, como en este caso, iguales.

En relación con los supuestos del ANOVA se comenta ahora brevemente que la independencia de los errores aleatorios resulta clave para el buen funcionamiento de la prueba sobre la igualdad de medias. Que los errores aleatorios sean independientes quiere decir por un lado que en un mismo tratamiento tales errores no se vinculen por ejemplo con el orden en que son registrados los datos o también que los errores atribuidos a un tratamiento no dependan en algún sentido de los errores de otro nivel. Probar esta condición puede no resultar sencillo pues ella suele ser consecuencia del diseño de la investigación y de la forma en que se recolectaron los datos.

8.5. Otros enfoques de ANOVA

Hemos trabajado hasta aquí con lo que se denomina un *diseño completamente aleatorio* del experimento en el cual la elección de cada unidad experimntal se ha realizado aleatoriamente. Así se ha supuesto además que el efecto de cada tratamiento es fijo. Si tal efecto fuera aleatorio el ANOVA requeriría variantes que no vamos a analizar[2]. También cuando las unidades experimentales no son homogéneas, por ejemplo en el río hay zonas de mayor contaminación que otras, esto debe tomarse en cuenta y realizar una selección aleatoria dentro de distintos bloques donde la homogeneidad esté asegurada. En este caso se dice que el diseño del experimento es en *bloque completamente aleatorizado*. Esto introduce en el modelo $x_{ij} = \mu + \tau_j + \epsilon_{ij}$ un efecto por bloque adicional que debe sumarse también, aunque tampoco para este caso avanzaremos en el análisis[3]. Además sólo hemos considerado un factor para determinar los tratamientos. En realidad podríamos también realizar un análisis multifactorial de la varianza al considerar la conjunción de varios factores lo que excede los alcances de nuestra presentación. Finalmente debemos aclarar que si las suposiciones realizadas no fueran de ningún modo satisfechas todavía sería posible realizar comparaciones por procedimientos no paramétricos.

[2] Véase por ejemplo George Canavos, ob.cit., pp. 418 a 419

[3] Al respecto puede verse George Canavos, ob.cit., pp. 420-426

8.6. Ejercicios

Ejercicio N°1*: Una fábrica de neumáticos produce tres tipos de cubiertas que, según anuncia, tienen similar duración. Con el objetivo de realizar un control de calidad ha tomado una muestra de 8 neumáticos de cada tipo y desea evaluar si efectivamente los promedios de duración pueden considerarse iguales o si los grupos revelan diferencias significativas entre los mismos. De ocurrir esto último se desea además saber entre que grupos se presentan tales diferencias. Los datos, expresados en kilómetros, son los siguientes.

Neumático Tipo A	Neumático Tipo B	Neumático Tipo C
33520	30120	34100
31340	29900	32700
32820	31300	31810
30210	32230	31900

Ejercicio N°2: Cuatro marcas de tubos fluorescentes compiten por un mismo mercado y afirman que ofrecen tubos de mayor duración media. Un constructor desea saber si esto es realmente así o si hay una marca que resulte en promedio más duradera. Con tal motivo ha registrado la duración en horas de tubos fluorescentes de las distintas marcas según se muestra en la tabla. ¿Cuál resulta su conclusión correcta?

Marca 1	Marca 2	Marca 3	Marca 4
20233	19322	20080	22000
21455	21040	21600	22560
19610	20387	22150	23310
19890	18900	21385	24115
20815		20990	20600
18654		21760	21550

Capítulo 9

Regresión y correlación lineal

9.1. Introducción

En las aplicaciones estadísticas es usual intentar describir, estimar y aún predecir el comportamiento de alguna característica poblacional en relación con otra u otras. Por ejemplo; si se desea explicar el consumo mensual de los hogares a partir de los ingresos que obtienen en tal período, es posible plantear una formulación matemática que vincule la *variable dependiente o explicada* "consumo" con la *variable independiente o explicativa* "ingreso". Se trata de establecer un *modelo matemático* que exprese, en este caso, la relación causal que la teoría económica supone existente entre el consumo de las personas y sus respectivos ingresos. En general tenemos entonces una función $f(x) = y$ pero no cualquier función representará adecuadamente el hecho económico que se pretende modelar. Un *modelo lineal* explicitado por la recta $y = \beta_1 x + \beta_0$ quizás resulte pertinente aunque, como enseguida veremos, no puede ser aplicado sin tener en cuenta cierto carácter aleatorio presente en el consumo hogareño. Los valores de los coeficientes β_1 y β_0 habrán de ser estimados a partir de datos experimentales obtenidos de unas cuantas familias y de ellos se inferirá, de acuerdo a ciertos supuestos, una función poblacional de consumo. Con todo, puede ocurrir que la expresión obtenida no sirva de mucho para representar la vinculación entre las variables y que deba recurrirse a otras formas de modelado no lineal. En síntesis se nos presentan dos problemas relacionados: por un lado estimar los parámetros del modelo lineal para definirlo adecuadamente y, por otro, evaluar que tan bien el modelo hallado "ajusta" a la relación causal que se postula entre ingreso y consumo. El primer aspecto es analizado por la teoría de regresión mientras que el segundo lo enfoca la teoría de correlación. Las aplicaciones

de ambas permiten estudiar una gran variedad de hechos que relacionan variables no sólo en la teoría económica, como en el ejemplo, sino también en dominios tales como la ingeniería, la sociología, la medicina y otros.

9.2. El modelo de regresión lineal

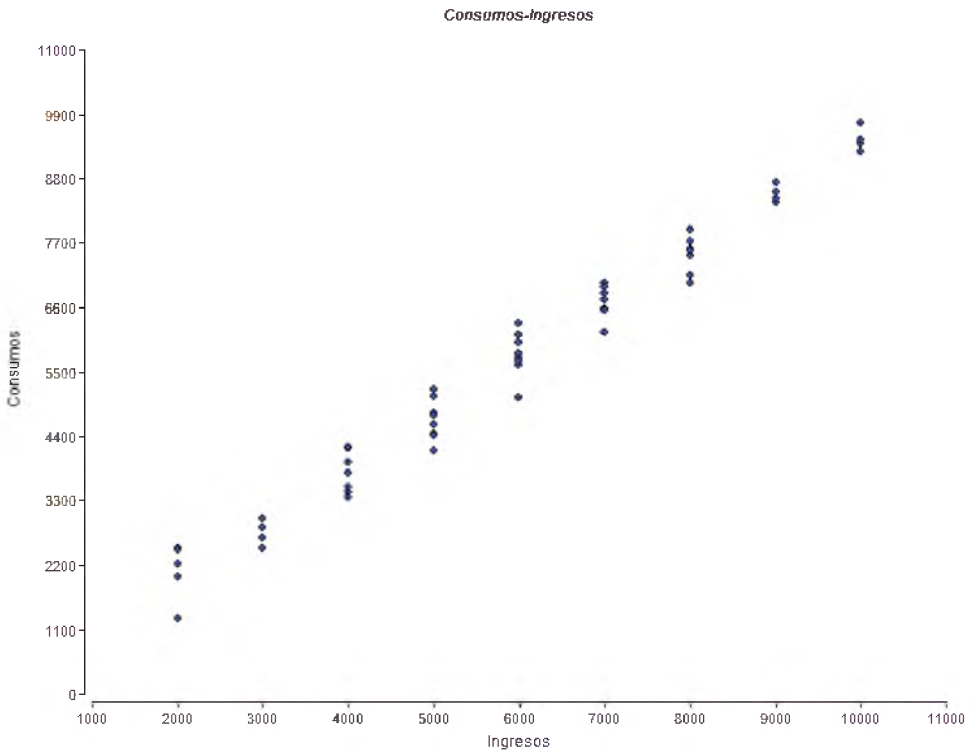
Supongamos que un censo desarrollado en una comunidad de 60 hogares ha permitido conocer cada ingreso, redondeado a miles, y cada consumo según se muestra en la tabla 1. Cada columna de la tabla corresponde al consumo de los hogares cuyo nivel de ingreso está rotulado en la primera fila.

Tabla 1.

2000	**3000**	**4000**	**5000**	**6000**	**7000**	**8000**	**9000**	**10000**
1299	2667	4198	4754	5078	6587	7029	8471	9746
2456	2490	3364	5211	5753	6838	7581	8576	9474
1996	2483	4222	4429	5696	6948	7733	8467	9408
2494	2855	3780	4746	5831	6566	7930	8740	9265
2224	3000	3439	4161	6127	7027	7150	8393	
	2482	3946	5091	5746	6172	7498		
		3536	4798	5626	6953	7595		
			4439	6007	6740			
			4593	6329				

En la figura 1 se vuelca la información de la tabla en un diagrama de dispersión:

Figura 1.

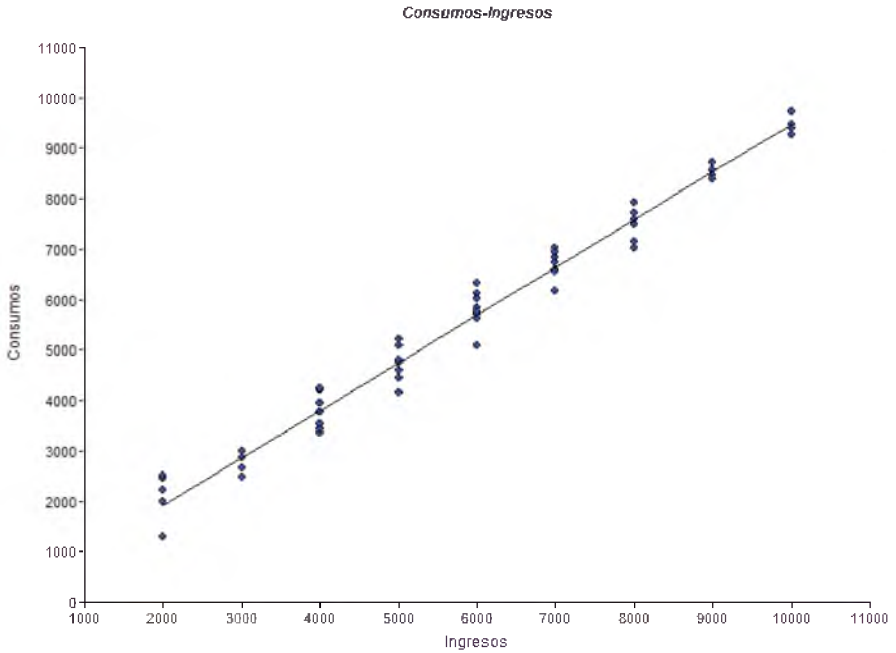


Claramente tenemos aquí los datos de toda la comunidad considerada la cual, desde el punto de vista estadístico, constituye entonces una población. Por supuesto que fijado un determinado ingreso x, el gasto de consumo y correspondiente a un hogar estará caracterizado por un punto ubicado sobre la recta respectiva. Por ejemplo; el consumo de $y = 3364$ pesos se ubicará sobre la recta $x = 4000$ pesos de ingreso. En cada hogar la parte del ingreso que se destina al consumo, y no al ahorro por ejemplo, depende de la interacción de muchos factores, algunos azarosos o no suficientemente explicitados, lo que permite considerar a la cantidad y como una variable aleatoria. De acuerdo a esto, dado un ingreso $x = x_i$ habrá un consumo esperado $E(y/x_i)$ cuyo valor se calculará a partir de la de probabilidad condicional $P(y/x_i)$. Se comprende entonces que existe una familia de distribuciones de la variable y, cada una bajo una condición x_i distinta. Esto habilita a pensar la esperanza de la variable y como una función del valor que asuma la variable x. Es decir; $E(y/x_i) = f(x_i)$ o mas generalmente aún $E(y/x) = f(x)$.

Ahora bien; la figura 1 autoriza a pensar que, para el caso que analizamos, la esperanza de la variable condicionada y, que representa el consumo, varía linealmente con los valores de ingreso simbolizados por x. De tal modo ponemos $E(y/x_i) = \beta_1 x_i + \beta_0$. El coeficiente β_0 es la ordenada al origen de la recta en cuestión y representa al consumo mínimo promedio que habrá de producirse aún sin existencia de ingreso. Por su parte β_1 es la pendiente y se interpreta como la proporción del ingreso que habrá de destinarse al consumo según las costumbres promedio de la población considerada. La recta imaginada deberá pasar entonces por todos los promedios de los valores obtenidos para y cada vez que se fijó un valor a x como se ve en la figura 2.

Podemos considerar el consumo de una familia en particular como el consumo promedio más (o menos) una cantidad aleatoria particular para cada caso. Así tenemos: $y_i = E(y/x_i) + u_i$ Entonces, dado el ingreso x_i, el valor esperado para y_i se calcula como $E(y_i/x_i) = E(E(y/x_i)) + E(u_i/x_i)$. Habida cuenta que una esperanza es una cantidad constante resulta $E(y_i/x_i) = E(y/x_i) + E(u_i/x_i)$ y como $E(y_i/x_i) = E(y/x_i)$ se tiene finalmente que $E(u_i/x_i) = 0$ Es decir; el valor esperado de la *perturbación aleatoria* es 0. Obsérvese ahora que si la recta que pasa por los promedios según cada x_i es $E(y_i/x_i) = \beta_1 x_i + \beta_0$ se tiene que $y_i = \beta_1 x_i + \beta_0 + u_i$ a lo que llamaremos *modelo de regresión lineal de la población.* El término regresión deviene de que al ser 0 el valor esperado de la perturbación se interpreta que los datos "regresan" en tendencia central hacia el promedio $E(y_i/x_i)$. Precisamente hay que notar que $E(u_i/x_i) = 0$ porque se ha supuesto que la recta pasa por los promedios. En este modelo, la perturbación u_i capta y totaliza los efectos aleatorios que desvían del promedio a cada observación.

Figura 2.



En general se tiene que un *modelo de regresión lineal poblacional* vincula dos variables, una *explicativa* y otra *explicada* a través de la ecuación $y_i = \beta_1 x_i + \beta_0 + u_i$. Se dice que este modelo es de regresión *lineal* simple pues involucra solo una variable independiente o explicativa. Debe observarse además que, tal cual está formulado, el modelo no solo es lineal respecto de dicha variable que aparece elevada a la 1, sino también respecto de sus coeficientes, los que también figuran con esa potencia. En el análisis de regresión a veces no importa tanto la linealidad respecto de las variables ya que ellas pueden ser sustituidas por otras cantidades de aspecto lineal, por ejemplo haciendo $z = x^2$ y planteando la ecuación $y_i = \beta_1 z_i + \beta_0 + u_i$. Más sustantivo resulta en estos casos cuidar la linealidad de los parámetros β_0 y β_1.

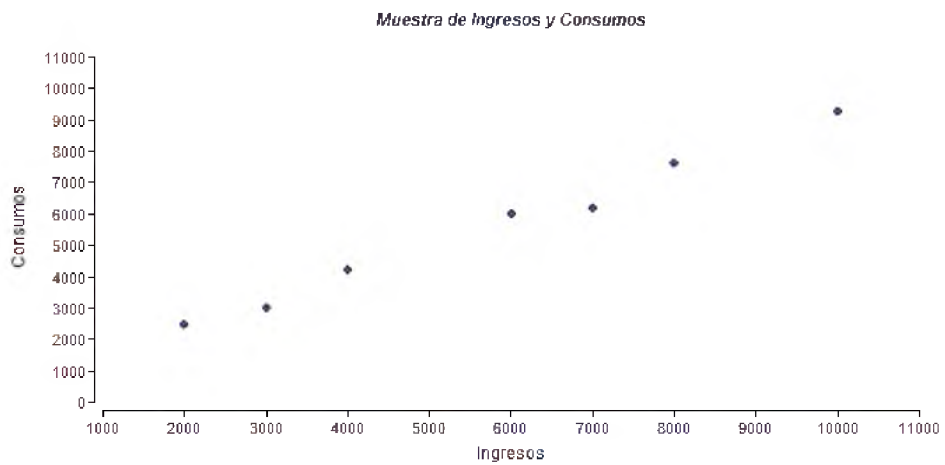
9.3. Supuestos y estimación de parámetros

Si se tienen los datos referidos al total de la población, en nuestro ejemplo todos los pares ingreso-consumo, el modelo posee solo carácter descriptivo,

sin capacidad de predicción sobre los elementos que forman parte de la población cuyo comportamiento es ya conocido. Pero esa no suele ser la situación. A menudo se conocen solo unos pocos pares de observaciones y a partir de ellos, suponiendo adecuado el modelo lineal para la población, se trata de estimar sus parámetros β_0 y β_1. Existen distintas formas de hacer esto pero hay un procedimiento que, bajo ciertas suposiciones previas, resulta óptimo. Veremos a continuación ése método de estimación y luego comentaremos brevemente sus supuestos.

Si existe una recta de regresión para una población cuyo comportamiento es desconocido, salvo el de algunas observaciones que constituyen una muestra, se puede realizar un diagrama de dispersión que represente tal situación como se ve en la figura 3:

Figura 3.

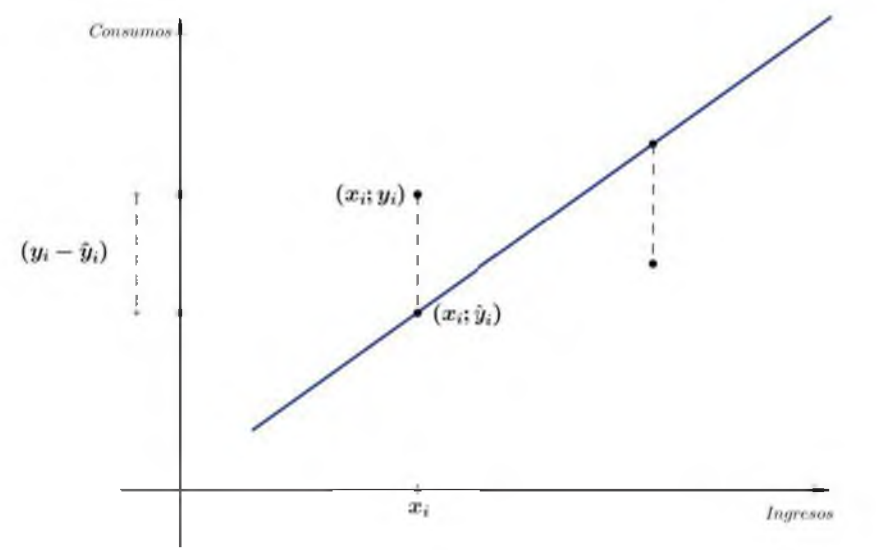


Aquí la muestra constituida por $n = 7$ casos permite intuir el crecimiento lineal del consumo al aumentar los ingresos. Sin embargo, como los puntos no están perfectamente alineados a partir de ellos no podrán hallarse con exactitud la pendiente y la ordenada al origen de la recta de regresión poblacional que, recordemos, supusimos pasando por los valores esperados de consumo según fueran los ingresos. En este caso la recta poblacional pasará por entre los puntos de la muestra, inclusive quizás conteniendo alguno, y, para estimarla, habrá que adoptar un criterio que garantice un "ajuste" eficiente con su comportamiento. Es decir; con base en las observaciones muestrales, hay que hallar una recta que se parezca lo más posible a la de regresión

poblacional desconocida. Para lograr esto se emplea el llamado *criterio de mínimos cuadrados* que estima la pendiente y la ordenada al origen de la recta utilizando las observaciones con las que se cuenta. Lo explicaremos a continuación.

Los puntos de ingreso y consumo que integran la muestra pueden denotarse por el par ordenado $(x_i; y_i)$. Como está claro, son observaciones mientras que los puntos $(x_i; \hat{y}_i)$, que tienen iguales ingresos que ellas, corresponden a los ubicados sobre la recta que se supone estimada. Por lo tanto la cantidad $\hat{y}_i$ es una estimación del consumo observado y_i cuando el ingreso es x_i. Esto se ilustra en la figura 4 que, a modo de ejemplo, exhibe la relación entre dos observaciones y sus respectivos valores estimados ubicados sobre la recta:

Figura 4.



Las diferencias entre valores observados y estimados algunas veces son positivas y otras negativas de modo que, como la recta buscada en teoría pasa por los promedios de consumo para cada nivel de ingreso, su suma resultará en promedio $\sum_{i=1}^{n} (y_i - \hat{y}_i) = 0$. La idea que está detrás de sumar los desvíos es hallar la recta $\hat{y}_i = \hat{\beta}_1 z_i + \hat{\beta}_0$ que minimice la suma, pues ella constituiría una buena estimación de la recta de regresión poblacional. Sin embargo vemos que, no importa cual sea la magnitud de cada desvío el resultado de la cuenta será 0. Podríamos considerar los valores absolutos $|y_i - \hat{y}_i|$ que al ser siempre positivos eliminarían la compensación apuntada

pero los módulos complicarían las cuentas ulteriores. Mejor que ello resulta utilizar la suma de cuadrados cuyos sumandos son todos positivos y facilitan el cálculo según veremos enseguida. En principio debe comprenderse que, como la recta en cuestión variará según los valores asignados a la pendiente $\hat{\beta}_1$ y a la ordenada al origen $\hat{\beta}_0$, la suma de cuadrados puede ser vista como una función de estas dos variables: $F\left(\hat{\beta}_1, \hat{\beta}_2\right) = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_1 x_i - \hat{\beta}_0\right)^2$ Habrá un valor de esta función para cada par de valores $\hat{\beta}_1$ y $\hat{\beta}_0$ ya que las otras cantidades x_i e y_i provienen de la muestra observada y son por ello conocidas. Entre los valores de la función F podrá haber alguno que sea mínimo. Es decir una cantidad que haga mínima la suma de cuadrados en la intención de ajustar lo mejor posible la recta a los puntos muestreados. Para hallar $\hat{\beta}_1$ y $\hat{\beta}_0$ que realizan este mínimo utilizamos las técnicas del cálculo diferencial y calculamos:

$$\frac{\partial F}{\partial \hat{\beta}_1} = 2\left(\sum_{i=1}^{n}\left(y_i - \hat{\beta}_1 x_i - \hat{\beta}_0\right)\right)(-x_i) = 0$$

$$\frac{\partial F}{\partial \hat{\beta}_0} = 2\left(\sum_{i=1}^{n}\left(y_i - \hat{\beta}_1 x_i - \hat{\beta}_0\right)\right)(-1) = 0$$

Operando obtenemos las llamadas ecuaciones normales:

$$\hat{\beta}_1 \sum_{i=1}^{n} x_i^2 + \hat{\beta}_0 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i$$

$$\hat{\beta}_1 \sum_{i=1}^{n} x_i + n\hat{\beta}_0 = \sum_{i=1}^{n} y_i$$

Ellas constituyen un sistema de dos ecuaciones con dos incógnitas $\hat{\beta}_1$ y $\hat{\beta}_0$, habida cuenta que las sumatorias se obtienen a partir de los datos aportados por la muestra.

