UNED

Introducción al análisis de datos

Ana Julia Garriga Trillo
Paula Lubin Pigouche
José M.ª Merino Merino
Miguel Padilla Suárez
Patricia Recio Saboya
Juan Carlos Suárez Falcón

Psicología

COLECCIÓN GRADO

Introducción al Análisis de Datos

ANA JULIA GARRIGA-TRILLO
PAULA LUBIN PIGOUCHE
JOSÉ MARÍA MERINO MERINO
MIGUEL PADILLA SUÁREZ
PATRICIA RECIO SABOYA
JUAN CARLOS SUÁREZ FALCÓN

UNIVERSIDAD NACIONAL DE EDUCACIÓN A DISTANCIA

(6201103GR01A01)
INTRODUCCIÓN AL ANÁLISIS DE DATOS

Librería UNED: c/ Bravo Murillo, 38 - 28015 Madrid
Tels.: 91 398 75 60 / 73 73
e-mail: libreria@adm.uned.es

ISBN: 978-84-362-6042-7
Depósito legal: M. 17.302-2010

Primera edición: septiembre de 2009
Primera reimpresión: abril de 2010

Impreso en España - Printed in Spain

ÍNDICE

PRÓLOGO

Esta obra ha sido concebida como un libro de texto dirigido a los estudiantes de primer curso del Grado en Psicología de la Universidad Nacional de Educación a Distancia, y ha sido elaborado por el equipo docente de la asignatura.

Las características de los alumnos a los que va dirigido y la metodología de la propia UNED hacen que el objetivo del texto no sea tratar nuevas técnicas ni recoger aportaciones originales sino presentar una serie de contenidos, que se encuentran ya en muchos otros libros, de manera clara y sencilla. En este sentido, hemos tratado de presentar los conceptos fundamentales seguidos de ejemplos concretos aplicados —en la medida de lo posible— a la Psicología y hemos prescindido de desarrollos matemáticos que no sean estrictamente necesarios.

Aunque en el análisis de datos resulta casi imprescindible la utilización del ordenador, no hacemos referencia en el texto a ningún software concreto (de las denominadas hojas de cálculo o programas estadísticos de los que existe una gran variedad en el mercado —Excel, SPSS, OpenStat...—) ni a todas las posibilidades que ofrece internet. Esto se debe a que creemos necesario que el alumno aprenda primero a resolver «manualmente» análisis de pequeños conjuntos de datos y a que la asignatura cuenta con un «curso virtual» en la red donde se tratarán todos estos aspectos.

Los contenidos presentados, a nivel introductorio en muchos casos, responden a las dos partes fundamentales que se consideran en el análisis de datos: descriptiva e inferencia. Están organizados en temas, donde los cuatro primeros, dedicados a la estadística descriptiva, recogen los conceptos fundamentales, la organización de datos y su representación gráfica (tema 1); los índices de tendencia central, posición y variabilidad (temas 2 y 3) y la correlación y regresión lineal (tema 4). Los temas restantes constituyen una primera aproximación a la inferencia y, puesto que ésta se realiza siempre en términos probabilísticos, comienzan con las nociones básicas de probabilidad (tema 5). Los temas 6 y 7 se dedican a las distribuciones discretas y continuas (haciendo especial hincapié en la distribución normal por su amplia utilización en Psicología) y, finalmente, se considera el muestreo y la estimación (tema 8).

Cada uno de los temas comienza con una introducción, donde se presentan los objetivos de aprendizaje a conseguir en su estudio, y se han señalado en negrita los términos fundamentales que aparecen a lo largo del texto. También se han colocado dentro de cuadros las fórmulas y definiciones más importantes, se han resaltado los ejemplos y se ha añadido un resumen. Por último, al final de cada tema, se presenta un gran número de ejercicios, con sus soluciones correspondientes, que permiten la autoevaluación del alumno.

Finalmente, queremos señalar que este texto es fruto de una amplia experiencia del equipo docente. Durante todos esos años han sido muchos los alumnos que nos han ayudado a intentar mejorar nuestros textos. A todos ellos nuestro agradecimiento. Los aciertos, si los hay, en este texto se deben a ellos, los errores son sólo nuestros.

Los autores,
Madrid, marzo de 2009

Tema 1

Conceptos básicos y organización de datos

1.1. INTRODUCCIÓN

En la actualidad, la estadística se aplica en casi todas las áreas del saber, y de una forma muy importante en las ciencias sociales y naturales. Sirva como ejemplo su utilización en estudios epidemiológicos (Medicina), en estudios toxicológicos relacionados con la eficacia de los medicamentos (Farmacia), en estudios genéticos y de impacto ambiental (Biología), en muestreos en las prospecciones petrolíferas o hidráulicas (Geología), en los censos de población e información demográfica (Sociología), y en estudios sobre la optimización del coste-beneficio (Economía).

Se puede hacer, por tanto, una distinción entre estadística teórica y aplicada; la primera se ocupa de los aspectos formales y normativos, y la segunda constituye la aplicación a un campo concreto, como los ejemplos que acabamos de ver. Esta estadística aplicada ha recibido distintas denominaciones según su campo de aplicación, tales como bioestadística, psicoestadística o socioestadística. Algunos autores han propuesto para la estadística aplicada la denominación de análisis de datos (Botella, León y San Martín, 1993; Merino y otros, 2007; Pardo y San Martín, 1998), término cuyo uso se está extendiendo en los nuevos planes de estudio de Psicología y que da nombre a este libro.

A pesar de su diversidad de aplicaciones, esta disciplina no es popular ni entre los estudiantes de ciencias sociales ni entre muchos profesionales de estas ciencias, debido posiblemente a la imagen de la estadística como una rama de las matemáticas de difícil comprensión y carácter abstracto. En contradicción directa con esta percepción, está la imagen de la estadística como una especie de instrumento mágico que impregna de carácter científico cualquier investigación que la utilice.

En este tema explicaremos el papel que juega la estadística en el análisis de los datos en Psicología, analizaremos los distintos niveles de medida (nominal, ordinal, de intervalo y de razón), definiremos el concepto de variable, así como su clasificación y notación simbólica, además de explicar cómo organizar los datos en una distribución de frecuencias y cómo construir representaciones gráficas de los datos para considerar, de un solo vistazo, las características del fenómeno estudiado.

Los objetivos marcados para este capítulo son:

- Ubicar la importancia de la materia Análisis de datos en el plan de estudios del grado en Psicología.
- Establecer las relaciones entre el análisis de datos y el método científico, reconociendo la utilidad de la estadística en el análisis de datos psicológicos.
- Diferenciar y manejar los conceptos básicos, la nomenclatura y las definiciones centrales de la estadística, a fin de poder aplicarlos en el estudio formal de la materia.
- Entender la importancia de la medición en el ámbito psicológico, distinguiendo entre las distintas escalas de medida (nominal, ordinal, de intervalo y de razón), y conociendo las relaciones que pueden establecerse en cada una de ellas.
- Manejar con soltura las distintas denominaciones y clasificaciones de las variables.
- Saber elaborar, a partir de un conjunto de datos, una distribución de frecuencias, adquiriendo y desarrollando la capacidad para recopilar, organizar, presentar, e interpretar datos numéricos.
- Aplicar las técnicas de representación gráfica adecuadas en función de los datos disponibles (diagrama de barras, diagrama de sectores, pictograma, histograma y polígono de frecuencias).

1.2. LA INVESTIGACIÓN EN PSICOLOGÍA

A lo largo de la historia, el hombre se ha servido de diversas formas de conocimiento, tales como la religión, el sentido común o el folclore popular. Con la aparición de la ciencia moderna en el siglo XVII, el método científico pasó a ser la fuente de conocimiento más utilizada, aunque no la única. La Psicología se sirve del método científico para acercarse a su objeto de estudio: la conducta. El ***método científico*** consiste en dar razón sistemática, empírica y en lo posible experimental, de los fenómenos (Yela, 1994). El método científico se caracteriza por ser sistemático y replicable. ***Sistemático*** porque es un procedimiento que tiene unas etapas definidas y ***replicable*** porque los datos obtenidos mediante su uso tienen que poder ser replicados o refutados por cualquier investigador interesado. El método científico, por tanto, proporciona una manera de actuar para afrontar una investigación, a través de las siguientes fases interdependientes:

1. Definición del problema.
2. Deducción de hipótesis contrastables.
3. Establecimiento de un procedimiento de recogida de datos.
4. Análisis de los resultados obtenidos.
5. Discusión de dichos resultados y búsqueda de conclusiones.
6. Elaboración de un informe de la investigación.

Esta asignatura se ocupa de la cuarta fase de una investigación, el análisis de los resultados obtenidos. En las asignaturas ***Fundamentos de Investigación*** y ***Diseños de Investigación*** se tratarán de manera detallada el resto de las fases de una investigación científica, así como los posibles diseños a utilizar y el análisis correspondiente a cada uno de ellos.

Veámoslo con un ejemplo:

Ejemplo 1.1. Hay evidencia en la literatura de la influencia del estrés en la hipertensión arterial. En particular, se considera que determinadas estrategias encaminadas a combatir el estrés pueden resultar beneficiosas para controlar la hipertensión arterial. Un investigador desea estudiar este fenómeno en un grupo de 40 pacientes con este trastorno, para lo cual divide su muestra en dos grupos: (1) pacientes que reciben el tratamiento estándar para la hipertensión con medicamentos y (2) pacientes que, además del tratamiento estándar, reciben una terapia de afrontamiento de situaciones estresantes. Al final del tratamiento se recogieron los datos de la tensión arterial de los pacientes, además de una serie de características sociodemográficas: *sexo*, *edad*, *estado civil*, *nivel de estudios*, *número de hijos*, *altura* y *peso*.

Las dos primeras fases son la definición del problema y la deducción de hipótesis contrastables. En nuestro ejemplo, el problema objeto de estudio es la posible influencia de las estrategias sobre el control del estrés en la hipertensión arterial y, puesto que una hipótesis no es más que la solución tentativa de un problema, nuestra hipótesis fundamental sería que el grupo (2) que recibe una terapia de afrontamiento de situaciones estresantes tendrá unos niveles menores de hipertensión arterial en comparación con el grupo que solo recibe el tratamiento estándar.

En la siguiente fase se encontraría la determinación de un plan de trabajo o procedimiento para la recogida de datos, es decir la elección de un diseño de investigación. En el ejemplo 1.1 el investigador decide escoger como muestra a los 40 pacientes que acuden a las consultas externas del hospital donde trabaja, asignándolos de manera aleatoria en los grupos (1) y (2), para que reciban un tratamiento diferente, comparando después sus resultados.

Comparar los resultados conlleva el análisis de los datos obtenidos y la discusión de dichos resultados (fases 4 y 5). Aquí se analizarían los niveles de tensión de ambos grupos para comprobar si realmente el grupo (2) que ha recibido el doble tratamiento psicológico y farmacológico obtiene niveles más bajos en tensión arterial. Por último, para difundir los resultados de la investigación se elabora un informe (fase 6).

En este libro de texto, se explicarán de manera detallada los análisis de datos básicos que pueden ser necesarios realizar tanto en la investigación psicológica como en el ejercicio profesional. Las técnicas estadísticas constituyen una parte integral no solo de la actividad investigadora, sino también del análisis de los datos que se originan en las actividades que desarrollan las instituciones y organizaciones. En este sentido, no hay que olvidar que el psicólogo que comprenda los conceptos estadísticos y su metodología sacará mejor provecho de ellos, ya que estará más preparado para evaluar los resultados de una investigación y podrá leer con mayor sentido crítico la literatura que, sobre su campo de acción, va día a día apareciendo.

1.3. CONCEPTO Y FUNCIONES DE LA ESTADÍSTICA: DESCRIPTIVA E INFERENCIAL

La ***Estadística*** se utiliza como tecnología al servicio de las Ciencias donde la variabilidad y la incertidumbre forman parte de su naturaleza. Así, se ocupa de la sistematización, recogida, ordenación y presentación de los datos referentes a un fenómeno que presenta variabilidad o incertidumbre para su estudio metódico, con objeto de hacer previsiones sobre los mismos, tomar decisiones u obtener conclusiones.

Teniendo en cuenta las funciones de la estadística, podemos considerar dos grandes áreas: la Estadística Descriptiva y la Estadística Inferencial.

Mediante la ***Estadística Descriptiva*** se organizan y resumen conjuntos de observaciones procedentes de una muestra o de la población total, en forma cuantitativa. Este resumen puede hacerse mediante tablas, gráficos o valores numéricos. Así, se dispone de distintos procedimientos que nos permiten estudiar las características de una o más variables:

— En el caso de una variable, podemos recurrir a índices que nos indicarán cuáles son los valores más habituales de esa variable (índices de ***tendencia central***), hasta qué punto esos valores son similares o diferentes entre sí (estadísticos de ***variabilidad***) y en qué grado las observaciones se reparten equilibradamente por encima y por debajo de la tendencia central (estadísticos de ***asimetría***). Estos conceptos se aprenderán de manera intuitiva al final de este tema, y de manera formal en los temas 2 y 3.

— En el caso de dos variables podemos utilizar índices que nos indiquen hasta qué punto están ambas variables relacionadas entre sí (***coeficientes de correlación***), así como procedimientos que nos permitirán predecir el valor de una variable en función de otra (***ecuaciones de regresión***). El tema 4 abordará de manera detallada ambos procedimientos.

Mediante la ***Estadística Inferencial*** se realizan inferencias acerca de una población basándose en los datos obtenidos a partir de una muestra. Para realizar estas generalizaciones de la muestra a la población total se utiliza el cálculo de probabilidades. Los últimos capítulos de este texto tratan sobre ***probabilidad*** e ***inferencia estadística***.

En una investigación cualquiera, lo habitual es que se desee conocer un ***parámetro*** o característica de los elementos de una ***población***; sin embargo, la población suele ser demasiado extensa para estudiarla al completo (conllevaría un coste inabordable). Por este motivo, se realiza un muestreo con el que se obtiene una muestra de elementos que representan a la población y se estudia la característica deseada en la ***muestra*** mediante ***estadísticos*** que se utilizarán para estimar los parámetros de la población.

En este sentido, en nuestro ejemplo es de esperar que el investigador esté interesado en estudiar si el tratamiento combinando es útil para tratar la hipertensión de las personas hipertensas en general. Por tanto su población objetivo serían las personas que padecen hipertensión. Dado que no es posible acceder a todas las personas hipertensas, escoge una muestra de 40 que son las que realmente participan en la investigación.

Un ejemplo harto conocido para todos es el de los sondeos electorales. Imaginemos, por ejemplo, que estamos interesados en predecir el resultado de un referéndum que se celebrará próximamente en España. La población objeto de estudio serían todos los españoles mayores de 18 años que son los que pueden votar; no sería posible preguntar a todos por su intención de voto por lo que escogemos una muestra representativa de 5.000 españoles y les preguntamos por el sentido de su voto en el referéndum. Deseamos conocer un ***parámetro***: el porcentaje de individuos de la población que responderían «sí»; eso no es posible, pero sí lo es conocer la estimación de ese parámetro, el ***estadístico*** o porcentaje de la muestra que responde «sí».

Estos conceptos pueden definirse de la siguiente manera:

Población es el conjunto de todos los elementos que cumplen una determinada característica objeto de estudio.

Muestra es un subconjunto cualquiera de una población.

Parámetro: es una propiedad descriptiva (una medida) de una población.

Estadístico: es una propiedad descriptiva (una medida) de una muestra.

Aunque cualquier subconjunto de una población recibe el nombre de muestra, las conclusiones obtenidas en una muestra solo servirán para el total de la población si la muestra es ***representativa***. Para seleccionar muestras que revelen adecuadamente las características de la población es necesario utilizar métodos de muestreo probabilístico, ya que una ***muestra probabilística*** se elige mediante reglas matemáticas, por lo que la probabilidad de selección de cada unidad es conocida de antemano. Por el contrario, una ***muestra no probabilística*** no se rige por las reglas matemáticas de la probabilidad. Ejemplos de éstas últimas son la muestra de conveniencia o incidental (que está conformada por personas de fácil acceso para el investigador como colegas o alumnos de su clase) y la obtenida mediante el muestreo «bola de nieve» (un elemento de la población lleva a otro y así sucesivamente).

1.4. MEDICIÓN Y ESCALAS DE MEDIDA

Cuando se trata de objetos físicos el proceso de medición es directo y generalmente sencillo porque es cuestión de seguir unas reglas prescritas expresadas mediante determinadas escalas. Así por ejemplo, es fácil medir la estatura de una persona asignando el número correspondiente de la cinta métrica a la distancia que hay desde sus pies hasta su cabeza. Cuando se trata de medir la timidez de un estudiante en una situación de interacción social, medir ya no es tan sencillo. El reto al que se enfrenta la Psicología es su necesidad de medir en muchas ocasiones variables que no son directamente observables.

> Medición es el proceso por el cual se asignan números a objetos o características según determinadas reglas.

Teniendo en cuenta que llamamos ***característica*** a cualquier propiedad de objetos o personas que deseamos estudiar y ***modalidad*** a las distintas formas de presentarse esta característica, esta definición implica asignar un número a cada una de las modalidades de una característica, convirtiendo algunas relaciones entre modalidades en sus correspondientes relaciones entre los números que representan su medida. Por ejemplo, a las dos modalidades de la variable *sexo* (hombre y mujer) se les puede asignar los números 1 y 2, y al *peso de una rata* se le puede asignar el número en gramos que da la balanza.

Hay que tener en cuenta que no es lo mismo medir el *número de hijos* de una familia nuclear monógama, que la *nacionalidad* de un conjunto de estudiantes en intercambio o el *tiempo* que tarda un roedor en recorrer un laberinto en forma de T. Por este motivo se utilizan diferentes escalas de medida en función de la variable a medir, entendiendo como escala de medida al conjunto de reglas o modelos desarrollados para la asignación de números a los valores de las variables. Un ejemplo de escala de medida es la escala centígrada de temperatura, que se basa en asignar 0° a la temperatura de congelación del agua y 100° a la de ebullición.

En función de las relaciones que puedan verificarse empíricamente entre las modalidades de las características, y siguiendo la clasificación de Stevens (1946), pueden distinguirse cuatro tipos de escala de medida: nominal, ordinal, de intervalo y de razón.

En la escala nominal solo distinguiremos la igualdad o desigualdad entre dos modalidades, la escala ordinal añade la posibilidad de establecer un orden, en la escala de intervalo se usa una unidad y tienen sentido las diferencias y, por último, en la escala de razón se pueden comparar dos medidas mediante un cociente.

a) Escala nominal

La escala de medida nominal consiste en la asignación, puramente arbitraria de números o símbolos a cada una de las diferentes modalidades de la característica. Por tanto, la única relación que se tiene en cuenta es la de ***igualdad*** (y la ***desigualdad***), que implica la pertenencia o no a una categoría determinada.

Usando una escala nominal podemos decidir si un sujeto es igual o diferente a otro, pero no podemos establecer relaciones de orden respecto a esa característica, ni de cantidad. Por ejemplo, si realizamos una distinción hipotética entre católicos: (1) «practicantes» y (2) «no practicantes», carece de sentido establecer relaciones entre estos dos números del tipo 1 + 1 = 2, o 2 – 1 = 1. En el primer caso estaríamos diciendo algo así como que dos católicos «practicantes» es igual a un católico «no practicante», y en el segundo que un católico «no practicante» menos uno «practicante es igual a otro «practicante».

En las variables nominales se puede asignar a cada modalidad cualquier tipo de símbolo. En el ejemplo anterior, en lugar de números podríamos haber utilizado (P) para designar a los «practicantes» y (No P) a los «no practicantes». En el ejemplo 1.1 sobre tratamiento de la hipertensión serían variables nominales el *grupo*, el *sexo* y el *estado civil*.

b) Escala ordinal

En la escala ordinal se asignan números a objetos para indicar la extensión relativa en que se posee una característica. Se clasifica a las personas, eventos u objetos en una posición con relación a cierto atributo, pero sin indicar la distancia que hay entre las posiciones. Cuando se asignan números es sólo para indicar el ***orden*** de las posiciones de lo que se está clasificando.

Esta escala no solo permite la identificación y diferenciación de los sujetos sino que además permite establecer relaciones del tipo «mayor que» o «menor que», aunque no se plantea una distancia entre unas medidas y otras. En este caso, la asignación de números a las distintas categorías no puede ser completamente arbitraria, debe hacerse atendiendo al orden existente entre éstas.

Un ejemplo sería la variable *estatus socioeconómico* con tres supuestas modalidades: (1) «bajo», (2) «medio» y (3) «alto»; en este caso los números no solo indican una diferencia de modalidades sino también verificar un orden entre ellas, de mayor a menor, o viceversa. En nuestro ejemplo sobre hipertensión, la variable *nivel de estudios* sería ordinal, porque además de ser estudios diferentes, podemos afirmar que la persona Diplomada tiene más estudios que la que estudió Secundaria.

c) Escala de intervalo

Las escalas de intervalos son aquellas que ordenan los objetos o eventos según la magnitud del atributo que representan y proveen intervalos iguales entre las unidades de medida. Con la escala de intervalo, los números asignados a los objetos, no solo permiten decidir si un objeto es igual o diferente a otro o si posee en mayor o menor grado la característica de interés, sino que estos números se pueden sumar y restar, y además, las diferencias entre esos números se pueden multiplicar y dividir.

Lo que caracteriza a una escala de intervalo es la existencia de una ***unidad de medición*** común y constante, que permite asignar un número real a todos los pares de objetos del conjunto ordenado. En la escala de intervalo el origen es ***arbitrario***, y no refleja en ningún momento ausencia de la magnitud que estamos midiendo.

La *inteligencia* medida con un test es un ejemplo de escala de intervalo. Si cuatro personas, A, B, C y D han obtenido 80, 90, 150 y 160 puntos en un test de inteligencia, podemos decir que la diferencia en inteligencia entre A y B es la misma que entre C y D (90 – 80 = 160 – 150), ya que el test proporciona una unidad de medida estable. Sin embargo, no podemos afirmar que C sea el doble de inteligente que A aunque tenga el doble de puntuación en el test, ya que para realizar una afirmación de ese tipo sería necesario que el cero de la escala fuera absoluto. En este caso es arbitrario porque obtener un cero en un test de inteligencia no refleja ausencia de la característica medida, no significa que no se posea ni un ápice de inteligencia.

d) Escala de razón

En la escala de razón los números asignados a los objetos admiten como válidas las relaciones de igualdad-desigualdad, orden, suma, resta, multiplicación y división.

Se caracteriza porque tiene todas las características de una medida de intervalo y, además, se le puede asignar un punto de origen verdadero de valor cero, es decir, el valor cero de esta escala significa ausencia de la magnitud que estamos midiendo. Dado que el cero ya no es arbitrario, sino un valor ***absoluto***, podemos decir que A tiene dos, tres o cuatro veces la magnitud de la propiedad presente en B. La *altura* y el *peso* son dos ejemplos típicos de escala de razón.

En el cuadro 1.1 se resumen los tipos de escalas, las características básicas de cada una de ellas, las relaciones que admiten, así como algunos ejemplos.

Cuadro 1.1. Resumen de las escalas de medida

Tipo	Características básicas	Relaciones válidas	Ejemplos
Nominal	Los números identifican y clasifican objetos	Relaciones del tipo «igual que» o «distinto que».	Sexo, estado civil, raza, diagnóstico clínico.
Ordinal	Además, los números indican las posiciones relativas de los objetos	Además, relaciones del tipo «mayor que» o «menor que»	Dureza, nivel socioeconómico, posición en el ranking de la ATP, grado de satisfacción.
Intervalo	Además, hay una unidad de medición común	Además, igualdad o desigualdad de diferencias	Temperatura, fecha de calendario, inteligencia
Razón	Además, el punto cero es absoluto	Además, igualdad o desigualdad de razones	Longitud, peso, altura, tiempo de reacción, coste económico

Es importante señalar que en muchas ocasiones el nivel de medida de una variable viene determinado por el papel que desempeña en la investigación. Así, una misma variable, por ejemplo la estatura, que en principio cumple las características necesararias para ser considerada de razón, puede resultar ordinal si el investigador se limita a operar con tres niveles de estatura: bajo, medio y alto.

1.5. VARIABLES: CLASIFICACIÓN Y NOTACIÓN

> Una variable es una representación numérica de una característica que presenta más de una modalidad (valor) de un conjunto determinado.

Si una característica tiene una única modalidad, se trata de una ***constante***.

Según el nivel de medición que les sea aplicado, podemos clasificar las variables en nominales, ordinales, de intervalo y de razón. Para cada tipo de variable existen unos procedimientos estadísticos apropiados para hacer el mejor uso de la información que contienen los valores de las variables.

Es habitual la distinción en la literatura científica de tres grandes tipos de variables: ***cualitativa***, ***cuasicuantitativa*** y ***cuantitativa***, perteneciendo las variables nominales al primer tipo, las ordinales al segundo, y las de intervalo y razón al tercer tipo.

Las variables cualitativas se clasifican además, en función del número de categorías o modalidades que presentan. Si una variable presenta sólo dos categorías se dice que es una ***variable dicotómica*** (por ejemplo, el *sexo*); si presenta más de dos categorías se dice que es una ***variable politómica*** (por ejemplo, la *nacionalidad*).

En ocasiones se categorizan variables que podrían medirse a un nivel superior; en este caso decimos que una variable se ha dicotomizado si se han establecido dos categorías, y politomizada si se han establecido más de dos categorías. Un ejemplo sería la variable *peso* del roedor de un experimento: aunque podríamos medir exactamente su peso en gramos, puede resultar útil en una investigación dicotomizar la variable *peso* clasificando a las ratas en peso alto y bajo, o politomizarla, estableciendo tres o más niveles de peso.

Las variables cuantitativas se clasifican, además, en función de los valores numéricos que pueden asignarse, en continuas y discretas.

Una ***variable continua*** es aquella para la que los individuos pueden tener valores en cualquier punto de una escala ininterrumpidamente. Es decir, para una variable continua, dados dos valores, siempre se puede encontrar un tercer valor que esté incluido entre los dos primeros. Un ejemplo de variable continua es el *peso*, ya que entre los valores 79 y 80 kg podemos considerar uno, dos, tres o todos los decimales que se quiera. Una ***variable discreta*** es aquella que adopta valores aislados. Por tanto, fijados dos consecutivos, no se puede tomar ningún valor intermedio. Un ejemplo de variable discreta es el *número de hijos* (huelga decir que se pueden tener dos hijos o tres, pero nunca un valor intermedio entre ambos).

La figura 1.1 recoge las clasificaciones de las variables comentadas en el texto.

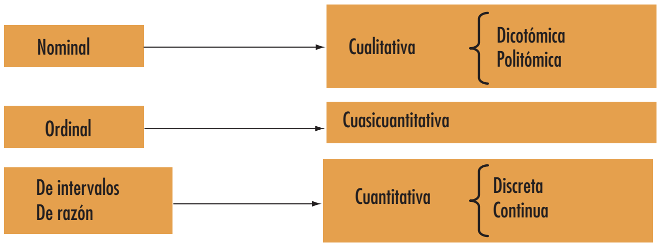


Figura 1.1. Clasificación de las variables (tomado de Merino y otros, 2007).

En el ámbito de la metodología científica se emplea otra clasificación del término variable, con un sentido diferente al propuesto en este libro, distinguiendo, por ejemplo, entre variable independiente, variable dependiente y variable extraña. Una ***variable independiente*** es cualquier suceso que sospechamos que pueda ser la causa de otro y que estamos interesados en estudiar. Llamamos ***variable dependiente*** a la medida utilizada para determinar los posibles efectos de la variable independiente. Por último, las ***variables extrañas*** son todas aquellas variables que pueden influir sobre la variable dependiente pero que no nos interesa estudiar sus efectos.

En el ejemplo 1.1 la variable independiente es el *tipo de tratamiento* y tiene dos valores posibles: (1) tratamiento estándar y (2) tratamiento estándar más terapia. Se espera que el utilizar uno u otro tratamiento influya en la *hipertensión arterial*, por lo que ésta será nuestra variable dependiente. En esta investigación no nos interesa estudiar la *obesidad*, aunque sabemos que se relaciona con la hipertensión, por lo que el peso puede considerarse una variable extraña.

En cuanto a la notación, para representar a las variables se utilizan letras latinas mayúsculas. Para referirnos a un valor cualquiera de la variable X se utiliza el subíndice i (X_i), siendo n el número de elementos que componen la muestra, por lo que, de manera genérica, designaremos la variable como:

$$X_i \text{ siendo } i = 1, 2, 3\ldots, n$$

1.6. DISTRIBUCIÓN DE FRECUENCIAS

En los apartados anteriores ha quedado de manifiesto que en Psicología se trabaja con datos de variables que pueden ser nominales, ordinales, de intervalo o de razón, con las peculiaridades propias de cada escala. Estos datos pueden provenir de la medición directa de estas variables (*peso de una rata*, *tiempo empleado en realizar una tarea*, *rendimiento académico*, etc.), o de frecuencias que provienen de un proceso de conteo (*número de matriculados en un curso académico*, *número de pacientes en un hospital psiquiátrico*, etc.).

En cualquier caso, una vez que el investigador ha recabado la información mediante el procedimiento de recogida de datos correspondiente, dispone de un listado de datos. Si tenemos muy pocos valores es posible que la simple inspección visual de los mismos sea suficiente para describir el fenómeno estudiado. Pero esto no es nada frecuente. Habitualmente necesitamos, por tanto, organizar la información mediante una distribución de frecuencias.

Una distribución de frecuencias es una representación de la relación entre un conjunto de medidas exhaustivas y mutuamente excluyentes y la frecuencia de cada una de ellas (Hays, 1988).

Además de la organización de los datos, la distribución de frecuencias cumple dos funciones fundamentales: ofrecer la información necesaria para realizar representaciones gráficas y facilitar los cálculos para obtener los estadísticos muestrales que serán objeto de atención en los próximos capítulos.

Veámoslo basándonos en el ejemplo 1.1. Se presentan los datos de los 40 pacientes en las variables: *sexo, edad, estado civil, nivel de estudios, número de hijos, altura, peso* e *hipertensión arterial mínima y máxima*. Los datos de todas estas personas, junto con un número de identificación *ID*, aparecen en la Tabla 1.1.

La tercera columna de la tabla nos informa sobre el *sexo* de los participantes; sin embargo, la simple inspección visual de estos datos no es suficiente para que el investigador se haga una idea precisa de los hombres y mujeres que hay; es necesario construir una distribución de frecuencias.

Para construir la tabla de distribución de frecuencias (Tabla 1.2) se inspeccionan en primer lugar los valores que toma la variable. En este caso se trata de una variable de carácter cualitativo (nominal) que puede adoptar dos valores distintos, hombre y mujer. En la primera columna especificamos los valores que adopta nuestra variable X. En la segunda columna aparece la ***frecuencia absoluta*** (n_i) que es el número de observaciones en cada categoría. En la siguiente columna aparece la ***frecuencia relativa o proporción de cada categoría*** (p_i), que se obtiene dividiendo la frecuencia absoluta, n_i, entre el número total de observaciones, que se representa por n. La frecuencia relativa también se expresa en términos de ***porcentaje*** (P_i) para lo cual hay que multiplicar cada una de las proporciones por cien (cuarta columna).

Tabla 1.1. Datos recogidos en la investigación del ejemplo 1.1

ID	Grupo	Sexo	Edad	Estado Civil	Nivel de Estudios	N.º Hijos	Altura	Peso	T.A. Mínima	T.A. Máxima
1	1	Hombre	50	Casado	Licenciatura	2	1,75	77	78	142
2	1	Mujer	26	Soltero	Diplomatura	0	1,7	69	87	137
3	1	Hombre	35	Casado	ESO	3	1,62	65	92	133
4	1	Hombre	64	Casado	ESO	4	1,68	85	81	147
5	1	Hombre	34	Soltero	Primaria	0	1,64	72	83	143
6	1	Mujer	28	Casado	ESO	1	1,65	67	72	154
7	1	Hombre	73	Viudo	ESO	4	1,62	69	87	148
8	1	Mujer	45	Casado	Primaria	2	1,68	70	88	118
9	1	Hombre	47	Divorciado	FP	2	1,82	84	84	123
10	1	Hombre	52	Casado	Licenciatura	1	1,76	85	82	129
11	1	Mujer	54	Casado	ESO	0	1,59	63	90	141
12	1	Mujer	67	Casado	Primaria	2	1,55	62	93	145
13	1	Mujer	74	Casado	ESO	1	1,63	68	95	149
14	1	Hombre	26	Soltero	Diplomatura	1	1,74	78	88	138
15	1	Hombre	35	Viudo	FP	0	1,73	83	86	133
16	1	Hombre	56	Casado	Licenciatura	3	1,77	79	83	158
17	1	Hombre	69	Divorciado	Primaria	1	1,59	70	85	144
18	1	Mujer	57	Casado	ESO	0	1,67	68	88	152
19	1	Hombre	48	Divorciado	FP	0	1,79	89	89	139
20	1	Mujer	29	Casado	Primaria	1	1,71	79	89	127
21	2	Hombre	35	Casado	Licenciatura	2	1,82	89	80	141
22	2	Hombre	52	Casado	ESO	1	1,63	75	84	145
23	2	Mujer	49	Casado	Primaria	2	1,58	72	87	142
24	2	Hombre	47	Casado	FP	1	1,83	89	89	135
25	2	Mujer	55	Casado	Primaria	1	1,59	72	80	136
26	2	Mujer	69	Divorciado	ESO	3	1,58	63	82	153
27	2	Hombre	75	Viudo	Licenciatura	2	1,62	71	83	146
28	2	Hombre	28	Soltero	Primaria	3	1,69	83	79	138
29	2	Mujer	48	Casado	FP	0	1,72	69	93	143
30	2	Hombre	64	Casado	Primaria	1	1,67	75	83	147
31	2	Mujer	71	Casado	Primaria	1	1,57	58	81	149
32	2	Mujer	29	Divorciado	FP	2	1,73	69	77	131
33	2	Hombre	44	Casado	Diplomatura	2	1,69	72	79	140
34	2	Hombre	48	Divorciado	Primaria	1	1,59	78	85	133
35	2	Hombre	59	Soltero	FP	0	1,68	72	77	143
36	2	Mujer	58	Casado	Primaria	1	1,73	80	78	138
37	2	Hombre	47	Viudo	ESO	3	1,63	75	72	150
38	2	Mujer	49	Soltero	Diplomatura	2	1,67	67	74	153
39	2	Hombre	37	Casado	Primaria	0	1,79	84	76	148
40	2	Hombre	57	Casado	ESO	3	1,73	86	76	144

Tabla 1.2. Distribución de frecuencias de la variable *sexo*

X	n_i	p_i	P_i
Hombre	24	0,6	60
Mujer	16	0,4	40
	n = 40	1	100

Pues bien, ahora sí podemos hacernos una idea de la distribución por sexo de los participantes, sabemos que hay más hombres que mujeres en la investigación (24 vs. 16), lo que en porcentaje corresponde al 60% de hombres frente al 40% de mujeres.

En el caso de variables ordinales se procede de la misma manera, aunque con las modalidades situadas en la tabla de acuerdo a un determinado orden. Por ejemplo, la variable *nivel de estudios* presenta los niveles «Primaria, ESO, FP, Diplomatura, Licenciatura» y en la distribución de frecuencias hay que preservar este orden:

Tabla 1.3. Distribución de frecuencias de la variable *nivel de estudios*

X	n_i	p_i	P_i	n_a	p_a	P_a
Primaria	13	0,32	32	13	0,32	32
ESO	11	0,28	28	24	0,6	60
FP	7	0,18	18	31	0,78	78
Diplomatura	4	0,1	10	35	0,88	88
Licenciatura	5	0,12	12	40	1	100
	40	1	100			

En esta tabla se han añadido tres columnas más: la ***frecuencia absoluta acumulada*** (n_a), la ***frecuencia relativa acumulada o proporción acumulada*** (p_a) y el ***porcentaje acumulado*** (P_a), para cada una de las categorías o modalidades de respuestas. Para obtener estos valores, simplemente hay que ir acumulando (sumando), desde la categoría de menor valor de la variable a la de mayor valor, las frecuencias absolutas, proporciones o porcentajes, de cada categoría de respuesta. Por ejemplo, la frecuencia absoluta acumulada en el caso de «Diplomatura» es 35, resultado de sumar las frecuencias de los órdenes anteriores (13 + 11 + 7 = 31) y la suya propia (31 + 4 = 35), indicando que 35 personas presentan un nivel de Diplomatura o menos. Nótese que en las variables nominales carece de sentido el cálculo de las frecuencias acumuladas.

Los conceptos explicados hasta el momento son:

Frecuencia absoluta **(n_i)**: número de veces que se repite cada uno de los valores de una variable. La suma de todas las frecuencias absolutas representa el total de la muestra (n).

Proporción o frecuencia relativa **(p_i)**: cociente entre la frecuencia absoluta de cada valor de la variable (n_i) y el número total de observaciones (n). Formalmente $p_i = n_i/n$.

Porcentaje **(P_i)**: valor de la frecuencia relativa (p_i) multiplicado por cien. Formalmente: $P_i = p_i \times 100$.

Frecuencia absoluta acumulada **(n_a)**: número de veces que se repita cada modalidad y cualquiera de las modalidades inferiores.

Proporción acumulada o frecuencia relativa acumulada **(p_a)**: cociente entre la frecuencia absoluta acumulada de cada clase y el total de observaciones. Formalmente $p_a = n_a/n$.

Porcentaje acumulado **(P_a)**: valor de la frecuencia relativa acumulada multiplicado por cien. Formalmente: $P_a = p_a \times 100$.

Cuando trabajamos con variables cuantitativas nos podemos encontrar con dos casos: el número de valores que toma la variable es reducido (como la variable *número de hijos* del ejemplo) o es muy amplio (*edad, altura, peso, nivel de hostilidad, nivel de estrés*). En el primer caso procederemos de la forma indicada para variables ordinales y en el segundo será necesario agrupar la variable en intervalos.

La variable *edad* del ejemplo 1 forma parte de este segundo caso. El participante de menor edad tiene 26 años y el mayor 75. Si construyésemos una tabla de distribución de frecuencias como la anterior tendríamos una lista demasiado extensa (50 filas) y muchas de las frecuencias serían cero. En estos casos se recurre a la agrupación en ***intervalos***, que consiste en formar grupos de valores consecutivos de la variable, situando cada uno de estos grupos en una fila, y calculando la frecuencia de cada grupo o intervalo de valores, y no de cada valor de la variable.

La variable *edad* toma valores entre 26 y 75 años ambos incluidos, por tanto, puede adoptar 50 valores distintos (75 – 26 + 1 = 50). En primer lugar hay que decidir qué número de intervalos tendrá la distribución de frecuencias. Siempre habrá varias posibilidades pudiendo optar desde establecer un número muy pequeño de intervalos muy amplios hasta muchos intervalos de muy pequeña ***amplitud***. A la hora de tomar esta decisión hay que tener presente que al establecer intervalos siempre se pierde información, ya que ahora la frecuencia no estará referida a un solo valor de la variable, sino a todos los contenidos en el intervalo. Por tanto, esta decisión dependerá del tratamiento que el investigador quiera dar a la variable en su estudio, tratando de encontrar el equilibrio entre la precisión que necesite y la manejabilidad de los datos.

En el ejemplo que nos ocupa, consideramos que unos intervalos de amplitud 10 serán apropiados para la variable *edad*. El valor más pequeño es 26, por lo que el primer intervalo contendrá las edades 26, 27, 28, 29, 30, 31, 32, 33, 34 y 35 y el último 66, 67, 68, 69, 70, 71, 72, 73, 74 y 75. La tabla de distribución de frecuencias agrupada sería la siguiente:

Tabla 1.4. Distribución de frecuencias con los datos agrupados en intervalos de la variable *edad* del ejemplo 1.1

Intervalos X_i	n_i	p_i	n_a	p_a
26-35	10	0,250	10	0,250
36-45	3	0,075	13	0,325
46-55	13	0,325	26	0,650
56-65	7	0,175	33	0,825
66-75	7	0,175	40	1
	40	1		

Según esta tabla de distribución de frecuencias agrupadas, la variable edad no puede tomar valores entre 45 y 46 años, lo cual no es cierto en la práctica. Debemos considerar esta variable como continua en el intervalo, es decir, tiene que poder adoptar cualquier valor entre 26 y 75. Por tanto, los límites exactos del intervalo 26-35 son 25,5-35,5, los del intervalo 36-45 son 35,5-45,5 y así sucesivamente, de forma que el ***límite superior exacto*** de un intervalo coincida con el ***límite inferior exacto*** del siguiente. Los límites de los intervalos que aparecen en la tabla 1.4 reciben el nombre de ***límites informados o aparentes***.

En este ejemplo, el cálculo de los límites exactos de los intervalos es trivial porque al ser la edad un número entero basta con sumar y restar 0,5 al límite superior e inferior respectivamente para su cálculo. Sin embargo, cuando los límites aparentes contienen decimales, puede resultar de utilidad aplicar la siguiente fórmula:

Límites exactos = Valor informado $\pm$ 0,5 $\times$ I.

siendo I la unidad del instrumento de medida.

Por ejemplo, si se mide el tiempo que se emplea en ejecutar una determinada tarea, y para ello se utiliza un cronómetro con precisión de centésimas de segundo (0,01), calcularemos el tiempo real de un tiempo aparente de 22,37, sumando y restando 0,005 a este valor. Así:

Intervalo valor real = 22,37 $\pm$ 0,5 $\times$ 0,01 = 22,37 $\pm$ 0,005 = 22,365 – 22,375.

Tabla 1.5. Distribución de frecuencias con los datos agrupados en intervalos de la variable *edad* del ejemplo 1.1

Límites exactos X_i	**Límites aparentes X_i**	**Punto medio**	n_i	p_i	n_a	p_a
25,5-35,5	26-35	30,5	10	0,250	10	0,250
35,5-45,5	36-45	40,5	3	0,075	13	0,325
45,5-55,5	46-55	50,5	13	0,325	26	0,650
55,5-65,5	56-65	60,5	7	0,175	33	0,825
65,5-75,5	66-75	70,5	7	0,175	40	1
			40	1		

A partir de los límites informados o de los límites exactos calculamos el punto medio del intervalo, que es la semisuma del límite superior e inferior del intervalo. Con estos datos, completamos la distribución de frecuencias de la variable *edad* del ejemplo 1.1.

A un intervalo que, por lo menos teóricamente, no tiene límite inferior o límite superior se le denomina ***intervalo abierto***. Por ejemplo, si en la variable *edad* del ejemplo 1.1 tuviéramos dos sujetos de 98 y 99 años podríamos optar por establecer el intervalo abierto «76 años o más», en lugar de añadir los tres intervalos correspondientes 76-85 y 86-95 y 96-105, dos de ellos con frecuencia nula.

Los nuevos conceptos que han aparecido son:

Intervalo: sinónimo del concepto de modalidad, es cada uno de los grupos de valores que ocupan una fila en una distribución de frecuencias.

Límites aparentes, virtuales o informados: son los valores mayor y menor de cada intervalo, teniendo en cuenta el nivel de precisión del instrumento de medida.

Límites reales o exactos: son los valores máximo y mínimo que tendría cada intervalo si el instrumento de medida tuviera una precisión perfecta.

Punto medio del intervalo: es la semisuma de los límites exactos o de los límites aparentes.

Amplitud del intervalo: es la diferencia entre el límite exacto superior y el límite exacto inferior.

1.7. REPRESENTACIONES GRÁFICAS

Un gráfico es una forma rápida e intuitiva de visualizar un conjunto de datos o una distribución de frecuencias.

En toda representación gráfica se encuentra subyacente la idea de un sistema de coordenadas, consistiendo el más habitual en dos líneas perpendiculares. La línea o eje vertical se llama ordenada o eje de las Y y la línea o eje horizontal se denomina abscisa o eje de las *X*. Ambos ejes dividen al plano en cuatro cuadrantes, y el punto donde se cruzan ambos ejes se denomina origen (véase Figura 1.2).

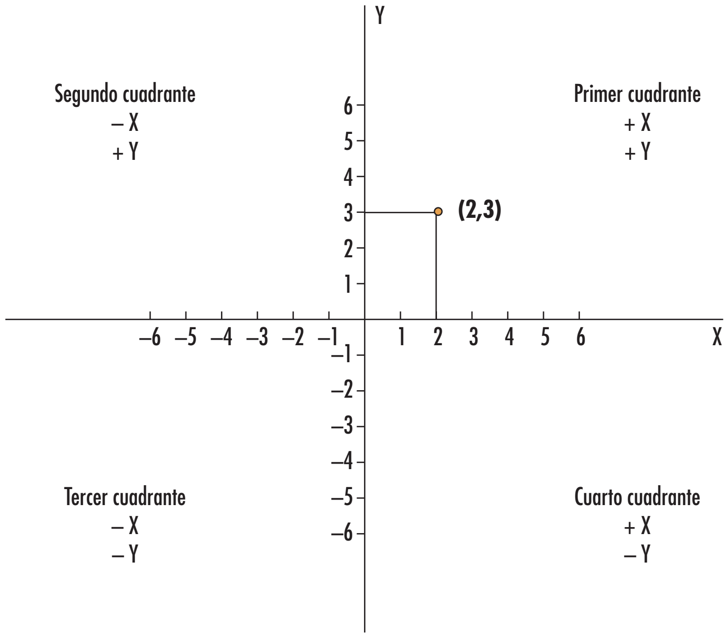


Figura 1.2. Sistema referencial de Coordenadas Cartesianas.

Cuando se representan variables cualitativas, o en puntuaciones positivas en general, lo habitual es representar únicamente el primer cuadrante.

Para elegir el tipo de gráfico más apropiado hay que tener en cuenta el nivel de medida de la variable. A continuación se describen las representaciones gráficas más utilizadas en Psicología para una y dos variables.

1.7.1. Representación gráfica de una variable

a) Diagrama de barras

Este tipo de representaciones se suele utilizar para ***variables nominales, ordinales y cuantitativas discretas***. En el eje de abscisas se colocan los distintos valores de la variable y en el eje de ordenadas las frecuencias.

Sobre cada valor de la variable se traza un rectángulo o barra perpendicular cuya altura debe ser igual a la frecuencia, ya sea absoluta o relativa (ver figura 1.3).

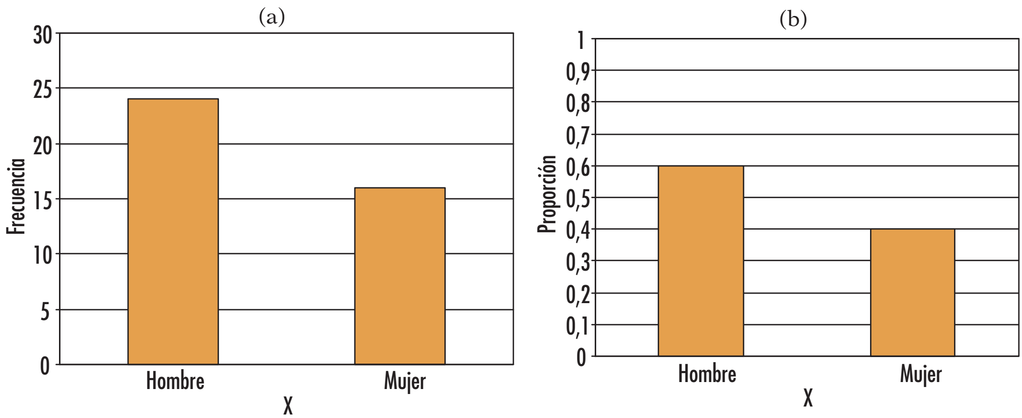


Figura 1.3. Diagrama de barras con frecuencias absolutas (a) y relativas (b) construido sobre la variable *sexo*.

En variables ordinales y cuantitativas discretas, se puede utilizar además un ***Diagrama de barras acumulativo***, que nos permite conocer cuántas observaciones se sitúan por debajo de un valor cualquiera de la variable. Hay que situar en el eje de abscisas los valores de la variable y en el de ordenadas las frecuencias acumuladas, ya sean absolutas o relativas, trazando sobre cada valor una perpendicular cuya longitud sea igual a la frecuencia acumulada. En la figura 1.4 se muestran un diagrama de barras (a)

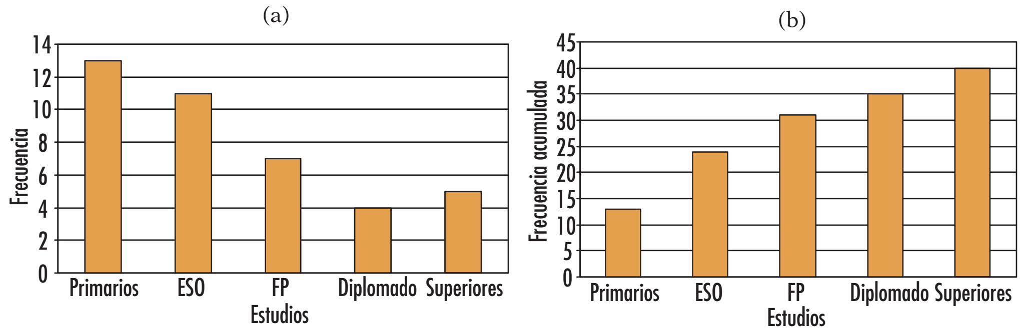


Figura 1.4. Diagrama de barras (a) y diagrama de barras acumulativo (b) de la variable *nivel de estudios*.

y diagrama de barras acumulativo (b) para la variable *nivel de estudios*. En el primer caso, cada barra representa únicamente un valor de la variable, mientras que en el segundo representa ese valor y los valores inferiores. Por ejemplo, la barra relativa a los estudios de Diplomatura representa únicamente a estos estudios en el gráfico (a) y a los estudios de Diplomatura, FP, ESO y Primaria en el gráfico (b).

b) Diagrama de sectores

Es una representación en forma de círculo en las que éste se divide en secciones cuya superficie es proporcional a la frecuencia de la modalidad correspondiente. El ángulo total del círculo, representa el número total de observaciones, y para determinar el ángulo de los sectores de cada modalidad se multiplica la frecuencia relativa (proporción) por 360, que es el número de grados de una circunferencia. En su representación, es habitual indicar el porcentaje obtenido en cada valor de la variable. Se utiliza para ***variables cualitativas*** y ***cuasicuantitativas***.

En la figura 1.5 representamos gráficamente la variable *sexo*, a la que se ha añadido una nueva columna en la distribución de frecuencias, con los grados correspondientes a cada categoría o modalidad para su representación mediante un diagrama de sectores.

X	n_i	p_i	P_i	**Grados**
Hombre	24	0,6	60	216
Mujer	16	0,4	40	144
	n = 40	1	100	360

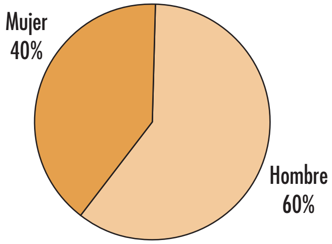


Figura 1.5. Distribución de frecuencias y diagrama de sectores de la variable *sexo* del ejemplo 1.1.

c) Pictograma

Los pictogramas expresan con dibujos, símbolos, mapas, etc. alusivos al objeto de estudio las frecuencias de las modalidades de la variable. Estos gráficos se hacen representando a diferentes escalas un mismo dibujo, de tal manera que el área de cada uno de ellos sea proporcional a la frecuencia de la modalidad que representa. Es un error hacer la representación con una escala tal que el perímetro del dibujo sea proporcional a la frecuencia, ya que a frecuencia doble, correspondería un dibujo de área cuádruple, lo que daría un efecto visual engañoso. Una solución práctica es incluir una referencia indicando la frecuencia a la que equivale cada símbolo del gráfico.

Los pictogramas se utilizan habitualmente para ***variables cualitativas***. A continuación mostramos una representación, mediante un pictograma, de la variable *sexo*.

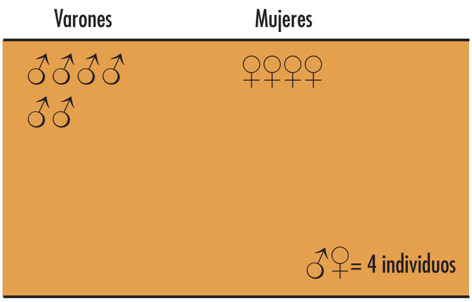


Figura 1.6. Pictograma de la variable sexo del ejemplo 1.1.

d) Histograma

El histograma se utiliza para variables ***cuantitativas continuas con datos agrupados en intervalos***. En el eje de abscisas se colocan los límites exactos de cada uno de los intervalos en que se han agrupado los datos (todos con la misma amplitud), o los puntos medios de los intervalos y sobre ellos se levantan rectángulos cuyas áreas sean proporcionales a la frecuencia correspondiente, absoluta o relativa, según se quiera representar una u otra. También se utiliza para la distribución de frecuencias acumuladas. En la figura 1.7 se muestra el histograma (a) e histograma acumulativo (b) para la variable *edad*. Al igual que en el diagrama de barras, en el primer caso, cada rectángulo representa únicamente un valor de la variable, mientras que en el segundo representa ese valor y los valores inferiores.

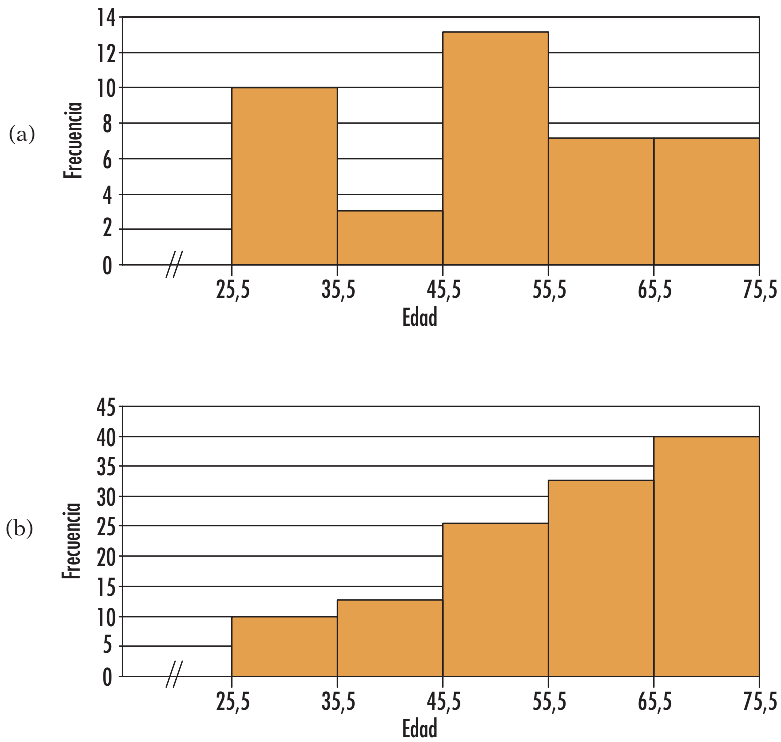


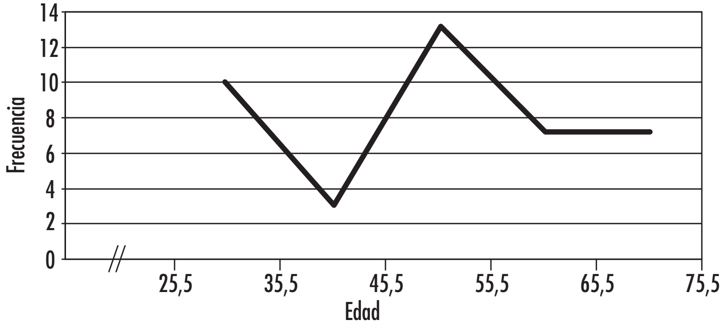
Figura 1.7. Histograma (a) e histograma acumulativo (b) de la variable *edad*.

e) Polígono de frecuencias

Para variables discretas, el polígono de frecuencias es la figura que resulta de unir los extremos superiores de las que hubieran sido las barras si se hubiera hecho un diagrama de barras. Si se trata de una variable continua, podemos decir lo mismo pero referido a los puntos medios de las bases superiores de los rectángulos correspondientes a un hipotético histograma construido con esos mismos datos. En la figura 1.8 (a) se presenta el polígono de frecuencias de la variable *edad* del ejemplo 1.1 Es habitual representar el histograma junto con el polígono de frecuencias en un mismo gráfico, tal y como se presenta en la figura 1.8(b).

En variables continuas también se utiliza el polígono de frecuencias acumuladas incluyendo en el eje de ordenadas las frecuencias acumuladas, ya sean absolutas o relativas. Para realizarlo, se une, mediante un segmen-

(a)



(b)

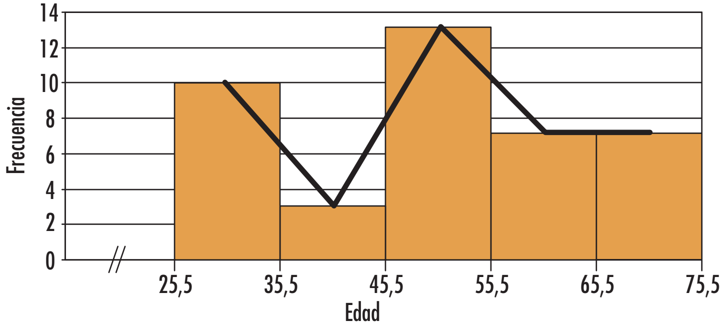


Figura 1.8. Polígono de frecuencias (a) y polígono de frecuencias realizado sobre la base del histograma (b) de la variable edad.

to rectilíneo, el vértice inferior izquierdo del primer rectángulo (el situado a la izquierda de todos) con su vértice superior derecho; este punto con el vértice superior derecho del siguiente rectángulo, y así sucesivamente. En la Figura 1.9(a) se muestra el polígono de frecuencias acumuladas para la variable *edad* y en la Figura 1.9(b) se puede comprobar cómo se ha construido a partir del histograma correspondiente.

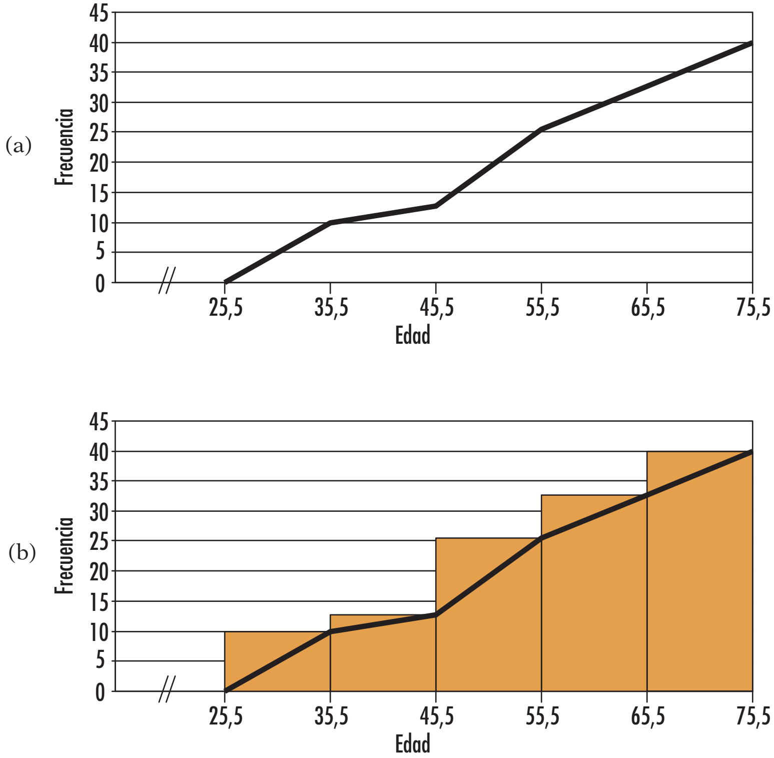


Figura 1.9. Polígono de frecuencias acumulado (a) y polígono de frecuencias acumulado realizado sobre la base del histograma (b) de la variable edad.

1.7.2. Representación gráfica de dos variables

a) Diagrama de barras conjunto

Se utiliza cuando ***al menos una de las dos variables es cualitativa***. Cuando ambas variables son cualitativas, antes de realizar la representación gráfica conviene organizar los datos en una ***tabla de doble entrada***. En este tipo de tablas se sitúan los valores de una de las variables en las filas y los valores de la otra variable en las columnas.

Continuando con el ejemplo 1.1, tenemos en la siguiente tabla (ver Tabla 1.6) los cuatro posibles valores (casado, divorciado, soltero y viudo) de la variable *estado civil* en las filas, y los dos posibles valores (hombre y mujer) de la variable *sexo* en las columnas. Cada celdilla representa la fre-

Tabla 1.6. Distribución conjunta de las variables *estado civil* y *sexo* del Ejemplo 1.1

	Hombre	Mujer	
Casado	12	12	24
Divorciado	4	2	6
Soltero	4	2	6
Viudo	4	0	4
	24	16	40

cuencia o número de elementos que reúne a la vez los valores de las dos variables que se cruzan en cada casilla. Como puede observarse en la tabla, doce es el número de personas que reúnen los requisitos de ser hombre y casado, cuatro es el número de personas que reúnen los requisitos de ser hombre y divorciado, y así sucesivamente.

En el siguiente gráfico se muestran dos formas de representar gráficamente las variables *estado civil* y *sexo* mediante un diagrama de barras:

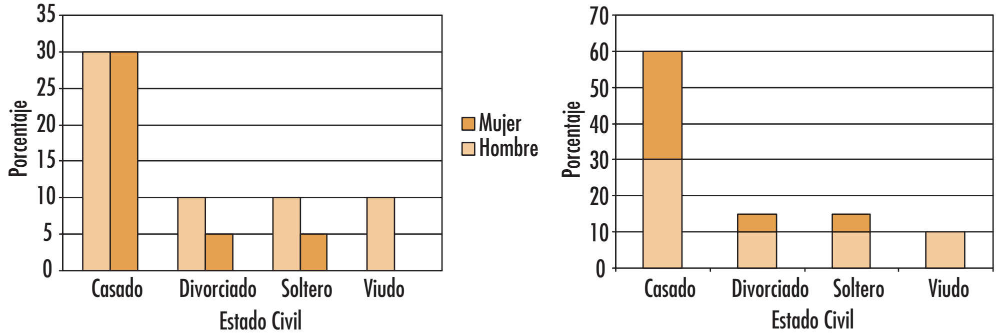


Figura 1.10. Diagramas de barras conjunto de las variables *estado civil* y *sexo*.

En ambos casos se ha situado la variable *estado civil* en el eje de abscisas, utilizando distinto trazo para cada modalidad de la variable *sexo*. Cuando se representan dos variables conjuntamente, hay que tener en cuenta que para utilizar las frecuencias absolutas es conveniente que el número de sujetos sea similar en las dos variables, siendo preferible en caso contrario utilizar las frecuencias relativas o porcentajes.

b) Diagrama de dispersión o nube de puntos

Se utiliza en el caso de **dos *variables cuantitativas***, dando una idea de la relación que existe entre ambas variables. Se sitúa una de las variables en el eje de abscisas y la otra en el eje de ordenadas. Para cada par de datos, se localiza la intersección de ambas variables y se marca con un punto.

En el gráfico se muestra la representación conjunta de las variables *altura* y *peso* del Ejemplo 1.1 (ver datos de ambas variables en Tabla 1.1).

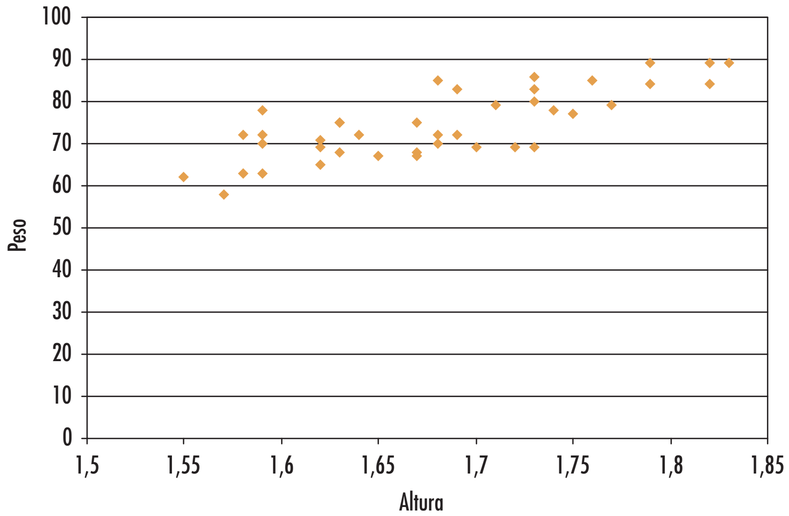


Figura 1.11. Diagrama de dispersión de las variables *altura* y *peso*.

Atendiendo al diagrama de dispersión, podemos observar que existe cierta ***relación lineal*** entre las variables *altura* y *peso*, correspondiendo, en mayor medida, pesos altos a alturas mayores y viceversa. En el tema cuatro tendremos la oportunidad de explicar la relación entre variables con más detenimiento.

1.8. PROPIEDADES DE LA DISTRIBUCIÓN DE FRECUENCIAS

Hemos visto cómo construir una distribución de frecuencias y cómo representar gráficamente sus datos; ahora mostraremos, también de forma gráfica, las posibles formas que adoptan las distribuciones de frecuencia teniendo en cuenta sus propiedades básicas: la ***tendencia central***, la ***variabilidad*** y la ***asimetría***. Estas propiedades tienen sus correspondientes índices, cuyo cálculo será objeto de estudio en los dos próximos capítulos.

Para ilustrar las propiedades de las distribuciones de frecuencia utilizaremos las distribuciones de frecuencia y las representaciones gráficas de las variables relativas a la tensión arterial del Ejemplo 1.1.

a) Tendencia central

La tendencia central de una distribución se refiere al lugar donde se centra una distribución particular en la escala de valores. Así, para la variable *tensión arterial máxima* del Ejemplo 1.1 la mayoría de los sujetos se encuentran entre los valores 136 y 150 mientras que, para la *tensión mínima*, se encuentran entre 76 y 90 (ver Figuras 1.12 y 1.13).

X	n_i
111-115	0
116-120	1
121-125	1
126-130	2
131-135	5
136-140	7
141-145	11
146-150	8
151-155	4
156-160	1
	40

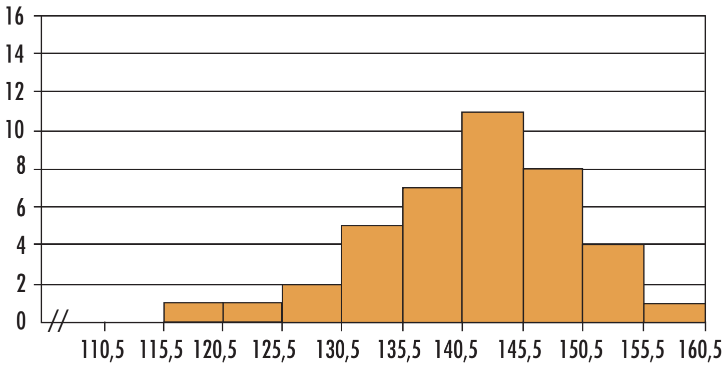


Figura 1.12. Histograma de la variable *T.A. Máxima* con su distribución de frecuencias.

X	n_i
71-75	3
76-80	10
81-85	12
86-90	11
91-95	4
	40

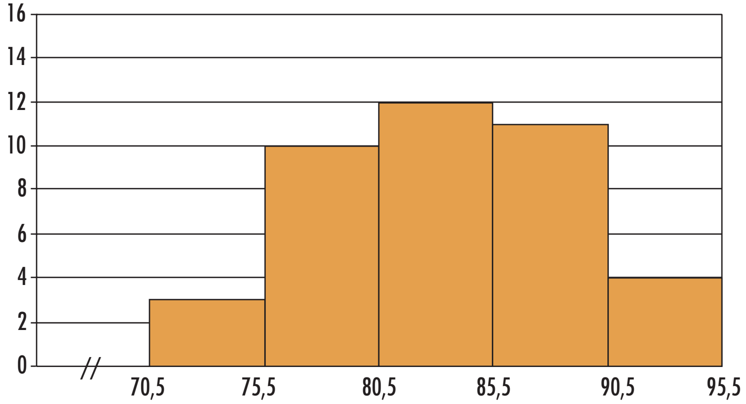


Figura 1.13. Histograma de la variable *T.A. Mínima* con su distribución de frecuencias.

b) Variabilidad

Esta propiedad se refiere al grado de concentración de las observaciones en torno al promedio. Una distribución de frecuencias es ***homogénea*** (tiene poca variabilidad) si los valores de la distribución están cercanos al promedio y es ***heterogénea*** (tiene mucha variabilidad) si los valores se dispersan mucho con respecto al promedio.

