

BEST.

PROBLEMAS DE BioESTADÍSTICA

Javier Llorca
Trinidad Dierssen Sotos
Inés Gómez Acebo
Jéssica Alonso Molero
Camilo Palazuelos Calderón
María Fernández Ortiz



Javier Llorca: Doctor en Medicina y licenciado en Ciencias Matemáticas. Catedrático de Medicina Preventiva y Salud Pública en la Universidad de Cantabria e investigador del Ciber de Epidemiología y Salud Pública (CIBERESP).

Trinidad Dierssen Sotos: Doctora en Medicina. Profesora Titular de Medicina Preventiva y Salud Pública en la Universidad de Cantabria, jefe de grupo de Epidemiología y Salud Pública del IDIVAL e investigadora del CIBERESP.

Inés Gómez Acebo: Doctora en Ciencias de la Salud. Profesora Titular de Medicina Preventiva y Salud Pública en la Universidad de Cantabria, investigadora del IDIVAL y del CIBERESP.

Jéssica Alonso Molero: Doctora en Biomedicina y Ciencias de la Salud. Profesora Ayudante Doctor de Medicina Preventiva y Salud Pública en la Universidad de Cantabria, investigadora del IDIVAL.

Camilo Palazuelos Calderón: Doctor en Ciencia y Tecnología. Investigador contratado en el IDIVAL.

María Fernández Ortiz: Ingeniera informática. Investigadora contratada en el IDIVAL. Confundadora y responsable SysAdmin/DevOps de Group4Layers®. Data manager del proyecto Cohorte Cantabria®.

BEST
PROBLEMAS DE BIOESTADÍSTICA



CONSEJO EDITORIAL

Dña. Silvia Tamayo Haya
Presidenta. Secretaria General,
Universidad de Cantabria

D. Vitor Abrantes
Facultad de Ingeniería,
Universidad de Oporto

D. Ramón Agüero Calvo
ETS de Ingenieros Industriales y
de Telecomunicación,
Universidad de Cantabria

D. Miguel Ángel Bringas Gutiérrez
Facultad de Ciencias Económicas y
Empresariales,
Universidad de Cantabria

D. Diego Ferreño Blanco
ETS de Ingenieros de Caminos, Canales y Puertos,
Universidad de Cantabria

Dña. Aurora Garrido Martín
Facultad de Filosofía y Letras,
Universidad de Cantabria

D. José Manuel Goñi Pérez
Modern Languages Department,
Aberystwyth University

D. Carlos Marichal Salinas
Centro de Estudios Históricos,
El Colegio de México

D. Salvador Moncada
Faculty of Biology, Medicine and Health,
The University of Manchester

D. Agustín Oterino Durán
Neurología (HUMV), investigador del IDIVAL

D. Luis Quindós Poncela
Radiología y Medicina Física,
Universidad de Cantabria

D. Marcelo Norberto Rougier
Historia Económica y Social Argentina, UBA y
CONICET (IIEP)

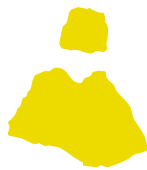
Dña. Claudia Sagastizábal
IMPA (Instituto Nacional de Matemática Pura e
Aplicada)

Dña. Belmar Gándara Sancho
Directora. Editorial de la
Universidad de Cantabria

BEST

PROBLEMAS DE BIOESTADÍSTICA

Javier Llorca
Trinidad Dierssen Sotos
Inés Gómez Acebo
Jéssica Alonso Molero
Camilo Palazuelos Calderón
María Fernández Ortiz



TEXTOS UNIVERSITARIOS, nº 31
CIENCIAS BIOMÉDICAS



Editorial
Universidad
Cantabria

Llorca Díaz, Javier, autor

BEST : problemas de BioESTadística / Javier Llorca, Trinidad Dierssen Sotos, Inés Gómez Acebo, Jéssica Alonso Molero, Camilo Palazuelos Calderón, María Fernández Ortiz. — Santander : Editorial de la Universidad de Cantabria, 2021.

450 páginas. — (Manuales. Textos universitarios. Ciencias biomédicas ; 31)

ISBN 978-84-17888-54-1

1. Biometría-Problemas, ejercicios, etc. 2. Biomatemáticas. I. Dierssen Sotos, Trinidad, autor. II. Gómez Acebo, Inés, autor. III. Alonso Molero, Jéssica, autor. IV. Palazuelos Calderón, Camilo, autor. V. Fernández Ortiz, María, autor.

57.087.1(076.2)

THEMA: MBNS, 4Z-ES-AF

Esta edición es propiedad de la EDITORIAL DE LA UNIVERSIDAD DE CANTABRIA, cualquier forma de reproducción, distribución, traducción, comunicación pública o transformación sólo puede ser realizada con la autorización de sus titulares, salvo excepción prevista por la ley. Diríjase a CEDRO (Centro Español de Derechos Reprográficos, www.cedro.org) si necesita fotocopiar o escanear algún fragmento de esta obra.

Ilustración de cubierta: Elena Gómez Menéndez

Digitalización: Manuel Ángel Ortiz Velasco [emeaov]

© Javier Llorca, Trinidad Dierssen Sotos, Inés Gómez Acebo, Jéssica Alonso Molero [Universidad de Cantabria], Camilo Palazuelos Calderón, María Fernández Ortiz [IDIVAL]

© Editorial de la Universidad de Cantabria
Avda. de los Castros, 52 - 39005 Santander. Cantabria (España)
Tlfno.-Fax: +34 942 201 087
www.editorial.unican.es

ISBN: 978-84-17888-54-1 (PDF)

DOI: <https://doi.org/10.22429/Euc2021.029>

Hecho en España-*Made in Spain*

Santander, 2021

Sumario

Prefacio	5
Material suplementario	9
Enunciados	11
1 Estadística descriptiva	13
1.1. Introducción	13
1.2. Descripción del estudio	13
1.3. Descripción de los datos	13
1.4. Problemas	15
2 Probabilidad condicionada e independencia de sucesos	25
2.1. Introducción	25
2.2. Notación	25
2.3. Descripción del estudio	26
2.4. Problemas	26
3 Distribuciones de probabilidad	35
3.1. Introducción	35
3.2. Notación	35
3.3. Fórmulas fundamentales	36
3.4. Problemas	37
4 Estimación de una media	55
4.1. Introducción	55
4.2. Fórmulas	56
4.3. Descripción de la base de datos	56
4.4. Descripción del estudio	57
4.5. Problemas	58

5	Comparación de dos medias	69
5.1.	Introducción	69
5.2.	Procedimiento	69
5.3.	Descripción de la base de datos	70
5.4.	Descripción del estudio	71
5.5.	Problemas - Sin emparejamiento	71
5.6.	Problemas - Con emparejamiento	78
6	Comparación de tres o más medias	83
6.1.	Introducción	83
6.2.	Procedimiento	83
6.3.	Descripción de la base de datos	84
6.4.	Descripción del estudio	85
6.5.	Problemas	85
7	Estimación de una proporción	93
7.1.	Introducción	93
7.2.	Fórmulas	93
7.3.	Descripción de la base de datos	93
7.4.	Descripción del estudio	94
7.5.	Problemas	94
8	Comparación de dos proporciones	101
8.1.	Introducción	101
8.2.	Notación	101
8.3.	Fórmulas	101
8.4.	Descripción de la base de datos	102
8.5.	Problemas	102
9	Tabla de contingencia	109
9.1.	Introducción	109
9.2.	Notación	109
9.3.	Fórmulas	110
9.4.	Descripción de la base de datos	110
9.5.	Problemas	111
10	Cálculo del tamaño de muestra	119
10.1.	Introducción	119
10.2.	Problemas	120
11	Regresión y correlación	129
11.1.	Introducción	129
11.2.	Descripción del estudio	129
11.3.	Notación	129
11.4.	Fórmulas	130

11.5. Descripción de la base de datos	130
11.6. Problemas	131
12 Regresión lineal múltiple	139
12.1. Introducción	139
12.2. Notación	139
12.3. Problemas	140
Soluciones	151
1 Estadística descriptiva	151
2 Probabilidad condicionada e independencia de sucesos	173
3 Distribuciones de probabilidad	189
4 Estimación de una media	233
5 Comparación de dos medias	271
6 Comparación de tres o más medias	311
7 Estimación de una proporción	347
8 Comparación de dos proporciones	361
9 Tabla de contingencia	381
10 Cálculo del tamaño de muestra	399
11 Regresión y correlación	417
12 Regresión lineal múltiple	437

Prefacio

La crisis provocada por la pandemia de COVID-19 nos va a obligar a modificar durante unos años la forma de enseñar en la universidad. La enseñanza cara a cara se va a reducir y está seriamente amenazada. Esto nos ha decidido a preparar un libro de problemas de bioestadística en el que las soluciones se presenten paso a paso, de forma que puedan ser seguidas por el alumno de forma autónoma.