Resolviendo, por ejemplo por determinantes, se obtienen las fórmulas:

$$\beta_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$\beta_0 = \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

Los valores así calculados de $\hat{\beta}_1$ y $\hat{\beta}_0$ son un punto crítico de la función $F\left(\hat{\beta}_1, \hat{\beta}_2\right) = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_1 x_i - \hat{\beta}_0\right)^2$. Aunque no abundaremos aquí,

un análisis más pormenorizado [1] de esta función, que resulta siempre mayor o igual que 0, revela que en tal punto alcanza un mínimo. De esta forma la recta hallada $y_i = \hat{\beta}_1 x_i + \hat{\beta}_0$ minimiza la suma de las diferencias al cuadrado y a partir de la muestra observada estima la recta de regresión poblacional. Por esta razón $\hat{\beta}_1$ y $\hat{\beta}_0$ son llamados *estimadores de mínimos cuadrados*. En el ejemplo que hemos utilizado hasta aquí para plantear el tema, la pendiente $\hat{\beta}_1$ estima la proporción destinada al consumo de cada peso ingresado. La ordenada al origen $\hat{\beta}_0$ estima el consumo mínimo, que igual se produce siendo el ingreso 0. En general, contamos entonces con un procedimiento para estimar los parámetros de una recta de regresión poblacional que intente modelar una relación entre variables de cualquier dominio de estudio. La recta obtenida es la que mejor ajusta, en *el sentido de mínimos cuadrados*, a los puntos que integran la muestra y por ello se denomina *recta de regresión muestral*.

Ejemplo 1: *Distintas teorías antropológicas sostienen que la altura de los hijos está en relación directamente proporcional con la altura de los padres. A efecto de comprobar esa relación entre padres e hijos para cierta comunidad se han tomado cinco observaciones de alturas según se ve en la tabla 2. A partir de ellas se desea construir un modelo matemático lineal que explique la altura de los hijos a partir de la altura de los padres:*

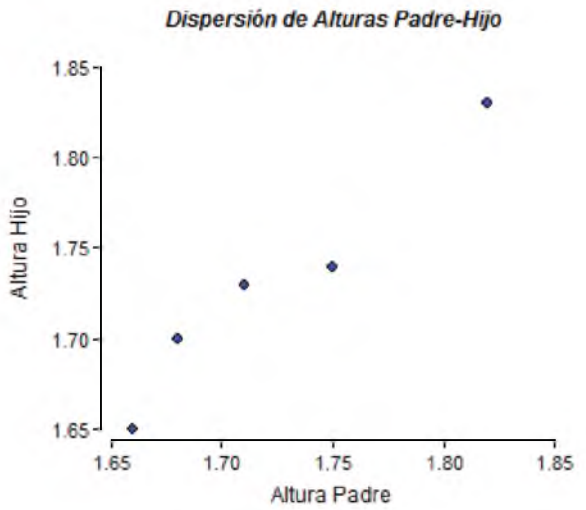
Tabla 2.

Altura del Padre (en metros)	*Altura del Hijo (en metros)*
1.68	1.70
1.82	1.83
1.75	1.74
1.66	1.65
1.71	1.73

Comenzamos por realizar un diagrama de dispersión de las observaciones como se ve en la figura 5:

[1] Véase por ejemplo Paul Meyer, ob.cit., p. 309.

Figura 5.



Se observa que hay una tendencia lineal de crecimiento de la altura de los hijos conforme crece la altura de sus padres. Esto hace pensar que sería razonable estimar un modelo de regresión lineal a partir de las observaciones muestrales aunque debemos adelantar que luego evaluaremos esta hipótesis de linealidad con mayor precisión. Por ahora calculamos los valores de la pendiente y ordenada al origen que surgen de los datos obtenidos.

$$n = 5$$

$$\sum_{i=1}^{5} x_i = 1{,}68 + 1{,}82 + 1{,}75 + 1{,}66 + 1{,}71 = 8{,}62$$

$$\sum_{i=1}^{5} x_i^2 = (1{,}68)^2 + (1{,}82)^2 + (1{,}75)^2 + (1{,}66)^2 + (1{,}71)^2 = 14{,}88$$

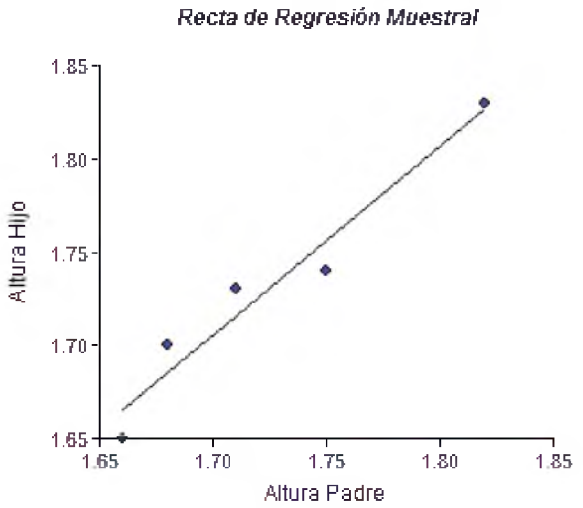
$$\sum_{i=1}^{5} y_i = 1{,}70 + 1{,}83 + 1{,}74 + 1{,}65 + 1{,}73 = 8{,}65$$

$$\sum_{i=1}^{5} x_i y_i = 1{,}68{\times}1{,}70{+}1{,}82{\times}1{,}83{+}1{,}75{\times}1{,}74{+}1{,}66{\times}1{,}65{+}1{,}71{\times}1{,}73 = 14{,}93$$

$$\hat{\beta}_1 = \frac{5 \times 14{,}93 - 8{,}62 \times 8{,}65}{5 \times 14{,}88 - (8{,}62)^2} = 0{,}91 \qquad \hat{\beta}_0 = \frac{14{,}88 \times 8{,}65 - 8{,}62 \times 14{,}93}{5 \times 14{,}88 - (8{,}62)^2} = 0{,}1$$

Obtenemos entonces la recta de regresión muestral $y = 0{,}91x + 0{,}16$ que estima, según mínimos cuadrados, la relación lineal poblacional que se supone existente entre alturas de padres e hijos. En la figura 6 puede apreciarse el ajuste de la recta con la nube de puntos:

Figura 6.



Establecida entonces la forma de estimación a través de la aplicación del método de cuadrados mínimos, cabe preguntarse sobre las propiedades de tal estimación. Y es aquí donde conviene explicitar ciertos supuestos realizados sobre el modelo de regresión lineal de la población. En primer lugar se supone que el modelo está *bien especificado*. Esto quiere decir que representa adecuadamente la relación poblacional entre las variables. Claramente, si no lo hiciera, la estimación de tal modelo carecería de objeto. Además se cuenta con que los datos muestreados son suficientemente representativos del comportamiento poblacional. En segundo término se supone que las perturbaciones aleatorias u_i son cantidades independientes de los valores de la variable explicativa x_i. Esto surge naturalmente del modelo poblacional que entiende a la variable explicada como suma del promedio de sus valores, condicionados al valor de x_i, más una perturbación aleatoria u_i. Es decir; $y_i = E(y/x_i) + u_i$ Si x_i y u_i estuvieran relacionadas no sería posible aislar sus efectos en dos sumandos separados. Por motivos similares se realiza una tercera suposición, también de independencia, entre las perturbaciones aleatorias u_i existentes bajo distintos valores de x_i. Esto significa que tales perturbaciones, consideradas como variables aleatorias, no están relaciona-

das y que por lo tanto los desvíos de la variable dependiente respecto de su promedio, ante un valor fijado x_i, son independientes de los desvíos dado cualquier otro valor x_j. La cuarta suposición ya se ha comentado y consiste en aceptar que la media de las perturbaciones aleatorias es 0. En fórmulas: $E(u_i/x_i) = 0$ Esto se acompaña con una quinta suposición que establece que la varianza de tales perturbaciones es la misma para cualquier valor de x_i. Es decir; las perturbaciones aleatorias deben tener media 0 y una varianza constante σ^2.Esta última condición de *homocedasticidad* implica dar a los datos muestrales observados una similar representatividad. Si para distintos valores de x_i las perturbaciones aleatorias tuviesen varianza distinta, no todos los puntos de la muestra serían igualmente confiables a efecto de la estimación[2]. Bajo las suposiciones realizadas es posible demostrar el importante resultado siguiente.

Propiedad 1: (*Gauss-Markov*) Si se cumplen las cinco suposiciones antedichas para el modelo de regresión lineal poblacional $y_i = \beta_1 z_i + \beta_0 + u_i$, las estimaciones $\hat{\beta}_1$ y $\hat{\beta}_0$, según mínimos cuadrados, son las mejores lineales e insesgadas (MELI).

Si bien la demostración de tal propiedad está fuera del alcance de este libro es importante destacar que el término lineal aplicado a los estimadores significa precisamente que resultan función lineal de las observaciones $y_1, y_2, \ldots, y_n$. Además son insesgados ya que puede probarse que $E(\hat{\beta}_1) = \beta_1$ y $E(\hat{\beta}_0) = \beta_0$. Finalmente puede demostarse también que si las perturbaciones aleatorias tienen distribución normal la distribución de los estadísticos $\hat{\beta}_1$ y $\hat{\beta}_0$ es también normal[3].

9.4. Correlación lineal

Dos variables aleatorias pueden estar relacionadas. Esto quiere decir que los valores que adopte una cualquiera de ellas están de alguna forma condicionados por los valores que adopte la otra. Una relación de tal tipo puede expresarse, con mayor o menor acierto, a través de una fórmula funcional que la represente y, en este sentido, por su simplicidad y fecundidad en las

[2]Para un desarrollo más extenso de estas suposiciones véase George Canavos, ob.cit., pp. 444-448

[3]Véase George Canavos, ob.cit., pp. 457-465

aplicaciones de la teoría, importa estudiar con detalle las relaciones lineales entre variables. Es decir; se trata de analizar el caso en que las variables aleatorias se asocian linealmente, aún cuando lo hagan solo en forma aproximada.

En primer lugar debe notarse que si las variables x e y no son independientes ocurre que no vale la fórmula general $P(xy) = P(x)P(y)$. Por otra parte la variabilidad conjunta de ambas variables podrá medirse por la llamada *covarianza* definida como $Cov(xy) = E[(x - E(x))(y - E(y))]$ que extiende naturalmente el concepto de varianza de una variable definido en el capítulo 2 según $Var(x) = \sum (x - E(x))^2 p(x) = E\left[(x - E(x))^2\right]$ Es claro que, si las variables fueran independientes, la variación de una de ellas no debería vincularse con la variación de la otra y entonces la varianza conjunta sería 0. Esto es precisamente lo que ocurre:

$$Cov(xy) = E[(x - E(x))(y - E(y))] = E[xy - xE(y) - yE(x) +$$

$$+E(x)E(y) = E(xy) - 2E(x)E(y) + E(x)E(y) = E(xy) - E(x)E(y) =$$

$$= E(x)E(y) - E(x)E(y) = 0$$

pues al ser independientes x e y se cumple, como ya se ha demostrado, que $E(xy) = E(x)E(y)$.

Es decir; si las variables son independientes la covarianza tendrá que ser 0. Sin embargo la recíproca no necesariamente es cierta. Puede darse el caso en que la covarianza sea 0 y, aún así, las variables estén relacionadas como se muestra en el siguiente ejemplo.

Ejemplo 2: *Sea x una variable aleatoria con la distribución de probabilidad dada por la tabla 3:*

Tabla 3.

x	$P(x)$
-2	0.25
-1	0.25
1	0.25
2	0.25

Sea la variable y relacionada en forma cuadrática con x a través de $y = x^2$. Hallar la covarianza entre ambas variables.

La distribución de probabilidad de la variable y resulta:

y	$P(y)$
1	0.50
4	0.50

Para calcular la distribución de la variable conjunta xy debe observarse que si $x = 1$ o $x = -1$ se tiene $y = 1$ mientras que si $x = 2$ o $x = -2$ se obtiene $y = 4$ de forma que los casos en que $x = \pm 1$ e $y = 4$ tanto como aquellos en que $x = \pm 2$ e $y = 1$ no pueden presentarse y deben tener probabilidad 0. En resumidas cuentas se tiene la distribución conjunta:

xy	$P(xy)$
$x = 1, y = 1$	0.25
$x = -1, y = 1$	0.25
$x = 2, y = 4$	0.25
$x = -2, y = 4$	0.25

De acuerdo con las distribuciones obtenidas calculamos:

$$E(x) = -1 \times 0{,}25 + 1 \times 0{,}25 + (-4) \times 0{,}25 + 4 \times 0{,}25 = 0$$

$$E(y) = 1 \times 0{,}5 + 4 \times 0{,}5 = 2{,}5$$

Y entonces obtenemos:

$Cov(xy) = E\left[(x - E(x))\,(y - E(y))\right] = E\left[xy - xE(y) - yE(x) + \right.$
$\left. + E(x)E(y) = E\left[xy - xE(y)\right] = E(xy) - E(x)E(E(y)) = E(xy) = \right.$
$= 1 \times 1 \times 0{,}25 + (-1) \times 1 \times 0{,}25 + 2 \times 4 \times 0{,}25 + (-2) \times 4 \times 0{,}25 = 0$

Es decir, tenemos covarianza 0 aún cuando las variables están relacionadas en forma cuadrática.

A la luz del ejemplo podemos preguntarnos que ocurre con la covarianza si las variables se relacionan linealmente siendo $y = \beta_1 x + \beta_0$. En tal caso resulta:

$$Cov(xy) = E\left[(x - E(x))\,((\beta_1 x + \beta_0) - E(\beta_1 x + \beta_0))\right] =$$

$$= \beta_1 E\left[(x - E(x))^2\right] = \beta_1 Var(x)$$

Por lo tanto, en caso de existir una relación lineal entre las variables, la covarianza quedará en función de la pendiente de esa expresión lineal.

Es claro que si tal pendiente fuera 0, es decir si no hubiera relación lineal entre las variables, la covarianza tendría que ser 0 como en el Ejemplo 2 donde la vinculación no es lineal sino cuadrática. Con el objetivo de obtener un coeficiente adimensional para medir el grado de correlación lineal entre variables, la covarianza puede dividirse por los desvíos estándar de cada una de ellas, consideradas aisladamente. Así se tiene la siguiente definición:

Coeficiente de correlación lineal: Se define el *coeficiente de correlación lineal* entre las variables x e y según:

$$\rho = \frac{Cov(xy)}{\sigma_x \sigma_y}$$

Propiedad 2: $-1 \leq \rho \leq 1$

Observemos primero que $Var(x) = E\left[(x - E(x))^2\right]$, $Var(y) =$
$= E\left[(y - E(y))^2\right]$ y $Cov(xy) = E\left[(x - E(x))\,(y - E(y))\right]$ Luego, para simplificar, anotamos $U = x - E(x)$ y $V = y - E(y)$. Vamos ahora a definir la expresión auxiliar $f(t) = E\left[(U + tV)^2\right]$ que es una función de t siempre mayor o igual que 0 pues es la esperanza del cuadrado $(U + tV)^2 \geq 0$. Es decir $f(t) \geq 0$ Ahora obsérvese que se trata de una función cuadrática dado que $f(t) = E\left[(U + tV)^2\right] = E\left[U^2 + 2tUV + t^2V^2\right] = E(U^2) + 2tE(UV) + t^2E(V^2)$

Entonces, al ser mayor o igual que cero tendrá raíces 0 o complejas por lo cual el discriminante de su formula resolutoria debe ser 0 o negativo. Es decir $4\left(E(UV)\right)^2 - 4E(U^2)E(V^2) \leq 0$ de donde resulta:

$$\frac{(E(UV))^2}{E(U^2)E(V^2)} = \frac{\left(E\left[(x - E(x))\,(y - E(y))\right]\right)^2}{E\left[(x - E(x))^2\right] E\left[(y - E(y))^2\right]} = \frac{(Cov(xy))^2}{Var(x)Var(y)} =$$

$$= \left(\frac{Cov(xy)}{\sigma_x \sigma_y}\right)^2 = \rho^2 \leq 1$$

Luego se tiene $-1 \leq \rho \leq 1$ que es lo que se quería demostrar.
Se tiene además el siguiente resultado:

Propiedad 3: Si las variables aleatorias x e y están en relación lineal $y = \beta_1 x + \beta_0$ entonces $\rho^2 = 1$ Además, en tal caso, si $\beta_1 > 0$ entonces $\rho = 1$ y si $\beta_1 < 0$ entonces $\rho = -1$ Recíprocamente si $\rho^2 = 1$, x e y están en

relación lineal con probabilidad 1.

No abundaremos aquí con la prueba de esta propiedad que se basa en la ya apuntada relación entre covarianza y pendiente[4]. Notemos sin embargo que cuando las variables están vinculadas linealmente con probabilidad 1 se habla de un suceso que ciertamente ocurre y, en tal caso, el coeficiente de correlación es 1 ó -1. Si el coeficiente adoptara un valor intermedio entre 1 y -1, la relación conjunta o *correlación* no sería perfectamente lineal y el valor de ρ obtenido resultaría una *medida del grado de asociación lineal* existente entre las variables. En el caso en que las variables no estuvieran relacionadas linealmente la pendiente de la recta tendría que ser 0 y, según ya se ha comentado, esto llevaría a una covarianza 0. Por lo tanto el coeficiente de correlación lineal resultaría $\rho = 0$.

Ejemplo 3: *La distribución de probabilidad conjunta de dos variables aleatorias y sus respectivas distribuciones marginales son las dadas por la tabla 4:*

Tabla 4.

x/y	$y=-3$	$y=2$	$y=4$	P_x
$x=1$	0.1	0.2	0.2	0.5
$x=3$	0.3	0.1	0.1	0.5
P_y	0.4	0.3	0.3	

Se desea establecer si las variables son independientes y, en caso de que no lo sean, analizar su grado de relación lineal.

La definición de independencia de variables requiere que para todo par de valores (x, y) se cumpla que $P(xy) = P(x)P(y)$. Es claro que, de acuerdo a la tabla 3, se tiene, por ejemplo, $0{,}3 = P(x = 3, y = -3) \neq P(x = 3) \times P(y = -3) = 0{,}5 \times 0{,}4 = 0{,}2$ De acuerdo a esto, las variables no son independientes. Ahora analicemos si observa algún grado de dependencia lineal utilizando el coeficiente ρ. Para ello tengamos presente que: $Cov(xy) = E\left[(x - E(x))\,(y - E(y))\right] = E\left[xy - xE(y) - yE(x) + E(x)E(y)\right] = E(xy) - E(x)E(y)$

Calculando tenemos:
$E(xy) = \sum_{i,j} x_i y_j P(x_i, y_j) = 1 \times (-3) \times 0{,}1 + 1 \times 2 \times 0{,}2 + 1 \times 4 \times 0{,}2 +$

[4]La demostración puede consultarse en Paul Meyer, ob.cit., pp. 150-151.

$3 \times (-3) \times 0{,}3 + 3 \times 2 \times 0{,}1 + 3 \times 4 \times 0{,}1 = 0$

$E(x) = 1 \times 0{,}5 + 3 \times 0{,}5 = 2 \quad E(y) = -3 \times 0{,}4 + 2 \times 0{,}3 + 4 \times 0{,}3 = 0{,}6$
y entonces $Cov(xy) = 0 - 2 \times 0{,}6 = -1{,}2$

Además como $E(x^2) = 1^2 \times 0{,}5 + 3^2 \times 0{,}5 = 0{,}5$ se tiene que $\sigma_x = \sqrt{E\left[(x - E(x))^2\right]} = \sqrt{(E(x)^2 - E^2(x))} = \sqrt{5 - 2^2} = 1$

Análogamente $E(y^2) = (-3)^2 \times 0{,}4 + 2^2 \times 0{,}3 + 4^2 \times 0{,}3 = 9{,}6$ y $\sigma_y = \sqrt{E\left[(y - E(y))^2\right]} = \sqrt{(E(y)^2 - E^2(y))} = \sqrt{9{,}6 - 0{,}6^2} = 3{,}04$
y entonces $\rho = \frac{Cov(xy)}{\sigma_x \sigma_y} = \frac{-1{,}2}{1 \times 3{,}04} = 0{,}39$

El valor del coeficiente de correlación lineal obtenido está a medio camino entre la correlación lineal perfecta que correspondería con $\rho = 1$ y la ausencia de correlación lineal que estaría indicada por $\rho = 0$. Por lo tanto debemos concluir que las variables consideradas tienen cierto grado de relación lineal entre ellas.