Siguiendo con el Ejemplo 1.1, si observamos los datos de la variable *tensión máxima* para el grupo 1 y para el grupo 2 (ver Figuras 1.14 y 1.15), comprobamos que en el primero hay más variabilidad que en el segundo en la muestra estudiada.

X	n_i
111-115	0
116-120	1
121-125	1
126-130	2
131-135	2
136-140	3
141-145	5
146-150	3
151-155	2
156-160	1
	20

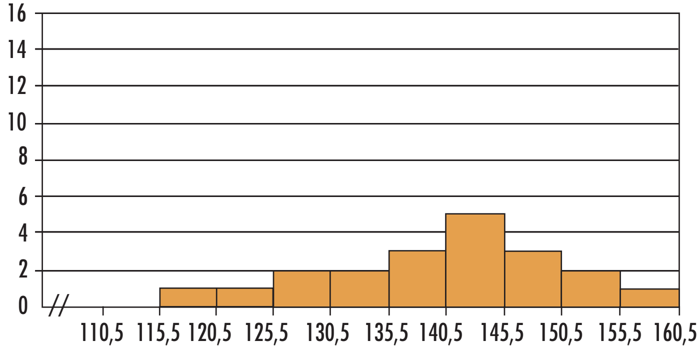


Figura 1.14. Histograma de la variable *T.A. Máxima* en el grupo 1 (pacientes que reciben el tratamiento estándar), con su distribución de frecuencias.

X	n_i
111-115	0
116-120	0
121-125	0
126-130	0
131-135	3
136-140	4
141-145	6
146-150	5
151-155	2
156-160	0
	20

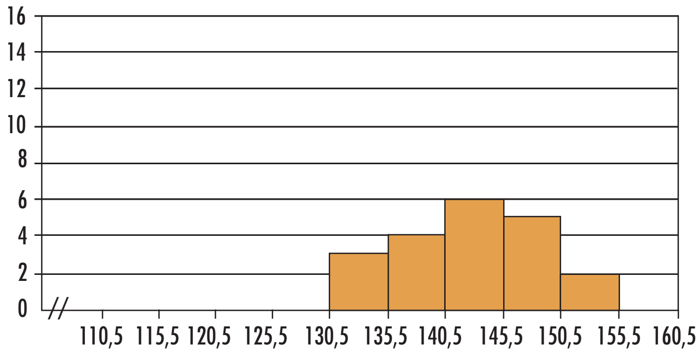


Figura 1.15. Histograma de la variable *T.A. Máxima* en el grupo 2 (pacientes que reciben además una terapia de afrontamiento del estrés) con su distribución de frecuencias.

c) Asimetría o sesgo

Esta propiedad se refiere al grado en que los datos se reparten equilibradamente por encima y por debajo de la tendencia central. Una distribución será simétrica cuando al dividirla en dos a la altura de la media, las dos mitades se superponen. Una distribución tiene ***asimetría positiva*** cuando la mayor concentración de puntuaciones se produce en la parte baja de la escala y ***asimetría negativa*** cuando la mayor parte de las puntuaciones se sitúan en la parte alta de la escala.

Las distribuciones con asimetría negativa son propias de tests fáciles, en los que la mayoría de los sujetos puntúan alto; en los tests difíciles la mayoría de los sujetos puntúan bajo, por lo que la distribución adopta una forma asimétrica positiva.

La distribución de los 40 sujetos en la variable *tensión arterial mínima* del Ejemplo 1.1 presenta un alto grado de simetría. Por el contrario, las distribuciones de esta variable en los grupos 1 y 2 por separado presentan asimetría negativa y positiva respectivamente, como puede observarse en las tablas y gráficas siguientes.

Tal y como se refleja en las Figuras 1.16 y 1.17 los pacientes que siguieron el tratamiento estándar para la hipertensión obtuvieron con mayor frecuencia valores altos de esta variable, mientras que los pacientes tratados además con la terapia de afrontamiento de situaciones estresantes obtuvieron valores más bajos en *tensión arterial mínima,* por lo que éste tratamiento parece resultar más eficaz que el estándar en pacientes con hipertensión arterial.

X	n_i
71-75	1
76-80	1
81-85	6
86-90	9
91-95	3
	20

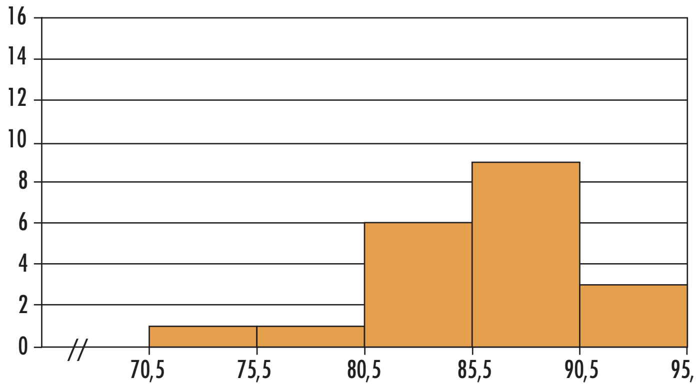


Figura 1.16. Histograma de la variable *T.A. Mínima* en el grupo 1 (pacientes que reciben el tratamiento estándar) con su distribución de frecuencias.

X	n_i
71-75	2
76-80	9
81-85	6
86-90	2
91-95	1
	20

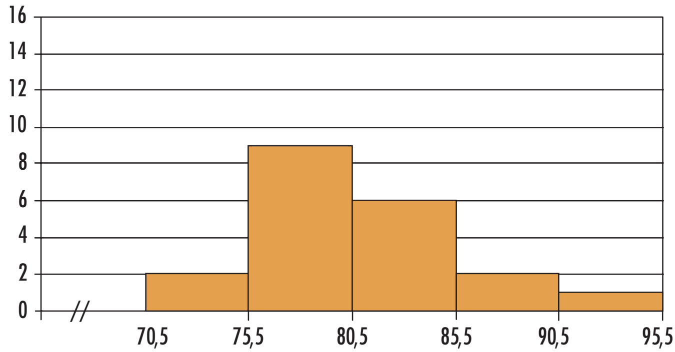


Figura 1.17. Histograma de la variable *T.A. Mínima* en el grupo 2 (pacientes que reciben además una terapia de afrontamiento del estrés) con su distribución de frecuencias.

1.9. RESUMEN

En este capítulo hemos tratado el papel que juega el Análisis de datos dentro del método general de la ciencia y hemos visto algunos conceptos importantes relacionados con la Estadística, además de tratar el problema de la medición y los distintos tipos de escala: nominal, ordinal, de intervalo y de razón. Posteriormente se ha abordado el concepto de variable, y su notación y clasificación de acuerdo a distintos criterios. También se ha tratado la organización y tabulación de los datos, mediante la confección de una distribución de frecuencias. Además, se han presentado algunas formas de representar gráficamente una distribución de frecuencias, de modo que su visión aporte una información de carácter general acerca de cómo se comporta el fenómeno objeto de estudio. Por último, hemos adelantado de manera intuitiva los aspectos más relevantes que se deben analizar en toda distribución de frecuencias: la tendencia central, la variabilidad y la asimetría, que serán objeto de estudio en los próximos temas.

1.10. EJERCICIOS DE AUTOEVALUACIÓN

1.1. El *número de aciertos* en un examen tipo test es una variable: A) nominal; B) ordinal; C) de razón.

1.2. Para poder concluir que un sujeto posee el doble que otro de la característica evaluada, es necesario disponer de una escala de: A) intervalo B) orden C) razón.

1.3. ¿En qué escala de medida el origen no es arbitrario? A) En la escala nominal; B) En la escala de intervalo; C) En la escala de razón.

1.4. ¿Cuál es el nivel de medida de una escala cuyas opciones de respuesta son: 1 = totalmente en desacuerdo, 2 = en desacuerdo, 3 = de acuerdo y 4 = totalmente de acuerdo? A) Nominal; B) Ordinal; C) De intervalo.

1.5. Se han asignado los valores 1, 2 y 3 a pacientes con un problema de claustrofobia muy leve, moderado y alto, respectivamente. ¿Qué nivel de medida tiene la variable grado de claustrofobia? A) Nominal; B) Ordinal; C) De razón.

1.6. Las variables dicotómicas: A) solo admiten dos valores posibles; B) admiten como mínimo dos valores posibles; C) admiten dos o más valores siempre y cuando se trate de una variable nominal.

1.7. El Centro de Investigaciones Sociológicas (CIS) realiza de manera regular una encuesta a los ciudadanos españoles mayores de edad. En una de ellas, preguntó a 1600 ciudadanos sobre el principal problema que existe actualmente en España, encontrando que la mayoría de los encuestados (el 52,5%) opinaron que el paro era el principal problema. ¿Cuál es la población objeto de estudio? A) 1600; B) La población española; C) La población española mayor de edad.

1.8. Continuando con el ejercicio anterior, 52,5% es el valor de: A) un parámetro; B) un estadístico; C) una muestra.

1.9. Se muestra a continuación la distribución de frecuencias de la variable *estado civil* del ejemplo 1.1

X	n_i
Soltero	6
Casado	24
Divorciado	6
Viudo	4
	40

¿Qué tipo de variable es? A) Nominal; B) Ordinal; C) De intervalo.

1.10. Con los datos del ejercicio anterior, ¿cuál es la proporción de pacientes casados? A) 0,6; B) 24; C) 60.

1.11. Con los datos del Ejercicio 1.9, ¿cuál es la frecuencia acumulada de los pacientes divorciados? A) 6; B) 36; C) No tiene sentido su cálculo.

1.12. ¿Qué gráfico sería apropiado utilizar para representar los datos del Ejercicio 1.9? A) Diagrama de barras; B) Histograma; C) Diagrama de dispersión.

1.13. ¿Cuáles son los límites exactos del valor 18,56? A) 18,55 – 18,56; B) 18,555 – 18,565; C) 18,565 – 18,565.

1.14. En un experimento de atención visual focalizada se ha utilizado como variable dependiente el *tiempo de reacción* en milisegundos a un determinado estímulo visual presentado en la pantalla de un ordenador. Los tiempos de reacción obtenidos han sido:

520, 487, 458, 399, 458, 465, 502, 389, 444, 478, 415, 501, 388, 466, 438, 474, 458, 468, 479, 511, 458, 499, 487, 468, 423, 415, 429, 473, 426, 409, 450, 410, 439, 490, 480, 417, 432, 491, 451, 382, 458, 510, 390, 433, 487, 429, 389, 477, 466, 520.

¿Qué nivel de medida tiene la variable tiempo de reacción? A) Ordinal; B) De intervalo; C) De razón.

1.15. La distribución de frecuencias de la variable *tiempo de reacción* del ejercicio anterior es:

A)

X	n_i
381-400	6
401-420	5
421-440	8
441-460	8
461-480	11
481-500	6
501-520	6

B)

X	n_i
400 o menos	6
401-425	6
426-450	9
451-475	13
476-500	10
más de 500	6

C) Cualquiera de las dos anteriores.

1.16. La amplitud de los intervalos de la distribución de frecuencias A del ejercicio anterior es: A) 19; B) 20; C) 25.

1.17. Según los datos del ejercicio 1.15, ¿Qué porcentaje de sujetos tardó 450,5 milisegundos o menos? A) 42%; B) 54%; C) 68%.

1.18. ¿Cuáles son los límites exactos del primer intervalo de la distribución de frecuencias A del ejercicio 1.16? A) 380,5 – 400,5; B) 380 – 401; C) 381,5 – 400,5.

1.19. Atendiendo a la distribución de frecuencias A del Ejercicio 1.15., el punto medio del primer intervalo es: A) 390; B) 390,5; C) 391.

1.20. ¿Qué gráfico representa de manera apropiada los valores de esta variable? A) Diagrama de barras; B) Histograma; C) Diagrama de dispersión.

1.11. SOLUCIONES A LOS EJERCICIOS DE AUTOEVALUACIÓN

1.1. Solución: C.
Es una variable de razón, ya que el cero es absoluto.

1.2. Solución: C.
Es necesario disponer de una escala de razón, que es la única que admite como válida la relación de división.

1.3. Solución: C.
En la escala de razón el origen de la escala no es arbitrario, sino que representa un origen real que corresponde a la ausencia (valor cero) de la característica que se está midiendo.

1.4. Solución: B.
El nivel de medida es ordinal, porque los números asignados a las opciones de respuesta solo nos permiten diferenciarlas y ordenarlas. Si una persona escoge la opción 4, solo podemos afirmar que está más de acuerdo con la cuestión planteada que otra persona que ha escogido la opción 3, pero no podemos saber cuánto más de acuerdo está.

1.5. Solución: B.
El nivel de medida es ordinal, ya que podemos diferenciar entre tres niveles de claustrofobia y ordenarlos en función de su gravedad, pero no podemos precisar la distancia entre un nivel y otro.

1.6. Solución: A.
Una variable dicotómica se define como aquella que solo puede presentar dos categorías o valores.

1.7. Solución: C
A) es el tamaño muestral y B) incluye a toda la población española, cuando en el estudio solo interesan los mayores de edad.

1.8. Solución: B
Es el valor de un estadístico, ya que 52,5 es un porcentaje realizado sobre los 1600 encuestados que forman parte de la muestra.

1.9. Solución: A.
Es nominal, ya que no podemos más que diferenciar entre las categorías existentes.

1.10. Solución: A.
$p_i = n_i/n = 24/40 = 0{,}6$

1.11. Solución: C.
No tiene sentido calcular la frecuencia acumulada, ya que la variable estado civil es nominal.

1.12. Solución: A.
El diagrama de barras es un gráfico apropiado para variables cualitativas, no como el histograma (opción B) que se utiliza para variables cuantitativas en intervalo y el diagrama de dispersión (opción C) que se utiliza en el caso de representar conjuntamente dos variables cuantitativas.

1.13. Solución: B.
Límites exactos = Valor informado $\pm\ 0{,}5 \times I = 18{,}56 \pm 0{,}5 \times 0{,}01 = 18{,}56 \pm 0{,}005 = 18{,}555 - 18{,}565$.

1.14. Solución: C.
De razón, porque el cero representa la ausencia total de la característica medida (del tiempo).

1.15. Solución: C.
Ambas reflejan adecuadamente los datos del ejercicio 1.14, diferenciándose únicamente en las decisiones tomadas respecto al número y amplitud de los intervalos.

1.16. Solución: B.
La amplitud es la diferencia entre el límite exacto superior y el límite exacto inferior, por tanto 400,5 – 380,5 = 20.

1.17. Solución: A.
Hay que calcular el porcentaje acumulado del intervalo 426-450 que se encuentra en la distribución de frecuencias B. Para facilitar este cálculo, añadimos además las columnas correspondientes a las frecuencias acumuladas (absolutas y relativas).

X	n_i	n_a	p_a	P_a
400 o menos	6	6	0,12	12
401-425	6	12	0,24	24
426-450	9	21	0,42	42
451-475	13	34	0,68	68
476-500	10	44	0,88	88
más de 500	6	50	1	100

El 42% de sujetos tardó 450,5 milisegundos o menos.

1.18. Solución: A.
Dado que la unidad de medida es la unidad, basta con restar y sumar 0,5 a los límites aparentes para obtener los límites exactos. Así, 381 – 0,5 = 380,5 y 400 + 0,5 = 400,5.

1.19. Solución: B.
El Punto medio del intervalo es la semisuma de los límites exactos o de los límites aparentes, (381 + 400)/2 = 390,5

1.20. Solución: B.
El histograma representa adecuadamente los valores de esta variable, ya que es cuantitativa. El diagrama de barras (opción A) no se puede utilizar en distribuciones de frecuencias agrupadas en intervalos y el diagrama de dispersión (opción C) se utiliza para representar conjuntamente dos variables.

Tema 2

Medidas de tendencia central y posición

2.1. INTRODUCCIÓN

Como se ha mencionado en el tema anterior, una de las propiedades más importantes a estudiar de una distribución de frecuencias es la tendencia central de las puntuaciones. Esta característica de la distribución se puede resumir en un valor o puntuación que refleje esa tendencia central de la distribución y que represente al conjunto de observaciones. Con el fin de cuantificar esta propiedad, se han desarrollado una serie de medidas o índices de tendencia central que indican sobre qué puntuación se concentran las observaciones.

En este tema se van a presentar los principales índices de tendencia central: la media aritmética, la mediana y la moda. Además de exponer el procedimiento de cálculo de los índices, se discuten las principales ventajas e inconvenientes de cada uno de ellos y se ofrecen criterios para su aplicación.

Posteriormente, se abordan las medidas de posición, las cuales son útiles para informar sobre la posición relativa en la que se encuentra un sujeto con respecto al conjunto al que pertenece, a partir de su puntuación en la variable. Se describen los tres índices de posición más utilizados en la práctica: los percentiles, los cuartiles y los deciles.

Los objetivos de aprendizaje que se persiguen en este tema son los siguientes:

- Conocer las características de las principales medidas de tendencia central (media aritmética, mediana y moda) y de posición (percentiles, cuartiles y deciles).
- Saber aplicar los índices de tendencia central y de posición.
- Seleccionar los índices de tendencia central y de posición adecuados en cada caso.
- Interpretar correctamente los valores obtenidos mediante los índices de tendencia central y de posición.

2.2. MEDIDAS DE TENDENCIA CENTRAL

En el análisis descriptivo de la distribución de frecuencias de una variable, es habitual que el número de observaciones sea grande y que nos planteemos resumir, mediante valores numéricos, las principales propiedades de esa distribución. En lo que respecta a la tendencia central de la distribución, nos interesa calcular un valor central que actúe como resumen numérico para representar al conjunto de datos. Estos valores centrales de la variable son las medidas o índices de tendencia central. Los índices de tendencia central permiten representar toda la distribución de frecuencias con un único valor y, además, facilitan la comparación de diferentes conjuntos de puntuaciones de una variable. Por ejemplo, si medimos el *nivel de autoestima* en una muestra de 200 niños (100 niños y 100 niñas), además de estudiar la tendencia central en niños y niñas de forma conjunta, los índices de tendencia central posibilitan la comparación de niños y niñas en su grado de autoestima. Así, podemos averiguar si el nivel medio de autoestima es mayor en los niños que en las niñas, o viceversa. Trabajando directamente con las 200 observaciones iniciales, no podríamos, de forma eficiente, ni describir la tendencia central de niños y niñas, ni comparar las distribuciones de ambos en su grado de autoestima.

A continuación se van a describir las tres medidas de tendencia central, representativas de la distribución, más utilizadas en el análisis de datos: ***la media aritmética, la mediana y la moda.***

2.2.1. La media aritmética

La media aritmética, también llamada promedio o simplemente media, es la medida de tendencia central más conocida y usada en la práctica debido, básicamente, a la sencillez de su cálculo y a que es el fundamento de un gran número de técnicas estadísticas.

La media aritmética indica la tendencia general de una distribución de frecuencias de una variable y es el valor central alrededor del cual están la mayoría de las observaciones. De hecho, desde una perspectiva geométrica, la media aritmética se puede interpretar como el «centro de gravedad» de la distribución de frecuencias —véase Amón (1999)—. Por otro lado, a diferencia de otros índices de tendencia central, sólo puede calcularse para variables cuantitativas.

La ***media aritmética*** de una variable X, denotada por $\bar{X}$, se define como la suma de todos los valores observados de la variable divididos por el número total de observaciones. Se expresa matemáticamente de la siguiente manera:

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n} = \frac{\sum X_i}{n}$$

donde:

X_i es el valor que toma la variable u observación del sujeto i.

n es el número total de observaciones.

Veamos un ejemplo con pocos datos en el que se obtiene la media aritmética.

Ejemplo 2.1. Se ha medido la variable *aptitud espacial* en 5 alumnos de Enseñanza Secundaria de un centro educativo. Calculemos la media aritmética.

Alumno	Aptitud espacial (X)
1	133
2	120
3	125
4	115
5	122

La media aritmética de estas observaciones es:

$$\bar{X} = \frac{\sum X_i}{5} = \frac{133+120+125+115+122}{5} = \frac{615}{5} = 123$$

Por lo general, el número de observaciones es mucho mayor que en el Ejemplo 2.1. Por ese motivo, es usual que los datos se presenten en tablas de distribución de frecuencias agrupados o no en intervalos. En este caso, la media aritmética se puede calcular a partir de las frecuencias absolutas (n_i) o de las frecuencias relativas o proporciones (p_i):

Cálculo de la media en tablas de distribución de frecuencias con datos agrupados o no en intervalos

Media aritmética a partir de una distribución de frecuencias absolutas:

$$\bar{X} = \frac{\sum n_i X_i}{\sum n_i} = \frac{\sum n_i X_i}{n}$$

donde:

n es el número total de observaciones.

X_i es el valor i en la variable X_i o el punto medio del intervalo.

n_i es la frecuencia absoluta del valor o intervalo i.

Media aritmética a partir de una distribución de frecuencias relativas:

$$\bar{X} = \sum p_i X_i$$

donde:

p_i es la frecuencia relativa o proporción de observaciones del valor o del intervalo i.

Como es de esperar, con una u otra fórmula se obtiene el mismo resultado para la media. Su cálculo se ilustra con los siguientes ejemplos.

Ejemplo 2.2. En la tabla adjunta se muestra la distribución de frecuencias de las puntuaciones en un examen tipo test de los alumnos que han obtenido una nota de cinco o superior. Calcula la media utilizando las frecuencias absolutas y las relativas.

Nota (X_i)	n_i
5	135
6	66
7	45
8	36
9	18
Σ	300

En la tabla nos dan las frecuencias absolutas. Si aplicamos la fórmula de la media para las frecuencias absolutas, obtenemos el siguiente resultado:

$$\bar{X} = \frac{\sum n_i X_i}{n} = \frac{135 \cdot 5 + 66 \cdot 6 + 45 \cdot 7 + 36 \cdot 8 + 18 \cdot 9}{135 + 66 + 45 + 36 + 18} = \frac{1836}{300} = 6{,}12$$

Para aplicar la segunda fórmula se deben obtener las frecuencias relativas de cada puntuación:

Nota (X_i)	n_i	$p_i = n_i/n$
5	135	135/300 = 0,45
6	66	66/300 = 0,22
7	45	45/300 = 0,15
8	36	36/300 = 0,12
9	18	18/300 = 0,06
Σ	300	1

$$\bar{X} = \sum p_i X_i = 0{,}45 \cdot 5 + 0{,}22 \cdot 6 + 0{,}15 \cdot 7 + 0{,}12 \cdot 8 + 0{,}06 \cdot 9 = 6{,}12$$

Se comprueba que con ambas fórmulas se obtiene el mismo valor para la media aritmética.

Ejemplo 2.3. En la tabla se muestran las puntuaciones agrupadas en intervalos de 50 personas en una prueba de inglés (formada por 15 ítems). Calcula la media utilizando las frecuencias absolutas y las relativas.

X	X_i	n_i	p_i
1-3	2	2	0,04
4-6	5	7	0,14
7-9	8	13	0,26
10-12	11	18	0,36
13-15	14	10	0,20
Σ		50	1

Con frecuencias absolutas:

$$\bar{X} = \frac{\sum n_i X_i}{n} = \frac{2 \cdot 2 + 7 \cdot 5 + 13 \cdot 8 + 18 \cdot 11 + 10 \cdot 14}{2 + 7 + 13 + 18 + 10} = \frac{481}{50} = 9{,}62$$

Con frecuencias relativas:

$$\bar{X} = \sum p_i X_i = 0{,}04 \cdot 2 + 0{,}14 \cdot 5 + 0{,}26 \cdot 8 + 0{,}36 \cdot 11 + 0{,}20 \cdot 14 = 9{,}62$$

Como podemos observar, la media aritmética aprovecha toda la información disponible en los datos, ya que para su cálculo es necesario utilizar todas las puntuaciones de los sujetos. Como veremos posteriormente, esto no ocurre con otros índices. Asimismo, la media aritmética presenta una serie de propiedades matemáticas, de las que podemos destacar las siguientes:

1. En una distribución, la suma de las desviaciones de cada valor con respecto a su media es igual a cero. Matemáticamente se expresa como:

$$\sum_{i=1}^{n} (X_i - \bar{X}) = 0$$

Esta propiedad se puede comprobar con los datos del ejemplo 2.1. La media es igual a $\bar{X}$ = 123, y el sumatorio de las desviaciones se obtiene de la siguiente manera:

$$\sum_{i=1}^{n}\left(X_i-\bar{X}\right)=\left(133-123\right)+\left(120-123\right)+\left(125-123\right)+\left(115-123\right)+\left(122-123\right)=$$
$$=10+\left(-3\right)+2+\left(-8\right)+\left(-1\right)=0$$

2. Si a los valores de la variable X les aplicamos la siguiente transformación lineal: $Y_i = bX_i + a$, la media de los nuevos valores Y será $\bar{Y} = b\bar{X} + a$.

En el ejemplo 2.1, si las puntuaciones de la variable X = *aptitud espacial*, se multiplican por 2 (b = 2) y se le suman 10 (a = 10), se obtienen las puntuaciones de Y que sigue siendo la puntuación en aptitud espacial pero en una nueva escala:

Alumno	**Aptitud espacial (*X*)**	**Aptitud espacial (*Y*)**
1	133	$Y_1 = 2 \cdot 133 + 10 = 276$
2	120	$Y_2 = 2 \cdot 120 + 10 = 250$
3	125	$Y_3 = 2 \cdot 125 + 10 = 260$
4	115	$Y_4 = 2 \cdot 115 + 10 = 240$
5	122	$Y_5 = 2 \cdot 122 + 10 = 254$

La media de Y calculada a partir de las puntuaciones es:

$$\bar{Y}=\frac{\sum_{i=1}^{5}Y_i}{5}=\frac{276+250+260+240+254}{5}=\frac{1280}{5}=256$$

Si aplicamos la propiedad de la media, podemos obtener la media de Y directamente con:

$$\bar{Y}=b\bar{X}+a=2\cdot 123+10=256$$

Por último, a la hora de utilizar la media como medida representativa de la tendencia central de la distribución, conviene tener en cuenta las siguientes limitaciones:

a) Cuando los datos están agrupados en intervalos, la media no se puede calcular si el intervalo máximo no tiene límite superior y/o el intervalo mínimo no lo tiene inferior. Por ejemplo, en la siguiente distribución de frecuencias:

X	n_i	X_i
10-14	2	12
15-19	6	17
20-24	12	22
25-29	8	27
$X \geq 30$	5	?

el intervalo máximo no tiene límite superior, por lo que no podemos determinar el punto medio de ese intervalo, necesario para el cálculo de la media aritmética.

b) La media es sensible a la existencia de unas pocas observaciones con valores extremos en la distribución de frecuencias. Esta circunstancia se da en distribuciones marcadamente asimétricas, por lo que no es recomendable la utilización de la media en este tipo de distribuciones debido a que afecta a su representatividad como valor central de la distribución. Estos valores extremos pueden ser: 1) producto de errores en la recogida o grabación de los datos, o 2) valores que aportan información relevante de la variable. En el primer caso, se eliminan estas observaciones y la distribución se vuelve más simétrica, por lo que podría calcularse la media aritmética. En el segundo caso, se recomienda aplicar otros índices de tendencia central menos sensibles a los valores extremos como la mediana, que la abordaremos en el siguiente epígrafe.

2.2.2. La mediana

Tal y como hemos mencionado en el apartado anterior, cuando la distribución es asimétrica una buena alternativa a la media aritmética para resumir la tendencia central de las puntuaciones es el índice denominado mediana. A diferencia de la media, la mediana no se ve afectada por los valores extremos que pueda adoptar la variable debido a que en su cálculo no intervienen todos los valores de la distribución sino únicamente los que ocupan las posiciones centrales. Por tanto, en este caso, la mediana es un valor más apropiado para representar la tendencia central de la distribución. Por otro lado, la mediana se puede obtener en todo tipo de variables, excepto en variables cualitativas.

La ***mediana*** de una variable *X*, representada por *Md*, se define como el valor que divide la distribución de frecuencias en dos partes iguales, conteniendo cada una el 50% de las observaciones.

Supongamos que hemos obtenido la puntuación de n sujetos en una variable de interés. Para el cálculo de la mediana con pocos casos se procede de la siguiente manera:

1. En primer lugar, se ordenan las n puntuaciones de menor a mayor.
2. En segundo lugar se observa si el número de observaciones n es impar o par.
   - Si n es impar, el valor de la mediana es el de la observación que ocupa la posición central, dentro de ese conjunto de observaciones ya ordenadas.
   - Sin embargo, si el número de observaciones n es par, la mediana es la media aritmética de los dos valores centrales de la distribución.

A continuación presentamos un ejemplo de cada caso:

Ejemplo 2.4. Cálculo de la mediana con n impar.

Calculemos la mediana en los datos del Ejemplo 2.1. siguiendo los dos pasos:

1.º Ordenamos las puntuaciones de los alumnos en *aptitud espacial* de menor a mayor valor:

115, 120, 122, 125, 133

2.º Dado que $n = 5$ es un número impar, la mediana es el valor o puntuación que ocupa la posición central en esa secuencia ordenada de observaciones, es decir, $Md = 122$.

Ejemplo 2.5. Cálculo de la mediana con n par.

Las puntuaciones de seis sujetos que han realizado un test de autoestima se presentan en la siguiente tabla:

Sujeto	X_i
1	18
2	16
3	24
4	20
5	28
6	30

En primer lugar, ordenamos las puntuaciones de menor a mayor:

16, 18, 20, 24, 28, 30

En segundo lugar, dado que $n = 6$ es un número par, la mediana es la media aritmética de los dos valores centrales de la distribución:

$$Md = \frac{20+24}{2} = 22$$

Como ocurría con la media aritmética, lo normal es que el número de observaciones no sea tan pequeño, aparezcan valores de observaciones repetidos y, por ello, los datos se presenten en tablas de distribución de frecuencias agrupados o no en intervalos. En el caso más general, cuando los datos están agrupados en intervalos, el intervalo en el que se encuentra la mediana se denomina ***intervalo crítico*** y se corresponde con aquél en el que la frecuencia absoluta acumulada n_a es igual o superior a $n/2$. La mediana se obtiene con la siguiente fórmula:

$$Md = L_i + \left(\frac{\frac{n}{2} - n_d}{n_c} \right) \cdot I$$

donde:

L_i = Límite exacto inferior del intervalo crítico.

n = Número de observaciones.

n_d = Frecuencia absoluta acumulada por debajo del intervalo crítico.

n_c = Frecuencia absoluta del intervalo crítico.

I = Amplitud del intervalo crítico.

Ejemplo 2.6. Calculemos la mediana en el Ejemplo 2.3. en el que los datos están agrupados en intervalos.

X	X_i	n_i
1-3	2	2
4-6	5	7
7-9	8	13
10-12	11	18
13-15	14	10
Σ		50

En este ejemplo los intervalos se presentan en orden creciente. Para facilitar el cálculo de la mediana y de cualquier medida de posición es recomendable invertir el orden de la tabla:

X	X_i	n_i	n_a
13-15	14	10	50
10-12	11	18	40
7-9	8	13	22
4-6	5	7	9
1-3	2	2	2
Σ		50	

Como se puede apreciar, $\frac{n}{2}=\frac{50}{2}=25,$ por lo que el intervalo crítico es el cuarto intervalo [10-12], con n_a = 40. Aplicando la fórmula:

$$Md = L_i + \left(\frac{\frac{n}{2} - n_d}{n_c} \right) \cdot I = 9,5 + \left(\frac{\frac{50}{2} - 22}{18} \right) \cdot 3 = 10$$

La fórmula planteada se basa en el método de interpolación en el que se asume la distribución homogénea de las puntuaciones dentro de cada intervalo. Veamos cómo podemos calcular directamente la mediana con este método utilizando los datos del ejemplo 2.6. Sabemos que el número de observaciones es n = 50 y que, por lo tanto, la mediana es el valor que deja por debajo de sí a 25 sujetos. Hemos identificado el intervalo crítico en [10-12] y el número de puntuaciones acumuladas hasta el límite superior del intervalo anterior al crítico [7-9] es de n_a = 22. Por tanto, faltan 25-22 = 3 observaciones para llegar al 50% en el que se encuentra la mediana (ver figura 2.1).

Si asumimos que las puntuaciones se reparten a lo largo de cada intervalo de forma homogénea, entonces podemos afirmar que las 18 observaciones del intervalo crítico se distribuyen homogéneamente en una amplitud de 3 unidades. Por lo tanto, si 18 observaciones se reparten en una amplitud de 3, ¿qué amplitud o unidades dentro del intervalo crítico ocuparán las 3 observaciones que faltan para llegar al 50%? Por una regla de tres:

$$\left.\begin{array}{r} 18 \text{ observaciones} \rightarrow 3 \text{ unidades} \\ 3 \text{ observaciones} \rightarrow x \text{ unidades} \end{array}\right\} \Rightarrow x = \frac{3 \cdot 3}{18} = 0,5$$

Estas 0,5 unidades debemos sumarlas al límite inferior del intervalo crítico obteniendo el mismo resultado que con la fórmula:

$$Md = 9,5 + 0,50 = 10$$

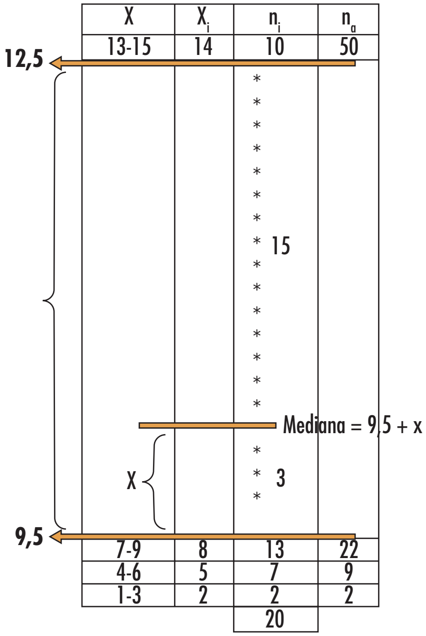


Figura 2.1. Representación del cálculo de la Mediana para los datos del ejemplo 2.6.

Por otra parte, cuando se trata de una distribución de frecuencias pero los datos no están agrupados en intervalos, el cálculo de la mediana es un caso particular de la fórmula anterior en la que la amplitud de los intervalos es igual a uno ($I = 1$). Veamos un ejemplo:

Ejemplo 2.7. Calcúlese la mediana en la distribución de frecuencias del ejemplo 2.2.

En primer lugar se obtienen las frecuencias absolutas acumuladas:

Nota (X_i)	n_i	n_a
9	18	300
8	36	282
7	45	246
6	66	201
5	135	135
Σ	300	

Como se puede apreciar, $\frac{n}{2} = \frac{300}{2} = 150$, por lo que el intervalo crítico es el segundo intervalo unitario [5,5-6,5], con $n_a = 201$. Aplicando la fórmula:

$$Md = L_i + \left(\frac{\frac{n}{2} - n_d}{n_c} \right) \cdot I = 5,5 + \left(\frac{\frac{300}{2} - 135}{66} \right) \cdot 1 = 5,727 \approx 5,73$$

La mediana, por lo general se puede calcular en cualquier distribución de frecuencias excepto cuando los datos están agrupados en intervalos y existe un intervalo abierto en el que se encuentra la mediana. Veamos un ejemplo:

Ejemplo 2.8.

Tabla 2.1

X	n_i	n_a
$X \geq 30$	14	—
25-29	18	76
20-24	29	58
15-19	20	29
10-14	9	9
Σ	90	

Tabla 2.2

X	n_i	n_a
$X \geq 31$	35	—
121-130	9	25
111-120	8	16
100-110	6	8
90-99	2	2
Σ	60	

En la distribución de frecuencias de la izquierda, $\frac{n}{2} = \frac{90}{2} = 45$ por lo que el intervalo crítico es [20-24] con $n_a = 58$. En este caso, como el intervalo crítico no es el intervalo abierto, se puede calcular la mediana que es igual a:

$$Md = L_i + \left(\frac{\frac{n}{2} - n_d}{n_c} \right) \cdot I = 19,5 + \left(\frac{\frac{90}{2} - 29}{29} \right) \cdot 5 = 22,26$$

Sin embargo, en la distribución de la tabla 2.2, $\frac{n}{2} = \frac{60}{2} = 30$, por lo que el intervalo crítico es el intervalo superior que está abierto y, por tanto, no se puede calcular la mediana.

2.2.3. La moda

Un tercer índice de tendencia central que se puede obtener en variables cualitativas, cuasicuantitativas y cuantitativas es la moda.

La ***moda*** de una distribución, que se representa por Mo, se define como el valor o categoría de la variable con mayor frecuencia absoluta.

En el caso de una distribución de una variable cualitativa, la moda es la categoría con la máxima frecuencia.

Ejemplo 2.9. En la tabla adjunta se muestra la distribución de frecuencias del idioma elegido por 200 alumnos en la Escuela oficial de Idiomas.
En esta variable, la categoría con mayor frecuencia absoluta es Inglés, y esa es, por tanto, la moda de esta distribución.

Idioma elegido	n_i
Alemán	35
Francés	52
Inglés	90
Italiano	23

En una distribución de una variable cuantitativa con los datos no agrupados en intervalos, la moda es el valor con la mayor frecuencia absoluta.

Ejemplo 2.10. En la distribución de frecuencias del Ejemplo 2.2, la moda es igual a $Mo = 5$, dado que este valor es el que presenta la frecuencia absoluta máxima.

Nota (X_i)	n_i
5	135
6	66
7	45
8	36
9	18
Σ	300

Finalmente, si se trata de una distribución de una variable cuantitativa con los datos agrupados en intervalos, se localiza el intervalo modal que es el intervalo con la frecuencia máxima y la moda es el punto medio de dicho intervalo.

Ejemplo 2.11. En la distribución de frecuencias del Ejemplo 2.3, el intervalo modal es [10, 12], por lo que la moda es $Mo = 11$.

X	X_i	n_i
1-3	2	2
4-6	5	7
7-9	8	13
10-12	11	18
13-15	14	10
Σ		50

Cuando en una variable existe un único valor con la frecuencia máxima, la distribución presenta una moda y es unimodal. Sin embargo, la distribución de una variable no tiene por qué tener una única moda. De hecho, si son dos los valores con la frecuencia más alta la distribución es bimodal, si son tres los valores sería trimodal, etc. En la figura 2.2, la distribución de la izquierda es unimodal y la moda es el valor X_3, mientras que la de la derecha es bimodal, siendo las dos modas los valores X_2 y X_3. También puede ocurrir que una distribución no tenga moda, lo que se denomina distribución amodal. Esto sucede cuando todos los valores tienen la misma frecuencia absoluta; en este caso no se puede calcular la moda.

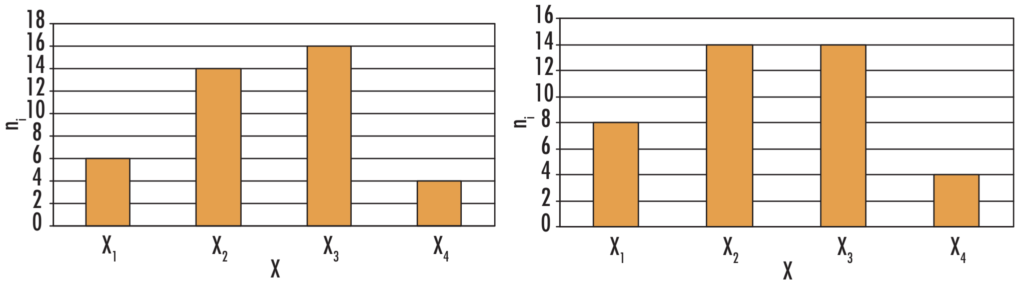


Figura 2.2. Distribución de frecuencias unimodal (izquierda) y bimodal (derecha).

Por último, completando lo dicho hasta aquí, las principales características de la moda son las siguientes:

a) Es un índice de cálculo sencillo y de fácil interpretación

b) De los tres índices de tendencia central estudiados, la moda es el único que, además de aplicarse a variables cuantitativas, se puede calcular en variables cualitativas.

c) Cuando los datos están agrupados en intervalos y existen intervalos abiertos, la moda se puede calcular excepto si el intervalo modal coincide con el intervalo abierto. Si nos fijamos en las tablas 2.1 y 2.2 del ejemplo 2.8, la moda se puede calcular en el primer caso y su valor es Mo = 22, mientras que no es posible calcularla en el segundo caso debido a que el intervalo modal (el intervalo superior) está abierto.

2.2.4. La elección de una medida de tendencia central

Cuando se ha medido una variable en una muestra de n sujetos y deseamos seleccionar un valor que resuma adecuadamente la tendencia central de la distribución de frecuencias, la primera pregunta que nos debemos plantear es: ¿qué medida de tendencia central debemos utilizar? Pues bien, como primera opción se recomienda la media aritmética porque en ella están basadas un gran número de estadísticos y técnicas estadísticas de gran importancia y de uso frecuente que se estudiarán posteriormente. Únicamente se desaconseja su utilización cuando la distribución es muy asimétrica con unos pocos valores extremos que pueden distorsionar la representatividad de la media como tendencia central de la distribución. Por último, la media no se puede aplicar: 1) cuando el nivel de medida de la variable es nominal u ordinal, y/o 2) en datos agrupados en los que existen intervalos abiertos en los extremos de la distribución.

Cuando la media no se puede aplicar, o no es recomendable su utilización, la siguiente opción puede ser la mediana. Como hemos señalado, la mediana es más resistente a los valores extremos que generan asimetría en la distribución, se puede obtener en variables con nivel de medida ordinal, y, además, se puede calcular en distribuciones con datos agrupados en intervalos con intervalos abiertos. Sin embargo, en ocasiones no se puede obtener la mediana. Esto puede ocurrir por dos motivos: 1) el nivel de medida de la variable es nominal y/o 2) con datos agrupados en intervalos, la mediana se encuentra en el intervalo abierto. En esa situación, la única alternativa posible es utilizar la moda. Por otro lado, como ya sabemos, la moda no se puede calcular cuando la distribución sea amodal (no tiene moda) o el intervalo abierto coincide con el intervalo modal.

Hoy en día, con el uso de programas informáticos para el análisis estadístico de los datos, se recomienda, siempre y cuando sea pertinente, el cálculo de los tres índices para el estudio de la tendencia central de la distribución. Cuando las variables son cualitativas únicamente puede utilizarse la moda como medida de tendencia central. Sin embargo, en el caso de variables con nivel de medida ordinal, se pueden obtener tanto la moda como la mediana. Por último, si la variable es cuantitativa se pueden calcular los tres índices de tendencia central, lo que implica que dispondremos de mayor información para estudiar esta propiedad de las distribuciones. Es interesante resaltar que cuando la distribución de una variable cuantitativa es simétrica y unimodal, coinciden los valores de la media, mediana y moda, como podemos apreciar en la figura 2.3.

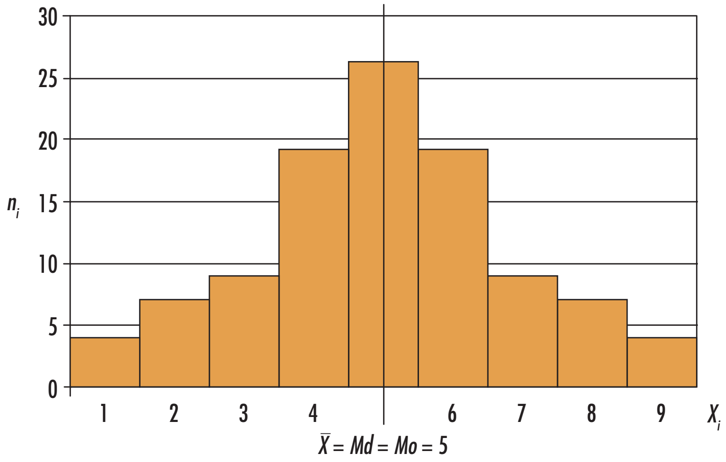


Figura 2.3. Medidas de tendencia central en una distribución de frecuencias simétrica y unimodal.

2.3. MEDIDAS DE POSICIÓN

En la primera parte de este tema hemos definido medidas que representaban al conjunto de datos. Interesaba disponer de un indicador o resumen numérico de la tendencia central de todas las puntuaciones. Pues bien, ahora la cuestión que nos planteamos va dirigida a un sujeto o dato particular y la podríamos formular de la siguiente manera: en una distribución de frecuencias de una variable, un sujeto s obtiene una puntuación X_s , ¿qué posición ocupa este sujeto en la distribución con respecto al resto de sujetos?, ¿qué puntuación tendría que obtener para superar a un porcentaje determinado de sujetos de la distribución? Por ejemplo, en un test de creatividad administrado a los 50 niños de una clase podemos plantearnos las siguientes cuestiones: ¿qué puntuación debe alcanzar un alumno para superar al 50% de sus compañeros?, ¿qué puntuación debe obtener para estar entre el 25% de los más creativos? Imaginemos que un alumno obtiene una puntuación de 15, ¿qué posición le corresponde a $X = 15$ en el conjunto de puntuaciones de los alumnos de la clase?, ¿está entre los más creativos de la clase?, ¿qué porcentaje de sus compañeros están por debajo de él en creatividad o qué porcentaje le superan en dicha variable?

Las ***medidas*** o ***índices de posición***, también denominados ***cuantiles***, que vamos a presentar responden a este tipo de preguntas. Informan acerca de la posición relativa de un sujeto con respecto a su grupo de referencia, dentro de la distribución de frecuencias de la variable. Es decir, indican la situación de una puntuación con respecto a un grupo, utilizando a éste como marco de referencia.

Dado que se trata de localizar la posición de un sujeto en una distribución, para construir un índice de posición, debemos dividir la distribución en un número de partes o secciones iguales entre sí en cuanto al número de observaciones. Por ejemplo, si queremos dividir una distribución en dos partes iguales, necesitamos un único valor para esa partición, que coincide con la mediana de la distribución (recuerda que la mediana divide la distribución en dos partes, cada una con el 50% de los sujetos). En el caso de querer dividirla en tres partes, cada una con un tercio de los sujetos, necesitamos dos valores de la variable, y así sucesivamente. Dependiendo de cuantos valores de la variable utilicemos para dividir la distribución, podemos hablar de diferentes medidas de posición.

A continuación vamos a describir tres medidas de posición o cuantiles: los percentiles, los cuartiles y los deciles. Estos cuantiles se utilizan con mucha frecuencia en la práctica, especialmente los dos primeros.

2.3.1. Percentiles

Los ***percentiles***, también denominados ***centiles***, son los 99 valores de la variable que dividen en 100 partes iguales la distribución de frecuencias.

El ***percentil k*** , denotado por $\boldsymbol{P_k}$, es un valor de la variable de interés que deja por debajo de sí un porcentaje k de sujetos, donde k = 1,2,...99.

Supongamos que en una distribución de frecuencias de la variable *inteligencia espacial*, la puntuación $X = 25$ deja por debajo de sí al 40% de los sujetos de la distribución. Entonces, podemos afirmar que el percentil 40 de esa distribución es $X = 25$, $P_{40} = 25$, y que los sujetos con $X = 25$ están por encima del 40% de los sujetos en *inteligencia espacial* y son superados por el 60% de los sujetos. Otra forma de expresarlo sería que un 40% de los sujetos no superan la puntuación 25 y un 60% sí superan dicha puntuación.

El percentil 50, P_{50}, de una distribución deja por debajo de sí al 50% de los sujetos y por encima al otro 50%. El lector puede percatarse que esa definición coincide con la de mediana de una distribución estudiada previamente. En efecto, el valor de la mediana coincide con el percentil 50 de la distribución. De este modo, la mediana es uno de los 99 posibles percentiles de una distribución, en concreto, el percentil 50. Por este motivo, el cálculo de los percentiles lo vamos a realizar utilizando una extensión del método expuesto para la mediana. La diferencia entre el cálculo de la mediana y de los percentiles, estriba en que, en la mediana se trataba de localizar la posición de $\frac{n}{2}$. En cambio, en los percentiles y de forma más general, se hace en base al número $\frac{n \cdot k}{100}$. Este número es igual a $\frac{n}{2}$ cuando calculamos el percentil 50. En efecto, $k = 50$ por lo que $\frac{n \cdot 50}{100} = \frac{n}{2}$.

Cálculo de los percentiles:

Los datos se presentan en tablas de distribución de frecuencias absolutas, agrupados en intervalos. Pues bien, el intervalo en el que se encuentra el percentil k se denomina intervalo crítico y se corresponde con aquél en el que la frecuencia absoluta acumulada n_a es igual o superior a $\frac{n \cdot k}{100}$. El percentil k se obtiene con la siguiente fórmula:

$$P_k = L_i + \left(\frac{\frac{n \cdot k}{100} - n_d}{n_c} \right) \cdot I$$

donde:

n_d = Frecuencia absoluta acumulada por debajo del intervalo crítico.
n_c = Frecuencia absoluta del intervalo crítico.
L_i = Límite inferior exacto del intervalo crítico.
I = Amplitud del intervalo.

Como ocurría con la mediana, cuando en la distribución de frecuencias los datos no están agrupados en intervalos, se aplica la misma fórmula pero con amplitud del intervalo igual a uno ($I = 1$).

Ejemplo 2.12. Calculemos el percentil 10 en el Ejemplo 2.3. cuyos datos figuran en la siguiente tabla.

X	X_i	n_i	n_a
13-15	14	10	50
10-12	11	18	40
7-9	8	13	22
4-6	5	7	9
1-3	2	2	2
Σ		50	

Como se puede apreciar, $\frac{n \cdot k}{100} = \frac{50 \cdot 10}{100} = 5$, por lo que el intervalo crítico es [4,6] con $n_a = 9$. Aplicando la fórmula:

$$P_{10} = L_i + \left(\frac{\frac{n \cdot 10}{100} - n_d}{n_c} \right) \cdot I = 3{,}5 + \left(\frac{\frac{50 \cdot 10}{100} - 2}{7} \right) \cdot 3 = 4{,}7857 \approx 4{,}79$$

Con el método descrito podemos calcular el valor de cualquiera de los 99 percentiles de una distribución. Sin embargo, puede suceder que tengamos un valor o puntuación de la variable, X_i, y nos interese saber qué percentil ocupa ese valor en la distribución. Es decir, ¿qué percentil le corresponde a la puntuación del sujeto s, X_s? Realmente nos están pidiendo el valor de k, dado el valor de X_i. Para realizar ese cálculo debemos despejar k de la ecuación anterior y obtenemos la siguiente fórmula:

Cálculo de k para X_i:

$$k = \left[\frac{\frac{(P_k - L_i) \cdot n_c}{I} + n_d}{n} \right] \cdot 100$$

Ejemplo 2.13. Con los datos del Ejemplo 2.3, si un sujeto obtiene una puntuación de $X = 11$, ¿qué percentil le corresponde?

X	X_i	n_i	n_a
13-15	14	10	50
10-12	11	18	40
7-9	8	13	22
4-6	5	7	9
1-3	2	2	2
Σ		50	

La puntuación $X = 11$ está en el intervalo [10-12] que va a ser, por tanto, el intervalo crítico. Se aplica la fórmula y se obtiene lo siguiente:

$$k = \left[\frac{\dfrac{(P_k - L_i)\cdot n_c}{I} + n_d}{n}\right]\cdot 100 = \left[\frac{\dfrac{(11-9,5)\cdot 18}{3} + 22}{50}\right]\cdot 100 = 62$$

Por lo tanto, a la puntuación $X = 11$, le corresponde el percentil 62, $P_{62} = 11$.

Cuando se calcula el percentil que corresponde a una puntuación determinada, puede ocurrir que obtengamos un valor con decimales. En este caso, y dado que los percentiles son 99 valores enteros, tomamos la cantidad entera más próxima. Por ejemplo si nos piden el percentil de $X = 9$ en el ejemplo anterior, el resultado es que $P_{39,67} = 9$, con $k = 39,67$. La cantidad entera más próxima es 40, por lo que el percentil es 40, $P_{40} = 9$.

2.3.2. Cuartiles y deciles

Los cuartiles y deciles son dos medidas de posición en las que las secciones o partes en las que se divide la distribución de frecuencias son muchas menos que en los percentiles.

Los ***cuartiles*** son tres valores de la distribución que dividen en cuatro partes de igual frecuencia a la distribución:

El ***primer cuartil***, que se representa por Q_1, deja por debajo de sí al 25% de los sujetos y por encima al 75% restante. Como se puede deducir fácilmente, se corresponde con el percentil 25 de la distribución, esto es, $Q_1 = P_{25}$.

El ***segundo cuartil***, Q_2, deja por debajo de sí al 50% de los sujetos y por encima al otro 50%. Es equivalente al percentil 50, y, por ende, a la mediana de la distribución, $Q_2 = P_{50} = Md$.

Por último, el ***tercer cuartil***, Q_3, deja por debajo de sí al 75% de los sujetos y por encima al 25% restante. Se corresponde con el percentil 75 de la distribución, $Q_3 = P_{75}$.

Debido a la equivalencia con los percentiles, para el cálculo de los tres cuartiles vamos a utilizar los métodos propuestos para los percentiles. En concreto, Q_1 lo calculamos mediante P_{25}, Q_2 con P_{50}, y Q_3 con P_{75}.

Por otra parte, los cuartiles se utilizan para construir índices para el estudio de la variabilidad de una distribución de frecuencias, como se verá en el próximo tema.

Por último, los deciles se definen de la siguiente manera:

Los ***deciles*** son nueve valores que dividen en diez partes iguales a la distribución. Se representan por D_i, donde $i = 1,2,\ldots,9$.

El primer decil, D_1 deja por debajo de sí al 10% de los sujetos, el D_2 al 20%, el D_3 al 30% y así hasta el D_9 que deja por debajo de sí al 90% de los sujetos. De este modo,

$$D_1 = P_{10},\ D_2 = P_{20},\ \ldots D_5 = P_{50} = Md,\ \ldots D_9 = P_{90}$$

Por lo tanto, también podemos calcular los deciles a partir de los percentiles correspondientes. En la figura 2.4 se representa la equivalencia entre los diferentes cuantiles estudiados de una distribución de frecuencias.

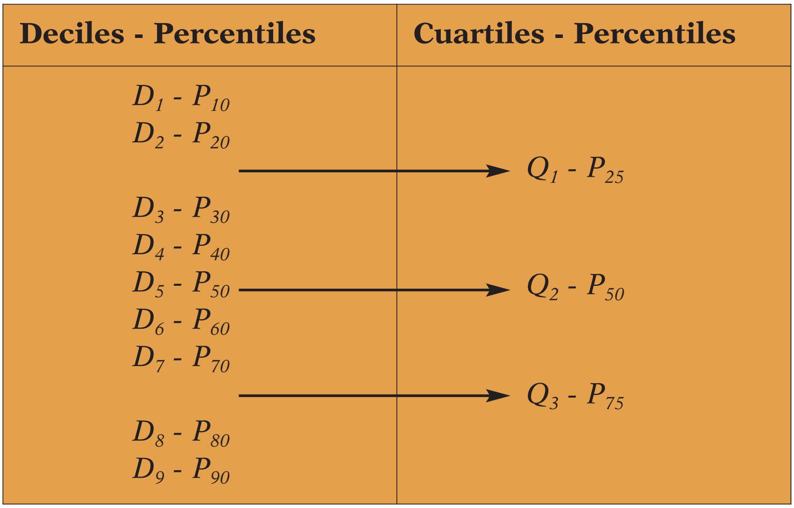


Figura 2.4. Representación de la relación entre medidas de posición.

2.4. RESUMEN

En este tema se ha presentado una de las propiedades o características más relevantes de una distribución de frecuencias como es el estudio de la tendencia central de las puntuaciones. Se han descrito las tres medidas de tendencia central más empleadas, que son la media aritmética, la mediana y la moda. De cada índice se han expuesto sus principales características, los métodos de cálculo y las ventajas y limitaciones de su aplicación al análisis de datos. Esta primera parte del tema se ha concluido con la discusión de una serie de criterios para la elección del índice más adecuado en cada caso.

Posteriormente, se han abordado las medidas de posición, con el fin de estudiar la posición relativa de los sujetos con respecto al conjunto de puntuaciones de la distribución. Dependiendo del número de partes en las que se divida la distribución de frecuencias, se pueden definir diferentes cuantiles. Se han descrito los tres índices de posición más relevantes (percentiles, cuartiles y deciles) y se ha explicado el procedimiento de cálculo según la configuración de los datos y el tipo de cuestión a la que se quiere responder. Asimismo, se ha enfatizado la equivalencia entre los tres tipos de cuantiles, y cómo, una vez definidos los percentiles, se pueden obtener los cuartiles y deciles como casos particulares de los percentiles.

2.5. EJERCICIOS DE AUTOEVALUACIÓN

2.1. La media aritmética no se puede aplicar cuando: A) la variable es continua; B) la distribución es simétrica; C) el intervalo superior está abierto.

2.2. El valor de una variable que siempre divide la distribución de frecuencias en dos partes con el mismo número de observaciones cada una se denomina: A) media aritmética; B) mediana; C) moda.

2.3. Para estudiar la tendencia central en una variable cualitativa con una gran asimetría, el índice adecuado es: A) la media; B) la moda; C) la mediana.

2.4. En una distribución de frecuencias de una variable medida a nivel ordinal, el índice que NO se puede aplicar es: A) la media; B) la moda; C) la mediana.

2.5. En una distribución unimodal se obtienen los mismos valores en los índices moda, media y mediana cuando: A) los datos están agrupados en intervalos; B) la distribución es simétrica; C) el número de observaciones es pequeño.

2.6. En un conjunto de observaciones de una variable, la puntuación que es superada por el 75% de los sujetos se corresponde con el: A) Q_1; B) P_{75}; C) D_2.

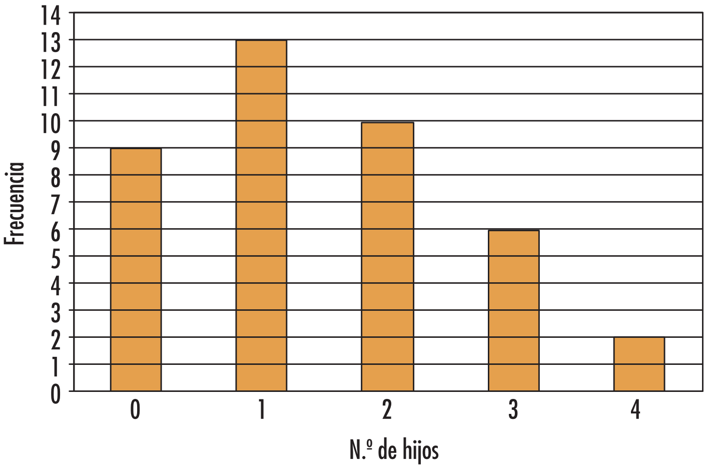
2.7. El quinto decil de una distribución es equivalente al: A) percentil 10; B) percentil 5; C) percentil 50.

2.8. En una distribución de frecuencias, el número de sujetos entre Q_1 y Q_2 es el mismo que entre: A) D_1 y D_2; B) P_{25} y P_{50}; C) Q_1 y Q_3.

2.9. La variable X toma los siguientes valores: 50, 26, 35, 64, 34, 28, 73, 45, 48, 52, 54, 67. La media aritmética es igual a: A) 48; B) 47; C) 49.

2.10. El valor de la mediana en los datos del ejercicio 2.9 es: A) 49; B) 50; C) 51.

2.11. En el siguiente diagrama de barras se representa la variable X: *número de hijos*.



La media del número de hijos es igual a: A) 2,32; B) 1; C) 1,48.

2.12. En los datos del ejercicio anterior, ¿cuál es la moda?: A) 13; B) 1; C) 2.

2.13. Continuando con el ejercicio 2.11, el valor de la mediana es igual a: A) 1,35; B) 1,50; C) 0,75.

2.14. Con los datos del ejercicio 2.11, a la puntuación $X = 2$, ¿qué percentil le corresponde?: A) P_{85} ; B) P_{68} ; C) P_{80}.

2.15. De acuerdo con los datos del ejercicio 2.11, el primer cuartil de la distribución es: A) 0,02; B) 0,50; C) 0,58.

2.16. En la tabla adjunta se muestra la variable *edad* agrupada en intervalos. La moda es: A) 55,5; B) 46; C) 50,5.

X	n_i
66-75	7
56-65	7
46-55	13
36-45	3
26-35	10

2.17. Continuando con la tabla del ejercicio anterior, la edad media de los sujetos es: A) 50,5 ; B) 50 ; C) 52.

2.18. Siguiendo con la tabla del ejercicio 2.16, ¿cuál es el valor mediano de la variable edad?: A) 50,88; B) 52,76; C) 48,24.

2.19. Con los datos del ejercicio 2.16, el percentil 90 es igual a: A) 70,50; B) 69,79; C)65,82.

2.20. De acuerdo a la distribución del ejercicio 2.16, el valor del cuarto decil es: A) 46,50; B) 47,81; C) 52,11.

2.6. SOLUCIONES A LOS EJERCICIOS DE AUTOEVALUACIÓN

2.1. Solución: C
Si el intervalo máximo no tiene límite superior, no podemos determinar el punto medio de ese intervalo, el cual es necesario para el cálculo de la media aritmética (ver apartado 2.2.1).

2.2. Solución: B
Véase la definición de mediana en el apartado 2.2.2.

2.3. Solución: B
Cuando la variable es cualitativa la única medida de tendencia central que se puede utilizar es la moda (ver apartado 2.2.4).

2.4. Solución: A
Cuando la variable está medida a nivel ordinal podemos utilizar la moda y la mediana, pero no la media que requiere que sea de intervalo o de razón (ver apartados 2.2.1 y 2.2.4).

2.5. Solución: B
Tal y como se señala en el apartado 2.2.4, cuando la distribución es unimodal y simétrica, los valores de la media, mediana y moda coinciden.

2.6. Solución: A
La puntuación que es superada por el 75% de los sujetos es aquella que supera al 25%, por lo que se corresponde con el percentil 25 o el primer cuartil, Q_1. (ver apartado 2.3.2)

2.7. Solución: C
Los deciles son nueve valores que dividen en diez partes iguales la distribución. De este modo el Decil 5, D_5, deja por debajo de sí al 50% de las observaciones, por lo que equivale al percentil 50 (ver apartado 2.3.2).

2.8. Solución: B

El número de sujetos entre Q_1 y el Q_2 es igual al 25% de la distribución.

Entre D_1 y D_2 es el 10%

Entre P_{25} y P_{50} es el 25%

Entre Q_1 y Q_2 es el 50%

2.9. Solución: A

$$\bar{X} = \frac{\sum X_i}{12} = \frac{50+26+35+64+34+28+73+45+48+52+54+67}{12} = \frac{576}{12} = 48$$

2.10. Solución: A

Para el cálculo de la mediana, primero se ordenan los datos de menor a mayor:

26, 28, 34, 35, 45, 48, 50, 52, 54, 64, 67, 73

dado que n = 12 es par, la mediana es la media aritmética de los dos valores centrales de la distribución:

$$Md = \frac{48+50}{2} = 49$$

2.11. Solución: C

X_i	n_i	n_a	n_iX_i
0	9	9	0
1	13	22	13
2	10	32	20
3	6	38	18
4	2	40	8
Σ	40		59

$$\bar{X} = \frac{\sum n_i X_i}{n} = \frac{59}{40} = 1{,}475 \approx 1{,}48$$

2.12. Solución: B

El valor de X_i con la frecuencia absoluta mayor es X_i = 1, por lo que Mo = 1.

2.13. Solución: A

$$\frac{n}{2} = \frac{40}{2} = 20,$$

por lo que el intervalo crítico es [0,5-1,5] con n_a = 22. Aplicando la fórmula:

$$Md = L_i + \left(\frac{\frac{n}{2} - n_d}{n_c} \right) \cdot I = 0,5 + \left(\frac{\frac{40}{2} - 9}{13} \right) \cdot 1 = 1,346 \approx 1,35$$

2.14. Solución: B

La puntuación $X = 2$ está en el intervalo unitario [1,5-2,5]

$$k = \left[\frac{\frac{(P_k - L_i) \cdot n_c}{I} + n_d}{n} \right] \cdot 100 = \left[\frac{\frac{(2 - 1,5) \cdot 10}{1} + 22}{40} \right] \cdot 100 = 67,5$$

Por lo tanto, a la puntuación $X = 2$, le corresponde el percentil 68, $P_{68} = 2$.

2.15. Solución: C

$$Q_1 = P_{25}, \; \frac{n \cdot 25}{100} = \frac{40 \cdot 25}{100} = 10,$$

por lo que el intervalo crítico es [0,5-1,5] con n_a = 22.

$$P_{25} = L_i + \left(\frac{\frac{n \cdot k}{100} - n_d}{n_c} \right) \cdot I = 0,5 + \left(\frac{\frac{40 \cdot 25}{100} - 9}{13} \right) \cdot 1 = 0,5769 \approx 0,58$$

2.16. Solución: C

El tercer intervalo es el intervalo modal (n_i = 13), y su punto medio es 50,5. Por lo tanto, Mo = 50,5.

2.17. Solución: B

X	X_i	n_i	$X_i n_i$
66-75	70,5	7	493,5
56-65	60,5	7	423,5
46-55	50,5	13	656,5
36-45	40,5	3	121,5
26-35	30,5	10	305
Σ		40	2000

$$\bar{X} = \frac{\sum n_i X_i}{n} = \frac{2000}{40} = 50$$

2.18. Solución: A

$\frac{n}{2} = \frac{40}{2} = 20$, por lo que el intervalo crítico es [46-55] con n_a = 26. Aplicando la fórmula:

$$Md = L_i + \left(\frac{\frac{n}{2} - n_d}{n_c} \right) \cdot I = 45,5 + \left(\frac{\frac{40}{2} - 13}{13} \right) \cdot 10 = 50,88$$

2.19. Solución: B

$\frac{n \cdot k}{100} = \frac{40 \cdot 90}{100} = 36$, por lo que el intervalo crítico es [66-75] con n_a = 40. Aplicando la fórmula:

$$P_{90} = L_i + \left(\frac{\frac{n \cdot k}{100} - n_d}{n_c} \right) \cdot I = 65,5 + \left(\frac{\frac{40 \cdot 90}{100} - 33}{7} \right) \cdot 10 = 69,79$$

2.20. Solución: B

$$D_4 = P_{40}, \quad \frac{n \cdot k}{100} = \frac{40 \cdot 40}{100} = 16,$$

por lo que el intervalo crítico es [46-55] con $n_a = 26$. Aplicando la fórmula:

$$P_{40} = L_i + \left(\frac{\frac{n \cdot k}{100} - n_d}{n_c} \right) \cdot I = 45,5 + \left(\frac{\frac{40 \cdot 40}{100} - 13}{13} \right) \cdot 10 = 47,81$$

Tema 3

Medidas de variabilidad y asimetría

3.1. INTRODUCCIÓN

En este tema se van a abordar dos nuevas propiedades de una distribución de puntuaciones: la variabilidad o dispersión y la asimetría o sesgo de la distribución.

La segunda propiedad de una distribución de frecuencias, y de la misma importancia que la tendencia central estudiada en el tema anterior, es la variabilidad o dispersión de los datos. La variabilidad hace referencia al grado en que las puntuaciones se asemejan o diferencian entre sí, o se aproximan o alejan de una medida de tendencia central como la media arimética. Se han propuesto numerosos índices para medir la variabilidad de una distribución. En este tema se describen los índices de dispersión más habituales en la práctica como son la amplitud total, la varianza, la desviación típica y la amplitud semi-intercuartil. Además, se presenta un índice, el coeficiente de variación, que resulta útil para comparar distintas distribuciones de frecuencias en términos de su variabilidad.

Posteriormente, se estudia un tercer aspecto de la distribución de frecuencias relacionado con su forma que es la asimetría o sesgo. Como se ha visto en el primer tema, mediante la representación gráfica se puede analizar si una distribución es más o menos simétrica o qué tipo de asimetría la caracteriza. En este tema se describe el índice de asimetría de Pearson que ofrece un resultado numérico sobre el grado y tipo de asimetría de la distribución.

Por último, con el fin de poder comparar a los sujetos entre sí y en diferentes variables, se describen dos puntuaciones que se derivan de las puntuaciones directas: las puntuaciones diferenciales y las típicas. Se presentan sus principales propiedades y la información que proporcionan ambos tipos de puntuaciones.

Los objetivos de aprendizaje que se persiguen en este tema son los siguientes:

- Conocer las características de los principales índices para medir la variabilidad en una distribución de frecuencias, con especial énfasis en la varianza y la desviación típica.
- Saber aplicar los índices de variabilidad o dispersión a una determinada distribución.
- Conocer y saber aplicar el índice de Pearson para analizar el grado y el tipo de asimetría de una distribución.
- Distinguir entre los distintos tipos de puntuaciones: directas, diferenciales y típicas, la información que proporcionan y sus propiedades fundamentales.