La enseñanza de la bioestadística forma parte del cuerpo fundamental de la formación en ciencias de la salud, pero tiene una peculiaridad que la diferencia de la enseñanza de la bioquímica, la genética u otros campos: los conceptos y herramientas que se enseñan son muy elementales y permanecen estables durante muchas décadas, hasta el punto de que en la mayoría de los programas de bioestadística apenas aparece algún concepto posterior a 1950. Por poner los ejemplos más claros, el teorema de Bayes se publicó en 1763, la regresión lineal por mínimos cuadrados se aplicó por primera vez en 1809, la ji cuadrado fue desarrollada en 1900, la t de Student en 1908, el análisis de la varianza (ANOVA) en 1918. Solo si el temario llega hasta modelos de regresión no lineales se encuentra la regresión logística en 1943 y la regresión de Cox en 1972. Es decir, lo que enseñamos es muy básico, está muy lejos de ser puntero y es apenas el barniz mínimo necesario para poder entender los artículos de investigación más elementales. Porque ese es el problema fundamental en bioestadística elemental: la inmensa mayoría de los artículos científicos que vale la pena leer requieren conocimientos básicos de estadística para comprenderlos.

Hay muchos buenos libros de bioestadística. La mayoría de ellos desarrollan los mismos conceptos con enfoques ligeramente distintos. Lo que les diferencia fundamentalmente es el estilo. Algunos son más teóricos que otros, algunos hacen más hincapié en la realización de problemas o en el uso de paquetes estadísticos. Lo que no se encuentra con facilidad es un libro dedicado exclusivamente a hacer problemas. Eso es lo que presentamos aquí.

El libro es repetitivo. En algunos capítulos se propone una y otra vez el mismo tipo de problema con diferentes datos. Puede que la primera vez un alumno quiera ir directamente a la solución, pero hacer lo mismo en los siguientes problemas no le aportará nada: tiene que enfrentarse al papel en blanco para intentar encontrar la solución por sí mismo.

Nuestro enfoque es que saberse los comandos de un paquete estadístico no es lo mismo que saber estadística por el mismo motivo que aprender el funcionamiento de un ecocardiógrafo no le convierte a uno en cardiólogo. Por eso creemos que los problemas hay que hacerlos al menos una vez con calculadora para comprender bien qué es lo que uno se trae entre manos. Pero a lo largo del libro aportamos también soluciones con Excel o con Stata. La mayor parte de los capítulos parten de una base de datos en Excel accesible mediante los enlaces indicados en el texto. En algunos problemas, el primer paso es usar esa base de datos para generar una tabla a partir de la cual se puede realizar el resto del problema con calculadora. El alumno que no esté interesado en el manejo de Excel o Stata, puede saltarse ese paso, consultar directamente la tabla en las soluciones y seguir con el problema “a mano”.

No queremos que nadie se llame a engaño: este no es un libro sobre cómo resolver problemas de bioestadística con Excel o con Stata. El eje del libro es la solución con calculadora. Las soluciones con Excel y Stata se aportan como complemento. El mensaje es: peléese usted con las fórmulas hasta que las entienda. Solo después de eso busque un programa con el que hacer las cuentas.

Las diferentes soluciones se identifican con un código de colores, rojo para la solución con calculadora, verde para Excel y azul para Stata, de la siguiente forma:

Solución con calculadora

En este recuadro se presentan las fórmulas y los resultados del problema con calculadora.

Solución con Excel

En este recuadro se presentan las fórmulas y los resultados del problema con Excel.

Solución con Stata

En este recuadro se presentan las fórmulas y los resultados del problema con Stata.

El libro está estructurado en doce capítulos. El primero se dedica a estadística descriptiva. El 2º y el 3º a las distribuciones de probabilidad (normal, binomial y de Poisson) y a conceptos elementales de probabilidad (probabilidad condicionada, sucesos independientes y teorema de Bayes). En los capítulos 4º a 6º se trabaja con la estimación de medias; los capítulos 7º y 8º se dedican al estudio de proporciones y el 9º al tema conexo de las tablas de contingencia. El cálculo del tamaño muestral se trata en el capítulo 10º. La relación entre dos variables continuas (correlación y regresión) se aborda en el capítulo 11º y la regresión lineal múltiple en el 12º. De momento, hemos excluido del libro los test no paramétricos y las regresiones no lineales.

La terminología estadística no siempre es uniforme. Hemos procurado ser coherentes a lo largo del libro, pero esta edición puede pecar de precipitada y probablemente tenga todavía cierto grado de incoherencia.

Otro aspecto de la terminología es el uso de los separadores de decimales y de miles. La Ortografía de la lengua española, de las Academias de la Lengua, -de acuerdo con el Sistema Internacional de Unidades- recomienda el uso del punto como separador de decimales, aunque sigue considerando válida la coma, y no admite el uso ni del punto ni de la coma como separador de miles. En este libro hemos seguido la recomendación y utilizamos siempre el punto como separador de decimales.

Este libro se elaboró durante el verano de 2020 y lo publicamos con **licencia Creative Commons**. Puede ser reproducido libremente sin finalidad comercial, reconociendo la autoría e indicando si se han realizado cambios. Puede encontrar más información sobre el significado de esta licencia en <https://creativecommons.org/licenses/by-nc/3.0/es/>. Agradecemos a la Editorial de la Universidad de Cantabria que lo haya acogido en el verano de 2021.

El resto de los autores queremos agradecer a María Fernández Ortiz su trabajo en la maquetación, edición y diseño del libro.

Santander, septiembre de 2021
Los autores

Material suplementario

- Bases de datos en Excel para realizar los problemas.
- Tablas estadísticas
- Tutoriales sobre el manejo de datos en Excel:
 1. Como instalar el analizador de datos
 2. Como aplicar filtros
 3. Autorellenar celdas (doble click mágico)
 4. Etiquetar variable completa
 5. Etiquetar variable parcialmente
 6. Administrador de nombres (de etiquetas)
 7. Formato de celdas
 8. Inmovilizar fila superior
 9. Trabajar con pestañas
 10. Tabla dinámica de una variable

Puede acceder al material suplementario en el Moodle de las asignaturas Bioestadística y Uso de Software Científico (Grado de Medicina, Universidad de Cantabria) y Bioestadística (Grado de Ciencias Biomédicas, Universidad de Cantabria). Si no tiene acceso a esos dispositivos, puede acceder mediante el siguiente link o puede solicitarlas por correo electrónico a mpsp.unican@gmail.com.

Enunciados

1 Estadística descriptiva

1.1. Introducción

La estadística descriptiva es la rama de la estadística que se encarga del análisis de las características de un conjunto de datos dado. El objetivo principal de la estadística descriptiva es la recolección, presentación y caracterización de los datos para poder llevar a cabo una correcta descripción de la muestra en estudio.

1.2. Descripción del estudio

Un hospital está repartido en dos edificios A y B. En el edificio A se están produciendo casos de una enfermedad inespecífica (conjuntivitis, dolor de garganta, rinitis, ...) desde hace tres años, lo que se achaca a las condiciones de deterioro del edificio, que deben producir algún tipo de contaminación ambiental.

La inspección de trabajo y el comité de empresa plantean que en el edificio A no deberían trabajar las mujeres embarazadas para no exponerse al agente contaminante.

Para aclarar si hay algún riesgo para el embarazo, se lleva a cabo un estudio de cohortes retrospectivo. En los 543 embarazos producidos en cualquiera de los dos edificios en los años 2017-2019, se obtiene información sobre antecedentes de la madre, desarrollo y control del embarazo, condiciones del parto y salud del recién nacido.