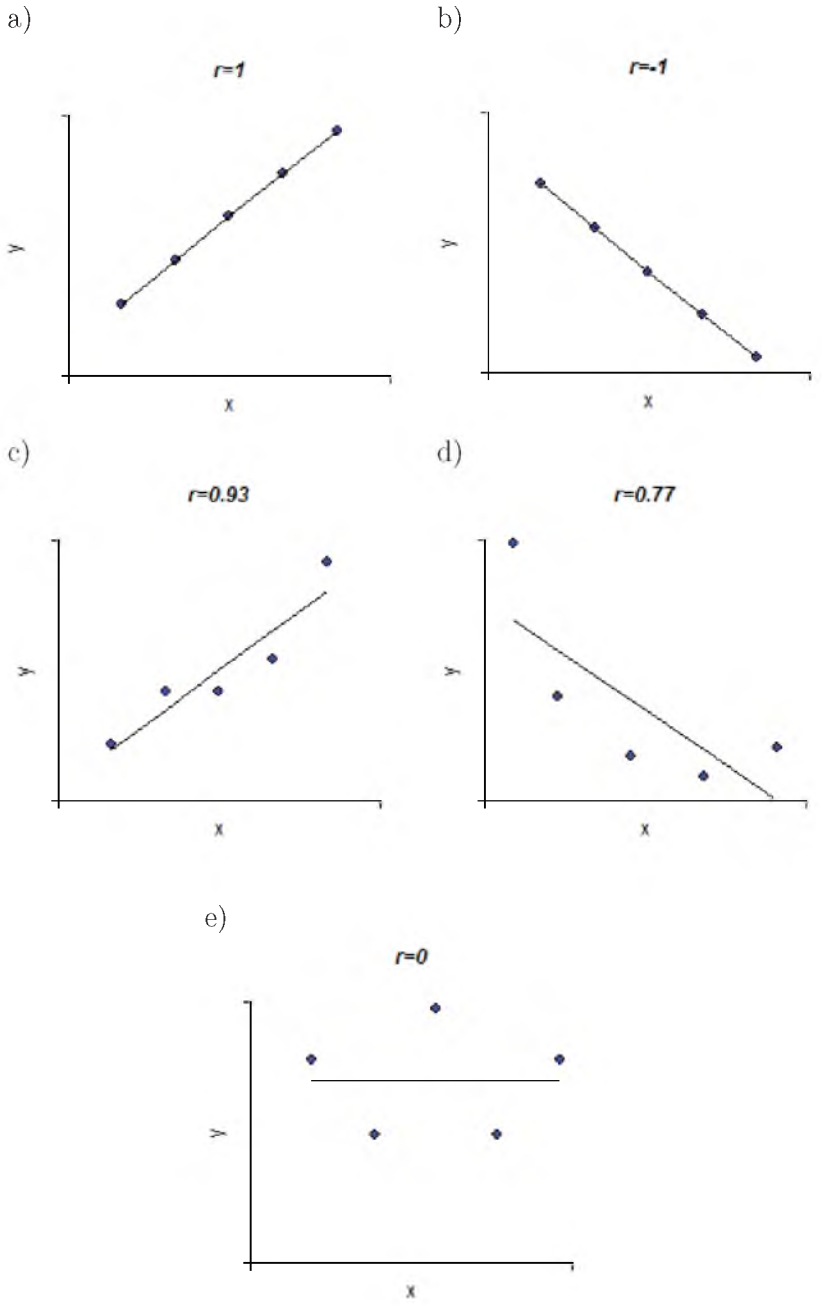
9.5. Regresión y correlación

De lo expuesto hasta aquí surge naturalmente la necesidad de analizar la vinculación entre la regresión lineal y la correlación lineal. Tal cual lo hemos explicitado la teoría de regresión busca establecer una fórmula lineal que explique, en términos "promedio", la relación entre las variables. Esta fórmula lineal se obtiene a partir de una muestra por consideraciones de tipo estadístico y está por ende sujeta a errores en la estimación que procedimientos como el de mínimos cuadrados procuran minimizar. Entre los supuestos de tal método se ha citado en primer término la necesidad de que el modelo lineal represente adecuadamente la dependencia estadística que hay entre las variables. Esto quiere decir que al emplear un modelo lineal para explicar una variable poblacional en función de otra debe esperarse un valor cercano a 1 o a -1 del coeficiente de correlación medido a nivel poblacional. Claramente si el modelo lineal elegido no pudiera expresar con propiedad la vinculación entre las variables, el coeficiente de correlación lineal poblacional tendría que acercarse hacia 0 tanto como dicho modelo dejara de explicar la relación entre variables. A esto hay que agregar además lo siguiente: la

relación estadística entre variables no implica necesariamente en los hechos una relación causal entre las magnitudes reales que ellas representan. Por ejemplo; pudiera ocurrir de un modo casual que la cantidad de automóviles que pasan por delante de un edificio comercial y la cantidad de palabras que se escriben en las computadoras de sus oficinas observaran una relación estadística lineal. Sin embargo, es muy difícil que se pueda afirmar que la causa de que se escriban cierta cantidad de palabras es que por delante del edificio pasa un determinado número de automóviles. En suma, la relación entre una causa real y un efecto real se asienta en una teoría, como por ejemplo la que vincula en la teoría económica el ingreso y el consumo y, en todo caso, solo después que se ha establecido el hecho teórico se intenta desarrollar un modelo estadístico que lo represente y compruebe. En general, razonar a la inversa, creyendo que el modelo justifica relaciones fácticas, no es válido.

Supongamos ahora que efectivamente estudiamos dos variables cuya relación causal se basa en una teoría pero que, como ocurre casi siempre, no tenemos todos los datos de su comportamiento poblacional y contamos sólo con una muestra. Si estamos interesados en modelar la relación por medio de una ecuación lineal podemos inferir su pendiente y la ordenada al origen por mínimos cuadrados tal como ya fue explicitado. Pero la cuestión es que la recta hallada a partir de la muestra asociará linealmente las variables de acuerdo al particular azar con que fue tomada y por ende la correlación medida en la muestra no necesariamente representará la correlación lineal entre las variables a nivel poblacional. Tenemos entonces dos problemas a resolver: evaluar la correlación lineal en la muestra e inferir a partir de ella la correlación lineal en la población.

Figura 7.



Llamaremos r al coeficiente de correlación lineal calculado a partir de

una muestra por medio de la fórmula:

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

que, puede demostrarse, corresponde al estimador de máxima verosimilitud del coeficiente de correlación lineal poblacional ρ definido en la sección anterior[5]. En forma análoga a la prueba realizada para el coeficiente de correlación lineal poblacional es posible demostrar que $-1 \leq r \leq 1$[6]. La figura 7 exhibe distintas muestras con su correspondiente recta de regresión calculada y el coeficiente de correlación muestral respectivo.

Los casos a) y b) muestran una relación lineal perfecta que se dice directa cuando $r = 1$ e inversa si $r = -1$. El caso c) exhibe una alta correlación lineal pues se observa claramente que la nube de puntos establece la tendencia de una recta casi en forma exacta. En cambio, en el caso d) la nube parece seguir más una línea parabólica que la recta de regresión hallada. Como es natural entonces el valor del coeficiente de correlación muestral se aleja de -1. En e) la disposición casi circular de los puntos fuerza la pendiente horizontal de la recta de regresión y el valor 0 para el coeficiente de correlación. Precisamente puede demostrarse[7] que si $\hat{y} = \hat{\beta}_1 x + \hat{\beta}_0$ es la recta de regresión calculada a partir de la muestra, se cumple la relación:

$$\hat{\beta}_1 = r \sqrt{\frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

de tal modo que cuando $r = 0$ se tiene $\hat{\beta}_1 = 0$ como se ha comentado en el caso poblacional.

Hasta aquí nuestro modelo de regresión muestral *explica* el comportamiento de la variable dependiente "y" precisamente como una función de la variable *regresora* "x". Esto en el terreno concreto de una aplicación se fundamenta en aspectos teóricos que postulan una relación causa-efecto que tiene esa forma. Pero desde el punto de vista estadístico, y aún sin significación en el campo de aplicación, puede muy bien hallarse una recta de regresión que exprese "x" como función de "y". En tal caso se tendrá la

[5] La deducción del estimador del coeficiente de correlación lineal puede verse en D. Montgomery, E. Peck y G. Vining: *Introducción a análisis de regresión lineal*, Grupo Editorial Patria, pp. 47-48

[6] Véase por ejemplo Fausto Toranzos: *Teoría estadística y aplicaciones*, Ediciones Macchi, 1997, pp. 249-250.

[7] Véase Fausto Toranzos, ob. cit.,pp. 248-249.

ecuación $\hat{x} = \hat{\beta}'_1 x + \hat{\beta}'_0$ y entonces:

$$\hat{\beta}'_1 = r\sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Si ahora efectuamos el producto:

$$\hat{\beta}_1\hat{\beta}'_1 = r\sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}r\sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = r^2$$

A veces para evaluar cuanto explica de las cantidades "y" el modelo lineal que las vincula con "x" se utiliza este *coeficiente de determinación* r^2 que toma valores entre 0 y 1 y puede expresarse por ello como porcentaje $r^2\,\%$.

En el supuesto que el valor del coeficiente de correlación r permita comprobar en la muestra una relación lineal, resta todavía establecer si puede concluirse que, en términos poblacionales, hay al menos cierto grado de asociación lineal entre las variables. Para esto podemos basarnos en la distribución del estadístico $\hat{\beta}_1$ y realizar una *prueba de significancia de la regresión.* $\hat{\beta}_1$ tiene una distribución normal y es insesgado[8], es decir $E(\hat{\beta}_1) = \beta_1$. Dado que usualmente se desconoce la varianza de los errores aleatorios entre los casos reales y los predichos por la recta de regresión, se realiza un test de hipótesis sobre la distribución t-student con $gl = n - 2$. La hipótesis nula es precisamente que no hay relación lineal entre las variables, lo que equivale a suponer que la pendiente de la recta de regresión poblacional es 0 pues así el valor de "y" no está vinculado linealmente al valor de "x". Se tiene:

$H_0 : \beta_1 = 0$

$H_a : \beta_1 \neq 0$

El estadístico de prueba en este caso es:

$$t_p = \frac{\hat{\beta}_1\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}{\sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{n-2}}}$$

dónde $\hat{y}_i = \hat{\beta}_1 x_i + \hat{\beta}_0$ Como en todo test de hipótesis si el estadístico de prueba cae dentro de la región de rechazo, se rechaza la hipótesis nula y se acepta la alterna. Al hacer tal cosa se acepta entonces que existe una

[8]Una prueba de tal resultado se realiza en D. Montgomery, E. Peck y G. Vining.ob.cit., pp. 24-25

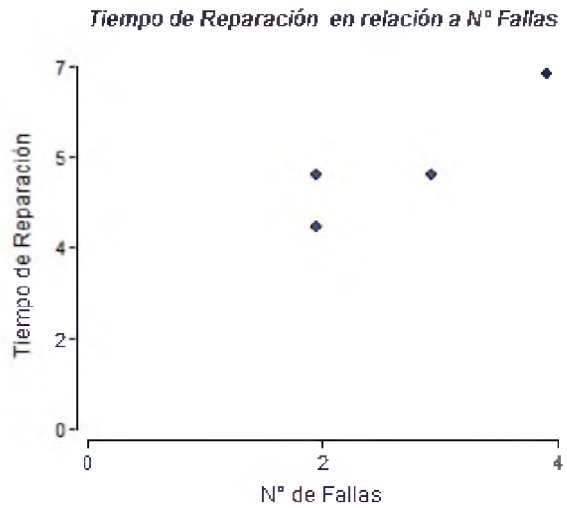
pendiente distinta de 0 de una recta de regresión lineal poblacional, que representará en algún grado la relación entre las variables[9].

Ejemplo 4: *Un equipo de mecánicos que prepara automóviles de carrera ha observado, para los últimos cuatro motores preparados, el número de fallas comprobadas y el tiempo en horas, empleado para la reparación completa de las mismas. Así ha resultado la tabla 5 con la que se pretende construir un modelo que explique en general el tiempo de reparación de motores como una función de la cantidad de fallas:*

	N°de Fallas	Tiempo de Reparación (hs)
Motor 1	3	5
Motor 2	2	4
Motor 3	2	5
Motor 4	4	7

Un diagrama de dispersión de los pares de valores observados para cada motor se muestra en la figura 8:

Figura 8.



La tendencia lineal de la nube de puntos sugiere pensar que en toda circunstancia, es decir con carácter poblacional, y no solo para esta mues-

[9]Una justificación y desarrollo más extenso del párrafo puede hallarse en George Canavos, ob.cit., pp. 465-488

tra, el tiempo de reparación se vincula linealmente con el número de fallas observadas en los motores. Sin embargo esto habrá que probarlo hallando primero la recta de regresión muestral y planteando luego la prueba de significancia de la regresión. Calculamos la recta de regresión muestral tomando como variable regresora al número de fallas y como variable dependiente al tiempo de reparación. Así resulta:

$\sum_{i=1}^{4} x_i = 11$, $\sum_{i=1}^{4} x_i^2 = 33$, $\sum_{i=1}^{4} y_i = 21$ y $\sum_{i=1}^{4} x_i y_i = 61$.

Además $\bar{x} = 2{,}75$ y $\bar{y} = 5{,}25$ por lo que se tiene:

$$\hat{\beta}_1 = \frac{4 \times 61 - 11 \times 21}{4 \times 33 - 11^2} = 1{,}18 \quad y \quad \hat{\beta}_0 = \frac{33 \times 21 - 11 \times 61}{4 \times 33 - 11^2} = 2$$

Se calcula:

$$\sum_{i=1}^{4} (x_i - \bar{x})^2 = (3-2{,}75)^2 + (2-2{,}75)^2 + (2-2{,}75)^2 + (3-2{,}75)^2 = 2{,}75$$

$$\sum_{i=1}^{4} (y_i - \bar{y})^2 = (5-5{,}25)^2 + (4-5{,}25)^2 + (5-5{,}25)^2 + (7-5{,}25)^2 = 4{,}75$$

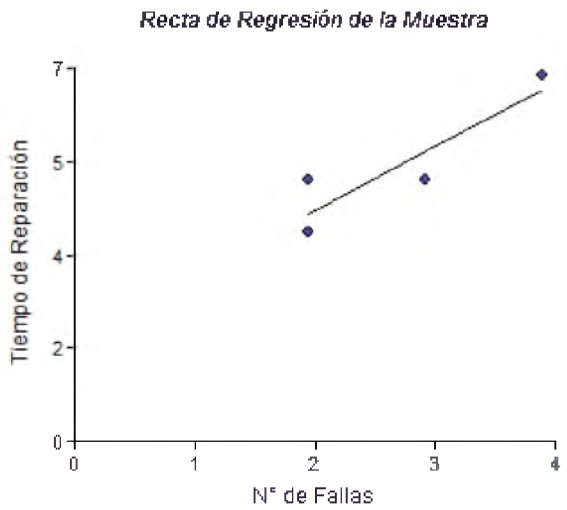
$\sum_{i=1}^{4} (x_i - \bar{x})(y_i - \bar{y}) = (3-2{,}75)(5-5{,}25) + (2-2{,}75)(4-5{,}25) + (2-2{,}75)(5-5{,}25) + (3-2{,}75)(7-5{,}25) = 3{,}25$

El coeficiente de correlación lineal de la muestra queda entonces:
$r = \frac{3{,}25}{\sqrt{2{,}75}\sqrt{4{,}75}} = 0{,}90$

La alta correlación lineal muestral se aprecia gráficamente al observar la figura 9 donde se ve la recta de regresión $\hat{y}_i = 1{,}18x_i + 2$:

Para utilizar este modelo lineal en la estimación de la demora en arreglar cualquier motor que presente entre 2 y 4 fallas realizamos la prueba de significancia sobre la regresión y observamos si corresponde rechazar la hipótesis de pendiente nula para la recta poblacional. En tal caso el modelo de regresión muestral será útil para predecir el comportamiento poblacional.

Figura 9.



$H_0 : \beta_1 = 0$
$H_a : \beta_1 \neq 0$
El estadístico de prueba en este caso es:

$$t_p = \frac{1{,}18 \times \sqrt{2{,}75}}{\sqrt{\frac{(5-(1{,}18\times 3)-2)^2+(4-(1{,}18\times 2)-2)^2+(5-(1{,}18\times 2)-2)^2+(7-(1{,}18\times 4)-2)^2}{4-2}}} = 2{,}90$$

En la t-student con $gl = 4 - 2$ el valor crítico correspondiente a una región de rechazo, que contiene $\alpha = 0{,}05$ de la probabilidad distribuida en dos colas, es $t_c = 4{,}303$ Se observa que el estadístico de prueba cae dentro de la región de aceptación. En efecto resulta $-4{,}303 < t_p = 2{,}90 < 4{,}303$ No puede entonces afirmarse que el modelo hallado sea significativo a nivel poblacional aunque en la muestra haya exhibido buena correlación lineal entre las variables.

9.6. Ejercicios

Ejercicio N°1*: Dados:

x	2	4	6	8	10
y	4	7	13	16	19

a) Realizar el diagrama de dispersión.

b) Hallar la recta de regresión que pone y en función de x.

c) Hallar el coeficiente de correlación.

Ejercicio N°2: La siguiente tabla presenta la producción mensual (en miles de unidades) de una fábrica y las utilidades (en millones de $) de los cinco primeros meses del año:

	Enero	Febrero	Marzo	Abril	Mayo
Producción	65	72	82	90	100
Utilidades	30	35	42	48	60

a) Realizar un diagrama de dispersión para mostrar las utilidades según el volumen de la producción.

b) Establecer un modelo lineal hallando la recta de regresión correspondiente.

c) Evaluar el grado de representatividad del fenómeno real que posee la recta hallada. (Considerar solo la muestra)

Ejercicio N°3*: Una compañía aeronáutica desea establecer una ecuación que le permita calcular la carga en dólares en función de las millas a recorrer. Para ello toma una muestra de 10 facturas de carga:

Distancia en Millas	14	23	9	17	10	22	5	12	6	16
Carga en Dólares	68	105	40	79	81	95	31	72	45	93

a) Realizar el diagrama de dispersión

b) Hallar la recta muestral y el coeficiente de correlación lineal respectivo.

c) Establecer que porcentaje de la variación queda explicado por el modelo lineal

d) ¿Los resultados son válidos en términos poblacionales?

Ejercicio N°4: Una serie de tiempo toma valores cada vez que pasa un lapso fijado de tiempo. Así es por ejemplo la serie que muestra el crecimiento de la población de acuerdo al paso del tiempo. Como toda serie, esta reconocerá una tendencia principal que explica esencialmente el fenómeno, sumada a tendencias estacionales, cíclicas y aleatorias que también se manifiestan en algún grado. Buscar en Internet los datos de los últimos 6 censos de Población Argentina y diseñar un procedimiento para evaluar la tendencia principal. En particular se desea establecer cual es la propensión decenal de crecimiento de la población.

Capítulo 10

Software estadístico

10.1. Introducción

El objetivo del capítulo es introducir al lector en el uso de software estadístico disponible para efectuar un análisis de datos. Por supuesto no se trata aquí de presentar un listado exhaustivo de los paquetes disponibles ni tampoco de desarrollar destreza en la utilización particular de alguno de ellos. Más bien procuramos mostrar las potencialidades que tiene su aplicación referida a los temas expuestos en los capítulos anteriores y motivar así el uso de la herramienta informática. Sin embargo, hay que tener en cuenta que lo habitual es que los datos sean multidimensionales. Esto quiere decir que cada unidad experimental, cada instancia de una base de datos tiene varios atributos o variables para registrar. Por ejemplo, en la base de datos de los estudiantes de una universidad por cada estudiante se registra el número de documento, domicilio, teléfono, género, y eventualmente algunos datos referidos a salud como altura y peso. Además se puede incorporar la nota por cada materia aprobada y el promedio de notas, el año de inscripción y el año de egreso. En suma se estarían considerando al menos 9 variables más las que correspondan a la nota de cada materia, digamos por ejemplo 40. En total se trata con 49 variables. Algunas de ellas, como las notas, serán cuantitativas pero otras resultan nominales o categóricas como el género. Algunas estarán correlacionadas, otras serán independientes entre si. Todas tendrán su tendencia central, su variabilidad, su grado de simetría en la distribución y el análisis conjunto de estos parámetros y relaciones corresponde a la estadística multivariada. Los contenidos desarrollados en los capítulos anteriores se ocupan mayoritariamente de estadística univariada y solo en algunos de ellos se estudian conjuntamente dos o más variables

como en Análisis de la variancia o en regresión y correlación. De modo que todo lo dicho es, en la práctica, sólo una imprescindible introducción para que el futuro ingeniero comprenda conceptualmente y conozca el lenguaje de las técnicas estadísticas profesionalmente utilizadas. En el mismo sentido entonces presentamos aquí el uso del software.

Antes que nada, hay que mencionar que existe software libre disponible directamente en Internet en forma legal y software que corre bajo licencia paga. Hay también una situación intermedia: versiones libres o estudiantiles con licencia casi sin costo de paquetes informáticos que en su versión profesional son de licencia paga. Esto permite apreciar las posibilidades del software sin necesidad de adquirirlo previamente. Además de las planillas de cálculo, que pueden realizar una cantidad importante pero básica de procesos estadísticos, entre el software que corre bajo licencia se puede citar a SPSS, SAS o SPAD paquetes muy potentes y especialmente diseñados para el proceso estadístico. En Argentina, el Grupo Infostat de la Facultad de Ciencias Agrarias de la Universidad Nacional de Córdoba ha desarrollado INFOSTAT con amplias e interesantes funcionalidades que dispone de una versión estudiantil liberada para uso de docentes y alumnos. En cuanto al software libre corresponde citar al R, desarrollado a partir de 1997 por Ihaka y Gentlemen que actualmente cuenta con cientos de miles de usuarios a nivel mundial y cuyos usos y funcionalidades continúan en expansión.