3.2. MEDIDAS DE VARIABILIDAD

En el tema anterior vimos que uno de los aspectos más relevantes a la hora de caracterizar una distribución de frecuencias es la tendencia central de los datos y se presentaron las tres principales medidas que resumen numéricamente esta característica. Sin embargo, el estudio de una distribución resultaría incompleto sin el análisis de una segunda propiedad tan importante como la tendencia central; esto es, la variabilidad de los datos. La ***variabilidad*** o ***dispersión*** hace referencia al grado de variación que hay en un conjunto de puntuaciones. Por ejemplo, en la figura 3.1 se muestra la representación gráfica de dos distribuciones que presentan la misma media aritmética pero que difieren en la variabilidad de sus puntuaciones.

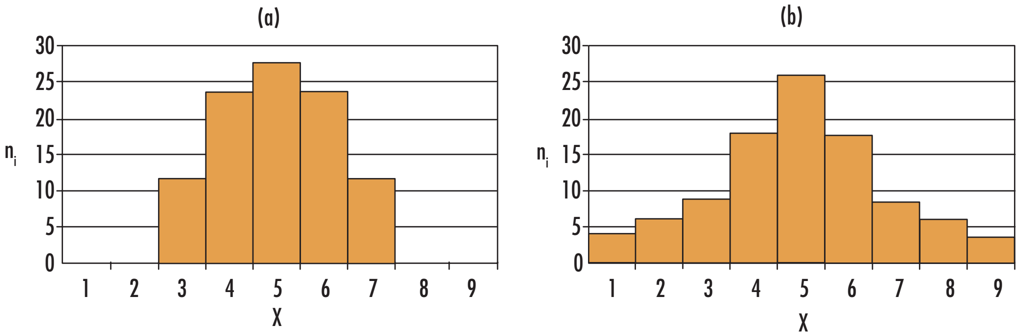


Figura 3.1. Representación gráfica de dos distribuciones; a)menos dispersión. b) más dispersión.

En la figura 3.1a) las puntuaciones están muy próximas entre sí y concentradas en torno al valor promedio por lo que parece que existe poca dispersión en los datos. En la figura 3.1b), las puntuaciones están más alejadas entre sí y no están tan concentradas alrededor de la media, existiendo mayor variabilidad. De este modo, cuanto menor es la variabilidad en una distribución, más homogénea es la muestra de sujetos en la variable que estamos midiendo. En el caso extremo y poco habitual de máxima homogeneidad, todos los valores de la variable serían iguales entre sí y a la media y no habría variabilidad en los datos. Por otro lado, cuando existe más o menos dispersión en los datos, la muestra es más o menos heterogénea y las puntuaciones difieren entre sí.

Con el fin de cuantificar la dispersión presente en los datos, se han definido numerosas medidas o índices de variabilidad. Dos tipos de índices se pueden distinguir: aquellos que miden el grado en el que las puntuaciones se asemejan o diferencian entre sí, y aquellos otros en los que la dispersión se mide con respecto a alguna medida de tendencia central como la media aritmética. En este tema se van a estudiar dos índices del primer tipo: la amplitud total o rango y la amplitud semi-intercuartil. Del segundo tipo, y de gran importancia en la estadística, se van a describir la varianza y la desviación típica.

Tanto unos como otros son útiles para el estudio de la variabilidad de una distribución de frecuencias, pero resultan poco adecuados cuando se trata de comparar la dispersión de dos o más distribuciones. Para realizar dicho análisis, un índice apropiado y que se presenta en este tema es el coeficiente de variación que se basa en la relación entre la desviación típica y la media de cada distribución de frecuencias.

3.2.1. Amplitud total o rango

Una primera aproximación a la dispersión de los datos es el índice de amplitud total, también denominado rango o recorrido, de las observaciones.

La ***amplitud total***, denotada como A_T, de un conjunto de puntuaciones es la distancia que hay en la escala numérica entre los valores que representan la puntuación máxima y la puntuación mínima. Es decir,

$$A_T = X_{\text{máx}} - X_{\text{mín}}$$

Cuando asumimos que trabajamos con variables continuas, la puntuación máxima es el límite exacto superior del intervalo máximo y la puntuación mínima es el límite exacto inferior del intervalo mínimo (véase el apartado 1.6 del primer tema).

Veamos un ejemplo.

Ejemplo 3.1. La siguiente distribución de frecuencias corresponde a las notas de los alumnos del ejemplo 2.2 del tema anterior.

Nota (X_i)	n_i
5	135
6	66
7	45
8	36
9	18
Σ	300

Calculemos la amplitud total o rango de la distribución.

Esta variable se asume que es continua con amplitud de intervalo igual a uno. Por lo tanto, la puntuación máxima es $X_{máx} = 9{,}5$ y la mínima es $X_{min} = 4{,}5$. La amplitud total es igual a $A_T = X_{máx} - X_{mín} = 9{,}5 - 4{,}5 = 5$.

Como se puede apreciar, este índice es muy sencillo de calcular y utiliza muy poca información del conjunto de puntuaciones, ya que se trata sólo de la diferencia entre el mayor valor (X_{max}) y el menor valor (X_{min}) de la variable. Por otro lado, y como consecuencia de lo anterior, su principal inconveniente es que es sensible únicamente a los valores extremos de la distribución. De esta manera, este índice no captura la poca o mucha dispersión que pueda existir entre los restantes valores, que son la gran mayoría de las puntuaciones. Aún así, en el análisis de datos se recomienda incluir el valor de la amplitud total como información complementaria de otras medidas de dispersión más relevantes como la varianza y desviación típica, que se estudiarán a continuación.

3.2.2. Varianza y desviación típica

La medida de variabilidad también se puede basar en la distancia observada entre las puntuaciones y un valor central de la distribución como la media aritmética. De este modo, una distribución con poca variabilidad es aquella en la que la mayoría de las puntuaciones están muy próximas a la media, mientras que en una distribución con mucha variabilidad, las puntuaciones están alejadas o muy alejadas del valor medio de la variable.

Un primer índice que nos podemos plantear de forma lógica es el promedio de las desviaciones o diferencias de cada puntuación con respecto a su media.

$$\bar{X}_d = \frac{\sum d_i}{n} = \frac{\sum (X_i - \bar{X})}{n}$$

El problema de este índice es que, según vimos en la primera propiedad matemática de la media en el tema anterior, el sumatorio del numerador $\sum (X_i - \bar{X})$, siempre es igual a cero, por lo que no sería una buena medida de variabilidad.

Con el fin de poder utilizar un índice con estas desviaciones evitando que sea igual a cero, se han propuesto dos soluciones. La primera consiste en calcular el valor absoluto de cada desviación antes de realizar la suma, obteniendo un índice denominado ***desviación media*** cuya expresión matemática es la siguiente:

$$DM = \frac{|X_1 - \bar{X}| + |X_2 - \bar{X}| + \quad + |X_n - \bar{X}|}{n} = \frac{\sum |X_i - \bar{X}|}{n}$$

La desviación media se emplea muy poco en la actualidad, debido a que es poco manejable matemáticamente por el uso del valor absoluto, lo que ha llevado a que apenas existan técnicas estadísticas basadas en este índice.

Una segunda alternativa al problema del signo de las desviaciones consiste en basarnos en el cuadrado de las diferencias y así obtenemos la varianza que se define de la siguiente manera:

La ***varianza*** de un conjunto de n puntuaciones en una variable X, denotada por S_X^2, se define como el promedio de los cuadrados de las desviaciones de las puntuaciones con respecto a la media. Matemáticamente se expresa como:

$$S_X^2 = \frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \quad + (X_n - \bar{X})^2}{n} = \frac{\sum \left(X_i - \bar{X}\right)^2}{n}$$

Otra forma alternativa de calcular la varianza que se deriva de la fórmula anterior y que simplifica los cálculos es la siguiente:

$$S_X^2 = \frac{\sum X_i^2}{n} - \bar{X}^2$$

Es importante darse cuenta de que, para el cálculo de la varianza, primero se elevan al cuadrado las diferencias y después se obtiene el promedio de esas desviaciones al cuadrado.

Ejemplo 3.2. En la tabla adjunta figuran las puntuaciones de 5 alumnos en la variable *aptitud espacial* del ejemplo 2.1 del tema anterior. La media que se obtuvo fue de $\bar{X}$ = 123. Calcúlese la varianza de las puntuaciones con las dos fórmulas propuestas.

Alumno	**Aptitud espacial (X_i)**	**(X_i – $\bar{X}$)**	**(X_i – $\bar{X}$)²**	**X_i^2**
1	133	10	100	17689
2	120	–3	9	14400
3	125	2	4	15625
4	115	–8	64	13225
5	122	–1	1	14884
Σ			178	75823

Según la primera fórmula:

$$S_X^2 = \frac{\sum (X_i - 123)^2}{5} = \frac{178}{5} = 35,6$$

Según la segunda fórmula:

$$S_X^2 = \frac{\sum X_i^2}{5} - 123^2 = \frac{75823}{5} - 15129 = 15164,6 - 15129 = 35,6$$

Por otra parte, cuando los datos se presentan en tablas de distribución de frecuencias agrupados o sin agrupar en intervalos, la varianza se puede obtener utilizando las dos expresiones equivalentes siguientes:

Cálculo de la varianza en tablas de distribución de frecuencias con datos agrupados o no en intervalos

Varianza a partir de una distribución de frecuencias absolutas:

$$S_X^2 = \frac{\sum n_i (X_i - \bar{X})^2}{\sum n_i} = \frac{\sum n_i (X_i - \bar{X})^2}{n}$$

ó

$$S_X^2 = \frac{\sum n_i X_i^2}{\sum n_i} - \bar{X}^2 = \frac{\sum n_i X_i^2}{n} - \bar{X}^2$$

donde:

n es el número total de observaciones.

X_i es el valor i en la variable X ó el punto medio del intervalo.

n_i es la frecuencia absoluta del valor o intervalo i.

Varianza a partir de una distribución de frecuencias relativas:

$$S_X^2 = \sum p_i X_i^2 - \bar{X}^2$$

donde:

p_i es la frecuencia relativa o proporción de observaciones del valor o del intervalo i.

Ejemplo 3.3. Calcúlese la varianza de la distribución de frecuencias del ejemplo 3.1, sabiendo que la media aritmética es igual a 6,12.

Nota (X_i)	n_i	p_i	$(X_i - \bar{X})$	$(X_i - \bar{X})^2$	$n_i(X_i - \bar{X})^2$	X_i^2	$n_i X_i^2$
5	135	0,45	–1,12	1,2544	169,344	25	3375
6	66	0,22	–0,12	0,0144	0,9504	36	2376
7	45	0,15	0,88	0,7744	34,848	49	2205
8	36	0,12	1,88	3,5344	127,2384	64	2304
9	18	0,06	2,88	8,2944	149,2992	81	1458
Σ	300	1			481,68		11718

Aplicando la primera fórmula:

$$S_X^2 = \frac{\sum n_i (X_i - 6{,}12)^2}{300} = \frac{481{,}68}{300} = 1{,}6056 \quad 1{,}61$$

Aplicando la segunda fórmula:

$$S_X^2 = \frac{\sum n_i X_i^2}{300} - (6{,}12)^2 = \frac{11718}{300} - 37{,}4544 = 39{,}06 - 37{,}4544 = 1{,}6056 \approx 1{,}61$$

Aplicando la tercera fórmula:

$$S_X^2 = \sum p_i X_i^2 - \bar{X}^2 = (0{,}45 \cdot 25 + 0{,}22 \cdot 36 + 0{,}15 \cdot 49 + 0{,}12 \cdot 64 + 0{,}06 \cdot 81) - (6{,}12)^2 =$$
$$= 39{,}06 - 37{,}4544 = 1{,}6056 \approx 1{,}61$$

Ejemplo 3.4. En la siguiente tabla se muestran las puntuaciones agrupadas en intervalos de la prueba de inglés del ejemplo 2.3 del tema anterior. La media calculada para esta distribución es igual a 9,62. Calcúlese la varianza de las puntuaciones con ambas fórmulas.

X	X_i	n_i	p_i	$(X_i - \bar{X})$	$(X_i - \bar{X})^2$	$n_i(X_i - \bar{X})^2$	X_i^2	$n_i X_i^2$
1-3	2	2	0,04	–7,62	58,0644	116,1288	4	8
4-6	5	7	0,14	–4,62	21,3444	149,4108	25	175
7-9	8	13	0,26	–1,62	2,6244	34,1172	64	832
10-12	11	18	0,36	1,38	1,9044	34,2792	121	2178
13-15	14	10	0,20	4,38	19,1844	191,844	196	1960
Σ		50	1,00			525,78		5153

Según la fórmula 1:

$$S_X^2 = \frac{\sum n_i (X_i - 9{,}62)^2}{50} = \frac{525{,}78}{50} = 10{,}5156 \approx 10{,}52$$

Según la fórmula 2:

$$S_X^2 = \frac{\sum n_i X_i^2}{50} - (9{,}62)^2 = \frac{5153}{50} - 92{,}5444 = 103{,}06 - 92{,}5444 = 10{,}5156 \approx 10{,}52$$

Aplicando la tercera fórmula:

$$S_X^2 = \sum p_i X_i^2 - \bar{X}^2 = (0{,}04 \cdot 4 + 0{,}14 \cdot 25 + 0{,}26 \cdot 64 + 0{,}36 \cdot 121 + 0{,}20 \cdot 196) - (9{,}62)^2 =$$
$$= 103{,}06 - 92{,}5444 = 10{,}5156 \approx 10{,}52$$

Como se puede observar, la varianza, al basarse en diferencias al cuadrado, es un número positivo que se expresa en las unidades de la variable al cuadrado. Por ejemplo, supongamos que la variable X se mide en metros. En este caso, las desviaciones de las puntuaciones con respecto a la media $(X_i - \bar{X})$, también vendrán expresadas en metros, mientras que al elevarlas al cuadrado, $(X_i - \bar{X})^2$, las unidades se elevan al cuadrado. Por lo tanto, la varianza viene expresada en las mismas unidades que la variable pero al cuadrado, en este ejemplo, en metros cuadrados. Con el fin de lograr una medida de dispersión en las mismas unidades que la variable y que sea más fácilmente interpretable, se calcula la raíz cuadrada de la varianza y se obtiene un índice que se denomina desviación típica.

La ***desviación típica*** de un conjunto de n puntuaciones, que se representa por S_X, es la raíz cuadrada de la varianza y la fórmula para calcularla es:

$$S_X = \sqrt{\frac{\sum (X_i - \bar{X})^2}{n}}$$

Tanto la varianza como la desviación típica son índices de dispersión muy útiles en el desarrollo posterior de la estadística inferencial estando en la base de numerosas técnicas estadísticas. Por lo general, a la hora de cuantificar la variabilidad de los datos, la desviación típica se suele utilizar más que la varianza debido a que se expresa en las mismas unidades de medida que la variable objeto de estudio. Asimismo, ambos índices presentan una serie de propiedades de las que pueden destacarse las siguientes:

1. El cálculo de la varianza y la desviación típica, a diferencia de otros índices de dispersión, requieren el uso de todas las puntuaciones observadas en la distribución.
2. La varianza y la desviación típica miden la variabilidad de los datos con respecto a la media aritmética, por lo que únicamente deben aplicarse si estamos utilizando la media como medida de tendencia central.
3. La varianza y la desviación típica siempre son no negativas, es decir, pueden ser iguales o mayores que cero. Son iguales a cero únicamente si todas las puntuaciones son iguales entre sí. En este caso, no habría variabilidad o dispersión en los datos. En el resto de los casos la varianza y la desviación típica son positivas, siendo sus valores mayores a medida que aumenta la variabilidad de las puntuaciones.
4. Si a las puntuaciones de la variable X les aplicamos la siguiente transformación lineal: $Y_i = bX_i + a$, la varianza de las nuevas puntuaciones y

será $S_Y^2 = b^2 S_X^2$ y la desviación típica igual a $S_Y = |b| S_X$. Es decir, si a una variable X se le suma o resta una constante a, la varianza y desviación típica de la variable original no se ven afectadas y siguen siendo las mismas. En cambio, cuando multiplicamos los valores de X por una constante b, la varianza queda multiplicada por la constante al cuadrado y la desviación típica por el valor absoluto de dicha constante.

Por último, otro índice de variabilidad relacionado con la varianza y que se aplicará en inferencia estadística es la ***cuasivarianza*** que se define como:

$$S_{n-1}^2 = \frac{\sum \left(X_i - \bar{X} \right)^2}{n-1}$$

donde se divide por $n - 1$, en lugar de n como en la varianza. De forma análoga, la cuasidesviación típica se define como la raíz cuadrada de la cuasivarianza.

$$S_{n-1} = \sqrt{S_{n-1}^2} = \sqrt{\frac{\sum \left(X_i - \bar{X} \right)^2}{n-1}}$$

3.2.3. Coeficiente de variación

Es frecuente que uno de los objetivos del análisis descriptivo de los datos sea la comparación del grado de variabilidad o dispersión entre dos conjuntos de puntuaciones en una misma o distintas variables. Debido a que, por lo general, las variables objeto de estudio se miden en unidades distintas no tiene sentido compararlas en base a los valores de sus varianzas o desviaciones típicas. Para paliar este inconveniente es necesario definir un índice de variabilidad relativa que no dependa de las unidades de medida. Un coeficiente que cumple con estos requisitos es el coeficiente de variación, que se expresa en porcentajes y se define como:

Coeficiente de Variación

$$CV = \frac{S_X}{\bar{X}} \cdot 100$$

El coeficiente de variación está definido para variables con $\bar{X} > 0$ y es recomendable que su resultado se acompañe de la media y desviación típica de la distribución a partir de las cuales ha sido calculado.

Es importante resaltar que cuando comparamos dos conjuntos de puntuaciones obtenidas de la misma variable, también es necesario el coeficiente de variación para comparar la dispersión de ambas distribuciones. Únicamente es posible utilizar la desviación típica cuando la media de ambos grupos es la misma, y, en ese caso, llegaríamos a las mismas conclusiones con ambos índices.

Veamos un ejemplo:

Ejemplo 3.5. Se desea saber si la distribución de frecuencias de las notas de los alumnos del ejemplo 3.1 presenta un mayor o menor grado de dispersión en comparación con las puntuaciones de una segunda clase de alumnos en un test de inteligencia general en el que han obtenido una media de 102 y una varianza de 16.

Las notas de los alumnos del ejemplo 3.1 presentan una media de 6,12 y una varianza de 1,61. Por tanto, la desviación típica es $S_X = \sqrt{1,61} = 1,27$ y el coeficiente de variación es igual a:

$$CV_1 = \frac{S_X}{\bar{X}} \cdot 100 = \frac{1,27}{6,12} \cdot 100 = 20,75$$

Los alumnos de la segunda clase, con una media de 102 y una desviación típica de $S_X = \sqrt{16} = 4$ obtienen un coeficiente de variación igual a:

$$CV_2 = \frac{S_X}{\bar{X}} \cdot 100 = \frac{4}{102} \cdot 100 = 3,92$$

El porcentaje de variación de la primera clase en las notas es del 20,75%, mientras que el de la segunda clase en inteligencia general es del 3,92%. Por lo tanto, dado que el coeficiente es mayor en el primer grupo podemos concluir que el grado de dispersión de los datos es mayor en el primer grupo, siendo el segundo grupo más homogéneo entre sí en las puntuaciones en inteligencia. Dicho de otra forma, los alumnos de la primera clase difieren más entre sí en las notas que los de la segunda clase en el test de inteligencia. Es interesante observar que si hubiésemos utilizado las desviaciones típicas o las varianzas, que son superiores en el segundo grupo, hubiésemos concluido erróneamente que la variabilidad es mayor en el segundo grupo.

3.2.4. Amplitud semi-intercuartil

La varianza y la desviación típica, junto con la media aritmética, son los estadísticos recomendados para estudiar la variabilidad y la tendencia central de una distribución de frecuencias. Sin embargo, como se ha mencionado previamente, en ocasiones, y debido a la asimetría de la distribución, no es aconsejable el uso de estos índices y debemos buscar una alternativa. En estas circunstancias, un índice resistente de dispersión adecuado, que se utilizaría junto con la mediana como medida de tendencia central, sería la amplitud semi-intercuartil.

La ***amplitud semi-intercuartil***, Q, o ***rango semi-intercuartil*** es la distancia media entre el tercer y el primer cuartil. Es decir,

$$Q = \frac{Q_3 - Q_1}{2} = \frac{P_{75} - P_{25}}{2}$$

Como se puede observar, este índice no informa de la variabilidad del conjunto de puntuaciones, sino del 50% de las mismas comprendidas entre el percentil 25 y el 75 de la distribución.

Veamos un ejemplo:

Ejemplo 3.6. Con los datos del ejemplo 3.4, calcúlese la amplitud semi-intercuartil de la distribución.

X	X_i	n_i	n_a
13-15	14	10	50
10-12	11	18	40
7-9	8	13	22
4-6	5	7	9
1-3	2	2	2
Σ		50	

Cálculo del percentil 75:

$\frac{n \cdot k}{100} = \frac{50 \cdot 75}{100} = 37,5,$ por lo que el intervalo crítico es [10,12] con $n_a = 40$.
Aplicando la fórmula:

$$P_{75} = L_i + \left(\frac{\frac{n \cdot 75}{100} - n_d}{n_c} \right) \cdot I = 9,5 + \left(\frac{\frac{50 \cdot 75}{100} - 22}{18} \right) \cdot 3 = 12,08$$

Cálculo del percentil 25:

$\frac{n \cdot k}{100} = \frac{50 \cdot 25}{100} = 12,5,$ por lo que el intervalo crítico es [7,9] con $n_a = 22$.
Aplicando la fórmula:

$$P_{25} = L_i + \left(\frac{\frac{n \cdot 25}{100} - n_d}{n_c} \right) \cdot I = 6,5 + \left(\frac{\frac{50 \cdot 25}{100} - 9}{13} \right) \cdot 3 = 7,30769 \approx 7,31$$

Por lo tanto, la amplitud semi-intercuartil es :

$$Q = \frac{P_{75} - P_{25}}{2} = \frac{12{,}08 - 7{,}31}{2} = 2{,}385 \approx 2{,}39$$

En Psicología, en concreto, en la construcción de escalas de actitudes, la amplitud intercuartil (P_{75}-P_{25}), se ha utilizado profusamente en aquellos procedimientos de selección de ítems en los que se tiene en cuenta la valoración de jueces o expertos en la materia.

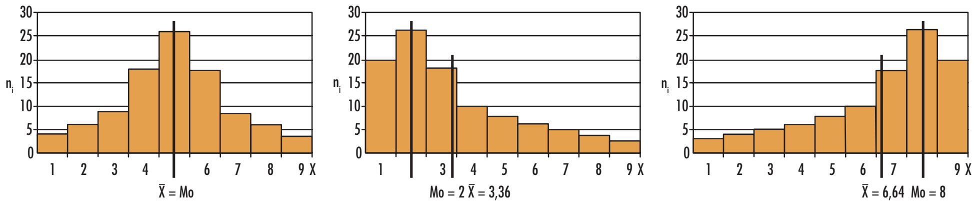
3.3. ÍNDICE DE ASIMETRÍA DE PEARSON

Tal y como se ha señalado en el primer tema, otra propiedad de una distribución de frecuencias relacionada con su forma es la asimetría o sesgo. La ***asimetría*** de una distribución nos indica el grado en el que las puntuaciones de los sujetos se reparten por debajo y por encima de la medida de tendencia central. Asimismo, en ese tema vimos cómo, mediante la representación gráfica de la distribución, podemos realizar un primer análisis sobre el grado de asimetría y observar si ésta es positiva o negativa. En este tema vamos a proponer un índice numérico que cuantifique esta propiedad. De entre los numerosos indicadores, hemos seleccionado el índice de asimetría de Pearson que se basa en la relación entre la media y la moda y matemáticamente se expresa de la siguiente manera:

Indice de asimetría de Pearson

$$A_s = \frac{\bar{X} - Mo}{S_X}$$

Se trata de un índice adimensional (no tiene unidades de medida) que se aplica a distribuciones unimodales. Cuando la distribución es simétrica, la media y la moda coinciden, por lo que el numerador se anula y el valor de $A_S = 0$. En distribuciones con asimetría positiva, la media es mayor que la moda, por lo que $A_S > 0$. Por otro lado, cuando la asimetría es negativa, el valor de la moda es superior al de la media y $A_S < 0$. En la figura 3.2 se establece la relación entre la representación gráfica de la asimetría de una distribución y el índice de asimetría de Pearson.



Distribución simétrica: $\bar{X} = Mo \Rightarrow A_s = 0$

Asimetría positiva: $\bar{X} > Mo \Rightarrow A_s > 0$

Asimetría negativa: $\bar{X} < Mo \Rightarrow A_s < 0$

Figura 3.2. Relación entre la asimetría de una distribución y el índice de Pearson.

Veamos un ejemplo:

Ejemplo 3.7. Con los datos del ejemplo 3.4, calcúlese el índice de asimetría de Pearson, sabiendo que la media es 9,62 y la varianza es igual a 10,52.

X	X_i	n_i
1-3	2	2
4-6	5	7
7-9	8	13
10-12	11	18
13-15	14	10
Σ		50

La desviación típica es $S_X = \sqrt{S_X^2} = \sqrt{10,52} = 3,24$, y la moda es el punto medio del intervalo modal [10-12] que es 11. Por lo tanto,

$$A_s = \frac{\bar{X} - Mo}{S_X} = \frac{9,62 - 11}{3,24} = \frac{-1,38}{3,24} = -0,4259 \approx -0,43$$

El resultado indica que la distribución presenta asimetría negativa, resultado que concuerda con la inspección visual de la gráfica de la distribución de frecuencias de la figura 3.3.

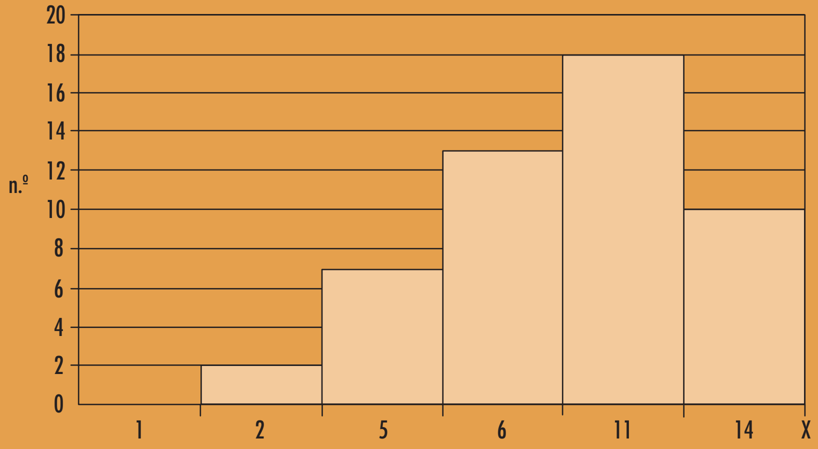


Figura 3.3. Representación gráfica de la distribución de frecuencias del Ejemplo 3.7.

3.4. PUNTUACIONES TÍPICAS

Hasta ahora hemos tratado fundamentalmente con puntuaciones directas (puntuaciones de un sujeto en un test, etc.). Son los primeros datos de los que habitualmente disponemos pero la comparación de las puntuaciones directas de un mismo sujeto en dos variables distintas puede llevarnos a confusión, ya que las puntuaciones directas nos ofrecen muy poca información. De hecho, conocida una puntuación directa no sabemos si se trata de un valor alto o bajo porque esto depende del promedio del grupo.

Si a una puntuación directa X_i le restamos la media de su grupo obtenemos una ***puntuación diferencial*** o de diferencia, que representamos por x_i (minúscula) y que, por tanto, viene definida así:

$$x_i = X_i - \overline{X}$$

Las puntuaciones diferenciales aportan más información: nos indican si la puntuación coincide con la media de su grupo, es inferior o es superior a ella. Estas puntuaciones presentan las siguientes propiedades:

a) Su media es cero: $\bar{x} = 0$

$$\bar{x} = \frac{\Sigma x_i}{n} = \frac{\Sigma\left(X_i - \overline{X}\right)}{n} = \frac{\Sigma X_i - \Sigma \overline{X}}{n} = \frac{\Sigma X_i}{n} - \frac{n\overline{X}}{n} = \overline{X} - \overline{X} = 0$$

b) La varianza de las puntuaciones diferenciales es igual a la varianza de las puntuaciones directas:

$$S_x^2 = \frac{\Sigma\left(x_i - \bar{x}\right)^2}{n} = \frac{\Sigma x_i^2}{n} = \frac{\Sigma\left(X_i - \overline{X}\right)^2}{n} = S_X^2$$

Por tanto, al restar a las puntuaciones directas su media hemos obtenido una nueva escala con media 0 y con idéntica varianza a las puntuaciones directas. Sin embargo, dos puntuaciones diferenciales idénticas pueden tener un significado muy diferente en función de la media y de la varianza de las distribuciones de las que proceden. Para eliminar este inconveniente se utilizan las puntuaciones típicas. Las puntuaciones típicas van más allá y nos permiten no sólo comparar las puntuaciones de un sujeto en dos variables distintas sino también comparar dos sujetos distintos en dos pruebas o variables distintas.

Una ***puntuación típica*** viene definida por:

$$z_x = \frac{x}{S_x} = \frac{X - \overline{X}}{S_x}$$

Al proceso de obtener puntuaciones típicas se llama tipificación, por este motivo estas puntuaciones se denominan también tipificadas.

En realidad una puntuación típica indica el número de desviaciones típicas que se aparta de la media una determinada puntuación.

Las puntuaciones típicas tienen las siguientes propiedades:

a) Su media es cero

$$\bar{z}_x = \frac{\sum z_x}{n} = \frac{\sum \left(\frac{x_i}{s_x} \right)}{n} = \frac{\frac{1}{s_x} \sum x_i}{n} = \frac{\sum x_i}{n s_x} = \frac{0}{n s_x} = 0$$

b) Su varianza es igual a 1

$$S^2_{z_x} = \frac{\sum \left(z_x - \bar{z}_x \right)^2}{n} = \frac{\sum \left(z_x \right)^2}{n} = \frac{\sum \left(\frac{x}{s_x} \right)^2}{n} = \frac{\frac{1}{S^2_x} \sum x^2}{n} = \frac{1}{S^2_x} \frac{\sum x^2}{n} = \frac{1}{S^2_x} S^2_x = 1$$

Las puntuaciones típicas reflejan las relaciones entre las puntuaciones con independencia de la unidad de medida. Por este motivo permiten hacer comparaciones entre distintos grupos e incluso entre distintas variables.

Ejemplo 3.8. Demostrar para las siguientes puntuaciones de 5 niños en la asignatura *X*: 6, 8, 7, 10 y 4 las propiedades de las puntuaciones diferenciales y típicas señaladas anteriormente.

X	x	$(x - \bar{x})^2$	z_x	$(z_x - \bar{z}_x)^2$
6	–1	1	–0,5	0,25
8	1	1	0,5	0,25
7	0	0	0	0
10	3	9	1,5	2,25
4	–3	9	–1,5	2,25
Σ	0	20	0	5

Puntuaciones Directas:

$$\bar{X} = 7 \,;\, S^2_X = 4$$

Puntuaciones Diferenciales:

$$\bar{x} = \frac{\sum x}{n} = \frac{0}{5} = 0$$

$$S_x^2 = \frac{\sum (x - \bar{x})^2}{n} = \frac{20}{5} = 4 = S_X^2$$

Puntuaciones Típicas:

$$\bar{z}_x = \frac{\sum z_x}{n} = \frac{0}{5} = 0$$

$$S_{z_x}^2 = \frac{\sum (z_x - \bar{z}_x)^2}{n} = \frac{5}{5} = 1$$

3.5. RESUMEN

Este tema se ha centrado en un aspecto fundamental en la caracterización de una distribución de frecuencias: la variabilidad o dispersión de los datos. Se han descrito las medidas de variabilidad que se emplean habitualmente, haciendo hincapié en las dos más relevantes en el campo de la estadística: la varianza y la desviación típica. Asimismo, se ha presentado un índice, el coeficiente de variación, que resulta útil para el estudio comparativo de la variabilidad en diferentes conjuntos de puntuaciones.

A continuación se ha analizado otra propiedad importante de una distribución relacionada con su forma como es la asimetría o sesgo. Con el fin de cuantificar el grado asimetría de una distribución y detectar el tipo de asimetría, se ha presentado el índice de asimetría de Pearson, basado en la relación entre la media y la moda del conjunto de las puntuaciones.

Por último, se han definido las puntuaciones diferenciales y las típicas que se derivan, a través de una transformación, de las puntuaciones directas de los sujetos. Se han estudiado las propiedades de cada tipo de puntuación así como la información que podemos obtener a partir de ellas para poder comparar entre sí a los sujetos, o al mismo sujeto en diferentes variables.

3.6. EJERCICIOS DE AUTOEVALUACIÓN

3.1. La varianza es una medida de dispersión que se basa en las desviaciones de cada puntuación con respecto a la: A) moda; B) mediana; C) media.

3.2. Si multiplicamos las puntuaciones de una variable por tres, la desviación típica de la nueva puntuación es: A) la misma que en la variable original; B) la desviación típica original multiplicada por tres; C) la desviación típica original multiplicada por nueve.

3.3. La desviación típica de una distribución de frecuencias: A) se expresa en las mismas unidades de medida que las puntuaciones; B) se expresa en las mismas unidades pero elevadas al cuadrado; C) no tiene unidades de medida.

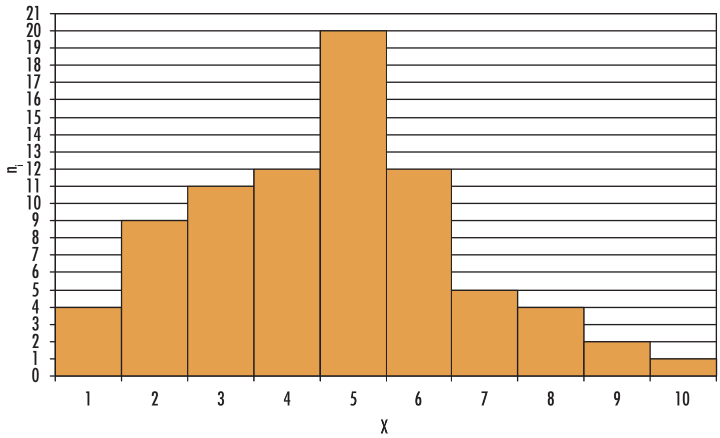
3.4. En una distribución marcadamente asimétrica, se recomienda medir la dispersión de los datos con: A) la amplitud semi-intercuartil; B) la varianza; C) el coeficiente de variación.

3.5. En el estudio de la asimetría de una distribución de frecuencias se ha observado un $A_s = 0{,}80$. La media de las puntuaciones es: A) igual que la moda; B) menor que la moda; C) mayor que la moda.

3.6. La variable X toma los siguientes valores: 50, 26, 35, 64, 34, 28, 73, 45, 48, 52, 54, 67. Sabiendo que la media es 48, la varianza es igual a: A) 15; B) 213; C) 115.

3.7. El valor del rango en los datos del ejercicio anterior es: A) 73; B) 23; C) 48.

3.8. La siguiente gráfica se corresponde con las notas en lengua de 80 niños de una clase de Primaria. Se sabe que la media es 4,63.



La desviación típica es igual a: A) 1,97; B) 2,53; C) 3,88.

3.9. En los datos del ejercicio anterior, ¿cuál es el valor de la amplitud total?: A) 8; B) 20; C)10.

3.10. Continuando con el ejercicio 3.8, el valor del índice de asimetría de Pearson es: A)–0,09; B)–0,19; C)–0,18.

3.11. Con los datos del ejercicio 3.8, a un sujeto con una puntuación de $X = 7$, ¿qué puntuación típica le corresponde?: A) 0,61; B) 1,20; C) 2,37.

3.12. ¿Cuál es el coeficiente de variación de la distribución de frecuencias del ejercicio 3.8?: A) 83,80; B) 46,32; C) 42,55.

3.13. De acuerdo con los datos del ejercicio 3.8, la amplitud semi-intercuartil es igual a: A) 3,56; B) 1,35; C) 2,69.

3.14. En la tabla adjunta se muestra la variable *edad* agrupada en intervalos cuya media es 50.

X	n_i
66-75	7
56-65	7
46-55	13
36-45	3
26-35	10

La desviación típica es: A) 13,96; B) 194,75; C) 6,50.

3.15. Continuando con la tabla del ejercicio anterior, la amplitud semi-intercuartil es igual a: A) 49; B) 12,86; C) 25,71.

3.16. Siguiendo con la tabla del ejercicio 3.14, ¿es exactamente simétrica la distribución?: A) sí; B) no, es ligeramente asimétrica positiva; C) no, es ligeramente asimétrica negativa.

3.17. De acuerdo a la distribución del ejercicio 3.14, un sujeto con 55 años, tiene una puntuación diferencial de: A) –5; B) 5; C) 0.

3.18. Si se compara la variabilidad de las distribuciones de frecuencias de los ejercicios 3.8 y 3.14, se concluye que la dispersión: A) es mayor en la puntuación en lengua; B) es mayor en la variable *edad*; C) es la misma en ambas variables.

3.19. El índice de asimetría de Pearson NO se puede calcular cuando: A) la variable es continua; B) la distribución es bimodal; C) la amplitud total es superior a diez.

3.20. Si realizamos la siguiente transformación lineal con las puntuaciones típicas, $V = 14 + 4z$, la varianza de la variable V será: A) 14; B) 4; C) 16.

3.7. SOLUCIONES A LOS EJERCICIOS DE AUTOEVALUACIÓN

3.1. Solución: C
La varianza mide la dispersión de los datos con respecto a la media, tal y como se puede apreciar en su fórmula (ver apartado 3.2.2).

3.2. Solución: B
$Y_i = 3X_i$
Según la cuarta propiedad de la varianza y la desviación típica (ver apartado 3.2.2) la desviación típica de las nuevas puntuaciones es $S_Y = |b|S_X$. En este caso, $S_Y = 3S_X$, es decir, es igual a la desviación típica original multiplicada por tres.

3.3. Solución: A
La desviación típica, a diferencia de la varianza, se expresa en las mismas unidades que la variable medida (ver apartado 3.2.2).

3.4. Solución: A
En una distribución asimétrica no es recomendable utilizar la media como medida de tendencia central. Como consecuencia, la varianza,

que se basa en la variabilidad con respecto a la media, tampoco es recomendable. Una alternativa es la amplitud semi-intercuartil, un índice resistente de dispersión (ver apartado 3.2.4).

3.5. Solución: C
Tal y como se indica en el apartado 3.3, cuando el índice de asimetría de Pearson es positivo (A_s = 0,80), la media es mayor que la moda.

3.6. Solución: B
$\bar{X}$ = 48

X_i	$(X_i - \bar{X})$	$(X_i - \bar{X})^2$	X_i^2
50	2	4	2500
26	–22	484	676
35	–13	169	1225
64	16	256	4096
34	–14	196	1156
28	–20	400	784
73	25	625	5329
45	–3	9	2025
48	0	0	2.304
52	4	16	2704
54	6	36	2916
67	19	361	4761
$\sum$	0	2.556	30204

$$S_X^2 = \frac{\sum (X_i - 48)^2}{12} = \frac{2556}{12} = 213$$

$$S_X^2 = \frac{\sum X_i^2}{12} - 48^2 = \frac{30204}{12} - 2304 = 2517 - 2304 = 213$$

3.7. Solución: C

$$X_{\text{máx}} = 73{,}5 \text{ y } X_{\text{mín}} = 25{,}5.\ A_T = X_{\text{máx}} - X_{\text{mín}} = 73{,}5 - 25{,}5 = 48$$

3.8. Solución: A

X_i	n_i	n_a	X_i^2	$n_i X_i^2$
10	1	80	100	100
9	2	79	81	162
8	4	77	64	256
7	5	73	49	245
6	12	68	36	432
5	20	56	25	500
4	12	36	16	192
3	11	24	9	99
2	9	13	4	36
1	4	4	1	4
Σ	80	—		2026

$$\bar{X} = 4,63$$

$$S_X^2 = \frac{\sum n_i X_i^2}{80} - (4,63)^2 = \frac{2026}{80} - 21,4369 = 3,8881$$

$$S_X = \sqrt{3,8881} = 1,9718 \approx 1,97$$

3.9. Solución: C

$$X_{\text{máx}} = 10,5 \text{ y } X_{\text{mín}} = 0,5.\ A_T = X_{\text{máx}} - X_{\text{mín}} = 10,5 - 0,5 = 10$$

3.10. Solución: B

$$\bar{X} = 4,63 \quad Mo = 5 \quad A_s = \frac{\bar{X} - Mo}{S_X} = \frac{4,63 - 5}{1,97} = \frac{-0,37}{1,97} = -0,18781 \approx -0,19$$

3.11. Solución: B

$$\bar{X} = 4,63 \quad S_X = 1,97 \quad z_x = \frac{X - \bar{X}}{S_x} = \frac{7 - 4,63}{1,97} = 1,2030 \approx 1,20$$

3.12. Solución: C

$$\bar{X} = 4,63 \quad S_X = 1,97 \quad CV = \frac{S_x}{\bar{X}} \cdot 100 = \frac{1,97}{4,63} \cdot 100 = 42,54859 \approx 42,55$$

3.13. Solución: B

$Q_1 = P_{25}, \quad \frac{n \cdot 25}{100} = \frac{80 \cdot 25}{100} = 20,$ por lo que el intervalo crítico es [2,5-3,5] con $n_a = 24$.

$$P_{25} = L_i + \left(\frac{\frac{n \cdot k}{100} - n_d}{n_c} \right) \cdot I = 2,5 + \left(\frac{\frac{80 \cdot 25}{100} - 13}{11} \right) \cdot 1 = 3,136 \approx 3,14$$

$Q_3 = P_{75}, \quad \frac{n \cdot 75}{100} = \frac{80 \cdot 75}{100} = 60,$ por lo que el intervalo crítico es [5,5-6,5] con $n_a = 68$.

$$P_{75} = L_i + \left(\frac{\frac{n \cdot k}{100} - n_d}{n_c} \right) \cdot I = 5,5 + \left(\frac{\frac{80 \cdot 75}{100} - 56}{12} \right) \cdot 1 = 5,833 \approx 5,83$$

$$Q = \frac{P_{75} - P_{25}}{2} = \frac{5,83 - 3,14}{2} = 1,345 \approx 1,35$$

3.14. Solución: A

X	X_i	n_i	n_a	X_i^2	$n_i\ X_i^2$
66-75	70,5	7	40	4970,25	34791,75
56-65	60,5	7	33	3660,25	25621,75
46-55	50,5	13	26	2550,25	33153,25
36-45	40,5	3	13	1640,25	4920,75
26-35	30,5	10	10	930,25	9302,50
Σ	—	40			107790

$$\bar{X} = 50$$

$$S_X^2 = \frac{\sum n_i X_i^2}{40} - (50)^2 =$$

$$= \frac{107790}{40} - 2500 = 194,75$$

$$S_X = \sqrt{194,75} = 13,95528 \approx 13,96$$

3.15. Solución: B

Percentil 25:

$\frac{n \cdot k}{100} = \frac{40 \cdot 25}{100} = 10,$ por lo que el intervalo crítico es [26-35] con n_a = 10.

Aplicando la fórmula:

$$P_{25} = L_i + \left(\frac{\frac{n \cdot k}{100} - n_d}{n_c} \right) \cdot I = 25,5 + \left(\frac{\frac{40 \cdot 25}{100} - 0}{10} \right) \cdot 10 = 35,5$$

Percentil 75:

$\frac{n \cdot k}{100} = \frac{40 \cdot 75}{100} = 30,$ por lo que el intervalo crítico es [56-65] con n_a = 33.

Aplicando la fórmula:

$$P_{75} = L_i + \left(\frac{\frac{n \cdot k}{100} - n_d}{n_c} \right) \cdot I = 55,5 + \left(\frac{\frac{40 \cdot 75}{100} - 26}{7} \right) \cdot 10 = 61,21$$

$$Q = \frac{P_{75} - P_{25}}{2} = \frac{61,21 - 35,5}{2} = 12,855 \approx 12,86$$

3.16. Solución: C

$$A_s = \frac{\bar{X} - Mo}{S_X} = \frac{50 - 50,5}{13,96} = -0,03581$$

3.17. Solución: B

$$\bar{X} = 50 \quad x_i = X_i - \bar{X} = 55 - 50 = 5$$

3.18. Solución: A

Coeficiente de variación de *lengua*: $CV_L = 42{,}55$.

Coeficiente de variación de *edad*:

$$\bar{X} = 50 \quad S_X = 13{,}96 \quad CV_E = \frac{S_X}{\bar{X}} \cdot 100 = \frac{13{,}96}{50} \cdot 100 = 27{,}92$$

Dado que $CV_L > CV_E$, existe mayor dispersión en la distribución de las puntuaciones en *lengua*.

3.19. Solución: B

El índice de asimetría de Pearson se puede calcular en variables continuas y con cualquier valor en su amplitud total. En cambio, no se puede calcular cuando la distribución es bimodal (ver Apartado 3.3).

3.20. Solución: C

$$V = 14 + 4z \;\Rightarrow\; S_V^2 = 4^2 \cdot S_z^2 = 16 \cdot 1 = 16$$

Tema 4

Análisis conjunto de dos variables

4.1. INTRODUCCIÓN

En los temas estudiados hasta ahora, nos hemos limitado al análisis y descripción de una variable. En Psicología, y en cualquier otra disciplina de las denominadas Ciencias de la Salud, es frecuente trabajar con varias variables, obteniendo en ello más información. En este tema nos limitaremos al estudio conjunto de dos variables. En primer lugar consideraremos el caso de dos variables cualitativas (clasificadoras o categóricas y medidas en escala nominal) y, posteriormente, analizaremos el caso de dos variables cuantitativas (medidas en una escala de intervalo o de razón).

No consideraremos, separadamente, las variables cuasi-cuantitativas, medidas en una escala ordinal. Estas se pueden elaborar con las mismas técnicas, que indicamos para las variables cualitativas. Trabajando así no se tiene en cuenta la información que aporta el hecho de que sus valores están ordenados. Por supuesto, existen métodos de análisis apropiados, que tienen en cuenta la información de orden de estas variables, pero no vamos a explicarlos aquí.

En el caso de dos variables cualitativas utilizaremos la tabla de datos, elaboraremos la tabla de contingencia, el diagrama de barras, y la tabla de diferencias entre frecuencias empíricas y teóricas, para su representación gráfica. Aprenderemos a calcular el coeficiente X^2, y el coeficiente de contingencia, para el estudio de su grado de asociación. Veremos las propiedades fundamentales de estas elaboraciones, su cálculo, su aplicación a casos concretos y su interpretación.

En el caso de dos variables cuantitativas, X e Y, utilizaremos la tabla de datos conjuntos. Elaboraremos el diagrama de dispersión como representación gráfica; aprenderemos a calcular la covarianza y el coeficiente de correlación de Pearson como los dos índices fundamentales para el análisis de la relación lineal entre ellas. Veremos las propiedades fundamentales de la covarianza y del coeficiente de correlación de Pearson, su cálculo, su aplicación a casos concretos y su interpretación.

Finalmente, si dos variables cuantitativas están relacionadas linealmente podemos utilizar una de ellas para efectuar predicciones o pronósticos sobre la otra. La recta de regresión será el instrumento adecuado para ello. Definiremos y analizaremos la ecuación de la recta que nos permita, en cada caso, efectuar las predicciones con el menor margen de error y estableceremos las relaciones entre esta ecuación y el coeficiente de correlación de Pearson.

Los objetivos que pretendemos son los siguientes:

- Distinguir entre variables cualitativas y cuantitativas, y saber elegir los métodos que hay que utilizar en cada caso.
- Conocer métodos gráficos y cuantitativos para analizar la relación existente entre dos variables.
- Adquirir la capacidad para saber si dos variables están más o menos relacionadas entre sí, la forma de esa relación, y el significado de que dos variables estén relacionadas.
- En el caso de dos variables cuantitativas, entre las que hay relación lineal, aprenderemos a hacer predicciones de los valores de la variable Y, correspondientes a cada valor de la variable X, mediante la recta de regresión.

4.2. CONCEPTOS PREVIOS

Vamos a iniciar el estudio conjunto de dos variables observando 100 sujetos. Un mismo sujeto tendrá dos medidas, una por cada una de las variables escogidas. Una de las variables es la variable X, Género, y se anota en cada caso si es Varón o Mujer. A cada sujeto le administraremos un test para medir el estrés, si padece estrés (Sí) o no (No), y anotaremos el resultado. Nos resulta una lista en la que en cada fila está la información de un sujeto y tenemos cuatro columnas. La primera columna aparece con el número del caso, o su identificador. En la segunda aparece el Nombre y Apellidos del sujeto, en la tercera Varón o Mujer, según proceda, y en la cuarta aparece «estrés» o «no estrés» (*Sí* o *No,* respectivamente), según el resultado del test. Al final de este trabajo, tendremos una lista de cuatro columnas y 100 filas. Con la información de esta lista, queremos saber si las dos variables están relacionadas y, si están relacionadas entre sí, cómo es esa relación. Ponemos, en el ejemplo 4.1, el principio y el fin de esta tabla:

Ejemplo 4.1. Hemos recogido los datos de esta muestra de 100 sujetos con dos variables: ***X***, representa el género o sexo e ***Y*** el grado de estrés con dos categorías (Sí padece estrés o No padece estrés). Se presenta a continuación el principio y el fin de esta lista de datos (una lista de datos, para recoger ordenadamente la información, puede tener más columnas si se evalúan más variables sobre cada sujeto):

Tabla 4.1. Tabla de datos

Caso	Nombre y Apellidos	Género (*X*)	Estrés (*Y*)
1	Francisco Pérez García	Varón	Sí
2	Lucía Revilla López	Mujer	No
...	...	...	...
99	Inés Ayala Ruiz	Mujer	Sí
100	David Ruipérez Rodríguez	Varón	Sí

Podemos definir el concepto de ***Asociación y/o Relación*** entre dos variables. Dos variables están relacionadas entre sí, cuando ciertos valores, de una de las variables, se asocian con ciertos valores de la otra variable. Por ejemplo, si tenemos en cuenta las variables Género y Estrés y... si sucede que cuando se tiene el valor «Varón» en la variable Género, hay una incidencia mayor del valor «Sí» en la variable Estrés, y además, cuando se tiene el valor «Mujer» en la variable Género, hay una incidencia mayor del valor «No» en la variable Estrés, decimos que las variables Género y Estrés están relacionadas.

Sin embargo, la asociación o relación entre las variable género y estrés también podría ser al revés: los varones tienden a no tener estrés y las mujeres tienden a tenerlo.

Aún existe otra opción posible que puede darse en esta situación: tanto los varones como las mujeres pueden estar equiparados en el estrés o no estrés. Así, no existiría asociación y/o relación entre las variables consideradas.

4.3. ASOCIACIÓN ENTRE DOS VARIABLES CUALITATIVAS

Hemos definido una variable como «cualitativa» cuando está medida en una escala nominal, o de clasificación (tema 1). Estas variables pueden ser a su vez dicótómicas, cuando sólo presentan dos categorías, o politómicas cuando presentan un mayor número. También consideraremos como cualitativas aquellas variables que, en un principio, presentan un mayor nivel de medida (intervalos o razón) pero, a posteriori, han sido categorizadas. Cuando se dispone de los datos de dos variables cualitativas para todos los sujetos de una muestra se puede elaborar la denominada ***Tabla de Contingencia.***

Consideremos el ejemplo 4.1, al que ya nos hemos referido, de 100 sujetos, en cada uno de los cuales se han recogido el valor de la variable género y estrés. A partir de la tabla con la información de toda la muestra del Ejemplo 4.1, contabilizamos los cuatro casos posibles (**V**arón, Sí), (**V**arón, No), (**M**ujer, Sí), (**M**ujer, No) y elaboramos la Tabla de Contingencia del Ejemplo 4.2.

Ejemplo 4.2. Hemos recogido los datos de esta muestra de 100 sujetos en dos variables: **X**, representa el *género o sexo* e **Y** el *grado de estrés* con dos categorías (Sí padece estrés o No padece estrés). Estos datos, contabilizando las cuatro combinaciones posibles, aparecen en la siguiente tabla de contingencia:

Tabla 4.2. Frecuencias observadas o empíricas (n_e) en *X* e *Y*

		Y		
		Sí	No	
X	V	30	10	40
	M	25	35	60
		55	45	100

donde:

V = «varón», M = «mujer»,

Sí = «padece estrés» y No = «no padece estrés».

Esta Tabla de Contingencia podemos representarla gráficamente mediante el siguiente diagrama de barras:

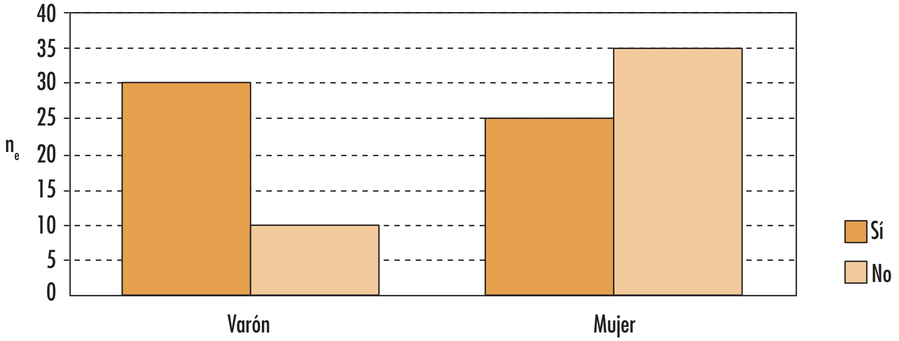


Figura 4.1. Diagrama de barras correspondiente al ejemplo 4.2.

Si observamos detenidamente la tabla 4.2 y la figura 4.1, podremos ver que:

- El grupo de varones tiene una incidencia mayor de «padece estrés», que el grupo de mujeres.
- El grupo de varones tiene una incidencia menor de «no padece estrés» que el grupo de mujeres.

Podemos preguntarnos: ¿existe alguna relación entre la variable **Género** y **Padecer o No Padecer Estrés**?, o ¿son independientes el **Género** y **Padecer o No Padecer Estrés**? Para responder a estas preguntas tendríamos que definir un estadístico, $\boldsymbol{X^2}$, asociado a una distribución de probabilidad (Chi cuadrado, χ^2) que estudiaréis en el Capítulo 7. Nuestro estadístico permite determinar si dos variables están relacionadas o son independientes. Así, $\boldsymbol{X^2}$ se define en función de las ***frecuencias empíricas*** (n_e) y las ***frecuencias teóricas*** (n_t). Las frecuencias teóricas se calculan asumiendo que ambas variables son independientes o no relacionadas. Las frecuencias teóricas, n_t, serán el producto del total de su fila por el total de su columna dividido por la frecuencia total, n. Es decir:

$$\text{Frecuencia teórica} = n_t = \frac{\text{Total fila} \times \text{Total columna}}{n}$$

Así, para la Tabla de Contingencia del ejemplo 4.2, tendríamos las siguientes frecuencias teóricas:

Tabla 4.3. Frecuencias teóricas (n_t) correspondientes a la tabla 4.2

		Y		
		Sí	No	
X	V	$\frac{40 \times 55}{100} = 22$	$\frac{40 \times 45}{100} = 18$	40
	M	$\frac{60 \times 55}{100} = 33$	$\frac{60 \times 45}{100} = 27$	60
		55	45	100

Obviamente, la suma de todas las frecuencias teóricas marginales es igual a la suma de todas las frecuencias empíricas marginales, e igual al total de todas las observaciones, n.

Calculadas las frecuencias teóricas, a continuación se puede elaborar la Tabla 4.4 de diferencias entre frecuencias empíricas menos frecuencias teóricas.

Tabla 4.4. Diferencias de las frecuencias observadas o empíricas, n_e, menos las frecuencias teóricas, n_t, en *X* e *Y* (la suma de filas y columnas en esta Tabla tiene que ser siempre igual a cero)

		Y	
		Sí	No
X	V	8	–8
	M	–8	8

donde:

V = «varón», M = «mujer»,

Sí «padece estrés» y

No «no padece estrés»

Para ver la forma que toma la relación entre las variables, es necesario observar la tabla 4.4. Los valores positivos de las diferencias (el **8**), nos indican una relación entre el Sí-V (*sí padecer estrés* y *ser varón*) y entre el No-M (*no padecer estrés y ser mujer*). Los valores negativos de las diferencias (el **–8)**, nos indican una relación negativa entre el No-V (*ser varón y la negación de no padecer estrés*) y entre el Sí-M (*ser mujer y la negación de padecer estrés*). Uniendo estos resultados se ve la forma de la relación entre las variables. En nuestro ejemplo, los varones tienen una mayor tendencia a tener estrés y las mujeres tienen menos tendencia a tener estrés.

Una vez vista la forma de la relación entre las dos variables, calculamos el estadístico **X²** (X Cuadrado) cuya expresión es:

$$\text{Estadístico } \boldsymbol{X^2} = \sum\sum \frac{(n_e - n_t)^2}{n_t}$$

donde,

n_e es la frecuencia empírica (o también llamada frecuencia conjunta observada).

n_t es la frecuencia teórica (o también llamada frecuencia conjunta esperada).

Aplicada la fórmula anterior a las Tablas 4.2 y 4.3, con las frecuencias empíricas y teóricas, respectivamente, el resultado es el siguiente:

$$X^2 = \left[\frac{(30\text{-}22)^2}{22} + \frac{(10\text{-}18)^2}{18} + \frac{(25-33)^2}{33} + \frac{(35\text{-}27)^2}{27} \right] = 2,91 + 3,56 + 1,94 + 2,37 = 10,78$$

Uno de los inconvenientes del estadístico $\mathbf{X^2}$ es su difícil interpretación puesto que desconocemos su límite superior. Sólo sabemos que tiene valor cero, cuando no hay relación entre las dos variables, es decir, cuando las frecuencias empíricas y teóricas son iguales en todos los casos. Por ello, las variables son independientes.

Para resolver el problema que conlleva la interpretación de la relación entre dos variables, de acuerdo al valor obtenido de $\mathbf{X^2}$ (*X* cuadrado), se ha definido el ***índice o Coeficiente de Contingencia***, ***C***. Este índice toma los valores $0 \leq C < 1$. Su fórmula es la siguiente:

$$\text{Coeficiente de Contingencia} = C = \sqrt{\frac{X^2}{X^2 + n}}$$

En el ejemplo anterior:

$$C = \sqrt{\frac{X^2}{X^2 + n}} = \sqrt{\frac{10,78}{10,78 + 100}} = \sqrt{0,097} = 0,312$$

El valor de *C* obtenido se puede comparar, dado que la Tabla de Contingencia tiene igual número de filas que de columnas (*k*), con un *C* máximo definido como:

$$C_{máx} = \sqrt{\frac{k-1}{k}}$$

En nuestro caso, para $k = 2$,

$$C_{máx} = \sqrt{\frac{2-1}{2}} = \sqrt{0,5} = 0,707$$

Nuestro valor obtenido, C = 0,312, comparado con 0,707 (el C máximo) es casi la mitad. En cuanto a la relación, ésta es cercana al tipo medio.

Hasta aquí hemos considerado la asociación entre dos variables cualitativas pero sólo con dos valores en cada una de ellas. Las tablas de los datos han sido de 2 × 2. Pasamos ahora a presentar la relación entre dos variables cualitativas pero cada una de ellas tendrá más de dos categorías. Los datos provienen de una investigación de Garriga-Trillo y Aguilera-Genicio (2007), de un Curso de Verano en el Centro Asociado de Denia y de una presentación y publicación en el libro Fechner Day 2005 (Garriga-Trillo y Aguilera-Genicio, 2005), sobre la relación entre el grado de deterioro cognitivo y la sensibilidad olfativa, pudiendo llegar a pronosticar el deterioro cognitivo en función de la sensibilidad olfativa utilizando varias medidas olfativas. En este caso, sólo presentamos una de las medidas de la sensibilidad olfativa como variable dependiente, **Y** (**Número de aciertos en la identificación de olores)**, con valores desde 0-5 aciertos. La variable independiente, **X**, incluye tres **Grupos de deterioro cognitivo** (Grupo Control, Deterioro Cognitivo Leve y Pacientes de Alzheimer).

Tabla 4.5. Tabla de frecuencias empíricas, n_e, en X e Y

		Y = Número de aciertos en la identificación de olores						
		0	**1**	**2**	**3**	**4**	**5**	
X = Grupos de deterioro cognitivo (GDC)	Grupo Control			18	42	54	12	126
	Deterioro Cognitivo Leve	6	54	30	30			120
	Pacientes de Alzheimer	72	43	17	12			144
		78	97	65	84	54	12	390

Tabla 4.6. Tabla de frecuencias teóricas, n_t, en X e Y

		Y = Número de aciertos en la identificación de olores						
		0	**1**	**2**	**3**	**4**	**5**	
X = Grupos de deterioro cognitivo (GDC)	Grupo Control	25,20	31,34	21,00	27,14	17,45	3,88	126
	Deterioro Cognitivo Leve	24,00	29,85	20,00	25,85	16,62	3,69	120
	Pacientes de Alzheimer	28,80	35,82	24,00	31,02	19,94	4,43	144
		78	97	65	84	54	12	390

Tabla 4.7. Tabla de diferencias entre n_e y n_t

		Y = Número de aciertos en la identificación de olores					
		0	**1**	**2**	**3**	**4**	**5**
X = Grupos de deterioro cognitivo (GDC)	Grupo Control	–25,20	–31,34	–3,00	14,86	36,55	8,12
	Deterioro Cognitivo Leve	–18,00	24,15	10,00	4,15	–16,62	–3,69
	Pacientes de Alzheimer	43,20	7,18	–7,00	–19,02	–19,94	–4,43

El valor de $X^2 = 322{,}05$ y el del Coeficiente de Contingencia, $C = 0{,}673$. Ambos estadísticos indican que existe una relación significativa entre los **Grupos de Deterioro Cognitivo** y el **Número de Aciertos en la Identificación de Olores**. El grupo control tiene el mayor número de aciertos y no tiene ningún sujeto con menos de 2 aciertos. Los grupos con deterioro no tienen ningún sujeto con más de tres aciertos.

Visto el ejemplo del coeficiente C utilizando tablas de contingencia de más de dos filas y dos columnas, pasamos a mencionar las ***Características del Coeficiente C.*** Éstas son:

— El coeficiente de contingencia, *C*, puede asumir valores mayores o iguales a cero y menores que 1, como ya hemos señalado. El valor cero lo alcanza cuando $X^2 = 0$ e indica que las dos variables no tienen relación entre ellas y, además, las frecuencias empíricas coinciden con las frecuencias teóricas. El valor uno, sólo se consigue si $n = 0$, lo que implica que no hay observaciones, por lo que nunca se puede dar.

— Cuanto mayor es el valor de *C*, mayor es la relación entre las dos variables, y al revés, cuanto menor es el valor de *C*, menor es la relación entre las dos variables. Si queremos utilizar el valor de *C* para comparar la relación entre las mismas dos variables, cuyos datos tenemos en dos tablas de contingencia diferentes y son resultado de dos investigaciones distintas, tenemos que vigilar que ambas tablas de contingencia tengan el mismo número de filas y de columnas y aproximadamente el mismo número de datos. Si no tienen el mismo número de filas, de columnas, y aproximadamente el mismo número de datos, los valores de *C* no permiten una comparación válida de la relación de las variables en ambas investigaciones.

— Otro aspecto más complejo es fundamentar la causalidad en un coeficiente de contingencia. Cuando existe un valor elevado en nuestro coeficiente de contingencia, no se puede afirmar que una de las variables es causa de la otra. Hay cantidad de variables que se relacionan entre sí, porque existe otra variable ajena que tiene una relación clara con ambas. Un ejemplo de esto es la influencia que aparece en muchos casos entre zona geográfica y la corrección en la forma de hablar. Esto no implica que la corrección en la forma de hablar sea causada por la geografía, sino, tal vez, por la influencia de diferentes procesos educativos.

— Se puede estimar, en casos en que la tabla de contingencia tenga igual número de filas que de columnas, un valor máximo que puede alcanzar C.

4.4. CORRELACIÓN ENTRE DOS VARIABLES CUANTITATIVAS

Cuando tenemos dos variables cuantitativas el primer paso a realizar es obtener una muestra grande, como la del ejemplo 4.1 (n = 100). En este Tema utilizaremos, con fines didácticos, un ejemplo muy sencillo y con un pequeño número de observaciones. Indicaremos, a continuación, cómo realizar ***el diagrama de dispersión***, calcular la ***covarianza*** y calcular el ***coeficiente de correlación de Pearson***. Terminaremos indicando cómo interpretar los resultados.

En primer lugar presentamos la Tabla de Datos del ejemplo, sobre el que realizaremos todas las elaboraciones de este apartado, y del apartado 4.5 sobre la recta de regresión.

Ejemplo 4.3. Sea la variable **X** la puntuación obtenida en un **test de razonamiento numérico** y sea la variable **Y** la **calificación obtenida en la asignatura de matemáticas**. Para un grupo de 5 niños hemos obtenido los resultados recogidos en esta Tabla.

Niño	*X*	*Y*
1	4	6
2	8	4
3	10	7
4	12	8
5	16	10

En primer lugar, vamos a considerar la representación gráfica de dos variables cuantitativas. Se trata del diagrama de dispersión, o «nube de puntos», que es la representación gráfica más utilizada, y más habitual, por la información que facilita.

La representación gráfica de la tabla del ejemplo 4.3 aparece en la figura 4.2. En ella puede apreciarse que existe una cierta ***relación lineal*** en las variables X e Y. En general, a medida que aumentan las puntuaciones en el test (variable X) aumentan también las calificaciones en matemáticas (variable Y).

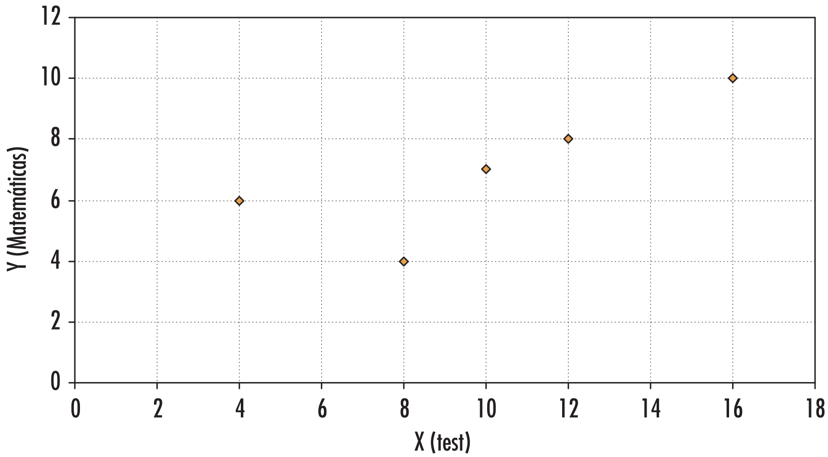


Figura 4.2. Diagrama de dispersión correspondiente al ejemplo 4.3.

Vamos a estudiar dos índices, relacionados entre sí, que permiten cuantificar la relación lineal que pueda haber entre dos variables cuantitativas.

La ***covarianza*** es un primer índice, que nos permitirá estudiar esa posible relación entre X e Y. El término covarianza hace referencia a la variación conjunta de dos variables, y tanto por su definición como por su cálculo, es un índice que cuantifica la variabilidad conjunta de dos variables. Se designa por ***Cov (X,Y)***, o por $\boldsymbol{S_{XY}}$. Se define así:

$$\text{Covarianza} = S_{XY} = \frac{\sum_{i=1}^{n} X_i Y_i}{n} - \bar{X}\,\bar{Y}$$

donde:

X_i = valor de la variable X en el caso i.

Y_i = valor de la variable Y en el caso i.

$\bar{X}$ = media de la variable X.

$\bar{Y}$ = media de la variable Y.

n = número de casos de la muestra.

Aplicando la fórmula a los datos de la tabla del ejemplo 4.3, elaboramos la siguiente tabla para el cálculo de la Covarianza:

Niño	***X***	***Y***	***XY***
1	4	6	24
2	8	4	32
3	10	7	70
4	12	8	96
5	16	10	160
	50	35	382

$$\bar{X} = \frac{50}{5} = 10 \qquad \bar{Y} = \frac{35}{5} = 7$$

$$S_{XY} = \frac{\sum_{i=1}^{n}(X_i Y_i)}{n} - \bar{X}\,\bar{Y} = \frac{382}{5} - (10 \cdot 7) = 76,4 - 70 = 6,4$$

El signo, positivo o negativo, de la covarianza nos indica si la relación lineal entre ambas variables es directa o inversa, respectivamente.