1.3. Descripción de los datos

La siguiente tabla presenta la descripción de todas las variables incluidas en la hoja **estudioEmbarazadas** del fichero **datosProblemasBest.xlsx**:

Variable	Descripción	Codificación
Id	Número de identificación del embarazo	Numérica
Building	Edificio en el que trabaja la embarazada	1: A 2: B
AgeMother	Edad de la madre en el momento del parto	Numérica (años cumplidos)
DateLM	Fecha de la última menstruación	Fecha
DateBirth	Fecha del parto	Fecha
DuraGest	Duración de la gestación	Numérica
DayFirst	Día del embarazo en que se produjo la primera consulta obstétrica	Numérica
Npregnanc	Número de embarazos previos	Numérica
Smokes	Tabaquismo durante el embarazo	0: No 1: Sí 2: Exfumadora
WeighNB	Peso del recién nacido	Numérica (gramos)
Hb	Menor nivel de hemoglobina en sangre durante el embarazo	Numérica (mg/dL)

Tabla 1.1: Descripción de las variables de estudio



1.4



PROBLEMAS

PROBLEMA 1.1

¿Cuál es la media, mediana y desviación típica de la duración de la gestación para las mujeres de esta cohorte? Para el cálculo a mano emplee los datos de las cinco primeras mujeres de la base de datos, que se presentan en la siguiente tabla.

Id	DuraGest
100228	246
100640	259
100706	236
101802	278
102373	270

Tabla 1.2: Datos de 5 mujeres de la cohorte

PROBLEMA 1.2

Se desea conocer la media, la desviación estándar y la mediana de la edad a la que las mujeres se quedaron embarazadas en cada uno de los edificios (A y B). Para el cálculo a mano emplee los datos de las diez primeras mujeres de la base de datos, que se presentan en la siguiente tabla.

Id	Building	AgeMother
100228	1	28
100640	2	28
100706	2	31
101802	2	40
102373	2	28
103044	1	44
105117	2	40
107040	1	28
107537	1	29
107572	1	42

Tabla 1.3: Datos de 10 mujeres de la cohorte.

PROBLEMA 1.3

¿Qué porcentaje de recién nacidos han sido prematuros? Prematuro se define como duración del embarazo menor a 37 semanas. Para el cálculo a mano emplee los datos de las cinco primeras mujeres de la base de datos, que se presentan en la siguiente tabla.

Id	DuraGest (días)
100228	246
100640	259
100706	236
101802	278
102373	270

Tabla 1.4: Datos de 5 mujeres de la cohorte

PROBLEMA 1.4

Se desea conocer la duración media, la desviación estándar y la mediana del tiempo de gestación en semanas para cada uno de los edificios y en general. Para el cálculo a mano emplee los datos de las diez primeras mujeres de la base de datos, que se presentan en la siguiente tabla.

Id	Building	DuraGest_semana
100228	1	35.14
100640	2	37.00
100706	2	33.71
101802	2	39.71
102373	2	38.57
103044	1	32.14
105117	2	39.29
107040	1	35.14
107537	1	33.86
107572	1	35.86

Tabla 1.5: Datos de 10 mujeres de la cohorte.

PROBLEMA 1.5

Se desea realizar un análisis descriptivo completo sobre los datos de los niveles de hemoglobina de todas las gestantes participantes en la cohorte. Para el cálculo a mano emplee los datos de las diez primeras mujeres de la base de datos, que se presentan en la siguiente tabla.

Id	Building	Hb
100228	1	11.3
100640	2	10.8
100706	2	12.9
101802	2	12.0
102373	2	10.4
103044	1	11.6
105117	2	13.2
107040	1	11.3
107537	1	12.3
107572	1	10.8

Tabla 1.6: Datos de 10 mujeres de la cohorte.

PROBLEMA 1.6

Se requiere conocer qué porcentaje de niños ha nacido con bajo peso ($\leq 2500\text{g}$), con peso normal ($2500\text{-}4000\text{ g}$) o con sobrepeso ($\geq 4000\text{g}$) en esta cohorte. Para el cálculo a mano emplee los datos de las mujeres que aparecen en la siguiente tabla.

Id	WeighNB	Peso_cat
100706	1540	Bajo
101802	3280	Normal
105117	3200	Normal
109690	3580	Normal
109727	4160	Alto
111457	2670	Normal
116331	3890	Normal
118315	3240	Normal
123125	2760	Normal
125242	3460	Normal

Tabla 1.7: Datos de 10 mujeres de la cohorte.

PROBLEMA 1.7

¿Qué proporción de las mujeres tiene una edad igual o menor a 34 años en la cohorte? ¿Y en cada uno de los edificios? ¿Y cuántas son mayores de 39 años? Para el cálculo a mano emplee los datos de las diez primeras mujeres de la base de datos. Expresé los resultados en porcentaje.

Id	Building	AgeMother
100228	1	28
100640	2	28
100706	2	31
101802	2	40
102373	2	28
103044	1	44
105117	2	40
107040	1	28
107537	1	29
107572	1	42

Tabla 1.8: Datos de 10 mujeres de la cohorte.

PROBLEMA 1.8

¿Qué proporción de mujeres padeció anemia durante el embarazo en el edificio A? ¿y en el B? ¿Qué media y desviación estándar de los niveles de hemoglobina presentaban las mujeres con anemia y sin anemia en cada uno de los edificios? Para el cálculo a mano emplee los datos de las mujeres que aparecen en la siguiente tabla. Nota: anemia se define como niveles de hemoglobina menores a 11 mg/dL. Expresé las proporciones en porcentaje.

Id	Building	Hb	Anemia
111428	1	11.7	No
111436	1	11.7	No
117651	1	10.4	Si
118661	1	10.3	Si
121262	1	10.4	Si
159658	2	13.3	No
159801	2	12.6	No
160336	2	10.6	Si
166116	2	10.7	Si
166256	2	13.6	No

Tabla 1.9: Datos de 10 mujeres de la cohorte

Probabilidad condicionada e independencia de sucesos

2.1 Introducción

En este capítulo se practicará el cálculo de:

- La probabilidad de un suceso.
- La probabilidad conjunta de dos sucesos.
- La probabilidad condicionada.
- La probabilidad esperada en caso de independencia de sucesos.
- Uso del teorema de Bayes para estimar una probabilidad condicionada.

Los problemas se realizarán sobre un estudio que relaciona el mesotelioma de pulmón con varios factores (la exposición profesional al asbesto y dos variantes genéticas).

2.2 Notación

Probabilidad de un suceso: $P(A)$. Por ejemplo, la probabilidad de que un paciente tenga mesotelioma se denotará como $P(\text{mesotelioma})$.

Probabilidad conjunta de dos sucesos: $P(A \cap B)$. Por ejemplo la probabilidad de que un paciente tenga mesotelioma y también exposición al asbesto se denotará como $P(\text{mesotelioma} \cap \text{exposición al asbesto})$.

Probabilidad de un suceso A condicionada a otro suceso B: $P(A | B)$. Por ejemplo, la probabilidad de que un paciente expuesto al asbesto (=la condición) tenga mesotelioma se denotará como $P(\text{mesotelioma} | \text{exposición al asbesto})$.

2.3 Descripción del estudio

Objetivo: Evaluar si la exposición al asbesto y dos polimorfismos genéticos se asocian a mesotelioma de pleura.

Ámbito: Estudio multicéntrico en una serie de empresas que trabajan con asbesto.

Criterios de inclusión:

- **Casos:** Todos los trabajadores de estas empresas a los que se ha diagnosticado un mesotelioma de pleura confirmado por anatomía patológica. La información del diagnóstico se extrajo de los registros médicos de las empresas, que se consideran exhaustivos. Un comité externo valoraba la aplicación de los criterios en anatomía patológica.
- **Controles:** Trabajadores de las mismas empresas sin diagnóstico de mesotelioma de pleura. La selección se realizó seleccionando los posibles controles al azar y emparejándolos por edad y sexo. El emparejamiento se hizo por frecuencia; es decir, se estima la distribución por edades de los casos y se seleccionan controles con la misma distribución; análogamente se procede con el sexo.

Recogida de información: La fecha reclutamiento de los casos es la del diagnóstico. De los registros de la empresa se obtuvo información sobre el tiempo de exposición a asbesto. De cada paciente se obtuvo una muestra de sangre o saliva para determinaciones genéticas. En el polimorfismo genético 1 (snp1), los dos alelos posibles son C y T. En el polimorfismo genético 2 (snp2), los alelos posibles son G y C. Como el material genético está duplicado, de cada polimorfismo hay dos copias por lo que pueden producirse tres genotipos según el número de mutaciones presentes (por ejemplo, en el snp1 puede haber 0, 1 o 2 copias del alelo T). Los dos polimorfismos se encuentran en distinto cromosoma.