Los paquetes existentes trabajan de formas distintas. Algunos exhiben una pantalla amigable en la que se muestra una tabla con los datos y se dispone de un menú de procesos y gráficos al estilo de la presentación de una planilla de cálculo usual. Otros presentan una consola con una interfaz de línea de instrucciones. En este caso permiten programar una secuencia de tales instrucciones que ejecute la lectura de los datos, los procesos de acuerdo a lo requerido y exhiba los resultados.

10.2. Distribuciones de frecuencias y gráficos

Para comenzar vamos a presentar un conjunto de datos muy famoso. Se trata de la base IRIS introducida por R. Fischer en 1936 para ejemplificar cierto análisis. Está constituida por datos de 150 flores de 3 especies diferentes y como se puede apreciar en la tabla 1 extractada del Anexo E, las variables consideradas son el nombre de la especie, el largo del sépalo, su ancho, el largo del pétalo y su ancho:

Tabla 1.

Especie	**Largo Sépalo**	**Ancho Sépalo**	**Largo Pétalo**	**Ancho Pétalo**
I. setosa	5.1	3.5	1.4	0.2
I. setosa	4.9	3	1.4	0.2
I. setosa	4.7	3.2	1.3	0.2
I. setosa	4.6	3.1	1.5	0.2
I. setosa	5	3.6	1.4	0.2
I. setosa	5.4	3.9	1.7	0.4
I. setosa	4.6	3.4	1.4	0.3
I. setosa	5	3.4	1.5	0.2
I. setosa	4.4	2.9	1.4	0.2
I. setosa	4.9	3.1	1.5	0.1

La figura 1 muestra una flor cualquiera con sus pétalos y sus sépalos:

Figura 1.

Al considerar la tabla 1 se aprecia que la variable Especie es categórica. Una de sus categorías es precisamente *I. setosa* pero, si se observa la tabla completa del Anexo E se ve que existen también las categorías *I. versicolor* e *I. virginica*. Las demás variables son, en este caso, cuantitativas. Hay que aclarar que en la terminología de las bases de datos cada fila es una *instancia* o un *caso* de los que integran la base y, como resulta claro, no es posible cambiar el orden en una sola variable sin reordenar también las

demás solidariamente para que todas las cantidades de la fila pertenezcan a la misma flor.

Si en un buscador de Internet colocamos, por ejemplo, las palabras "Fischer Conjunto Iris" obtendremos un listado de direcciones desde las cuales podemos bajar o simplemente copiar la base IRIS. Supongamos entonces que así lo hemos hecho y que sobre una planilla de cálculo tenemos ya los datos de las 150 flores. Ahora probablemente deseemos usar un software específico de estadística para graficar y analizar estos datos. Vamos a utilizar la versión estudiantil de INFOSTAT que puede obtenerse en forma gratuita de la referencia citada en la bibliografía.

La pantalla de INFOSTAT ofrece un menú de opciones ubicado en la parte superior de una planilla que inicialmente no contiene datos. Para cargar en ella la base IRIS, siempre que se encuentre almacenada en un formato compatible, basta con seleccionar la opción Archivo-Abrir y señalar el archivo correspondiente. Con los datos visibles en la pantalla se puede entonces comenzar a trabajar.

Si consideramos una variable, por ejemplo Largo Sépalo, establecemos su distribución de frecuencias seleccionando la opción Estadísticas-Tablas de Frecuencias. La tabla 2 muestra el resultado respectivo dónde la cantidad de clases han sido seleccionadas en forma automática por INFOSTAT.

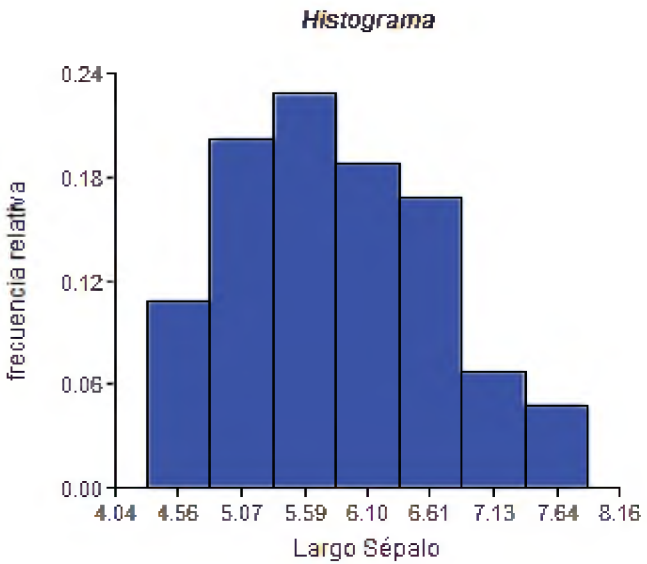
Tabla 2.

Tablas de frecuencias

Variable	Clase	LI	LS	MC	FA	FR
Largo Sépalo	1	[4.30	4.81]	4.56	16	0.11
Largo Sépalo	2	(4.81	5.33]	5.07	30	0.20
Largo Sépalo	3	(5.33	5.84]	5.59	34	0.23
Largo Sépalo	4	(5.84	6.36]	6.10	28	0.19
Largo Sépalo	5	(6.36	6.87]	6.61	25	0.17
Largo Sépalo	6	(6.87	7.39]	7.13	10	0.07
Largo Sépalo	7	(7.39	7.90]	7.64	7	0.05

LI y LS son los límites inferior y superior de cada una de las 7 clases, MC la marca de clase, FA la frecuencia absoluta y FR la frecuencia relativa. Podemos también graficar la distribución utilizando la opción Gráficos-Histograma para la misma variable Largo Sépalo según se muestra en la figura 2:

Figura 2.



Como se podrá apreciar al ensayar tales procedimientos con el software INFOSTAT, existen una serie de variantes que permitirían por ejemplo disminuir el número de clases en la distribución de frecuencias o cambiar las escalas en los gráficos, por citar sólo dos de una gran cantidad de opciones. Sin embargo nuestro objetivo no es aquí enseñar a utilizar todas las alternativas que ofrece el software sino simplemente enfocar su uso en relación con los contenidos de los capítulos anteriores. Se deja entonces abierta al lector la posibilidad de explorar variantes por si mismo y adquirir familiaridad con la herramienta. Mantendremos este criterio a lo largo del presente capítulo.

Consideremos ahora la variable cualitativa Especie. Su distribución de frecuencias es la de la tabla 3:

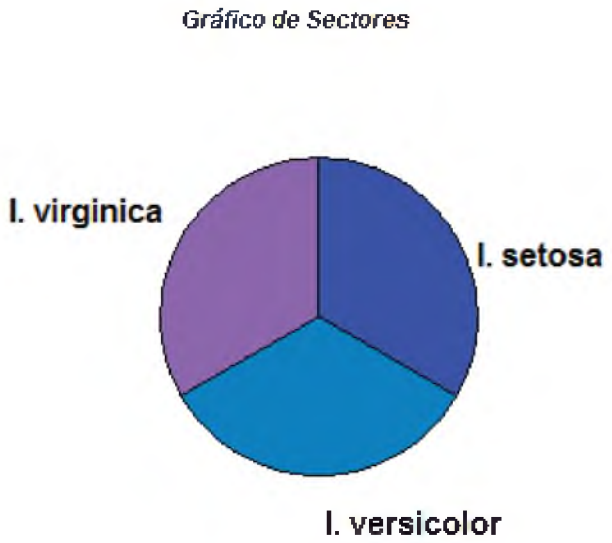
Tabla 3.

Tablas de frecuencias

Variable	Clase	FA	FR
Especie	1	50	0.33
Especie	2	50	0.33
Especie	3	50	0.33

El respectivo diagrama en forma de pastel se aprecia en la figura 3:

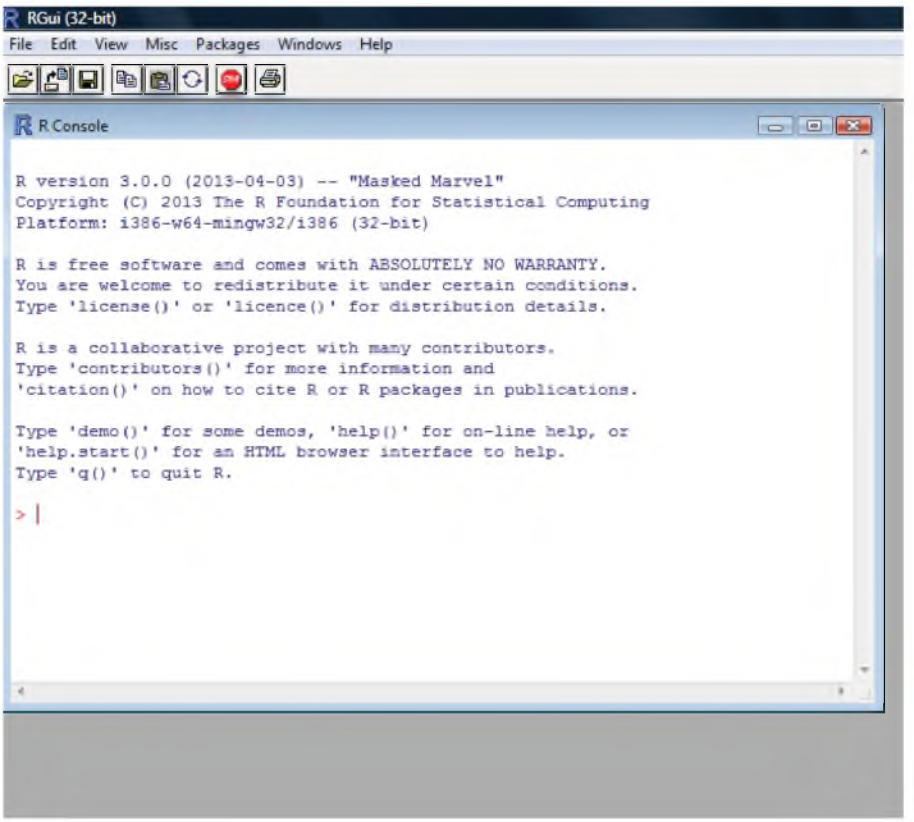
Figura 3.



Otra alternativa para realizar análisis estadísticos es la ofrecida por el software libre R que puede bajarse de la dirección señalada en la bibliografía. R opera en forma un poco distinta a lo que hemos visto hasta aquí. En efecto, luego de cargarlo queda disponible en el directorio correspondiente un archivo que al ser ejecutado abre la *consola* del soft. En esa consola habrá que escribir las instrucciones acerca de lo que quiere realizarse, los comandos R, y en ella también podrán leerse los resultados de estas acciones.

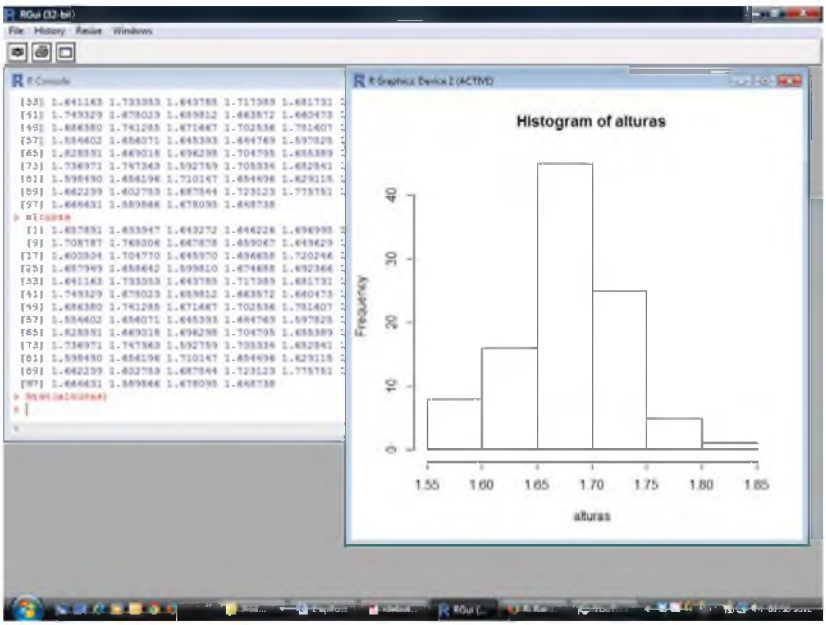
Ejecutamos entonces el archivo mencionado, cuyo nombre varía según la versión de R que usemos (por ejemplo R i386 3.0.0), y aparece la pantalla sobre la cual trabajaremos como se ve en la figura 4:

Figura 4.



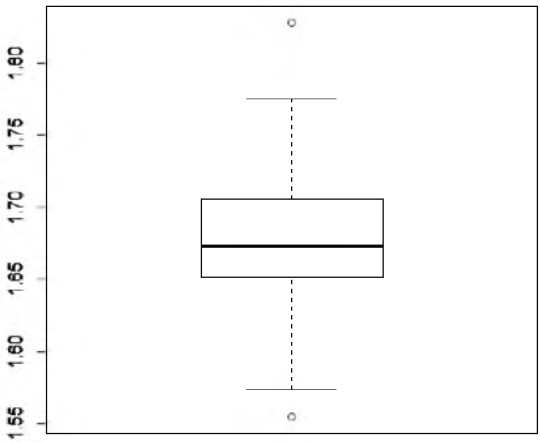
Vamos a generar ahora un conjunto de datos para trabajar con R. Si en la consola escribimos `datos <- 1:50` generamos una sucesión de números naturales desde el 1 hasta el 50 que se guarda en el archivo "datos". Para ver el contenido de "datos" basta con escribir la instrucción `datos` y aparecerá en la consola el listado completo de los números. Supongamos ahora que deseamos contar con 100 datos de altura de personas de forma tal que su distribución de frecuencias tenga un perfil aproximadamente normal con un promedio de 1.68 metros y un desvío estándar de 0.05 metros. Podemos escribir en la consola `alturas <- rnorm(100,1.68,0.05)` y quedarán almacenados 100 datos de altura distribuidos en forma normal con la media y el desvío estándar requeridos. Podemos graficar el histograma de frecuencias de las alturas mediante la instrucción `hist(alturas)`. En R los gráficos se muestran en una ventana auxiliar como se ve en la figura 5:

Figura 5.



Si escribimos en la consola de R el comando boxplot(alturas) obtenemos un diagrama de caja que permite visualizar la forma de la distribución en la figura 6:

Figura 6.



En el diagrama, la caja tiene por lado superior al tercer cuartil de los datos y por lado inferior al primer cuartil. La línea central representa a la mediana. Las rama inferior del bigote culmina en otra línea colocada al restar al primer cuartil 1.5 veces el el rango intercuartílico (diferencia entre el cuartil 3 y el 1). La rama superior termina en la línea trazada a 1.5 veces el rango intercuartílico medido desde el tercer cuartil. Los circulitos por fuera son datos que se denominan *atípicos* o *outliers*. Este tipo de datos puede influir severamente en el cálculo de estimaciones, por ejemplo en el caso de la media que es muy sensible a los extremos, razón por la cual a veces se los elimina del análisis.

10.3. Estadística descriptiva

Las distintas medidas de tendencia central, de variabilidad, asimetría y otras pueden calcularse fácilmente al utilizar software. Para el caso de la variable Largo Sépalo de la base IRIS calculamos por ejemplo el número total de casos, la media, el desvío estándar, la mediana y la cantidad de datos faltantes. Este último dato es relevante en los estudios estadísticos multivariados pues bien puede ocurrir que para una determinada flor se hayan registrado la longitud y el ancho del pétalo, el ancho del sépalo, la especie correspondiente y sin embargo no se cuente con el largo del sépalo, en cuyo caso ése es precisamente un dato faltante. La tabla 4 muestra los valores calculados con INFOSTAT al seleccionar la opción `Estadísticas-Medidas Resumen`.

Tabla 4.

`Medidas resumen`

Variable	n	Media	D.E.	Mediana	Datos faltantes
Largo Sépalo	150	5.84	0.83	5.84	0

Con R podemos hacer cálculos similares. Para nuestro conjunto de 100 alturas, al escribir `mean(alturas)` , `sd(alturas)`, `median(alturas)` obtenemos la media, el desvío estándar y la mediana respectivamente. La imagen de la figura 7 nos muestra la escritura de los comandos y sus resultados sobre

la consola.

Figura 7.

```
   >mean(alturas)
[1] 1.678084
>sd(alturas)
[1] 0.04772495
>median(alturas)
[1] 1.673177
```

Como se ve, los valores obtenidos no son exactamente los requeridos al generar el conjunto pues esos parámetros eran media 1.68 y desvío estándar 0.05. La razón de estos pequeños desvíos es que la cantidad de 100 observaciones es una muestra de una población distribuida normalmente con tales parámetros, no la población completa. Se trata de una muestra que no es suficientemente grande como para que los estadísticos coincidan exactamente con los valores de media y desvío estándar poblacional. Como la distribución de los datos se aproxima a la normal sin alcanzarla aún, también es natural que difieran la media y la mediana.

Un aspecto interesante de R es que puede programarse la ejecución de varias instrucciones mediante la confección de un *script*. Este consiste en un archivo de texto, un .txt escrito con el Bloc de Notas por ejemplo, donde se anotan los comandos en una sucesión de líneas. Así por ejemplo construiríamos el archivo Script.txt listando las instrucciones:

```
   mean(alturas)
sd(alturas)
median(alturas)
```

A continuación podríamos pegar su contenido en la consola obteniendo como respuesta todos los valores pedidos simultáneamente. éste es uno de los procedimientos básicos para desarrollar programas en R pues los scripts pueden almacenarse y ejecutarse cada vez que se desee.

10.4. Test de hipótesis

Si, como vimos, el conjunto de 100 alturas se considera una muestra de las alturas de una población, es posible testear hipótesis que se realicen, por ejemplo, sobre el valor de la media poblacional. En la consola de R podemos escribir entonces `t.test(alturas, alternative = â€œtwo.sidedâ€, mu=1.68, conf.level=.95)` y la ejecución nos devuelve la salida de la figura 8:

Figura 8.

```
One Sample t-test
data: alturas
t = -0.4014, df = 99, p-value = 0.689
alternative hypothesis: true mean is not equal to 1.68
95 percent confidence interval:
1.668615 1.687554
sample estimates:
mean of x
1.678084
```

En este caso para la ejecución del comando se han colocado valores a los argumentos que dan forma a su acción. Luego de citar el conjunto `alturas`, el argumento `alternative="two.sided"` establece que el test de dos colas, `mu=1.68` fija la hipótesis nula en ése valor y `conf.level=.95` establece el nivel de confianza del intervalo de estimación.

La convergencia de la distribución t-student a la normal conforme crece el número de grados de libertad permite efectuar el test t cualquiera sea el tamaño muestral de modo que en todos los software se utiliza la t- student para testear hipótesis sobre la media poblacional. En la salida R de la figura 8 la cantidad `t = -0.4014` es el estadístico de prueba calculado a partir de la muestra de 100 observaciones de alturas, los grados de libertad son `df = 99` y el valor p `p-value = 0.689` nos indica la probabilidad de cometer un Error de Tipo 1, rechazar la hipótesis nula cuando es verdadera, si decidimos efectivamente rechazarla. Por supuesto en este caso y dado el valor de p no la rechazamos. Finalmente el intervalo de confianza de 95 % para la media poblacional es `1.668615 1.687554` y la media muestral `1.678084` como por otra parte ya habíamos calculado.

Consideremos ahora los datos de IRIS y realicemos la hipótesis de que

la media de Largo Sépalo es mayor o igual que 6.5. Para testearla empleamos INFOSTAT seleccionando `Estadísticas-Inferencia basada en una muestra-Prueba t para una media`. En el cuadro de diálogo colocamos Largo Sépalo como variable activa y luego `Prueba unilateral izquierda` y `Parámetro en 6.5`. Al ejecutar obtenemos los resultados de la figura 9:

Figura 9.

```
Prueba t para una media
Valor de la media bajo la hipótesis nula: 6.5
```

Variable	n	Media	D.E.	LS(95)	T	p(Unilateral I)
Largo Sépalo	150	5.84	0.83	5.96	-9.71	<0.0001

Luego de anotar el tamaño, la media y el desvío estándar muestrales, INFOSTAT nos informa que el límite superior para el intervalo de confianza del 95 % (porcentaje default) es 5.96. El valor p obtenido aconseja rechazar la hipótesis nula. En cambio si nuestra hipótesis nula fuera, por ejemplo, $\mu \geq 5{,}5$ los resultados informados serían los de la figura 10 y entonces, de acuerdo al valor p calculado no debiéramos rechazar la hipótesis nula.

Figura 10.

```
Prueba t para una media
Valor de la media bajo la hipótesis nula: 0
```

Variable	n	Media	D.E.	LS(95)	T	p(Unilateral I)
Largo Sépalo	150	5.84	0.83	5.96	86.43	>0.9999

Supongamos ahora que pretendemos realizar un test de Bondad de Ajuste para saber si podemos aceptar la hipótesis que la variable poblacional Largo Sépalo tiene distribución normal. Debemos realizar entonces el test Chi-Cuadrado correspondiente. Con INFOSTAT seleccionamos Estadísticas-Tablas de frecuencias y en el cuadro de diálogo colocamos Largo Sépalo como la variable activa. A continuación en la solapa Ajustes elegimos las opciones normal y chi-cuadrado. Los resultados se ven en la figura 11.

Figura 11.