Podemos definir que la **relación lineal directa** es la que asume que a valores mayores en una de las variables, corresponden también valores mayores en la otra variable y los valores menores en una variable se corresponden con los valores menores en la otra variable.

Igualmente, definimos que la **relación lineal inversa** es la que asume que a valores mayores en una de las variables, corresponden valores menores en la otra variable y viceversa.

En nuestro caso existe, como habíamos pronosticado a partir del diagrama de dispersión, una relación directa entre la puntuación en el test, variable *X*, y la calificación en matemáticas, variable *Y*.

Sin embargo, la covarianza presenta un grave problema, al igual que habíamos visto para el coeficiente X^2 (con variables cualitativas), desconocemos el rango de la covarianza. En este caso son los valores máximos y mínimos que pueda tener. Para evitar este problema disponemos del ***Coeficiente de Correlación de Pearson***, r_{XY}.

El coeficiente de correlación de Pearson entre dos variables *X* e *Y*, que designaremos por r_{XY}, viene definido de la siguiente manera:

$$\text{Coeficiente de Correlación de Pearson} = r_{XY} = \frac{S_{XY}}{S_X S_Y}$$

donde:

S_X = desviación típica de la variable *X*.

S_Y = desviación típica de la variable *Y*.

S_{XY} = covarianza entre *X* e *Y*.

Es decir, el coeficiente de correlación de Pearson es el cociente entre la covarianza entre *X* e *Y* y el producto de la desviación típica de *X* y la desviación típica de *Y*. Las desviaciones típicas de *X* e *Y* son, respectivamente, 4 y 2.

Siguiendo con los mismos datos de la tabla del ejemplo 4.3, y calculadas las desviaciones típicas de las variables X e Y, como ya sabemos:

$$r_{XY} = \frac{S_{XY}}{S_X S_Y} = \frac{6{,}4}{4 \times 2} = \frac{6{,}4}{8} = 0{,}8$$

El coeficiente de correlación de Pearson, r_{XY}, presenta —entre otras— las siguientes propiedades.

Propiedades:

1) $-1 \leq r_{XY} \leq 1$. Es decir, sólo toma valores comprendidos entre –1 y 1. Valdrá 0 cuando no exista relación lineal entre *X* e *Y*.

2) $r_{XY} = \pm\, 1$, si una variable es una transformación lineal de la otra.

Una fórmula para calcular r_{XY}, utilizada con mucha frecuencia, y alternativa a la presentada a partir de la covarianza de X e Y y de las desviaciones típicas de X y de Y, es la siguiente:

$$r_{XY} = \frac{\mathrm{n}\sum(\mathrm{XY}) - \sum \mathrm{X} \sum \mathrm{Y}}{\sqrt{\mathrm{n}\sum \mathrm{X}^2 - \left(\sum \mathrm{X}\right)^2}\sqrt{\mathrm{n}\sum \mathrm{Y}^2 - \left(\sum \mathrm{Y}\right)^2}}$$

Veamos su aplicación a partir de los datos de la tabla del ejemplo 4.3, elaborando la siguiente Tabla y efectuando los cálculos oportunos:

Niño	***X***	***Y***	***XY***	***X²***	***Y²***
1	4	6	24	16	36
2	8	4	32	64	16
3	10	7	70	100	49
4	12	8	96	144	64
5	16	10	160	256	100
Σ	50	35	382	580	265

$$r_{XY} = \frac{n\sum(XY) - \sum X \sum Y}{\sqrt{n\sum X^2 - \left(\sum X\right)^2}\sqrt{n\sum Y^2 - \left(\sum Y\right)^2}} =$$

$$= \frac{5 \times 382 - 50 \times 35}{\sqrt{5 \times 580 - (50)^2}\sqrt{5 \times 265 - (35)^2}} =$$

$$= \frac{1910 - 1750}{\sqrt{400}\sqrt{100}} = \frac{160}{200} = 0{,}8$$

Para interpretar los resultados que se obtienen con el coeficiente de correlación de Pearson hay que tener en cuenta, en primer lugar, el valor absoluto. Cuanto mayor es el valor absoluto el coeficiente nos está indicando que la relación lineal entre las dos variables es más fuerte. En segundo lugar, hay que tener en cuenta el signo del coeficiente de correlación de Pearson. Cuando el signo es positivo, indica que a valores mayores de la variable X tienden a corresponder, en media, valores mayores de la variable Y, y a valores menores de la variable X tienden a corresponder, en media, valores menores de la variable Y. Esta es una *relación directa*. Cuando el signo es negativo, indica que a valores mayores de la variable X tienden a corresponder, en media, valores menores de la variable Y, y a valores menores de la variable X tienden a corresponder, en media, valores mayores de la variable Y. Esta es una *relación inversa*.

En la figura 4.3, aparecen cuatro diagramas de dispersión, o nubes de puntos, que nos van a servir para presentar cuatro situaciones posibles, que nos podemos encontrar, cuando analizamos la relación entre dos variables cuantitativas.

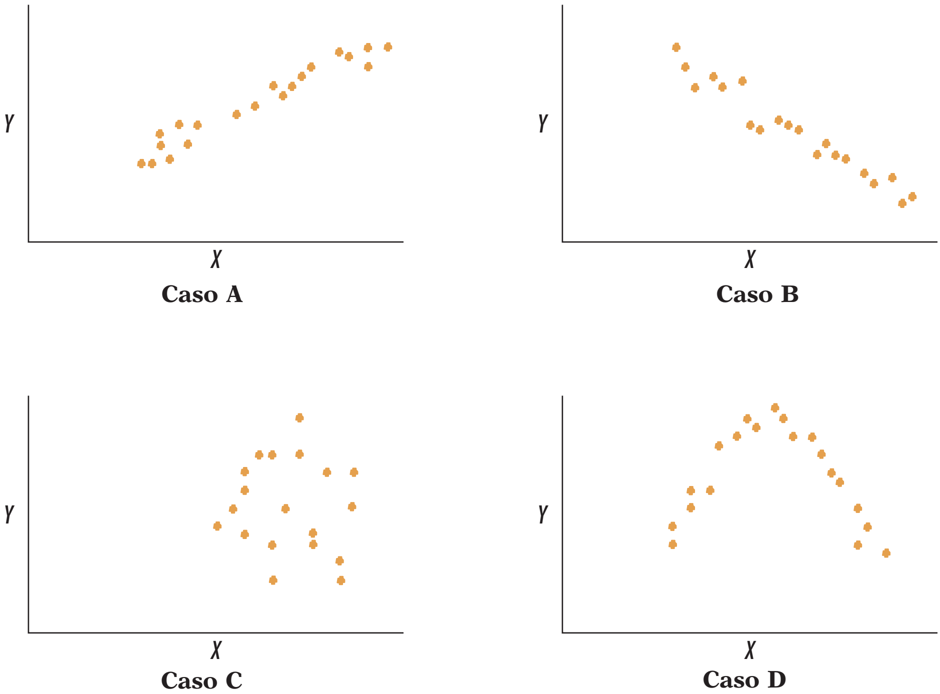


Figura 4.3. Cuatro nubes de puntos, para comentar sus coeficientes de correlación de Pearson.

En las cuatro nubes de puntos, que vemos en la figura 4.3, nos encontramos situaciones típicamente diferentes. El caso que está arriba a la izquierda (**Caso A**), responde a un coeficiente de correlación positivo, indicando una relación lineal directa bastante clara. A valores mayores de la variable *X*, corresponden valores mayores de la variable *Y*. El caso que está arriba a la derecha (**Caso B**), responde a un coeficiente de correlación negativo, indicando una relación lineal inversa. A mayores valores de la variable X, corresponden menores valores de la variable Y. El caso que está abajo a la izquierda (**Caso C**), responde a un coeficiente de correlación lineal de valor cercano a cero. No existe correlación lineal. El caso que está abajo a la derecha (**Caso D**), responde también a un coeficiente de correlación lineal de valor cercano a cero. De hecho, no existe relación lineal, pero sí existe una relación curvilínea entre las dos variables. Esto nos indica también una limitación importante del coeficiente de correlación lineal. El coeficiente de correlación lineal sólo detecta relaciones lineales entre dos variables. Por tanto, un coeficiente de correlación lineal cercano a cero, indica que no existe relación lineal entre las variables, pero no excluye la posibilidad de que las variables tengan otras relaciones entre sí de carácter no lineal.

Para analizar los valores de los coeficientes de correlación de Pearson que elaboremos, tenemos que tener en cuenta que no tienen una comparación directa entre resultados de estudios diferentes. Sabemos que $r_{XY} = \pm 1$ indica la correlación lineal perfecta en cualquier caso, y que $r_{XY} = 0$ indica la ausencia total de correlación lineal.

Es menos clara la situación, cuando nos encontramos con un valor intermedio cualquiera, por ejemplo, 0,55. No se puede afirmar que ese valor indica correlación alta, o baja. Depende del tipo de datos que estemos analizando. Será baja, si se trata de dos test similares, que estemos aplicando a los mismos sujetos, o si tenemos pocos sujetos. Podría ser muy alta, si se trata de tests bastante diferenciados entre sí, o si tenemos muchos sujetos. Un número grande de sujetos en la muestra pueden tender a bajar el valor de los coeficientes de correlación que se obtienen. Los resultados de otros investigadores, con variables similares y muestras equivalentes, son los que nos sirven de comparación para evaluar los resultados que obtengamos con nuestros datos. El coeficiente de correlación evaluado por nosotros será bajo, si los coeficientes de correlación que obtienen otros investigadores en circunstancias similares, son mucho más altos. Y lo mismo se puede afirmar en la dirección contraria. Si nosotros obtenemos unos coeficientes de correlación mucho mayores que los encontrados por otros investigadores en circunstancias similares, los nuestros serán muy elevados.

Otro aspecto más complejo es fundamentar la causalidad en un coeficiente de correlación. Cuando existe un coeficiente de correlación elevado entre dos variables, no se puede afirmar que una de las variables es causa de la otra. Hay cantidad de variables que evolucionan conjuntamente. El número de televisores y el número de neveras, por ejemplo, en una muestra de ciudades. Las ciudades con más televisores, suelen tener más neveras y las ciudades con menos televisores suelen tener menos neveras. En realidad, existe la variable *nivel de vida de la ciudad*, que lleva a que haya más televisores y neveras, cuando el nivel de vida de la ciudad aumenta.

Otro caso real, que sirve para ver la complejidad de deducir la «causalidad» entre dos variables, cuando el coeficiente de correlación es elevado. Por ejemplo, si se encuentra una alta correlación negativa entre el número de niños por mujer y los años de escolarización de la mujer en distintos países. No se puede afirmar que la causa del tener menos hijos es que la mujer tiene muchos años de escolarización. No tiene que existir una relación causal en la correlación. Puede existir una variable interviniente entre el número de hijos por mujer y los años de escolarización de ella. Una posible variable interviniente podría ser el tener las mujeres una mayor libertad. Hay que ser muy cuidadoso en este aspecto para no afirmar como relaciones causales las relaciones entre variables.

4.5. REGRESIÓN LINEAL

Cuando existe una relación lineal podemos utilizar la denominada ***recta de regresión*** para efectuar pronósticos de los valores de una variable a partir de la otra variable. La ecuación general de una recta es de la forma: $Y = a + bX$, donde «b» es la pendiente y «a» es la ordenada en el origen.

La ecuación de regresión lineal de Y sobre X, es decir, la que sirve para pronosticar las puntuaciones en Y a partir de las puntuaciones en X, es la siguiente:

$$Y_i' = a + bX_i$$

donde:

$$b = \frac{n\sum(XY) - \sum X \sum Y}{n\sum X^2 - \left(\sum X\right)^2}$$

$$a = \bar{Y} - b\bar{X}$$

Aplicando estas fórmulas al cálculo de la recta de regresión con los datos del ejemplo 4.3, nos resulta la tabla y los resultados siguientes:

Niño	***X***	***Y***	***XY***	***X²***
1	4	6	24	16
2	8	4	32	64
3	10	7	70	100
4	12	8	96	144
5	16	10	160	256
Σ	50	35	382	580

$$b = \frac{n\sum(XY) - \sum X \sum Y}{n\sum X^2 - \left(\sum X\right)^2} = \frac{5\times 382 - 50\times 35}{5\times 580 - (50)^2} = \frac{1910 - 1750}{2900 - 2500} = \frac{160}{400} = 0{,}4$$

$$a = \bar{Y} - b\bar{X} = 7 - 0{,}4\times 10 = 3$$

Por tanto la recta de regresión es: $Y' = 3 + 0{,}4X$.

Podemos observar la representación de esta recta de regresión sobre el diagrama de dispersión en la figura 4.2.

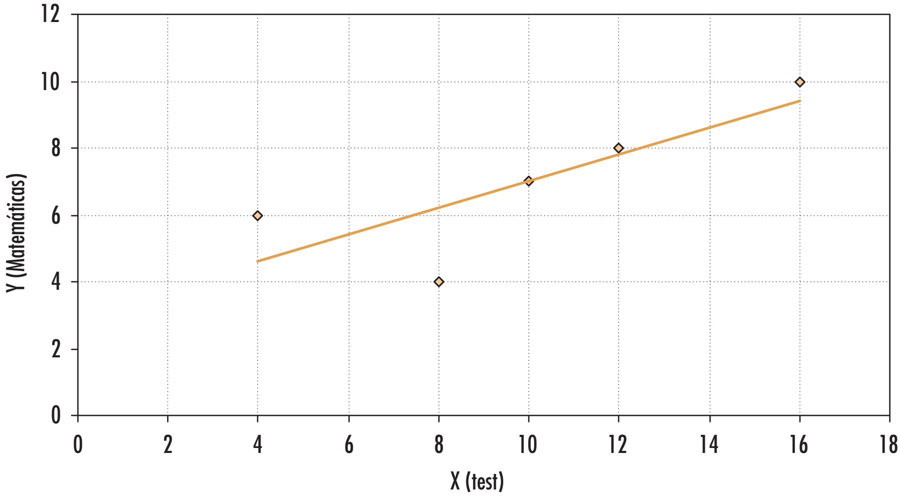


Figura 4.4. La recta de regresión de *Y* sobre *X*, con los datos de la Tabla del ejemplo 4.3.

Esta recta pasa por el punto ($\bar{X}$, $\bar{Y}$), cuyas coordenadas son (10,7). A las puntuaciones, Y_i', obtenidas mediante la recta de regresión las denominamos ***puntuaciones pronosticadas***. A la diferencia entre la puntuación real o verdadera, Y_i, y su pronóstico, Y_i', la llamamos «error» y lo representaremos por E_i.

Vamos a calcular las puntuaciones pronosticadas y los errores en nuestro caso, así como sus medias y sus varianzas, elaborando la tabla 4.12:

Tabla 4.12. Tabla para el cálculo de las estimaciones, *Y'*; errores, *E*, y varianzas de los mismos, S_Y^2 y $S_{Y \cdot X}^2$

Niño	X	Y	Y^2	$Y' = 3 + 0{,}4\,X$	$E = (Y - Y')$	$(Y')^2$	E^2
1	4	6	36	4,6	1,4	21,16	1,96
2	8	4	16	6,2	-2,2	38,44	4,84
3	10	7	49	7,0	0,0	49,00	0,00
4	12	8	64	7,8	0,2	60,84	0,04
5	16	10	100	9,4	0,6	88,36	0,36
Σ	50	35	265	35	0,0	257,8	7,20

Conviene reseñar las siguientes propiedades ejemplificadas con los datos de la tabla 4.12:

1. La media de los errores es 0. $\bar{E} = \dfrac{\sum E_i}{n} = \dfrac{\sum (Y_i - Y_i')}{n} = \dfrac{0}{n} = 0$
2. La media de las puntuaciones pronosticadas coincide con la media de las verdaderas puntuaciones en Y:

$$\bar{Y}' = \frac{\sum Y_i'}{n} = \frac{35}{5} = 7, \quad \bar{Y} = \frac{\sum Y_i}{n} = \frac{35}{5} = 7$$

3. La varianza de las puntuaciones en Y,

$$S_Y^2 = \frac{\sum Y_i^2}{n} - \bar{Y}^2 = \frac{265}{5} - 7^2 = 53 - 49 = 4$$

es igual a la suma de la varianza de los pronósticos,

$$S^2_{Y'} = \frac{\sum (Y'_i)^2}{n} - \left(\bar{Y}'\right)^2 = \frac{257,8}{5} - 7^2 = 51,56 - 49 = 2,56$$

más la varianza de los errores, que representaremos por S^2_E o $S^2_{Y \cdot X}$

$$S^2_{Y \cdot X} = \frac{\sum E^2}{n} - \bar{E}^2 = \frac{7,2}{5} - 0 = 1,44$$

Es decir:

$$S^2_Y = 4 = S^2_{Y'} + S^2_{Y \cdot X} = 2,56 + 1,44$$

Por otro lado se puede comprobar que:

1. La pendiente de la recta de regresión es: $b = r_{XY} \frac{S_Y}{S_X}$. El signo del coeficiente de correlación de Pearson nos dice si la relación lineal entre las variables es directa, o inversa, pues el signo del coeficiente de X en la fórmula de regresión es el mismo del coeficiente de correlación de Pearson. Las desviaciones típicas siempre son positivas.

2. $r^2_{XY} = \frac{S^2_{Y'}}{S^2_Y}$, nos explica que podemos tomar el cuadrado del coeficiente de correlación de Pearson como el tanto por uno de varianza explicada (o proporción de varianza explicada).

3. $1 - r^2_{XY} = \frac{S^2_{Y \cdot X}}{S^2_Y}$, nos explica que podemos tomar el resto a uno del cuadrado del coeficiente de correlación de Pearson como el tanto por uno, o proporción, de la varianza no explicada en la regresión lineal.

4.6. RESUMEN

En este tema se han definido dos formas, o métodos, de analizar la relación entre dos variables.

El primer método está preparado para analizar la relación entre dos variables cualitativas. Como cualquier variable cuantitativa se puede transformar en una variable cualitativa, clasificando sus valores en dos o más grupos de valores consecutivos, este método sirve para analizar la relación entre cualquier par de variables. Para ello hemos explicado las tablas de contingencia, las tablas de diferencias entre las frecuencias empíricas y las teóricas, y el coeficiente de contingencia, C.

En el segundo método, se ha definido una forma de analizar la relación entre dos variables cuantitativas. En esta parte, se han explicado la nube de puntos, el coeficiente de correlación de Pearson, la recta de regresión, y la relación entre el coeficiente de correlación de Pearson y la recta de regresión.

4.7. EJERCICIOS DE AUTOEVALUACIÓN

4.1. Para una muestra de 100 personas hemos obtenido la siguiente tabla:

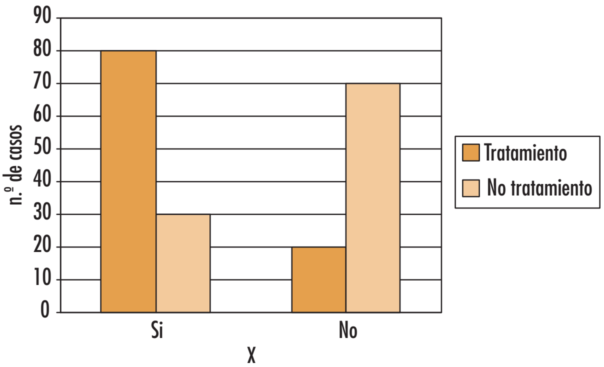
		Y	
		A	B
X	≤50	12	38
	>50	4	46

En ella se recogen los datos de la variable X: Edad, (que se ha dicotomizado en iguales o menores de 50 años y mayores de 50), e Y: estrés (que toma los valores A: no tener estrés, y B: sí tener estrés). Si deseamos conocer si existe relación entre X e Y debemos utilizar: A) la covarianza; B) X^2; C) r_{XY}.

4.2. Para los datos del ejercicio anterior, el valor de X^2 está comprendido entre: A) 0 y 10; B) 10 y 20; C) 20 y 30.

4.3. Para los datos del ejercicio 4.1, el coeficiente de contingencia, C, está comprendido entre: A) 0 y 0,3; B) 0,4 y 0,7; C) 0,8 y 1.

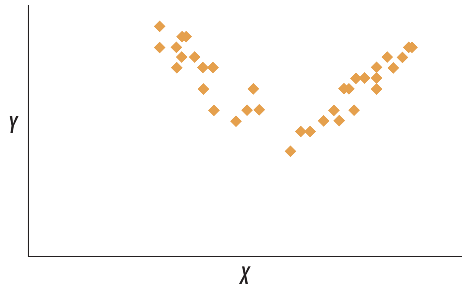
4.4. En la siguiente gráfica, se recogen los datos de un grupo de 200 fumadores (en el que la mitad han sido sometidas a tratamiento para dejar de fumar y la otra no) y su resultado (Sí = han dejado de fumar, No = no han dejado de fumar).



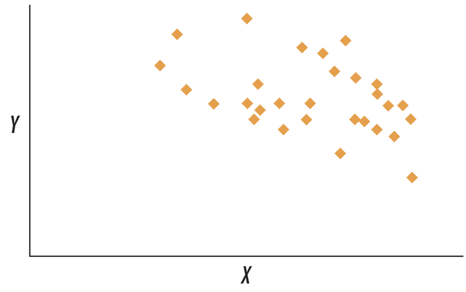
Con estos datos, el coeficiente de contingencia entre las dos variables consideradas está comprendido entre: A) 0 y 0,3; B) 0,4 y 0,7; C) 0,8 y 1.

4.5. Con los datos y el resultado del ejercicio anterior, podemos considerar: A) siendo fumador, no merece la pena someterse al tratamiento; B) no tratarse tiene casi la misma relación con el resultado «dejar de fumar» que el tratamiento; C) existe una relación media-alta entre utilizar el tratamiento y dejar de fumar.

4.6. Con los siguientes diagramas de dispersión,



Gráfica 1



Gráfica 2

correspondientes a dos variables cuantitativas, X e Y, ¿en qué caso debería utilizarse el coeficiente de correlación de Pearson para estudiar la relación entre X e Y?: A) En la Gráfica 1 porque la relación «tiene forma de V»; B) En la gráfica 2 porque la relación es «inversa»; C) En ninguna de las dos.

4.7. En la siguiente tabla se recogen las puntuaciones en dos tests (uno de razonamiento abstracto *X*, y otro de razonamiento espacial, *Y*) de cinco niños. Con estos datos, la covarianza entre *X* e *Y* vale: A) 36; B) 6; C) 63.

Niños	***X***	***Y***
Amaya	92,50	100,50
Carlos	77,50	103,50
Lucía	100,00	105,00
Inés	107,50	106,50
David	122,50	109,50

4.8. Con los datos del ejemplo anterior (4.7), la correlación de Pearson entre *X* e *Y* toma el valor: A) 0,6; B) 0,8; C) 0,4.

4.9. Con los datos del ejercicio 4.7, la pendiente de la ecuación de la recta de regresión que permite pronosticar las puntuaciones en *Y*, *Y'*, a partir de las puntuaciones en *X* vale: A) 2; B) 0,50; C) 0,16.

4.10. La ordenada en el origen de la ecuación de la recta de regresión de *Y* sobre *X* para los datos del ejercicio 4.7 vale: A) 20; B) 60; C) 89.

4.11. La proporción de la varianza de *Y* explicada por la varianza de *X* para los datos del ejercicio 4.7 vale: A) 0,36; B) 0,64; C) 0,80.

4.12. Los datos de la siguiente tabla,

X	***Y***	***XY***
$\sum X = 50000$	$\sum Y = 3500$	$\sum XY = 368000$
$\sum X^2 = 5112500$	$\sum Y^2 = 29000$	

corresponden a las puntuaciones de 500 niños en un test de razonamiento numérico (*X*) y en la asignatura de matemáticas (*Y*). El valor de la covarianza entre *X* e *Y* es: A) 25; B) 36; C) 40.

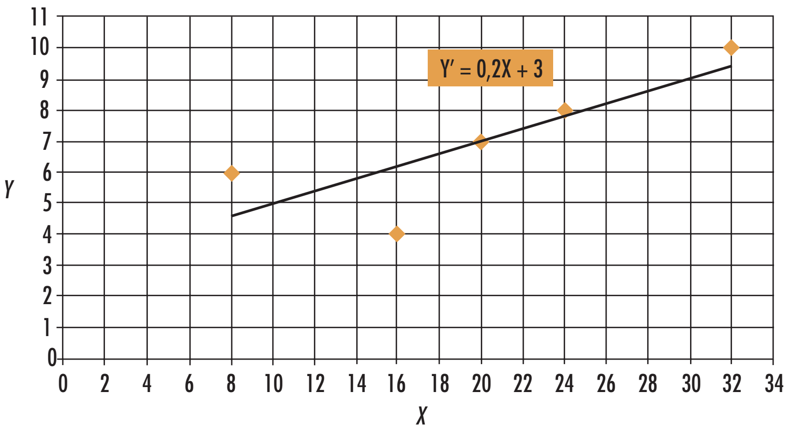
4.13. Con los datos del ejercicio anterior, el coeficiente de correlación de Pearson entre *X* e *Y* es: A) 0,8; B) 0,6; C) 0,9.

4.14. Teniendo en cuenta el resultado obtenido en el ejercicio anterior, podemos decir que a: A) puntuaciones altas en el test se corresponden con puntuaciones bajas en matemáticas; B) puntuaciones bajas en el test se corresponden con puntuaciones bajas en matemáticas; C) puntuaciones bajas en el test se corresponden con puntuaciones altas en matemáticas.

4.15. Con los datos del ejercicio 4.12, la ecuación de la recta de regresión que nos permite pronosticar las puntuaciones en matemáticas a partir de las puntuaciones en el test de razonamiento numérico es: A) $Y' = 0{,}2\,X - 10$; B) $Y' = 0{,}16\,X - 9$; C) $Y' = -0{,}16\,X + 20$.

4.16. A principio de curso, hemos pasado el test de razonamiento numérico a Jaimito y ha obtenido una puntuación de 90. Si este niño reúne las características del grupo utilizado en el ejercicio 4.12, ¿qué puntuación le pronosticaremos en la asignatura de matemáticas a final de curso?: A) 9,0; B) 5,4; C) 5,6.

4.17. En la siguiente gráfica:



se recogen las puntuaciones obtenidas por 5 niños en dos variables, X e Y, y se presenta también la ecuación de regresión de Y sobre X. ¿Cuánto vale la pendiente de la recta de regresión?: A) 3; B) 0,2; C) 3,2.

4.18. Teniendo en cuenta la gráfica del ejercicio anterior (4.17), ¿qué puntuación pronosticamos en Y a un niño que ha tenido en X una puntuación de 20?: A) 10; B) 4; C) 7.

4.19. ¿Cuánto vale la varianza de las puntuaciones pronosticadas para los datos del ejercicio 4.17?: A) 2,56; B) 4,25; C) 5,36.

4.20. ¿Cuánto vale el coeficiente de correlación de Pearson para los datos representados en la gráfica del ejercicio 4.17?: A) 0,6; B) 0,8; C) 0,7.

4.8. SOLUCIONES DE LOS EJERCICIOS DE AUTOEVALUACIÓN

4.1. Solución: B
Al ser X e Y variables cualitativas se aplica X^2.

4.2. Solución: A

Frecuencias empíricas:

		Y		
		A	*B*	
X	≤50	12	38	50
	>50	4	46	50
		16	84	100

Frecuencias teóricas:

		Y		
		A	*B*	
X	≤50	$\frac{16 \cdot 50}{100} = 8$	$\frac{84 \cdot 50}{100} = 42$	50
	>50	$\frac{16 \cdot 50}{100} = 8$	$\frac{84 \cdot 50}{100} = 42$	50
		16	84	100

Diferencias:

		Y		
		A	*B*	
X	≤50	4	–4	0
	>50	–4	4	0
		0	0	0

$$X^2 = \frac{16}{8} + \frac{16}{42} + \frac{16}{8} + \frac{16}{42} = 2 + 0{,}38 + 2 + 0{,}38 = 4{,}76$$

4.3. Solución: A

$$C = \sqrt{\frac{4{,}76}{4{,}76 + 100}} = \sqrt{0{,}045} = 0{,}213$$

4.4. Solución: B

Frecuencias empíricas:

	Sí	No	
Tratamiento	80	20	100
No tratamiento	30	70	100
	110	90	200

Frecuencias teóricas:

	Sí	No	
Tratamiento	55	45	100
No tratamiento	55	45	100
	110	90	200

Diferencias:

	Sí	No	
Tratamiento	25	–25	0
No tratamiento	–25	25	0
	0	0	0

$$X^2 = \frac{625}{55} + \frac{625}{45} + \frac{625}{55} + \frac{625}{45} =$$
$$= 11,364 + 13,889 + 11,364 + 13,889 = 50,506$$

$$C = \sqrt{\frac{50,506}{250,506}} = \sqrt{0,2016} = 0,45$$

4.5. Solución: C

Puesto que $C = 0,45$ y $C_{\max} = \sqrt{\frac{1}{2}} = \sqrt{0,5} = 0,71.$

4.6. Solución: C

No debe utilizarse en ninguno de los dos casos porque no existe relación lineal.

4.7. Solución: A

$$\bar{X} = \frac{500}{5} = 100$$

$$\bar{Y} = \frac{525}{5} = 105$$

Niños	***X***	***Y***	***XY***
Amaya	92,50	100,50	9296,25
Carlos	77,50	103,50	8021,25
Lucía	100,00	105,00	10500,00
Inés	107,50	106,50	11448,75
David	122,50	109,50	13413,75
Σ	**500,00**	**525,00**	**52680,00**

$$S_{XY} = \frac{52680}{5} - (100 \cdot 105) = 10536 - 10500 = 36$$

4.8. Solución: B

Niños	***X***	***Y***	***XY***	***X²***	***Y²***
Amaya	92,50	100,50	9296,25	8556,25	10100,25
Carlos	77,50	103,50	8.021,25	6006,25	10712,25
Lucía	100,00	105,00	10500,00	10000,00	11025,00
Inés	107,50	106,50	11448,75	11556,25	11342,25
David	122,50	109,50	13413,75	15006,25	11990,25
Σ	**500,00**	**525,00**	**52680,00**	**51125,00**	**55170,00**

$$S_X^2 = \frac{51125}{5} - 100^2 = 225 \qquad S_X = \sqrt{225} = 15$$

$$S_Y^2 = \frac{55170}{5} - 105^2 = 9 \qquad S_Y = \sqrt{9} = 3$$

$$r_{XY} = \frac{S_{XY}}{S_X S_Y} = \frac{36}{15 \cdot 3} = \frac{36}{45} = 0{,}8$$

También:

$$r_{XY} = \frac{n\sum XY - \sum X \sum Y}{\sqrt{n\sum X^2 - \left(\sum X\right)^2}\sqrt{n\sum Y^2 - \left(\sum Y\right)^2}} = \frac{5 \cdot 52680 - 500 \cdot 525}{\sqrt{5 \cdot 51125 - 500^2}\sqrt{5 \cdot 55170 - 525^2}} =$$

$$= \frac{900}{\sqrt{5625}\sqrt{225}} = \frac{900}{75 \cdot 15} = \frac{900}{1125} = 0{,}8$$

4.9. Solución: C

$$b = \frac{n\sum (XY) - \sum X \sum Y}{n\sum X^2 - \left(\sum X\right)^2} = \frac{5 \cdot 52680 - 500 \cdot 525}{5 \cdot 51125 - 500^2} = \frac{900}{5625} = 0{,}16$$

También:

$$b = r_{XY}\frac{S_Y}{S_X} = 0{,}8 \cdot \frac{3}{15} = 0{,}16$$

4.10. Solución: C

$$a = \bar{Y} - b\bar{X} = 105 - (0{,}16 \cdot 100) = 105 - 16 = 89$$

4.11. Solución: B

$$\frac{S_{Y'}^2}{S_Y^2} = r_{XY}^2 = 0{,}8^2 = 0{,}64$$

4.12. Solución: B

$$S_{XY} = \frac{\sum XY}{n} - (\bar{X}\bar{Y}) = \frac{368000}{500} - \left(\frac{50000}{500} \cdot \frac{3500}{500}\right) = 736 - (100 \cdot 7) = 736 - 700 = 36$$

4.13. Solución: A

$$S_X^2 = \frac{5112500}{500} - 100^2 = 225 \qquad S_X = \sqrt{225} = 15$$

$$S_Y^2 = \frac{29000}{500} - 7^2 = 9 \qquad S_Y = \sqrt{9} = 3$$

$$r_{XY} = \frac{S_{XY}}{S_X \cdot S_Y} = \frac{36}{15 \cdot 3} = \frac{36}{45} = 0{,}8$$

4.14. Solución: B

Puesto que $r_{XY} = 0{,}8$, puntuaciones altas en X se corresponden con puntuaciones altas en Y y puntuaciones bajas en X se corresponden con puntuaciones bajas en Y.

4.15. Solución: B

$$Y' = bX + a$$

$$b = \frac{n\sum(XY) - \sum X \sum Y}{n\sum X^2 - \left(\sum X\right)^2} = \frac{500 \cdot 368000 - 50000 \cdot 3500}{500 \cdot 5112500 - 50000^2} =$$

$$= \frac{184000000 - 175000000}{2556250000 - 2500000000} = \frac{9000000}{56250000} = 0{,}16$$

También:

$$b = r_{XY}\frac{S_Y}{S_X} = 0,8\cdot\frac{3}{15} = 0,16$$

$$a = \bar{Y} - b\bar{X} = 7 - (0,16\cdot 100) = 7 - 16 = -9$$

Por tanto:

$$Y' = 0,16X - 9$$

4.16. Solución: B

$$Y' = 0,16X - 9 \quad \Rightarrow \quad Y' = 0,16\cdot 90 - 9 = 14,4 - 9 = 5,4$$

4.17. Solución: B

Es el término que multiplica a X en la ecuación que aparece en la gráfica.

4.18. Solución: C

Puede observarse directamente en la gráfica que para X = 20 el pronóstico, utilizando la recta de regresión, es 7. También puede calcularse así:

$$Y' = 0,2X + 3 \quad \Rightarrow \quad Y' = 0,2\cdot 20 + 3 = 4 + 3 = 7$$

4.19. Solución: A

Sujeto	***X***	***Y***	***Y' = 0,2X + 3***	**(Y')²**
1	8	6	4,6	21,16
2	16	4	6,2	38,44
3	20	7	7,0	49,00
4	24	8	7,8	60,84
5	32	10	9,4	88,36
Σ	**100**	**35**	**35**	**257,8**

$$\bar{Y}' = \frac{\sum Y'}{n} = \frac{35}{5} = 7 \qquad \text{(Obsérvese que } \bar{Y}' = \bar{Y}\text{)}$$

$$S^2_{Y'} = \frac{257,8}{5} - 7^2 = 51,56 - 49 = 2,56$$

4.20. Solución: B

Sujeto	***X***	***Y***	***X*²**	***Y*²**
1	8	6	64	36
2	16	4	256	16
3	20	7	400	49
4	24	8	576	64
5	32	10	1.024	100
Σ	**100**	**35**	**2.320**	**265**

$$\bar{X} = \frac{100}{5} = 20 \qquad S_X = \sqrt{\frac{2320}{5} - 20^2} = \sqrt{64} = 8$$

$$\bar{Y} = \frac{35}{5} = 7 \qquad S_Y = \sqrt{\frac{265}{5} - 7^2} = \sqrt{4} = 2$$

$$b = 0,2 = r_{XY} \cdot \frac{S_Y}{S_X} \quad \Rightarrow \quad r_{XY} = \frac{0,2 \cdot S_X}{S_Y} = \frac{0,2 \cdot 8}{2} = \frac{1,6}{2} = 0,8$$

Tema 5

Nociones básicas de probabilidad

5.1. INTRODUCCIÓN

En los temas estudiados hasta ahora, el análisis estadístico se ha limitado a la descripción de un conjunto pequeño de datos denominado muestra. Sin embargo, en cualquier investigación es importante poder generalizar o inferir nuestros resultados a un colectivo mucho más amplio, al que hemos denominado población, y al que no podemos acceder por diferentes motivos (tiempo, economía...). En este caso, la extensión de nuestras conclusiones requiere llevar a cabo una inferencia que siempre será probabilística o formular una hipótesis que aceptaremos o rechazaremos con una determinada probabilidad. Por esta razón es necesario abordar el estudio de la probabilidad.

Tenemos una idea aproximada de lo que significa probabilidad en nuestra vida cotidiana. Así, todos sabemos que es muy poco probable que nos toque un premio de la lotería, que es muy probable que suspendamos una asignatura si hemos dedicado poco tiempo a estudiarla o que, en el nacimiento de nuestro «primer retoño» es casi tan probable que sea niño como que sea niña.

En este tema, y guiados a través de ejemplos concretos, vamos a introducirnos de una forma más rigurosa en el estudio de la probabilidad. Para ello, primero introduciremos unos conceptos fundamentales (experimento aleatorio, suceso...), posteriormente, trataremos de definir el concepto de probabilidad, y finalmente consideraremos el estudio de las probabilidades condicionadas.

Los objetivos que pretendemos son los siguientes:

- Conocer los conceptos de experimento aleatorio y espacio muestral.
- Distinguir los distintos tipos de sucesos que forman parte del espacio muestral y las operaciones fundamentales que con ellos pueden realizarse.
- Adquirir un concepto de probabilidad más preciso.

- Saber resolver aquellos problemas en que se nos presentan probabilidades condicionadas.

5.2. CONCEPTOS PREVIOS

Vamos a iniciar el estudio de la probabilidad recordando algunos conceptos básicos, necesarios para definir el ***concepto de probabilidad***, y vamos a hacerlo con un ejemplo.

> **Ejemplo 5.1.** Imaginemos que lanzamos al aire, una vez, un dado cuyas caras están numeradas del 1 al 6.

Este hecho, el lanzamiento de un dado, constituye un «experimento aleatorio» porque representa un proceso mediante el cual podemos obtener un resultado (experimento) y es aleatorio porque interviene el azar. Todo experimento aleatorio presenta tres características:

— Todos los resultados posibles son conocidos con anterioridad a su realización

— No se puede predecir con certeza el resultado que vamos a obtener

— El experimento puede repetirse, todas las veces que se desee, en idénticas condiciones

Por tanto,

> Un ***experimento aleatorio*** es un proceso, que se puede repetir indefinidamente en las mismas condiciones, cuyo resultado no se puede predecir con certeza.

El conjunto de todos los resultados posibles de un experimento aleatorio se denomina ***espacio muestral*** y se representa, habitualmente, por la letra mayúscula E. Así, el espacio muestral para el lanzamiento del dado es:

E = {⚀, ⚁, ⚂, ⚃, ⚄, ⚅}

A los resultados de un experimento aleatorio, o subconjuntos del espacio muestral, se les denomina ***sucesos*** y se representan por letras mayúsculas: *A*, *B*... Los sucesos, a su vez, pueden ser ***elementales o compuestos***. El suceso simple o elemental consta de un solo resultado del espacio muestral, *E*, mientras que el suceso compuesto consta de dos o más resultados del espacio muestral. Así, por ejemplo, al lanzar un dado, el suceso A = «obtener un cuatro» es elemental mientras que los sucesos B = «obtener un número par» y C = «obtener un múltiplo de 3» son sucesos compuestos:

$$A = \{\text{⚃}\}$$
$$B = \{\text{⚁}, \text{⚃}, \text{⚅}\}$$
$$C = \{\text{⚂}, \text{⚅}\}$$

A todo el espacio muestral de un experimento se le denomina también *suceso seguro* porque siempre ocurre. Al suceso que no puede ocurrir nunca se le denomina ***Suceso Imposible*** y se representa por ϕ ó conjunto vacío.

Vamos a definir ahora algunas operaciones con sucesos, basadas en la correspondencia con la teoría de conjuntos, que nos resultarán útiles más adelante:

- **UNIÓN:**

Llamaremos unión de dos sucesos *A* y *B*, y lo representaremos por $A \cup B$, al subconjunto de *E* formado por los sucesos elementales que pertenecen a *A*, a *B* o a ambos a la vez:

En el ejemplo: $A \cup B = \{\text{⚁}, \text{⚃}, \text{⚅}\}$

- **INTERSECCIÓN:**

Llamaremos intersección de dos sucesos *A* y *B*, y lo representaremos por $A \cap B$, al subconjunto de E formado solamente por los sucesos elementales que pertenecen a *A* y a *B*.

$$A \cap B \ \{\text{⚃}\}$$

Cuando la intersección de dos sucesos no contiene ningún elemento se dice que son incompatibles o excluyentes y, por tanto, no pueden verificarse simultáneamente. Siguiendo con el ejemplo, los sucesos A y C lo son.

$$A \cap C = \{\ \} = \phi$$

■ **COMPLEMENTARIO:**

Llamaremos complementario de un suceso A, y lo representaremos por $\bar{A}$, al subconjunto de E formado por los sucesos elementales que no pertenecen a A:

$$\bar{A} = \{ ⚀, ⚁, ⚂, ⚄, ⚅ \}$$

Para su representación podemos utilizar los ***diagramas de Venn***, ampliamente utilizados en la Teoría de Conjuntos. En la figura 5.1 se representan gráficamente los sucesos anteriores.

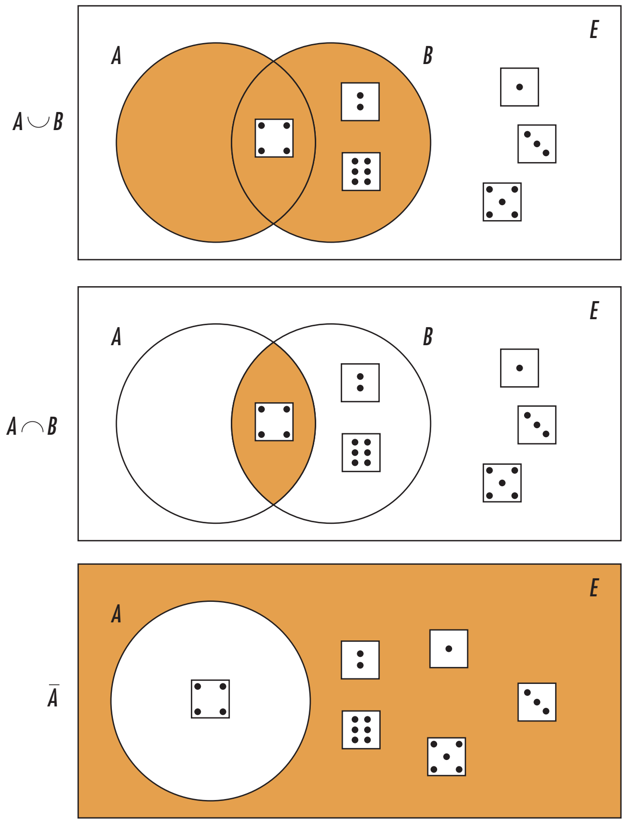


Figura 5.1. Representación gráfica de la unión, intersección y complementario de sucesos.

Las operaciones de unión e intersección pueden extenderse al caso de dos o más sucesos, por ejemplo $A \cup B \cup C$, e igualmente puede hablarse de complementario de la unión o de la intersección de dos sucesos.

5.3. DEFINICIÓN DE PROBABILIDAD

Con los conceptos previos señalados anteriormente, vamos a iniciar el análisis de la probabilidad. Expondremos tres definiciones diferentes del concepto de probabilidad (la definición clásica, la estadística y la axiomática) encaminadas a un mismo fin: calcular la «posibilidad» de ocurrencia de un suceso. Veremos que con cualquiera de estas tres definiciones, la probabilidad se cuantifica con un número comprendido entre cero y uno: cero para el suceso imposible y 1 para el suceso seguro. Cualquier otro suceso tendrá asignado un número entre 0 y 1 en función de la cuantía de su probabilidad de ocurrencia.

La ***definición clásica***, formulada por Laplace, indica que: «*La probabilidad de un suceso es igual al cociente entre el número de casos favorables de que ocurra ese suceso y el número de casos posibles en el supuesto de que todos los casos tengan la misma oportunidad de ocurrir (es decir, sean igualmente probables)*».

Es decir:

$$\text{Probabilidad de un suceso} = \frac{\text{Número de casos favorables}}{\text{Número de casos posibles}}$$

Veamos un ejemplo:

Ejemplo 5.2. Retomando el ejemplo 5.1, lanzamos un dado imparcial una sola vez. ¿Cuál es la probabilidad de obtener un 2? ¿Cuál es la probabilidad de obtener un número par?

Tenemos que calcular la probabilidad de obtener el suceso $A = \{⚁\}$.
El espacio muestral, conjunto de todos los resultados posibles, es:

$$E = \{⚀, ⚁, ⚂, ⚃, ⚄, ⚅\}$$

donde cada uno de los seis resultados tiene la misma probabilidad de salir. Sobre estos 6 casos posibles y equiprobables sólo hay un caso favorable (obtener un dos). Por tanto: $P(A)=\frac{1}{6}$.

Llamemos B al suceso «obtener un número par». La probabilidad de B será igual a:

$$P(B)=\frac{3}{6}=\frac{1}{2}=0{,}5$$

Nótese que tenemos, en este caso, tres resultados favorables sobre el total de seis posibles.

Esta definición de probabilidad, y su aplicación, requiere que los sucesos sean equiprobables (cosa que no siempre ocurre) y, en muchos casos, puede resultar díficil la clasificación de los sucesos como favorables y posibles.

Si repetimos un experimento aleatorio (por ejemplo lanzar un dado al aire) muchas veces, y anotamos las frecuencias relativas de un suceso, podemos observar que tienden a estabilizarse en un valor comprendido entre 0 y 1. Este valor se denomina probabilidad del suceso. Por tanto, podemos definir la probabilidad de un suceso A como:

«el límite al que tiende la frecuencia relativa de aparición de un suceso A cuando el número de ensayos, n, tiende a infinito»:

$$\boxed{P(A)=\lim_{n\to\infty}\frac{n_A}{n}}$$

Esta definición de probabilidad, denominada ***estadística***, aunque es correcta presenta un grave problema: muchas veces no es posible repetir un experimento aleatorio un gran número de veces y, si lo es, no es práctico.

Los graves problemas con las definiciones de probabilidad presentadas (clásica y estadística) llevaron a los matemáticos a establecer una nueva definición, denominada *axiomática*:

Dado un espacio muestral E, *llamamos probabilidad de un suceso A, definido en el espacio muestral E y que designamos por P(A), a un número real que asignamos al suceso A, tal que cumple las siguientes propiedades:*

- $0 \leq P(A) \leq 1$
- $P(E) = 1$
- $P(A) = 1 - P(\bar{A})$

Las dos primeras propiedades indican que la probabilidad es cuantificable numéricamente con un número comprendido entre cero y uno. Como se dijo en la introducción de este tema, asignamos un cero a un suceso que no puede ocurrir nunca y un uno al suceso que se produce con seguridad. La tercera, indica que la probabilidad de un suceso A puede obtenerse también restando de uno la probabilidad de su complementario, $\bar{A}$, puesto que ambos son exhaustivos y mutuamente excluyentes (si no ocurre A necesariamente lo hará su complementario).

A estas propiedades, podemos añadir el denominado «***Teorema de la Suma***»:

Este teorema establece que la probabilidad de que ocurra el suceso A ó el suceso B es igual a la probabilidad de que ocurra A más la probabilidad de que ocurra B menos la probabilidad de que ocurran ambos, A y B. Es decir:

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

Cuando los sucesos A y B son incompatibles, es decir, no pueden ocurrir simultaneamente o la ocurrencia de uno implica la no ocurrencia del otro, la regla de la suma se simplifica a:

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

puesto que su intersección es vacía: $(A \cap B) = \phi$.

Vamos a comprobar que estas propiedades se mantienen utilizando el ejemplo del lanzamiento de un dado que presentamos anteriormente:

Recordemos que en ese caso: E = {⚀, ⚁, ⚂, ⚃, ⚄, ⚅} y que habíamos definido los sucesos: A = «obtener un 4», B = «obtener un número par» y C = «obtener un múltiplo de 3».

El suceso $\bar{A}$ = {⚀, ⚁, ⚂, ⚄, ⚅}

Pues bien, utilizando la definición de probabilidad como el cociente entre casos favorables y casos posibles, tenemos:

$$P(A)=\frac{1}{6}, P(B)=\frac{3}{6}, P(A\cup B)=\frac{3}{6}, P(A\cap B)=\frac{1}{6}, P(\bar{A})=\frac{5}{6}$$

En todos los casos tenemos valores comprendidos entre 0 y 1 y, aplicando la propiedad del complementario:

$$P(A)=1-P(\bar{A})=1-\frac{5}{6}=\frac{1}{6}$$

Aplicando el Teorema de la Suma tenemos:

$$P(A\cup B)=\frac{1}{6}+\frac{3}{6}-\frac{1}{6}=\frac{3}{6}$$

5.4. PROBABILIDAD CONDICIONADA

Hay situaciones, muy frecuentes en la vida cotidiana, donde la aparición de un suceso A depende de la aparición de otro suceso B. Diremos en estos casos que los sucesos A y B son ***dependientes*** porque la probabilidad de A depende, o está condicionada, al suceso B.

La probabilidad de A condicionado a B, o dependiente de la aparición de B, se escribe $P(A|B)$ donde B es la condición requerida.

Consideremos el ejemplo siguiente:

Ejemplo 5.3. En una determinada universidad tenemos la siguiente distribución de sus 5.000 alumnos:

	Medicina (*Md*)	**Enfermería (*E*)**	**Psicología (*Ps*)**	
Varones (**V**)	400	400	1200	2000
Mujeres (**M**)	600	1100	1300	3000
	1000	1500	2500	5000

Si elegimos aleatoriamente uno de ellos,

A) ¿Cuál es la probabilidad de que sea varón?

B) ¿Cuál es la probabilidad de que estudie psicología?

C) ¿Cuál es la probabilidad de que estudie psicología y sea varón?

A partir de la tabla, y considerando la probabilidad como el cociente entre casos favorables y casos posibles, tenemos:

A) $P(V)=\dfrac{2000}{5000}=0,4$

B) $P(Ps)=\dfrac{2500}{5000}=0,5$

C) $P(Ps \cap V)=\dfrac{1200}{5000}=0,24$

Supongamos ahora que hemos elegido al azar un alumno y ha resultado ser varón, ¿cuál es la probabilidad de que estudie psicología?

Nos estamos preguntando por la probabilidad $P\,(Ps|V)$, es decir: supuesto varón (V) ¿cuál es la probabilidad de que estudie psicología (Ps)?

A partir de la tabla, y considerando nuevamente el cociente entre casos favorables y casos posibles tenemos, tenemos que:

$$P(Ps|V)=\frac{1200}{2000}=0,6$$

A partir de los resultados anteriores, podemos comprobar que:

$$P(Ps|V)=\frac{P(Ps \cap V)}{P(V)}=\frac{0,24}{0,4}=0,6$$

De esta forma, hemos expuesto el concepto de probabilidad condicionada que en general, como ya hemos indicado anteriormente, lo escribimos P(A|B) y se lee «probabilidad de *A* condicionada a B». Veamos su definición:

Para dos sucesos cualesquiera *A* y B, **la *probabilidad de A condicionado a B*** *—o de A supuesto B— es igual a la probabilidad de la intersección dividido por la probabilidad de la condición B.* Es decir:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \text{siempre que } P(B) \neq 0$$

De la misma forma:

$$P(B|A) = \frac{P(B \cap A)}{P(A)} \quad \text{siempre que } P(A) \neq 0$$

Nótese que si los sucesos *A* y *B* son *independientes*:

$$P(A|B) = P(A) \text{ y } P(B|A) = P(B)$$

5.5. LA REGLA DEL PRODUCTO Y EL TEOREMA DE BAYES

Hasta ahora, hemos considerado la realización de un único experimento aleatorio o hemos considerado una sola extracción, o ensayo, en un proceso en el que interviene el azar (lanzar un dado al aire una sola vez, extraer o seleccionar una persona dentro de un grupo...). Podemos extender lo dicho hasta ahora al caso en que realizamos varios experimentos simultáneamente (por ejemplo, lanzar un dado y una moneda al aire), repetimos un experimento varias veces (por ejemplo, lanzar una moneda al aire en varias ocasiones o ensayos) o, en general, al caso en que realizamos un proceso varias veces (por ejemplo, extraer de una en una varias bolas de una urna).

Hemos visto, en el epígrafe anterior sobre la probabilidad condicionada, que:

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

Pues bien, si despejamos $P(A \cap B)$ nos quedaría:

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

A esta fórmula se la conoce como «***regla o Teorema del Producto***» y establece lo siguiente:

La probabilidad de ocurrencia de A y B es igual a la probabilidad de ocurrencia de A por la probabilidad de ocurrencia de B, dado que A ha ocurrido previamente. Es decir:

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

donde $P(B|A)$ se lee como «la probabilidad de que ocurra B dado que ha ocurrido A».

Cuando los sucesos A y B son *independientes*:

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

Consideremos el siguiente ejemplo:

Ejemplo 5.4: Supongamos una urna con cinco bolas de las cuales tres son verdes y dos son rojas. Introducimos la mano en la urna y extraemos una bola (primera extracción). Sin devolver la bola que hemos extraído a la urna, volvemos a introducir la mano y extraemos otra bola (segunda extracción); o sea la extracción es «sin reposición».

A) ¿Cuál es la probabilidad de que las dos bolas extraídas sean verdes?

Llamemos V a «bola verde». La probabilidad pedida será igual a «la probabilidad de que en la primera extracción la bola sea verde por la probabilidad de que en la segunda extracción la bola sea verde supuesto que en la primera también lo ha sido».

Para la primera extracción tenemos 5 bolas, 3 verdes y 2 rojas. Por tanto, la probabilidad de que la bola extraída sea verde es 3/5.

Para la segunda extracción sólo disponemos de 4 bolas, 2 verdes y 2 rojas, puesto que una verde ya ha sido extraída en la primera. Ahora la probabilidad de bola verde es 2/4.

Es decir:

$$P(V_1 \cap V_2) = P(V_1) \cdot P(V_2 | V_1) = \frac{3}{5} \cdot \frac{2}{4} = \frac{6}{20} = 0,3$$

donde los subíndices 1 y 2 hacen referencia a la extracción (primera y segunda, respectivamente).

B) ¿Cuál es la probabilidad de que las dos bolas extraídas sean rojas?

Llamando R a «bola roja», y por un razonamiento análogo al anterior:

$$P(R_1 \cap R_2) = P(R_1) \cdot P(R_2 | R_1) = \frac{2}{5} \cdot \frac{1}{4} = \frac{2}{20} = 0,1$$

C) ¿Cuál es la probabilidad de que las bolas sean de distinto color?

La probabilidad de que las bolas sean de distinto color es la probabilidad de que una sea verde y la otra roja (VyR) pero esto puede ocurrir de dos maneras: que la primera sea verde y la segunda roja (V_1R_2) o que la primera sea roja y la segunda verde (R_1V_2). Es decir:

$$P(V_1 \cap R_2) + P(R_1 \cap V_2) = \left[P(V_1) \cdot P(R_2 | V_1)\right] + \left[P(R_1) \cdot P(V_2 | R_1)\right] =$$
$$= \left(\frac{3}{5} \cdot \frac{2}{4}\right) + \left(\frac{2}{5} \cdot \frac{3}{4}\right) = \frac{6}{20} + \frac{6}{20} = \frac{12}{20} = 0,6$$

Observese que la suma de las tres probabilidades anteriormente calculadas vale 1.

En la aplicación de la regla del producto, como ya hemos señalado, debemos tener en cuenta que cuando los sucesos A y B son independientes (la aparición de uno de ellos no depende de la aparición o no del otro), la regla del producto queda reducida a:

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

Ejemplo 5.5. Lanzamos al aire una moneda imparcial en dos ocasiones, ¿cuál es la probabilidad de obtener dos caras?

Si denominamos «C» a «salir cara», la probabilidad de obtener dos caras es:

$$P(C \cap C) = P(C) \cdot P(C) = 0,5 \cdot 0,5 = 0,25$$

puesto que la probabilidad de obtener cara en el segundo lanzamiento no depende del resultado obtenido en el primer lanzamiento (son ensayos independientes).

Consideremos ahora el siguiente ejemplo:

Ejemplo 5.6. En una planta psiquiátrica hay 5 pacientes de los cuales 1 padece psicosis, 2 neurosis y 2 esquizofrenia. Se sabe además que la probabilidad de que un paciente responda favorablemente al tratamiento es 0,6 si padece psicosis, 0,9 si padece neurosis y 0,8 si padece esquizofrenia.

Elegido aleatoriamente un paciente, hemos observado que ha respondido favorablemente al tratamiento ¿cuál es la probabilidad de que ese sujeto padezca neurosis?

Llamemos P a ser psicótico, N a ser neurótico y E a ser esquizofrénico. Llamemos también F a responder favorablemente al tratamiento y $\overline{F}$ a no responder favorablemente Los datos de los que disponemos, o podemos calcular a partir de la información que se nos ofrece, son los siguientes:

$$P(P) = \frac{1}{5} = 0,2 \qquad P(F|P) = 0,6$$

$$P(N) = \frac{2}{5} = 0,4 \qquad P(F|N) = 0,9$$

$$P(E) = \frac{2}{5} = 0,4 \qquad P(F|E) = 0,8$$

y lo que queremos saber es cuánto vale la probabilidad $P(N|F)$

Veamos: Por la definición de probabilidad condicionada:

$$P(N|F) = \frac{P(N \cap F)}{P(F)}$$

y, sabemos que:

$$P(F) = P(P \cap F) + P(N \cap F) + P(E \cap F)$$

Este tipo de problemas puede representarse gráficamente mediante un ***diagrama de árbol*** como el de la figura 5.2. donde los números corresponden a las probabilidades condicionadas al suceso que aparece inmediatamente antes (a la izquierda en el árbol).

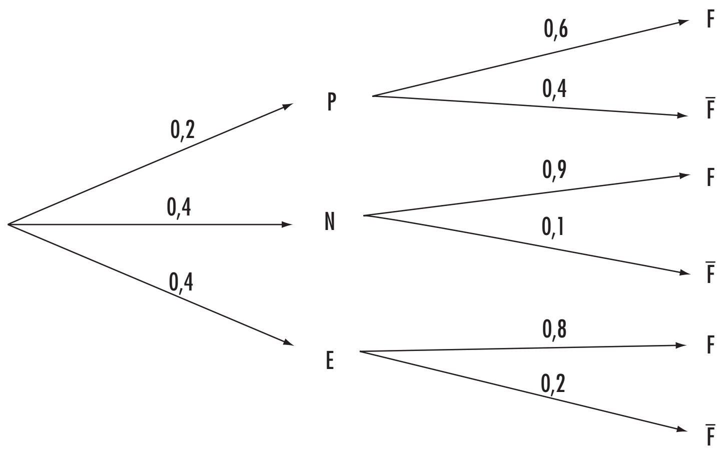


Figura 5.2. Representación gráfica de las probabilidades condicionadas.

Así, por ejemplo, la probabilidad 0,6 corresponde a $P(F|P)$ que es la probabilidad condicionada a responder favorablemente al tratamiento supuesto que es un paciente psicótico, mientras que el 0,2 anterior a él corresponde a la probabilidad de ser psicótico. Se debe cumplir siempre que la suma de las probabilidades que salen del mismo punto deben sumar 1.

Para calcular las probabilidades de intersección de dos sucesos hay que ir multiplicando las probabilidades de cada «rama» hasta que se llegue al extremo del árbol. Por ejemplo, para determinar la probabilidad de «ser psicótico» y «responder favorablemente al tratamiento», $P(F \cap P)$, se multiplica el valor 0,2 por 0,6. Para calcular la probabilidad de «ser neurótico» y «responder favorablemente al tratamiento» $P(N \cap F) = 0{,}4 \cdot 0{,}9 = 0{,}36$, y así sucesivamente.

Por tanto:

$$P(N|F) = \frac{P(N \cap F)}{P(F)} = \frac{P(N) \cdot P(F|N)}{P(F)}$$

que se denomina fórmula o Teorema de Bayes.*Y*, puesto que:

$$\begin{aligned} P(F) &= P(P \cap F) + P(N \cap F) + P(E \cap F) = \\ &= P(P) \cdot P(F|P) + P(N) \cdot P(F|N) + P(E) \cdot P(F|E) = \\ &= 0{,}2 \cdot 0{,}6 + 0{,}4 \cdot 0{,}9 + 0{,}4 \cdot 0{,}8 = 0{,}8 \end{aligned}$$

entonces:

$$P(N|F) = \frac{P(N \cap F)}{P(F)} = \frac{P(N) \cdot P(F|N)}{P(F)} = \frac{0{,}4 \cdot 0{,}9}{0{,}8} = \frac{0{,}36}{0{,}8} = 0{,}45$$

Obsérvese que inicialmente $P(N) = 0{,}4$ y que cuando hemos añadido una nueva información (ha respondido favorablemente al tratamiento) la probabilidad ha subido a 0,45.

La importancia de esta regla no está en su formulación, puesto que se puede obtener a partir de las probabilidades condicionadas de A sobre B y de B sobre A para dos sucesos cualesquiera A y B.

Despejando $P(A \cap B)$ a partir de las fórmulas de probabilidad condicionada

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \Rightarrow P(A \cap B) = P(B) \cdot P(A|B)$$

$$P(B|A) = \frac{P(B \cap A)}{P(A)} \quad \Rightarrow P(B \cap A) = P(A) \cdot P(B|A)$$

Puesto que $P(B \cap A) = P(A \cap B)$, podemos igualar las siguientes expresiones:

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

y, despejando tenemos:

$$P(A|B) = \frac{P(A) \cdot P(B|A)}{P(B)}$$

Por tanto, el ***Teorema de Bayes*** podemos expresarlo de la siguiente manera:

$$\boxed{P(A|B) = \frac{P(A) \cdot P(B|A)}{P(B)}}$$

Su importancia radica en los trabajos que ha generado y en la «corriente» denominada bayesiana, que tiene amplias aplicaciones y cuyo estudio sobrepasa los objetivos de este texto.

5.6. RESUMEN

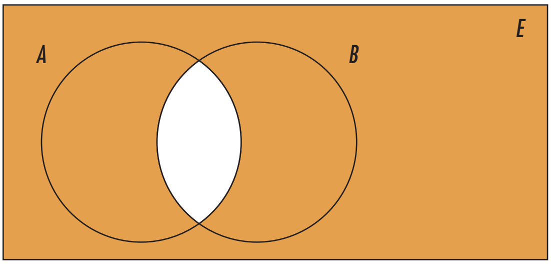
En este tema se han definido los conceptos básicos de experimento aleatorio, suceso y algunas operaciones con sucesos (unión, intersección...). Posteriormente se han dado algunas definiciones de probabilidad (clásica, estadística y axiomática). En cualquier caso, estas definiciones no se excluyen y, de hecho, el cociente entre los casos favorables y los posibles no es más que una frecuencia relativa de aparición de un suceso que se va aproximando a un valor constante a medida que el número de ensayos aumenta (definición estadística). Por otra parte, la probabilidad de un suceso, obtenido por cualquiera de estos dos procedimientos tiene que cumplir los axiomas de la definición formal o axiomática de la probabilidad que hemos expuesto y que se resumen en dos ideas fundamentales: la probabilidad de un suceso es un número comprendido entre 0, para el suceso imposible o conjunto vacío, y 1, para el suceso seguro o espacio muestral, E.

Hemos expuesto el Teorema de la Suma y, posteriormente, hemos definido la probabilidad condicionada: $P(A|B) = \frac{P(A \cap B)}{P(B)}$. Considerando el caso de sucesos independientes hemos visto que la Ley del Producto presenta la forma $P(A \cap B) = P(A) \cdot P(B)$ y hemos finalizado, presentado el teorema de Bayes en un caso concreto.

5.7. EJERCICIOS DE AUTOEVALUACIÓN

5.1. En un experimento aleatorio: A) no conocemos su espacio muestral; B) no interviene el azar; C) no podemos predecir con certeza el resultado que se va a producir

5.2. En la siguiente figura, donde *A* y *B* representan dos sucesos y E el espacio muestral,



la zona coloreada representa: A) la unión de A y B; B) el complementario de la interseción de *A* y *B*; C) el complementario de *B*.

5.3. En la definición clásica, la probabilidad de un suceso es: A) el cociente entre casos favorables y casos posibles; B) la suma entre casos favorables y casos posibles: C) la resta entre casos favorables y casos posibles.

5.4. La frase «En una serie larga de tiradas (o realizaciones de un experimento), la frecuencia relativa observada de un suceso se aproxima a su probabilidad», se corresponde con: A) la definición clásica de la probabilidad: B) la definición estadística de la probabilidad; C) la definición axiomática de la probabilidad.

5.5. Sabiendo que $P(A)=0{,}40$, que $P(B)=0{,}30$ y que $P(A \cap \mathrm{B})=0{,}15$ entonces $P(A \cup B)$ es: A) 0,55; B) 0,85; C) 0,70.

5.6. En un espacio muestral E hay dos sucesos A y B tales que $\mathrm{P}(\bar{A}) = 2/3$; $P(B) = 1/2$; $P(A \cap B) = 1/5$, ¿cuál es la probabilidad de $(A \cup B)$?: A) 13/30; B) 17/30; C) 19/30.

5.7. Si dos sucesos A y B son independientes: A) $P(A \cap B) = P(A) + P(B)$; B) $P(A \cap B) = P(A) - P(B)$; C) $P(A \cap B) = P(A) \cdot P(B)$.

5.8. Lanzamos simultáneamente un dado y una moneda ¿cuál es la probabilidad de obtener un número par en el dado y una cara en la moneda?: A) 0,5; B) 0,25; C) 0,75.

5.9. En la tabla se recoge la composición de un colectivo profesional en función del sexo (varón y mujer) y de si padece (S) o no ($\bar{S}$) algún tipo de estrés.

	S	$\bar{S}$	
V	10	30	40
M	20	40	60
	30	70	100

Donde: V = «varón», M = «mujer»,
S = «estrés» y $\bar{S}$ = «no estrés».

La probabilidad de que escogida una persona al azar padezca estrés vale: A) 0,70; B) 0,30; C) 0,10.

5.10. Con los datos de la tabla del ejercicio anterior, elegida una persona al azar, ¿cuál es la probabilidad de que sea «varón»?: A) 0,10; B) 0,40; C) 0,20.

5.11. Continuando con los datos de la tabla del ejercicio 5,9, elegida una persona al azar ¿cuál es la probabilidad de que «padezca estrés y sea varón»?: A) 0,40; B) 0,70; C) 0,10.

5.12. Con los datos de la tabla del ejercicio 5.9, elegida una persona al azar ha resultado ser varón. La probabilidad de que padezca estrés vale: A) 0,29; B) 0,67; C) 0,25.

5.13. Con los datos de la tabla del ejercicio 5.9, ¿podemos decir que los sucesos «padecer estrés» y «ser varón» son independientes?: A) No; B) Sí; C) No se puede determinar con los datos de la tabla.

5.14. Por la sintomatología se sabe que la probabilidad de contraer una enfermedad *A* en un hospital «afectado» es de 0,4 y la de contraer una enfermedad *B* es de 0,6. Un paciente es sometido a análisis clínico conociéndose que quienes padecen la enfermedad *A* dan resultado positivo (P) con probabilidad 0,90 y quienes padecen la enfermedad *B*, dan resultado positivo (*P*) en el análisis con probabilidad 0,05. Si a un enfermo se le hizo un análisis y el resultado fue positivo ¿cuál es la probabilidad de que padezca la enfermedad *A*?: A) 0,725; B) 0,923; C) 0,532.

5.15. Continuando con el ejercicio anterior, ¿cuál es la probabilidad de que padezca la enfermedad *B* dado que ha sido positivo el análisis?: A) 0,077; B) 0,247; C) 0,532.

5.16. En un experimento de detección de estímulos, presentamos la mitad de veces el estímulo *A* y la otra mitad el estímulo B. El *A* es detectado el 80% de las veces y el *B* el 70 %. En un ensayo determinado sabemos que se ha presentado el estímulo *A*. ¿Cuál es la probabilidad de que no sea detectado?: A) 0,80; B) 0,53; C) 0,20.

5.17. Con los datos del ejercicio anterior, cuando un estímulo no es detectado ¿cuál es la probabilidad de que sea el estímulo B?: A) 0,60; B) 0,30; C) 0,25.

5.18. Con los datos del ejercicio 5.16, ¿cuál es la probabilidad de que un estímulo sea detectado o sea el estímulo *B*?: A) 0,90; B) 0,35; C) 0,70.