2.4



PROBLEMAS

PROBLEMA 2.1

	Mesotelioma	No mesotelioma	Total
Exposición al asbesto	226	189	415
No exposición al asbesto	191	238	429
Total	417	427	844

Tabla 2.1: Relación entre exposición al asbesto y mesotelioma

Pregunta 2.1. Probabilidad de estar expuesto al asbesto.

Pregunta 2.2. Probabilidad de tener mesotelioma.

Pregunta 2.3. Probabilidad conjunta de tener mesotelioma y estar expuesto al asbesto.

Pregunta 2.4. Probabilidad conjunta de no tener mesotelioma y no estar expuesto al asbesto.

Pregunta 2.5. Probabilidad de estar expuesto al asbesto condicionada a tener mesotelioma.

Pregunta 2.6. Probabilidad de no estar expuesto al asbesto condicionada a tener mesotelioma.

Pregunta 2.7. Probabilidad de tener mesotelioma condicionada a estar expuesto al asbesto.

Pregunta 2.8. Probabilidad de no tener mesotelioma condicionada a no estar expuesto al asbesto.

PROBLEMA 2.2

Genotipo del snp1	Mesotelioma	No mesotelioma	Total
TT	180	119	299
TC	189	223	412
CC	48	85	133
Total	417	427	844

Tabla 2.2: Relación entre genotipos del snp1 y mesotelioma

Pregunta 2.9. Probabilidad de tener cada genotipo.

Pregunta 2.10. Probabilidad de tener cada genotipo condicionada a tener mesotelioma.

Pregunta 2.11. Probabilidad de tener cada genotipo condicionada a no tener mesotelioma.

PROBLEMA 2.3

Genotipo del snp2	Mesotelioma	No mesotelioma	Total
CC	56	46	102
CG	223	197	420
GG	138	184	322
Total	417	427	844

Tabla 2.3: Relación entre genotipos del snp2 y mesotelioma

Pregunta 2.12. Probabilidad de cada genotipo.

Pregunta 2.13. Probabilidad de cada genotipo condicionada a tener mesotelioma.

Pregunta 2.14. Probabilidad de cada genotipo condicionada a no tener mesotelioma.

PROBLEMA 2.4

Se reproduce a continuación la Tabla 2.2, pero indicando solo los totales de cada fila y cada columna.

Genotipo del snp1	Mesotelioma	No mesotelioma	Total
TT			299
TC			412
CC			133
Total	417	427	844

Tabla 2.4: Relación entre genotipos del snp1 y mesotelioma

Pregunta 2.15. Calcule las siguientes probabilidades conjuntas asumiendo que el genotipo y la presencia o no de mesotelioma son independientes:

- $P(TT \cap \text{mesotelioma})$
- $P(TC \cap \text{mesotelioma})$
- $P(CC \cap \text{mesotelioma})$
- $P(TT \cap \text{no mesotelioma})$
- $P(TC \cap \text{no mesotelioma})$
- $P(CC \cap \text{no mesotelioma})$

Pregunta 2.16. Como consecuencia, complete la tabla anterior con el número de pacientes que habría en cada casilla si hubiera independencia.

PROBLEMA 2.5

El mesotelioma es una enfermedad rara. En la base de datos que estamos usando en este capítulo se ha hecho una selección de pacientes especial (lo que se conoce como un estudio de casos y controles) de forma que el porcentaje de pacientes con mesotelioma es anormalmente alta (casi el 50%). Esta forma de seleccionar hace que las siguientes probabilidades no tengan sentido:

$$P(\text{mesotelioma} | TT)$$

$$P(\text{mesotelioma} | TC)$$

$$P(\text{mesotelioma} | CC)$$

$$P(\text{no mesotelioma} | TT)$$

$$P(\text{no mesotelioma} | TC)$$

$$P(\text{no mesotelioma} | CC)$$

Es decir, por la forma especial de obtener la muestra, las probabilidades de tener o de no tener mesotelioma no se pueden calcular directamente.

En cambio, sí es correcto calcular:

$$P(TT | \text{mesotelioma})$$

$$P(TC | \text{mesotelioma})$$

$$P(CC | \text{mesotelioma})$$

$$P(TT | \text{no mesotelioma})$$

$$P(TC | \text{no mesotelioma})$$

$$P(CC | \text{no mesotelioma})$$

A partir de éstas y conociendo la probabilidad de mesotelioma es posible aplicar el **teorema de Bayes**.

Pregunta 2.17. En una población en la que 1 de cada 1000 personas tienen mesotelioma [es decir: $P(\text{mesotelioma}) = 0.001$], calcule las siguientes probabilidades usando el teorema de Bayes.

- $P(\text{mesotelioma} | TT)$
- $P(\text{mesotelioma} | TC)$
- $P(\text{mesotelioma} | CC)$
- $P(\text{no mesotelioma} | TT)$
- $P(\text{no mesotelioma} | TC)$
- $P(\text{no mesotelioma} | CC)$

PROBLEMA 2.6

Este problema no se hace con los datos que hemos usado en todo el capítulo.

Durante la pandemia de COVID-19 en la primavera de 2020, se ha discutido mucho sobre la validez de los test rápidos. Vamos a analizarla.

Se habló de test que tenían sensibilidad = 0.70 y especificidad = 0.99. Antes de seguir, hay que definir estos dos conceptos en forma de probabilidad condicionada. La sensibilidad se define como:

$$\text{Sensibilidad} = P(\text{obtener un resultado positivo} \mid \text{haber pasado COVID-19})$$

Para simplificar, lo escribiremos así:

$$S = P(+ \mid \text{COVID-19})$$

Análogamente, la especificidad se define como:

$$\text{Especificidad} = P(\text{obtener un resultado negativo} \mid \text{no haber pasado COVID-19})$$

Que simplificado se escribirá:

$$E = P(- \mid \text{no COVID-19})$$

Pregunta 2.18. En España, se informó el 13 de mayo de que el 5 % de los habitantes había pasado COVID-19 (es decir: $P(\text{COVID-19}) = 0.05$). Aplicamos un test rápido a una persona que vive en España y se obtiene un resultado positivo, ¿cuál es la probabilidad de que haya pasado COVID-19?

Pregunta 2.19. Por las mismas fechas, se informó de que en Nueva York $P(\text{COVID-19}) = 0.20$. Si una persona que vive en Nueva York da positivo, ¿cuál es la probabilidad de que haya pasado COVID-19?

Pregunta 2.20. Si un residente en España da negativo, ¿cuál es la probabilidad de que no haya pasado COVID-19?

Pregunta 2.21. Si un residente en Nueva York da negativo, ¿cuál es la probabilidad de que no haya pasado COVID-19?

Distribuciones de probabilidad

3.1 Introducción

En este capítulo no se utilizará ninguna base de datos. Se harán problemas sobre:

- Manejo de la tabla de la distribución normal.
- Uso de la tabla de la distribución normal para conocer cómo está distribuida una población.
- Cálculos sobre distribuciones binomiales.
- Cálculos sobre distribuciones de Poisson.
- Cálculos sobre distribuciones binomiales usando la aproximación de Poisson.
- Cálculos sobre distribuciones binomiales y de Poisson usando la aproximación normal.

3.2 Notación

Distribución normal de media (μ) y varianza (σ^2): $N(\mu, \sigma^2)$

Distribución normal estandarizada: $N(0, 1)$

Distribución binomial de parámetros n y p : $B(n, p)$, donde n es el número de repeticiones (por ejemplo, el número de sujetos) y p la probabilidad de que ocurra el evento de interés.

Distribución de Poisson de parámetro μ : $P(\mu)$, donde μ es la media de eventos esperados en un periodo fijo de tiempo.

3.3 Fórmulas fundamentales

- Distribución de densidad de la distribución $N(\mu, \sigma^2)$.

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

Nota: esta fórmula es fundamental, pero no de utilidad práctica para los problemas de este libro.