Tablas de frecuencias

Ajuste: Normal con estimación de parámetros: Media= 5.84333 y varianza= 0.68569

Variable	Clase	LI	LS	MC	FA	FR	E(FA)	E(FR)	Chi-Cuadrado p
Largo Sépalo	1	[4.30	4.81]	4.56	16	0.11	16.05	0.11	1.4E-04
Largo Sépalo	2	(4.81	5.33]	5.07	30	0.20	24.02	0.16	1.49
Largo Sépalo	3	(5.33	5.84]	5.59	34	0.23	34.90	0.23	1.51
Largo Sépalo	4	(5.84	6.36]	6.10	28	0.19	34.91	0.23	2.88
Largo Sépalo	5	(6.36	6.87]	6.61	25	0.17	24.04	0.16	2.92
Largo Sépalo	6	(6.87	7.39]	7.13	10	0.07	11.39	0.08	3.09
Largo Sépalo	7	(7.39	7.90]	7.64	7	0.05	4.69	0.03	4.23
									0.3755

Para testar el ajuste el software eligió en forma automática 7 clases cuyos límites inferior y superior son LI y LS respectivamente, con MC marca de clase. FA es la frecuencia absoluta observada en cada intervalo y FR la relativa. Los valores E(FA) y E(FR) son los que debieran esperarse para ambos tipos de frecuencia en cada clase si la distribución fuera normal con media 5.84333 y varianza 0.68569. El valor $p = 0{,}3755$ obtenido es la probabilidad de rechazar la hipótesis nula cuando esta es verdadera y es demasiado alto como para efectuar el rechazo. Se concluye entonces que la variable Largo Sépalo se distribuye normalmente con la media y la varianza estimadas.

10.5. Análisis de la varianza

Consideremos Largo Sépalo como la variable respuesta que representa la característica de interés al estudiar las flores de una región donde es posible encontrar las tres especies *I. setosa, I. versicolor e I. virginica.* En el lenguaje del análisis de la varianza podemos interpretar que la variable Especies es un factor y que cada una de sus categorías es un tratamiento cuyo efecto es fijo sobre la longitud del sépalo. En ese contexto cabe entonces preguntarse si las medias de Largo Sépalo para cada tratamiento son iguales, es decir si es la misma para cada especie de flor en particular. En INFOSTAT seleccionamos Estadísticas-Análisis de la Varianza y en el cuadro de diálogo colocamos Largo Sépalo como variable activa y Especie como variable de clasificación. Si en el cuadro de diálogo siguiente, en la solapa Comparaciones elegimos la opción Ninguna para Método de Comparación, no realizaremos

ningún análisis post-hoc. Los resultados son los de la figura 12.

Figura 12.

Análisis de la varianza
Cuadro de Análisis de la Varianza (SC tipo III)

F.V.	SC	gl	CM	F	p-valor
Modelo	63.21	2	31.61	119.26	<0.0001
Especie	63.21	2	31.61	119.26	<0.0001
Error	38.96	147	0.27		
Total	102.17	149			

La columna SC exhibe las sumas de cuadrados, los grados de libertad de numerador y denominador y también los respectivos cuadrados medios con los que se calcula el valor del estadístico de Fischer. El valor p obtenido indica que debe rechazarse la hipótesis de que los efectos de cada uno de los tratamientos son nulos ó, lo que es equivalente, rechazar que las medias de la variable Largo Sépalo para las tres especies son las mismas. Esto motiva la necesidad de un análisis post-hoc pues no sabemos para que par de ellas las diferencias son significativas. En INFOSTAT procedemos como hicimos hasta aquí, pero al llegar al segundo cuadro de diálogo elegimos en `Métodos de Comparación` de la solapa `Comparaciones` la opción `LSD Fischer` que corresponde al método de análisis post-hoc explicado en el capítulo 8. Además allí tildamos `Gráfico de Barras` para obtener los resultados de las figuras 13 a) y 13 b):

Figura 13 a)

Análisis de la varianza
Cuadro de Análisis de la Varianza (SC tipo III)

F.V.	SC	gl	CM	F	p-valor
Modelo	63.21	2	31.61	119.26	<0.0001
Especie	63.21	2	31.61	119.26	<0.0001
Error	38.96	147	0.27		
Total	102.17	149			

Test:LSD Fisher Alfa=0.05 DMS=0.20347

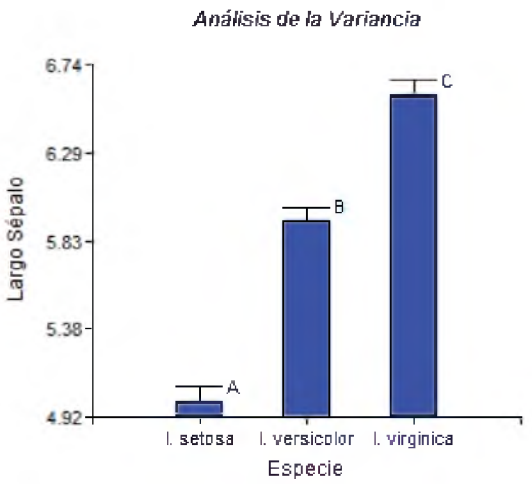
Error: 0.2650 gl: 147

Especie	Medias	n	E.E.			
I. setosa	5.01	50	0.07	A		
I. versicolor	5.94	50	0.07		B	
I. virginica	6.59	50	0.07			C

Medias con una letra común no son significativamente diferentes (p >0.05)

Las letras A, B y C indican que cada media resulta significativamente diferente de las otras dos. Bien podría haber ocurrido que dos de ellas pudieran considerarse iguales y otra diferente con lo cual hubieran aparecido, por ejemplo, dos letras A y una B.

Figura 13 b)



En la figura 13 b) desde un valor tomado como base las barras muestran la variación entre las medias.

En el análisis de la varianza aquí presentado se ha supuesto que los errores aleatorios de las observaciones son independientes. Además se ha supuesto que las variables Largo Sépalo en cada tratamiento son normales y que también tienen igual varianza. Se sugiere al lector analizar el grado de cumplimiento de estos dos últimos supuestos utilizando INFOSTAT.

10.6. Regresión y correlación lineal

Dado el conjunto IRIS nos proponemos ahora estudiar una posible asociación lineal entre el largo del sépalo y su ancho para las flores de la región dentro la cual se tomó la muestra. Pero intentamos trabajar con R por lo que primero debemos leer los datos con este software. Un camino sencillo es, por ejemplo, seleccionar con INFOSTAT ambas variables y guardarlas como un archivo de texto que luego sea leído por R. Para hacer esto primero eliminamos las columnas Especie, Largo Pétalo y Ancho Pétalo seleccionando cada columna y luego optando por `Eliminar columna` en el menú del botón derecho del mouse. Además hay que corregir el nombre de las columnas restantes juntando sus dos partes constituyentes para no confundir la forma de separación de datos en R. Quedarán entonces LargoSépalo y AnchoSépalo. El archivo resultante debe ser guardado en el mismo directorio que la consola de R con el nombre IRIS, pero ahora con la extensión .txt Para ello se utiliza la selección `Archivo-Guardar tabla como` y el separador default `tabulador`. Una vez hecho esto nuestros datos están listos para ser leídos por R y alojados en un archivo conveniente. Ejecutamos en la consola la siguiente sucesión de instrucciones.

```
   >Iris<-read.table ( Iris.txt, header=TRUE)
names(Iris)
Iris
```

El comando `names(Iris)` permitirá ver los nombres de las variables ahora cargadas en R e `Iris` listará todo el archivo en la consola lo que servirá para asegurarnos que hemos obrado sin cometer errores en la sintaxis de los comandos. Ahora estamos en condiciones de calcular con R una recta de regresión por mínimos cuadrados que vincule ambas variables. En la consola escribimos:

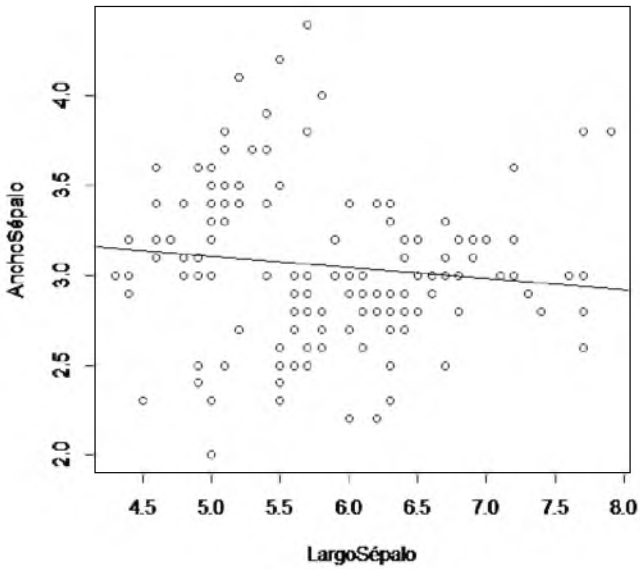
```
   >lm(formula = AnchoSépalo ~ LargoSépalo, data = Iris)
```

La instrucción detalla que se utilizará un modelo lineal (`lm`=linear model), que la variable AnchoSépalo será la explicada por LargoSépalo (`y~x`) y que los datos se tomarán de Iris. A continuación haciendo `summary(Iris)` en la consola se obtiene la salida de la figura 14:

Figura 14.

```
Call:
lm(formula = AnchoSépalo ~ LargoSépalo, data = Iris)
Residuals:
 Min      1Q     Median      3Q     Max
-1.1095 -0.2454 -0.0167 0.2763 1.3338
Coefficients:
 Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.41895 0.25356 13.48 <2e-16 ***
LargoSépalo -0.06188 0.04297 -1.44 0.152
--
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.4343 on 148 degrees of freedom
Multiple R-squared: 0.01382, Adjusted R-squared: 0.007159
F-statistic: 2.074 on 1 and 148 DF, p-value: 0.1519
```

`Call` significa que a la llamada se ejecutó el comando lm. `Residuals` da cuenta del máximo, mínimo, cuartil 1, mediana y cuartil 3 de los residuos. `Coefficients` aporta la ordenada al origen 3.41895 y la pendiente -0.06188. El valor p 0.152 indica que debe rechazarse la hipótesis nula de que la pendiente de la recta es 0 y por lo tanto se concluye que existe relación lineal. La estimación del coeficiente de determinación es `R-squared`: 0.01382 que corresponde a un coeficiente de correlación lineal de -0.1176 pues la correlación es inversa. La recta de regresión entre ambas variables es $y = 0{,}06168x + 3{,}41895$ dónde y es AnchoSépalo y x LargoSépalo. Podemos graficar el diagrama de dispersión y la recta escribiendo las instrucciones: `>plot(Iris$LargoSépalo, Iris$AnchoSépalo, xlab = "LargoSépalo", ylab ="AnchoSépalo" >abline(regresion)` La figura 15 muestra la gráfica de salida:



El nombre Iris seguido del símbolo $ y el nombre de la variable respectiva llama a la variable correspondiente de Iris. El atributo `xlab="Largo Sépalo"` indica que el nombre de la variable explicativa x en la gráfica debe ser ése.

10.7. Ayudas y consultas

Hemos mostrado un rápido panorama de las aplicaciones de dos paquetes de software distintos referidas a los temas que abordamos a lo largo del libro. Con todo solo hemos visto apenas una pequeña parte del total de posibilidades de ambas herramientas. Para aquel lector interesado, la práctica será el camino más directo para profundizar y lograr destreza en el uso de software, conforme vaya avanzando también en la comprensión de aspectos teóricos que hacen al trabajo profesional en estadística. Para ello, en el caso de INFOSTAT cuenta con un excelente Manual al cual se accede al seleccionar `Ayuda-Manual` e inclusive si ejecuta `Ayuda-Como instalar R` hallará explicaciones on-line sobre como vincular ambos paquetes. En el caso de R existe además la posibilidad de programar secuencias de procedimientos y algoritmos en general. Pueden consultarse los manuales on-line accediendo desde la página principal del software y también, para conocer la sintaxis y

los alcances de algún comando en particular, ejecutar en la consola la instrucción `help(comando)` que habilita la ayuda on-line. En la bibliografía se proporcionan las direcciones principales de consulta para las dos herramientas estadísticas.

10.8. Ejercicios

Ejercicio N°1: Utilizar R para generar 50 observaciones de una distribución uniforme entre 10 y 40. Graficar el histograma de frecuencias y hallar media, desvío estándar, mediana, máximo y mínimo de la muestra obtenida.

Ejercicio N°2: Evaluar la bondad del ajuste normal para la variable Largo Pétalo del conjunto IRIS. Luego testear la hipótesis $\mu \geq 4$ acerca de la media poblacional.

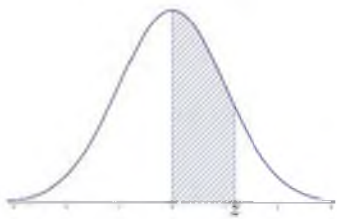
Ejercicio N°3: Hallar la recta de regresión poblacional para las variables Largo Pétalo y Largo Sépalo. Realizar el diagrama de dispersión y graficar la recta. Calcular el coeficiente de correlación lineal.

Ejercicio N°4: Evaluar la igualdad de medias de Largo Pétalo para cada una de las especies consideradas.

Anexo A

Distibución Normal estándar

Áreas bajo la curva entre o y z.



Z	**0.00**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990

Anexo B

Distribución t-Student

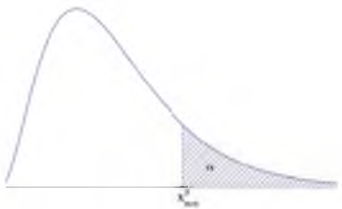
La tabla da áreas $1-\alpha$ y valores $c = t_{1-\alpha,r}$ donde $P[T \leq c] = 1-\alpha$, y T tiene distribución *t-Student* con r grados de libertad.

r	**0.75**	**0.80**	**0.85**	**0.90**	**0.95**	**0.975**	**0.99**	**0.995**
1	1.000	1.376	1.963	3.078	6.314	12.706	31.821	63.657
2	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925
3	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841
4	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604
5	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032
6	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707
7	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499
8	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355
9	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250
10	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169
11	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106
12	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055
13	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012
14	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977
15	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947
16	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921
17	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898
18	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878
19	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861
20	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845
21	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831
22	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819
23	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807
24	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797
25	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787
26	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779
27	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771
28	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763
29	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756
30	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750
40	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704
60	0.679	0.848	1.046	1.296	1.671	2.000	2.390	2.660
120	0.677	0.845	1.041	1.289	1.658	1.980	2.358	2.617
∞	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576

Anexo C

Distribución ji-cuadrado

- α = Área de cola.
- ν = Número de grados de libertad.



La tabla corresponde a los valores críticos χ^2 para un área α.

Para un área de cola: 0.995-0.500

ji-cuadrado	**Área de la cola, α**							
α/ν	**0.995**	**0.990**	**0.975**	**0.950**	**0.900**	**0.800**	**0.700**	**0.500**
1	0.00	0.00	0.00	0.00	0.02	0.06	0.15	0.45
2	0.01	0.02	0.05	0.10	0.21	0.45	0.71	1.39
3	0.07	0.11	0.22	0.35	0.58	1.01	1.42	2.37
4	0.21	0.30	0.48	0.71	1.06	1.65	2.19	3.36
5	0.41	0.55	0.83	1.15	1.61	2.34	3.00	4.35
6	0.68	0.87	1.24	1.64	2.20	3.07	3.83	5.35
7	0.99	1.24	1.69	2.17	2.83	3.82	4.67	6.35
8	1.34	1.65	2.18	2.73	3.49	4.59	5.53	7.34
9	1.73	2.09	2.70	3.33	4.17	5.38	6.39	8.34
10	2.16	2.56	3.25	3.94	4.87	6.18	7.27	9.34
11	2.60	3.05	3.82	4.57	5.58	6.99	8.15	10.34
12	3.07	3.57	4.40	5.23	6.30	7.81	9.03	11.34
13	3.57	4.11	5.01	5.89	7.04	8.63	9.93	12.34
14	4.07	4.66	5.63	6.57	7.79	9.47	10.82	13.34
15	4.60	5.23	6.26	7.26	8.55	10.31	11.72	14.34
16	5.14	5.81	6.91	7.96	9.31	11.15	12.62	15.34
17	5.70	6.41	7.56	8.67	10.09	12.00	13.53	16.34
18	6.26	7.01	8.23	9.39	10.86	12.86	14.44	17.34
19	6.84	7.63	8.91	10.12	11.65	13.72	15.35	18.34
20	7.43	8.26	9.59	10.85	12.44	14.58	16.27	19.34
21	8.03	8.90	10.28	11.59	13.24	15.44	17.18	20.34
22	8.64	9.54	10.98	12.34	14.04	16.31	18.10	21.34
23	9.26	10.20	11.69	13.09	14.85	17.19	19.02	22.34
24	9.89	10.86	12.40	13.85	15.66	18.06	19.94	23.34
25	610.52	11.52	13.12	14.61	16.47	18.94	20.87	24.34
26	11.16	12.20	13.84	15.38	17.29	19.82	21.79	25.34
27	11.81	12.88	14.57	16.15	18.11	20.70	22.72	26.34
28	12.46	13.56	15.31	16.93	18.94	21.59	23.65	27.34
29	13.12	14.26	16.05	17.71	19.77	22.48	24.48	28.34
30	13.79	14.95	16.79	18.49	20.60	23.36	25.51	29.34
40	20.71	22.16	24.43	26.51	29.05	32.34	34.87	39.34
50	27.99	29.71	32.36	34.76	37.69	41.45	44.31	49.33
60	35.53	37.48	40.48	43.19	46.46	50.64	53.81	59.33
70	43.28	45.44	48.76	51.74	55.33	59.90	63.35	69.33
80	51.17	53.54	57.15	60.39	64.28	69.21	72.92	79.33
90	59.20	61.75	65.65	69.13	73.29	78.56	82.51	89.33
100	67.33	70.06	74.22	77.93	82.36	87.95	92.13	99.33

Para un área de cola: 0.300-0.001

ji-cuadrado	**Área de la cola, α**							
α/ν	**0.300**	**0.200**	**0.100**	**0.050**	**0.025**	**0.010**	**0.005**	**0.001**
1	1.07	1.64	2.71	3.84	5.02	6.63	7.88	10.83
2	2.41	3.22	4.61	5.99	7.38	9.21	10.60	13.82
3	3.66	4.64	6.25	7.81	9.35	11.34	12.84	16.27
4	4.88	5.99	7.78	9.49	11.14	13.28	14.86	18.47
5	6.06	7.29	9.24	11.07	12.83	15.09	16.75	20.51
6	7.23	8.56	10.64	12.59	14.45	16.81	18.55	22.46
7	8.38	9.80	12.02	14.07	16.01	18.48	20.28	24.32
8	9.52	11.03	13.36	15.51	17.53	20.09	21.95	26.12
9	10.66	12.24	14.68	16.92	19.02	21.67	23.59	27.88
10	11.78	13.44	15.99	18.31	20.48	23.21	25.19	29.59
11	12.90	14.63	17.28	19.68	21.92	24.73	26.76	31.26
12	14.01	15.81	18.55	21.03	23.34	26.22	28.30	32.91
13	15.12	16.98	19.81	22.36	24.74	27.69	29.82	34.53
14	16.22	18.15	21.06	23.68	26.12	29.14	31.32	36.12
15	17.32	19.31	22.31	25.00	27.49	30.58	32.80	37.70
16	18.42	20.47	23.54	26.30	28.85	32.00	34.27	39.25
17	19.51	21.61	24.77	27.59	30.19	33.41	35.72	40.79
18	20.60	22.76	25.99	28.87	31.53	34.81	37.16	42.31
19	21.69	23.90	27.20	30.14	32.85	36.19	38.58	43.82
20	22.77	25.04	28.41	31.41	34.17	37.57	40.00	45.31
21	23.86	26.17	29.62	32.67	35.48	38.93	41.40	46.80
22	24.94	27.30	30.81	33.92	36.78	40.29	42.80	48.27
23	26.02	28.43	32.01	35.17	38.08	41.64	44.18	49.73
24	27.10	29.55	33.20	36.42	39.36	42.98	45.56	51.18
25	28.17	30.68	34.38	37.65	40.65	44.31	46.93	52.62
26	29.25	31.79	35.56	38.89	41.92	45.64	48.29	54.05
27	30.32	32.91	36.74	40.11	43.19	46.96	49.65	55.48
28	31.39	34.03	37.92	41.34	44.46	48.28	50.99	56.89
29	32.46	35.14	39.09	42.56	45.72	49.59	52.34	58.30
30	33.53	36.25	40.26	43.77	46.98	50.89	53.67	59.70
40	44.16	47.27	51.81	55.76	59.34	63.69	66.77	73.40
50	54.72	58.16	63.17	67.50	71.42	76.15	79.49	86.66
60	65.23	68.97	74.40	79.08	83.30	88.38	91.95	99.61
70	75.69	79.71	85.53	90.53	95.02	100.43	104.21	112.32
80	86.12	90.41	96.58	101.88	106.63	112.33	116.32	124.84
90	96.52	101.05	107.57	113.15	118.14	124.12	128.30	137.21
100	106.91	111.67	118.50	124.34	129.56	135.81	140.17	149.45

Anexo D

Valores F de la distribución F de Fisher

$1-\alpha = 0{,}9$ $\quad\quad$ ν_1 = grados de libertad del numerador

$1-\alpha = P(F \leq f_{\alpha,\nu_1,\nu_2})$ $\quad\quad$ ν_2 = grados de libertad del denominador