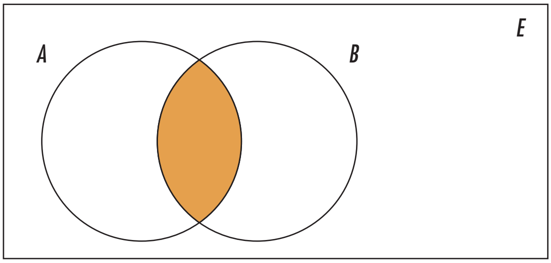
5.19. Sobre 500 alumnos, 100 pertenecen al Plan Antiguo y el resto al Plan Nuevo. Del Plan Nuevo aprueban 240 y del Plan Antiguo aprueban 60. Elegido un alumno al azar, la probabilidad de que haya aprobado es: A) 0,1; B) 0,6; C) 0,4.

5.20. Con los datos del ejercicio 5.19, son independientes los sucesos «aprobar» y «pertenecer al Plan Antiguo»?: A) si; B) no; C) no podemos saberlo.

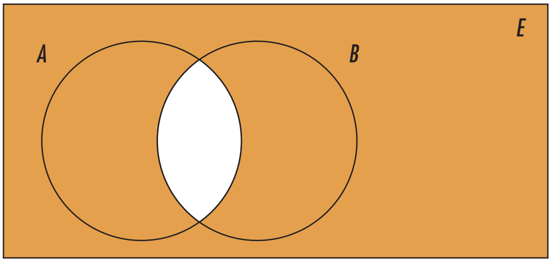
5.8. SOLUCIONES A LOS EJERCICIOS DE AUTOEVALUACIÓN

5.1. Solución: C
Un experimento aleatorio es aquel en que no podemos predecir con certeza el resultado que se va a producir.

5.2. Solución: B
Puesto que $A \cap B$ es:



el complementario es:



5.3. Solución: A
En la definición clásica, la probabilidad de un suceso es el cociente entre casos favorables y casos posibles.

5.4. Solución: B
En la definición estadística, la probabilidad de un suceso A se corresponde con la frecuencia relativa observada cuando el experimento se realiza en un gran número de ocasiones.

5.5. Solución: A

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0{,}40 + 0{,}30 - 0{,}15 = 0{,}55$$

5.6. Solución: C

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = \left[1 - P(\bar{A})\right] + P(B) - P(A \cap B) =$$

$$= (1 - \frac{2}{3}) + \frac{1}{2} - \frac{1}{5} = \frac{1}{3} + \frac{1}{2} - \frac{1}{5} = \frac{10 + 15 - 6}{30} = \frac{19}{30}$$

5.7. Solución: C
Por el teorema del producto, si dos sucesos son independientes la probabilidad de su intersección es igual al producto de sus probabilidades.

5.8. Solución: B
Sea P = «número par» y C = «salir cara»

$$P(P) = \frac{3}{6} = 0,5 \quad P(C) = \frac{1}{2} = ,05$$

$$P(P \cap C) = P(P) \cdot P(C) = 0,5 \cdot 0,5 = 0,25$$

(Son independientes))

5.9. Solución: B

$$P(S) = \frac{30}{100} = 0,30$$

5.10. Solución: B

$$P(V) = \frac{40}{100} = 0,40$$

5.11. Solución: C

$$P(S \cap V) = \frac{10}{100} = 0,10$$

5.12. Solución: C

$$P(S|V) = \frac{P(S \cap V)}{P(V)} = \frac{0,1}{0,4} = 0,25$$

5.13. Solución: A

Para que los sucesos «padecer estrés» (S) y «ser varón» (V) fueran independientes tendría que cumplirse que $P(S \cap V)$ sea igual a $P(S) \cdot P(V)$. Veamos:

$$P(S \cap V) = 0{,}10$$

$$P(S) \cdot P(V) = 0{,}3 \cdot 0{,}4 = 0{,}12$$

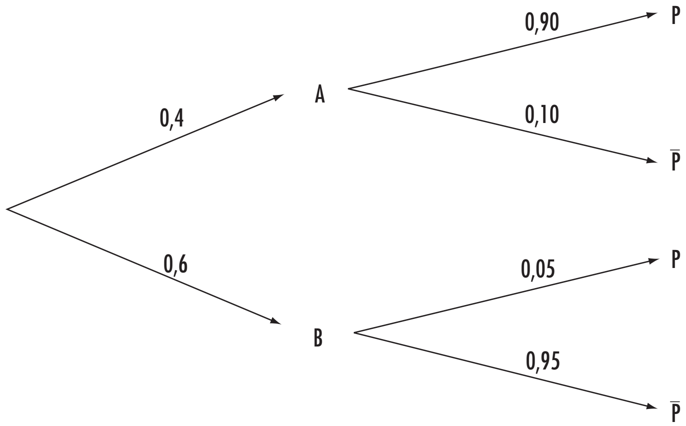
Puesto que $0{,}10 \neq 0{,}12$ no podemos decir que sean independientes.

También:

Si S y V fueran independientes tendría que ocurrir: $P(S|V) = P(S)$. Como podemos comprobar: $0{,}25 \neq 0{,}30$.

5.14. Solución: B

La representación gráfica de los datos de los que disponemos es:



$$P(A|P) = \frac{P(A \cap P)}{P(P)} = \frac{P(A) \cdot P(P|A)}{P(A \cap P) + P(B \cap P)} = \frac{P(A) \cdot P(P|A)}{P(A) \cdot P(P|A) + P(B) \cdot P(P|B)} =$$

$$= \frac{0{,}4 \cdot 0{,}9}{(0{,}4 \cdot 0{,}9) + (0{,}6 \cdot 0{,}05)} = \frac{0{,}36}{0{,}36 + 0{,}03} = \frac{0{,}36}{0{,}39} = 0{,}923$$

5.15. Solución: A

$$P(B|P) = \frac{P(B \cap P)}{P(P)} = \frac{0{,}03}{0{,}39} = 0{,}077$$

También:

$$P(B|P) = 1 - P(A|P) = 1 - 0,923 = 0,077$$

5.16. Solución: C

$$P(A) = 0,5 \quad P(D|A) = 0,80$$
$$P(B) = 0,5 \quad P(D|B) = 0,70$$

$$P(\bar{D}|A) = 1 - P(D|A) = 1 - 0,80 = 0,20$$

5.17. Solución: A

$$P(A) = 0,5 \qquad P(D|A) = 0,8 \Rightarrow P(\bar{D}|A) = 0,2$$
$$P(B) = 0,5 \qquad P(D|B) = 0,7 \Rightarrow P(\bar{D}|B) = 0,3$$

$$P(B|\bar{D}) = \frac{P(B \cap \bar{D})}{P(\bar{D})} = \frac{P(B) \cdot P(\bar{D}|B)}{P(A \cap \bar{D}) + P(B \cap \bar{D})} = \frac{P(B) \cdot P(\bar{D}|B)}{P(A) \cdot P(\bar{D}|A) + P(B) \cdot P(\bar{D}|B)} =$$
$$= \frac{0,5 \cdot 0,3}{0,5 \cdot 0,2 + 0,5 \cdot 0,3} = \frac{0,15}{0,10 + 0,15} = \frac{0,15}{0,25} = 0,60$$

5.18. Solución: A

$$P(D \cup B) = P(D) + P(B) - P(D \cap B) = \left[P(D \cap A) + P(D \cap B)\right] + P(B) - P(D \cap B) =$$
$$= P(D \cap A) + P(B) = P(A) \cdot P(D|A) + P(B) = 0,5 \cdot 0,8 + 0,5 = 0,40 + 0,5 = 0,90$$

5.19. Solución: B
PA = Plan Antiguo
PN = Plan Nuevo

$$P(PA) = 100/500 = 0,2 \qquad P(A|PA) = \frac{60}{100} = 0,6$$

$$P(PN) = 400/500 = 0,8 \qquad P(A|PN) = \frac{240}{400} = 0,6$$

También:

Los datos que tenemos son:

	A	*NA*	
PA	**60**	40	**100**
PN	**240**	160	400
	300	200	**500**

Donde: *PA* = Plan antiguo
PN = Plan nuevo
A = Aprobar
NA = No aprobar

Nota: en la Tabla aparecen en negrita los datos que nos dan en el ejercicio y en «normal» los que se han completado.

A partir de la tabla, tenemos:

$$P(A) = \frac{300}{500} = 0,6$$

5.20. Solución: A

$P(PA) = 0,2$

$P(A|PA) = 0,6$

$P(A) = 0,6$ (ver ejercicio anterior)

$$P(A \cap PA) = P(PA) \cdot P(A|PA) = 0,2 \cdot 0,6 = 0,12$$

$$P(PA) \cdot P(A) = 0,2 \cdot 0,6 = 0,12$$

Si son independientes

También:

Los datos que tenemos son:

	A	*NA*	
PA	**60**	40	**100**
PN	**240**	160	400
	300	200	**500**

Donde: *PA* = Plan antiguo
PN = Plan nuevo
A = Aprobar
NA = No aprobar

Nota: en la tabla aparecen en negrita los datos del enunciado.

A partir de la tabla, tenemos:

$$P(A) = \frac{300}{500} = 0,6$$

$$P(PA) = \frac{100}{500} = 0,2$$

$$P(A \cap PA) = \frac{60}{500} = 0{,}12$$

Puesto que:

$P(A \cap PA) = P(A) \cdot P(PA) = 0{,}6 \cdot 0{,}2 = 0{,}12$, los dos sucesos son independientes.

Tema 6

Distribuciones discretas de probabilidad

6.1. INTRODUCCIÓN

Hemos visto, en el tema anterior, que un experimento cuyo resultado no podemos predecir con certeza se denomina aleatorio. Si el experimento aleatorio se realiza una sola vez se obtendrá un único resultado del espacio muestral, pero a medida que aumenta el número de ensayos irán apareciendo todos los resultados posibles, cada uno de ellos con su correspondiente probabilidad.

Veremos en este tema que, para cada experimento, podemos definir una o varias variables que pueden ser de naturaleza discreta o continua (de acuerdo a los mismos conceptos vistos en el tema 1, referidos a las variables estadísticas) y que denominamos variable, o variables, aleatorias. En este tema nos limitaremos a una variable aleatoria discreta, dejando para el tema siguiente el caso de las variables continuas.

Mostraremos también como, para una variable aleatoria discreta, podemos construir su función de probabilidad y de distribución.

El siguiente paso será describir su función de probabilidad mediante la obtención de unos valores numéricos que representen su tendencia central y su dispersión o variabilidad. Estos conceptos son similares a los ya vistos en los primeros temas cuando se estudiaban variables estadísticas.

Finalmente, dedicaremos especial atención a la distribución binomial que recoge el caso en que una variable aleatoria presenta solamente dos alternativas. Analizaremos sus características fundamentales y veremos como su aplicación a la práctica es muy sencilla utilizando las tablas del Apéndice.

Los objetivos a conseguir con el estudio de este tema son los siguientes:

- Ser capaz de definir correctamente una o más variables aleatorias sobre los resultados de un experimento aleatorio y determinar los valores que toma una determinada variable aleatoria previamente definida.

- Conocer las propiedades que deben cumplir la función de probabilidad y de distribución de una variable aleatoria discreta.
- Obtener la función de probabilidad y la función de distribución de una variable aleatoria discreta y realizar su representación gráfica.
- Calcular e interpretar la media y la varianza de una variable aleatoria discreta.
- Conocer las condiciones de aplicación de la distribución binomial, su media y su varianza.
- Manejar con soltura las tablas de la distribución binomial para la resolución de problemas concretos.

6.2. VARIABLE ALEATORIA: DEFINICIÓN

A los conceptos ya conocidos de espacio muestral y probabilidad estudiados en el tema anterior, vamos a añadir el de «variable aleatoria». Comenzaremos definiendo qué es una variable aleatoria y, a continuación, expondremos algunos ejemplos.

> Una ***variable aleatoria*** es una función que asigna un número real, y sólo uno, a cada uno de los resultados de un experimento aleatorio.

Nótese que sobre un experimento aleatorio, podemos definir una variable de la manera que consideremos oportuna. Así, por ejemplo, sobre el experimento de «lanzar una moneda al aire en tres ocasiones» podemos definir una variable aleatoria como «número de caras obtenidas», como «número de cruces obtenidas», o también como una variable que « toma el valor 1 cuando el número de caras obtenido es mayor que el número de cruces y toma el valor 0 en otro caso». Pues bien, definida la variable y una vez obtenido un resultado en el experimento aleatorio, la función asigna un valor numérico inequívoco a ese resultado. Lo que es aleatorio, en lo que interviene el azar, es el resultado que obtenemos al realizar el experimento aleatorio y no la variable o función.

Las variables aleatorias las vamos a representar por letras mayúsculas de nuestro alfabeto latino y utilizaremos las letras minúsculas, con subíndice, para referirnos a los valores concretos que toman estas variables aleatorias. Así: *X*, *Y*, ... representan variables aleatorias, en tanto que x_1, x_2,.. y_1, y_2... representan los valores concretos que toman esas variables, respectivamente.

Las variables aleatorias, así definidas y representadas, pueden ser ***discretas o continuas***. Cuando la variable aleatoria, *X*, sólo puede tomar un conjunto infinito y numerable de valores (por ejemplo, el conjunto de números naturales) o finito de valores (por ejemplo, la variable *X* definida como «número de caras obtenidas al lanzar una moneda imparcial en tres ocasiones» sólo puede tomar los valores 0, 1, 2 y 3) decimos que es discreta. A los valores concretos que puede tomar una variable aleatoria *X* los designaremos por $x_1, x_2, ..., x_n$ y, en general, por x_i. Si una variable puede tomar infinitos valores (o un conjunto de valores «no numerable») decimos que es continua.

En este tema nos limitaremos exclusivamente al caso «discreto» y dejaremos el estudio de las variables continuas para el tema siguiente.

Ejemplo 6.1. Definimos la variable aleatoria *X* como «el número de caras obtenidas» al lanzar al aire una moneda imparcial en tres ocasiones. En este caso la variable *X* puede tomar los valores 0, 1, 2 y 3. Es decir:

$$X = \begin{cases} 0 & \text{si no sale ninguna cara} \\ 1 & \text{si sale una cara} \\ 2 & \text{si salen dos caras} \\ 3 & \text{si salen las tres caras} \end{cases}$$

Por tanto $x_1 = 0;\ x_2 = 1;\ x_3 = 2\ y\ x_4 = 3.$

6.3. VARIABLES ALEATORIAS DISCRETAS

En el estudio de las variables aleatorias discretas, vamos a prestar especial atención al análisis de sus funciones de probabilidad y de distribución, su media y su varianza. A lo largo de todo el tema vamos a desarrollar un caso muy sencillo, el presentado en el Ejemplo 6.1, para facilitar el cálculo y ayudar a fijar los conceptos fundamentales.

6.3.1. Función de probabilidad

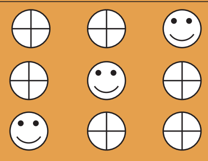
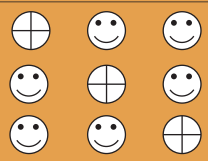
La descripción del comportamiento matemático de una variable aleatoria discreta lo haremos de forma similar a como se hizo, en los temas iniciales del texto, con las variables estadísticas. En ese caso su distribución venía dada por los valores que tomaba la variable y su correspondiente frecuencia relativa o proporción. En el caso de una variable aleatoria discreta X, vendrá dada por los valores que puede tomar la variable (que denominaremos habitualmente por $x_1, x_2, ..., x_n$, una vez ordenados) y su correspondiente probabilidad. Siempre que sea posible, y de aquí en adelante, prescindiremos de los subíndices para una mayor claridad.

Se llama ***función de probabilidad*** de una variable aleatoria discreta, X, y se representa por $f(x)$, a aquella función que asocia a cada valor de la variable la probabilidad de que ésta adopte ese valor. Es decir:

$$f(x) = P(X = x)$$

Veamos algún ejemplo:

Ejemplo 6.2. Consideremos un experimento aleatorio consistente en lanzar una moneda al aire en tres ocasiones. Si definimos una variable aleatoria X como «número de caras obtenidas» obtenemos la siguiente tabla:

E	***x***	***P***
	$x_1 = 0$	1/8 = 0,125
	$x_2 = 1$	3/8 = 0,375
	$x_3 = 2$	3/8 = 0,375
	$x_4 = 3$	1/8 = 0,125

La primera columna recoge el espacio muestral del experimento, la segunda los valores que puede tomar la variable X anteriormente definida y la tercera sus correspondientes probabilidades.

Por tanto, la función de probabilidad de X es:

x	0	1	2	3
***f*(*x*)**	0,125	0,375	0,375	0,125

o, también:

x	***f*(*x*)**
0	0,125
1	0,375
2	0,375
3	0,125

La función de probabilidad de una variable aleatoria discreta puede representarse mediante un diagrama de barras donde se recogen los valores que toma la variable en el eje de abscisas y, en el eje de ordenadas, las correspondientes probabilidades. En la figura. 6.1 se recoge la representación gráfica de la función de probabilidad correspondiente al ejemplo 6.2.

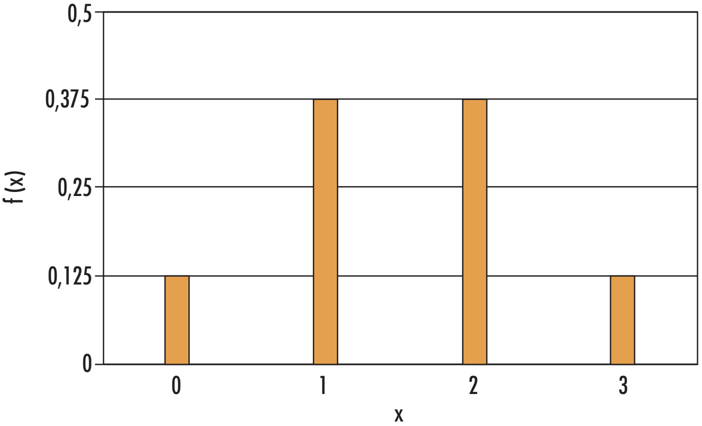


Figura 6.1. Representación gráfica de la función de probabilidad del ejemplo 6.2.

Las dos propiedades fundamentales que debe cumplir la función de probabilidad son:

- Para cualquier valor de x, siempre toma valores positivos o nulos, es decir:

$$\forall x \in X \;\; f(x) \geq 0$$

- La suma de todas las probabilidades correspondientes a cada valor de x es igual a uno:

$$\sum f(x) = f(x_1) + f(x_2) + \ldots + f(x_n) = 1$$

Observará, también, que estas propiedades no son más que una adaptación de la definición axiomática de la probabilidad al caso de variables aleatorias.

6.3.2. FUNCIÓN DE DISTRIBUCIÓN

La función de distribución, o función de distribución de probabilidad de una variable aleatoria X, se representa con la misma letra que su función de probabilidad, pero en mayúscula: $F(x)$, y nos indica cuál es la probabilidad de que la variable aleatoria tome un valor menor o igual que un valor concreto x. Su definición es la siguiente:

Se llama ***función de distribución*** de una variable aleatoria discreta, X, y se representa por $F(x)$, a aquella función que asocia a cada valor de la variable la probabilidad de que ésta adopte ese valor o cualquier otro inferior. Es decir:

$$F(x) = P(X \leq x)$$

Dicho de otra manera: si ordenamos de menor a mayor los valores x de la variable aleatoria discreta, la función de distribución se obtiene acumulando (o sumando) los valores de la función de probabilidad, de forma que aplicaremos la siguiente expresión:

$$F(x) = P(X \leq x) = f(x_1) + f(x_2) + ... + f(x)$$

Ejemplo 6.3. Con los mismos datos del ejemplo 6.1, donde el experimento aleatorio consistía en lanzar una moneda al aire en tres ocasiones y hemos definido la variable X como «número de caras» vamos a calcular su función de distribución.

Calcularemos $F(0)$, $F(1)$, $F(2)$ y $F(3)$. Comenzamos por $F(0)$ que es la probabilidad de que la variable aleatoria X: «número de caras» tome un valor menor o igual a cero, esto es:

$$F(0) = P(X \leq 0) = P(X = 0) = 0{,}125$$

De forma similar, F(1) es la probabilidad de que el «número de caras» sea menor o igual a 1, esto es que sea cero y uno:

$$F(1) = P(X \leq 1) = P(X = 0) + P(X = 1) = f(0) + f(1) = 0{,}125 + 0{,}375 = 0{,}50$$

Para calcular la probabilidad de que el «número de caras» sea menor o igual a dos:

$$F(2) = P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2) = f(0) + f(1) + f(2) =$$
$$= 0{,}125 + 0{,}375 + 0{,}375 = 0{,}875$$

Finalmente $F(3)$, o la probabilidad de que el número de caras sea menor o igual que tres, es:

$$F(3) = P(X \leq 3) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) =$$
$$= f(0) + f(1) + f(2) + f(3) = 0{,}125 + 0{,}375 + 0{,}375 + 0{,}125 = 1$$

Los valores obtenidos se suelen presentar resumidos en una tabla como la siguiente:

x	0	1	2	3
***F*(*x*)**	0,125	0,500	0,875	1

O, de forma similar a las variables estadísticas:

x	***F*(*x*)**
3	1
2	0,875
1	0,500
0	0,125

La representación gráfica de la función de distribución anterior aparece recogida en la figura 6.2.

Nótese que $F(x)$ va «dando saltos» precisamente en los valores de la variable (0, 1, 2 y 3) y el círculo «blanco», de la gráfica, no incluye esos valores. Así, por ejemplo $F(2) = 0{,}875$ pero $F(1{,}9999...) = F(1) = 0{,}5$.

Observando la gráfica de la figura 6.2 se deducen, sin necesidad de recurrir a demostraciones matemáticas, las propiedades fundamentales que debe cumplir la función de distribución de probabilidad. Éstas son:

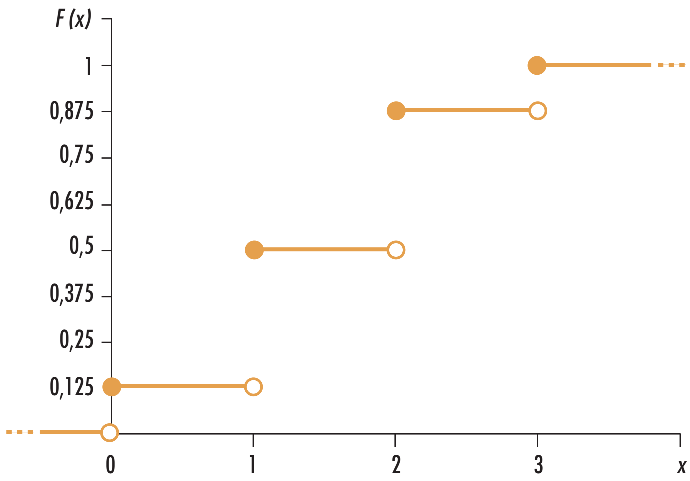


Figura 6.2. Representación gráfica de la Función de Distribución del ejemplo 6.3.

- Todos los valores que toma la función de distribución de probabilidad son positivos o nulos, es decir:

$$\forall x \qquad F(x) \geq 0$$

- $F(x)$ es nula, vale 0, para todo valor inferior al menor valor de la variable aleatoria, x_1:

$$F(x) = 0 \;\text{ si }\; x < x_1$$

- $F(x)$ es igual a uno para todo valor igual o superior al mayor valor de la variable aleatoria. Si llamamos «x_k» al mayor valor de la variable:

$$F(x) = 1 \;\text{ si }\; x \geq x_k$$

- La función $F(x)$ es no decreciente ya que es una acumulación o suma de probabilidades que son siempre positivas o nulas.
- La probabilidad, P, de que la variable aleatoria X tome valores x superiores a x_1 e inferiores o iguales a x_2 ($x_1 < x \leq x_2$) es la diferencia entre los valores de la función de distribución correspondientes a su valor superior menos su valor inferior. Es decir:

$$P(x_1 < x \leq x_2)= F(x_2) - F(x_1)$$

6.3.3. Media y varianza de una variable aleatoria

En los primeros temas aprendimos a describir una distribución de frecuencias de una variable estadística mediante los índices de tendencia central y de dispersión. Lo mismo podemos hacer ahora con una variable aleatoria: calcular su media y su varianza.

Recuerde el lector que para una variable discreta X podíamos calcular su media sencillamente obteniendo el sumatorio del producto de cada uno de los valores de la variable por su frecuencia relativa o proporción. Pues bien, para calcular la media, que designaremos por la letra griega «μ», de una variable aleatoria discreta X calcularemos el sumatorio de cada uno de los valores que toma la variable por su correspondiente función de probabilidad. Es decir:

La ***media***, μ, de una variable aleatoria discreta X viene definida por la siguiente expresión:

$$\mu = \sum x \cdot f(x)$$

La media de una variable X, también se denomina ***esperanza matemática o valor esperado*** de X y se representa por $E(X)$. Este término tiene sus raíces en los juegos de azar y fue introducido con el fin de poder estimar las ganancias esperadas, si se repitiese el juego un elevado número de veces. Referido a una variable aleatoria representa el promedio teórico que tomaría la variable aleatoria si se repitiese el experimento aleatorio infinitas veces.

Ejemplo 6.4. La media o esperanza matemática de la variable X que venimos considerando desde el ejemplo 6.1. es:

x	$f(x)$	$x \cdot f(x)$
0	0,125	0
1	0,375	0,375
2	0,375	0,750
3	0,125	0,375
		1,5

$$\mu = E(X) = \sum xf(x) = x_1 f(x_1) + x_2 f(x_2) + x_3 f(x_3) + x_4 f(x_4) =$$
$$= 0 \cdot 0{,}125 + 1 \cdot 0{,}375 + 2 \cdot 0{,}375 + 3 \cdot 0{,}125 = 1{,}5$$

Para obtener la varianza de una variable aleatoria X, que designaremos por σ^2 (letra griega «sigma» elevada al cuadrado) ó $V(X)$, debemos calcular el sumatorio del producto de cada uno de los valores que toma la variable menos su media elevados al cuadrado por su correspondiente valor de la función de probabilidad. Es decir:

La ***varianza*** $\boldsymbol{\sigma^2}$, de una variable aleatoria discreta X viene definida por la siguiente expresión:

$$\sigma^2 = \sum (x - \mu)^2 \cdot f(x)$$

Nota: Una fórmula alternativa, que puede resultar muy útil en diversas ocasiones, para calcular la varianza es:

$$\sigma^2 = E(X^2) - \left[E(X)\right]^2$$

donde $E(X^2) = \sum x^2 f\left(x\right)$ y $\left[E(X)\right]^2$ es la media de la variable elevada al cuadrado. Por tanto, la varianza puede definirse también como la esperanza de los cuadrados de X, $E(X^2)$, menos el cuadrado de la esperanza de X, $[E(X)]^2$.

De manera análoga a las variables estadísticas, la desviación típica de una variable aleatoria será la raíz cuadrada de la varianza. Es decir: $\sigma = \sqrt{\sigma^2}$.

Ejemplo 6.5. Continuando con el ejemplo 6.1, vamos a calcular su varianza (sabiendo que μ = 1,5). Para ello completamos la siguiente tabla:

x	$f(x)$	$(x-\mu)$	$(x-\mu)^2$	$(x-\mu)^2 \cdot f(x)$	x^2	$x^2 \cdot f(x)$
0	0,125	–1,5	2,25	0,28125	0	0
1	0,375	–0,5	0,25	0,09375	1	0,375
2	0,375	0,5	0,25	0,09375	4	1,500
3	0,125	1,5	2,25	0,28125	9	1,125
				0,75		3

Por tanto:

$$\sigma^2 = \sum (x-\mu)^2 \cdot f(x) = 0,75$$

y, también:

$$\sigma^2 = E\left(X^2\right) - \left[E\left(X\right)\right]^2 = 3 - (1,5)^2 = 3 - 2,25 = 0,75$$

6.4. DISTRIBUCIONES DISCRETAS DE PROBABILIDAD

Existen algunas distribuciones discretas que, por utilizarse frecuentemente como modelo o por su interés como instrumento estadístico, son especialmente importantes. De muchas de ellas se han elaborado una serie de tablas que facilitan su aplicación a problemas concretos.

En Ciencias Sociales y de la Salud se trabaja, en muchas ocasiones, con variables aleatorias discretas que sólo pueden tomar dos valores (dicotómicas) y que habitualmente representaremos por 1 y 0. En estos casos, resulta muy útil la utilización de la ***distribución binomial*** que analizamos en el siguiente epígrafe.

6.4.1. La distribución binomial

La realización de un experimento aleatorio, como lanzar una moneda al aire, admite sólo dos resultados posibles. En este caso concreto los resultados posibles son «cara» o «cruz». Se trata de un experimento o ensayo denominado Bernouilli, por ser este autor uno de los pioneros en su estudio. El nacimiento de varón o mujer, el acierto o fallo a una pregunta con dos alternativas respondida al azar, el lado izquierdo o derecho de un laberinto en forma de *T* elegido por una rata no entrenada en el laboratorio... son algunos de los múltiples ejemplos en que sólo se presentan dos alternativas posibles. Pues bien, a una de ellas la denominaremos «éxito ó acierto» (que, habitualmente, codificaremos con «1») y a la otra «fracaso o error» (que codificaremos como «0»), sin que estos términos tengan connotaciones ni positivas ni negativas, respectivamente.

Un experimento binomial consiste en repetir «n» veces, y de forma independiente, un ensayo Bernouilli. Una variable aleatoria X sigue una distribución binomial (con parámetros n y p) si expresa el número de éxitos en «n» realizaciones independientes de un experimento con probabilidad «p» de obtener «éxito» y, por tanto, $(1 - p)$ de obtener fracaso. Esta distribución suele representarse por la expresión $B(n, p)$ donde B indica «binomial», n (número de ensayos o veces que se repite un experimento Bernouilli) y p (probabilidad de «éxito»).

Ejemplo 6.7. Si lanzamos una moneda imparcial al aire en tres ocasiones y definimos la variable aleatoria X como «número de caras obtenidas», ésta variable seguirá el modelo de distribución binomial con parámetros $n = 3$ y $p = 0{,}5$. Diremos que X sigue un modelo $B(3; 0{,}5)$. Esto es así porque en cada lanzamiento sólo son posibles dos resultados: «éxito» (salir cara) y «fracaso» (salir cruz); los ensayos son independientes entre sí (el resultado en un ensayo no depende de lo que haya salido o no en los ensayos anteriores) y la probabilidad de «éxito» (en este caso «salir cara») se mantiene constante a lo largo de los ensayos (en este caso $p = 0{,}5$).

Pues bien, una variable X que sigue un modelo de distribución binomial, con parámetros «n» y «p», y que simbolizamos por $X \to B(n, p)$, presenta las características fundamentales recogidas en el siguiente recuadro:

Características fundamentales de una distribución B(n,p)

- Función de probabilidad:

$$f(x) = P(X = x) = \binom{n}{x} p^x q^{n-x}$$

- Función de distribución:

$$F(x) = P(X \leq x) = \sum \binom{n}{x} p^x q^{n-x}$$

- Media: $\mu = np$
- Varianza: $\sigma^2 = npq$

donde: «x» es el «número de aciertos», «n» es el número de ensayos, «p» es la probabilidad de éxito en cada uno de los ensayos, «q» es la probabilidad de fracaso $(1-p)$ y el número combinatorio $\binom{n}{x}$, que se lee «n sobre x» es igual a $\dfrac{n!}{x!\,(n-x)!}$.

Aunque es relativamente fácil deducir las características anteriores no lo vamos a hacer aquí formalmente, sino que recurriremos a su aplicación en ejemplos concretos.

Ejemplo 6.8. Siguiendo con el experimento aleatorio de lanzar una moneda en tres ocasiones, presentado en el ejemplo 6.1, y definida X como «número de caras» se pregunta: A) ¿cuál es la probabilidad de obtener exactamente 2 caras?; B) ¿cuál es la probabilidad de obtener dos caras o menos? y C) ¿cuál es la probabilidad de obtener más de dos caras?

Como se ha indicado en el Tema 5, podemos responder a estas preguntas desarrollando el espacio muestral y aplicando, en cada caso, la conocida fórmula de Laplace (cociente entre casos favorables y casos posibles) pero podemos hacerlo también recurriendo a la función de probabilidad y de distribución binomial. Veamos:

A)

$$f(2) = P(X=2) = \binom{3}{2} \cdot 0,5^2 \cdot 0,5^{3-2} = \binom{3}{2} \cdot 0,5^2 \cdot 0,5 = \left(\frac{3!}{2! \cdot 1!}\right) \cdot 0,5^2 \cdot 0,5 =$$
$$= 3 \cdot 0,5^2 \cdot 0,5 = 3 \cdot 0,25 \cdot 0,5 = 0,375$$

B)

$$F(2) = P(X \leq 2) = f(0) + f(1) + f(2) = 0,125 + 0,375 + 0,375 = 0,875$$

puesto que:

$$f(0) = P(X=0) = \binom{3}{0} \cdot 0,5^0 \cdot 0,5^{3-0} = \binom{3}{0} \cdot 1 \cdot 0,5^3 = \left(\frac{3!}{0! \cdot 3!}\right) \cdot 1 \cdot 0,5^3 =$$
$$= 1 \cdot 1 \cdot 0,125 = 0,125$$

$$f(1) = P(X=1) = \binom{3}{1} \cdot 0,5^1 \cdot 0,5^{3-1} = \binom{3}{1} \cdot 0,5 \cdot 0,5^2 = \left(\frac{3!}{1! \cdot 2!}\right) \cdot 0,5 \cdot 0,5^2 =$$
$$= 3 \cdot 0,5 \cdot 0,25 = 0,375$$

$f(2) = P(X = 2) = 0,375$ (Véase el apartado A).

C)

$$P(X > 2) = 1 - P(X \leq 2) = 1 - F(2) = 1 - 0,875 = 0,125$$

Puesto que ya lo hemos calculado en el apartado B)

Puede observarse también que la media y la varianza coinciden con la calculada en los Ejemplos 6.4 y 6.5, respectivamente:

$$\mu = np = 3 \cdot 0,5 = 1,5$$
$$\sigma^2 = npq = 3 \cdot 0,5 \cdot 0,5 = 0,75$$

La utilización de las funciones de probabilidad y de distribución requieren cálculos tediosos. Las tablas de la función de probabilidad y de la función de distribución binomial, tablas I y II que se incluyen en el Apéndice al final del texto, nos evitan, en muchos casos, el cálculo de las probabilidades a partir de la ecuación de esas funciones, facilitando considerablemente su obtención cuando tenemos un elevado número de ensayos («*n*»).

En la Tabla I, para la función de probabilidad binomial, la primera columna encabezada con la letra «*n*» se refiere al número de ensayos e incluye los valores desde 1 hasta 20. La segunda columna recoge el número de éxitos («*x*») que esperamos obtener para ese número de ensayos y que abarcan desde 0 hasta ese número de ensayos. La primera fila de la tabla recoge algunos valores de la probabilidad de «éxito» («*p*») que van desde 0,01 a 0,5. En el interior de la tabla se encuentran las probabilidades correspondientes. La probabilidad buscada, para unos valores concretos de «n» y «x», se encuentra en la intersección de su fila con la correspondiente columna de «p». Así, por ejemplo, la probabilidad de obtener dos éxitos en tres ensayos con una probabilidad de éxito de 0,3 se encuentra en la tabla en la posición que se recoge en la figura 6.3 y vale 0,1890.

		Probabilidad de éxito "p"							
n	x	0,01	0,05	0,10	.....	0,30		0,45	0,50
1	0	....	....	......	....		....	......	......
1	1	....	....	......	....		....	.....	......
2	0	....	.....	....	......		.....	......	....
..	...	....	.....	.....	......		.....	.....	....
3	1	....	.....	.....	......	......	......	......	....
3	2	....	.....	.....	.....	0,1890	......	0,3341	0,3750
..	..	....	.....	.....	.....	......	......	......	......

Figura 6.3. Obtención de las probabilidades a partir de la tabla de la función de probabilidad binomial.

La utilización de la tabla II, función de distribución binomial, es idéntica a la anterior. Unicamente resaltar que las probabilidades que aparecen en el interior de la tabla son acumuladas. Veamos un ejemplo de la utilización de estas dos tablas.

Ejemplo 6.9. Con los mismos datos del ejemplo anterior:

A)

$f(2) = P(X = 2) = 0{,}375$ Utilizando la tabla I y recogiendo el valor que aparece en la intersección de la fila $n = 3$ $x = 2$ con la columna $p = 0{,}5$.

B)

$F(2) = P(X \leq 2) = 0{,}875$ Utilizando la tabla II y recogiendo el valor que aparece en la intersección de la fila $n = 3$ $x = 2$ con la columna $p = 0{,}5$.

C)

$P(X > 2) = 1 - P(X \leq 2) = 1 - F(2) = 1 - 0{,}875 = 0{,}125$

Puesto que $F(2)$ ya lo hemos obtenido en el apartado anterior utilizando la tabla II.

Sin embargo las tablas I y II sólo contienen valores de p desde 0,1 hasta 0,5, entonces ¿qué hacer cuando tengamos una $p > 0{,}5$? En casos como éste, lo que haremos será «intercambiar» las condiciones de «éxito» y «fracaso». Veámoslo con un ejemplo.

Ejemplo 6.10. Sabemos, por la experiencia de años anteriores en el Servicio de Psiquiatria y Psicología Clínica, que un 60% de los pacientes son tratados con Técnicas de Modificación de Conducta. Si un determinado día acuden 5 personas a consulta ¿Cuál es la probabilidad de que tres sean tratadas con Técnicas de Modificación de Conducta?

En este caso, si la probabilidad de ser tratado con Técnicas de Modificación de Conducta es $p = 0.6$, la probabilidad de no ser tratado con tales técnicas es $q = 1 - p = 0{,}4$. Por otro lado, que tres personas, de un total de cinco, sean tratados con Técnicas de Modificación de Conducta, es lo mismo que dos personas, de las cinco, no sean tratados con tales técnicas. Por tanto, el valor correspondiente, en la tabla I, a la intersección de la fila $n = 5$ y $x = 2$ con la columna $p = 0{,}4$ nos dará respuesta a la pregunta planteada. El resultado es 0,3456.

Finalmente podemos observar que, en las tablas I y II, el número de ensayos «n» sólo llega hasta 20. Este hecho no plantea ningún problema porque para valores superiores a ese podemos hacer una aproximación de la Binomial a la distribución Normal, como veremos en el próximo tema.

6.4.2. Otras distribuciones discretas

Hemos prestado, en las páginas anteriores, especial atención a la distribución Binomial por su amplia utilización en distintos ámbitos de las Ciencias Sociales y de la Salud pero existen otros muchos modelos de distribución para variables aleatorias discretas. El modelo de Poisson o de «los sucesos raros» se utiliza, bajo las mismas condiciones de la binomial, para variables dicotómicas pero con un elevado número de ensayos y un valor de «p» muy pequeño. La distribución multinomial se utiliza para ensayos que ofrecen más de dos resultados posibles y, en cierto sentido, supone una generalización de la binomial o ésta puede considerarse un caso particular de áquella. No desarrollaremos ninguno de estos modelos y dejamos abierta la posibilidad, al lector interesado, de consultar la bibliografía.

6.5. RESUMEN

En este tema hemos introducido el concepto de variable aleatoria, hemos distinguido entre variables aleatorias discretas y continuas y hemos establecido el paralelismo entre la función de probabilidad de una variable aleatoria discreta y la distribución de proporciones (o frecuencias relativas) de una variable estadística. El mismo paralelismo se produce entre la tabla de la función de distribución y la tabla de proporciones acumuladas.

Hemos estudiado la función de probabilidad de una variable aleatoria discreta y la hemos descrito haciendo uso de su media y su varianza. Finalmente, hemos comprobado la sencillez de manejo de las tablas de la distribución binomial y la utilidad para resolver los problemas planteados.

6.6. EJERCICIOS DE AUTOEVALUACIÓN

6.1. La expresión, $f(x)$, en el contexto de las variables aleatorias discretas, representa: A) la probabilidad de que la variable aleatoria X tome un valor menor o igual que x; B) la probabilidad de que la variable aleatoria X tome un valor concreto, x; C) la probabilidad de que la variable aleatoria X tome un valor menor que x.

6.2. ¿Cuál de las siguientes afirmaciones es una propiedad básica de toda función de probabilidad de una variable aleatoria X discreta?: A) para cualquier valor de la variable aleatoria, su función de probabilidad puede tomar valores negativos; B) la función de probabilidad es, siempre, no decreciente; C) para cualquier valor de la variable aleatoria x, la función de probabilidad siempre toma valores positivos o nulos

6.3. En la siguiente tabla:

x	1	2	3	4	5
$f(x)$	0	10/60	24/60	20/60	4/60

se muestra la función asignada a una variable aleatoria discreta X. La función: A) es una función de probabilidad porque $f(x) \geq 0$; B) no es una función de probabilidad porque $f(1)$ es nula; C) no es una función de probabilidad porque no cumple alguna de las propiedades fundamentales.

6.4. Para el diseño de un experimento de discriminación disponemos de tres cuadros grises y dos azules. Seleccionamos de forma sucesiva y sin reposición dos de estos cinco estímulos y definimos la variable aleatoria X: «número de estímulos grises seleccionados». La función de probabilidad de esta variable aleatoria es:

A)

x	1	1	2
$f(x)$	1/3	1/3	1/3

B)

x	0	1	2	3
$f(x)$	0,2	0,3	0,3	0,2

C)

x	0	1	2
$f(x)$	0,1	0,6	0,3

6.5. En un experimento aleatorio cualquiera para denotar la expresión «la probabilidad de que una variable aleatoria, X, tome valores menores o iguales que 4» utilizamos: A) $f(4)$; B) $F(4)$; $P(4)$.

6.6. Los valores de una variable aleatoria discreta X son 0, 1, 2, 3, 4 y 5. Si se sabe que $P(X \leq 4) = 0{,}974$ y que $P(X \leq 3) = 0{,}963$, entonces $P(X = 4)$ será: A) 0,011; B) 0,022; C) 0,001.

6.7. Una urna contiene dos bolas negras y dos blancas. Se extraen dos bolas, una a una, con reposición. Sea la variable aleatoria X: «Número de bolas blancas extraídas». La función de distribución de esta variable para $x = 0$, $x = 1$ y $x = 2$ será, respectivamente: A) 0,25; 0,75 y 1; B) 0,25; 0,50 y 1; C) 0,25; 0,50 y 0,25.

6.8. A partir de la tabla, con la función de probabilidad de una variable aleatoria X discreta, la media es:

A) 2,4

B) 2,2

C) 2,6

x	$f(x)$
–1	0,2
2	0,4
4	0,4

6.9. Una variable aleatoria discreta X toma los valores 0; 1 y 2, con probabilidades 0,7; 0,2; 0,1, respectivamente. La media o esperanza matemática de X vale: A) 0,2; B) 0,24; C) 0,4.

6.10. Sea Y una variable aleatoria discreta con valores 0, 1, 2, 3 y 4. Si los cinco valores de Y son equiprobables, su media es: A) 1,2; B) 2,0; C) 1,5.

6.11. La variable aleatoria X toma dos valores (cero y uno). Sabiendo que $E(X) = 0{,}2$ ¿Cuánto vale la probabilidad de que X tome el valor cero?: A) 0,8; B) 0,2; C) 0,5.

6.12. Teniendo en cuenta los datos de la tabla, la media de la variable aleatoria X vale:

A) 2,7

B) 7

C) 2,4

x	$F(x)$
1	0,2
2	0,5
3	0,9
4	1

6.13. Con los datos de la tabla 1 del ejercicio 6.8, la varianza de la variable X vale: A) 6,3; B) 3,36; C) 1,63.

6.14. ¿Cuál de las siguientes condiciones no forma parte de los requisitos imprescindibles para la aplicación de la binomial: A) sólo son posibles dos resultados («acierto» y «error»); B) el resultado de un determinado ensayo es función de los resultados obtenidos en los ensayos anteriores; C) la probabilidad de «éxito», p, se mantiene constante.

6.15. Con los datos del ejercicio 6.4, pero siendo la selección «con reposición», y considerando «éxito» obtener «cuadro gris», la probabilidad de que la variable X allí definida tome el valor 2 es: A) 0,50; B) 0,75; C) 0,36.

6.16. Se sabe que un 10% de la población española padece algún tipo de estrés. Si elegimos aleatoriamente una muestra de 8 personas, la probabilidad de que sólo una de ellas padezca estrés vale: A) 0,2638; B) 0,0026; C) 0,3826.

6.17. Continuando con los datos del problema anterior, la probabilidad de que más de una de ellas padezca estrés vale: A) 0,1869; B) 0,3826; C) 0,4305.

6.18. El examen de PIR (Psicólogo Interno Residente) consta de cientos de preguntas tipo test con 5 alternativas cada una de la que una sola es correcta. Si un aspirante a la admisión en el PIR contesta al azar 20 de ellas, la probabilidad de que acierte más de 5 vale: A) 0,6296; B) 0,1958; C) 0,9133.

6.19. Continuando con el ejercicio anterior (6.18) ¿Cuál sería el número de aciertos más probable, en esas 20 preguntas?: A) 5; B) 4; C) 2.

6.20. Con los mismos datos del Ejercicio 6.18 ¿Cuál sería la probabilidad de que falle 13 o más preguntas?: A) 0,9679; B) 0,8265; C) 0,4114.

6.7. SOLUCIÓN A LOS EJERCICIOS DE AUTOEVALUACIÓN

6.1. Solución B

La expresión $f(x)$ se utiliza para representar la probabilidad de una variable aleatoria X tome un valor concreto que representamos por x, es decir: $f(x) = P(X = x)$.

6.2. Solución: C
(Ver apartado 6.3.1).

6.3. Solución: C
La función propuesta no cumple que $\sum f(x) = 1$, que es una de las propiedades fundamentales de la función de probabilidad. En efecto:

$$\sum f(x) = 0 + \frac{10}{60} + \frac{24}{60} + \frac{20}{60} + \frac{4}{60} = \frac{58}{60} \neq 1$$

6.4. Solución: C
A) no es correcta porque la variable no adopta el valor 0.
B) no es correcta porque X no puede tomar el valor 3.

Por tanto, y por exclusión, la respuesta correcta es C. (Puede comprobar el lector que efectivamente esta es la solución correcta efectuando los cálculos oportunos).

6.5. Solución: B
La probabilidad de que una variable aleatoria, *X*, tome «valores menores o iguales que 4» o, lo que es lo mismo, $P(X \leq 4)$ se representa por $F(4)$.

6.6. Solución: A

$$P(X = 4) = F(4) - F(3) = P(X \leq 4) - P(X \leq 3) = 0,974 - 0,963 = 0,011$$

6.7. Solución: A

$$f(0) = P(X = 0) = \frac{2}{4} \cdot \frac{2}{4} = \frac{4}{16} = 0,25$$

$$f(1) = P(X = 1) = 2 \cdot \left(\frac{2}{4} \cdot \frac{2}{4} \right) = 2 \cdot \left(\frac{4}{16} \right) = \frac{8}{16} = 0,5$$

$$f(2) = P(X = 2) = \frac{2}{4} \cdot \frac{2}{4} = \frac{4}{16} = 0,25$$

Por tanto:

$$F(0) = f(0) = 0,25$$

$$F(1) = f(0) + f(1) = 0,25 + 0,5 = 0,75$$

$$F(2) = f(0) + f(1) + f(2) = 0,25 + 0,5 + 0,25 = 1$$

6.8. Solución: B

$$\mu = \sum x f(x) = (-1) \cdot 0,2 + 2 \cdot 0,4 + 4 \cdot 0,4 = -0,2 + 0,8 + 1,6 = 2,2$$

6.9. Solución: C

$$\mu = E(X) = \sum x f(x) = 0 \cdot 0,7 + 1 \cdot 0,2 + 2 \cdot 0,1 = 0 + 0,2 + 0,2 = 0,4$$

6.10. Solución: B

La función de probabilidad es:

x	0	1	2	3	4
$f(x)$	0,2	0,2	0,2	0,2	0,2

Por tanto,

$$\mu_y = \sum y f(y) = 0 \cdot 0,2 + 1 \cdot 0,2 + 2 \cdot 0,2 + 3 \cdot 0,2 + 4 \cdot 0,2 = 0 + 0,2 + 0,4 + 0,6 + 0,8 = 2,0$$

6.11. Solución: A

$E(X) = 0,2$ es la Esperanza o Media de la variable X (también se representa por μ). Su fórmula es:

$$E(X) = \sum x \cdot f(x)$$

la función de probabilidad de la variable X es la siguiente:

x	0	1
$f(x)$	$f(0)$	$f(1)$

puesto que los valores que toma la variable X son 0 y 1 y, donde, $f(x)$ representa las probabilidades asociadas a esos valores. Por tanto, $f(0)$ es la probabilidad de que X tome el valor 0 y $f(1)$ es la probabilidad de que X tome el valor 1.

Entonces:

$$E(X) = \sum x \cdot f(x) = 0 \cdot f(0) + 1 \cdot f(1) = 0 + f(1) = f(1) = 0,2$$

Al tratarse de una función de probabilidad:

$$\sum f(x) = 1$$

y, por tanto, $f(0)+f(1)=1 \Rightarrow f(0)+0,2=1 \Rightarrow f(0)=1-0,2=0,8.$

La probabilidad de que X tome el valor 0 es 0,8.

6.12. Solución: C

Para calcular la media de X necesitamos conocer su función de probabilidad. Esta función la obtenemos (ver la tercera columa de la tabla) «desacumulando» las probabilidades que aparecen acumuladas en la función de distribución:

x	***F*(*x*)**	***f*(*x*)**	***xf*(*x*)**
1	0,2	$f(1) = F(1) = 0,2$	0,2
2	0,5	$f(2) = F(2) - F(1) = 0,5 - 0,2 = 0,3$	0,6
3	0,9	$f(3) = F(3) - F(2) = 0,9 - 0,5 = 0,4$	1,2
4	1	$f(4) = F(4) - F(3) = 1 - 0,9 = 0,1$	0,4
			2,4

$$E(X)=\sum x \cdot f(x)=2,4$$

6.13. Solución: B

Hay dos fórmulas equivalentes para calcular la varianza de una variable aleatoria X:

$$\sigma^2 = V(X) = \sum (x-\mu)^2 \cdot f(x)$$
$$\sigma^2 = V(X) = E(X^2) - \left[E(X)\right]^2$$

Vamos a utilizar las dos en la siguiente tabla:

x	$f(x)$	$x \cdot f(x)$	$(x-\mu)$	$(x-\mu)^2$	$(x-\mu)^2 \cdot f(x)$	x^2	$x^2 \cdot f(x)$
–1	0,2	–0,2	–3,2	10,24	2,048	1	0,2
2	0,4	0,8	–0,2	0,04	0,016	4	1,6
4	0,4	1,6	1,8	3,24	1,296	16	6,4
		2,2			3,36		8,2

Por tanto:

$$\sigma^2 = V(X) = \sum (x-\mu)^2 \cdot f(x) = 2,048 + 0,016 + 1,296 = 3,36$$
$$\sigma^2 = V(X) = E(X^2) - \left[E(X)\right]^2 = 8,2 - (2,2)^2 = 8,2 - 4,84 = 3,36$$

6.14. Solución: B

La utilización de la binomial requiere que los «ensayos» sean independientes, es decir, que el resultado obtenido en un determinado ensayo no dependa del resultado obtenido en los ensayos anteriores.

6.15. Solución: C

Como la selección es «con reposición» («*p*» se mantiene constante a lo largo de los ensayos) podemos utilizar la binomial:

$$f(2) = P(X = 2) = \binom{2}{2} \cdot 0,6^2 \cdot 0,4^0 = 1 \cdot 0,36 \cdot 1 = 0,36$$

Este mismo resultado lo podemos obtener mirando el valor de la tabla I. Obtener 2 cuadros grises con $p = 0,6$ es lo mismo que obtener 0 cuadros azules con $p = 0,4$. Mirando la tabla para $n = 2$, $x = 0$ y $p = 0,4$ obtenemos 0,36.

6.16. Solución: C

Para resolver este ejercicio podemos aplicar la fórmula de la función de probabilidad de la binomial. Lo más práctico, sin embargo, es utilizar la Tabla I.

El valor 0,3826 que se encuentra en la intersección de la fila $n = 8$ y $x = 1$ con la columna $p = 0,1$ es la solucción correcta.

6.17. Solución: A

Se nos pide $P(X > 1)$ y sabemos que $P(X > 1) = 1 - P(X \leq 1)$.

Por otro lado: $P(X \leq 1) = P(X = 0) + P(X = 1)$.

Mirando la Tabla I, comprobamos que $P(X = 0) = 0,4305$ y que $P(X = 1) = 0,3826$. Por tanto:

$$P(X > 1) = 1 - (0,4305 + 0,3826) = 1 - 0,8131 = 0,1869$$

Nótese que $P(X \leq 1)$ podemos obtenerlo directamente a partir de la Tabla II (fila $n = 8$, $x = 1$ y columna $p = 0,1$) haciendo más cómoda la resolución del ejercicio.

6.18. Solución: B

$$P(X > 5) = 1 - P(X \leq 5)$$

Utilizando la tabla II comprobamos (para $n = 20$, $x = 5$ y $p = 0,2$) que $P(X \leq 5) = 0,8042$. Por tanto, $P(X > 5) = 1 - 0,8042 = 0,1958$.

6.19. Solución: B
El número de respuestas acertadas más probable será la media o esperanza matemática de la variable para $n = 20$ y $p = 0,2$. Por tanto: $\mu = np = 20 \cdot 0,2 = 4$.

(Nota: Obsérvese que en la tabla I, para $n = 20$ y $p = 0,2$, el mayor valor de la probabilidad corresponde, efectivamente, a $x = 4$).

6.20. Solución: A
La probabilidad de fallar 13 o más preguntas es la misma que la de acertar 7 preguntas ó menos. Por tanto, se trata de obtener el valor $P(X \leq 7)$ para $n = 20$ y $p = 0,2$. Utilizando la tabla II obtenemos el resultado 0,9679.

Tema 7

Distribuciones continuas de probabilidad

7.1. INTRODUCCIÓN

En este tema vamos a estudiar los modelos de distribución de una variable aleatoria continua más ampliamente utilizados en el área de las ciencias sociales y de la salud. Conviene distinguir entre aquellos modelos a los que frecuentemente se ajustan las variables con las que trabajamos y, aquellos modelos que tienen una gran aplicación como instrumentos estadísticos. Entre los primeros, se encuentra el modelo normal y, entre los segundos, chi-cuadrado de Pearson, *t* de Sudent y *F* de Snedecor.

En todos los modelos seguiremos, aproximadamente, el mismo esquema: primero veremos su definición, posteriormente su media y su varianza y, finalmente, veremos la forma práctica de trabajar con ellos utilizando las tablas estandarizadas existentes que incluimos en el Apéndice.

Dedicaremos especial atención a la distribución normal porque, además de su relevancia como instrumento estadístico, responde al tipo de distribución que sigue la mayoría de variables físicas y psicológicas (la estatura, el peso, la extraversión, el CI —Cociente Intelectual—... son algunas de ellas). Además, como señalábamos al final del tema anterior, las tablas I y II del Anexo no nos permiten resolver un problema binomial con más de 20 ensayos y recurririamos a la aproximación de la binomial a la normal.

Los objetivos que se pretender alcanzar son los siguientes:

- Conocer las características de la distribución normal como distribución de probabilidad de una variable y la aproximación de la binomial a dicha distribución.
- Utilizar las tablas de la distribución normal para obtener probabilidades.
- Conocer las características de las distribuciones *t*, chi-cuadrado y *F*: su media y varianza.

- Utilizar las tablas de las distribuciones t, chi-cuadrado y F, para obtener probabilidades asociadas a unos determinados valores e, inversamente, obtener los valores de estas variables asociados a unas determinadas probabilidades.

7.2. LA DISTRIBUCIÓN NORMAL

En el tema anterior hemos tratado las variables aleatorias discretas. Cuando una variable aleatoria puede tomar infinitos valores diremos que se trata de una variable aleatoria continua. En este caso no tiene sentido hablar de la probabilidad de que la variable tome un valor concreto (que es cero) sino que dicha variable se encuentre en un determinado intervalo.

La *distribución normal*, campana de Gauss o, sencillamente, curva normal como también se conoce a esta distribución fue definida por De Moivre en un intento de encontrar las probabilidades acumulas en una distribución binomial cuando «n» (el número de ensayos) es grande. Nosotros, vamos a señalar sus características fundamentales, la utilización de las tablas y, posteriormente veremos una aproximación intuitiva desde el histograma hasta la curva normal.

7.2.1. Caracteristicas y propiedades

La siguiente fórmula recoge la función, que para variables continuas se denomina de densidad de probabilidad, para una variable X que tiene una distribución normal:

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

donde μ y σ, media y desviación típica, son sus parámetros, π = 3,1416 y e = 2,718 y (base de los logaritmos neperianos).

Si una variable X tiene una distribución que se ajusta a la fórmula anterior, diremos que se distribuye normalmente y lo expresaremos por: $X \rightarrow N(\mu, \sigma)$, indicando que tiene una distribución normal (N) con parámetros «μ» y «σ».

En realidad, como señalaremos también para otras distribuciones, no se trata de una única distribución sino que corresponde a toda una familia caracterizada por sus parámetros media, μ, y desviación típica, σ. Como puede observarse en la figura 7.1 su forma de «campana» es más apuntada cuanto menor es su desviación típica.

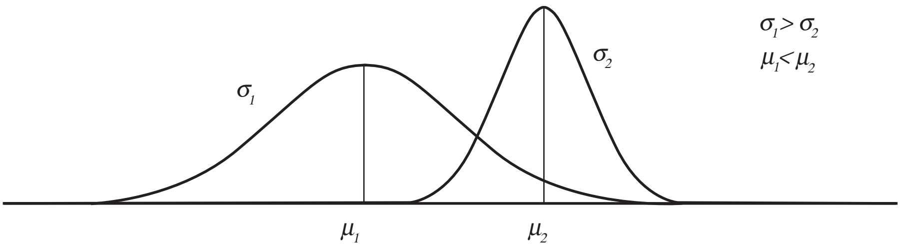


Figura 7.1. Curva normal o campana de Gauss en función de sus parámetros.

Su figura nos indica que la puntuación de la mayoria de los individuos, en una variable que sigue esta distribución, se encuentra entorno a la media y, a medida que nos alejamos de esa puntuación, por su lado izquierdo y derecho, va disminuyendo la frecuencia.

Según una de sus propiedades fundamentales, si a una variable X que se distribuye normalmente, con media μ y varianza σ^2, le aplicamos una transformación lineal de la forma $Y = bX+a$ la nueva variable Y también se distribuirá normalmente pero con media $b\mu_x + a$ y desviación típica $|b|\sigma_x$. Si restamos la media y dividimos por la desviación típica obtenemos una nueva variable que designamos por «z». Esta nueva variable «z» se distribuirá, por tanto:

$$z \rightarrow N(0,1)$$

Y su función de densidad de probabilidad vendrá dada por:

$$\boxed{f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \quad \text{para} \quad -\infty < z < \infty}$$

Su representación gráfica es la siguiente:

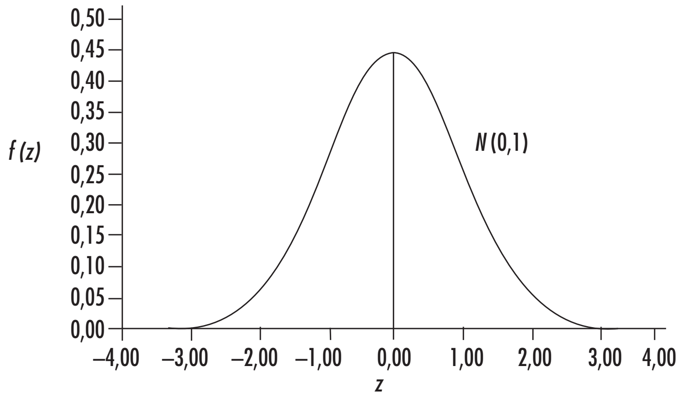


Figura 7.2. Distribución normal típificada o estándar, $N(0,1)$.

Esta distribución se denomina *normal tipificada o normal estandarizada*. Nosotros no vamos a trabajar directamente con su función de densidad de probabilidad. Para la aplicación a problemas concretos, en que se siga esta distribución, recurriremos a las tablas III y IV del Apéndice.

Si observamos la figura 7.2, entre las propiedades fundamentales de una distribución normal podemos destacar las siguientes:

- Es *simétrica* en torno a su media, μ, que coincide con su mediana y su moda.
- La curva normal *tiene dos puntos de inflexión*, es decir, dos puntos donde la curva pasa de ser cóncava a convexa. Estos puntos están situados a una distancia de una desviación típica de la media.
- Es *asintótica* en el eje de abscisas, se extiende desde $-\infty$ hasta $+\infty$ sin llegar nunca a tocar el eje.

Su función de distribución aparece recogida en la siguiente figura:

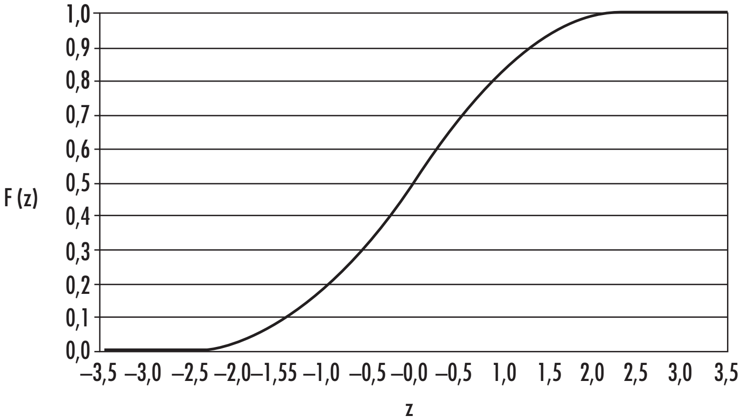


Figura 7.3. Función de distribución.

7.2.2. Utilización de las tablas

En las tablas III y IV se recoge la función de distribución de la distribución normal estándar. En ellas se presentan todas las puntuaciones típicas desde –3,59 hasta +3,59 con intervalos de 0,01. La primera columna encabezada con la letra z consta de un número con un decimal que corresponde a la puntuación típica y la primera fila (a la derecha de la letra *z*) corresponde al segundo decimal de la puntuación *z*. Todos los valores interiores representan probabilidades y llevan, obviamente, un cero delante de la coma. La tabla III corresponde a las puntuaciones típicas negativas (por debajo de la media) y la tabla IV a las positivas (por encima de la media).

Tabla III del Apéndice

x	0,00	0,01	0,02	....	0,05	....	......	0,09
-3,5	....	....	......	....		....	......	......
-3,4	....	....	......	....		....	.....	......
....	....	.....	....	......		.....	......	....
....		.....	.....	......		.....	.....	....
...	....	.....	.....	......	......	......	......	....
-0,2	....	.....	.....	.....	0,4013	......		
....	....	.....	.....	.....	......	......	......	......
-0,0								

Así, por ejemplo la puntuación típica $z = -0{,}25$ (tabla III) deja por debajo de sí una probabilidad de 0,4013.

La puntuación típica $z = 0{,}25$ (tabla IV) deja por debajo de sí una proporción de 0,5987. Al ser una distribución simétrica puede comprobarse que la proporción que queda por debajo de $z = -0{,}25$ es igual a la proporción que queda por encima de $z = 0{,}25$ ($1 - 0{,}5987 = 0{,}4013$). Si la tabla no recoge el valor exacto de z que deseamos podemos utilizar el más próximo. Veamos algunos casos concretos:

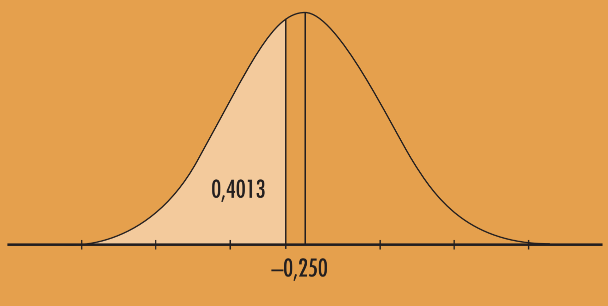
1. Cálculo de la probabilidad para valores menores o iguales que una determinada puntuación típica

En este caso se busca directamente en la tabla.

Ejemplo 7.2. Si una variable se distribuye normalmente, ¿cuál es la probabilidad de obtener valores menores o iguales que $z = -0{,}25$?

Como el valor es negativo se encuentra a la izquierda de la media (ver zona gris de la gráfica). En la tabla III, buscamos en la primera columna el valor de $-0{,}2$ y en la primera fila el valor 0,05.

La probabilidad que deja por debajo de sí esa puntuación es precisamente el valor que se encuentra en la intersección de esa fila y esa columna. En este caso 0,4013.

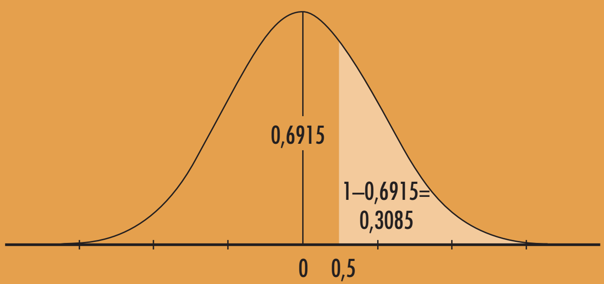


2. *Cálculo de la probabilidad para valores mayores que una determinada puntuación*

En este caso se mira en la tabla la probabilidad que esa puntuación deja por debajo y se resta de 1.

Ejemplo 7.3. Si una variable se distribuye normalmente, ¿cuál es la probabilidad de obtener valores mayores que $z = 0{,}50$?

Si miramos en la tabla IV, la puntuación típica 0,50 deja por debajo de sí una probabilidad de 0,6915.



A nosotros nos interesa la probabilidad que queda por encima y, para calcularla, restaremos de 1 (probabilidad total incluida en la distribución normal): $1 - 0{,}6915 = 0{,}3085$.

3. *Cálculo de la probabilidad entre dos puntuaciones determinadas*

En este caso se restan las probabilidades que dejan por debajo de sí las dos puntuaciones típicas.

Ejemplo 7.4. Si una variable se distribuye normalmente, ¿cuál es la probabilidad de obtener valores comprendidos entre $z = -0{,}25$ y $z = 0{,}50$?

Podemos determinarlo a partir de las puntuaciones típicas y las probabilidades ya obtenidas: bastará con restar a 0,6915 (probabilidad que deja por debajo de sí la puntuación típica 0,50) 0,4013 (probabilidad que deja por debajo de sí la puntuación típica –0,25). El resultado sería 0,2902.

7.2.3. HISTOGRAMA Y DISTRIBUCIÓN NORMAL

Imaginemos que disponemos de los datos de una muestra en una variable *X* (figura 7.4.A). Si hacemos los intervalos más pequeños (figura 7.4.B) y dibujamos el polígono de frecuencias (figura 7.4.C) llegamos a una distribución similar a la normal.

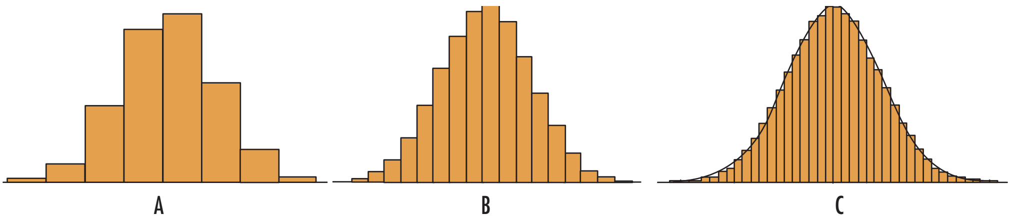


Figura 7.4. Representación gráfica desde el histograma a la curva normal.

Su figura nos indica también que la puntuación de la mayoria de los casos, en una variable que sigue esta distribución, se encuentra entorno a la media y, a medida que nos alejamos de la media, por su lado izquierdo o derecho, va disminuyendo la frecuencia de casos.

Este hecho nos va a permitir aplicar las propiedades de la curva normal a nuestros datos y utilizar las tablas de la misma forma que hemos visto anteriormente.

Si disponemos de los datos originales, en una determinada variable *X*, de un grupo de sujetos y ésta se distribuye normalmente, para resolver determinados cálculos podemos utilizar, como ya se ha señalado, las tablas III y IV de la distribución normal estándar. Para ello deberemos transformar las puntuaciones directas en puntuaciones típicas mediante la siguiente expresión ya utilizada:

$$z_i = \frac{X_i - \bar{X}}{S_x}$$

Para aplicar las tablas de la curva normal a casos concretos que siguen una distribución normal vamos a considerar tres ejemplos prácticos:

Ejemplo 7.5. Imaginemos que las puntuaciones en una determinada asignatura, X, de un grupo de 500 niños se distribuyen normalmente con media 6 y desviación típica 2, ¿cuántos niños no han alcanzado la puntuación 5?