- Fórmula para estandarizar una distribución normal (es decir, para convertir $N(\mu, \sigma^2)$ en $N(0, 1)$).

$$z = \frac{x - \mu}{\sigma}$$

- En una distribución $B(n, p)$: probabilidad de que ocurran *exactamente* k eventos.

$$P(x = k|n, p) = \binom{n}{k} p^k (1-p)^{n-k}$$

- En una distribución $B(n, p)$: probabilidad de que ocurran k eventos o *menos*.

$$P(x \leq k|n, p) = \sum_{i=0}^k \binom{n}{i} p^i (1-p)^{n-i}$$

- En una distribución $B(n, p)$: probabilidad de que ocurran *más* de k eventos

$$P(x > k|n, p) = \sum_{i=k+1}^n \binom{n}{i} p^i (1-p)^{n-i} = 1 - P(x \leq k|n, p)$$

- En una distribución $P(\mu)$: probabilidad de que ocurran *exactamente* k eventos.

$$P(x = k|\mu) = \frac{\mu^k e^{-\mu}}{k!}$$

- En una distribución $P(\mu)$: probabilidad de que ocurran k eventos o *menos*.

$$P(x \leq k|\mu) = \sum_{i=0}^k \frac{\mu^i e^{-\mu}}{i!}$$

- En una distribución $P(\mu)$: probabilidad de que ocurran *más* de k eventos.

$$P(x > k|\mu) = 1 - P(x \leq k|\mu) = 1 - \sum_{i=0}^k \frac{\mu^i e^{-\mu}}{i!}$$

Media, varianza y desviación típica de las distribuciones:

Distribución	Media	Varianza	Desviación típica
$N(\mu, \sigma^2)$	μ	σ^2	σ
$B(n, p)$	np	$np(1 - p)$	$\sqrt{np(1 - p)}$
$P(\mu)$	μ	μ	$\sqrt{\mu}$

3.4

PROBLEMAS

PROBLEMA 3.1

El colesterol HDL se suele presentar en mg/dL. Supongamos que en la población española tiene una distribución $N(38, 25)$.

Responda a las siguientes preguntas:

Pregunta 3.1. ¿Qué porcentaje de la población tiene más de 38 mg/dL?

Pregunta 3.2. ¿Qué porcentaje de la población tiene más de 43 mg/dL?

Pregunta 3.3. ¿Qué porcentaje de la población tiene entre 33 y 43 mg/dL?

Pregunta 3.4. ¿Qué porcentaje de la población tiene más de 30 mg/dL?

Pregunta 3.5. ¿Qué porcentaje de la población tiene entre 30 y 35 mg/dL?

PROBLEMA 3.2

Con la misma distribución de la pregunta anterior, responda a las siguientes preguntas:

Pregunta 3.6. ¿Qué valor HDL tiene el percentil 70 de la población?

Pregunta 3.7. ¿Qué valor HDL tiene el percentil 40 de la población?

Pregunta 3.8. Se decide considerar HDL bajo cuando esté por debajo del percentil 10. ¿Por debajo de que valor de HDL se considerará que es bajo?

PROBLEMA 3.3

Se realiza la determinación de osteoprotegerina en suero en 200 sujetos y se encuentra que tiene distribución $N(0.17 \text{ ng/mL}, 0.0009)$.

Responda a las siguientes preguntas:

Pregunta 3.9. ¿Qué porcentaje de la población tiene más de 0.23 ng/mL?

Pregunta 3.10. ¿Qué porcentaje de la población tiene entre 0.17 y 0.23 ng/mL?

Pregunta 3.11. ¿Qué porcentaje de la población tiene entre 0.12 y 0.20 ng/mL?

Pregunta 3.12. ¿Qué porcentaje de la población tiene menos de 0.14 ng/mL?

Pregunta 3.13. ¿Qué porcentaje de la población tiene entre 0.14 y 0.16 ng/mL?

PROBLEMA 3.4

Con la misma distribución de la pregunta anterior, responda a las siguientes preguntas:

Pregunta 3.14. ¿Qué valor de osteoprotegerina deja por debajo al 45 % de la población?

Pregunta 3.15. ¿Qué valor de osteoprotegerina deja por encima al 30 % de la población?

Pregunta 3.16. ¿Qué valor de osteoprotegerina deja por encima al 5 % de la población?

Pregunta 3.17. ¿Qué valores de osteoprotegerina dejan en el centro el 80 % de la población y un 10 % en cada cola?

PROBLEMA 3.5

En un paso posterior a los experimentos de Mendel, vamos a seguir trabajando con plantas de guisantes amarillos o verdes y lisos o rugosos. Estas dos características (color y lisura) dependen cada una de un gen. El gen color puede tener dos alelos: A -de amarillo- y v -de verde. Representamos el A en mayúsculas porque es dominante. El gen lisura puede tener los alelos L -de liso- y r -de rugoso. Los genes color y lisura se transmiten de forma independiente. En cada planta, el gen color y el gen lisura tienen dos alelos; esto hace que para color pueda haber tres genotipos (AA , Av , vv) y para lisura otros tres (LL , Lr , rr).

Tenemos una población de 100 plantas de guisantes que se han obtenido por sucesivos cruzamientos al azar. El número de cruzamientos se considera suficientemente alto como para que la distribución de los alelos sea estable.

Las proporciones que se encuentran de los alelos son:

$$P(A) = 0.6$$

$$P(v) = 0.4$$

$$P(L) = 0.8$$

$$P(r) = 0.2$$

Con esta información responda a las siguientes preguntas:

Pregunta 3.18. ¿Qué proporción de plantas tendrá cada genotipo del gen color?

- $P(AA) =$
- $P(Av) =$
- $P(vv) =$

Pregunta 3.19. ¿Qué proporción de plantas tendrá cada genotipo del gen lisura?

- $P(LL) =$
- $P(Lr) =$
- $P(rr) =$

Pregunta 3.20. Sabiendo que los alelos amarillo y liso son dominantes y que los dos genes se transmiten de forma independiente, ¿qué proporción de plantas tendrá cada fenotipo?

- $P(\text{plantas amarillas y lisas}) =$
- $P(\text{plantas amarillas y rugosas}) =$
- $P(\text{plantas verdes y lisas}) =$
- $P(\text{plantas verdes y rugosas}) =$

PROBLEMA 3.6

En este problema queremos generalizar los resultados obtenidos en las preguntas 3.18 y 3.19.

Un gen puede tener los alelos A y B. Se dispone de una población en la que los cruza-
mientos se producen al azar y hace tiempo que no se incorporan sujetos nuevos. En esta
población, la frecuencia del alelo A es $P(A) = p$ y, por lo tanto, la frecuencia del alelo B es
 $P(B) = 1 - p$.

Pregunta 3.21. ¿Cuál será la frecuencia de cada genotipo?

- $P(AA) =$
- $P(AB) =$
- $P(BB) =$

PROBLEMA 3.7

La fibrosis quística es una enfermedad genética autosómica recesiva; es decir, solo se tiene la enfermedad si las dos copias del gen tienen el alelo mutado. Se debe a mutaciones en el gen CFTR, que es un gen grande (con unas 180000 pares de bases). Se han descrito cerca de 2000 mutaciones en CFTR. En este problema simplificaremos mucho y asumiremos que solo hay una mutación. Llamemos A al alelo normal y a al alelo mutado. En España, la frecuencia de la mutación es de 1 de cada 30 personas.

Pregunta 3.22. ¿Cuál será la frecuencia de fibrosis quística en los recién nacidos?

Pregunta 3.23. ¿Cuál será la frecuencia de portadores en los recién nacidos? (Llamamos portador al que solo tiene una copia del alelo mutado).

PROBLEMA 3.8

Tras un trasplante de pulmón, se produce la enfermedad injerto contra huésped (EICH) en el 20% de los pacientes. En el hospital en el que usted trabaja, se pone en marcha un nuevo programa de trasplante pulmonar. Tras realizar 10 trasplantes, se han producido 4 casos de EICH. La pregunta clave es si algo falla en su programa y por eso están teniendo demasiados casos de EICH. Pero antes hay que saber si 4 casos de 10 son realmente demasiados o si podrían haber ocurrido al azar.

Pregunta 3.24. Si su probabilidad de tener un caso de EICH fuera realmente del 20%, ¿cuál sería la probabilidad de que en 10 pacientes hubieran ocurrido 4 casos?

Pregunta 3.25. ¿Y cuál sería la probabilidad de que en 10 pacientes hubieran ocurrido 4 casos o más?

PROBLEMA 3.9

El 7 % de los pacientes ingresados por COVID-19 han fallecido. Usted descubre que de 15 pacientes portadores de la mutación ApoE- ϵ 4 han muerto 3 (un 20 %). Le preocupa saber si los portadores de esa mutación tienen más riesgo de morir por COVID-19.

Pregunta 3.26. Si la mortalidad es 7 %, ¿cuál es la probabilidad de que de 15 pacientes mueran 3?