	1	2	3	4	5	6	7	8	9	10	11	12	13
1	39.864	49.500	53.593	55.833	57.240	58.204	58.906	59.439	59.857	60.195	60.473	60.705	60.902
2	8.526	9.000	9.162	9.243	9.293	9.326	9.349	9.367	9.381	9.392	9.401	9.408	9.415
3	5.538	5.462	5.391	5.343	5.309	5.285	5.266	5.252	5.240	5.230	5.222	5.216	5.210
4	4.545	4.325	4.191	4.107	4.051	4.010	3.979	3.955	3.936	3.920	3.907	3.896	3.886
5	4.060	3.780	3.619	3.520	3.453	3.405	3.368	3.339	3.316	3.297	3.282	3.268	3.257
6	3.776	3.463	3.289	3.181	3.108	3.055	3.014	2.983	2.958	2.937	2.920	2.905	2.892
7	3.589	3.257	3.074	2.961	2.883	2.827	2.785	2.752	2.725	2.703	2.684	2.668	2.654
8	3.458	3.113	2.924	2.806	2.726	2.668	2.624	2.589	2.561	2.538	2.519	2.502	2.488
9	3.360	3.006	2.813	2.693	2.611	2.551	2.505	2.469	2.440	2.416	2.396	2.379	2.364
10	3.285	2.924	2.728	2.605	2.522	2.461	2.414	2.377	2.347	2.323	2.302	2.284	2.269
11	3.225	2.860	2.660	2.536	2.451	2.389	2.342	2.304	2.274	2.248	2.227	2.209	2.193
12	3.177	2.807	2.606	2.480	2.394	2.331	2.283	2.245	2.214	2.188	2.166	2.147	2.131
13	3.136	2.763	2.560	2.434	2.347	2.283	2.234	2.195	2.164	2.138	2.116	2.097	2.080
14	3.102	2.726	2.522	2.395	2.307	2.243	2.193	2.154	2.122	2.095	2.073	2.054	2.037
15	3.073	2.695	2.490	2.361	2.273	2.208	2.158	2.119	2.086	2.059	2.037	2.017	2.000
16	3.048	2.668	2.462	2.333	2.244	2.178	2.128	2.088	2.055	2.028	2.005	1.985	1.968
17	3.026	2.645	2.437	2.308	2.218	2.152	2.102	2.061	2.028	2.001	1.978	1.958	1.940
18	3.007	2.624	2.416	2.286	2.196	2.130	2.079	2.038	2.005	1.977	1.954	1.933	1.916
19	2.990	2.606	2.397	2.266	2.176	2.109	2.058	2.017	1.984	1.956	1.932	1.912	1.894
20	2.975	2.589	2.380	2.249	2.158	2.091	2.040	1.999	1.965	1.937	1.913	1.892	1.875
21	2.961	2.575	2.365	2.233	2.142	2.075	2.023	1.982	1.948	1.920	1.896	1.875	1.857
22	2.949	2.561	2.351	2.219	2.128	2.060	2.008	1.967	1.933	1.904	1.880	1.859	1.841
23	2.937	2.549	2.339	2.207	2.115	2.047	1.995	1.953	1.919	1.890	1.866	1.845	1.827
24	2.927	2.538	2.327	2.195	2.103	2.035	1.983	1.941	1.906	1.877	1.853	1.832	1.814
25	2.918	2.528	2.317	2.184	2.092	2.024	1.971	1.929	1.895	1.866	1.841	1.820	1.802
26	2.909	2.519	2.307	2.174	2.082	2.014	1.961	1.919	1.884	1.855	1.830	1.809	1.790
27	2.901	2.511	2.299	2.165	2.073	2.005	1.952	1.909	1.874	1.845	1.820	1.799	1.780
28	2.894	2.503	2.291	2.157	2.064	1.996	1.943	1.900	1.865	1.836	1.811	1.790	1.771
29	2.887	2.495	2.283	2.149	2.057	1.988	1.935	1.892	1.857	1.827	1.802	1.781	1.762
30	2.881	2.489	2.276	2.142	2.049	1.980	1.927	1.884	1.849	1.819	1.794	1.773	1.754
40	2.835	2.440	2.226	2.091	1.997	1.927	1.873	1.829	1.793	1.763	1.737	1.715	1.695
50	2.809	2.412	2.197	2.061	1.966	1.895	1.840	1.796	1.760	1.729	1.703	1.680	1.660
60	2.791	2.393	2.177	2.041	1.946	1.875	1.819	1.775	1.738	1.707	1.680	1.657	1.637
70	2.779	2.380	2.164	2.027	1.931	1.860	1.804	1.760	1.723	1.691	1.665	1.641	1.621
80	2.769	2.370	2.154	2.016	1.921	1.849	1.793	1.748	1.711	1.680	1.653	1.629	1.609
90	2.762	2.363	2.146	2.008	1.912	1.841	1.785	1.739	1.702	1.670	1.643	1.620	1.599
100	2.756	2.356	2.139	2.002	1.906	1.834	1.778	1.732	1.695	1.663	1.636	1.612	1.592
200	2.731	2.329	2.111	1.973	1.876	1.804	1.747	1.701	1.663	1.631	1.603	1.579	1.558
500	2.716	2.313	2.095	1.956	1.859	1.786	1.729	1.683	1.644	1.612	1.583	1.559	1.537
1000	2.711	2.308	2.089	1.950	1.853	1.780	1.723	1.676	1.638	1.605	1.577	1.552	1.531

$1 - \alpha = 0{,}9$

	14	15	16	17	18	19	20	21	22	23	24	25	26
1	61.073	61.220	61.350	61.465	61.566	61.658	61.740	61.815	61.883	61.945	62.002	62.055	62.103
2	9.420	9.425	9.429	9.433	9.436	9.439	9.441	9.444	9.446	9.448	9.450	9.451	9.453
3	5.205	5.200	5.196	5.193	5.190	5.187	5.184	5.182	5.180	5.178	5.176	5.175	5.173
4	3.878	3.870	3.864	3.858	3.853	3.848	3.844	3.841	3.837	3.834	3.831	3.828	3.826
5	3.247	3.238	3.230	3.223	3.217	3.212	3.207	3.202	3.198	3.194	3.191	3.187	3.184
6	2.881	2.871	2.863	2.855	2.848	2.842	2.836	2.831	2.827	2.822	2.818	2.815	2.811
7	2.643	2.632	2.623	2.615	2.607	2.601	2.595	2.589	2.584	2.580	2.575	2.571	2.568
8	2.475	2.464	2.454	2.446	2.438	2.431	2.425	2.419	2.414	2.409	2.404	2.400	2.396
9	2.351	2.340	2.330	2.320	2.312	2.305	2.298	2.292	2.287	2.282	2.277	2.272	2.268
10	2.255	2.244	2.233	2.224	2.215	2.208	2.201	2.194	2.189	2.183	2.178	2.174	2.170
11	2.179	2.167	2.156	2.147	2.138	2.130	2.123	2.117	2.111	2.105	2.100	2.095	2.091
12	2.117	2.105	2.094	2.084	2.075	2.067	2.060	2.053	2.047	2.041	2.036	2.031	2.027
13	2.066	2.053	2.042	2.032	2.023	2.014	2.007	2.000	1.994	1.988	1.983	1.978	1.973
14	2.022	2.010	1.998	1.988	1.978	1.970	1.962	1.955	1.949	1.943	1.938	1.933	1.928
15	1.985	1.972	1.961	1.950	1.941	1.932	1.924	1.917	1.911	1.905	1.899	1.894	1.889
16	1.953	1.940	1.928	1.917	1.908	1.899	1.891	1.884	1.877	1.871	1.866	1.860	1.855
17	1.925	1.912	1.900	1.889	1.879	1.870	1.862	1.855	1.848	1.842	1.836	1.831	1.826
18	1.900	1.887	1.875	1.864	1.854	1.845	1.837	1.829	1.823	1.816	1.810	1.805	1.800
19	1.878	1.865	1.852	1.841	1.831	1.822	1.814	1.807	1.800	1.793	1.787	1.782	1.777
20	1.859	1.845	1.833	1.821	1.811	1.802	1.794	1.786	1.779	1.773	1.767	1.761	1.756
21	1.841	1.827	1.815	1.803	1.793	1.784	1.776	1.768	1.761	1.754	1.748	1.742	1.737
22	1.825	1.811	1.798	1.787	1.777	1.768	1.759	1.751	1.744	1.737	1.731	1.726	1.720
23	1.811	1.796	1.784	1.772	1.762	1.753	1.744	1.736	1.729	1.722	1.716	1.710	1.705
24	1.797	1.783	1.770	1.759	1.748	1.739	1.730	1.722	1.715	1.708	1.702	1.696	1.691
25	1.785	1.771	1.758	1.746	1.736	1.726	1.718	1.710	1.702	1.695	1.689	1.683	1.678
26	1.774	1.760	1.747	1.735	1.724	1.715	1.706	1.698	1.690	1.683	1.677	1.671	1.666
27	1.764	1.749	1.736	1.724	1.714	1.704	1.695	1.687	1.680	1.673	1.666	1.660	1.655
28	1.754	1.740	1.726	1.715	1.704	1.694	1.685	1.677	1.669	1.662	1.656	1.650	1.644
29	1.745	1.731	1.717	1.705	1.695	1.685	1.676	1.668	1.660	1.653	1.647	1.640	1.635
30	1.737	1.722	1.709	1.697	1.686	1.676	1.667	1.659	1.651	1.644	1.638	1.632	1.626
40	1.678	1.662	1.649	1.636	1.625	1.615	1.605	1.596	1.588	1.581	1.574	1.568	1.562
50	1.643	1.627	1.613	1.600	1.588	1.578	1.568	1.559	1.551	1.543	1.536	1.529	1.523
60	1.619	1.603	1.589	1.576	1.564	1.553	1.543	1.534	1.526	1.518	1.511	1.504	1.498
70	1.603	1.587	1.572	1.559	1.547	1.536	1.526	1.517	1.508	1.500	1.493	1.486	1.479
80	1.590	1.574	1.559	1.546	1.534	1.523	1.513	1.503	1.495	1.487	1.479	1.472	1.465
90	1.581	1.564	1.550	1.536	1.524	1.513	1.503	1.493	1.484	1.476	1.468	1.461	1.455
100	1.573	1.557	1.542	1.528	1.516	1.505	1.494	1.485	1.476	1.468	1.460	1.453	1.446
200	1.539	1.522	1.507	1.493	1.480	1.468	1.458	1.448	1.438	1.430	1.422	1.414	1.407
500	1.518	1.501	1.485	1.471	1.458	1.446	1.435	1.425	1.416	1.407	1.399	1.391	1.384
1000	1.511	1.494	1.478	1.464	1.451	1.439	1.428	1.418	1.408	1.399	1.391	1.383	1.376

$1 - \alpha = 0{,}9$

	27	**28**	**29**	**30**	**40**	**50**	**60**	**70**	**80**	**90**	**100**	**200**	**500**
1	62.148	62.189	62.229	62.265	62.529	62.688	62.794	62.871	62.927	62.972	63.007	63.167	63.264
2	9.454	9.456	9.457	9.458	9.466	9.471	9.475	9.477	9.479	9.480	9.481	9.486	9.489
3	5.172	5.170	5.169	5.168	5.160	5.155	5.151	5.149	5.147	5.145	5.144	5.139	5.136
4	3.823	3.821	3.819	3.817	3.804	3.795	3.790	3.786	3.782	3.780	3.778	3.769	3.764
5	3.181	3.179	3.176	3.174	3.157	3.147	3.140	3.135	3.132	3.129	3.126	3.116	3.109
6	2.808	2.805	2.803	2.800	2.781	2.770	2.762	2.756	2.752	2.749	2.746	2.734	2.727
7	2.564	2.561	2.558	2.555	2.535	2.523	2.514	2.508	2.504	2.500	2.497	2.484	2.476
8	2.392	2.389	2.386	2.383	2.361	2.348	2.339	2.333	2.328	2.324	2.321	2.307	2.298
9	2.265	2.261	2.258	2.255	2.232	2.218	2.208	2.202	2.196	2.192	2.189	2.174	2.165
10	2.166	2.162	2.159	2.155	2.132	2.117	2.107	2.100	2.095	2.090	2.087	2.071	2.062
11	2.087	2.083	2.080	2.076	2.052	2.036	2.026	2.019	2.013	2.009	2.005	1.989	1.979
12	2.022	2.019	2.015	2.011	1.986	1.970	1.960	1.952	1.946	1.942	1.938	1.921	1.911
13	1.969	1.965	1.961	1.958	1.931	1.915	1.904	1.896	1.890	1.886	1.882	1.864	1.853
14	1.923	1.919	1.916	1.912	1.885	1.869	1.857	1.849	1.843	1.838	1.834	1.816	1.805
15	1.885	1.880	1.876	1.873	1.845	1.828	1.817	1.808	1.802	1.797	1.793	1.774	1.763
16	1.851	1.847	1.843	1.839	1.811	1.793	1.782	1.773	1.766	1.761	1.757	1.738	1.726
17	1.821	1.817	1.813	1.809	1.781	1.763	1.751	1.742	1.735	1.730	1.726	1.706	1.694
18	1.795	1.791	1.787	1.783	1.754	1.736	1.723	1.714	1.707	1.702	1.698	1.678	1.665
19	1.772	1.767	1.763	1.759	1.730	1.711	1.699	1.690	1.683	1.677	1.673	1.652	1.639
20	1.751	1.746	1.742	1.738	1.708	1.690	1.677	1.667	1.660	1.655	1.650	1.629	1.616
21	1.732	1.728	1.723	1.719	1.689	1.670	1.657	1.647	1.640	1.634	1.630	1.608	1.595
22	1.715	1.711	1.706	1.702	1.671	1.652	1.639	1.629	1.622	1.616	1.611	1.590	1.576
23	1.700	1.695	1.691	1.686	1.655	1.636	1.622	1.613	1.605	1.599	1.594	1.572	1.558
24	1.686	1.681	1.676	1.672	1.641	1.621	1.607	1.597	1.590	1.584	1.579	1.556	1.542
25	1.672	1.668	1.663	1.659	1.627	1.607	1.593	1.583	1.576	1.569	1.565	1.542	1.527
26	1.660	1.656	1.651	1.647	1.615	1.594	1.581	1.570	1.562	1.556	1.551	1.528	1.514
27	1.649	1.645	1.640	1.636	1.603	1.583	1.569	1.558	1.550	1.544	1.539	1.515	1.501
28	1.639	1.634	1.630	1.625	1.592	1.572	1.558	1.547	1.539	1.533	1.528	1.504	1.489
29	1.630	1.625	1.620	1.616	1.583	1.562	1.547	1.537	1.529	1.522	1.517	1.493	1.478
30	1.621	1.616	1.611	1.606	1.573	1.552	1.538	1.527	1.519	1.512	1.507	1.482	1.467
40	1.556	1.551	1.546	1.541	1.506	1.483	1.467	1.455	1.447	1.439	1.434	1.406	1.389
50	1.517	1.512	1.507	1.502	1.465	1.441	1.424	1.412	1.402	1.395	1.388	1.359	1.340
60	1.492	1.486	1.481	1.476	1.437	1.413	1.395	1.382	1.372	1.364	1.358	1.326	1.306
70	1.473	1.467	1.462	1.457	1.418	1.392	1.374	1.361	1.350	1.342	1.335	1.302	1.281
80	1.459	1.453	1.448	1.443	1.403	1.377	1.358	1.344	1.334	1.325	1.318	1.284	1.261
90	1.448	1.442	1.437	1.432	1.391	1.365	1.346	1.332	1.321	1.312	1.304	1.269	1.245
100	1.440	1.434	1.428	1.423	1.382	1.355	1.336	1.321	1.310	1.301	1.293	1.257	1.232
200	1.400	1.394	1.388	1.383	1.339	1.310	1.289	1.273	1.261	1.250	1.242	1.199	1.168
500	1.377	1.370	1.364	1.358	1.313	1.282	1.260	1.243	1.229	1.218	1.209	1.160	1.122
1000	1.369	1.362	1.356	1.350	1.304	1.273	1.250	1.232	1.218	1.207	1.197	1.145	1.103

Valores F de la distribución F de Fisher

$1-\alpha = 0{,}95$ $\qquad$ ν_1 = grados de libertad del numerador

$1-\alpha = P(F \leq f_{\alpha,\nu_1,\nu_2})$ $\qquad$ ν_2 = grados de libertad del denominador

	1	2	3	4	5	6	7	8	9	10	11	12	13
1	161.446	199.499	215.707	224.583	230.160	233.988	236.767	238.884	240.543	241.882	242.981	243.905	244.690
2	18.513	19.000	19.164	19.247	19.296	19.329	19.353	19.371	19.385	19.396	19.405	19.412	19.419
3	10.128	9.552	9.277	9.117	9.013	8.941	8.887	8.845	8.812	8.785	8.763	8.745	8.729
4	7.709	6.944	6.591	6.388	6.256	6.163	6.094	6.041	5.999	5.964	5.936	5.912	5.891
5	6.608	5.786	5.409	5.192	5.050	4.950	4.876	4.818	4.772	4.735	4.704	4.678	4.655
6	5.987	5.143	4.757	4.534	4.387	4.284	4.207	4.147	4.099	4.060	4.027	4.000	3.976
7	5.591	4.737	4.347	4.120	3.972	3.866	3.787	3.726	3.677	3.637	3.603	3.575	3.550
8	5.318	4.459	4.066	3.838	3.688	3.581	3.500	3.438	3.388	3.347	3.313	3.284	3.259
9	5.117	4.256	3.863	3.633	3.482	3.374	3.293	3.230	3.179	3.137	3.102	3.073	3.048
10	4.965	4.103	3.708	3.478	3.326	3.217	3.135	3.072	3.020	2.978	2.943	2.913	2.887
11	4.844	3.982	3.587	3.357	3.204	3.095	3.012	2.948	2.896	2.854	2.818	2.788	2.761
12	4.747	3.885	3.490	3.259	3.106	2.996	2.913	2.849	2.796	2.753	2.717	2.687	2.660
13	4.667	3.806	3.411	3.179	3.025	2.915	2.832	2.767	2.714	2.671	2.635	2.604	2.577
14	4.600	3.739	3.344	3.112	2.958	2.848	2.764	2.699	2.646	2.602	2.565	2.534	2.507
15	4.543	3.682	3.287	3.056	2.901	2.790	2.707	2.641	2.588	2.544	2.507	2.475	2.448
16	4.494	3.634	3.239	3.007	2.852	2.741	2.657	2.591	2.538	2.494	2.456	2.425	2.397
17	4.451	3.592	3.197	2.965	2.810	2.699	2.614	2.548	2.494	2.450	2.413	2.381	2.353
18	4.414	3.555	3.160	2.928	2.773	2.661	2.577	2.510	2.456	2.412	2.374	2.342	2.314
19	4.381	3.522	3.127	2.895	2.740	2.628	2.544	2.477	2.423	2.378	2.340	2.308	2.280
20	4.351	3.493	3.098	2.866	2.711	2.599	2.514	2.447	2.393	2.348	2.310	2.278	2.250
21	4.325	3.467	3.072	2.840	2.685	2.573	2.488	2.420	2.366	2.321	2.283	2.250	2.222
22	4.301	3.443	3.049	2.817	2.661	2.549	2.464	2.397	2.342	2.297	2.259	2.226	2.198
23	4.279	3.422	3.028	2.796	2.640	2.528	2.442	2.375	2.320	2.275	2.236	2.204	2.175
24	4.260	3.403	3.009	2.776	2.621	2.508	2.423	2.355	2.300	2.255	2.216	2.183	2.155
25	4.242	3.385	2.991	2.759	2.603	2.490	2.405	2.337	2.282	2.236	2.198	2.165	2.136
26	4.225	3.369	2.975	2.743	2.587	2.474	2.388	2.321	2.265	2.220	2.181	2.148	2.119
27	4.210	3.354	2.960	2.728	2.572	2.459	2.373	2.305	2.250	2.204	2.166	2.132	2.103
28	4.196	3.340	2.947	2.714	2.558	2.445	2.359	2.291	2.236	2.190	2.151	2.118	2.089
29	4.183	3.328	2.934	2.701	2.545	2.432	2.346	2.278	2.223	2.177	2.138	2.104	2.075
30	4.171	3.316	2.922	2.690	2.534	2.421	2.334	2.266	2.211	2.165	2.126	2.092	2.063
40	4.085	3.232	2.839	2.606	2.449	2.336	2.249	2.180	2.124	2.077	2.038	2.003	1.974
50	4.034	3.183	2.790	2.557	2.400	2.286	2.199	2.130	2.073	2.026	1.986	1.952	1.921
60	4.001	3.150	2.758	2.525	2.368	2.254	2.167	2.097	2.040	1.993	1.952	1.917	1.887
70	3.978	3.128	2.736	2.503	2.346	2.231	2.143	2.074	2.017	1.969	1.928	1.893	1.863
80	3.960	3.111	2.719	2.486	2.329	2.214	2.126	2.056	1.999	1.951	1.910	1.875	1.845
90	3.947	3.098	2.706	2.473	2.316	2.201	2.113	2.043	1.986	1.938	1.897	1.861	1.830
100	3.936	3.087	2.696	2.463	2.305	2.191	2.103	2.032	1.975	1.927	1.886	1.850	1.819
200	3.888	3.041	2.650	2.417	2.259	2.144	2.056	1.985	1.927	1.878	1.837	1.801	1.769
500	3.860	3.014	2.623	2.390	2.232	2.117	2.028	1.957	1.899	1.850	1.808	1.772	1.740
1000	3.851	3.005	2.614	2.381	2.223	2.108	2.019	1.948	1.889	1.840	1.798	1.762	1.730