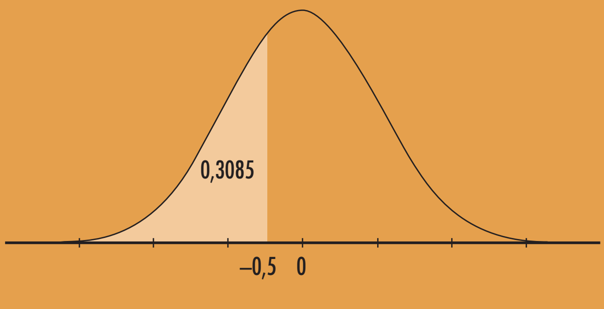
Transformamos la puntuación directa 5 en puntuación típica:

$$\frac{5-6}{2} = \frac{-1}{2} = -0,5$$

Observamos, en la tabla III, que esta puntuación deja por debajo de sí una proporción de 0,3085.

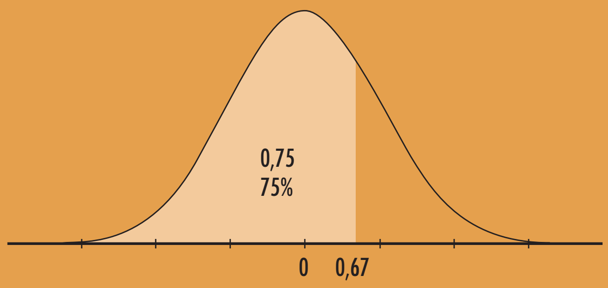
Por tanto:

$$0,3085 \cdot 500 = 154,25 \cong 154 \text{ niños}$$



Ejemplo 7.6. Con los mismos datos del ejemplo anterior (ejemplo 7.5.), ¿cuál será el Percentil 75, P_{75}, de la distribución?

Tal como se definió en el tema 2, el P_{75} será una puntuación directa que dejará por debajo de sí el 75% de los casos. A este percentil le corresponderá una puntuación típica que deja por debajo de sí una proporción de casos de 0,75.



Ahora debemos buscar en el interior de la tabla la proporción 0,75, o en su defecto la más próxima (en este caso 0,7486), y ver a que puntuación típica corresponde: 0,67 (lógicamente se trata de una puntuación típica positiva porque el percentil 75 deja por debajo de sí más del 50% que corresponde a la media).

A partir de esta puntuación típica calculamos el P_{75} de la siguiente manera:

$$z = \frac{P_{75} - \bar{X}}{S_x} \Rightarrow 0{,}67 = \frac{P_{75} - 6}{2} \Rightarrow P_{75} = (0{,}67 \cdot 2) + 6 = 7{,}34$$

Ejemplo 7.7. El peso de un grupo de 1.000 niños se distribuye normalmente con un Coeficiente de variación de 10 (CVx = 10). Si el 84,13% de ellos no supera los 33 kg. ¿Cuánto vale la media y la desviación típica de la distribución?

Establecemos el sistema de ecuaciones y resolvemos:

$$\left.\begin{array}{l} \dfrac{S_x}{\bar{X}}100 = 10 \\[2ex] 84{,}13\% \rightarrow z = 1 \\[1ex] 1 = \dfrac{33-\bar{X}}{S_x} \end{array}\right\} \Rightarrow \left.\begin{array}{l} S_x 100 = 10\bar{X} \\ S_x = 33-\bar{X} \end{array}\right\} \Rightarrow (33-\bar{X})100 = 10\bar{X} \Rightarrow \left\{\begin{array}{l} \bar{X} = 30 \\ S_x = 3 \end{array}\right.$$

7.2.4. APROXIMACIÓN DE LA BINOMIAL A LA NORMAL

Al finalizar el tema anterior nos habiamos preguntado qué hacer cuando, para la distribución binomial, tenemos un «n» superior a 20 (las tablas de la binomial no recogen valores superiores a éste). La opción, para valores grandes de «n», es aproximar la distribución binomial a la normal. La aproximación de la binomial a la normal mejora a medida que «p» (la probabilidad de éxito) se aproxima a 0,5 y «n» (número de ensayos) es grande, como podemos observar en la figura 7.5:

Sabemos que para una variable, X, que sigue una distribución binomial su media es $\mu = np$ y su desviación típica es $\sigma = \sqrt{npq}$. Por tanto, podemos transformar su función de probabilidad, que es discreta, a la normal de la siguiente manera:

$$P(X = x) = P\left[\frac{(x-0,5)-\mu}{\sigma} \leq \frac{x-\mu}{\sigma} \leq \frac{(x+0,5)-\mu}{\sigma}\right]$$

$$P(X = x) = P\left[\frac{(x - 0{,}5) - np}{\sqrt{npq}} \leq z \leq \frac{(x + 0{,}5) - np}{\sqrt{npq}}\right]$$

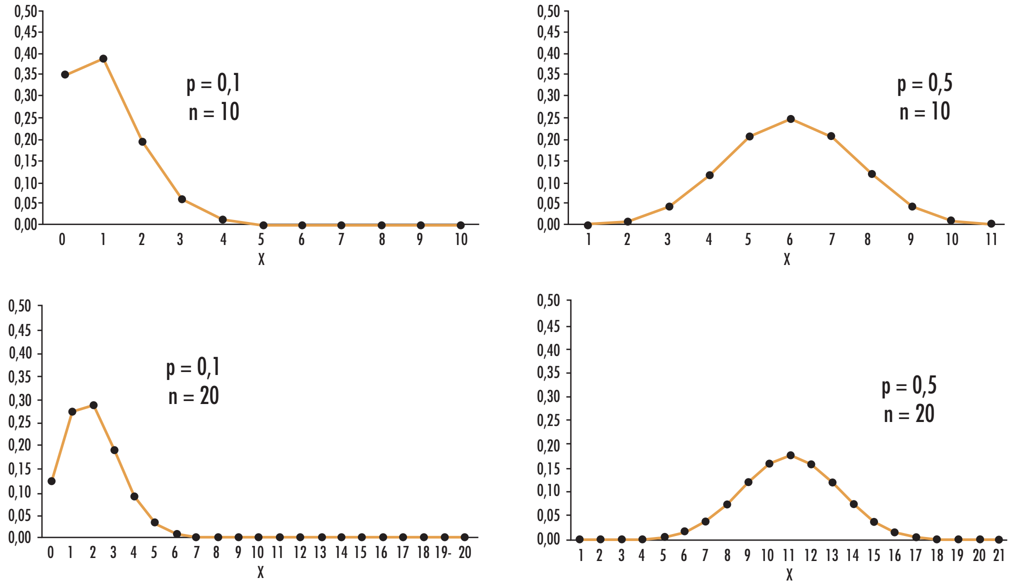


Figura 7.5. Distribución binomial para distintos valores de p (0,1 y 0,5) y n (10 y 20).

Ejemplo 7.8. Supongamos que lanzamos una moneda al aire en 20 ocasiones ¿cuál es la probabilidad de obtener 12 caras?

Para contestar a esta pregunta tenemos que recurrir a la distribución binomial y buscar la probabilidad de que la variable aleatoria, X: «número de caras» tome el valor 12 ($x = 12$) con n = 20 y $p = 0{,}5$. Mirando en la tabla I obtenemos el valor 0,1201

Vamos a responder a la pregunta haciendo una aproximación de la binomial a la normal.

La media y la desviación típica de esta distribución binomial son:

$\mu = 20 \cdot 0{,}5 = 10$ y $\sigma = \sqrt{npq} = \sqrt{20{\cdot}0{,}5{\cdot}0{,}5} = \sqrt{5} = 2{,}24.$

La distribución Normal es continua y, como para cualquier distribución continua, la probabilidad de que la variable X tome un valor concreto es cero: $P(X = 12) = 0$. Para aproximar la distribución binomial a la normal estableceremos un intervalo entre 0,5 unidades a la izquierda y a la derecha de la puntuación, es decir:

$$P[(12 - 0{,}5) \leq x \leq (12{+}0{,}5)$$

A continuación, transformamos las puntuaciones en típicas:

$$P\left[\frac{(12-0,5)-\mu}{\sigma} \leq \frac{x-\mu}{\sigma} \leq \frac{(12+0,5)-\mu}{\sigma}\right]$$

y nos quedaría:

$$P\left[\frac{(12-0,5)-\mu}{\sigma} \leq z \leq \frac{(12+0,5)-\mu}{\sigma}\right]$$

Puesto que $\mu = 10$ y $\sigma = 2{,}24$:

$$P\left[\frac{(12-0,5)-10}{2,24} \leq z \leq \frac{(12+0,5)-10}{2,24}\right] = P(0,67 \leq z \leq 1,12)$$

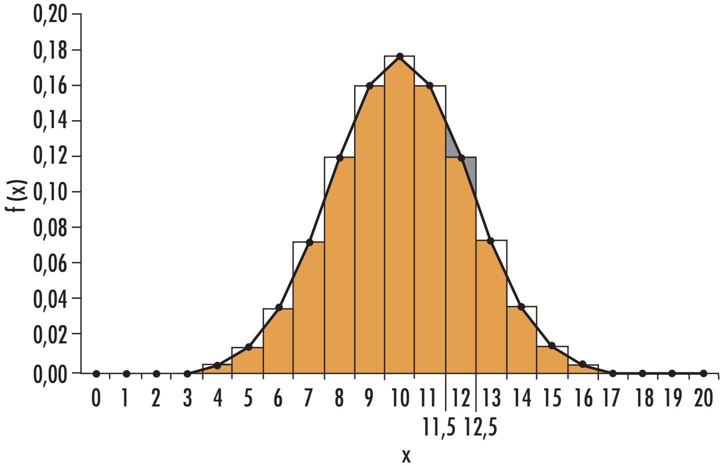
Utilizando las tablas de la distribución Normal:

$$P\,(0{,}67 \leq z \leq 1{,}12) = 0{,}8686 - 0{,}7486 = 0{,}12$$

Como puede observarse en este caso, la aproximación es «muy buena» (hay una diferencia de solo una diezmilésima) para $n = 20$. A medida que «n» aumenta mejora la aproximación.

Sumar y restar, en el caso anterior, el valor 0,5 se llama «*corrección por continuidad*» y nos va a permitir utilizar las puntuaciones discretas, X, como si fuesen continuas. Para ello, interpretamos cada puntuación, X, como si fuesen los puntos medios de sus intervalos. Con este procedimiento tratamos de asegurar que el intervalo incluya los valores discretos de la binomial.

Gráficamente:



Veamos otro ejemplo:

Ejemplo 7.9. De todas las preguntas del examen para el PIR (cada pregunta consta de cinco alternativas de la que sólo una es correcta), un aspirante desconoce completamente 40 de ellas y las responde al azar.

A) ¿Cuál es la probabilidad de que acierte entre 10 y 12 de esas preguntas?

B) ¿Cuál es la probabilidad de que acierte más de 10?

Para esta distribución binomial:

$$\mu = np = 40 \cdot 0,20 = 8 \qquad \text{y} \qquad \sigma = \sqrt{npq} = \sqrt{40 \cdot 0,20 \cdot 0,80} = 2,53$$

Por tanto:

A)

$$P(9,5 \leq X \leq 12,5) = P\left[\left(\frac{9,5-8}{2,53}\right) \leq z \leq \left(\frac{12,5-8}{2,53}\right)\right] =$$

$$= P(0,59 \leq z \leq 1,78) = 0,9625 - 0,7224 = 0,2401$$

B)

La probabilidad de que acierte más de 10 es igual a la probabilidad de que acierte 11, 12... Por tanto:

$$P(X > 10{,}5) = P\left(z > \frac{10{,}5 - 8}{2{,}53}\right) = P(z > 0{,}99) =$$
$$= 1 - P(z \le 0{,}99) = 1 - 0{,}8389 = 0{,}1611$$

7.3. LA DISTRIBUCIÓN «CHI CUADRADO» DE PEARSON

Hemos analizado, en las páginas anteriores, la distribución normal. Siguiendo con las distribuciones continuas, en este epígrafe consideraremos la distribución chi-cuadrado (χ^2) de Pearson íntimamente relacionada con ella.

Sean X_1, X_2,..., X_n, un conjunto de n variables aleatorias independientes con una distribución $N(0,1)$, entonces una nueva variable aleatoria $X = X_1^2 + X_2^2 + \ldots X_n^2$ sigue una distribución χ_n^2 (chi-cuadrado con n grados de libertad) y se representa así: $X \to \chi_n^2$. Su media y su varianza valdrán $\mu = n$ y $\sigma^2 = 2n$, respectivamente.

Los grados de libertad (n) indican que cada una de las n variables aleatorias puede tomar cualquier valor, de sus posibles valores, sean cuales sean los valores tomados por las $n - 1$ restantes. Su análisis más detallado escapa a los objetivos de este texto.

Esta distribución se usa fundamentalmente en pruebas de bondad de ajuste (para contrastar si la distribución de una variable se ajusta a una distribución determinada, por ejemplo la normal). En realidad, al igual que otras distribuciones, es una familia de curvas, en función de los grados de libertad, como las presentadas en la siguiente figura 7.6:

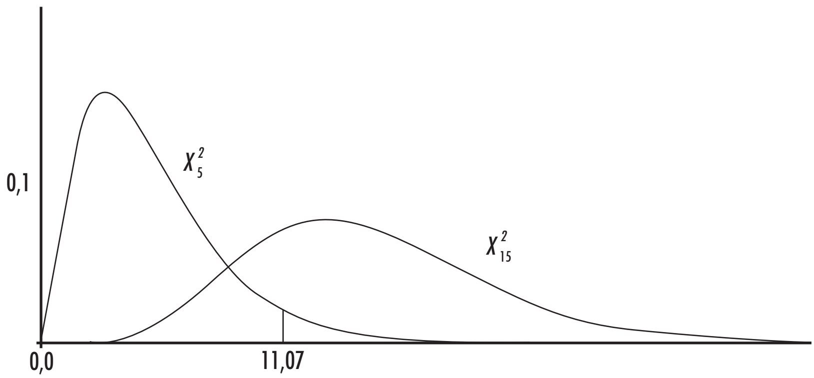


Figura 7.6. Representación gráfica de la distribución Chi-cuadrado en función de sus grados de libertad (5 y 15).

Entre sus propiedades, ver figura 7.6, podemos señalar las siguientes:

- Nunca adopta valores menores de 0.
- Es asimétrica positiva pero a medida que aumentan sus grados de libertad se va aproximando a la distribución normal.
- Para $n > 100$ la podemos aproximar a una distribución $N(n, \sqrt{2n})$.

La tabla V nos permite obtener las probabilidades asociadas a algunos valores de toda la familia de distribuciones χ^2, entre los que se encuentran los más usados habitualmente.

La primera fila recoge las probabilidades o proporciones y la primera columna los grados de libertad correspondientes. En el interior de la tabla se encuentran los valores de la variable. Así, por ejemplo, para una variable que sigue una distribución chi-cuadrado con 5 grados de libertad, $X \rightarrow \chi^2_5$, el valor 11,07 deja por debajo de sí una proporción de 0,95. Es decir, $P(X \leq 11{,}07) = 0{,}95$. Esta puntuación se corresponde, por tanto, con el percentil 95. Suele presentarse de la siguiente manera: ${}_{0,95}\chi^2_5 = 11{,}07$. En la siguiente gráfica puede observarse su situación en la tabla:

Tabla V del Apéndice

g.l.	0,001	0,005	0,02	....	0,950	....	......	0,999
1	....	....	......	....		....	......	......
2	....	....	......	....		....	....	......
3								
4	....	.....	.....	......	......	......	......	....
5	....	....	....	....	11,07	......		
.... 100	....	.....	.....	.....	......	......	......	......

Ahora bien, si interesara hallar $P(X > 11{,}07)$ haríamos lo siguiente:

$$P(X > 11{,}07) = 1 - P(X \leq 11{,}07) = 1 - 0{,}95 = 0{,}05$$

7.4. LA DISTRIBUCIÓN «*t*» DE STUDENT

A la hora de definir este tipo de distribución de probabilidad, al igual que hicimos anteriormente con chi-cuadrado, lo haremos en función de otras distribuciones ya conocidas.

Sean X e Y dos variables aleatorias independientes, donde X sigue una distribución $N(0, 1)$ e Y una distribución χ^2_n. Entonces, la variable aleatoria $T = \dfrac{X}{\sqrt{Y/n}}$ sigue una distribución «*t*» con n grados de libertad y se expresa por: $T \to t_n$

Su media siempre vale 0 ($\mu = 0$) y su varianza $\sigma^2 = \dfrac{n}{n-2}$

Podemos definir una distribución «*t*» como el cociente entre una variable $N(0,1)$ y la raíz cuadrada de una variable χ^2_n dividida por sus grados de libertad. Su nombre se debe a su descubridor, el matemático Gosset, que publicó sus trabajos bajo el seudonimo de «Student».

En la figura 7.7 se representa la distribución «*t*», con dos grados de libertad, junto a la distribución normal estándar.

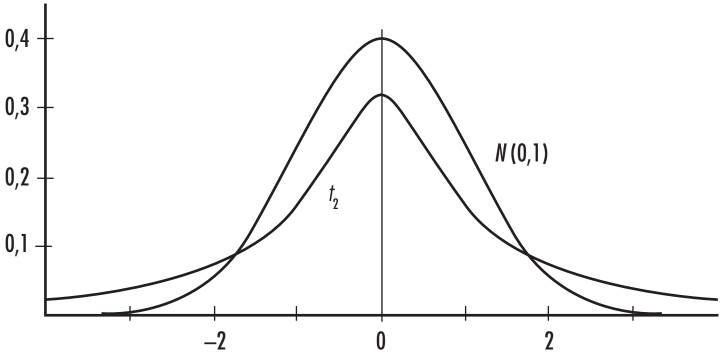


Figura 7.7. Representación gráfica de la distribución «*t*» con 2 grados de libertad.

A partir de su definición y de su representación gráfica podemos señalar las siguientes características:

— Es simétrica, con $\mu = 0$. Su forma es muy parecida a la $N(0,1)$, aunque menos apuntada.

— Puede tomar cualquier valor entre $-\infty$ y $+\infty$.

— A medida que aumentan los grados de libertad, la distribución se aproxima más y más a una distribución normal.

— La curva es asintótica al eje de abscisas.

Se utiliza, fundamentalmente, en estadística inferencial en las pruebas de contraste. En la tabla VI se presentan los valores positivos para esta distribución. En la primera columna se presentan los grados de libertad y en la primera fila las distintas probabilidades o proporciones de valores menores o iguales que un valor positivo dado. Como se trata de una distribución simétrica podemos hallar las probabilidades asociadas a valores negativos a partir de los valores positivos de la tabla VI. Veámoslo con un ejemplo.

Ejemplo 7.10. Sea X una variable que se distribuye según t con 5 grados de libertad.

A) Calcular la probabilidad de obtener valores menores o iguales a 2,015, $P(X \leq 2{,}015)$.

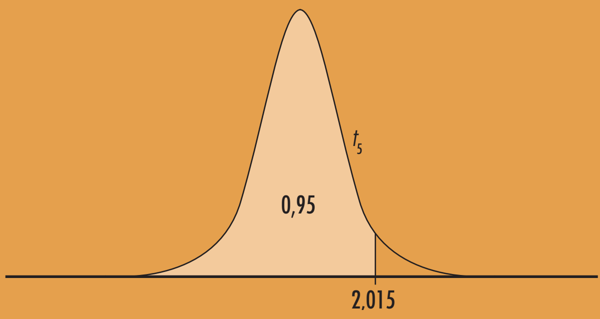
B) Calcular $P(X > 0{,}920)$.

C) Calcular $P(X \leq -2{,}571)$.

A)

Corresponde a la zona coloreada de la figura. Para ello, consultamos la tabla VI. En la primera columna (grados de libertad) localizamos el valor 5. Los valores incluidos en su fila correspondiente son valores de t, localizamos 2,015 y vemos que en la primera fila se corresponde con 0,95. Por tanto:

$$P(X \leq 2{,}015) = 0{,}95$$



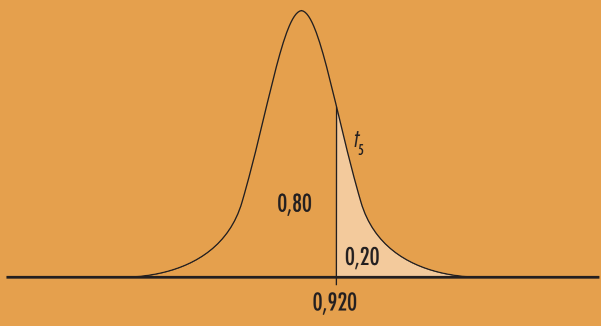
B)

En la tabla VI vemos que para t_5:

$$P(X \leq 0{,}920) = 0{,}80$$

Por tanto,

$$P(X > 0{,}920) = 1 - 0{,}80 = 0{,}20$$



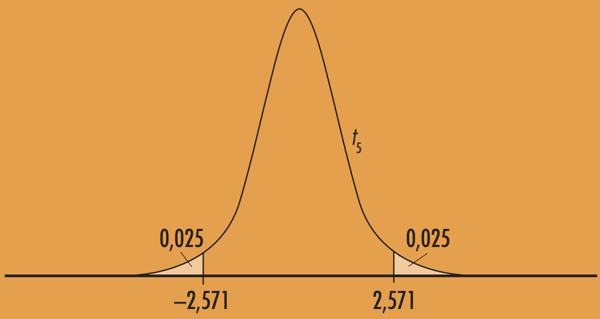
C)

Como se ve en la gráfica, los valores negativos se encuentran a la izquierda de la media (que vale cero) y los positivos, a la derecha. Al ser simétrica:

$$P(X \leq -2{,}571) = P(X > 2{,}571) \text{ y } P(X > 2{,}571) =$$
$$= 1 - P(X \leq -2{,}571) = 1 - 0{,}975 = 0{,}025$$

Por tanto:

$$P(X \leq -2{,}571) = 0{,}025$$



7.5. LA DISTRIBUCIÓN «*F*» DE SNEDECOR

Al igual que con las distribuciones anteriores, nos limitaremos a presentar su definición, algunas de sus propiedades y la utilización de las tablas.

Definición:

Si X_1 y X_2 son variables aleatorias independientes, con distribución chi-cuadrado con n_1 y n_2 grados de libertad respectivamente, entonces una nueva variable F definida por $F = \dfrac{X_1 / n_1}{X_2 / n_2}$ sigue una distribución F con n_1 y n_2 grados de libertad (F_{n_1, n_2}).

Siendo «n_1» los grados de libertad del numerador y «n_2» los grados de libertad del denominador, su media y su varianza vienen definidas por:

$$\mu = \frac{n_2}{n_2 - 2} \quad \text{para } n_2 > 2, \text{ y } \sigma^2 = \frac{2n_2^2(n_1 + n_2 - 2)}{n_1(n_2 - 4)(n_2 - 2)^2} \quad \text{para } n_2 > 4$$

Se la conoce habitualmente como F de Fisher o de Snedecor, se emplea fundamentalmente en el contraste de hipótesis (Análisis de Varianza...), y en la figura 7.8 aparece su representación según distintos grados de libertad.

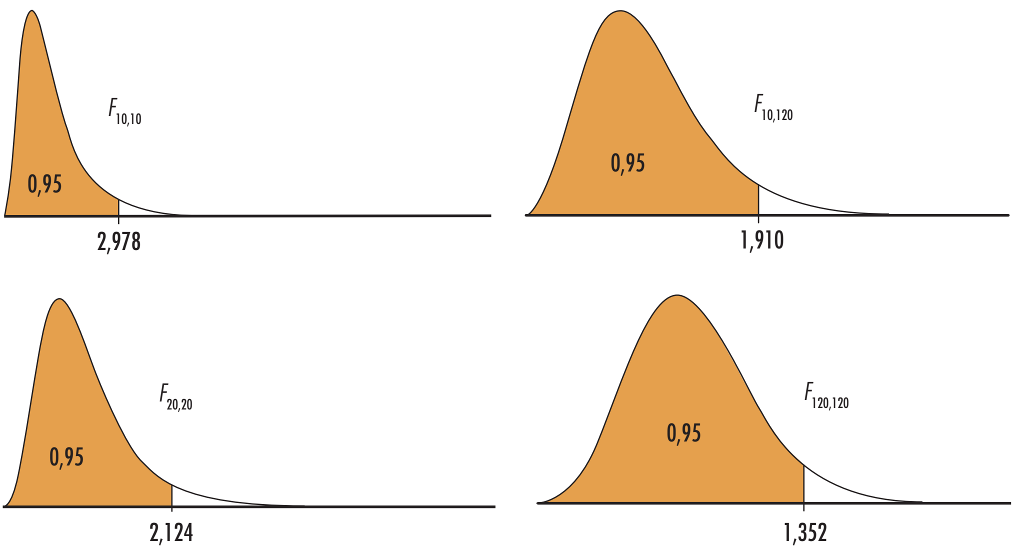


Fig. 7.8. Distribuciones F con distintos grados de libertad.

Sus características más importantes son:

- Es asimétrica positiva por lo que nunca toma valores menores que 0.
- Una importante propiedad de esta distribución es la llamada propiedad recíproca y dice que si X es una variable con distribución F con n_1 y n_2 grados de libertad, entonces la variable $Y = 1/X$ es también una

distribución F con n_1 y n_2 grados de libertad. Esta propiedad la podemos también expresar de la siguiente forma:

$$_{1-p}F_{n_1,n_2} = \frac{1}{_{p}F_{n_2,n_1}}$$

donde p es la probabilidad asociada al valor de la variable. Esta propiedad es de enorme importancia para obtener algunos percentiles o probabilidades que no aparecen en la tabla, tal y como se verá en ejemplos posteriores.

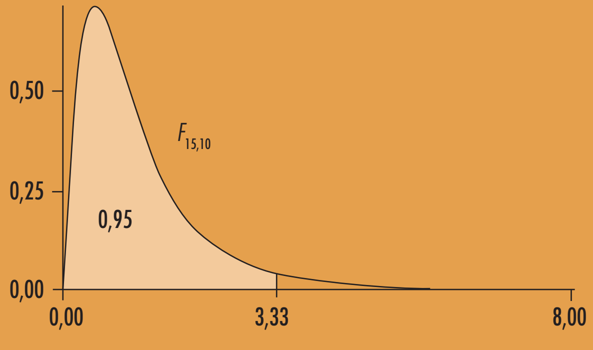
La tabla VII recoge solamente la probabilidad de que X sea menor o igual que 0,900; 0,950; 0,975 y 0,990 que son los valores utilizados habitualmente. Para comprender el manejo de esta tabla, vamos a considerar el siguiente ejemplo.

Ejemplo 7.11. Sea X una variable que se distribuye según $F_{5,10}$: A) Calcular $P(X \leq 3{,}33)$ y B) Determinar el valor del percentil 5 de X, es decir: $_{0,05}F_{5,10}$

A)

Buscamos en la tabla VII, para 5 grados de libertad en el numerados y 10 para el denominador, donde se encuentra el valor 3,33. Vemos en la parte superior de la tabla que se corresponde con una probabilidad de 0,95.

Por tanto, 3,33 se corresponde con el percentil 95.



B)

En este caso, tenemos que hacer uso de la propiedad recíproca. Es decir:

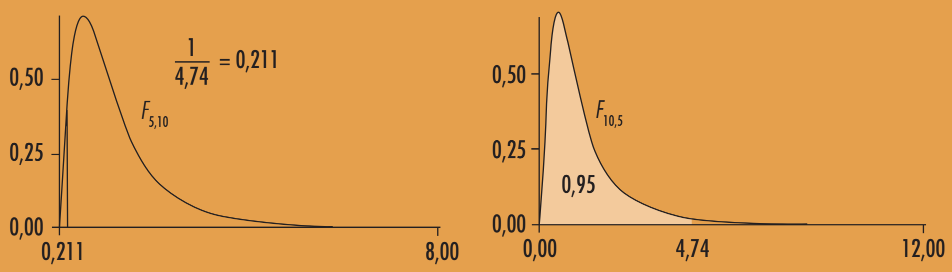
$$_{0,05}F_{5,10} = \frac{1}{_{(1-0,05)}F_{10,5}} = \frac{1}{_{0,95}F_{10,5}}$$

A partir de la tabla VII vemos que: $_{0,95}F_{10,5}$ es igual a 4,74.

Por tanto:

$$_{0,05}F_{5,10} = \frac{1}{_{0,95}F_{10,5}} = \frac{1}{4,74} = 0,211$$

Puede verse gráficamente en la siguiente figura:



7.6. RESUMEN

Hemos visto, en este tema, algunas de las distribuciones continuas de probabilidad: la distribución normal, la distribución chi-cuadrado, la distribución *t* de Student y la distribución *F* de Snedecor.

Se han definido cada una de estas distribuciones. La distribución chi-cuadrado se define en función de otras con distribución normal. La distribución *t* se ha definido en función de otras dos: una normal y otra chi-cuadrado y, por último, la distribución *F* se ha definido en función de dos chi-cuadrado, que a su vez se definen en función de la normal. Por tanto, no debe sorprender que todas ellas convergan, en algún momento, en la distribución normal.

Se ha prescindido de incluir la ecuación de sus respectivas funciones de densidad de probabilidad y distribución, por su complejidad y porque podemos servirnos de unas tablas donde se recogen estas probabilidades. Además de la conveniencia de saber utilizar estas tablas, por su relevancia en los temas de inferencia, es necesario conocer sus características más importantes: el rango de valores en el que la función está definida, su media, varianza y aproximación a la normal, en su caso, bajo determinadas circunstancias.

7.7. EJERCICIOS DE AUTOEVALUACIÓN

7.1. En una distribución normal: A) la media es mayor que la mediana; B) la media es menor que la mediana; C) media y mediana coinciden.

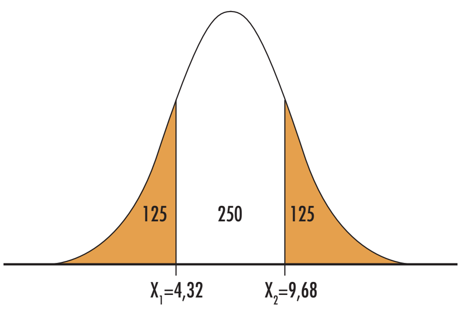
7.2. En una distribución normal ¿entre qué puntuaciones típicas se encuentra el 60 % de los casos centrales de la distribución?: A) –0,84 y 0,84; B) –1,96 y 1,96; C) –1,64 y 1,64.

7.3. Las puntuaciones de 1.000 niños en un test de inteligencia, *X*, se distribuyen normalmente con media 100 y desviación típica 15 ¿Cuál es la probabilidad de obtener puntuaciones menores o iguales que 85? : A) 0,8413; B) 0,1587; C) 0,6826.

7.4. Con los datos del ejercicio anterior (7.3.) ¿Cuántos niños obtienen puntuaciones superiores a 115?: A) 115; B) 200; C) 159.

7.5. Continuando con los datos del ejercicio 7.3 ¿Cuánto vale el Percentil 75 de la distribución?: A) 110,05; B) 75,00; C) 89,95.

7.6. Las calificaciones obtenidas en el examen en una asignatura (*X*), de un grupo de 500 alumnos, se distribuyen normalmente. Como se muestra en la figura, de ellos 125 no alcanzan la puntuación 4,32 y otros 125 superan la puntuación 9,68. ¿Cuánto vale la media de *X*?: A) 7,00; B) 5,00; C) 6,00.



7.7. Con los datos del ejercicio anterior, ¿cuánto vale la desviación típica de X?: A) 3; B) 2; C) 4.

7.8. Con los datos del ejercicio 7.6. ¿Cuál será el Percentil 33?: A) 5,24; B) 8,76; C) 5,67.

7.9. Siguiendo con los datos del ejercicio 7.6 y considerando suspendidos aquellos alumnos que no alcanzan la puntuación 5. ¿Cuántos alumnos han suspendido?: A) 250; B) 200; C) 154.

7.10. Sabiendo que X se distribuye normalmente, que $\overline{X} = 60$ y que la puntuación directa 40,8 es superada por el 89,97 % de la distribución, la desviación típica vale: A) 15; B) 1,28; C) 17,87.

7.11. Una variable X se distribuye normalmente, con desviación típica 5. Sabiendo que la puntuación 45 deja por encima de sí el 84,13 % de los casos, su media valdrá: A) 40; B) 50; C) 60.

7.12. Las puntuaciones de 10.000 niños españoles en una prueba de inteligencia (X) se distribuyen normalmente con media 100. Sabemos que 668 niños no alcanzan la puntuación 85 y otros 668 niños obtienen puntuaciones superiores a 115. Su varianza vale: A) 10; B) 200; C) 100.

7.13. El 20% de los niños en edad escolar presenta problemas de adaptación al Colegio. Si en un determinado centro hay 225 niños, ¿Cuál es la probabilidad de que 30 o menos presenten algún problema de adaptación?: A) 0,0080; B) 0,3026; C) 0,0263.

7,14. Con los datos del ejercicio anterior, ¿cuál es la probabilidad de que más de 55 presenten algún problema de adaptación?: A) 0,1040; B) 0,0401; C) 0,4010.

7.15. Con los mismos datos del ejercicio 7.13, ¿Cuál es la probabilidad de que entre 40 y 50 niños presenten problemas de adaptación?: A) 0,4642; B) 0,2446; C) 0,6424.

7.16. En una distribución chi-cuadrado con 28 grados de libertad, el valor 41,34 es: A) el percentil 5; B) el percentil 90; C) el percentil 95.

7.17. En una distribución F con 10 grados de libertad en el numerador y 20 grados de libertad en el denominador, ¿cuál es el valor del percentil 90?: A) 2,20; B) 2,35; C) 1,94.

7.18. En una distribución F con 10 grados de libertad en el numerador y 20 en el denominador, ¿cuál es el valor del percentil 10 (el que deja por debajo al 10% de los casos)?: A) 1,940; B) 2,200; C) 0,455.

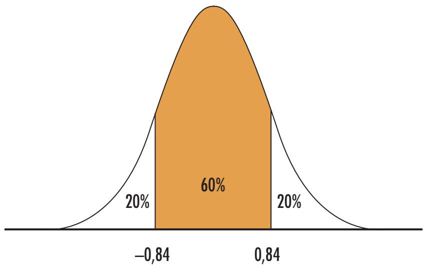
7.19. ¿Cuál de las siguientes distribuciones NO es simétrica?: A) normal con media 5 y desviación típica 2; B) chi-cuadrado con 10 grados de libertad; C) t de Student con 10 grados de libertad.

7.20. El valor 0,86 se corresponde con: A) el percentil 80 de una distribución t de Student con 20 grados de libertad; B) el percentil 5 de una distribución chi-cuadrado con 19 grados de libertad; C) el percentil 20 de una distribución t de Student con 20 grados de libertad.

7.8. SOLUCIONES A LOS EJERCICIOS DE AUTOEVALUACIÓN

7.1. Solución: C

(ver apartado 7.2.1).



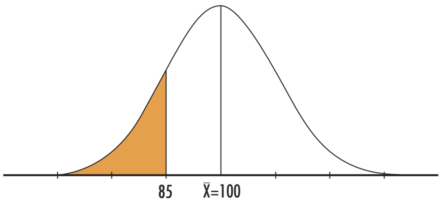
7.2. Solución: A

(ver Tabla de la Curva Normal).

7.3. Solución: B

$$z = \frac{X - \bar{X}}{S_X} = \frac{85 - 100}{15} = -1$$

Tabla III: 0,1587.



7.4. Solución: C

$$z = \frac{X - \bar{X}}{S_X} = \frac{115 - 100}{15} = 1$$

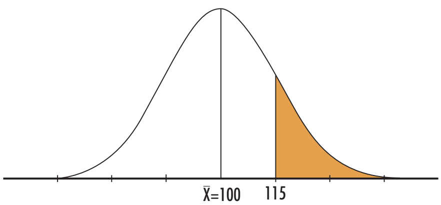


Tabla IV: 0,8413

1 – 0,8413 = 0,1587

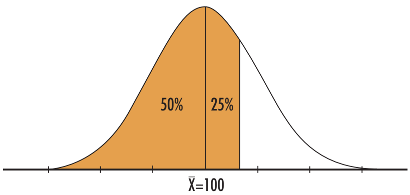
0,1587 × 1000 = 158,7 ≅ 159

7.5. Solución: A

$P_{75} \Rightarrow z = 0,67$ Tabla IV)

$$0,67 = \frac{P_{75} - 100}{15} \Rightarrow 0,67 \cdot 15 = P_{75} - 100 \Rightarrow$$

$$P_{75} = (0,67 \cdot 15) + 100 = 110,05$$



7.6. Solución: A

$$\bar{X} = \frac{4,32 + 9,68}{2} = \frac{14}{2} = 7$$

7.7. Solución: C

$$\left.\begin{array}{r} -0,67 = \dfrac{4,32 - \bar{X}}{S_x} \\ 0,67 = \dfrac{9,68 - \bar{X}}{S_x} \end{array}\right\} \Rightarrow \left.\begin{array}{r} -0,67 S_x = 4,32 - \bar{X} \\ 0,67 S_x = 9,68 - \bar{X} \end{array}\right\} \Rightarrow \bar{X} = 7 \quad S_x = 4$$

7.8. Solución: A

$$-0,44 = \frac{P_{33} - 7}{4} \Rightarrow P_{33} = 7 - 1,76 = 5,24$$

7.9. Solución: C

$$\frac{5-7}{4} = -0,5 \Rightarrow (\text{Tablas})\, 0,3085$$

$$0,3085 \cdot 500 = 154,25 \cong 154$$

7.10. Solución: A

$$1 - 0,8997 = 0,1003 \Rightarrow z = -1,28$$

$$-1,28 = \frac{40,8 - 60}{S_x} \Rightarrow S_x = \frac{40,8 - 60}{-1,28} = 15$$

7.11. Solución: B

$$1-0,8413=0,1587 \Rightarrow z=-1$$

$$-1=\frac{45-\bar{X}}{5} \Rightarrow \bar{X}=5+45=50$$

7.12. Solución: C

Puesto que las puntuaciones son simétricas, su media vale:

$$\bar{X}=\frac{85+115}{2}=100$$

y su desviación típica:

$$-1,5=\frac{85-100}{S_x} \Rightarrow S_x=\frac{-15}{-1,5}=10$$

Por tanto, su varianza es $10^2 = 100$.

7.13. Solución: A

$$n = 225 \quad p = 0,2 \quad q = 1 - \mathrm{p} = 0,8$$

$$P(X \leq 30)=P\left(z \leq \frac{30,5-np}{\sqrt{npq}}\right)=P\left(z \leq \frac{30,5-(225 \cdot 0,2)}{\sqrt{225 \cdot 0,2 \cdot 0,8}}\right)=$$

$$=P\left(z \leq \frac{30,5-45}{6}\right)=P(z \leq -2,41)=0,0080$$

(Utilizando la tabla III de la curva normal).

7.14. Solución: B

$$P(X>55)=P\left(z>\frac{55,5-np}{\sqrt{npq}}\right)=P\left(z>\frac{55,5-(225 \cdot 0,2)}{\sqrt{225 \cdot 0,2 \cdot 0,8}}\right)=$$

$$=P\left(z>\frac{55,5-45}{6}\right)=P(z>1,75)=1-P(\leq 1,75)=1-0,9599=0,0401$$

(Utilizando la tabla IV de la curva normal).

7.15. Solución: C

$$P(40 \le X \le 50) = P\left(\frac{39,5-np}{\sqrt{npq}} \le z \le \frac{50,5-np}{\sqrt{npq}}\right) =$$

$$= P\left(\frac{39,5-(225\cdot 0,2)}{\sqrt{225\cdot 0,2\cdot 0,8}} \le z \le \frac{50,5-(225\cdot 0,2)}{\sqrt{225\cdot 0,2\cdot 0,8}}\right) =$$

$$= P\left(\frac{39,5-45}{6} \le z \le \frac{50,5-45}{6}\right) =$$

$$= P(-0,92 \le z \le 0,92) = 0,8212 - 0,1788 = 0,6424$$

(Utilizando las tablas III y IV de la curva normal).

7.16. Solución: C

(Ver tabla V)

7.17. Solución: C

(Ver tabla VII)

7.18. Solución: C

$${}_{0,10}F_{10,20} = \frac{1}{{}_{0,90}F_{20,10}} = \frac{1}{2,20} \cong 0,455$$

7.19. Solución: B

Las distribuciones $N(5,2)$ y t_{10} son simétricas

7.20. Solución: A

(Ver tablas correspondientes)

Tema 8

Estimación

8.1. INTRODUCCIÓN

En este tema, iniciamos el estudio de la parte del análisis de datos denominada ***inferencia estadística*** que básicamente consiste en estimar, con cierta probabilidad, un parámetro desconocido a partir de una muestra aleatoria extraída de una población. Así, a partir de las características (media, proporción...) de una muestra inferiremos esas mismas características a la población. El proceso a seguir, en cinco fases, aparece recogido en la figura 8.1.

Especificación de la población

$\Downarrow$

Obtención de la muestra

$\Downarrow$

Medición de la variable

$\Downarrow$

Análisis descriptivo y verificación de los datos

$\Downarrow$

Inferencia estadística sobre un parámetro

Figura 8.1. Proceso estadístico para inferir un parámetro a partir de una muestra.

Fase 1. Consiste en una especificación clara de la población de interés, dado que el procedimiento permite realizar inferencias únicamente a la población de la que procede la muestra. La población que se utilice dependerá de los objetivos de la investigación.

Fase 2. La muestra es el conjunto de elementos en el que se realizará la investigación. Se obtiene mediante un método de selección y el número de elementos que la componen es una característica esencial de la muestra.

Fase 3. Se mide la variable de interés a todos los elementos de la muestra y en las mismas condiciones, obteniendo una medida por cada elemento.

Fase 4. Se realiza un análisis descriptivo de los datos, tanto analítico como gráfico, que dará una descripción detallada de la muestra. Además, se verifican los datos para detectar posibles errores en la recogida de los mismos.

Fase 5. Se aplican las herramientas de inferencia. Distinguiremos dos procedimientos de inferencia estadística, la estimación por intervalo (intervalos de confianza) y el contraste de hipótesis. En esta asignatura, únicamente estudiaremos el primer procedimiento.

Los objetivos que se pretenden en este tema son:

- Saber relacionar los conceptos de población, muestra, análisis estadístico descriptivo y análisis estadístico inferencial.
- Distinguir los conceptos de muestra aleatoria y muestra representativa así como las características fundamentales de algunos tipos de muestreo.
- Conocer los aspectos básicos de la inferencia estadística (distribución muestral...).
- Realizar inferencias con intervalos de confianza para responder a problemas de investigación. Dos aspectos fundamentales de este procedimiento son el error de estimación y el tamaño de la muestra.

8.2. CONCEPTOS PREVIOS

En este apartado pretendemos estudiar, en primer lugar, los conceptos de población y muestra y su relación con el análisis estadístico descriptivo e inferencial, y en segundo lugar, el de muestreo.

8.2.1. Población y muestra

En el contexto estadístico, el término ***población*** se refiere al conjunto total de elementos en el que se quiere estudiar una o más características. Debe estar claramente definida (por ejemplo, los pacientes de un determinado hospital, las personas mayores de 65 años de una determinada comunidad autónoma...), lo que significa que cualquier elemento de la población puede clasificarse como perteneciente o no a ella. Llamaremos N al número total de los elementos de una población. Así, si la población está formada por 100.000 elementos, $N = 100.000$.

Por lo general, en la investigación psicológica, la población está formada por personas, pero la definición anterior de población contempla a cualquier conjunto de cosas o animales. Por ejemplo, la población podría estar formada por ratas de laboratorio como suele ocurrir en la investigación psicobiológica. Por otra parte, se suele utilizar los términos ***individuos***, ***sujetos*** y ***casos*** para referirse a los elementos de la población.

Cuando se dispone de un ***censo*** de la población, es decir, de un listado de todos los elementos de la población, se puede estudiar a todos ellos. En el ámbito profesional es común trabajar con la población entera: el profesor que está interesado en conocer la opinión de sus alumnos, el psicólogo que quiere estudiar la evolución de sus pacientes, el directivo de una empresa que quiere estudiar la satisfacción laboral de sus empleados, etcétera.

No obstante, no siempre es factible estudiar a la población en su totalidad ya sea porque la población a estudiar es muy grande, por motivos económicos, por el riesgo que implica (por ejemplo, aplicar un nuevo tratamiento a todos los enfermos de una determinada enfermedad para estudiar su eficacia podría tener consecuencias muy graves si el tratamiento resulta no ser eficaz) o para una mayor rapidez en la recogida de los datos. En esos casos, se estudia sólo un subconjunto del total de los elementos, es decir, una ***muestra*** de la población. Llamaremos n al número de los elementos de una muestra. Así, si la muestra está formada por 50 elementos, $n = 50$.

Por tanto, nos encontramos con dos contextos diferentes:

1. Investigar la población entera.
2. Investigar una muestra extraída de la población y luego inferir a la población.

Veamos la diferencia entre ambas estrategias de investigación y su relación con el análisis estadístico descriptivo e inferencial.

1. Un psicólogo desea conocer la efectividad de una terapia que está aplicando a los pacientes depresivos de su consulta, por lo que cal-

cula la proporción de pacientes curados. Dado que el estudio se hace con todos los pacientes depresivos de la consulta (población) no hay que aplicar técnicas de tipo inferencial, basta un índice descriptivo como la proporción de pacientes curados. Esta proporción es el parámetro que le indica la efectividad de la terapia a nivel de toda la población (los depresivos de la consulta).

2. Un investigador está interesado en conocer la efectividad de una nueva terapia para curar la depresión. Como es arriesgado aplicar la nueva terapia a todos los depresivos, el investigador la aplica a un subconjunto de la población (muestra) y obtiene la proporción de pacientes curados de la muestra (estadístico). Para inferir la proporción de pacientes que se curarían si se aplicara la terapia a la población entera (parámetro), el investigador necesitará herramientas estadísticas de tipo inferencial.

En este tema estudiaremos la segunda estrategia de investigación, es decir, aprenderemos a inferir un parámetro de una población a partir de una muestra aleatoria extraída de la población.

8.2.2. Muestreo

Para que las inferencias de la muestra a la población tengan sentido, no vale cualquier muestra. El ***muestreo*** es el proceso mediante el que se selecciona una muestra de una población con el fin de obtener una muestra lo más semejante posible a la población y así obtener estimaciones precisas. Hay que tener en cuenta que una muestra debe ser lo suficientemente amplia para representar adecuadamente las propiedades de la población y lo suficientemente reducida para que pueda ser examinada en la práctica. Por lo tanto, el tamaño es una característica esencial de una muestra.

Hay dos tipos de muestreo: el ***probabilístico*** y el ***no probabilístico***. En el probabilístico se conoce, o puede calcularse, la probabilidad asociada a una determinada muestra y cada elemento de la población tiene una probabilidad conocida, o calculable, de pertenecer a la muestra. En el muestreo no-probabilístico se desconoce, o no se tiene en cuenta, la probabilidad asociada a cada una de las muestras posibles y se selecciona la muestra que más representativa le parece al investigador o aquella que puede obtenerse más fácilmente (personas voluntarias, alumnos de una determinada clase...). Esto no quiere decir que con un muestreo no-probabilístico no podamos obtener, en determinados casos, muestras representativas de la población pero no tendremos ninguna garantía de que eso sea así y no podremos realizar inferencias a la población.

Una forma de obtener una muestra representativa es utilizar un procedimiento que garantice a todos y cada uno de los elementos de la población la misma probabilidad de formar parte de la muestra. En este principio se basa el ***muestreo aleatorio simple***.

Decimos que hemos extraído una muestra aleatoria simple cuando:

- Cada elemento de la población tiene la misma probabilidad de ser elegido.
- Los elementos se seleccionan de uno en uno, y con reposición, por lo que la población permanece idéntica en todas las extracciones. No obstante, cuando el tamaño de la población (N) es grande es indiferente que el muestreo sea con o sin reposición.

El procedimiento suele ser el siguiente: primero se asigna un número a cada elemento de la población y después mediante algún medio mecánico (papeletas en un cajón, bolas en una bolsa, tablas de números aleatorios, números aleatorios generados en un ordenador...) se elijen tantos elementos como sea necesario para completar el tamaño de la muestra.

Cuando los elementos de la población están ordenados o pueden ordenarse (por ejemplo, los alumnos de un determinado centro) podemos utilizar el ***muestreo sistemático*** en lugar del muestreo aleatorio simple. Supongamos, para simplificar, que la población tiene un tamaño $N = 100$ y se desea obtener una muestra de $n = 5$ entonces el muestreo se realizaría de la siguiente forma:

1. Seleccionamos al azar un elemento entre el primero y el que ocupa el lugar $\frac{N}{n} = \frac{100}{5} = 20.$ Imaginemos que obtenemos el número 15.
2. Completamos la lista sumando de 20 en 20, al valor obtenido anteriormente (15) hasta completar la muestra. Así, el resto de los elementos de la muestra serían: 35, 55, 75, 95.

El riesgo de este tipo de muestreo está en aquellos casos en que se dan periodicidades en la población ya que, al elegir con una periodicidad constante, los elementos seleccionados para la muestra pueden no ser representativos del conjunto total de elementos.

Los muestreos señalados anteriormente deben utilizarse cuando existe homogeneidad en la población. Cuando existen grupos o subpoblaciones heterogéneos, y disponemos de información suficiente podemos utilizar el ***muestreo estratificado***. Por ejemplo, si queremos estudiar alguna característica de los alumnos de un centro en el que se imparten las enseñanzas de Infantil, Primaria y Secundaria podemos elegir una muestra en función del número de alumnos en cada nivel de enseñanza o estrato.

Los métodos anteriores requieren disponer de un listado de los elementos de la población o poder elaborarlo fácilmente. Cuando esto no es posible, podemos utilizar el ***muestreo por conglomerados***. Si quisiéramos, por ejemplo, extraer una muestra de los universitarios españoles podemos proceder de la siguiente manera: seleccionaríamos al azar primero algunas universidades, luego algunas facultades dentro de cada universidad, después algunos cursos y, finalmente, todos los alumnos de los cursos seleccionados. Llamamos «conglomerados» a estas unidades en que se clasifican los elementos de la población. Si los conglomerados son heterogéneos, este método puede llevarnos a muestras poco representativas puesto que sólo se analizan algunos de ellos.

Hay otro tipo de muestreo, denominado ***polietápico***, que es una combinación de los dos anteriores (estratificado y por conglomerados).

En ocasiones, el muestreo probabilístico resulta demasiado costoso y se acude a métodos no probabilísticos. Entre ellos se encuentran los siguientes:

- El muestreo ***por cuotas*** (o accidental): se basa en un buen conocimiento de los estratos o individuos «más representativos» o adecuados para los fines de la investigación a realizar. Es por tanto, semejante al muestreo estratificado pero carece del carácter aleatorio de éste.
- El muestreo ***opinático*** (o intencional): se caracteriza por el interés de incluir en la muestra a grupos supuestamente típicos. Su uso es frecuente, por ejemplo, en sondeos preelectorales de zonas que en anteriores ocasiones han marcado la tendencia de voto.

- El muestreo ***casual*** (o incidental): se selecciona directamente a individuos o elementos de la población a los que se tiene fácil acceso (por ejemplo, los profesores emplean a sus alumnos).
- El denominado «***bola de nieve***»: se caracteriza porque un elemento de la población lleva a otro y este, a su vez, a otro... hasta completar la muestra. Suele utilizarse en estudios con poblaciones de difícil acceso (delincuentes, sectas, determinado tipo de enfermos...).

Por último, hemos de señalar que las nociones de ***muestra representativa*** y ***muestra aleatoria*** se refieren a aspectos distintos aunque ambos deseables de una muestra. Una muestra es representativa si exhibe internamente el mismo grado de diversidad que la población y una muestra es aleatoria si los elementos han sido extraídos al azar de la población.

8.3. INFERENCIA ESTADÍSTICA

Hemos obtenido una muestra aleatoria de una población. Ahora bien, en investigación interesa estudiar ciertas características de los elementos de la población, como puede ser la inteligencia emocional, la agresividad, el tiempo de reacción a un estímulo, el nivel de colesterol, el nivel de las defensas del sistema inmunológico, la opinión (sí/no) sobre algún tema, etcétera.

Las medidas de estas características obtenidas en una muestra pueden resumirse mediante estadísticos como la media (por ejemplo, el tiempo de reacción medio), la proporción (por ejemplo, la proporción de respuestas afirmativas), etc.

Pero una muestra es sólo un subconjunto de la población por lo que el valor del estadístico obtenido en la muestra (como la media) no será igual, por lo general, al valor del parámetro de la población. Para inferir un parámetro a partir de un estadístico hay que aplicar herramientas estadísticas de tipo inferencial como la ***estimación por intervalo*** (intervalos de confianza) o el contraste de hipótesis.

8.4. ESTIMACIÓN DE LA MEDIA

La media[1] muestral es una variable aleatoria (concepto ya estudiado en el tema 6) que toma un valor u otro según la muestra concreta que se obtenga. En realidad, tendremos tantas medias como posibles muestras del mismo tamaño podamos extraer de la población. Se denomina ***distribución muestral de la media*** a su función de probabilidad.

Queremos señalar que la ***distribución muestral de un estadístico*** es un concepto central de la inferencia estadística, tanto de la estimación por intervalo como del contraste de hipótesis.

8.4.1. Distribución muestral de la media

Ilustraremos con un ejemplo sencillo, de carácter exclusivamente didáctico, cómo obtener la distribución muestral de la media y las principales características de dicha distribución: su media, varianza (y desviación típica) y forma.

Ejemplo 8.1. Sea una población formada por 5 sujetos ($N = 5$) que en la variable X tienen las siguientes puntuaciones: 1, 2, 3, 4 y 5. La figura 8.2 recoge, junto a la tabla de frecuencias, la representación gráfica de la distribución de esa variable en la población, su media (μ) y su varianza (σ^2).

X	n_i	$f(X)$
1	1	0,2
2	1	0,2
3	1	0,2
4	1	0,2
5	1	0,2

$\mu = 3$
$\sigma^2 = 2$

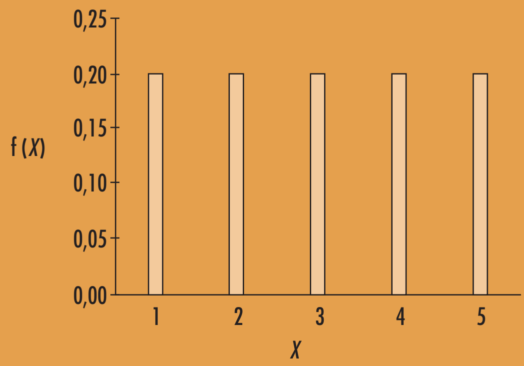


Figura 8.2. Tabla de frecuencias y distribución de la variable X en la población.

[1] Cuando hablemos de media en este tema, nos referiremos siempre a la media aritmética.

Extraemos de esa población, al azar y con reposición, todas las muestras posibles de tamaño $n = 2$. Cada uno de los elementos de esta población tiene una probabilidad $\frac{1}{5} = 0{,}2$ de ser elegido tanto en la primera como en la segunda extracción (es con reposición). El conjunto de muestras posibles es:

1,1	2,1	3,1	4,1	5,1
1,2	2,2	3,2	4,2	5,2
1,3	2,3	3,3	4,3	5,3
1,4	2,4	3,4	4,4	5,4
1,5	2,5	3,5	4,5	5,5

Para cada una de estas muestras podemos calcular su media. Esta media varía para las distintas muestras, como puede observarse en la siguiente tabla donde se recoge también la probabilidad de cada una de ellas:

Muestra	**Valores de X en la muestra**	$\bar{X}$	**Probabilidad**
1	1,1	1	1/25
2	1,2	1,5	1/25
3	1,3	2	1/25
4	1,4	2,5	1/25
5	1,5	3	1/25
6	2,1	1,5	1/25
7	2,2	2	1/25
8	2,3	2,5	1/25
9	2,4	3	1/25
10	2,5	3,5	1/25
11	3,1	2	1/25
12	3,2	2,5	1/25
13	3,3	3	1/25
14	3,4	3,5	1/25
15	3,5	4	1/25
16	4,1	2,5	1/25

(Continúa)

(Continuación)

Muestra	Valores de *X* en la muestra	$\bar{X}$	Probabilidad
17	4,2	3	1/25
18	4,3	3,5	1/25
19	4,4	4	1/25
20	4,5	4,5	1/25
21	5,1	3	1/25
22	5,2	3,5	1/25
23	5,3	4	1/25
24	5,4	4,5	1/25
25	5,5	5	1/25

Por tanto, podemos considerar el estadístico media como una variable aleatoria que toma una serie de valores, cada uno de ellos con su correspondiente probabilidad. Pues bien, la distribución muestral de la media será su correspondiente función de probabilidad (Figura 8.3.).

$\bar{X}$	n_i	$f(\bar{X})$
1	1	0,04
1,5	2	0,08
2	3	0,12
2,5	4	0,16
3	5	0,20
3,5	4	0,16
4	3	0,12
4,5	2	0,08
5	1	0,04
		1

$$\mu_{\bar{X}} = 3$$

$$\sigma^2_{\bar{X}} = 1$$

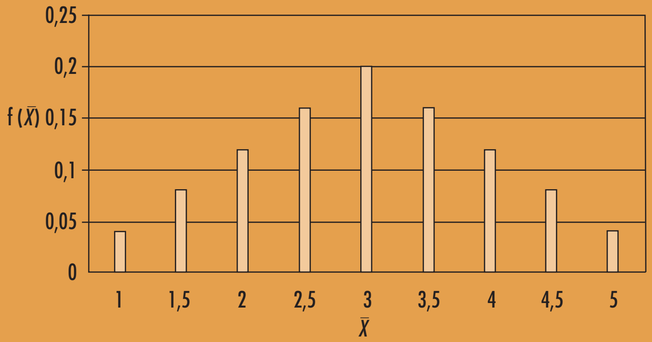


Figura 8.3. Tabla de frecuencias y distribución muestral de la media ($\bar{X}$) para n = 2.

Como ya hemos señalado en el tema 6, una función de probabilidad queda caracterizada por su forma, su media y su varianza. Siguiendo con el ejemplo, la media de la distribución muestral de la media (que designaremos por $\mu_{\bar{X}}$) es 3 y la varianza (que designaremos por $\sigma^2_{\bar{X}}$) es 1. En el ejemplo observamos que:

1. La media de la distribución muestral de la media ($\mu_{\bar{X}}$) es igual a la media de la población (μ).
2. La varianza de la distribución muestral de la media es $\frac{\sigma^2}{n}$ y la desviación típica de la distribución muestral de la media, denominada ***error típico de la media***, es $\sigma_{\bar{X}} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$.
3. La forma de la distribución muestral de la media (figura 8.3) se «parece» a una distribución normal (estudiada en el tema anterior) aunque la distribución original de la variable en la población no es normal (figura 8.2).

Como ejercicio, comprueben a partir de las tablas de frecuencias de la variable X y de la $\bar{X}$ que:

— La media, la varianza y la desviación típica de la población son respectivamente: $\mu = 3$, $\sigma^2 = 2$ y $\sigma = 1{,}41$.

— La media, la varianza y la desviación típica de la distribución muestral de la media son respectivamente: $\mu_{\bar{X}} = 3$, $\sigma^2_{\bar{X}} = 1$ y $\sigma_{\bar{X}} = 1$.

Por lo tanto, hemos verificado las dos primeras propiedades:

$$\mu_{\bar{X}} = \mu = 3$$

$$\sigma^2_{\bar{X}} = \frac{\sigma^2}{n} = \frac{2}{2} = 1 \rightarrow \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{\sqrt{2}}{\sqrt{2}} = \sqrt{\frac{2}{2}} = 1$$

Para ilustrar mejor la tercera propiedad, supongamos que tenemos una población grande cuya distribución no es normal y extraemos muchas muestras al azar de tamaño $n = 30$. Observen en la figura 8.4, que aunque la distribución de la variable X en la población no es normal (es uniforme), la distribución muestral de la media ($\bar{X}$), para $n = 30$, es muy próxima a la normal.

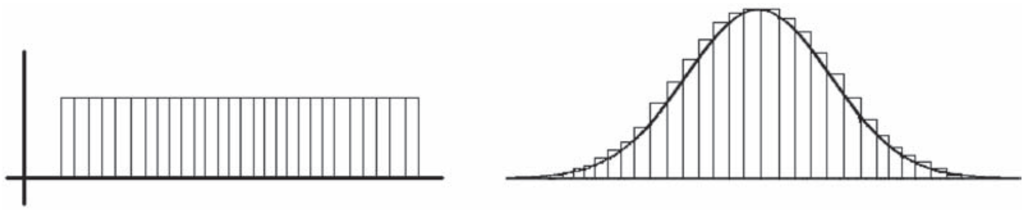


Figura 8.4. Distribución de la variable X y distribución muestral de la media ($\bar{X}$) para muestras de tamaño n = 30.

Volviendo al ejemplo 8.1, hemos obtenido empíricamente la distribución muestral de la media para una población N = 5 siendo n = 2. En realidad las poblaciones son mucho más grandes y las muestras también son más grandes, por lo que en la práctica no es posible (ni es necesario) obtener la distribución muestral como en el ejemplo expuesto[2]. De hecho, podemos conocer las características de la distribución muestral de la media a partir de ciertos teoremas que resumimos a continuación.

Dado el muestreo aleatorio simple:

- Si la distribución de X en la población es normal con media μ y desviación típica σ, entonces la distribución muestral de la $\bar{X}$ es normal $\left(\mu, \dfrac{\sigma}{\sqrt{n}}\right)$.
- Si la distribución de X en la población no es normal con media μ y desviación típica σ, entonces la distribución muestral de la $\bar{X}$ tiende a la normal $\left(\mu, \dfrac{\sigma}{\sqrt{n}}\right)$ a medida que n crece ***(Teorema Central del Límite),*** siendo la aproximación buena para $n \geq 30$.

Hemos estudiado la distribución del estadístico media. Su conocimiento permitirá realizar inferencias sobre la media poblacional, con cierta probabilidad, a partir de la media muestral. Veremos más adelante cómo se realizan las inferencias mediante intervalos de confianza.

[2] El propósito del ejemplo es que el estudiante comprenda qué es y cómo se origina la distribución muestral de un estadístico, en este caso, la media.

Para una mayor claridad, recogemos en la tabla 8.1. la media, la varianza y la desviación típica de la variable X en la población y en la muestra, y de la distribución muestral de la media.

Tabla 8.1. Media, varianza y desviación típica de la variable cuantitativa X en la población y en la muestra, y de la distribución muestral de la media ($\bar{X}$).

	Población	**Muestra**	**Distribución muestral de la media**
Media	$\mu = \frac{\sum X}{N}$	$\bar{X} = \frac{\sum X}{n}$	$\mu_{\bar{X}} = \mu$
Varianza	$\sigma^2 = \frac{\sum (X-\mu)^2}{N}$	$S_{n-1}^2 = \frac{\sum (X-\bar{X})^2}{n-1}$ Cuasivarianza[3]	$\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}$
Desviación típica	$\sigma = \sqrt{\frac{\sum (X-\mu)^2}{N}}$	$S_{n-1} = \sqrt{\frac{\sum (X-\bar{X})^2}{n-1}}$ Cuasidesviación típica	$\sigma_{\bar{X}} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$ Error típico de la media

Observen la diferencia entre desviación típica de la población, desviación típica de la muestra (cuasidesviación típica) y desviación típica de la distribución muestral de la media (error típico de la media).

La desviación típica de la población es una medida de la variabilidad de la variable X en la población.

La desviación típica de la muestra (cuasidesviación típica) es una medida de la variabilidad de la variable X en la muestra. Como veremos en las próximas páginas, cuando desconozcamos σ utilizaremos S_{n-1}.

La desviación típica de la distribución muestral de la media (error típico de la media) representa el grado de variabilidad entre los valores de las medias muestrales. Cuanto mayor es el error típico de la media, más imprecisa es la estimación.

[3] Se utiliza S_{n-1}^2 en lugar de S^2 porque S_{n-1}^2 es un estimador insesgado de la varianza poblacional (σ^2) mientras que S^2 no lo es.

8.4.2. La media como estimador

Cuando se utiliza un estadístico para estimar un parámetro se le llama ***estimador***. En este sentido, la media de la muestra es un estimador de la media poblacional. Y el valor que toma el estimador en una muestra concreta se denomina ***estimación*** o ***estimación puntual***.

Vimos en el apartado anterior que la media de la distribución muestral de la media es igual a la media poblacional ($\mu_{\bar{X}} = \mu$). Esta circunstancia indica que la media muestral $\bar{X}$ es un ***estimador insesgado*** de la media poblacional (μ).

La desviación típica de la distribución muestral de la media, es decir, el ***error típico de la media*** es un indicador de la ***precisión de la estimación de la media:*** cuanto menor es el error típico mayor es la precisión. Dado el muestreo aleatorio simple, es igual a:

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

Por lo tanto, el error típico de la media depende de la desviación típica de la población σ y del tamaño de la muestra n. Observen:

- Cuanto menor es la desviación típica de la población, menor será el error típico de la media.
- Cuanto mayor es n, menor será el error típico de la media.

8.5. ESTIMACIÓN DE LA PROPORCIÓN

La proporción muestral es una variable aleatoria que toma un valor u otro según la muestra concreta que se obtenga. Por ello, podríamos obtener empíricamente la distribución muestral de la proporción de una forma similar a la que hicimos para la media, pero lo omitimos porque no aporta nada nuevo en cuanto al procedimiento. Presentaremos, en el siguiente apartado, únicamente las principales características de dicha distribución: su media, varianza (y desviación típica) y forma.

8.5.1. Distribución muestral de la proporción

Sea X una variable que sólo toma valores 0 y 1, la proporción de la muestra P se define como:

$$P = \frac{\sum X}{n}$$

Dado el muestreo aleatorio simple (por lo que π permanece constante en cada extracción), el estadístico proporción (P) se distribuye según una binomial con $\mu_P = \pi$ y $\sigma_P^2 = \frac{\pi(1-\pi)}{n}$.

Como P es la media de los valores de X en la muestra (donde X toma valores 0 y 1), entonces según el ***Teorema Central del Límite***, a medida que el tamaño de la muestra crece, la distribución muestral de la proporción tiende a la normal con media π y varianza $\frac{\pi(1-\pi)}{n}$.

Cuanto más alejado esté π de 0,5, más elementos debe tener la muestra para realizar la aproximación a la normal (figura 8.5).

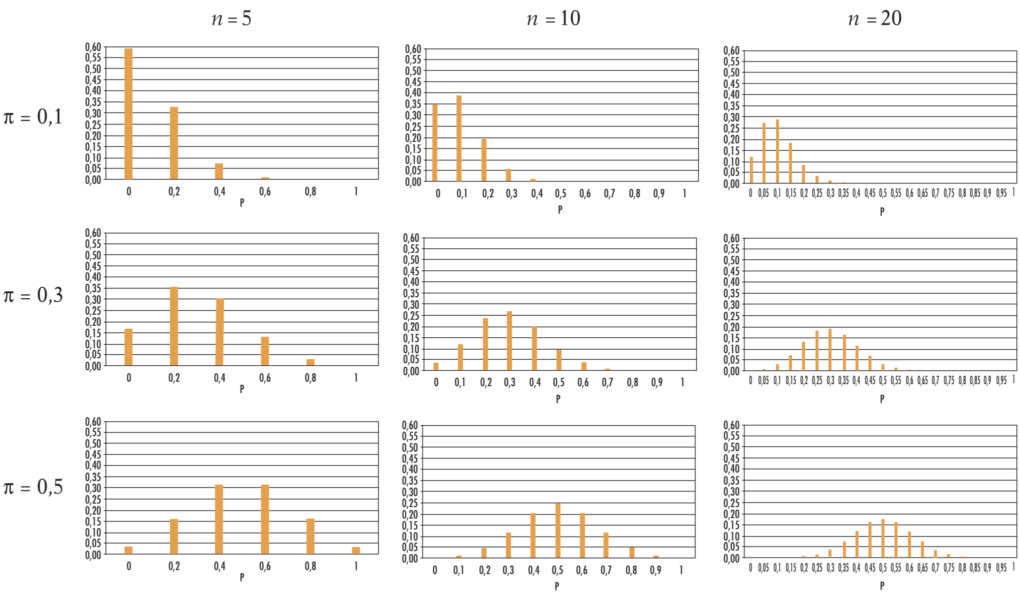


Figura 8.5. Distribución muestral de la proporción en función de n y π.