Pregunta 3.27. ¿Y cuál es la probabilidad de que mueran 3 o más?

PROBLEMA 3.10

Un laboratorio farmacéutico lanza un nuevo medicamento para tratar el COVID-19. Según sus estudios, el laboratorio afirma que la mortalidad bajará al 4%. Usted quiere comprobar este resultado y para ello selecciona a 7 pacientes y los trata con el nuevo medicamento. Si la mortalidad es del 4%.

Pregunta 3.28. ¿Cuál es la probabilidad de que no muera ninguno?

Pregunta 3.29. ¿Cuál es la probabilidad de que muera 1?

Pregunta 3.30. ¿Cuál es la probabilidad de que mueran 2 o más?

PROBLEMA 3.11

Con el mismo medicamento del problema anterior, se lleva a cabo un estudio con 2000 pacientes.

Pregunta 3.31. ¿Cuál es la media y la desviación típica del número de fallecidos?

Pregunta 3.32. ¿Cuál es la probabilidad de que ocurran menos de 60 fallecimientos?

Pregunta 3.33. ¿Cuál es la probabilidad de que ocurran menos de 80 fallecimientos?

Pregunta 3.34. ¿Cuál es la probabilidad de que ocurran 90 fallecimientos o más?

PROBLEMA 3.12

En una población de 125000 habitantes, la media anual de muertes por leucemia mieloide aguda (LMA) es 5. En 2021 ocurren 8 muertes por LMA, lo que produce una gran alarma social. El responsable de sanidad tiene que hacer un estudio para saber si 8 muertes en un año es un fenómeno normal cuando la media es 5 o es un fenómeno tan infrecuente que debería hacerse una investigación para saber qué está ocurriendo. Para ello, tiene que responder a estas dos preguntas:

Pregunta 3.35. Si la media es 5 muertes al año, ¿cuál es la probabilidad de que ocurran 8 muertes en un año?

Pregunta 3.36. Si la media es 5 muertes al año, ¿cuál es la probabilidad de que ocurran 8 muertes o más en un año?

PROBLEMA 3.13

En una provincia de tamaño medio se produce cada año una media de 9 ingresos por enfermedad neumocócica invasiva (ENI). Para evitarlo, se pone en marcha la vacunación contra el neumococo en los grupos de riesgo (recién nacidos, mayores de 65 años, enfermos crónicos, esplenectomizados, inmunodeprimidos). Al año siguiente ocurren solo 6 ingresos por ENI. Queremos saber si el descenso de 9 ingresos a 6 se debe a la vacunación o si puede deberse a una variación al azar. Para ello, responda a las siguientes preguntas:

Pregunta 3.37. Si la media es 9 ingresos al año, ¿cuál es la probabilidad de que ocurran 6 ingresos en un año?

Pregunta 3.38. Si la media es 9 ingresos al año, ¿cuál es la probabilidad de que ocurran 6 ingresos o menos en un año?

PROBLEMA 3.14

En la misma provincia del problema 3.13, el número medio anual de casos de varicela es 27. Un año se producen 33. Necesitamos saber si este número es anormalmente alto (por lo que deberíamos llevar a cabo una investigación para saber qué está pasando) o pudiera ser debido a una variación al azar.

Pregunta 3.39. Si la media es 27 casos al año, ¿cuál es la probabilidad de que un año ocurran 33 casos o más?

PROBLEMA 3.15

El 27 de junio de 2020, Chu et al. Publicaron en The Lancet un metaanálisis sobre, entre otros, los efectos de la distancia interpersonal en la transmisión del SARS-CoV-2¹. Estimaron que la probabilidad de infección entre personas a menos de un metro de distancia es del 12.8 %, mientras que a más de un metro es del 2.6 %.

Pregunta 3.40. ¿Cuál sería la probabilidad de infección en una población en que la distancia fuera menor que un metro entre el 20 % de las personas?

Pregunta 3.41. Si en Nochebuena todas las cenas familiares reunieran a exactamente 6 personas a una distancia de menos de un metro, ¿qué distribución seguiría el número de infectados en las cenas de Nochebuena?

Pregunta 3.42. ¿Cuál es la probabilidad de que, en una reunión de 20 personas en que todas están a menos de un metro de distancia, se infecten al menos 2?

Pregunta 3.43. ¿Y la de que, en una reunión de 1000 personas en que todas están a más de un metro de distancia, se infecten exactamente 10?

¹Chu, D. K. et al. “Physical distancing, face masks, and eye protection to prevent person-to-person transmission of SARS-CoV-2 and COVID-19: a systematic review and meta-analysis”. The Lancet, 395. 1973-1987 (2020)

Estimación de una media

4

4.1 Introducción

En este capítulo se practicará el procedimiento para calcular una media con su intervalo de confianza y saber posteriormente interpretarla.

En la estimación de una media vamos a seguir el siguiente **procedimiento**:

1. Calcular la media muestral ($\bar{m}$)
2. Calcular la desviación estándar muestral (s)
3. Calcular el error estándar de la media (desviación estándar de la distribución de las medias)
4. Calcular el intervalo de confianza al 95 %
5. Interpretar el intervalo de confianza al 95 %

A través del siguiente ejemplo, se muestran los **problemas** que pueden existir de **interpretación** del intervalo de confianza: $\text{Media} = 170.0$; $\text{IC}95\% = (166.8; 173.2)$

Interpretaciones correctas:

- Tenemos una confianza del 95 % de que la media de la población está entre 166.8 y 173.2.
- En mi muestra he obtenido una media de 170, pero Si repitiésemos el estudio 100 veces en 95 ocasiones los valores estarían comprendidos entre 167 y 173.

Interpretaciones erróneas:

- El verdadero valor de la media en la población estará entre 166.8 y 173.2 el 95 % de las veces.
- El 95 % de la población tiene una altura entre 166.8 y 173.2.

Los problemas se realizarán sobre un estudio que pretende saber si el balance hídrico, el tipo de suero utilizado y la hipercloremia en las primeras 24 horas son factores de riesgo del daño renal agudo.

4.2 Fórmulas

Concepto	Fórmula
<i>Muestras grandes</i>	
Media	$m = \frac{\sum_{i=1}^n x_i}{n}$
Varianza	$s^2 = \frac{\sum_{i=1}^n (x_i - m)^2}{n-1}$
Desviación estándar	$s = \sqrt{\frac{\sum_{i=1}^n (x_i - m)^2}{n-1}}$
Estimación de una media	$EEM = \frac{s}{\sqrt{n}}$
	$m \pm z_{\alpha/2} \times \frac{s}{\sqrt{n}}$
	$m \pm t_{\alpha/2; n-1} \times \frac{s}{\sqrt{n}}$
	La distribución normal no funciona bien para muestras pequeñas ($n < 60$) y en su lugar se utiliza la distribución t de Student
Valores de z para diferentes intervalos de confianza (consejo: compruebe estos valores en la tabla de la distribución normal del material suplementario)	Para IC 80 %: $z_{\alpha/2} = 1.28$ Para IC 90 %: $z_{\alpha/2} = 1.64$ Para IC 95 %: $z_{\alpha/2} = 1.96$ Para IC 99 %: $z_{\alpha/2} = 2.58$

4.3 Descripción de la base de datos

La siguiente tabla presenta la descripción de todas las variables incluidas en la hoja **estudioDañoRenal** del fichero **datosProblemasBest.xlsx**:

Nombre	Descripción	Codificación
Id	Número de identificación del sujeto	Numérica
Caso	Variable identificadora del caso	1: Necesita diálisis 0: No necesita diálisis
Edad	Edad en años cumplidos	Numérica
Sexo	Sexo	1: Varón 0: Mujer
Apache	Puntuación Apache II	Numérica

Nombre	Descripción	Codificación
Hta	Variable indicadora de hipertensión arterial	1: Sí 0: No
DiabMell	Variable indicadora de diabetes mellitus	1: Sí 0: No
Epoc	Variable indicadora de enfermedad pulmonar obstructiva crónica	1: Sí 0: No
Cardiopatía	Variable indicadora de cardiopatía	1: Sí 0: No
Sepsis	Variable indicadora de sepsis	1: Sí 0: No
CreatIng	Creatinina en sangre al ingreso en UCI	Numérica
UreaIng	Urea en sangre al ingreso en UCI	Numérica
Cl24h	Cloro en sangre a las 24 horas del ingreso	Numérica
Balance	Balance hídrico en las primeras 24 horas del ingreso (en centímetros cúbicos)	Numérica
TipoFluido	Tipo de fluidos recibidos	1: Suero salino 2: Suero salino + suero balanceado 3: Suero salino + coloide 4: Suero salino + balanceado + coloide 5: Ninguno
AccMet	Variable indicadora de acidosis metabólica por hipercloremia a las 24 horas del ingreso en UCI	1: Sí 0: No
Submuestra1	Variable que incluye un 70 % de los datos	1: Sí 0: No
Submuestra2	Variable que incluye un 30 % de los datos	1: Sí 0: No

4.4 Descripción del estudio

El daño renal agudo es una complicación frecuente en los pacientes que han necesitado reanimación en UCI.