$1 - \alpha = 0{,}95$

	14	15	16	17	18	19	20	21	22	23	24	25	26
1	245.363	245.949	246.466	246.917	247.324	247.688	248.016	248.307	248.579	248.823	249.052	249.260	249.453
2	19.424	19.429	19.433	19.437	19.440	19.443	19.446	19.448	19.450	19.452	19.454	19.456	19.457
3	8.715	8.703	8.692	8.683	8.675	8.667	8.660	8.654	8.648	8.643	8.638	8.634	8.630
4	5.873	5.858	5.844	5.832	5.821	5.811	5.803	5.795	5.787	5.781	5.774	5.769	5.763
5	4.636	4.619	4.604	4.590	4.579	4.568	4.558	4.549	4.541	4.534	4.527	4.521	4.515
6	3.956	3.938	3.922	3.908	3.896	3.884	3.874	3.865	3.856	3.849	3.841	3.835	3.829
7	3.529	3.511	3.494	3.480	3.467	3.455	3.445	3.435	3.426	3.418	3.410	3.404	3.397
8	3.237	3.218	3.202	3.187	3.173	3.161	3.150	3.140	3.131	3.123	3.115	3.108	3.102
9	3.025	3.006	2.989	2.974	2.960	2.948	2.936	2.926	2.917	2.908	2.900	2.893	2.886
10	2.865	2.845	2.828	2.812	2.798	2.785	2.774	2.764	2.754	2.745	2.737	2.730	2.723
11	2.739	2.719	2.701	2.685	2.671	2.658	2.646	2.636	2.626	2.617	2.609	2.601	2.594
12	2.637	2.617	2.599	2.583	2.568	2.555	2.544	2.533	2.523	2.514	2.505	2.498	2.491
13	2.554	2.533	2.515	2.499	2.484	2.471	2.459	2.448	2.438	2.429	2.420	2.412	2.405
14	2.484	2.463	2.445	2.428	2.413	2.400	2.388	2.377	2.367	2.357	2.349	2.341	2.333
15	2.424	2.403	2.385	2.368	2.353	2.340	2.328	2.316	2.306	2.297	2.288	2.280	2.272
16	2.373	2.352	2.333	2.317	2.302	2.288	2.276	2.264	2.254	2.244	2.235	2.227	2.220
17	2.329	2.308	2.289	2.272	2.257	2.243	2.230	2.219	2.208	2.199	2.190	2.181	2.174
18	2.290	2.269	2.250	2.233	2.217	2.203	2.191	2.179	2.168	2.159	2.150	2.141	2.134
19	2.256	2.234	2.215	2.198	2.182	2.168	2.155	2.144	2.133	2.123	2.114	2.106	2.098
20	2.225	2.203	2.184	2.167	2.151	2.137	2.124	2.112	2.102	2.092	2.082	2.074	2.066
21	2.197	2.176	2.156	2.139	2.123	2.109	2.096	2.084	2.073	2.063	2.054	2.045	2.037
22	2.173	2.151	2.131	2.114	2.098	2.084	2.071	2.059	2.048	2.038	2.028	2.020	2.012
23	2.150	2.128	2.109	2.091	2.075	2.061	2.048	2.036	2.025	2.014	2.005	1.996	1.988
24	2.130	2.108	2.088	2.070	2.054	2.040	2.027	2.015	2.003	1.993	1.984	1.975	1.967
25	2.111	2.089	2.069	2.051	2.035	2.021	2.007	1.995	1.984	1.974	1.964	1.955	1.947
26	2.094	2.072	2.052	2.034	2.018	2.003	1.990	1.978	1.966	1.956	1.946	1.938	1.929
27	2.078	2.056	2.036	2.018	2.002	1.987	1.974	1.961	1.950	1.940	1.930	1.921	1.913
28	2.064	2.041	2.021	2.003	1.987	1.972	1.959	1.946	1.935	1.924	1.915	1.906	1.897
29	2.050	2.027	2.007	1.989	1.973	1.958	1.945	1.932	1.921	1.910	1.901	1.891	1.883
30	2.037	2.015	1.995	1.976	1.960	1.945	1.932	1.919	1.908	1.897	1.887	1.878	1.870
40	1.948	1.924	1.904	1.885	1.868	1.853	1.839	1.826	1.814	1.803	1.793	1.783	1.775
50	1.895	1.871	1.850	1.831	1.814	1.798	1.784	1.771	1.759	1.748	1.737	1.727	1.718
60	1.860	1.836	1.815	1.796	1.778	1.763	1.748	1.735	1.722	1.711	1.700	1.690	1.681
70	1.836	1.812	1.790	1.771	1.753	1.737	1.722	1.709	1.696	1.685	1.674	1.664	1.654
80	1.817	1.793	1.772	1.752	1.734	1.718	1.703	1.689	1.677	1.665	1.654	1.644	1.634
90	1.803	1.779	1.757	1.737	1.720	1.703	1.688	1.675	1.662	1.650	1.639	1.629	1.619
100	1.792	1.768	1.746	1.726	1.708	1.691	1.676	1.663	1.650	1.638	1.627	1.616	1.607
200	1.742	1.717	1.694	1.674	1.656	1.639	1.623	1.609	1.596	1.583	1.572	1.561	1.551
500	1.712	1.686	1.664	1.643	1.625	1.607	1.592	1.577	1.563	1.551	1.539	1.528	1.518
1000	1.702	1.676	1.654	1.633	1.614	1.597	1.581	1.566	1.553	1.540	1.528	1.517	1.507

$1 - \alpha = 0{,}95$

	27	28	29	30	40	50	60	70	80	90	100	200	500
1	249.631	249.798	249.951	250.096	251.144	251.774	252.196	252.498	252.723	252.898	253.043	253.676	254.062
2	19.459	19.460	19.461	19.463	19.471	19.476	19.479	19.481	19.483	19.485	19.486	19.491	19.494
3	8.626	8.623	8.620	8.617	8.594	8.581	8.572	8.566	8.561	8.557	8.554	8.540	8.532
4	5.759	5.754	5.750	5.746	5.717	5.699	5.688	5.679	5.673	5.668	5.664	5.646	5.635
5	4.510	4.505	4.500	4.496	4.464	4.444	4.431	4.422	4.415	4.409	4.405	4.385	4.373
6	3.823	3.818	3.813	3.808	3.774	3.754	3.740	3.730	3.722	3.716	3.712	3.690	3.678
7	3.391	3.386	3.381	3.376	3.340	3.319	3.304	3.294	3.286	3.280	3.275	3.252	3.239
8	3.095	3.090	3.084	3.079	3.043	3.020	3.005	2.994	2.986	2.980	2.975	2.951	2.937
9	2.880	2.874	2.869	2.864	2.826	2.803	2.787	2.776	2.768	2.761	2.756	2.731	2.717
10	2.716	2.710	2.705	2.700	2.661	2.637	2.621	2.609	2.601	2.594	2.588	2.563	2.548
11	2.5818	2.582	2.576	2.570	2.531	2.507	2.490	2.478	2.469	2.462	2.457	2.431	2.415
12	2.484	2.478	2.472	2.466	2.426	2.401	2.384	2.372	2.363	2.356	2.350	2.323	2.307
13	2.398	2.392	2.386	2.380	2.339	2.314	2.297	2.284	2.275	2.267	2.261	2.234	2.218
14	2.326	2.320	2.314	2.308	2.266	2.241	2.223	2.210	2.201	2.193	2.187	2.159	2.142
15	2.265	2.259	2.253	2.247	2.204	2.178	2.160	2.147	2.137	2.130	2.123	2.095	2.078
16	2.212	2.206	2.200	2.194	2.151	2.124	2.106	2.093	2.083	2.075	2.068	2.039	2.022
17	2.167	2.160	2.154	2.148	2.104	2.077	2.058	2.045	2.035	2.027	2.020	1.991	1.973
18	2.126	2.119	2.113	2.107	2.063	2.035	2.017	2.003	1.993	1.985	1.978	1.948	1.929
19	2.090	2.084	2.077	2.071	2.026	1.999	1.980	1.966	1.955	1.947	1.940	1.910	1.891
20	2.059	2.052	2.045	2.039	1.994	1.966	1.946	1.932	1.922	1.913	1.907	1.875	1.856
21	2.030	2.023	2.016	2.010	1.965	1.936	1.916	1.902	1.891	1.883	1.876	1.845	1.825
22	2.004	1.997	1.990	1.984	1.938	1.909	1.889	1.875	1.864	1.856	1.849	1.817	1.797
23	1.981	1.973	1.967	1.961	1.914	1.885	1.865	1.850	1.839	1.830	1.823	1.791	1.771
24	1.959	1.952	1.945	1.939	1.892	1.863	1.842	1.828	1.816	1.808	1.800	1.768	1.747
25	1.939	1.932	1.926	1.919	1.872	1.842	1.822	1.807	1.796	1.787	1.779	1.746	1.725
26	1.921	1.914	1.907	1.901	1.853	1.823	1.803	1.788	1.776	1.767	1.760	1.726	1.705
27	1.905	1.898	1.891	1.884	1.836	1.806	1.785	1.770	1.758	1.749	1.742	1.708	1.686
28	1.889	1.882	1.875	1.869	1.820	1.790	1.769	1.754	1.742	1.733	1.725	1.691	1.669
29	1.875	1.868	1.861	1.854	1.806	1.775	1.754	1.738	1.726	1.717	1.710	1.675	1.653
30	1.862	1.854	1.847	1.841	1.792	1.761	1.740	1.724	1.712	1.703	1.695	1.660	1.637
40	1.766	1.759	1.751	1.744	1.693	1.660	1.637	1.621	1.608	1.597	1.589	1.551	1.526
50	1.710	1.702	1.694	1.687	1.634	1.599	1.576	1.558	1.544	1.534	1.525	1.484	1.457
60	1.672	1.664	1.656	1.649	1.594	1.559	1.534	1.516	1.502	1.491	1.481	1.438	1.409
70	1.646	1.637	1.629	1.622	1.566	1.530	1.505	1.486	1.471	1.459	1.450	1.404	1.374
80	1.626	1.617	1.609	1.602	1.545	1.508	1.482	1.463	1.448	1.436	1.426	1.379	1.347
90	1.610	1.601	1.593	1.586	1.528	1.491	1.465	1.445	1.429	1.417	1.407	1.358	1.326
100	1.598	1.589	1.581	1.573	1.515	1.477	1.450	1.430	1.415	1.402	1.392	1.342	1.308
200	1.542	1.533	1.524	1.516	1.455	1.415	1.386	1.364	1.346	1.332	1.321	1.263	1.221
500	1.508	1.499	1.490	1.482	1.419	1.376	1.345	1.322	1.303	1.288	1.275	1.210	1.159
1000	1.497	1.488	1.479	1.471	1.406	1.363	1.332	1.308	1.289	1.273	1.260	1.190	1.134

Valores F de la distribución F de Fisher

$1-\alpha = 0{,}975$ $\qquad$ ν_1 = grados de libertad del numerador

$1-\alpha = P(F \leq f_{\alpha,\nu_1,\nu_2})$ $\qquad$ ν_2 = grados de libertad del denominador

	1	2	3	4	5	6	7	8	9	10	11	12	13
1	1 647.793	799.482	864.151	899.599	921.835	937.114	948.203	956.643	963.279	968.634	973.028	976.725	979.839
2	38.506	39.000	39.166	39.248	39.298	39.331	39.356	39.373	39.387	39.398	39.407	39.415	39.421
3	17.443	16.044	15.439	15.101	14.885	14.735	14.624	14.540	14.473	14.419	14.374	14.337	14.305
4	12.218	10.649	9.979	9.604	9.364	9.197	9.074	8.980	8.905	8.844	8.794	8.751	8.715
5	10.007	8.434	7.764	7.388	7.146	6.978	6.853	6.757	6.681	6.619	6.568	6.525	6.488
6	8.813	7.260	6.599	6.227	5.988	5.820	5.695	5.600	5.523	5.461	5.410	5.366	5.329
7	8.073	6.542	5.890	5.523	5.285	5.119	4.995	4.899	4.823	4.761	4.709	4.666	4.628
8	7.571	6.059	5.416	5.053	4.817	4.652	4.529	4.433	4.357	4.295	4.243	4.200	4.162
9	7.209	5.715	5.078	4.718	4.484	4.320	4.197	4.102	4.026	3.964	3.912	3.868	3.831
10	6.937	5.456	4.826	4.468	4.236	4.072	3.950	3.855	3.779	3.717	3.665	3.621	3.583
11	6.724	5.256	4.630	4.275	4.044	3.881	3.759	3.664	3.588	3.526	3.474	3.430	3.392
12	6.554	5.096	4.474	4.121	3.891	3.728	3.607	3.512	3.436	3.374	3.321	3.277	3.239
13	6.414	4.965	4.347	3.996	3.767	3.604	3.483	3.388	3.312	3.250	3.197	3.153	3.115
14	6.298	4.857	4.242	3.892	3.663	3.501	3.380	3.285	3.209	3.147	3.095	3.050	3.012
15	6.200	4.765	4.153	3.804	3.576	3.415	3.293	3.199	3.123	3.060	3.008	2.963	2.925
16	6.115	4.687	4.077	3.729	3.502	3.341	3.219	3.125	3.049	2.986	2.934	2.889	2.851
17	6.042	4.619	4.011	3.665	3.438	3.277	3.156	3.061	2.985	2.922	2.870	2.825	2.786
18	5.978	4.560	3.954	3.608	3.382	3.221	3.100	3.005	2.929	2.866	2.814	2.769	2.730
19	5.922	4.508	3.903	3.559	3.333	3.172	3.051	2.956	2.880	2.817	2.765	2.720	2.681
20	5.871	4.461	3.859	3.515	3.289	3.128	3.007	2.913	2.837	2.774	2.721	2.676	2.637
21	5.827	4.420	3.819	3.475	3.250	3.090	2.969	2.874	2.798	2.735	2.682	2.637	2.598
22	5.786	4.383	3.783	3.440	3.215	3.055	2.934	2.839	2.763	2.700	2.647	2.602	2.563
23	5.750	4.349	3.750	3.408	3.183	3.023	2.902	2.808	2.731	2.668	2.615	2.570	2.531
24	5.717	4.319	3.721	3.379	3.155	2.995	2.874	2.779	2.703	2.640	2.586	2.541	2.502
25	5.686	4.291	3.694	3.353	3.129	2.969	2.848	2.753	2.677	2.613	2.560	2.515	2.476
26	5.659	4.265	3.670	3.329	3.105	2.945	2.824	2.729	2.653	2.590	2.536	2.491	2.452
27	5.633	4.242	3.647	3.307	3.083	2.923	2.802	2.707	2.631	2.568	2.514	2.469	2.429
28	5.610	4.221	3.626	3.286	3.063	2.903	2.782	2.687	2.611	2.547	2.494	2.448	2.409
29	5.588	4.201	3.607	3.267	3.044	2.884	2.763	2.669	2.592	2.529	2.475	2.430	2.390
30	5.568	4.182	3.589	3.250	3.026	2.867	2.746	2.651	2.575	2.511	2.458	2.412	2.372
40	5.424	4.051	3.463	3.126	2.904	2.744	2.624	2.529	2.452	2.388	2.334	2.288	2.248
50	5.340	3.975	3.390	3.054	2.833	2.674	2.553	2.458	2.381	2.317	2.263	2.216	2.176
60	5.286	3.925	3.343	3.008	2.786	2.627	2.507	2.412	2.334	2.270	2.216	2.169	2.129
70	5.247	3.890	3.309	2.975	2.754	2.595	2.474	2.379	2.302	2.237	2.183	2.136	2.095
80	5.218	3.864	3.284	2.950	2.730	2.571	2.450	2.355	2.277	2.213	2.158	2.111	2.071
90	5.196	3.844	3.265	2.932	2.711	2.552	2.432	2.336	2.259	2.194	2.140	2.092	2.051
100	5.179	3.828	3.250	2.917	2.696	2.537	2.417	2.321	2.244	2.179	2.124	2.077	2.036
200	5.100	3.758	3.182	2.850	2.630	2.472	2.351	2.256	2.178	2.113	2.058	2.010	1.969
500	5.054	3.716	3.142	2.811	2.592	2.434	2.313	2.217	2.139	2.074	2.019	1.971	1.929
1000	5.039	3.703	3.129	2.799	2.579	2.421	2.300	2.204	2.126	2.061	2.006	1.958	1.916

$1 - \alpha = 0{,}975$

	14	15	16	17	18	19	20	21	22	23	24	25	26
1	982.545	984.874	986.911	988.715	990.345	991.800	993.081	994.303	995.351	996.341	997.272	998.087	998.843
2	39.427	39.431	39.436	39.439	39.442	39.446	39.448	39.450	39.452	39.455	39.457	39.458	39.459
3	14.277	14.253	14.232	14.213	14.196	14.181	14.167	14.155	14.144	14.134	14.124	14.115	14.107
4	8.684	8.657	8.633	8.611	8.592	8.575	8.560	8.546	8.533	8.522	8.511	8.501	8.492
5	6.456	6.428	6.403	6.381	6.362	6.344	6.329	6.314	6.301	6.289	6.278	6.268	6.258
6	5.297	5.269	5.244	5.222	5.202	5.184	5.168	5.154	5.141	5.128	5.117	5.107	5.097
7	4.596	4.568	4.543	4.521	4.501	4.483	4.467	4.452	4.439	4.426	4.415	4.405	4.395
8	4.130	4.101	4.076	4.054	4.034	4.016	3.999	3.985	3.971	3.959	3.947	3.937	3.927
9	3.798	3.769	3.744	3.722	3.701	3.683	3.667	3.652	3.638	3.626	3.614	3.604	3.594
10	3.550	3.522	3.496	3.474	3.453	3.435	3.419	3.403	3.390	3.377	3.365	3.355	3.345
11	3.359	3.330	3.304	3.282	3.261	3.243	3.226	3.211	3.197	3.184	3.173	3.162	3.152
12	3.206	3.177	3.152	3.129	3.108	3.090	3.073	3.057	3.043	3.031	3.019	3.008	2.998
13	3.082	3.053	3.027	3.004	2.983	2.965	2.948	2.932	2.918	2.905	2.893	2.882	2.872
14	2.979	2.949	2.923	2.900	2.879	2.861	2.844	2.828	2.814	2.801	2.789	2.778	2.767
15	2.891	2.862	2.836	2.813	2.792	2.773	2.756	2.740	2.726	2.713	2.701	2.689	2.679
16	2.817	2.788	2.761	2.738	2.717	2.698	2.681	2.665	2.651	2.637	2.625	2.614	2.603
17	2.753	2.723	2.697	2.673	2.652	2.633	2.616	2.600	2.585	2.572	2.560	2.548	2.538
18	2.696	2.667	2.640	2.617	2.596	2.576	2.559	2.543	2.529	2.515	2.503	2.491	2.481
19	2.647	2.617	2.591	2.567	2.546	2.526	2.509	2.493	2.478	2.465	2.452	2.441	2.430
20	2.603	2.573	2.547	2.523	2.501	2.482	2.464	2.448	2.434	2.420	2.408	2.396	2.385
21	2.564	2.534	2.507	2.483	2.462	2.442	2.425	2.409	2.394	2.380	2.368	2.356	2.345
22	2.528	2.498	2.472	2.448	2.426	2.407	2.389	2.373	2.358	2.344	2.332	2.320	2.309
23	2.497	2.466	2.440	2.416	2.394	2.374	2.357	2.340	2.325	2.312	2.299	2.287	2.276
24	2.468	2.437	2.411	2.386	2.365	2.345	2.327	2.311	2.296	2.282	2.269	2.257	2.246
25	2.441	2.411	2.384	2.360	2.338	2.318	2.300	2.284	2.269	2.255	2.242	2.230	2.219
26	2.417	2.387	2.360	2.335	2.314	2.294	2.276	2.259	2.244	2.230	2.217	2.205	2.194
27	2.395	2.364	2.337	2.313	2.291	2.271	2.253	2.237	2.222	2.208	2.195	2.183	2.171
28	2.374	2.344	2.317	2.292	2.270	2.251	2.232	2.216	2.201	2.187	2.174	2.161	2.150
29	2.355	2.325	2.298	2.273	2.251	2.231	2.213	2.196	2.181	2.167	2.154	2.142	2.131
30	2.338	2.307	2.280	2.255	2.233	2.213	2.195	2.178	2.163	2.149	2.136	2.124	2.112
40	2.213	2.182	2.154	2.129	2.107	2.086	2.068	2.051	2.035	2.020	2.007	1.994	1.983
50	2.140	2.109	2.081	2.056	2.033	2.012	1.993	1.976	1.960	1.945	1.931	1.919	1.907
60	2.093	2.061	2.033	2.008	1.985	1.964	1.944	1.927	1.911	1.896	1.882	1.869	1.857
70	2.059	2.028	1.999	1.974	1.950	1.929	1.910	1.892	1.876	1.861	1.847	1.833	1.821
80	2.035	2.003	1.974	1.948	1.925	1.904	1.884	1.866	1.850	1.835	1.820	1.807	1.795
90	2.015	1.983	1.955	1.929	1.905	1.884	1.864	1.846	1.830	1.814	1.800	1.787	1.774
100	2.000	1.968	1.939	1.913	1.890	1.868	1.849	1.830	1.814	1.798	1.784	1.770	1.758
200	1.932	1.900	1.870	1.844	1.820	1.798	1.778	1.759	1.742	1.726	1.712	1.698	1.685
500	1.892	1.859	1.830	1.803	1.779	1.757	1.736	1.717	1.700	1.684	1.669	1.655	1.641
1000	1.879	1.846	1.816	1.789	1.765	1.743	1.722	1.703	1.686	1.670	1.654	1.640	1.627