Como pueden observar, la aproximación a la normal es buena para $\pi = 0{,}5$ y $n = 20$. En la práctica, se suele aplicar el criterio siguiente: $n\,\pi\,(1 - \pi) \geq 5$, por lo que el tamaño muestral mínimo requerido se obtiene a partir de: $n \geq \dfrac{5}{\pi(1-\pi)}$.

Así, si $\pi = 0{,}5$,

$$n = \frac{5}{(0{,}50)(0{,}50)} = 20$$

Comprueben que según el valor de π, el tamaño mínimo de la muestra para realizar la aproximación a la normal estará entre 20 y 56 (tabla 8.2).

Tabla 8.2. Relación entre n y π para la aproximación a la normal

π	0,10	0,20	0,30	0,40	0,50	0,60	0,70	0,80	0,90
n	56	32	24	21	20	21	24	32	56

En la tabla 8.3, se recogen la media, varianza y desviación típica de la variable X en la población y en la muestra, y de la distribución muestral de la proporción. Respecto a esta última, vean que:

1. La media de la distribución muestral de la proporción (μ_P) es igual a la proporción de la población (π).
2. La varianza de la distribución muestral de la proporción es $\sigma_P^2 = \dfrac{\pi(1-\pi)}{n}$.
3. La desviación típica de la distribución muestral de la proporción, llamada ***error típico de la proporción***, es $\sigma_P = \sqrt{\dfrac{\pi(1-\pi)}{n}}$.

Tabla 8.3. Media, varianza y desviación típica de la variable dicotómica o dicotomizada (*X*) en la población y en la muestra, y de la distribución muestral de la proporción (*P*)

	Población	Muestra	Distribución muestral de la proporción (*P*)
Media	$\pi = \frac{\sum X}{N}$ donde X: 0,1	$P = \frac{\sum X}{n}$ donde X: 0,1	$\mu_P = \pi$
Varianza	$\sigma^2 = \pi(1-\pi)$	$S^2 = P(1-P)$	$\sigma_P^2 = \frac{\pi(1-\pi)}{n}$
Desviación típica	$\sigma = \sqrt{\pi(1-\pi)}$	$S = \sqrt{P(1-P)}$	$\sigma_P = \sqrt{\frac{\pi(1-\pi)}{n}}$ Error típico de la proporción

8.5.2. La proporción como estimador

Vimos en el apartado anterior que la media de la distribución muestral de la proporción es igual a la proporción poblacional ($\mu_p = \pi$), por lo que la proporción muestral (P) es un ***estimador insesgado*** de la proporción poblacional (π).

La desviación típica de la distribución muestral de la proporción, es decir, el ***error típico de la proporción,*** es un indicador de la ***precisión de la estimación de la proporción***: cuanto menor es el error típico mayor es la precisión. Dado el muestreo aleatorio simple, es igual a:

$$\sigma_P = \sqrt{\frac{\pi(1-\pi)}{n}}$$

El error típico de la proporción depende de la desviación típica de la población $\sqrt{\pi(1-\pi)}$ y el tamaño de la muestra n. Observen:

- Cuanto menor es la desviación típica de la población, menor será el error típico de la proporción.
- Cuanto mayor es n, menor será el error típico de la proporción.

8.6. INTERVALOS DE CONFIANZA

La inferencia estadística básicamente consiste en estimar, con cierta probabilidad, el parámetro desconocido a partir de una muestra aleatoria extraída de la población. En este apartado estudiaremos la forma de realizar inferencias sobre un parámetro mediante intervalos de confianza.

Empezaremos por exponer el concepto de intervalo de confianza, después las cuestiones relacionadas con el tamaño de la muestra y finalmente veremos distintas aplicaciones del intervalo de confianza.

8.6.1. Concepto

Para entender el concepto de ***intervalo de confianza***, lo expondremos aplicado a la media, dados los siguientes supuestos: muestreo aleatorio simple, variable cuantitativa, distribución de la variable en la población normal, σ conocida.

La finalidad de un intervalo de confianza es estimar un parámetro desconocido de una población a partir de una muestra.

Al estimar la media de la población a partir de una muestra, podemos cometer un ***error de estimación*** que se define como $|\bar{X} - \mu|$. Desconocemos ese error dado que no conocemos μ, que es lo que precisamente queremos estimar.

La estimación por intervalo consiste en acotar el error de estimación con una alta probabilidad $1-\alpha$ (llamada ***nivel de confianza***) de tal manera que $|\bar{X} - \mu|$ no sea superior a un error de estimación máximo ($E_{\text{máx}}$) fijado por el investigador: $|\bar{X} - \mu| \leq E_{\text{máx}}$.

El ***error de estimación máximo*** ($E_{\text{máx}}$) es función de la variabilidad de la variable en la población, del nivel de confianza (n.c.) y del tamaño de la muestra:

$$E_{\text{máx}} = z_{1\text{-}\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Donde:

- $z_{1-\alpha/2}$ es función del n.c. = $1 - \alpha$ y se obtiene en la tabla de la distribución normal tipificada (tabla IV). Los valores más comunes del n.c. son: 0,95, 0,99 y 0,999.
- $\dfrac{\sigma}{\sqrt{n}}$ es la desviación típica de la distribución muestral de la media, es decir, el error típico de la media: $\sigma_{\bar{X}}$.

- σ es la desviación típica de la población que es conocida.
- n es el tamaño de la muestra.

La ecuación $E_{\text{máx}} = z_{1\text{-}\alpha/2}\dfrac{\sigma}{\sqrt{n}}$ es básica, a partir de ella deduciremos tanto el tamaño de la muestra como los límites del intervalo de confianza. Empezaremos por el tamaño de la muestra.

Cuanto mayor sea el tamaño de la muestra, mayor será la precisión de las estimaciones de los parámetros. No obstante, hay razones como las expuestas en el apartado 8.2.1, que imponen límites al tamaño de la muestra. Por ello, interesa saber cuál debe ser el tamaño de la muestra para un $E_{\text{máx}}$ dado. El ***tamaño de la muestra*** se obtiene despejando n de la ecuación:

$$E_{\text{máx}} = z_{1\text{-}\alpha/2}\frac{\sigma}{\sqrt{n}} \rightarrow n = \frac{z^2_{1-\alpha/2}\,\sigma^2}{E^2_{máx}}$$

Veamos un ejemplo:

Ejemplo 8.2. Un investigador quiere conocer el tiempo de reacción en una tarea de discriminación (en la que hay que elegir entre dos alternativas de respuesta) en niños de 12 años. La variable *tiempo de reacción en la tarea de discriminación* se distribuye normalmente en la población con $\sigma = 3$. Decide realizar una estimación por intervalo del parámetro μ desconocido (el tiempo de reacción medio en la tarea de discriminación de la población) y fija un error de estimación máximo de 1 segundo para un n.c. = 0,95. ¿Cuál debe ser el tamaño de la muestra para estimar la media?:

Solución:

n.c. = 0,95 $\rightarrow z_{1-\alpha/2} = z_{1-0,05/2} = z_{0,975} = 1,96$ (Tabla IV)

$$n = \frac{z^2_{1-\alpha/2}\,\sigma^2}{E^2_{máx}} = \frac{1,96^2 \cdot 3^2}{1^2} = 34,57 \rightarrow 35$$

Observen que n = 35 es el tamaño muestral mínimo para no superar un $E_{\text{máx}}$ = 1, con un n.c. = 0,95.

Los ***límites inferior (L_i) y superior (L_s) del intervalo de confianza*** se obtienen a partir del $E_{\text{máx}}$:

$$L_i = \bar{X} - E_{\text{máx}}$$
$$L_s = \bar{X} + E_{\text{máx}}$$

Es decir,

$$L_i = \bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$$L_s = \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Ejemplo 8.3. Continuando con el ejemplo anterior, el investigador extrae una muestra aleatoria simple de n = 35 niños de 12 años, les mide el tiempo de reacción medio en la tarea de discriminación y obtiene $\bar{X}$ = 4 segundos. ¿Cuál es el intervalo de confianza?

Solución:

El investigador tiene la siguiente información de la muestra $\bar{X}$ = 4 (el tiempo de reacción medio de su muestra) y quiere saber cuál sería el tiempo de reacción medio si se aplicara la tarea de discriminación a todos los niños de 12 años.

Como ya sabíamos, el error de estimación máximo es 1:

$$E_{\text{máx}} = z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} = 1{,}96\frac{3}{\sqrt{35}} = 0{,}99 \approx 1$$

Los límites del intervalo de confianza son:

$$L_i = \bar{X} - z_{1\text{-}\alpha/2}\frac{\sigma}{\sqrt{n}} = 4 - 1 = 3$$

$$L_s = \bar{X} + z_{1\text{-}\alpha/2}\frac{\sigma}{\sqrt{n}} = 4 + 1 = 5$$

Y, por tanto, la probabilidad de obtener un intervalo de confianza que contenga al parámetro μ es:

$$P(\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1-\alpha$$

$$P(\bar{X} - E_{\text{máx}} \leq \mu \leq \bar{X} + E_{\text{máx}}) = 1-\alpha$$

$$P(\bar{X} - 1 \leq \mu \leq \bar{X} + 1) = 0,95$$

El nivel de confianza o ***probabilidad 1 – α*** asociado al intervalo de confianza significa que si extrajésemos todas las muestras posibles de una población mediante muestreo aleatorio simple, calculásemos la media en cada una de ellas (recuerde la distribución muestral de la media) y para cada media calculáramos el intervalo de confianza, una proporción $1 - \alpha$ de todos los intervalos de confianza contendrá la media poblacional y una proporción α no lo contendrá.

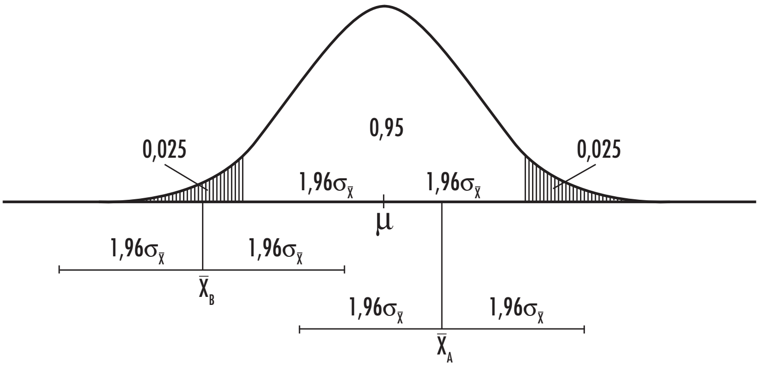


Figura 8.6. Distribución muestral de la media y dos posibles intervalos de confianza del 95%, uno que contiene el parámetro μ y otro que no lo contiene.

Vemos en la figura 8.6. que el intervalo de confianza asociado a $\bar{X}_A$ contiene el parámetro μ y el intervalo de confianza asociado a $\bar{X}_B$ no lo contiene. Imagínese que representáramos en la figura todos los posibles intervalos de confianza. Pues bien, $1 - \alpha = 0,95$ significa que el 95% de los intervalos de confianza contendrá el parámetro μ y el 5% no lo contendrá.

Por otra parte, vemos en la figura 8.6. que $\bar{X}_A$ está dentro de la zona no rayada y el intervalo de confianza contiene al parámetro μ mientras que $\bar{X}_B$ está en la zona rayada y el intervalo de confianza no contiene al parámetro m. Pues bien, para un n.c. = 0,95:

1. Cualquier valor de la media que pertenezca a la zona rayada (tanto en un lado como en el otro de la distribución) proporcionará un intervalo de confianza que no contendrá al parámetro y la probabilidad de que ello ocurra es $\alpha = 0{,}025 + 0{,}025 = 0{,}05$.
2. Cualquier valor de la media que pertenezca a la zona no rayada proporcionará un intervalo de confianza que contendrá al parámetro y la probabilidad de que ello ocurra es $1 - \alpha = 0{,}95$.

Observe en la figura 8.6. que la ***amplitud del intervalo*** es dos veces el error de estimación máximo ($2E_{\text{máx}}$) y es constante. Es decir, la amplitud es siempre la misma independientemente del intervalo de confianza que se obtenga. Lo que varía es el valor de la media y es lo que hace variar los límites del intervalo de confianza pero no su amplitud.

Respecto a la ***precisión de la estimación***, es fácil imaginar en la figura 8.6. que cuanto menor es el error de estimación máximo, menor es la amplitud del intervalo y más precisa es la estimación, dado que el rango de valores que recoge el intervalo es más estrecho.

Hemos visto el concepto de intervalo de confianza referido a la media, el concepto puede generalizarse a otros parámetros.

8.6.2. Tamaño de la muestra

Como vimos en el apartado anterior, el tamaño de la muestra para la estimación por intervalo de la media dada una variable X distribuida normalmente con σ conocida, se obtiene despejando n de la ecuación del error de estimación máximo de la media:

$$E_{\text{máx}} = z_{1\text{-}\alpha/2}\frac{\sigma}{\sqrt{n}} \rightarrow n = \frac{z^2_{1-\alpha/2}\,\sigma^2}{E^2_{máx}}$$

Vemos que n depende de tres factores:

- La desviación típica de la población.
- El nivel de confianza.
- El error de estimación máximo.

En este apartado, estudiaremos las relaciones entre el tamaño de la muestra y estos tres factores.

Ejemplo 8.4. Supongamos que en una investigación la variable se distribuye normalmente en la población con $\sigma = 4$ y queremos que el error de estimación máximo no sea mayor que 2,5 con un nivel de confianza de 0,95. ¿Qué tamaño debe tener la muestra para estimar la media?

Solución:

Supuestos: muestreo aleatorio simple, variable cuantitativa, distribución de la variable en la población normal, σ conocida.

n.c. = 0,95 $\rightarrow z_{1-\alpha/2} = z_{0,975} = 1{,}96$ (tabla IV)

$$n = \frac{1{,}96^2 \cdot 4^2}{2{,}5^2} = 9{,}84 \rightarrow 10$$

Interesa que un intervalo de confianza sea lo más estrecho posible y que la probabilidad del intervalo sea la más alta posible. Lamentablemente, a mayor nivel de confianza mayor es el error de estimación máximo, por lo que más amplio será el intervalo y menos precisa será la estimación. Una forma de mantener un error de estimación máximo dado (por ejemplo, $E_{\text{máx}} = 2{,}5$) y aumentar el n.c., es aumentando n.

Ejemplo 8.5. Supongamos que en el ejemplo 8.4. el nivel de confianza es 0,99.

Solución:

n.c. = 0,99 $\rightarrow z_{1-\alpha/2} = z_{0,995} = 2{,}58$ (Tabla IV)

$$n = \frac{2{,}58^2 \cdot 4^2}{2{,}5^2} = 17$$

Observamos que para un mismo $E_{\text{máx}} = 2{,}5$, es necesario un tamaño muestral mayor si el nivel de confianza es 0,99 que si es 0,95.

Una forma de reducir el error de estimación máximo y por lo tanto aumentar la precisión de la estimación es aumentando n.

Ejemplo 8.6. Supongamos que en otro estudio, el investigador del ejemplo 8.4. quiere reducir el error de estimación máximo a la mitad ¿cuánto debe valer n?

Solución:

n.c. = 0,95 $\rightarrow z_{1-\alpha/2} = z_{0,975} = 1,96$ (Tabla IV)

$$n = \frac{1,96^2 \cdot 4^2}{(2,5/2)^2} = 39,33 \rightarrow 40$$

Observamos que al reducir $E_{\text{máx}}$, el tamaño muestral requerido es mayor que el requerido en el ejemplo 8.4.

Hemos visto cómo interactúan el tamaño muestral, el nivel de confianza y el error de estimación máximo. Otro factor que interviene en la precisión de la estimación es la variabilidad de la variable, cuanto mayor sea la desviación típica de la población mayor debe ser n para alcanzar una misma precisión.

Ejemplo 8.7. Supongamos que en el ejemplo 8.4, σ hubiese sido igual a 5.

Solución:

n.c. = 0,95 $\rightarrow z_{1-\alpha/2} = z_{0,975} = 1,96$ (Tabla IV)

$$n = \frac{1,96^2 \cdot 5^2}{2,5^2} = 15,37 \rightarrow 16$$

Observamos que el tamaño muestral requerido es mayor que el requerido en el ejemplo 8.4. donde σ valía 4.

Para simplificar la exposición, hemos asumido que se conocía σ pero lo común es que cuando se desconoce la media de la población (μ) también se desconoce la desviación típica de la población (σ). Para calcular el tamaño de la muestra cuando se desconoce σ, hay que sustituir en la ecuación $n = \frac{z^2_{1-\alpha/2}\sigma^2}{E^2_{\text{máx}}}$, la desviación típica de la población (σ) por la cuasidesviación típica (S_{n-1}) de un estudio previo y $z_{1-\alpha/2}$ por $t_{n-1,\,1-\alpha/2}$ (tabla VI). En este caso, la obtención de n es algo más laboriosa[4]. Sepa el estudiante que hay programas informáticos que permiten calcular el tamaño de la muestra con mucha facilidad para situaciones como ésta u otras más complejas y para otros parámetros.

8.6.3. Aplicaciones

En las páginas anteriores hemos visto el concepto de intervalo de confianza y la forma de obtener el tamaño de la muestra y sus relaciones con la variabilidad, el nivel de confianza y el error de estimación máximo. Ahora, veremos distintas aplicaciones del intervalo de confianza.

Los pasos para aplicar un intervalo de confianza son los siguientes:

1. Establecer un error de estimación máximo para un nivel de confianza $1 - \alpha$.
2. Obtener el tamaño de la muestra n para el error de estimación máximo especificado.
3. Extraer una muestra aleatoria de tamaño n y medir la variable.
4. Calcular el estadístico (el estimador del parámetro) con las medidas obtenidas.
5. Calcular los límites del intervalo de confianza.

[4] Ocurre que en la ecuación, n (lo que queremos calcular) aparece en $t_{n-1,\,1-\alpha/2}$. Es decir, para buscar $t_{n-1,\,1-\alpha/2}$ en la tabla VI, necesitamos conocer n, que es precisamente lo que queremos calcular. Una solución es calcular n por aproximaciones sucesivas.

8.6.3.1. Intervalo de confianza para la media

Aplicaremos el intervalo de confianza para la media, para el caso de una variable X con distribución normal y σ conocida. Así que retomamos el enunciado del Ejemplo 8.2.

Ejemplo 8.8. Un investigador quiere conocer el tiempo de reacción en una tarea de discriminación (en la que hay que elegir entre dos alternativas de respuesta) en niños de 12 años. La variable *tiempo de reacción en la tarea de discriminación* se distribuye normalmente en la población con σ = 3. Decide realizar una estimación por intervalo del parámetro μ desconocido (el tiempo de reacción medio en la tarea de discriminación de la población) y fija un error de estimación máximo de 1 segundo para un n.c. = 0,95. ¿Cuál debe ser el tamaño de la muestra para estimar la media?:

Solución:

Supuestos: muestreo aleatorio simple, variable cuantitativa, distribución de la variable normal en la población, σ conocida.

n.c. = 0,95 $\rightarrow z_{1-\alpha/2} = z_{1-0,05/2} = z_{0,975} = 1,96$ (Tabla IV)

$$n = \frac{z^2_{1-\alpha/2}\sigma^2}{E^2_{\text{máx}}} = \frac{1,96^2.3^2}{1^2} = 34,57 \rightarrow 35$$

El investigador extrae una muestra aleatoria simple de n = 35 niños de 12 años, les mide el tiempo de reacción en la tarea de discriminación y obtiene $\bar{X}$ = 4 segundos. ¿Cuál es el intervalo de confianza?

El investigador tiene la siguiente información de la muestra $\bar{X}$ = 4 (el tiempo de reacción medio de su muestra) y quiere saber cuál sería el tiempo de reacción medio si se aplicara la tarea de discriminación a todos los niños de 12 años.

Como ya sabíamos, el error de estimación máximo es 1:

$$E_{\text{máx}} = z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} = 1,96\frac{3}{\sqrt{35}} = 0,99 \approx 1$$

Los límites del intervalo de confianza son:

$$L_i = \bar{X} - z_{1\text{-}\alpha/2}\frac{\sigma}{\sqrt{n}} = 4 - 1 = 3$$

$$L_s = \bar{X} + z_{1\text{-}\alpha/2}\frac{\sigma}{\sqrt{n}} = 4 + 1 = 5$$

Para un nivel de confianza del 95%, se estima que si se aplicara la tarea de discriminación a todos los niños de 12 años, el tiempo de reacción medio estaría entre 3 y 5 segundos.

Si la distribución no fuera normal (pero con $n \geq 30$) con σ conocida, los límites del intervalo de confianza se obtendrían de la misma manera que en el caso que acabamos de exponer.

Aplicaremos ahora el intervalo de confianza para la media, para el caso de una variable X distribuida normalmente y σ desconocida.

Ejemplo 8.9. Un investigador quiere conocer el nivel de defensas del sistema inmunológico de los varones españoles sometidos a risoterapia. Sabe que la variable *nivel de defensas* se distribuye normalmente pero desconoce tanto la media como la desviación típica de la población. Decide calcular el intervalo de confianza para el nivel medio de defensas y establece un error de estimación máximo de 5 unidades para un nivel de confianza del 99%. Calcula el tamaño de la muestra para ese error de estimación máximo y obtiene $n = 13$. A continuación, el investigador extrae una muestra aleatoria simple de 13 varones, les somete a risoterapia durante un mes, les mide el nivel de defensas y obtiene el nivel medio de defensas, $\bar{X} = 25$, y la cuasidesviación típica, $S_{n-1} = 6$. ¿Cuál es el intervalo de confianza?

Solución:

Supuestos: muestreo aleatorio simple, variable cuantitativa, distribución de la variable normal en la población, σ desconocida.

n.c. = 0,99 → como se desconoce σ → Distribución t de Student con $n - 1$ grados de libertad → $t_{n-1;1-\alpha/2} = t_{12;1-0,01/2} = t_{12;0,995} = 3{,}055$ (Tabla VI).

El investigador ha obtenido $\bar{X} = 25$ y $S_{n-1} = 6$ en su muestra y quiere saber cuál sería el nivel medio de defensas si se aplicara la risoterapia a todos los varones españoles.

Como ya sabíamos, el error de estimación máximo es 5:

$$E_{máx} = t_{12;0,995}\frac{S_{n-1}}{\sqrt{n}} = 3{,}055\frac{6}{\sqrt{13}} \approx 5$$

Nota: $E_{\text{máx}}$ no es exactamente 5 por el redondeo.

Los límites del intervalo de confianza son:

$$L_i = \bar{X} - t_{12;0,995}\frac{S_{n-1}}{\sqrt{n}} = 25 - 5 = 20$$

$$L_s = \bar{X} + t_{12;0,995}\frac{S_{n-1}}{\sqrt{n}} = 25 + 5 = 30$$

Para un nivel de confianza del 99%, si todos los varones españoles estuviesen sometidos a risoterapia, se estima que el nivel medio de defensas de su sistema inmunológico estaría entre 20 y 30 unidades.

Para cualquier distribución (normal o no normal), con desviación típica desconocida, podemos calcular el intervalo de confianza para la media, por aproximación de la distribución t de Student a la normal siempre que n sea grande ($n \geq 30$).

Ejemplo 8.10. Un investigador quiere estimar el tiempo que los estudiantes de primero de Psicología de la UNED dedican diariamente al estudio. Desconoce la forma, la desviación típica y la media de la distribución de la variable *tiempo diario de estudio* en la población. Decide calcular un intervalo de confianza para la media y establece un error de estimación máximo igual a 1 hora para un nivel de confianza del 99%. Calcula el tamaño de la muestra para ese error de estimación máximo y obtiene n = 30. A continuación, el investigador extrae una muestra aleatoria simple de 30 estudiantes de primero de Psicología de la UNED, obtiene el tiempo diario de estudio de los 30 estudiantes, $\bar{X}$ = 7, y la cuasidesviación típica, S_{n-1} = 2,2. ¿Cuál es el intervalo de confianza?

Solución:

Supuestos: muestreo aleatorio simple, variable cuantitativa, distribución de la variable en la población desconocida, σ desconocida.

n.c. = 0,99 $\rightarrow z_{1-\alpha/2} = z_{1-0{,}01/2} = z_{0{,}995}$ = 2,58 (Tabla IV) $\rightarrow$ Aproximación a la normal de la distribución t de Student.

A partir de la información de la muestra, $\bar{X}$ = 7 y S_{n-1} = 2,2, el investigador quiere saber el tiempo medio diario de estudio de todos los estudiantes de primero de Psicología de la UNED.

Como ya sabíamos, el error de estimación máximo es 1:

$$E_{máx} = z_{0{,}995}\frac{S_{n-1}}{\sqrt{n}} = 2{,}58\frac{2{,}2}{\sqrt{30}} = 1$$

Los límites del intervalo de confianza son:

$$L_i = \bar{X} - z_{1\text{-}\alpha/2}\frac{S_{n-1}}{\sqrt{n}} = 7 - 1 = 6$$

$$L_s = \bar{X} + z_{1\text{-}\alpha/2}\frac{S_{n-1}}{\sqrt{n}} = 7 + 1 = 8$$

Con una probabilidad de 0,99, se estima que el tiempo medio diario de estudio de todos los estudiantes de primero de Psicología está entre 6 y 8 horas.

A continuación, se recogen los intervalos de confianza para la media en función de los supuestos.

Tabla 8.4. Límites de los intervalos de confianza y supuestos para la estimación de la media.

Supuestos	**Límites del intervalo de confianza para la media**
• Muestreo aleatorio simple. • σ conocida. • Distribución normal o no normal con $n \geq 30$ (aprox. a la normal).	$L_i = \bar{X} - z_{1-\alpha/2}\sigma_{\bar{X}} \quad L_s = \bar{X} + z_{1-\alpha/2}\sigma_{\bar{X}}$ $z_{1\text{-}\alpha/2} \rightarrow$ Tabla IV $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$
• Muestreo aleatorio simple. • σ desconocida. • Distribución normal. • $n < 30$[5].	$L_i = \bar{X} - t_{n-1;1-\alpha/2}S_{\bar{X}} \quad L_s = \bar{X} + t_{n-1;1-\alpha/2}S_{\bar{X}}$ $t_{n-1;1-\alpha/2} \rightarrow$ Tabla VI $S_{\bar{X}} = \frac{S_{n-1}}{\sqrt{n}}$
• Muestreo aleatorio simple. • σ desconocida. • Distribución normal o no normal con $n \geq 30$ (aprox. a la normal).	$L_i = \bar{X} - z_{1-\alpha/2}S_{\bar{X}} \quad L_s = \bar{X} + z_{1-\alpha/2}S_{\bar{X}}$ $z_{1-\alpha/2} \rightarrow$ Tabla IV $S_{\bar{X}} = \frac{S_{n-1}}{\sqrt{n}}$
S_{n-1} es la cuasidesviación típica calculada en una muestra.	

8.6.3.2. Intervalo de confianza para la proporción

Dado el muestreo aleatorio simple, una variable dicotómica o dicotomizada, el error de estimación máximo de la proporción es:

$$E_{\text{máx}} = z_{1-\alpha/2}\sqrt{\frac{\pi(1-\pi)}{n}}$$

[5] Como a partir de $n = 30$ podemos realizar la aproximación de la distribución t de Student a la normal, a efectos prácticos, sólo utilizaremos la distribución t de Student para n < 30.

Donde:

- $z_{1-\alpha/2}$ es función del nivel de confianza $1-\alpha$ (Tabla IV).
- $\sqrt{\dfrac{\pi(1-\pi)}{n}}$ es el error típico de la proporción: σ_P.
- π es la proporción de la población que no es conocida.
- n es el tamaño de la muestra y se debe cumplir $n\pi(1-\pi) \geq 5$ para la aproximación a la normal.

Los límites inferior y superior del intervalo de confianza se obtienen a partir del error de estimación máximo. Como desconocemos π, que es lo que precisamente queremos estimar, operamos con la proporción muestral P. Así, si en $E_{\text{máx}}$ sustituimos π por la proporción muestral P, los límites inferior y superior del intervalo de confianza son:

$$L_i = P - z_{1-\alpha/2}\sqrt{\frac{P(1-P)}{n}} = P - E_{\text{máx}}$$

$$L_s = P + z_{1-\alpha/2}\sqrt{\frac{P(1-P)}{n}} = P + E_{\text{máx}}$$

Y la probabilidad de obtener un intervalo de confianza que contenga al parámetro π es:

$$P\left(P - z_{1-\alpha/2}\sqrt{\frac{P(1-P)}{n}} \leq \pi \leq P + z_{1-\alpha/2}\sqrt{\frac{P(1-P)}{n}}\right) = 1-\alpha$$

Ejemplo 8.11. Se ha propuesto un tratamiento para curar una determinada enfermedad y la comunidad científica quiere estimar la proporción de pacientes que se curarían si se aplicara el tratamiento a todos los pacientes. Deciden aplicar un intervalo de confianza para la proporción de pacientes curados, y fijan un error de estimación máximo de 0,20 para un n.c. = 0,95. Empiezan por calcular el tamaño de la muestra para ese error de estimación máximo y obtienen n = 24. A continuación, extraen una muestrea aleatoria simple de 24 pacientes, les aplican el tratamiento y obtienen 13 pacientes curados. ¿Cuál es el intervalo de confianza?

Supuestos: muestreo aleatoria simple, variable dicotómica, $n\pi(1-\pi) \geq 5$ para la aproximación a la normal.

Solución:

n.c. = 0,95 → $z_{1-\alpha/2} = z_{0{,}975} = 1{,}96$ (Tabla IV)

$P = 13/24 = 0{,}54$

Observen que se cumple la condición para la aproximación a la normal: $nP(1-P) \geq 5 \rightarrow 24\,(0{,}54)\,(0{,}46) = 5{,}96$.

Nota: Hemos sustituido π por P en la ecuación $n\pi(1-\pi) \geq 5$.

A partir de la información de la muestra, $P = 0{,}54$, la comunidad científica quiere saber cuál sería la proporción de pacientes que se curarían si se aplicara el tratamiento a la población entera de pacientes.

Como ya sabíamos, el error de estimación máximo es 0,20:

$$E_{\text{máx}} = 1{,}96\sqrt{\frac{(0{,}54)(0{,}46)}{24}} = 0{,}20$$

Los límites del intervalo de confianza son:

$$L_i = 0{,}54 - 0{,}20 = 0{,}34$$
$$L_s = 0{,}54 + 0{,}20 = 0{,}74$$

Para un nivel de confianza del 95%, la proporción de pacientes que se estima se curarían con el tratamiento propuesto si se aplicara a toda la población de pacientes está entre 0,34 y 0,74.

8.7. RESUMEN

En este tema, hemos empezado tratando el papel de la estadística descriptiva y de la estadística inferencial en relación con las poblaciones y muestras. A continuación, hemos estudiado diversos métodos de muestreo y hemos introducido los conceptos de muestra aleatoria y muestra representativa. Luego, hemos estudiado la distribución muestral, un concepto central de la estadística inferencial. Posteriormente, hemos visto el concepto de intervalo de confianza, hemos aprendido a calcular el tamaño de la muestra y hemos estudiado las relaciones entre el tamaño de la muestra y el nivel de confianza, el error de estimación máximo y la variabilidad. Finalmente, hemos aprendido a hacer inferencias mediante intervalos de confianza para distintas aplicaciones de la media y la proporción.

8.8. EJERCICIOS DE AUTOEVALUACIÓN

8.1. ¿Cuál de los siguientes tipos de muestreo NO es probabilístico?: A) el muestreo sistemático; B) el muestreo casual o incidental; C) el muestreo por conglomerados

8.2. Aunque desconocemos el número de ancianos de las residencias privadas españolas, deseamos realizar un estudio sobre la atención que reciben dichos ancianos. ¿Cuál de los siguientes tipos de muestreo podríamos utilizar?: A) muestreo aleatorio simple; B) muestreo sistemático; C) muestreo por conglomerados

8.3. Un psicólogo mide una variable cuantitativa al conjunto total de la población. El psicólogo: A) puede conocer el valor del parámetro μ de la población sin recurrir a técnicas inferenciales; B) sólo puede conocer una estimación del parámetro μ de la población; C) tendrá que estimar el parámetro μ con técnicas inferenciales.

8.4. Elija la afirmación correcta: A) una muestra aleatoria es siempre representativa de la población; B) un parámetro de una población es una variable aleatoria; C) una muestra aleatoria simple garantiza que todos los elementos de la población tienen la misma probabilidad de ser elegidos.

8.5. Un estimador: A) es un estadístico; B) es un parámetro; C) NO es una variable aleatoria.

8.6. Elija la afirmación correcta: A) una estimación es igual al parámetro; B) mediante el análisis descriptivo de una muestra podemos inferir a la población; C) la estimación por intervalo es un procedimiento inferencial.

8.7. La media y la cuasidesviación típica de una variable cuantitativa X en la muestra son: A) μ y σ; B) $\mu_{\bar{X}}$ y $\frac{\sigma}{\sqrt{n}}$; C) $\bar{X}$ y S_{n-1}.

8.8. Supongamos una población grande con distribución uniforme, extraemos muchas muestras de tamaño $n = 30$ y calculamos la media en cada muestra. La distribución muestral de la media es: A) uniforme; B) claramente no normal; C) aproximadamente normal

8.9. La media muestral, $\bar{X}$, es un estimador insesgado de la media poblacional (μ) porque: A) $\bar{X} = \mu$; B) $\mu_{\bar{X}} = \mu$; C) $\bar{X} = \mu_{\bar{X}}$.

8.10. ¿En qué circunstancias es útil calcular un intervalo de confianza para un determinado parámetro?: A) cuando se conoce el parámetro de interés; B) cuando se quiere estimar un parámetro desconocido; C) cuando se estudia a toda la población.

8.11. En un estudio donde la media de la muestra es 35, los límites inferior y superior del intervalo de confianza son respectivamente 30 y 40. ¿Cuál es el error de estimación máximo de la media?: A) 2,5; B) 5; C) 10.

8.12. En una muestra, hemos obtenido los límites 3,5 y 6,5 de un intervalo de confianza para la media. La amplitud del intervalo de confianza: A) es 2 $E_{máx}$; B) NO es 3; C) es $3 \cdot 2 = 6$.

8.13. En el ejercicio anterior, ¿cuánto vale el error máximo de estimación?: A) $3/2 = 1,5$; B) 3; C) $3 \cdot 2 = 6$.

8.14. Supongamos que en una investigación, la variable se distribuye normalmente en la población con $\sigma = 4$ y queremos que $E_{máx}$ no sea mayor que 2 con un nivel de confianza de 0,99 ¿Qué tamaño debe tener la muestra para estimar la media? A) 15; B) 20; C) 27.

8.15. Un investigador quiere inferir la autoestima media de los reclusos penitenciarios. En este estudio, $\sigma = 5$, n.c. = 0,95, $n = 43$ y $\bar{X} = 10$. ¿Cuáles son los límites del intervalo de confianza entre los cuales se espera esté la autoestima media de todos los reclusos penitenciarios? A) 8,5 y 10; B) 8,5 y 11,5; C) 10 y 11,5.

8.16. Los límites del intervalo de confianza para la media en el caso de distribución normal y varianza conocida son 5,85 y 10,15, y $E_{máx}$ es 2,15. ¿Cuánto vale la media de la muestra? A) 4; B) 5; C) 8.

8.17. En el ejercicio anterior, cuánto vale el nivel de confianza si $\sigma = 5$ y $n = 36$: A) 0,95; B) 0,99; C) 0,999.

8.18. Una persona dice que es capaz de predecir el resultado en el lanzamiento de una moneda y queremos comprobarlo. De 50 lanzamientos, falla 30 veces. ¿Cuál es el error de estimación máximo, fijado en este experimento, para la proporción de aciertos, dado que el nivel de confianza es 0,95?: A) 0,14; B) 0,95; C) 1,96.

8.19. En el ejercicio anterior, los límites del intervalo de confianza para la proporción de aciertos son: A) 0,20 y 0,40; B) 0,21 y 0,79; C) 0,26 y 0,54.

8.20. En el ejercicio 8.18., los límites del intervalo de confianza para la proporción de fallos son: A) 0,15 y 0,85; B) 0,20 y 0,60; C) 0,46 y 0,74.

8.9. SOLUCIONES A LOS EJERCICIOS DE AUTOEVALUACIÓN

8.1. Solución: B
El muestreo casual porque se seleccionan los elementos de la población a los que se tiene fácil acceso.

8.2. Solución C
Podríamos utilizar el muestreo por conglomerados dado que no requiere un listado de los elementos de la población, basta un listado de los conglomerados. El muestreo aleatorio simple y el muestreo sistemático sí lo requieren.

8.3. Solución: A
Dado que el psicólogo trabaja con toda la población, no debe usar técnicas inferenciales. Basta calcular la media con todos los datos y ésta será el parámetro μ.

8.4. Solución: C
No todas las muestras aleatorias son representativas de la población. Un parámetro tiene un único valor, no es una variable aleatoria.

8.5. Solución: A
Un estimador es un estadístico, no es un parámetro y es una variable aleatoria.

8.6. Solución: C
Una estimación no es igual al parámetro sino que es un valor que toma el estimador en una muestra concreta. Mediante el análisis descriptivo de una muestra no podemos inferir a la población.

8.7. Solución: C
μ y σ son la media y la desviación típica de la variable X en la población.

$\mu_{\bar{X}}$ y $\dfrac{\sigma}{\sqrt{n}}$ son la media y la desviación típica de la distribución muestral de la media.

8.8. Solución: C
Es aproximadamente normal (ver Teorema Central del Límite en el apartado 8.4.1.)

8.9. Solución: B
La media muestral, $\bar{X}$, es un estimador insesgado de la media poblacional (μ) porque $\mu_{\bar{X}} = \mu$. $\bar{X}$ no es igual a μ ni a $\mu_{\bar{X}}$.

8.10. Solución: B
Si se conoce el parámetro, no hay nada que inferir por lo que no necesitamos calcular ningún intervalo de confianza. Un intervalo de confianza es útil cuando se quiere estimar un parámetro desconocido. Cuando se estudia a toda la población no hay que inferir nada, se calcula el parámetro sin recurrir a técnicas inferenciales.

8.11. Solución: B

$$L_i = \bar{X} - E_{\text{máx}} \rightarrow E_{\text{máx}} = \bar{X} - L_i \quad \rightarrow \quad E_{\text{máx}} = 35 - 30 = 5 \text{ O bien,}$$
$$L_s = \bar{X} + E_{\text{máx}} \rightarrow E_{\text{máx}} = L_s - \bar{X} \rightarrow \quad E_{\text{máx}} = 40 - 35 = 5$$

8.12. Solución: A
La amplitud del intervalo de confianza es dos veces el error de estimación máximo: 6,5-3,5 = 3.

8.13. Solución: A
El error de estimación máximo es la mitad de la amplitud del intervalo de confianza: 3/2 = 1,5.

8.14. Solución: C

n.c. = 0,99 → $z_{1-\alpha/2} = z_{0,995} = 2,58$ (Tabla IV)

$$n = \frac{2,58^2 \cdot 4^2}{2^2} = 26,62 \rightarrow 27$$

8.15. Solución: B

Se desconoce la forma de la distribución de la autoestima en la población de reclusos pero $n = 43$, por lo que podemos hacer la aproximación a la normal.

n.c. = 0,95 → $z_{1-\alpha/2} = z_{0,975} = 1,96$ (Tabla IV)

$$E_{\text{máx}} = z_{1-\alpha/2}\sigma_{\bar{X}} = 1,96\frac{5}{\sqrt{43}} \approx 1,5$$

$L_i = 10\text{-}1,50 = 8,50$ $\quad L_s = 10 + 1,50 = 11,50$

8.16. Solución: C

$$L_i = \bar{X} - E_{\text{máx}} \rightarrow \bar{X} = L_i + E_{\text{máx}} \quad \rightarrow \quad \bar{X} = 5,85 + 2,15 = 8$$

O bien,

$$L_s = \bar{X} + E_{\text{máx}} \rightarrow \bar{X} = L_s - E_{\text{máx}} \quad \rightarrow \quad \bar{X} = 10,15 - 2,15 = 8$$

8.17. Solución: B

$$E_{\text{máx}} = z_{1\text{-}a/2}\sigma_{\bar{X}} \rightarrow \sigma_{\bar{X}} = 5/\sqrt{36} = 0,833 \rightarrow 2,15 = z_{1\text{-}\alpha/2}0,833 \rightarrow z_{1\text{-}\alpha/2} = \\ = 2,58 \rightarrow \text{n.c.} = 0,99 \text{ (Tabla IV)}$$

8.18. Solución: A

n.c. = 0,95 → $z_{1-\alpha/2} = z_{0,975} = 1,96$ (Tabla IV)

Probabilidad de aciertos: $p_{\text{aciertos}} = 20/50 = 0,40$

$$E_{\text{máx}} = 1,96\sqrt{\frac{0,40(1\text{-}0,40)}{50}} = 0,14$$

8.19. Solución: C

$L_i = 0,40 - 0,14 = 0,26$ $\quad L_s = 0,40 + 0,14 = 0,54$

8.20. Solución: C

Como $p_{\text{fallos}} = 1 - p_{\text{aciertos}}$, si p_{aciertos} está entre 0,26 y 0,54, p_{fallos} está entre 1 – 0,54 y 1 – 0,26, o sea, entre 0,46 y 0,74.

REFERENCIAS BIBLIOGRÁFICAS

AMÓN, J. (1999): *Estadística para psicólogos. Estadística descriptiva*. Vol. 1. Madrid. Ed. Pirámide (15.ª edición).

BOTELLA, J.; LEÓN, O. G. y SAN MARTÍN, R. (1993): *Análisis de Datos en Psicología I*. Madrid: Editorial Pirámide.

GARRIGA-TRILLO, A.; AGUILERA-GENICIO, F. (2005): Olfactory sensitivity measures can predict cognitive impairment: A parametric and non-parametric approach. In J. S. MONAHAN, S. M. SHEFFERT & J. T. TOWNSEND (eds.) *Fechner Day 2005* (pp. 101-106). Traverse City, MI: Central Michigan University Printing Services.

GARRIGA-TRILLO, A.; AGUILERA-GENICIO, F. (2007): Predicting cognitive impairment from olfactory sensitivity measures: A continuous and discontinuous approach. In A. GARRIGA-TRILLO (Ed.) *Converging research on predictors of cognitive impairment and neurodegenerative diseases* (pp. 15-28). Sevilla: Publidisa.

HAYS, W. (1988): *Statistics* (4th ed.) New York: Holt Rinehart & Winston.

MERINO, J. M. y otros (2007): *Análisis de datos en Psicología I*. Madrid: UNED. 5.ª reimpresión.

PARDO, A. y SAN MARTÍN, R. (1998): *Análisis de datos en Psicología II*. Madrid: Editorial Pirámide.

STEVENS, S. S. (1946): On the theory of scales of measurement. *Science,* 103, 677-680.

YELA, M. (1994): El problema del método científico en Psicología. *Anuario de Psicología,* 60, 3-12.

Apéndice

Tablas

TABLA I: FUNCIÓN DE PROBABILIDAD BINOMIAL

$$f(x) = P(X = x) = \binom{n}{x} p^x \, q^{n-x}$$

n	*x*	Probabilidad de éxito (*p*)										
		0,01	**0,05**	**0,10**	**0,15**	**0,2**	**0,25**	**0,30**	**0,35**	**0,40**	**0,45**	**0,50**
1	**0**	0,9900	0,9500	0,9000	0,8500	0,8000	0,7500	0,7000	0,6500	0,6000	0,5500	0,5000
1	**1**	0,0100	0,0500	0,1000	0,1500	0,2000	0,2500	0,3000	0,3500	0,4000	0,4500	0,5000
2	**0**	0,9801	0,9025	0,8100	0,7225	0,6400	0,5625	0,4900	0,4225	0,3600	0,3025	0,2500
2	**1**	0,0198	0,0950	0,1800	0,2550	0,3200	0,3750	0,4200	0,4550	0,4800	0,4950	0,5000
2	**2**	0,0001	0,0025	0,0100	0,0225	0,0400	0,0625	0,0900	0,1225	0,1600	0,2025	0,2500
3	**0**	0,9703	0,8574	0,7290	0,6141	0,5120	0,4219	0,3430	0,2746	0,2160	0,1664	0,1250
3	**1**	0,0294	0,1354	0,2430	0,3251	0,3840	0,4219	0,4410	0,4436	0,4320	0,4084	0,3750
3	**2**	0,0003	0,0071	0,0270	0,0574	0,0960	0,1406	0,1890	0,2389	0,2880	0,3341	0,3750
3	**3**	0,0000	0,0001	0,0010	0,0034	0,0080	0,0156	0,0270	0,0429	0,0640	0,0911	0,1250
4	**0**	0,9606	0,8145	0,6561	0,5220	0,4096	0,3164	0,2401	0,1785	0,1296	0,0915	0,0625
4	**1**	0,0388	0,1715	0,2916	0,3685	0,4096	0,4219	0,4116	0,3845	0,3456	0,2995	0,2500
4	**2**	0,0006	0,0135	0,0486	0,0975	0,1536	0,2109	0,2646	0,3105	0,3456	0,3675	0,3750
4	**3**	0,0000	0,0005	0,0036	0,0115	0,0256	0,0469	0,0756	0,1115	0,1536	0,2005	0,2500
4	**4**	0,0000	0,0000	0,0001	0,0005	0,0016	0,0039	0,0081	0,0150	0,0256	0,0410	0,0625
5	**0**	0,9510	0,7738	0,5905	0,4437	0,3277	0,2373	0,1681	0,1160	0,0778	0,0503	0,0313
5	**1**	0,0480	0,2036	0,3281	0,3915	0,4096	0,3955	0,3602	0,3124	0,2592	0,2059	0,1563
5	**2**	0,0010	0,0214	0,0729	0,1382	0,2048	0,2637	0,3087	0,3364	0,3456	0,3369	0,3125
5	**3**	0,0000	0,0011	0,0081	0,0244	0,0512	0,0879	0,1323	0,1811	0,2304	0,2757	0,3125
5	**4**	0,0000	0,0000	0,0005	0,0022	0,0064	0,0146	0,0284	0,0488	0,0768	0,1128	0,1563
5	**5**	0,0000	0,0000	0,0000	0,0001	0,0003	0,0010	0,0024	0,0053	0,0102	0,0185	0,0313
6	**0**	0,9415	0,7351	0,5314	0,3771	0,2621	0,1780	0,1176	0,0754	0,0467	0,0277	0,0156
6	**1**	0,0571	0,2321	0,3543	0,3993	0,3932	0,3560	0,3025	0,2437	0,1866	0,1359	0,0938
6	**2**	0,0014	0,0305	0,0984	0,1762	0,2458	0,2966	0,3241	0,3280	0,3110	0,2780	0,2344
6	**3**	0,0000	0,0021	0,0146	0,0415	0,0819	0,1318	0,1852	0,2355	0,2765	0,3032	0,3125
6	**4**	0,0000	0,0001	0,0012	0,0055	0,0154	0,0330	0,0595	0,0951	0,1382	0,1861	0,2344
6	**5**	0,0000	0,0000	0,0001	0,0004	0,0015	0,0044	0,0102	0,0205	0,0369	0,0609	0,0938
6	**6**	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0007	0,0018	0,0041	0,0083	0,0156
7	**0**	0,9321	0,6983	0,4783	0,3206	0,2097	0,1335	0,0824	0,0490	0,0280	0,0152	0,0078
7	**1**	0,0659	0,2573	0,3720	0,3960	0,3670	0,3115	0,2471	0,1848	0,1306	0,0872	0,0547
7	**2**	0,0020	0,0406	0,1240	0,2097	0,2753	0,3115	0,3177	0,2985	0,2613	0,2140	0,1641
7	**3**	0,0000	0,0036	0,0230	0,0617	0,1147	0,1730	0,2269	0,2679	0,2903	0,2918	0,2734
7	**4**	0,0000	0,0002	0,0026	0,0109	0,0287	0,0577	0,0972	0,1442	0,1935	0,2388	0,2734
7	**5**	0,0000	0,0000	0,0002	0,0012	0,0043	0,0115	0,0250	0,0466	0,0774	0,1172	0,1641
7	**6**	0,0000	0,0000	0,0000	0,0001	0,0004	0,0013	0,0036	0,0084	0,0172	0,0320	0,0547
7	**7**	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0006	0,0016	0,0037	0,0078

Los valores interiores de la tabla indican la probabilidad de obtener «x» éxitos en «n» ensayos de un experimento binomial, donde «p» es la probabilidad de éxito en un ensayo.

TABLA I (Cont.)

n	*x*	Probabilidad de éxito (*p*)										
		0,01	**0,05**	**0,10**	**0,15**	**0,20**	**0,25**	**0,30**	**0,35**	**0,40**	**0,45**	**0,50**
8	**0**	0,9227	0,6634	0,4305	0,2725	0,1678	0,1001	0,0576	0,0319	0,0168	0,0084	0,0039
8	**1**	0,0746	0,2793	0,3826	0,3847	0,3355	0,2670	0,1977	0,1373	0,0896	0,0548	0,0313
8	**2**	0,0026	0,0515	0,1488	0,2376	0,2936	0,3115	0,2965	0,2587	0,2090	0,1569	0,1094
8	**3**	0,0001	0,0054	0,0331	0,0839	0,1468	0,2076	0,2541	0,2786	0,2787	0,2568	0,2188
8	**4**	0,0000	0,0004	0,0046	0,0185	0,0459	0,0865	0,1361	0,1875	0,2322	0,2627	0,2734
8	**5**	0,0000	0,0000	0,0004	0,0026	0,0092	0,0231	0,0467	0,0808	0,1239	0,1719	0,2188
8	**6**	0,0000	0,0000	0,0000	0,0002	0,0011	0,0038	0,0100	0,0217	0,0413	0,0703	0,1094
8	**7**	0,0000	0,0000	0,0000	0,0000	0,0001	0,0004	0,0012	0,0033	0,0079	0,0164	0,0313
8	**8**	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0007	0,0017	0,0039
9	**0**	0,9135	0,6302	0,3874	0,2316	0,1342	0,0751	0,0404	0,0207	0,0101	0,0046	0,0020
9	**1**	0,0830	0,2985	0,3874	0,3679	0,3020	0,2253	0,1556	0,1004	0,0605	0,0339	0,0176
9	**2**	0,0034	0,0629	0,1722	0,2597	0,3020	0,3003	0,2668	0,2162	0,1612	0,1110	0,0703
9	**3**	0,0001	0,0077	0,0446	0,1069	0,1762	0,2336	0,2668	0,2716	0,2508	0,2119	0,1641
9	**4**	0,0000	0,0006	0,0074	0,0283	0,0661	0,1168	0,1715	0,2194	0,2508	0,2600	0,2461
9	**5**	0,0000	0,0000	0,0008	0,0050	0,0165	0,0389	0,0735	0,1181	0,1672	0,2128	0,2461
9	**6**	0,0000	0,0000	0,0001	0,0006	0,0028	0,0087	0,0210	0,0424	0,0743	0,1160	0,1641
9	**7**	0,0000	0,0000	0,0000	0,0000	0,0003	0,0012	0,0039	0,0098	0,0212	0,0407	0,0703
9	**8**	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0004	0,0013	0,0035	0,0083	0,0176
9	**9**	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0008	0,0020
10	**0**	0,9044	0,5987	0,3487	0,1969	0,1074	0,0563	0,0282	0,0135	0,0060	0,0025	0,0010
10	**1**	0,0914	0,3151	0,3874	0,3474	0,2684	0,1877	0,1211	0,0725	0,0403	0,0207	0,0098
10	**2**	0,0042	0,0746	0,1937	0,2759	0,3020	0,2816	0,2335	0,1757	0,1209	0,0763	0,0439
10	**3**	0,0001	0,0105	0,0574	0,1298	0,2013	0,2503	0,2668	0,2522	0,2150	0,1665	0,1172
10	**4**	0,0000	0,0010	0,0112	0,0401	0,0881	0,1460	0,2001	0,2377	0,2508	0,2384	0,2051
10	**5**	0,0000	0,0001	0,0015	0,0085	0,0264	0,0584	0,1029	0,1536	0,2007	0,2340	0,2461
10	**6**	0,0000	0,0000	0,0001	0,0012	0,0055	0,0162	0,0368	0,0689	0,1115	0,1596	0,2051
10	**7**	0,0000	0,0000	0,0000	0,0001	0,0008	0,0031	0,0090	0,0212	0,0425	0,0746	0,1172
10	**8**	0,0000	0,0000	0,0000	0,0000	0,0001	0,0004	0,0014	0,0043	0,0106	0,0229	0,0439
10	**9**	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0016	0,0042	0,0098
10	**10**	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0010
11	**0**	0,8953	0,5688	0,3138	0,1673	0,0859	0,0422	0,0198	0,0088	0,0036	0,0014	0,0005
11	**1**	0,0995	0,3293	0,3835	0,3248	0,2362	0,1549	0,0932	0,0518	0,0266	0,0125	0,0054
11	**2**	0,0050	0,0867	0,2131	0,2866	0,2953	0,2581	0,1998	0,1395	0,0887	0,0513	0,0269
11	**3**	0,0002	0,0137	0,0710	0,1517	0,2215	0,2581	0,2568	0,2254	0,1774	0,1259	0,0806
11	**4**	0,0000	0,0014	0,0158	0,0536	0,1107	0,1721	0,2201	0,2428	0,2365	0,2060	0,1611
11	**5**	0,0000	0,0001	0,0025	0,0132	0,0388	0,0803	0,1321	0,1830	0,2207	0,2360	0,2256
11	**6**	0,0000	0,0000	0,0003	0,0023	0,0097	0,0268	0,0566	0,0985	0,1471	0,1931	0,2256
11	**7**	0,0000	0,0000	0,0000	0,0003	0,0017	0,0064	0,0173	0,0379	0,0701	0,1128	0,1611
11	**8**	0,0000	0,0000	0,0000	0,0000	0,0002	0,0011	0,0037	0,0102	0,0234	0,0462	0,0806
11	**9**	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0018	0,0052	0,0126	0,0269
11	**10**	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0007	0,0021	0,0054
11	**11**	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0005

TABLA I (Cont.)

n	x	Probabilidad de éxito (p)										
		0,01	**0,05**	**0,10**	**0,15**	**0,20**	**0,25**	**0,30**	**0,35**	**0,40**	**0,45**	**0,50**
12	0	0,8864	0,5404	0,2824	0,1422	0,0687	0,0317	0,0138	0,0057	0,0022	0,0008	0,0002
12	1	0,1074	0,3413	0,3766	0,3012	0,2062	0,1267	0,0712	0,0368	0,0174	0,0075	0,0029
12	2	0,0060	0,0988	0,2301	0,2924	0,2835	0,2323	0,1678	0,1088	0,0639	0,0339	0,0161
12	3	0,0002	0,0173	0,0852	0,1720	0,2362	0,2581	0,2397	0,1954	0,1419	0,0923	0,0537
12	4	0,0000	0,0021	0,0213	0,0683	0,1329	0,1936	0,2311	0,2367	0,2128	0,1700	0,1208
12	5	0,0000	0,0002	0,0038	0,0193	0,0532	0,1032	0,1585	0,2039	0,2270	0,2225	0,1934
12	6	0,0000	0,0000	0,0005	0,0040	0,0155	0,0401	0,0792	0,1281	0,1766	0,2124	0,2256
12	7	0,0000	0,0000	0,0000	0,0006	0,0033	0,0115	0,0291	0,0591	0,1009	0,1489	0,1934
12	8	0,0000	0,0000	0,0000	0,0001	0,0005	0,0024	0,0078	0,0199	0,0420	0,0762	0,1208
12	9	0,0000	0,0000	0,0000	0,0000	0,0001	0,0004	0,0015	0,0048	0,0125	0,0277	0,0537
12	10	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0008	0,0025	0,0068	0,0161
12	11	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0010	0,0029
12	12	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002
13	0	0,8775	0,5133	0,2542	0,1209	0,0550	0,0238	0,0097	0,0037	0,0013	0,0004	0,0001
13	1	0,1152	0,3512	0,3672	0,2774	0,1787	0,1029	0,0540	0,0259	0,0113	0,0045	0,0016
13	2	0,0070	0,1109	0,2448	0,2937	0,2680	0,2059	0,1388	0,0836	0,0453	0,0220	0,0095
13	3	0,0003	0,0214	0,0997	0,1900	0,2457	0,2517	0,2181	0,1651	0,1107	0,0660	0,0349
13	4	0,0000	0,0028	0,0277	0,0838	0,1535	0,2097	0,2337	0,2222	0,1845	0,1350	0,0873
13	5	0,0000	0,0003	0,0055	0,0266	0,0691	0,1258	0,1803	0,2154	0,2214	0,1989	0,1571
13	6	0,0000	0,0000	0,0008	0,0063	0,0230	0,0559	0,1030	0,1546	0,1968	0,2169	0,2095
13	7	0,0000	0,0000	0,0001	0,0011	0,0058	0,0186	0,0442	0,0833	0,1312	0,1775	0,2095
13	8	0,0000	0,0000	0,0000	0,0001	0,0011	0,0047	0,0142	0,0336	0,0656	0,1089	0,1571
13	9	0,0000	0,0000	0,0000	0,0000	0,0001	0,0009	0,0034	0,0101	0,0243	0,0495	0,0873
13	10	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0006	0,0022	0,0065	0,0162	0,0349
13	11	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0012	0,0036	0,0095
13	12	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0016
13	13	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001
14	0	0,8687	0,4877	0,2288	0,1028	0,0440	0,0178	0,0068	0,0024	0,0008	0,0002	0,0001
14	1	0,1229	0,3593	0,3559	0,2539	0,1539	0,0832	0,0407	0,0181	0,0073	0,0027	0,0009
14	2	0,0081	0,1229	0,2570	0,2912	0,2501	0,1802	0,1134	0,0634	0,0317	0,0141	0,0056
14	3	0,0003	0,0259	0,1142	0,2056	0,2501	0,2402	0,1943	0,1366	0,0845	0,0462	0,0222
14	4	0,0000	0,0037	0,0349	0,0998	0,1720	0,2202	0,2290	0,2022	0,1549	0,1040	0,0611
14	5	0,0000	0,0004	0,0078	0,0352	0,0860	0,1468	0,1963	0,2178	0,2066	0,1701	0,1222
14	6	0,0000	0,0000	0,0013	0,0093	0,0322	0,0734	0,1262	0,1759	0,2066	0,2088	0,1833
14	7	0,0000	0,0000	0,0002	0,0019	0,0092	0,0280	0,0618	0,1082	0,1574	0,1952	0,2095
14	8	0,0000	0,0000	0,0000	0,0003	0,0020	0,0082	0,0232	0,0510	0,0918	0,1398	0,1833
14	9	0,0000	0,0000	0,0000	0,0000	0,0003	0,0018	0,0066	0,0183	0,0408	0,0762	0,1222
14	10	0,0000	0,0000	0,0000	0,0000	0,0000	0,0003	0,0014	0,0049	0,0136	0,0312	0,0611
14	11	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0010	0,0033	0,0093	0,0222
14	12	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0019	0,0056
14	13	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0009
14	14	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001
15	0	0,8601	0,4633	0,2059	0,0874	0,0352	0,0134	0,0047	0,0016	0,0005	0,0001	0,0000
15	1	0,1303	0,3658	0,3432	0,2312	0,1319	0,0668	0,0305	0,0126	0,0047	0,0016	0,0005
15	2	0,0092	0,1348	0,2669	0,2856	0,2309	0,1559	0,0916	0,0476	0,0219	0,0090	0,0032
15	3	0,0004	0,0307	0,1285	0,2184	0,2501	0,2252	0,1700	0,1110	0,0634	0,0318	0,0139
15	4	0,0000	0,0049	0,0428	0,1156	0,1876	0,2252	0,2186	0,1792	0,1268	0,0780	0,0417
15	5	0,0000	0,0006	0,0105	0,0449	0,1032	0,1651	0,2061	0,2123	0,1859	0,1404	0,0916

TABLA I (Cont.)

n	x	Probabilidad de éxito (p)										
		0,01	**0,05**	**0,10**	**0,15**	**0,20**	**0,25**	**0,30**	**0,35**	**0,40**	**0,45**	**0,50**
15	**6**	0,0000	0,0000	0,0019	0,0132	0,0430	0,0917	0,1472	0,1906	0,2066	0,1914	0,1527
15	**7**	0,0000	0,0000	0,0003	0,0030	0,0138	0,0393	0,0811	0,1319	0,1771	0,2013	0,1964
15	**8**	0,0000	0,0000	0,0000	0,0005	0,0035	0,0131	0,0348	0,0710	0,1181	0,1647	0,1964
15	**9**	0,0000	0,0000	0,0000	0,0001	0,0007	0,0034	0,0116	0,0298	0,0612	0,1048	0,1527
15	**10**	0,0000	0,0000	0,0000	0,0000	0,0001	0,0007	0,0030	0,0096	0,0245	0,0515	0,0916
15	**11**	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0006	0,0024	0,0074	0,0191	0,0417
15	**12**	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0004	0,0016	0,0052	0,0139
15	**13**	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0010	0,0032
15	**14**	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005
15	**15**	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
16	**0**	0,8515	0,4401	0,1853	0,0743	0,0281	0,0100	0,0033	0,0010	0,0003	0,0001	0,0000
16	**1**	0,1376	0,3706	0,3294	0,2097	0,1126	0,0535	0,0228	0,0087	0,0030	0,0009	0,0002
16	**2**	0,0104	0,1463	0,2745	0,2775	0,2111	0,1336	0,0732	0,0353	0,0150	0,0056	0,0018
16	**3**	0,0005	0,0359	0,1423	0,2285	0,2463	0,2079	0,1465	0,0888	0,0468	0,0215	0,0085
16	**4**	0,0000	0,0061	0,0514	0,1311	0,2001	0,2252	0,2040	0,1553	0,1014	0,0572	0,0278
16	**5**	0,0000	0,0008	0,0137	0,0555	0,1201	0,1802	0,2099	0,2008	0,1623	0,1123	0,0667
16	**6**	0,0000	0,0001	0,0028	0,0180	0,0550	0,1101	0,1649	0,1982	0,1983	0,1684	0,1222
16	**7**	0,0000	0,0000	0,0004	0,0045	0,0197	0,0524	0,1010	0,1524	0,1889	0,1969	0,1746
16	**8**	0,0000	0,0000	0,0001	0,0009	0,0055	0,0197	0,0487	0,0923	0,1417	0,1812	0,1964
16	**9**	0,0000	0,0000	0,0000	0,0001	0,0012	0,0058	0,0185	0,0442	0,0840	0,1318	0,1746
16	**10**	0,0000	0,0000	0,0000	0,0000	0,0002	0,0014	0,0056	0,0167	0,0392	0,0755	0,1222
16	**11**	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0013	0,0049	0,0142	0,0337	0,0667
16	**12**	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0011	0,0040	0,0115	0,0278
16	**13**	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0008	0,0029	0,0085
16	**14**	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0018
16	**15**	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002
16	**16**	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
17	**0**	0,8429	0,4181	0,1668	0,0631	0,0225	0,0075	0,0023	0,0007	0,0002	0,0000	0,0000
17	**1**	0,1447	0,3741	0,3150	0,1893	0,0957	0,0426	0,0169	0,0060	0,0019	0,0005	0,0001
17	**2**	0,0117	0,1575	0,2800	0,2673	0,1914	0,1136	0,0581	0,0260	0,0102	0,0035	0,0010
17	**3**	0,0006	0,0415	0,1556	0,2359	0,2393	0,1893	0,1245	0,0701	0,0341	0,0144	0,0052
17	**4**	0,0000	0,0076	0,0605	0,1457	0,2093	0,2209	0,1868	0,1320	0,0796	0,0411	0,0182
17	**5**	0,0000	0,0010	0,0175	0,0668	0,1361	0,1914	0,2081	0,1849	0,1379	0,0875	0,0472
17	**6**	0,0000	0,0001	0,0039	0,0236	0,0680	0,1276	0,1784	0,1991	0,1839	0,1432	0,0944
17	**7**	0,0000	0,0000	0,0007	0,0065	0,0267	0,0668	0,1201	0,1685	0,1927	0,1841	0,1484
17	**8**	0,0000	0,0000	0,0001	0,0014	0,0084	0,0279	0,0644	0,1134	0,1606	0,1883	0,1855
17	**9**	0,0000	0,0000	0,0000	0,0003	0,0021	0,0093	0,0276	0,0611	0,1070	0,1540	0,1855
17	**10**	0,0000	0,0000	0,0000	0,0000	0,0004	0,0025	0,0095	0,0263	0,0571	0,1008	0,1484
17	**11**	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0026	0,0090	0,0242	0,0525	0,0944
17	**12**	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0006	0,0024	0,0081	0,0215	0,0472
17	**13**	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0021	0,0068	0,0182
17	**14**	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0004	0,0016	0,0052
17	**15**	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003	0,0010
17	**16**	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001
17	**17**	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
18	**0**	0,8345	0,3972	0,1501	0,0536	0,0180	0,0056	0,0016	0,0004	0,0001	0,0000	0,0000
18	**1**	0,1517	0,3763	0,3002	0,1704	0,0811	0,0338	0,0126	0,0042	0,0012	0,0003	0,0001
18	**2**	0,0130	0,1683	0,2835	0,2556	0,1723	0,0958	0,0458	0,0190	0,0069	0,0022	0,0006
18	**3**	0,0007	0,0473	0,1680	0,2406	0,2297	0,1704	0,1046	0,0547	0,0246	0,0095	0,0031
18	**4**	0,0000	0,0093	0,0700	0,1592	0,2153	0,2130	0,1681	0,1104	0,0614	0,0291	0,0117
18	**5**	0,0000	0,0014	0,0218	0,0787	0,1507	0,1988	0,2017	0,1664	0,1146	0,0666	0,0327
18	**6**	0,0000	0,0002	0,0052	0,0301	0,0816	0,1436	0,1873	0,1941	0,1655	0,1181	0,0708

TABLA I (Cont.)

n	x	Probabilidad de éxito (p)										
		0,01	**0,05**	**0,10**	**0,15**	**0,20**	**0,25**	**0,30**	**0,35**	**0,40**	**0,45**	**0,50**
18	**7**	0,0000	0,0000	0,0010	0,0091	0,0350	0,0820	0,1376	0,1792	0,1892	0,1657	0,1214
18	**8**	0,0000	0,0000	0,0002	0,0022	0,0120	0,0376	0,0811	0,1327	0,1734	0,1864	0,1669
18	**9**	0,0000	0,0000	0,0000	0,0004	0,0033	0,0139	0,0386	0,0794	0,1284	0,1694	0,1855
18	**10**	0,0000	0,0000	0,0000	0,0001	0,0008	0,0042	0,0149	0,0385	0,0771	0,1248	0,1669
18	**11**	0,0000	0,0000	0,0000	0,0000	0,0001	0,0010	0,0046	0,0151	0,0374	0,0742	0,1214
18	**12**	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0012	0,0047	0,0145	0,0354	0,0708
18	**13**	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0012	0,0045	0,0134	0,0327
18	**14**	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0011	0,0039	0,0117
18	**15**	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0009	0,0031
18	**16**	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0006
18	**17**	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001
18	**18**	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
19	**0**	0,8262	0,3774	0,1351	0,0456	0,0144	0,0042	0,0011	0,0003	0,0001	0,0000	0,0000
19	**1**	0,1586	0,3774	0,2852	0,1529	0,0685	0,0268	0,0093	0,0029	0,0008	0,0002	0,0000
19	**2**	0,0144	0,1787	0,2852	0,2428	0,1540	0,0803	0,0358	0,0138	0,0046	0,0013	0,0003
19	**3**	0,0008	0,0533	0,1796	0,2428	0,2182	0,1517	0,0869	0,0422	0,0175	0,0062	0,0018
19	**4**	0,0000	0,0112	0,0798	0,1714	0,2182	0,2023	0,1491	0,0909	0,0467	0,0203	0,0074
19	**5**	0,0000	0,0018	0,0266	0,0907	0,1636	0,2023	0,1916	0,1468	0,0933	0,0497	0,0222
19	**6**	0,0000	0,0002	0,0069	0,0374	0,0955	0,1574	0,1916	0,1844	0,1451	0,0949	0,0518
19	**7**	0,0000	0,0000	0,0014	0,0122	0,0443	0,0974	0,1525	0,1844	0,1797	0,1443	0,0961
19	**8**	0,0000	0,0000	0,0002	0,0032	0,0166	0,0487	0,0981	0,1489	0,1797	0,1771	0,1442
19	**9**	0,0000	0,0000	0,0000	0,0007	0,0051	0,0198	0,0514	0,0980	0,1464	0,1771	0,1762
19	**10**	0,0000	0,0000	0,0000	0,0001	0,0013	0,0066	0,0220	0,0528	0,0976	0,1449	0,1762
19	**11**	0,0000	0,0000	0,0000	0,0000	0,0003	0,0018	0,0077	0,0233	0,0532	0,0970	0,1442
19	**12**	0,0000	0,0000	0,0000	0,0000	0,0000	0,0004	0,0022	0,0083	0,0237	0,0529	0,0961
19	**13**	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0024	0,0085	0,0233	0,0518
19	**14**	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0006	0,0024	0,0082	0,0222
19	**15**	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0022	0,0074
19	**16**	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0005	0,0018
19	**17**	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0003
19	**18**	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
19	**19**	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
20	**0**	0,8179	0,3585	0,1216	0,0388	0,0115	0,0032	0,0008	0,0002	0,0000	0,0000	0,0000
20	**1**	0,1652	0,3774	0,2702	0,1368	0,0576	0,0211	0,0068	0,0020	0,0005	0,0001	0,0000
20	**2**	0,0159	0,1887	0,2852	0,2293	0,1369	0,0669	0,0278	0,0100	0,0031	0,0008	0,0002
20	**3**	0,0010	0,0596	0,1901	0,2428	0,2054	0,1339	0,0716	0,0323	0,0123	0,0040	0,0011
20	**4**	0,0000	0,0133	0,0898	0,1821	0,2182	0,1897	0,1304	0,0738	0,0350	0,0139	0,0046
20	**5**	0,0000	0,0022	0,0319	0,1028	0,1746	0,2023	0,1789	0,1272	0,0746	0,0365	0,0148
20	**6**	0,0000	0,0003	0,0089	0,0454	0,1091	0,1686	0,1916	0,1712	0,1244	0,0746	0,0370
20	**7**	0,0000	0,0000	0,0020	0,0160	0,0545	0,1124	0,1643	0,1844	0,1659	0,1221	0,0739
20	**8**	0,0000	0,0000	0,0004	0,0046	0,0222	0,0609	0,1144	0,1614	0,1797	0,1623	0,1201
20	**9**	0,0000	0,0000	0,0001	0,0011	0,0074	0,0271	0,0654	0,1158	0,1597	0,1771	0,1602
20	**10**	0,0000	0,0000	0,0000	0,0002	0,0020	0,0099	0,0308	0,0686	0,1171	0,1593	0,1762
20	**11**	0,0000	0,0000	0,0000	0,0000	0,0005	0,0030	0,0120	0,0336	0,0710	0,1185	0,1602
20	**12**	0,0000	0,0000	0,0000	0,0000	0,0001	0,0008	0,0039	0,0136	0,0355	0,0727	0,1201
20	**13**	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0010	0,0045	0,0146	0,0366	0,0739
20	**14**	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0012	0,0049	0,0150	0,0370
20	**15**	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0003	0,0013	0,0049	0,0148
20	**16**	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0003	0,0013	0,0046
20	**17**	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002	0,0011
20	**18**	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0002
20	**19**	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
20	**20**	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000

TABLA II: FUNCIÓN DE DISRIBUCIÓN BINOMIAL

$$F(x) = P(X \leq x) = \sum \binom{n}{x} p^x \, q^{n-x}$$

n	**x**	**Probabilidad de éxito (p)**										
		0,01	**0,05**	**0,10**	**0,15**	**0,20**	**0,25**	**0,30**	**0,35**	**0,40**	**0,45**	**0,50**
1	**0**	0,9900	0,9500	0,9000	0,8500	0,8000	0,7500	0,7000	0,6500	0,6000	0,5500	0,5000
1	**1**	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
2	**0**	0,9801	0,9025	0,8100	0,7225	0,6400	0,5625	0,4900	0,4225	0,3600	0,3025	0,2500
2	**1**	0,9999	0,9975	0,9900	0,9775	0,9600	0,9375	0,9100	0,8775	0,8400	0,7975	0,7500
2	**2**	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
3	**0**	0,9703	0,8574	0,7290	0,6141	0,5120	0,4219	0,3430	0,2746	0,2160	0,1664	0,1250
3	**1**	0,9997	0,9928	0,9720	0,9393	0,8960	0,8438	0,7840	0,7183	0,6480	0,5748	0,5000
3	**2**	1,0000	0,9999	0,9990	0,9966	0,9920	0,9844	0,9730	0,9571	0,9360	0,9089	0,8750
3	**3**		1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
4	**0**	0,9606	0,8145	0,6561	0,5220	0,4096	0,3164	0,2401	0,1785	0,1296	0,0915	0,0625
4	**1**	0,9994	0,9860	0,9477	0,8905	0,8192	0,7383	0,6517	0,5630	0,4752	0,3910	0,3125
4	**2**	1,0000	0,9995	0,9963	0,9880	0,9728	0,9492	0,9163	0,8735	0,8208	0,7585	0,6875
4	**3**		1,0000	0,9999	0,9995	0,9984	0,9961	0,9919	0,9850	0,9744	0,9590	0,9375
4	**4**			1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
5	**0**	0,9510	0,7738	0,5905	0,4437	0,3277	0,2373	0,1681	0,1160	0,0778	0,0503	0,0313
5	**1**	0,9990	0,9774	0,9185	0,8352	0,7373	0,6328	0,5282	0,4284	0,3370	0,2562	0,1875
5	**2**	1,0000	0,9988	0,9914	0,9734	0,9421	0,8965	0,8369	0,7648	0,6826	0,5931	0,5000
5	**3**		1,0000	0,9995	0,9978	0,9933	0,9844	0,9692	0,9460	0,9130	0,8688	0,8125
5	**4**			1,0000	0,9999	0,9997	0,9990	0,9976	0,9947	0,9898	0,9815	0,9688
5	**5**				1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
6	**0**	0,9415	0,7351	0,5314	0,3771	0,2621	0,1780	0,1176	0,0754	0,0467	0,0277	0,0156
6	**1**	0,9985	0,9672	0,8857	0,7765	0,6554	0,5339	0,4202	0,3191	0,2333	0,1636	0,1094
6	**2**	1,0000	0,9978	0,9842	0,9527	0,9011	0,8306	0,7443	0,6471	0,5443	0,4415	0,3438
6	**3**		0,9999	0,9987	0,9941	0,9830	0,9624	0,9295	0,8826	0,8208	0,7447	0,6563
6	**4**		1,0000	0,9999	0,9996	0,9984	0,9954	0,9891	0,9777	0,9590	0,9308	0,8906
6	**5**			1,0000	1,0000	0,9999	0,9998	0,9993	0,9982	0,9959	0,9917	0,9844
6	**6**					1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
7	**0**	0,9321	0,6983	0,4783	0,3206	0,2097	0,1335	0,0824	0,0490	0,0280	0,0152	0,0078
7	**1**	0,9980	0,9556	0,8503	0,7166	0,5767	0,4449	0,3294	0,2338	0,1586	0,1024	0,0625
7	**2**	1,0000	0,9962	0,9743	0,9262	0,8520	0,7564	0,6471	0,5323	0,4199	0,3164	0,2266
7	**3**		0,9998	0,9973	0,9879	0,9667	0,9294	0,8740	0,8002	0,7102	0,6083	0,5000
7	**4**		1,0000	0,9998	0,9988	0,9953	0,9871	0,9712	0,9444	0,9037	0,8471	0,7734
7	**5**			1,0000	0,9999	0,9996	0,9987	0,9962	0,9910	0,9812	0,9643	0,9375
7	**6**				1,0000	1,0000	0,9999	0,9998	0,9994	0,9984	0,9963	0,9922
7	**7**						1,0000	1,0000	1,0000	1,0000	1,0000	1,0000

Los valores interiores de la tabla indican las probabilidadades acumuladas de obtener de 0 a «x» éxitos en «n» ensayos de un experimento binomial, donde «p» es la probabilidad de éxito en un ensayo.