Objetivo: Evaluar si el balance hídrico, el tipo de suero utilizado y la hipercloremia en las primeras 24 horas son factores de necesitar diálisis.

Ámbito: Estudio multicéntrico en una serie de hospitales.

Criterios de inclusión:

- **Casos:** Los casos son pacientes ingresados en UCI y que después han requerido diálisis
- **Controles:** Los controles se seleccionan entre los pacientes ingresados en UCI y que no han requerido diálisis.

Tanto de casos como de controles se excluyeron los pacientes con antecedentes de enfermedad renal.

Recogida de información: Los pacientes fueron reclutados entre el 1 de enero de 2016 y el 31 de diciembre de 2018. La información se obtuvo revisando la historia clínica hospitalaria. Apache II es una escala de gravedad de la enfermedad; puede estar entre 0 y 71. A mayor puntuación, más gravedad. Apache II y la presencia de comorbilidades (diabetes, EPOC, hipertensión, cardiopatía) se refieren al momento de ingreso en UCI. Sepsis se refiere a si el motivo del ingreso en UCI fue la sepsis.



4.5



PROBLEMAS

PROBLEMA 4.1

Para resolver las preguntas 4.1 a 4.3 de forma manual utilice la siguiente tabla:

	Edad		
	Muestra	Controles	Casos
n	636	318	318
$\sum_{i=1}^n x_i$	41808	20423	21385
$\sum_{i=1}^n (x_i - m)^2$	133007.6	63989.15	67563.37

Pregunta 4.1. Calcule la media y el intervalo de confianza al 95 % de la edad de la muestra.

Pregunta 4.2. Calcule la media y el intervalo de confianza al 95 % de la edad de los controles.

Pregunta 4.3. Calcule la media y el intervalo de confianza al 95 % de la edad de los casos.

PROBLEMA 4.2

Para resolver las preguntas 4.4 a 4.6 de forma manual utilice la siguiente tabla:

	Apache II		
	Muestra	Controles	Casos
n	634	318	316
$\sum_{i=1}^n x_i$	15970	7901	8069
$\sum_{i=1}^n (x_i - m)^2$	28781.29	12221.45	16484.62

Pregunta 4.4. Calcule la media y el intervalo de confianza al 90% de la puntuación Apache II de la muestra.

Pregunta 4.5. Calcule la media y el intervalo de confianza al 90% de la puntuación Apache II de los controles.

Pregunta 4.6. Calcule la media y el intervalo de confianza al 90% de la puntuación Apache II de los casos.

PROBLEMA 4.3

Para resolver las preguntas 4.7 a 4.9 de forma manual utilice la siguiente tabla:

Creatinina en sangre			
	Muestra	Controles	Casos
n	627	313	314
$\sum_{i=1}^n x_i$	1522.71	571.32	951.39
$\sum_{i=1}^n (x_i - m)^2$	2621.293	597.065	951.774

Pregunta 4.7. Calcule la media y el intervalo de confianza al 80 % de la creatinina en sangre de la muestra.

Pregunta 4.8. Calcule la media y el intervalo de confianza al 80 % de la creatinina en sangre de los controles.

Pregunta 4.9. Calcule la media y el intervalo de confianza al 80 % de la creatinina en sangre de los casos.

PROBLEMA 4.4

Para resolver las preguntas 4.10 a 4.12 de forma manual utilice la siguiente tabla:

	Urea		
	Muestra	Controles	Casos
n	625	311	314
$\sum_{i=1}^n x_i$	54258.59	21868.83	32489.76
$\sum_{i=1}^n (x_i - m)^2$	3082005	901446.9	2008827

Pregunta 4.10. Calcule la media y el intervalo de confianza al 99% de la urea en sangre de la muestra.

Pregunta 4.11. Calcule la media y el intervalo de confianza al 99% de la urea en sangre de los controles.

Pregunta 4.12. Calcule la media y el intervalo de confianza al 99% de la urea en sangre de los casos.

PROBLEMA 4.5

Para resolver las preguntas 4.13 a 4.15 de forma manual utilice la siguiente tabla:

	Cloro en sangre		
	Muestra	Controles	Casos
n	458	225	233
$\sum_{i=1}^n x_i$	48425.4	23952	24473.4
$\sum_{i=1}^n (x_i - m)^2$	15626.22	6977.76	8418.537

Pregunta 4.13. Calcule la media y el intervalo de confianza al 95% del cloro en sangre de la muestra.

Pregunta 4.14. Calcule la media y el intervalo de confianza al 95% del cloro en sangre de los controles.

Pregunta 4.15. Calcule la media y el intervalo de confianza al 95% del cloro en sangre de los casos.

PROBLEMA 4.6

Con la información obtenida en los problemas anteriores cumplimente la siguiente tabla. A continuación, interprete los intervalos de confianza indicados en las preguntas.

	Muestra		Controles		Casos	
	media	IC 95 %	media	IC 95 %	media	IC 95 %
Edad						
Puntuación Apache II						
Creatinina en sangre						
Urea en sangre						
Cloro en sangre						

Pregunta 4.16. De la edad de los controles.

Pregunta 4.17. La edad de los casos.

Pregunta 4.18. La puntuación Apache II de los controles.

Pregunta 4.19. La puntuación Apache II de los casos.

Pregunta 4.20. La urea en sangre de los controles.

Pregunta 4.21. La urea en sangre de los casos.

Pregunta 4.22. El cloro en sangre de los controles.

Pregunta 4.23. El cloro en sangre de los casos.

PROBLEMA 4.7

Para ver la importancia del tamaño muestral en la estimación de una media se han generado dos submuestras al azar en la base de datos. La variable submuestra 1 incluye el 70 % de los datos ($N=447$) y en la submuestra 2 incluye el 30 % de los datos ($N=184$).

Para resolver las preguntas 4.24 a 4.26 de forma manual utilice la siguiente tabla:

Edad			
	Muestra	Submuestra 1	Submuestra 2
n	636	447	184
$\sum_{i=1}^n x_i$	41808	29127	12143
$\sum_{i=1}^n (x_i - m)^2$	133007.6	94274.4	40176.99

Pregunta 4.24. Calcular la media, la desviación típica y el error estándar de la media de la edad para toda la muestra.

Pregunta 4.25. Calcular la media, la desviación típica y el error estándar de la media de la edad para la submuestra 1.

Pregunta 4.26. Calcular la media, la desviación típica y el error estándar de la media de la edad para la submuestra 2.

Para resolver las preguntas 4.27 a 4.29 de forma manual utilice la siguiente tabla:

Urea en sangre			
	Muestra	Submuestra 1	Submuestra 2
n	625	437	181
$\sum_{i=1}^n x_i$	54358.59	37015.16	16062.29
$\sum_{i=1}^n (x_i - m)^2$	3082005	2168226	1115155

Pregunta 4.27. Calcular la media, la desviación típica y el error estándar de la media de la urea en sangre para toda la muestra.

Pregunta 4.28. Calcular la media, la desviación típica y el error estándar de la media de la urea en sangre para la submuestra 1.

Pregunta 4.29. Calcular la media, la desviación típica y el error estándar de la media de la urea en sangre para la submuestra 2.

Pregunta 4.30. Con la información obtenida en las preguntas 4.24 a 4.29 completa la siguiente tabla y reflexiona sobre los cambios que se producen la media, la desviación típica y el error estándar de la media al disminuir el tamaño muestral.