$1 - \alpha = 0{,}975$

	27	28	29	30	40	50	60	70	80	90	100	200	500	1000
1	999.54	1000.24	1000.82	1001.41	1005.59	1008.10	1009.79	1011.01	1011.91	1012.61	1013.16	1015.72	1017.24	
2	39.461	39.462	39.463	39.465	39.473	39.478	39.481	39.484	39.486	39.487	39.488	39.493	39.496	
3	14.100	14.093	14.086	14.081	14.036	14.010	13.992	13.979	13.970	13.962	13.956	13.929	13.913	
4	8.483	8.475	8.468	8.461	8.411	8.381	8.360	8.346	8.335	8.326	8.319	8.288	8.270	
5	6.250	6.242	6.234	6.227	6.175	6.144	6.123	6.107	6.096	6.087	6.080	6.048	6.028	
6	5.088	5.080	5.072	5.065	5.012	4.980	4.959	4.943	4.932	4.923	4.915	4.882	4.862	
7	4.386	4.378	4.370	4.362	4.309	4.276	4.254	4.239	4.227	4.218	4.210	4.176	4.156	
8	3.918	3.909	3.901	3.894	3.840	3.807	3.784	3.768	3.756	3.747	3.739	3.705	3.684	
9	3.584	3.576	3.568	3.560	3.505	3.472	3.449	3.433	3.421	3.411	3.403	3.368	3.347	
10	3.335	3.327	3.319	3.311	3.255	3.221	3.198	3.182	3.169	3.160	3.152	3.116	3.094	
11	3.142	3.133	3.125	3.118	3.061	3.027	3.004	2.987	2.974	2.964	2.956	2.920	2.898	
12	2.988	2.979	2.971	2.963	2.906	2.871	2.848	2.831	2.818	2.808	2.800	2.763	2.740	
13	2.862	2.853	2.845	2.837	2.780	2.744	2.720	2.703	2.690	2.680	2.671	2.634	2.611	
14	2.758	2.749	2.740	2.732	2.674	2.638	2.614	2.597	2.583	2.573	2.565	2.526	2.503	
15	2.669	2.660	2.652	2.644	2.585	2.549	2.524	2.506	2.493	2.482	2.474	2.435	2.411	
16	2.594	2.584	2.576	2.568	2.509	2.472	2.447	2.429	2.415	2.405	2.396	2.357	2.333	
17	2.528	2.519	2.510	2.502	2.442	2.405	2.380	2.362	2.348	2.337	2.329	2.289	2.264	
18	2.471	2.461	2.453	2.445	2.384	2.347	2.321	2.303	2.289	2.278	2.269	2.229	2.204	
19	2.420	2.411	2.402	2.394	2.333	2.295	2.270	2.251	2.237	2.226	2.217	2.176	2.150	
20	2.375	2.366	2.357	2.349	2.287	2.249	2.223	2.205	2.190	2.179	2.170	2.128	2.103	
21	2.335	2.325	2.317	2.308	2.246	2.208	2.182	2.163	2.148	2.137	2.128	2.086	2.060	
22	2.299	2.289	2.280	2.272	2.210	2.171	2.145	2.125	2.111	2.099	2.090	2.047	2.021	
23	2.266	2.256	2.247	2.239	2.176	2.137	2.111	2.091	2.077	2.065	2.056	2.013	1.986	
24	2.236	2.226	2.217	2.209	2.146	2.107	2.080	2.060	2.045	2.034	2.024	1.981	1.954	
25	2.209	2.199	2.190	2.182	2.118	2.079	2.052	2.032	2.017	2.005	1.996	1.952	1.924	
26	2.184	2.174	2.165	2.157	2.093	2.053	2.026	2.006	1.991	1.979	1.969	1.925	1.897	
27	2.161	2.151	2.142	2.133	2.069	2.029	2.002	1.982	1.966	1.954	1.945	1.900	1.872	
28	2.140	2.130	2.121	2.112	2.048	2.007	1.980	1.959	1.944	1.932	1.922	1.877	1.848	
29	2.120	2.110	2.101	2.092	2.028	1.987	1.959	1.939	1.923	1.911	1.901	1.855	1.827	
30	2.102	2.092	2.083	2.074	2.009	1.968	1.940	1.920	1.904	1.892	1.882	1.835	1.806	
40	1.972	1.962	1.952	1.943	1.875	1.832	1.803	1.781	1.764	1.751	1.741	1.691	1.659	
50	1.895	1.885	1.875	1.866	1.796	1.752	1.721	1.698	1.681	1.667	1.656	1.603	1.569	
60	1.845	1.835	1.825	1.815	1.744	1.699	1.667	1.643	1.625	1.611	1.599	1.543	1.507	
70	1.810	1.799	1.789	1.779	1.707	1.660	1.628	1.604	1.585	1.570	1.558	1.500	1.463	
80	1.783	1.772	1.762	1.752	1.679	1.632	1.599	1.574	1.555	1.540	1.527	1.467	1.428	
90	1.763	1.752	1.741	1.731	1.657	1.610	1.576	1.551	1.531	1.516	1.503	1.441	1.401	
100	1.746	1.735	1.725	1.715	1.640	1.592	1.558	1.532	1.512	1.496	1.483	1.420	1.378	
200	1.673	1.661	1.650	1.640	1.562	1.511	1.474	1.447	1.425	1.407	1.393	1.320	1.269	
500	1.629	1.617	1.606	1.596	1.515	1.462	1.423	1.394	1.370	1.351	1.336	1.254	1.192	
1000	1.614	1.603	1.591	1.581	1.499	1.445	1.406	1.376	1.352	1.332	1.316	1.230	1.162	

Anexo E

Base IRIS

Especie	Largo Sépalo	Ancho Sépalo	Largo Pétalo	Ancho Pétalo
I. setosa	5.1	3.5	1.4	0.2
I. setosa	4.9	3	1.4	0.2
I. setosa	4.7	3.2	1.3	0.2
I. setosa	4.6	3.1	1.5	0.2
I. setosa	5	3.6	1.4	0.2
I. setosa	5.4	3.9	1.7	0.4
I. setosa	4.6	3.4	1.4	0.3
I. setosa	5	3.4	1.5	0.2
I. setosa	4.4	2.9	1.4	0.2
I. setosa	4.9	3.1	1.5	0.1
I. setosa	5.4	3.7	1.5	0.2
I. setosa	4.8	3.4	1.6	0.2
I. setosa	4.8	3	1.4	0.1
I. setosa	4.3	3	1.1	0.1
I. setosa	5.8	4	1.2	0.2
I. setosa	5.7	4.4	1.5	0.4
I. setosa	5.4	3.9	1.3	0.4
I. setosa	5.1	3.5	1.4	0.3
I. setosa	5.7	3.8	1.7	0.3
I. setosa	5.1	3.8	1.5	0.3
I. setosa	5.4	3.4	1.7	0.2
I. setosa	5.1	3.7	1.5	0.4
I. setosa	4.6	3.6	I	0.2
I. setosa	5.1	3.3	1.7	0.5
I. setosa	4.8	3.4	1.9	0.2
I. setosa	5	3	1.6	0.2
I. setosa	5	3.4	1.6	0.4
I. setosa	5.2	3.5	1.5	0.2
I. setosa	5.2	3.4	1.4	0.2
I. setosa	4.7	3.2	1.6	0.2
I. setosa	4.8	3.1	1.6	0.2
I. setosa	5.4	3.4	1.5	0.4
I. setosa	5.2	4.1	1.5	0.1
I. setosa	5.5	4.2	1.4	0.2
I. setosa	4.9	3.1	1.5	0.2
I. setosa	5	3.2	1.2	0.2
I. setosa	5.5	3.5	1.3	0.2
I. setosa	4.9	3.6	1.4	0.1
I. setosa	4.4	3	1.3	0.2
I. setosa	5.1	3.4	1.5	0.2
I. setosa	5	3.5	1.3	0.3
I. setosa	4.5	2.3	1.3	0.3
I. setosa	4.4	3.2	1.3	0.2
I. setosa	5	3.5	1.6	0.6
I. setosa	5.1	3.8	1.9	0.4
I. setosa	4.8	3	1.4	0.3
I. setosa	5.1	3.8	1.6	0.2
I. setosa	4.6	3.2	1.4	0.2
I. setosa	5.3	3.7	1.5	0.2
I. setosa	5	3.3	1.4	0.2

Especie	Largo Sépalo	Ancho Sépalo	Largo Pétalo	Ancho Pétalo
I. versicolor	7	3.2	4.7	1.4
I. versicolor	6.4	3.2	4.5	1.5
I. versicolor	6.9	3.1	4.9	1.5
I. versicolor	5.5	2.3	4	1.3
I. versicolor	6.5	2.8	4.6	1.5
I. versicolor	5.7	2.8	4.5	1.3
I. versicolor	6.3	3.3	4.7	1.6
I. versicolor	4.9	2.4	3.3	1
I. versicolor	6.6	2.9	4.6	1.3
I. versicolor	5.2	2.7	3.9	1.4
I. versicolor	5	2	3.5	1
I. versicolor	5.9	3	4.2	1.5
I. versicolor	6	2.2	4	1
I. versicolor	6.1	2.9	4.7	1.4
I. versicolor	5.6	2.9	3.6	1.3
I. versicolor	6.7	3.1	4.4	1.4
I. versicolor	5.6	3	4.5	1.5
I. versicolor	5.8	2.7	4.1	1
I. versicolor	6.2	2.2	4.5	1.5
I. versicolor	5.6	2.5	3.9	1.1
I. versicolor	5.9	3.2	4.8	1.8
I. versicolor	6.1	2.8	4	1.3
I. versicolor	6.3	2.5	4.9	1.5
I. versicolor	6.1	2.8	4.7	1.2
I. versicolor	6.4	2.9	4.3	1.3
I. versicolor	6.6	3	4.4	1.4
I. versicolor	6.8	2.8	4.8	1.4
I. versicolor	6.7	3	5	1.7
I. versicolor	6	2.9	4.5	1.5
I. versicolor	5.7	2.6	3.5	1
I. versicolor	5.5	2.4	3.8	1.1
I. versicolor	5.5	2.4	3.7	1
I. versicolor	5.8	2.7	3.9	1.2
I. versicolor	6	2.7	5.1	1.6
I. versicolor	5.4	3	4.5	1.5
I. versicolor	6	3.4	4.5	1.6
I. versicolor	6.7	3.1	4.7	1.5
I. versicolor	6.3	2.3	4.4	1.3
I. versicolor	5.6	3	4.1	1.3
I. versicolor	5.5	2.5	4	1.3
I. versicolor	5.5	2.6	4.4	1.2
I. versicolor	6.1	3	4.6	1.4
I. versicolor	5.8	2.6	4	1.2
I. versicolor	5	2.3	3.3	1
I. versicolor	5.6	2.7	4.2	1.3
I. versicolor	5.7	3	4.2	1.2
I. versicolor	5.7	2.9	4.2	1.3
I. versicolor	6.2	2.9	4.3	1.3
I. versicolor	5.1	2.5	3	1.1
I. versicolor	5.7	2.8	4.1	1.3

Especie	Largo Sépalo	Ancho Sépalo	Largo Pétalo	Ancho Pétalo
I. virginica	6.3	3.3	6	2.5
I. virginica	5.8	2.7	5.1	1.9
I. virginica	7.1	3	5.9	2.1
I. virginica	6.3	2.9	5.6	1.8
I. virginica	6.5	3	5.8	2.2
I. virginica	7.6	3	6.6	2.1
I. virginica	4.9	2.5	4.5	1.7
I. virginica	7.3	2.9	6.3	1.8
I. virginica	6.7	2.5	5.8	1.8
I. virginica	7.2	3.6	6.1	2.5
I. virginica	6.5	3.2	5.1	2
I. virginica	6.4	2.7	5.3	1.9
I. virginica	6.8	3	5.5	2.1
I. virginica	5.7	2.5	5	2
I. virginica	5.8	2.8	5.1	2.4
I. virginica	6.4	3.2	5.3	2.3
I. virginica	6.5	3	5.5	1.8
I. virginica	7.7	3.8	6.7	2.2
I. virginica	7.7	2.6	6.9	2.3
I. virginica	6	2.2	5	1.5
I. virginica	6.9	3.2	5.7	2.3
I. virginica	5.6	2.8	4.9	2
I. virginica	7.7	2.8	6.7	2
I. virginica	6.3	2.7	4.9	1.8
I. virginica	6.7	3.3	5.7	2.1
I. virginica	7.2	3.2	6	1.8
I. virginica	6.2	2.8	4.8	1.8
I. virginica	6.1	3	4.9	1.8
I. virginica	6.4	2.8	5.6	2.1
I. virginica	7.2	3	5.8	1.6
I. virginica	7.4	2.8	6.1	1.9
I. virginica	7.9	3.8	6.4	2
I. virginica	6.4	2.8	5.6	2.2
I. virginica	6.3	2.8	5.1	1.5
I. virginica	6.1	2.6	5.6	1.4
I. virginica	7.7	3	6.1	2.3
I. virginica	6.3	3.4	5.6	2.4
I. virginica	6.4	3.1	5.5	1.8
I. virginica	6	3	4.8	1.8
I. virginica	6.9	3.1	5.4	2.1
I. virginica	6.7	3.1	5.6	2.4
I. virginica	6.9	3.1	5.1	2.3
I. virginica	5.8	2.7	5.1	1.9
I. virginica	6.8	3.2	5.9	2.3
I. virginica	6.7	3.3	5.7	2.5
I. virginica	6.7	3	5.2	2.3
I. virginica	6.3	2.5	5	1.9
I. virginica	6.5	3	5.2	2
I. virginica	6.2	3.4	5.4	2.3
I. virginica	5.9	3	5.1	1.8

Respuesta a Ejercicios

Capítulo 1

Ejercicio 3*:

120

Ejercicio 5*:

a) 120

b) 24

c) 48

Ejercicio 7*:

a) 45

b) 21

c) 35

Ejercicio 9*:

a) $S = (1,c);(1,s);(2,c);(2,s);(3,c);(3,s);(4,c)(4,s);(5,c);(5,s);(6,c);(6,s)$

b) 1- $A = (2,c);(4,c);(6,c)$

2- $B = (1,c);(1,s);(2,c);(2,s);(3,c);(3,s);(5,c);(5,s)$

3- $C = (1,s);(3,s);(5,s)$

c) 1- $A \cup B = (1,c);(1,s);(2,c);(4,c);(6,c);(2,s);(3,c);(3,s);(5,c);(5,s)$

2- $A \cap B = (2,c)$

3- $B - (A \cap C) = (1,c);(2,s);(3,c);(5,c)$

d) A y C

Ejercicio 11*:

a) 0.2637

b) 0.4945

c) 0.7363

Ejercicio 13*:

a) 0.1667

b) 0.3333

c) 0.2727

Ejercicio 15*:
0.175
Ejercicio 17*:
0.5
Ejercicio 18*:
0.25 , No
Ejercicio 19*:
0,978

Capítulo 2
Ejercicio 1*:

x_i	2	3	4	5	6	7	8	9	10	11	12
$f(x)$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

x_i	1	2	3	4	5	6
$f(y)$	11/36	9/36	7/36	5/36	3/36	1/36

Ejercicio 4*:
1.9375
Ejercicio 5*:

x_i	0	1	2
$f(x)$	640/125	48/125	12/125

Ejercicio 6*:

a) 0.125

b) 0.375

c) 0.375

d) 0.875

Ejercicio 8*:

a) 0.2963

b) 0.9877

c) 0.5926

Ejercicio 9*:
$E(x) = 4$, $p(x = 4) = 0{,}2734$
Ejercicio 10*:
$E(x) = 200$, $\sigma = 14$
Ejercicio 12*:
Binomial: $n > 229{,}1$
Poisson: $n > 230{,}26$
Ejercicio 13*:
0.180
Ejercicio 16*:
0.201

Capítulo 3

Ejercicio 1*:

a) Cumple $f(x) \geq 0$ y $\int_{-\infty}^{+\infty} f(x)dx = 1$

b) $F(x) = \begin{cases} 0 & \text{si } x < 0 \\ x^2 & \text{si } 0 \leq x \leq 1 \\ 1 & \text{si} x > 1 \end{cases}$

c) 0.25

Ejercicio 3*:

a) $k = 1/12$

b) $E(x) = 15$

c) $F(x) = \begin{cases} 0 & \text{si } x < 10 \\ \frac{x-10}{10} & \text{si } 10 \leq x \leq 20 \\ 1 & \text{si} x > 20 \end{cases}$

Ejercicio 5*:

a) 0.2212

b) 0.2865

Ejercicio 6*:

a) 1) -0.75

2) 0

3) 1

4) 1.5

b) 1) 6.2

2) 8.9

3) 8

4) 9.5

Ejercicio 7*:

a) 0.4222

b) 0.2673

c) 0.8925

d) 0.6569

e) 0.6687

f) 0.1292

g) 0.3830

Ejercicio 8*:

a) 1.65

b) -0.85

c) -1.28

d) 0.67

e) 0.475

f) 2.575

Ejercicio 10*:

a) 294

b) 92

Capítulo 4

Ejercicio 1*:
0.69
Ejercicio 2*:
No
Ejercicio 3*:

a) 0.733

b) 0.734

Ejercicio 4*:
0.0495

Capítulo 6

Ejercicio 1*:

a) (493,36; 509,04)

b) (490,88; 511,32)

Ejercicio 2*:
$n > 24{,}01$
Ejercicio 3*:
$\alpha = 1$
Ejercicio 4*:
(9,76; 11,20)
Ejercicio 5*:

a) (79,64; 134,42)

b) (74, 44; 148,67)

Capítulo7

Ejercicio 1*:

a) $h_0 : \mu = 2600$, $h_a : \mu \neq 2600$

b) $z_c = \pm 1{,}96$

c) No, no puede concluirse.

Ejercicio 2*:

a) $h_0 : \mu \leq 1000$, $h_a : \mu > 1000$

b) Sí

Ejercicio 5*:
No
Ejercicio 7*:
Sí
Ejercicio 9*:
No

Capítulo 9

Ejercicio 1*:
b) $y = 1{,}95x + 0{,}1$
c) $r = 0{,}99$
Ejercicio 3*:
b) $y = 3{,}61x + 22{,}4$, $r = 0{,}895$
c) $80{,}1\,\%$
d) Sí

Bibliografía

[1] Bass, J.: *Elementos de Cálculo de Probabilidades Teórico y Aplicado*, Editorial Toray-Masson, (1975).

[2] Canavos, G.C.: *Probabilidad y Estadística. Aplicaciones y Métodos*, Editorial Mc Graw-Hill(1987).

[3] Chung, K.L.: *Teoría Elemental de los Procesos Estocásticos*, Editorial Reverté.

[4] Cramér, H.: *Métodos Matemáticos de Estadística*, Editorial Aguilar. S.A.,(1970).

[5] Cramér, H.: *Teoría de Probabilidades y Aplicaciones*, Editorial Aguilar. S.A.,(1977).

[6] Feller, W.: *An Introduction to Probability Theory and its Application*, John Wiley Sons, Inc.,(1971).

[7] Gmurman, V.E. : *Teoría de las Probabilidades y Estadística Matemática*,(1975).

[8] Jacovskis, P. y Perazzo, R.: *Azar, Ciencia y Sociedad*, Editorial EUDEBA,(2012).

[9] Kolmogorov, A.N.: *Foundations of the Theory of Probability*, Chelsea Publishing Company,(1950).

[10] Landro, A.H. y Gonzalez, M.L.: *Cuadernos de Teoría de la Probabilidad y sus Aplicaciones*, Editorial EUDEBA,(1993).

[11] Meyer, P.: *Probabilidad y Aplicaciones Estadísticas*, Fondo Educativo Interamericano. S. A.,(1973).

[12] Montgomery, D. ; Peck,E. y Vining,G.: *Introducción al Análisis de Regresión Lineal*, Grupo Editorial Patria.

[13] Rosenthal, J.: *A First Look at Rigorous Probability Theory*, World Scientific,(2006).

[14] Toranzos, F.I.: *Teoría Estadística y Aplicaciones*, Editorial EUDEBA,(1997).

[15] http://www.infostat.com.ar

[16] http://www.r-project.org

Índice de contenidos

editora
UNIVERSIDAD NACIONAL DE MORENO

Este libro es el resultado del trabajo realizado por los docentes a cargo de la asignatura Probabilidad y Estadística (2027), correspondiente al Ciclo Inicial de la carrera de Ingeniería en Electrónica, al cabo del cual los estudiantes acceden al título intermedio de Técnico Universitario en Electrónica de la Universidad Nacional de Moreno.

Se trata de un trabajo de gran utilidad para los estudiantes de ingeniería en general, ya que reúne los conocimientos básicos imprescindibles para entender los conceptos fundamentales de Probabilidad y Estadística y tiene la riqueza y el acervo científico pertinente para que pueda ser utilizado en las otras tecnicaturas y carreras del campo de las Ciencias Aplicadas y Tecnología.

EL Mg. Cristóbal R. SANTA MARÍA es Licenciado en Matemática y Magister en Explotación de Datos y Descubrimiento del Conocimiento de la UBA, y docente-investigador de grado y posgrado en varias universidades, incluida la Universidad Nacional de Moreno.

La Lic. Claudia S. BUCCINO es Licenciada en Enseñanza de la Matemática de la UTN y docente del nivel medio, superior y universitario, incluida la Universidad Nacional de Moreno.

ISBN 978-987-782-010-2

UNM 2010
UNIVERSIDAD
DEL BICENTENARIO
ARGENTINO

DEPARTAMENTO DE
CIENCIAS APLICADAS Y TECNOLOGÍA
COLECCIÓN:
BIBLIOTECA UNIVERSITARIA