TABLA II (Cont.)

n	x	Probabilidad de éxito (p)										
		0,01	0,05	0,10	0,15	0,20	0,25	0,30	0,35	0,40	0,45	0,50
8	0	0,9227	0,6634	0,4305	0,2725	0,1678	0,1001	0,0576	0,0319	0,0168	0,0084	0,0039
8	1	0,9973	0,9428	0,8131	0,6572	0,5033	0,3671	0,2553	0,1691	0,1064	0,0632	0,0352
8	2	0,9999	0,9942	0,9619	0,8948	0,7969	0,6785	0,5518	0,4278	0,3154	0,2201	0,1445
8	3	1,0000	0,9996	0,9950	0,9786	0,9437	0,8862	0,8059	0,7064	0,5941	0,4770	0,3633
8	4		1,0000	0,9996	0,9971	0,9896	0,9727	0,9420	0,8939	0,8263	0,7396	0,6367
8	5			1,0000	0,9998	0,9988	0,9958	0,9887	0,9747	0,9502	0,9115	0,8555
8	6				1,0000	0,9999	0,9996	0,9987	0,9964	0,9915	0,9819	0,9648
8	7					1,0000	1,0000	0,9999	0,9998	0,9993	0,9983	0,9961
8	8							1,0000	1,0000	1,0000	1,0000	1,0000
9	0	0,9135	0,6302	0,3874	0,2316	0,1342	0,0751	0,0404	0,0207	0,0101	0,0046	0,0020
9	1	0,9966	0,9288	0,7748	0,5995	0,4362	0,3003	0,1960	0,1211	0,0705	0,0385	0,0195
9	2	0,9999	0,9916	0,9470	0,8591	0,7382	0,6007	0,4628	0,3373	0,2318	0,1495	0,0898
9	3	1,0000	0,9994	0,9917	0,9661	0,9144	0,8343	0,7297	0,6089	0,4826	0,3614	0,2539
9	4		1,0000	0,9991	0,9944	0,9804	0,9511	0,9012	0,8283	0,7334	0,6214	0,5000
9	5			0,9999	0,9994	0,9969	0,9900	0,9747	0,9464	0,9006	0,8342	0,7461
9	6			1,0000	1,0000	0,9997	0,9987	0,9957	0,9888	0,9750	0,9502	0,9102
9	7					1,0000	0,9999	0,9996	0,9986	0,9962	0,9909	0,9805
9	8						1,0000	1,0000	0,9999	0,9997	0,9992	0,9980
9	9								1,0000	1,0000	1,0000	1,0000
10	0	0,9044	0,5987	0,3487	0,1969	0,1074	0,0563	0,0282	0,0135	0,0060	0,0025	0,0010
10	1	0,9957	0,9139	0,7361	0,5443	0,3758	0,2440	0,1493	0,0860	0,0464	0,0233	0,0107
10	2	0,9999	0,9885	0,9298	0,8202	0,6778	0,5256	0,3828	0,2616	0,1673	0,0996	0,0547
10	3	1,0000	0,9990	0,9872	0,9500	0,8791	0,7759	0,6496	0,5138	0,3823	0,2660	0,1719
10	4		0,9999	0,9984	0,9901	0,9672	0,9219	0,8497	0,7515	0,6331	0,5044	0,3770
10	5		1,0000	0,9999	0,9986	0,9936	0,9803	0,9527	0,9051	0,8338	0,7384	0,6230
10	6			1,0000	0,9999	0,9991	0,9965	0,9894	0,9740	0,9452	0,8980	0,8281
10	7				1,0000	0,9999	0,9996	0,9984	0,9952	0,9877	0,9726	0,9453
10	8					1,0000	1,0000	0,9999	0,9995	0,9983	0,9955	0,9893
10	9							1,0000	1,0000	0,9999	0,9997	0,9990
10	10									1,0000	1,0000	1,0000
11	0	0,8953	0,5688	0,3138	0,1673	0,0859	0,0422	0,0198	0,0088	0,0036	0,0014	0,0005
11	1	0,9948	0,8981	0,6974	0,4922	0,3221	0,1971	0,1130	0,0606	0,0302	0,0139	0,0059

TABLA II (Cont.)

n	*x*	Probabilidad de éxito (*p*)										
		0,01	**0,05**	**0,10**	**0,15**	**0,20**	**0,25**	**0,30**	**0,35**	**0,40**	**0,45**	**0,50**
11	**2**	0,9998	0,9848	0,9104	0,7788	0,6174	0,4552	0,3127	0,2001	0,1189	0,0652	0,0327
11	**3**	1,0000	0,9984	0,9815	0,9306	0,8389	0,7133	0,5696	0,4256	0,2963	0,1911	0,1133
11	**4**		0,9999	0,9972	0,9841	0,9496	0,8854	0,7897	0,6683	0,5328	0,3971	0,2744
11	**5**		1,0000	0,9997	0,9973	0,9883	0,9657	0,9218	0,8513	0,7535	0,6331	0,5000
11	**6**			1,0000	0,9997	0,9980	0,9924	0,9784	0,9499	0,9006	0,8262	0,7256
11	**7**				1,0000	0,9998	0,9988	0,9957	0,9878	0,9707	0,9390	0,8867
11	**8**					1,0000	0,9999	0,9994	0,9980	0,9941	0,9852	0,9673
11	**9**						1,0000	1,0000	0,9998	0,9993	0,9978	0,9941
11	**10**								1,0000	1,0000	0,9998	0,9995
11	**11**										1,0000	1,0000
12	**0**	0,8864	0,5404	0,2824	0,1422	0,0687	0,0317	0,0138	0,0057	0,0022	0,0008	0,0002
12	**1**	0,9938	0,8816	0,6590	0,4435	0,2749	0,1584	0,0850	0,0424	0,0196	0,0083	0,0032
12	**2**	0,9998	0,9804	0,8891	0,7358	0,5583	0,3907	0,2528	0,1513	0,0834	0,0421	0,0193
12	**3**	1,0000	0,9978	0,9744	0,9078	0,7946	0,6488	0,4925	0,3467	0,2253	0,1345	0,0730
12	**4**		0,9998	0,9957	0,9761	0,9274	0,8424	0,7237	0,5833	0,4382	0,3044	0,1938
12	**5**		1,0000	0,9995	0,9954	0,9806	0,9456	0,8822	0,7873	0,6652	0,5269	0,3872
12	**6**			0,9999	0,9993	0,9961	0,9857	0,9614	0,9154	0,8418	0,7393	0,6128
12	**7**			1,0000	0,9999	0,9994	0,9972	0,9905	0,9745	0,9427	0,8883	0,8062
12	**8**				1,0000	0,9999	0,9996	0,9983	0,9944	0,9847	0,9644	0,9270
12	**9**					1,0000	1,0000	0,9998	0,9992	0,9972	0,9921	0,9807
12	**10**							1,0000	0,9999	0,9997	0,9989	0,9968
12	**11**								1,0000	1,0000	0,9999	0,9998
12	**12**										1,0000	1,0000
13	**0**	0,8775	0,5133	0,2542	0,1209	0,0550	0,0238	0,0097	0,0037	0,0013	0,0004	0,0001
13	**1**	0,9928	0,8646	0,6213	0,3983	0,2336	0,1267	0,0637	0,0296	0,0126	0,0049	0,0017
13	**2**	0,9997	0,9755	0,8661	0,6920	0,5017	0,3326	0,2025	0,1132	0,0579	0,0269	0,0112
13	**3**	1,0000	0,9969	0,9658	0,8820	0,7473	0,5843	0,4206	0,2783	0,1686	0,0929	0,0461
13	**4**		0,9997	0,9935	0,9658	0,9009	0,7940	0,6543	0,5005	0,3530	0,2279	0,1334
13	**5**		1,0000	0,9991	0,9925	0,9700	0,9198	0,8346	0,7159	0,5744	0,4268	0,2905
13	**6**			0,9999	0,9987	0,9930	0,9757	0,9376	0,8705	0,7712	0,6437	0,5000
13	**7**			1,0000	0,9998	0,9988	0,9944	0,9818	0,9538	0,9023	0,8212	0,7095
13	**8**				1,0000	0,9998	0,9990	0,9960	0,9874	0,9679	0,9302	0,8666
13	**9**					1,0000	0,9999	0,9993	0,9975	0,9922	0,9797	0,9539
13	**10**						1,0000	0,9999	0,9997	0,9987	0,9959	0,9888
13	**11**							1,0000	1,0000	0,9999	0,9995	0,9983
13	**12**									1,0000	1,0000	0,9999
13	**13**											1,0000

TABLA II (Cont.)

n	*x*	Probabilidad de éxito (*p*)										
		0,01	**0,05**	**0,10**	**0,15**	**0,20**	**0,25**	**0,30**	**0,35**	**0,40**	**0,45**	**0,50**
14	**0**	0,8687	0,4877	0,2288	0,1028	0,0440	0,0178	0,0068	0,0024	0,0008	0,0002	0,0001
14	**1**	0,9916	0,8470	0,5846	0,3567	0,1979	0,1010	0,0475	0,0205	0,0081	0,0029	0,0009
14	**2**	0,9997	0,9699	0,8416	0,6479	0,4481	0,2811	0,1608	0,0839	0,0398	0,0170	0,0065
14	**3**	1,0000	0,9958	0,9559	0,8535	0,6982	0,5213	0,3552	0,2205	0,1243	0,0632	0,0287
14	**4**		0,9996	0,9908	0,9533	0,8702	0,7415	0,5842	0,4227	0,2793	0,1672	0,0898
14	**5**		1,0000	0,9985	0,9885	0,9561	0,8883	0,7805	0,6405	0,4859	0,3373	0,2120
14	**6**			0,9998	0,9978	0,9884	0,9617	0,9067	0,8164	0,6925	0,5461	0,3953
14	**7**			1,0000	0,9997	0,9976	0,9897	0,9685	0,9247	0,8499	0,7414	0,6047
14	**8**				1,0000	0,9996	0,9978	0,9917	0,9757	0,9417	0,8811	0,7880
14	**9**					1,0000	0,9997	0,9983	0,9940	0,9825	0,9574	0,9102
14	**10**						1,0000	0,9998	0,9989	0,9961	0,9886	0,9713
14	**11**							1,0000	0,9999	0,9994	0,9978	0,9935
14	**12**								1,0000	0,9999	0,9997	0,9991
14	**13**									1,0000	1,0000	0,9999
14	**14**											1,0000
15	**0**	0,8601	0,4633	0,2059	0,0874	0,0352	0,0134	0,0047	0,0016	0,0005	0,0001	0,0000
15	**1**	0,9904	0,8290	0,5490	0,3186	0,1671	0,0802	0,0353	0,0142	0,0052	0,0017	0,0005
15	**2**	0,9996	0,9638	0,8159	0,6042	0,3980	0,2361	0,1268	0,0617	0,0271	0,0107	0,0037
15	**3**	1,0000	0,9945	0,9444	0,8227	0,6482	0,4613	0,2969	0,1727	0,0905	0,0424	0,0176
15	**4**		0,9994	0,9873	0,9383	0,8358	0,6865	0,5155	0,3519	0,2173	0,1204	0,0592
15	**5**		0,9999	0,9978	0,9832	0,9389	0,8516	0,7216	0,5643	0,4032	0,2608	0,1509
15	**6**		1,0000	0,9997	0,9964	0,9819	0,9434	0,8689	0,7548	0,6098	0,4522	0,3036
15	**7**			1,0000	0,9994	0,9958	0,9827	0,9500	0,8868	0,7869	0,6535	0,5000
15	**8**				0,9999	0,9992	0,9958	0,9848	0,9578	0,9050	0,8182	0,6964
15	**9**				1,0000	0,9999	0,9992	0,9963	0,9876	0,9662	0,9231	0,8491
15	**10**					1,0000	0,9999	0,9993	0,9972	0,9907	0,9745	0,9408
15	**11**						1,0000	0,9999	0,9995	0,9981	0,9937	0,9824
15	**12**							1,0000	0,9999	0,9997	0,9989	0,9963
15	**13**								1,0000	1,0000	0,9999	0,9995
15	**14**									1,0000	1,0000	1,0000
15	**15**											
16	**0**	0,8515	0,4401	0,1853	0,0743	0,0281	0,0100	0,0033	0,0010	0,0003	0,0001	0,0000
16	**1**	0,9891	0,8108	0,5147	0,2839	0,1407	0,0635	0,0261	0,0098	0,0033	0,0010	0,0003
16	**2**	0,9995	0,9571	0,7892	0,5614	0,3518	0,1971	0,0994	0,0451	0,0183	0,0066	0,0021
16	**3**	1,0000	0,9930	0,9316	0,7899	0,5981	0,4050	0,2459	0,1339	0,0651	0,0281	0,0106
16	**4**		0,9991	0,9830	0,9209	0,7982	0,6302	0,4499	0,2892	0,1666	0,0853	0,0384
16	**5**		0,9999	0,9967	0,9765	0,9183	0,8103	0,6598	0,4900	0,3288	0,1976	0,1051
16	**6**		1,0000	0,9995	0,9944	0,9733	0,9204	0,8247	0,6881	0,5272	0,3660	0,2272
16	**7**			0,9999	0,9989	0,9930	0,9729	0,9256	0,8406	0,7161	0,5629	0,4018
16	**8**			1,0000	0,9998	0,9985	0,9925	0,9743	0,9329	0,8577	0,7441	0,5982
16	**9**				1,0000	0,9998	0,9984	0,9929	0,9771	0,9417	0,8759	0,7728
16	**10**					1,0000	0,9997	0,9984	0,9938	0,9809	0,9514	0,8949
16	**11**						1,0000	0,9997	0,9987	0,9951	0,9851	0,9616

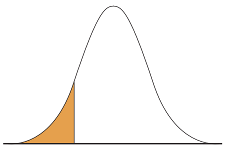
TABLA II (Cont.)

n	*x*	Probabilidad de éxito (*p*)										
		0,01	**0,05**	**0,10**	**0,15**	**0,20**	**0,25**	**0,30**	**0,35**	**0,40**	**0,45**	**0,50**
16	**12**							1,0000	0,9998	0,9991	0,9965	0,9894
16	**13**								1,0000	0,9999	0,9994	0,9979
16	**14**									1,0000	0,9999	0,9997
16	**15**										1,0000	1,0000
16	**16**											
17	**0**	0,8429	0,4181	0,1668	0,0631	0,0225	0,0075	0,0023	0,0007	0,0002	0,0000	0,0000
17	**1**	0,9877	0,7922	0,4818	0,2525	0,1182	0,0501	0,0193	0,0067	0,0021	0,0006	0,0001
17	**2**	0,9994	0,9497	0,7618	0,5198	0,3096	0,1637	0,0774	0,0327	0,0123	0,0041	0,0012
17	**3**	1,0000	0,9912	0,9174	0,7556	0,5489	0,3530	0,2019	0,1028	0,0464	0,0184	0,0064
17	**4**		0,9988	0,9779	0,9013	0,7582	0,5739	0,3887	0,2348	0,1260	0,0596	0,0245
17	**5**		0,9999	0,9953	0,9681	0,8943	0,7653	0,5968	0,4197	0,2639	0,1471	0,0717
17	**6**		1,0000	0,9992	0,9917	0,9623	0,8929	0,7752	0,6188	0,4478	0,2902	0,1662
17	**7**			0,9999	0,9983	0,9891	0,9598	0,8954	0,7872	0,6405	0,4743	0,3145
17	**8**			1,0000	0,9997	0,9974	0,9876	0,9597	0,9006	0,8011	0,6626	0,5000
17	**9**				1,0000	0,9995	0,9969	0,9873	0,9617	0,9081	0,8166	0,6855
17	**10**					0,9999	0,9994	0,9968	0,9880	0,9652	0,9174	0,8338
17	**11**					1,0000	0,9999	0,9993	0,9970	0,9894	0,9699	0,9283
17	**12**						1,0000	0,9999	0,9994	0,9975	0,9914	0,9755
17	**13**							1,0000	0,9999	0,9995	0,9981	0,9936
17	**14**								1,0000	0,9999	0,9997	0,9988
17	**15**									1,0000	1,0000	0,9999
17	**16**											1,0000
17	**17**											
18	**0**	0,8345	0,3972	0,1501	0,0536	0,0180	0,0056	0,0016	0,0004	0,0001	0,0000	0,0000
18	**1**	0,9862	0,7735	0,4503	0,2241	0,0991	0,0395	0,0142	0,0046	0,0013	0,0003	0,0001
18	**2**	0,9993	0,9419	0,7338	0,4797	0,2713	0,1353	0,0600	0,0236	0,0082	0,0025	0,0007
18	**3**	1,0000	0,9891	0,9018	0,7202	0,5010	0,3057	0,1646	0,0783	0,0328	0,0120	0,0038
18	**4**		0,9985	0,9718	0,8794	0,7164	0,5187	0,3327	0,1886	0,0942	0,0411	0,0154
18	**5**		0,9998	0,9936	0,9581	0,8671	0,7175	0,5344	0,3550	0,2088	0,1077	0,0481
18	**6**		1,0000	0,9988	0,9882	0,9487	0,8610	0,7217	0,5491	0,3743	0,2258	0,1189
18	**7**			0,9998	0,9973	0,9837	0,9431	0,8593	0,7283	0,5634	0,3915	0,2403
18	**8**			1,0000	0,9995	0,9957	0,9807	0,9404	0,8609	0,7368	0,5778	0,4073
18	**9**				0,9999	0,9991	0,9946	0,9790	0,9403	0,8653	0,7473	0,5927
18	**10**				1,0000	0,9998	0,9988	0,9939	0,9788	0,9424	0,8720	0,7597
18	**11**					1,0000	0,9998	0,9986	0,9938	0,9797	0,9463	0,8811
18	**12**						1,0000	0,9997	0,9986	0,9942	0,9817	0,9519
18	**13**							1,0000	0,9997	0,9987	0,9951	0,9846
18	**14**								1,0000	0,9998	0,9990	0,9962
18	**15**									1,0000	0,9999	0,9993
18	**16**										1,0000	0,9999
18	**17**											1,0000
18	**18**											

TABLA II (Cont.)

n	x	0,01	0,05	0,10	0,15	0,20	0,25	0,30	0,35	0,40	0,45	0,50
		Probabilidad de éxito (p)										
19	0	0,8262	0,3774	0,1351	0,0456	0,0144	0,0042	0,0011	0,0003	0,0001	0,0000	0,0000
19	1	0,9847	0,7547	0,4203	0,1985	0,0829	0,0310	0,0104	0,0031	0,0008	0,0002	0,0000
19	2	0,9991	0,9335	0,7054	0,4413	0,2369	0,1113	0,0462	0,0170	0,0055	0,0015	0,0004
19	3	1,0000	0,9868	0,8850	0,6841	0,4551	0,2631	0,1332	0,0591	0,0230	0,0077	0,0022
19	4		0,9980	0,9648	0,8556	0,6733	0,4654	0,2822	0,1500	0,0696	0,0280	0,0096
19	5		0,9998	0,9914	0,9463	0,8369	0,6678	0,4739	0,2968	0,1629	0,0777	0,0318
19	6		1,0000	0,9983	0,9837	0,9324	0,8251	0,6655	0,4812	0,3081	0,1727	0,0835
19	7			0,9997	0,9959	0,9767	0,9225	0,8180	0,6656	0,4878	0,3169	0,1796
19	8			1,0000	0,9992	0,9933	0,9713	0,9161	0,8145	0,6675	0,4940	0,3238
19	9				0,9999	0,9984	0,9911	0,9674	0,9125	0,8139	0,6710	0,5000
19	10				1,0000	0,9997	0,9977	0,9895	0,9653	0,9115	0,8159	0,6762
19	11					1,0000	0,9995	0,9972	0,9886	0,9648	0,9129	0,8204
19	12						0,9999	0,9994	0,9969	0,9884	0,9658	0,9165
19	13						1,0000	0,9999	0,9993	0,9969	0,9891	0,9682
19	14							1,0000	0,9999	0,9994	0,9972	0,9904
19	15								1,0000	0,9999	0,9995	0,9978
19	16									1,0000	0,9999	0,9996
19	17										1,0000	1,0000
19	18											
19	19											
20	0	0,8179	0,3585	0,1216	0,0388	0,0115	0,0032	0,0008	0,0002	0,0000	0,0000	0,0000
20	1	0,9831	0,7358	0,3917	0,1756	0,0692	0,0243	0,0076	0,0021	0,0005	0,0001	0,0000
20	2	0,9990	0,9245	0,6769	0,4049	0,2061	0,0913	0,0355	0,0121	0,0036	0,0009	0,0002
20	3	1,0000	0,9841	0,8670	0,6477	0,4114	0,2252	0,1071	0,0444	0,0160	0,0049	0,0013
20	4		0,9974	0,9568	0,8298	0,6296	0,4148	0,2375	0,1182	0,0510	0,0189	0,0059
20	5		0,9997	0,9887	0,9327	0,8042	0,6172	0,4164	0,2454	0,1256	0,0553	0,0207
20	6		1,0000	0,9976	0,9781	0,9133	0,7858	0,6080	0,4166	0,2500	0,1299	0,0577
20	7			0,9996	0,9941	0,9679	0,8982	0,7723	0,6010	0,4159	0,2520	0,1316
20	8			0,9999	0,9987	0,9900	0,9591	0,8867	0,7624	0,5956	0,4143	0,2517
20	9			1,0000	0,9998	0,9974	0,9861	0,9520	0,8782	0,7553	0,5914	0,4119
20	10				1,0000	0,9994	0,9961	0,9829	0,9468	0,8725	0,7507	0,5881
20	11					0,9999	0,9991	0,9949	0,9804	0,9435	0,8692	0,7483
20	12					1,0000	0,9998	0,9987	0,9940	0,9790	0,9420	0,8684
20	13						1,0000	0,9997	0,9985	0,9935	0,9786	0,9423
20	14							1,0000	0,9997	0,9984	0,9936	0,9793
20	15								1,0000	0,9997	0,9985	0,9941
20	16									1,0000	0,9997	0,9987
20	17										1,0000	0,9998
20	18											1,0000
20	19											
20	20											

TABLA III: DISTRIBUCIÓN NORMAL TIPIFICADA

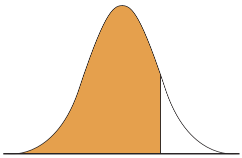


$P(Z \leq z)$

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
–3,5	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002
–3,4	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0003	0,0002
–3,3	0,0005	0,0005	0,0005	0,0004	0,0004	0,0004	0,0004	0,0004	0,0004	0,0003
–3,2	0,0007	0,0007	0,0006	0,0006	0,0006	0,0006	0,0006	0,0005	0,0005	0,0005
–3,1	0,0010	0,0009	0,0009	0,0009	0,0008	0,0008	0,0008	0,0008	0,0007	0,0007
–3,0	0,0013	0,0013	0,0013	0,0012	0,0012	0,0011	0,0011	0,0011	0,0010	0,0010
–2,9	0,0019	0,0018	0,0018	0,0017	0,0016	0,0016	0,0015	0,0015	0,0014	0,0014
–2,8	0,0026	0,0025	0,0024	0,0023	0,0023	0,0022	0,0021	0,0021	0,0020	0,0019
–2,7	0,0035	0,0034	0,0033	0,0032	0,0031	0,0030	0,0029	0,0028	0,0027	0,0026
–2,6	0,0047	0,0045	0,0044	0,0043	0,0041	0,0040	0,0039	0,0038	0,0037	0,0036
–2,5	0,0062	0,0060	0,0059	0,0057	0,0055	0,0054	0,0052	0,0051	0,0049	0,0048
–2,4	0,0082	0,0080	0,0078	0,0075	0,0073	0,0071	0,0069	0,0068	0,0066	0,0064
–2,3	0,0107	0,0104	0,0102	0,0099	0,0096	0,0094	0,0091	0,0089	0,0087	0,0084
–2,2	0,0139	0,0136	0,0132	0,0129	0,0125	0,0122	0,0119	0,0116	0,0113	0,0110
–2,1	0,0179	0,0174	0,0170	0,0166	0,0162	0,0158	0,0154	0,0150	0,0146	0,0143
–2,0	0,0228	0,0222	0,0217	0,0212	0,0207	0,0202	0,0197	0,0192	0,0188	0,0183
–1,9	0,0287	0,0281	0,0274	0,0268	0,0262	0,0256	0,0250	0,0244	0,0239	0,0233
–1,8	0,0359	0,0351	0,0344	0,0336	0,0329	0,0322	0,0314	0,0307	0,0301	0,0294
–1,7	0,0446	0,0436	0,0427	0,0418	0,0409	0,0401	0,0392	0,0384	0,0375	0,0367
–1,6	0,0548	0,0537	0,0526	0,0516	0,0505	0,0495	0,0485	0,0475	0,0465	0,0455
–1,5	0,0668	0,0655	0,0643	0,0630	0,0618	0,0606	0,0594	0,0582	0,0571	0,0559
–1,4	0,0808	0,0793	0,0778	0,0764	0,0749	0,0735	0,0721	0,0708	0,0694	0,0681
–1,3	0,0968	0,0951	0,0934	0,0918	0,0901	0,0885	0,0869	0,0853	0,0838	0,0823
–1,2	0,1151	0,1131	0,1112	0,1093	0,1075	0,1056	0,1038	0,1020	0,1003	0,0985
–1,1	0,1357	0,1335	0,1314	0,1292	0,1271	0,1251	0,1230	0,1210	0,1190	0,1170
–1,0	0,1587	0,1562	0,1539	0,1515	0,1492	0,1469	0,1446	0,1423	0,1401	0,1379
–0,9	0,1841	0,1814	0,1788	0,1762	0,1736	0,1711	0,1685	0,1660	0,1635	0,1611
–0,8	0,2119	0,2090	0,2061	0,2033	0,2005	0,1977	0,1949	0,1922	0,1894	0,1867
–0,7	0,2420	0,2389	0,2358	0,2327	0,2296	0,2266	0,2236	0,2206	0,2177	0,2148
–0,6	0,2743	0,2709	0,2676	0,2643	0,2611	0,2578	0,2546	0,2514	0,2483	0,2451
–0,5	0,3085	0,3050	0,3015	0,2981	0,2946	0,2912	0,2877	0,2843	0,2810	0,2776
–0,4	0,3446	0,3409	0,3372	0,3336	0,3300	0,3264	0,3228	0,3192	0,3156	0,3121
–0,3	0,3821	0,3783	0,3745	0,3707	0,3669	0,3632	0,3594	0,3557	0,3520	0,3483
–0,2	0,4207	0,4168	0,4129	0,4090	0,4052	0,4013	0,3974	0,3936	0,3897	0,3859
–0,1	0,4602	0,4562	0,4522	0,4483	0,4443	0,4404	0,4364	0,4325	0,4286	0,4247
–0,0	0,5000	0,4960	0,4920	0,4880	0,4840	0,4801	0,4761	0,4721	0,4681	0,4641

Los valores interiores representan la probabilidad de obtener valores de *Z* menores o iguales que la puntuación típica, «*z*», definida por el cruce de la fila con la columna indicativa del segundo decimal. Así, por ejemplo, la probabilidad de obtener puntuaciones menores o iguales que –1,05 es 0,1469. Es decir $P(Z \leq -1{,}05) = 0{,}1469$.

TABLA IV: DISTRIBUCIÓN NORMAL TIPIFICADA

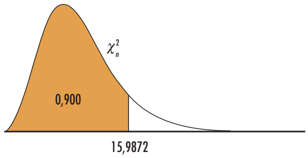


$P(Z \leq z)$

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990
3,1	0,9990	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,2	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995
3,3	0,9995	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997
3,4	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998
3,5	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998

Los valores interiores representan la probabilidad de obtener valores de *Z* menores o iguales que la puntuación típica, «*z*», definida por el cruce de la fila con la columna indicativa del segundo decimal. Así, por ejemplo, la probabilidad de obtener puntuaciones menores o iguales que 1,05 es 0,8531. Es decir $P(Z \leq 1{,}05) = 0{,}8531$.

TABLA V: DISTRIBUCIÓN CHI-CUADRADO

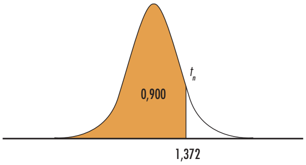


$$P(X \leq \chi^2_{gl})$$

g.l.	Probabilidad									
	0,005	**0,010**	**0,025**	**0,050**	**0,100**	**0,900**	**0,950**	**0,975**	**0,990**	**0,995**
1	0,0000	0,0002	0,0010	0,0039	0,0158	2,7055	3,8415	5,0239	6,6349	7,8794
2	0,0100	0,0201	0,0506	0,1026	0,2107	4,6052	5,9915	7,3778	9,2103	10,5966
3	0,0717	0,1148	0,2158	0,3518	0,5844	6,2514	7,8147	9,3484	11,3449	12,8382
4	0,2070	0,2971	0,4844	0,7107	1,0636	7,7794	9,4877	11,1433	13,2767	14,8603
5	0,4117	0,5543	0,8312	1,1455	1,6103	9,2364	11,0705	12,8325	15,0863	16,7496
6	0,6757	0,8721	1,2373	1,6354	2,2041	10,6446	12,5916	14,4494	16,8119	18,5476
7	0,9893	1,2390	1,6899	2,1673	2,8331	12,0170	14,0671	16,0128	18,4753	20,2777
8	1,3444	1,6465	2,1797	2,7326	3,4895	13,3616	15,5073	17,5345	20,0902	21,9550
9	1,7349	2,0879	2,7004	3,3251	4,1682	14,6837	16,9190	19,0228	21,6660	23,5894
10	2,1559	2,5582	3,2470	3,9403	4,8652	15,9872	18,3070	20,4832	23,2093	25,1882
11	2,6032	3,0535	3,8157	4,5748	5,5778	17,2750	19,6751	21,9200	24,7250	26,7568
12	3,0738	3,5706	4,4038	5,2260	6,3038	18,5493	21,0261	23,3367	26,2170	28,2995
13	3,5650	4,1069	5,0088	5,8919	7,0415	19,8119	22,3620	24,7356	27,6882	29,8195
14	4,0747	4,6604	5,6287	6,5706	7,7895	21,0641	23,6848	26,1189	29,1412	31,3193
15	4,6009	5,2293	6,2621	7,2609	8,5468	22,3071	24,9958	27,4884	30,5779	32,8013
16	5,1422	5,8122	6,9077	7,9616	9,3122	23,5418	26,2962	28,8454	31,9999	34,2672
17	5,6972	6,4078	7,5642	8,6718	10,0852	24,7690	27,5871	30,1910	33,4087	35,7185
18	6,2648	7,0149	8,2307	9,3905	10,8649	25,9894	28,8693	31,5264	34,8053	37,1565
19	6,8440	7,6327	8,9065	10,1170	11,6509	27,2036	30,1435	32,8523	36,1909	38,5823
20	7,4338	8,2604	9,5908	10,8508	12,4426	28,4120	31,4104	34,1696	37,5662	39,9968
21	8,0337	8,8972	10,2829	11,5913	13,2396	29,6151	32,6706	35,4789	38,9322	41,4011
22	8,6427	9,5425	10,9823	12,3380	14,0415	30,8133	33,9244	36,7807	40,2894	42,7957
23	9,2604	10,1957	11,6886	13,0905	14,8480	32,0069	35,1725	38,0756	41,6384	44,1813
24	9,8862	10,8564	12,4012	13,8484	15,6587	33,1962	36,4150	39,3641	42,9798	45,5585
25	10,5197	11,5240	13,1197	14,6114	16,4734	34,3816	37,6525	40,6465	44,3141	46,9279
26	11,1602	12,1981	13,8439	15,3792	17,2919	35,5632	38,8851	41,9232	45,6417	48,2899
27	11,8076	12,8785	14,5734	16,1514	18,1139	36,7412	40,1133	43,1945	46,9629	49,6449
28	12,4613	13,5647	15,3079	16,9279	18,9392	37,9159	41,3371	44,4608	48,2782	50,9934
29	13,1211	14,2565	16,0471	17,7084	19,7677	39,0875	42,5570	45,7223	49,5879	52,3356
30	13,7867	14,9535	16,7908	18,4927	20,5992	40,2560	43,7730	46,9792	50,8922	53,6720
40	20,7065	22,1643	24,4330	26,5093	29,0505	51,8051	55,7585	59,3417	63,6907	66,7660
50	27,9907	29,7067	32,3574	34,7643	37,6886	63,1671	67,5048	71,4202	76,1539	79,4900
60	35,5345	37,4849	40,4817	43,1880	46,4589	74,3970	79,0819	83,2977	88,3794	91,9517
70	43,2752	45,4417	48,7576	51,7393	55,3289	85,5270	90,5312	95,0232	100,4252	104,2149
80	51,1719	53,5401	57,1532	60,3915	64,2778	96,5782	101,8795	106,6286	112,3288	116,3211
90	59,1963	61,7541	65,6466	69,1260	73,2911	107,5650	113,1453	118,1359	124,1163	128,2989
100	67,3276	70,0649	74,2219	77,9295	82,3581	118,4980	124,3421	129,5612	135,8067	140,1695

Los números interiores representan valores de la variable chi-cuadrado para una probabilidad menor o igual que la especificada, con g.l. grados de libertad. Por ejemplo, con 10 g.l. la probabilidad de obtener valores menores o iguales que 15,9872 es 0,900.

TABLA VI: DISTRIBUCIÓN *t* DE STUDENT

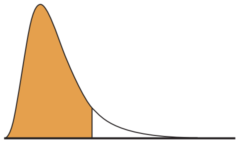


$P(T \leq t_{gl})$

g.l.	Probabilidad											
	0,550	**0,600**	**0,650**	**0,700**	**0,750**	**0,800**	**0,850**	**0,900**	**0,950**	**0,975**	**0,990**	**0,995**
1	0,158	0,325	0,510	0,727	1,000	1,376	1,963	3,078	6,314	12,706	31,821	63,657
2	0,142	0,289	0,445	0,617	0,816	1,061	1,386	1,886	2,920	4,303	6,965	9,925
3	0,137	0,277	0,424	0,584	0,765	0,978	1,250	1,638	2,353	3,182	4,541	5,841
4	0,134	0,271	0,414	0,569	0,741	0,941	1,190	1,533	2,132	2,776	3,747	4,604
5	0,132	0,267	0,408	0,559	0,727	0,920	1,156	1,476	2,015	2,571	3,365	4,032
6	0,131	0,265	0,404	0,553	0,718	0,906	1,134	1,440	1,943	2,447	3,143	3,707
7	0,130	0,263	0,402	0,549	0,711	0,896	1,119	1,415	1,895	2,365	2,998	3,499
8	0,130	0,262	0,399	0,546	0,706	0,889	1,108	1,397	1,860	2,306	2,896	3,355
9	0,129	0,261	0,398	0,543	0,703	0,883	1,100	1,383	1,833	2,262	2,821	3,250
10	0,129	0,260	0,397	0,542	0,700	0,879	1,093	1,372	1,812	2,228	2,764	3,169
11	0,129	0,260	0,396	0,540	0,697	0,876	1,088	1,363	1,796	2,201	2,718	3,106
12	0,128	0,259	0,395	0,539	0,695	0,873	1,083	1,356	1,782	2,179	2,681	3,055
13	0,128	0,259	0,394	0,538	0,694	0,870	1,079	1,350	1,771	2,160	2,650	3,012
14	0,128	0,258	0,393	0,537	0,692	0,868	1,076	1,345	1,761	2,145	2,624	2,977
15	0,128	0,258	0,393	0,536	0,691	0,866	1,074	1,341	1,753	2,131	2,602	2,947
16	0,128	0,258	0,392	0,535	0,690	0,865	1,071	1,337	1,746	2,120	2,583	2,921
17	0,128	0,257	0,392	0,534	0,689	0,863	1,069	1,333	1,740	2,110	2,567	2,898
18	0,127	0,257	0,392	0,534	0,688	0,862	1,067	1,330	1,734	2,101	2,552	2,878
19	0,127	0,257	0,391	0,533	0,688	0,861	1,066	1,328	1,729	2,093	2,539	2,861
20	0,127	0,257	0,391	0,533	0,687	0,860	1,064	1,325	1,725	2,086	2,528	2,845
21	0,127	0,257	0,391	0,532	0,686	0,859	1,063	1,323	1,721	2,080	2,518	2,831
22	0,127	0,256	0,390	0,532	0,686	0,858	1,061	1,321	1,717	2,074	2,508	2,819
23	0,127	0,256	0,390	0,532	0,685	0,858	1,060	1,319	1,714	2,069	2,500	2,807
24	0,127	0,256	0,390	0,531	0,685	0,857	1,059	1,318	1,711	2,064	2,492	2,797
25	0,127	0,256	0,390	0,531	0,684	0,856	1,058	1,316	1,708	2,060	2,485	2,787
26	0,127	0,256	0,390	0,531	0,684	0,856	1,058	1,315	1,706	2,056	2,479	2,779
27	0,127	0,256	0,389	0,531	0,684	0,855	1,057	1,314	1,703	2,052	2,473	2,771
28	0,127	0,256	0,389	0,530	0,683	0,855	1,056	1,313	1,701	2,048	2,467	2,763
29	0,127	0,256	0,389	0,530	0,683	0,854	1,055	1,311	1,699	2,045	2,462	2,756
30	0,127	0,256	0,389	0,530	0,683	0,854	1,055	1,310	1,697	2,042	2,457	2,750
40	0,126	0,255	0,388	0,529	0,681	0,851	1,050	1,303	1,684	2,021	2,423	2,704
50	0,126	0,255	0,388	0,528	0,679	0,849	1,047	1,299	1,676	2,009	2,403	2,678
60	0,126	0,254	0,387	0,527	0,679	0,848	1,045	1,296	1,671	2,000	2,390	2,660
70	0,126	0,254	0,387	0,527	0,678	0,847	1,044	1,294	1,667	1,994	2,381	2,648
80	0,126	0,254	0,387	0,526	0,678	0,846	1,043	1,292	1,664	1,990	2,374	2,639
90	0,126	0,254	0,387	0,526	0,677	0,846	1,042	1,291	1,662	1,987	2,368	2,632
100	0,126	0,254	0,386	0,526	0,677	0,845	1,042	1,290	1,660	1,984	2,364	2,626

Los números interiores representan valores de la variable T para una probabilidad menor o igual que la especificada, con g.l. grados de libertad. Por ejemplo, con 10 g.l. la probabilidad de obtener valores menores o iguales que 1,372 es 0,900.

TABLA VII: DISTRIBUCIÓN *F*



$$P(F_{n_1,n_2} \leq f_{n_1,n_2}) = 0{,}90$$

Grados de libertad del denominador (n_2)	Grados de libertad del numerador (n_1)															
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**20**	**30**	**40**	**50**	**60**	**120**
1	39,863	49,500	53,593	55,833	57,240	58,204	58,906	59,439	59,858	60,195	61,740	62,265	62,529	62,688	62,794	63,061
2	8,526	9,000	9,162	9,243	9,293	9,326	9,349	9,367	9,381	9,392	9,441	9,458	9,466	9,471	9,475	9,483
3	5,538	5,462	5,391	5,343	5,309	5,285	5,266	5,252	5,240	5,230	5,184	5,168	5,160	5,155	5,151	5,143
4	4,545	4,325	4,191	4,107	4,051	4,010	3,979	3,955	3,936	3,920	3,844	3,817	3,804	3,795	3,790	3,775
5	4,060	3,780	3,619	3,520	3,453	3,405	3,368	3,339	3,316	3,297	3,207	3,174	3,157	3,147	3,140	3,123
6	3,776	3,463	3,289	3,181	3,108	3,055	3,014	2,983	2,958	2,937	2,836	2,800	2,781	2,770	2,762	2,742
7	3,589	3,257	3,074	2,961	2,883	2,827	2,785	2,752	2,725	2,703	2,595	2,555	2,535	2,523	2,514	2,493
8	3,458	3,113	2,924	2,806	2,726	2,668	2,624	2,589	2,561	2,538	2,425	2,383	2,361	2,348	2,339	2,316
9	3,360	3,006	2,813	2,693	2,611	2,551	2,505	2,469	2,440	2,416	2,298	2,255	2,232	2,218	2,208	2,184
10	3,285	2,924	2,728	2,605	2,522	2,461	2,414	2,377	2,347	2,323	2,201	2,155	2,132	2,117	2,107	2,082
11	3,225	2,860	2,660	2,536	2,451	2,389	2,342	2,304	2,274	2,248	2,123	2,076	2,052	2,036	2,026	2,000
12	3,177	2,807	2,606	2,480	2,394	2,331	2,283	2,245	2,214	2,188	2,060	2,011	1,986	1,970	1,960	1,932
13	3,136	2,763	2,560	2,434	2,347	2,283	2,234	2,195	2,164	2,138	2,007	1,958	1,931	1,915	1,904	1,876
14	3,102	2,726	2,522	2,395	2,307	2,243	2,193	2,154	2,122	2,095	1,962	1,912	1,885	1,869	1,857	1,828
15	3,073	2,695	2,490	2,361	2,273	2,208	2,158	2,119	2,086	2,059	1,924	1,873	1,845	1,828	1,817	1,787
16	3,048	2,668	2,462	2,333	2,244	2,178	2,128	2,088	2,055	2,028	1,891	1,839	1,811	1,793	1,782	1,751
17	3,026	2,645	2,437	2,308	2,218	2,152	2,102	2,061	2,028	2,001	1,862	1,809	1,781	1,763	1,751	1,719
18	3,007	2,624	2,416	2,286	2,196	2,130	2,079	2,038	2,005	1,977	1,837	1,783	1,754	1,736	1,723	1,691
19	2,990	2,606	2,397	2,266	2,176	2,109	2,058	2,017	1,984	1,956	1,814	1,759	1,730	1,711	1,699	1,666
20	2,975	2,589	2,380	2,249	2,158	2,091	2,040	1,999	1,965	1,937	1,794	1,738	1,708	1,690	1,677	1,643
30	2,881	2,489	2,276	2,142	2,049	1,980	1,927	1,884	1,849	1,819	1,667	1,606	1,573	1,552	1,538	1,499
60	2,791	2,393	2,177	2,041	1,946	1,875	1,819	1,775	1,738	1,707	1,543	1,476	1,437	1,413	1,395	1,348
120	2,748	2,347	2,130	1,992	1,896	1,824	1,767	1,722	1,684	1,652	1,482	1,409	1,368	1,340	1,320	1,265

Los números interiores corresponden a los valores de la variable F con n_1 grados de libertad del numerador y n_2 grados de libertad del denominador. Por ejemplo, $P(F_{10,20} \leq 1{,}937) = 0{,}90$.

TABLA VII: DISTRIBUCIÓN *F*

$$P(F_{n_1,n_2} \leq f_{n_1,n_2}) = 0{,}95$$

Grados de libertad del denominador (n_2)	Grados de libertad del numerador (n_1): 1	2	3	4	5	6	7	8	9	10	20	30	40	50	60	120
1	161,448	199,500	215,707	224,583	230,162	233,986	236,768	238,883	240,543	241,882	248,013	250,095	251,143	251,774	252,196	253,253
2	18,513	19,000	19,164	19,247	19,296	19,330	19,353	19,371	19,385	19,396	19,446	19,462	19,471	19,476	19,479	19,487
3	10,128	9,552	9,277	9,117	9,013	8,941	8,887	8,845	8,812	8,786	8,660	8,617	8,594	8,581	8,572	8,549
4	7,709	6,944	6,591	6,388	6,256	6,163	6,094	6,041	5,999	5,964	5,803	5,746	5,717	5,699	5,688	5,658
5	6,608	5,786	5,409	5,192	5,050	4,950	4,876	4,818	4,772	4,735	4,558	4,496	4,464	4,444	4,431	4,398
6	5,987	5,143	4,757	4,534	4,387	4,284	4,207	4,147	4,099	4,060	3,874	3,808	3,774	3,754	3,740	3,705
7	5,591	4,737	4,347	4,120	3,972	3,866	3,787	3,726	3,677	3,637	3,445	3,376	3,340	3,319	3,304	3,267
8	5,318	4,459	4,066	3,838	3,687	3,581	3,500	3,438	3,388	3,347	3,150	3,079	3,043	3,020	3,005	2,967
9	5,117	4,256	3,863	3,633	3,482	3,374	3,293	3,230	3,179	3,137	2,936	2,864	2,826	2,803	2,787	2,748
10	4,965	4,103	3,708	3,478	3,326	3,217	3,135	3,072	3,020	2,978	2,774	2,700	2,661	2,637	2,621	2,580
11	4,844	3,982	3,587	3,357	3,204	3,095	3,012	2,948	2,896	2,854	2,646	2,570	2,531	2,507	2,490	2,448
12	4,747	3,885	3,490	3,259	3,106	2,996	2,913	2,849	2,796	2,753	2,544	2,466	2,426	2,401	2,384	2,341
13	4,667	3,806	3,411	3,179	3,025	2,915	2,832	2,767	2,714	2,671	2,459	2,380	2,339	2,314	2,297	2,252
14	4,600	3,739	3,344	3,112	2,958	2,848	2,764	2,699	2,646	2,602	2,388	2,308	2,266	2,241	2,223	2,178
15	4,543	3,682	3,287	3,056	2,901	2,790	2,707	2,641	2,588	2,544	2,328	2,247	2,204	2,178	2,160	2,114
16	4,494	3,634	3,239	3,007	2,852	2,741	2,657	2,591	2,538	2,494	2,276	2,194	2,151	2,124	2,106	2,059
17	4,451	3,592	3,197	2,965	2,810	2,699	2,614	2,548	2,494	2,450	2,230	2,148	2,104	2,077	2,058	2,011
18	4,414	3,555	3,160	2,928	2,773	2,661	2,577	2,510	2,456	2,412	2,191	2,107	2,063	2,035	2,017	1,968
19	4,381	3,522	3,127	2,895	2,740	2,628	2,544	2,477	2,423	2,378	2,155	2,071	2,026	1,999	1,980	1,930
20	4,351	3,493	3,098	2,866	2,711	2,599	2,514	2,447	2,393	2,348	2,124	2,039	1,994	1,966	1,946	1,896
30	4,171	3,316	2,922	2,690	2,534	2,421	2,334	2,266	2,211	2,165	1,932	1,841	1,792	1,761	1,740	1,683
60	4,001	3,150	2,758	2,525	2,368	2,254	2,167	2,097	2,040	1,993	1,748	1,649	1,594	1,559	1,534	1,467
120	3,920	3,072	2,680	2,447	2,290	2,175	2,087	2,016	1,959	1,910	1,659	1,554	1,495	1,457	1,429	1,352

Los números interiores corresponden a los valores de la variable F con n_1 grados de libertad del numerador y n_2 grados de libertad del denominador. Por ejemplo, $P(F_{10,20} \leq 2{,}348) = 0{,}95$.

TABLA VII: DISTRIBUCIÓN *F*

$$P(F_{n_1,n_2} \leq f_{n_1,n_2}) = 0{,}975$$

Grados de libertad del denominador (n_2)	Grados de libertad del numerador (n_1)															
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**20**	**30**	**40**	**50**	**60**	**120**
1	647,789	799,500	864,163	899,583	921,848	937,111	948,217	956,656	963,285	968,627	993,103	1001,414	1005,598	1008,117	1009,800	1014,020
2	38,506	39,000	39,165	39,248	39,298	39,331	39,355	39,373	39,387	39,398	39,448	39,465	39,473	39,478	39,481	39,490
3	17,443	16,044	15,439	15,101	14,885	14,735	14,624	14,540	14,473	14,419	14,167	14,081	14,037	14,010	13,992	13,947
4	12,218	10,649	9,979	9,605	9,364	9,197	9,074	8,980	8,905	8,844	8,560	8,461	8,411	8,381	8,360	8,309
5	10,007	8,434	7,764	7,388	7,146	6,978	6,853	6,757	6,681	6,619	6,329	6,227	6,175	6,144	6,123	6,069
6	8,813	7,260	6,599	6,227	5,988	5,820	5,695	5,600	5,523	5,461	5,168	5,065	5,012	4,980	4,959	4,904
7	8,073	6,542	5,890	5,523	5,285	5,119	4,995	4,899	4,823	4,761	4,467	4,362	4,309	4,276	4,254	4,199
8	7,571	6,059	5,416	5,053	4,817	4,652	4,529	4,433	4,357	4,295	3,999	3,894	3,840	3,807	3,784	3,728
9	7,209	5,715	5,078	4,718	4,484	4,320	4,197	4,102	4,026	3,964	3,667	3,560	3,505	3,472	3,449	3,392
10	6,937	5,456	4,826	4,468	4,236	4,072	3,950	3,855	3,779	3,717	3,419	3,311	3,255	3,221	3,198	3,140
11	6,724	5,256	4,630	4,275	4,044	3,881	3,759	3,664	3,588	3,526	3,226	3,118	3,061	3,027	3,004	2,944
12	6,554	5,096	4,474	4,121	3,891	3,728	3,607	3,512	3,436	3,374	3,073	2,963	2,906	2,871	2,848	2,787
13	6,414	4,965	4,347	3,996	3,767	3,604	3,483	3,388	3,312	3,250	2,948	2,837	2,780	2,744	2,720	2,659
14	6,298	4,857	4,242	3,892	3,663	3,501	3,380	3,285	3,209	3,147	2,844	2,732	2,674	2,638	2,614	2,552
15	6,200	4,765	4,153	3,804	3,576	3,415	3,293	3,199	3,123	3,060	2,756	2,644	2,585	2,549	2,524	2,461
16	6,115	4,687	4,077	3,729	3,502	3,341	3,219	3,125	3,049	2,986	2,681	2,568	2,509	2,472	2,447	2,383
17	6,042	4,619	4,011	3,665	3,438	3,277	3,156	3,061	2,985	2,922	2,616	2,502	2,442	2,405	2,380	2,315
18	5,978	4,560	3,954	3,608	3,382	3,221	3,100	3,005	2,929	2,866	2,559	2,445	2,384	2,347	2,321	2,256
19	5,922	4,508	3,903	3,559	3,333	3,172	3,051	2,956	2,880	2,817	2,509	2,394	2,333	2,295	2,270	2,203
20	5,871	4,461	3,859	3,515	3,289	3,128	3,007	2,913	2,837	2,774	2,464	2,349	2,287	2,249	2,223	2,156
30	5,568	4,182	3,589	3,250	3,026	2,867	2,746	2,651	2,575	2,511	2,195	2,074	2,009	1,968	1,940	1,866
60	5,286	3,925	3,343	3,008	2,786	2,627	2,507	2,412	2,334	2,270	1,944	1,815	1,744	1,699	1,667	1,581
120	5,152	3,805	3,227	2,894	2,674	2,515	2,395	2,299	2,222	2,157	1,825	1,690	1,614	1,565	1,530	1,433

Los números interiores corresponden a los valores de la variable F con n_1 grados de libertad del numerador y n_2 grados de libertad del denominador. Por ejemplo, $P(F_{10,20} \leq 2{,}774) = 0{,}975$.

TABLA VII: DISTRIBUCIÓN *F*

$$P(F_{n_1,n_2} \leq f_{n_1,n_2}) = 0{,}990$$

	Grados de libertad del numerador (n_1)															
Grados de libertad del denominador (n_2)	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**20**	**30**	**40**	**50**	**60**	**120**
1	4052,181	4999,500	5403,352	5624,583	5763,650	5858,986	5928,356	5981,070	6022,473	6055,847	6208,730	6260,649	6286,782	6302,517	6313,030	6339,391
2	98,503	99,000	99,166	99,249	99,299	99,333	99,356	99,374	99,388	99,399	99,449	99,466	99,474	99,479	99,482	99,491
3	34,116	30,817	29,457	28,710	28,237	27,911	27,672	27,489	27,345	27,229	26,690	26,505	26,411	26,354	26,316	26,221
4	21,198	18,000	16,694	15,977	15,522	15,207	14,976	14,799	14,659	14,546	14,020	13,838	13,745	13,690	13,652	13,558
5	16,258	13,274	12,060	11,392	10,967	10,672	10,456	10,289	10,158	10,051	9,553	9,379	9,291	9,238	9,202	9,112
6	13,745	10,925	9,780	9,148	8,746	8,466	8,260	8,102	7,976	7,874	7,396	7,229	7,143	7,091	7,057	6,969
7	12,246	9,547	8,451	7,847	7,460	7,191	6,993	6,840	6,719	6,620	6,155	5,992	5,908	5,858	5,824	5,737
8	11,259	8,649	7,591	7,006	6,632	6,371	6,178	6,029	5,911	5,814	5,359	5,198	5,116	5,065	5,032	4,946
9	10,561	8,022	6,992	6,422	6,057	5,802	5,613	5,467	5,351	5,257	4,808	4,649	4,567	4,517	4,483	4,398
10	10,044	7,559	6,552	5,994	5,636	5,386	5,200	5,057	4,942	4,849	4,405	4,247	4,165	4,115	4,082	3,996
11	9,646	7,206	6,217	5,668	5,316	5,069	4,886	4,744	4,632	4,539	4,099	3,941	3,860	3,810	3,776	3,690
12	9,330	6,927	5,953	5,412	5,064	4,821	4,640	4,499	4,388	4,296	3,858	3,701	3,619	3,569	3,535	3,449
13	9,074	6,701	5,739	5,205	4,862	4,620	4,441	4,302	4,191	4,100	3,665	3,507	3,425	3,375	3,341	3,255
14	8,862	6,515	5,564	5,035	4,695	4,456	4,278	4,140	4,030	3,939	3,505	3,348	3,266	3,215	3,181	3,094
15	8,683	6,359	5,417	4,893	4,556	4,318	4,142	4,004	3,895	3,805	3,372	3,214	3,132	3,081	3,047	2,959
16	8,531	6,226	5,292	4,773	4,437	4,202	4,026	3,890	3,780	3,691	3,259	3,101	3,018	2,967	2,933	2,845
17	8,400	6,112	5,185	4,669	4,336	4,102	3,927	3,791	3,682	3,593	3,162	3,003	2,920	2,869	2,835	2,746
18	8,285	6,013	5,092	4,579	4,248	4,015	3,841	3,705	3,597	3,508	3,077	2,919	2,835	2,784	2,749	2,660
19	8,185	5,926	5,010	4,500	4,171	3,939	3,765	3,631	3,523	3,434	3,003	2,844	2,761	2,709	2,674	2,584
20	8,096	5,849	4,938	4,431	4,103	3,871	3,699	3,564	3,457	3,368	2,938	2,778	2,695	2,643	2,608	2,517
30	7,562	5,390	4,510	4,018	3,699	3,473	3,304	3,173	3,067	2,979	2,549	2,386	2,299	2,245	2,208	2,111
60	7,077	4,977	4,126	3,649	3,339	3,119	2,953	2,823	2,718	2,632	2,198	2,028	1,936	1,877	1,836	1,726
120	6,851	4,787	3,949	3,480	3,174	2,956	2,792	2,663	2,559	2,472	2,035	1,860	1,763	1,700	1,656	1,533

Los números interiores corresponden a los valores de la variable F con n_1 grados de libertad del numerador y n_2 grados de libertad del denominador. Por ejemplo, $P(F_{10,20} \leq 3{,}368) = 0{,}990$.

TABLA VII: DISTRIBUCIÓN *F*

$$P(F_{n_1,n_2} \leq f_{n_1,n_2}) = 0{,}995$$

Grados de libertad del denominador (n_2)	Grados de libertad del numerador (n_1)															
	1	2	3	4	5	6	7	8	9	10	20	30	40	50	60	120
1	16210,723	19999,500	21614,741	22499,583	23055,798	23437,111	23714,566	23925,406	24091,004	24224,487	24835,971	25043,628	25148,153	25211,089	25253,137	25358,573
2	198,501	199,000	199,166	199,250	199,300	199,333	199,357	199,375	199,388	199,400	199,450	199,466	199,475	199,480	199,483	199,491
3	55,552	49,799	47,467	46,195	45,392	44,838	44,434	44,126	43,882	43,686	42,778	42,466	42,308	42,213	42,149	41,989
4	31,333	26,284	24,259	23,155	22,456	21,975	21,622	21,352	21,139	20,967	20,167	19,892	19,752	19,667	19,611	19,468
5	22,785	18,314	16,530	15,556	14,940	14,513	14,200	13,961	13,772	13,618	12,903	12,656	12,530	12,454	12,402	12,274
6	18,635	14,544	12,917	12,028	11,464	11,073	10,786	10,566	10,391	10,250	9,589	9,358	9,241	9,170	9,122	9,001
7	16,236	12,404	10,882	10,050	9,522	9,155	8,885	8,678	8,514	8,380	7,754	7,534	7,422	7,354	7,309	7,193
8	14,688	11,042	9,596	8,805	8,302	7,952	7,694	7,496	7,339	7,211	6,608	6,396	6,288	6,222	6,177	6,065
9	13,614	10,107	8,717	7,956	7,471	7,134	6,885	6,693	6,541	6,417	5,832	5,625	5,519	5,454	5,410	5,300
10	12,826	9,427	8,081	7,343	6,872	6,545	6,302	6,116	5,968	5,847	5,274	5,071	4,966	4,902	4,859	4,750
11	12,226	8,912	7,600	6,881	6,422	6,102	5,865	5,682	5,537	5,418	4,855	4,654	4,551	4,488	4,445	4,337
12	11,754	8,510	7,226	6,521	6,071	5,757	5,525	5,345	5,202	5,085	4,530	4,331	4,228	4,165	4,123	4,015
13	11,374	8,186	6,926	6,233	5,791	5,482	5,253	5,076	4,935	4,820	4,270	4,073	3,970	3,908	3,866	3,758
14	11,060	7,922	6,680	5,998	5,562	5,257	5,031	4,857	4,717	4,603	4,059	3,862	3,760	3,698	3,655	3,547
15	10,798	7,701	6,476	5,803	5,372	5,071	4,847	4,674	4,536	4,424	3,883	3,687	3,585	3,523	3,480	3,372
16	10,575	7,514	6,303	5,638	5,212	4,913	4,692	4,521	4,384	4,272	3,734	3,539	3,437	3,375	3,332	3,224
17	10,384	7,354	6,156	5,497	5,075	4,779	4,559	4,389	4,254	4,142	3,607	3,412	3,311	3,248	3,206	3,097
18	10,218	7,215	6,028	5,375	4,956	4,663	4,445	4,276	4,141	4,030	3,498	3,303	3,201	3,139	3,096	2,987
19	10,073	7,093	5,916	5,268	4,853	4,561	4,345	4,177	4,043	3,933	3,402	3,208	3,106	3,043	3,000	2,891
20	9,944	6,986	5,818	5,174	4,762	4,472	4,257	4,090	3,956	3,847	3,318	3,123	3,022	2,959	2,916	2,806
30	9,180	6,355	5,239	4,623	4,228	3,949	3,742	3,580	3,450	3,344	2,823	2,628	2,524	2,459	2,415	2,300
60	8,495	5,795	4,729	4,140	3,760	3,492	3,291	3,134	3,008	2,904	2,387	2,187	2,079	2,010	1,962	1,834
120	8,179	5,539	4,497	3,921	3,548	3,285	3,087	2,933	2,808	2,705	2,188	1,984	1,871	1,798	1,747	1,606

Los números interiores corresponden a los valores de la variable F con n_1 grados de libertad del numerador y n_2 grados de libertad del denominador. Por ejemplo, $P(F_{10,20} \leq 3{,}847) = 0{,}995$.

Esta obra ha sido concebida con un objetivo concreto: formar a los estudiantes de primer curso del Grado en Psicología en la UNED y a todos aquellos que se inician en la materia sin contar necesariamente con la ayuda de un profesor. Por este motivo, se presentan los conceptos fundamentales de manera breve y sencilla, utilizando ejemplos concretos aplicados –en la medida de lo posible– a la Psicología, prescindiendo de desarrollos matemáticos que no sean estrictamente necesarios. Los contenidos, presentados a nivel introductorio, coinciden con el programa de la asignatura, y se adaptan a la metodología de la enseñanza a distancia que permite el estudio independiente por parte del alumno. Tratarán, por tanto, sobre la organización de datos y su representación gráfica, los índices descriptivos, la correlación y regresión lineal, conceptos básicos de probabilidad, distribuciones discretas y continuas de probabilidad, y algunas nociones de muestreo y estimación. Aunque en el análisis de datos resulta imprescindible la utilización del ordenador, no se hace referencia a ningún software concreto ni a las posibilidades que ofrece la red. Estos aspectos serán considerados en el curso virtual de la asignatura.

El texto ha sido elaborado por el equipo docente de "Introducción al análisis de datos" de la UNED. Los miembros del equipo son psicólogos y todos cuentan con una amplia experiencia docente e investigadora en la utilización de modelos cuantitativos en las distintas áreas de la Psicología y la aplicación de nuevas tecnologías.

UNED Editorial

colección Grado

ISBN: 978-84-362-6042-7
01103
9 788436 260427

6201103GR01A01