Edad	Todos	Submuestra1=1	Submuestra2=1
N	636	447	184
Media			
Desviación típica			
Error estándar de la media			
Urea en sangre	Todos	Submuestra1=1	Submuestra2=1
N	625	437	181
Media			
Desviación típica			
Error estándar de la media			

PROBLEMA 4.8

Pregunta 4.31. En nuestra muestra de 418 hombres que ingresaron en la UCI, la creatinina en sangre tenía $m = 2.6$ y $s = 2.18$. ¿Se puede asumir que la muestra procedía de una población de hombres procedentes de cuidados intensivos con $m = 2.4$?

Pregunta 4.32. En nuestra muestra de 209 mujeres que ingresaron en la UCI, la creatinina en sangre tenía $m = 2.1$ y $s = 1.70$. ¿Se puede asumir que la muestra procedía de una población de mujeres procedentes de cuidados intensivos con $m = 1.7$?

Pregunta 4.33. En nuestra muestra la urea en sangre al ingreso en UCI de los 311 pacientes que no requirieron diálisis tenía $m = 70.3$ y $s = 53.9$. ¿Se puede asumir que la muestra procedía de una población de pacientes procedentes de la UCI con $m = 69$ y $s = 50.2$?

Pregunta 4.34. En nuestra muestra de pacientes, el cloro en sangre a las 24 horas del ingreso en UCI de los 225 pacientes que no requirieron diálisis tenía una $m = 106.5$ y $s = 5.58$. ¿Se puede asumir que la muestra procede de una población con una media mayor de 106?

5

Comparación de dos medias

5.1 Introducción

En este capítulo se practicará el uso de la t de Student para:

- Comparar la media de dos muestras con datos independientes y varianzas homogéneas.
- Comparar la media de dos muestras con datos independientes y varianzas heterogéneas.
- Comparar la media de dos muestras con datos dependientes.

Los problemas se realizarán sobre un estudio que relaciona la enfermedad cardiovascular y diabetes con varios factores.

5.2 Procedimiento

Concepto	Fórmula t de Student	
Media	$m = \frac{\sum_{i=1}^n x_i}{n}$	
Varianza	$s^2 = \frac{\sum_{i=1}^n (x_i - m)^2}{n - 1}$	
Desviación estándar	$s = \sqrt{\frac{\sum_{i=1}^n (x_i - m)^2}{n - 1}}$	
Cálculo de varianza conjunta	Cálculo de la t de Student si varianzas homogéneas*	Cálculo de la t de Student si varianzas no homogéneas*
	$s_p^2 = \frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{(n_A - 1) + (n_B - 1)}$	

Concepto	Fórmula t de Student	
Grados de libertad	$g.l. = n_A + n_B - 2$	$gl^* = \frac{\left(\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}\right)^2}{\frac{\left(\frac{s_A^2}{n_A}\right)^2}{n_A-1} + \frac{\left(\frac{s_B^2}{n_B}\right)^2}{n_B-1}}$ g.l. corregidos (Test de Welch)
Error estándar de la diferencia de medias	$EEDM = s_p \sqrt{\frac{1}{n_A} + \frac{1}{n_B}}$	$EEDM = \sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}$
Cálculo de t	$t = \frac{dm}{EEDM} = \frac{m_A - m_B}{s_p \sqrt{\frac{1}{n_A} + \frac{1}{n_B}}}$	$t = \frac{m_A - m_B}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}}$
Intervalo de confianza	$IC\ 95\% : dm \pm t_{0.025; n_A+n_B-2} \times EEDM$	

* Se asume que las varianzas son homogéneas si $\frac{s_A^2}{s_B^2} < 4$ y $\frac{s_B^2}{s_A^2} < 4$

En estos problemas vamos a considerar la regla de Box, es decir, si una varianza no es cuatro veces mayor que la otra, se puede usar directamente la t. Solo si es al menos cuatro veces mayor, usaremos el test de Welch. Observe que esto es equivalente a que la desviación típica mayor sea el doble que la desviación típica menor.

5.3 Descripción de la base de datos

La siguiente tabla presenta la descripción de todas las variables incluidas en la hoja **estudioEfCardiovascularDiabetes** del fichero **datosProblemasBest.xlsx**:

Nombre	Descripción	Codificación
ID	Número de identificación	Numérica
SEXO	Sexo	0: mujer, 1: varón
EDAD	Edad	Numérica (años cumplidos)
MUERTE	Muerte durante el estudio	0: No, 1: Si
ANGINA	Angina de pecho	0: No, 1: Si
ACV	Accidente cerebrovascular	0: No, 1: Si
FUMADOR	Fumador	0: No, 1: Si
IMC	Índice de masa corporal	Numérica
COLESTEROL	Nivel de colesterol (mg/dl)	Numérica

Nombre	Descripción	Codificación
COLESTEROL2	Nivel de colesterol (mg/dl) un mes después del diagnóstico	Numérica
GLUCEMIA	Glucemia	Numérica
GLUCEMIA2	Glucemia un mes después del diagnóstico	Numérica
TAS	tensión arterial sistólica	Numérica (mm de Hg)
TAS2	tensión arterial sistólica un mes después del diagnóstico	Numérica (mm de Hg)

5.4 Descripción del estudio

Objetivo: Valorar los factores de riesgo de enfermedad cardiovascular y diabetes se realiza un estudio transversal con personas mayores de 30 años que acuden a un centro de salud.

Recogida de información: Los pacientes fueron reclutados entre el 1 de enero de 2012 y el 31 de diciembre de 2017. La información demográfica y clínica se obtuvo revisando la historia clínica.

Las respuestas se darán en dos formatos Excel y Stata.

Los resultados se darán con una precisión de 3 decimales.

5.5

PROBLEMAS - SIN EMPAREJAMIENTO

PROBLEMA 5.1

Se cree que el nivel de colesterol está relacionado con el riesgo de muerte. Para comprobarlo realice un test de hipótesis sobre si la media de colesterol era la misma en los pacientes que murieron y en los que sobrevivieron.

	Colesterol	
	Fallecidos	Vivos
n	329	671
$\sum_{i=1}^n x_i$	87562	150312
$\sum_{i=1}^n (x_i - m)^2$	629084.997	978355.905

Pregunta 5.1. ¿Cuál es el colesterol medio en los fallecidos?

Pregunta 5.2. ¿Cuál es el valor de la t de Student?

Pregunta 5.3. ¿Cuál es el valor de la p de esta t de Student?

Pregunta 5.4. ¿Calcule el Intervalo de confianza al 95%?

Pregunta 5.5. ¿Qué conclusión saca del resultado obtenido?

PROBLEMA 5.2

Se cree que la tensión arterial sistólica aumenta el riesgo de muerte. Compruebe esta hipótesis. Para comprobarlo realice un test de hipótesis sobre si la media de la tensión arterial sistólica era mayor en los pacientes que murieron que en los que sobrevivieron.

Pregunta 5.6. ¿Cuál es la tensión arterial sistólica en los dos grupos a comparar? (vivos y muertos)

	Tensión Arterial Sistólica (TAS)	
	Fallecidos	Vivos
n	329	671
$\sum_{i=1}^n x_i$	44054	86593
$\sum_{i=1}^n (x_i - m)^2$	173745.388	256876.777

Pregunta 5.7. ¿Cuál es el valor de la t de Student?

Pregunta 5.8. ¿Cuál es el valor de la p de esta t de Student?

Pregunta 5.9. ¿Calcula el Intervalo de confianza al 95 %?

Pregunta 5.10. ¿Qué conclusión sacas del resultado obtenido?

PROBLEMA 5.3

Se cree que la media de la tensión arterial sistólica es mayor en los fumadores que en los no fumadores. Para comprobarlo realice un test de hipótesis.

	Tensión Arterial Sistólica (TAS)	
	Fumadores	No fumadores
n	486	514
$\sum_{i=1}^n x_i$	62506.5	68140.5
$\sum_{i=1}^n (x_i - m)^2$	219421.912	212490.298

Pregunta 5.11. ¿Podemos afirmar que en los fumadores la tensión arterial sistólica es mayor que en los no fumadores?

Pregunta 5.12. Calcula el intervalo de confianza al 95% de la diferencia de medias.

Pregunta 5.13. ¿Qué conclusión sacas del resultado obtenido?

PROBLEMA 5.4

Se cree que el índice de masa corporal (IMC) es mayor en los no fumadores. Para comprobarlo realice un test de hipótesis sobre si la media del IMC era la misma en los pacientes que fumaban y en los que no fumaban.

	Índice de masa corporal (IMC)	
	Fumadores	No fumadores
n	486	514
$\sum_{i=1}^n x_i$	11014.28	14786.44
$\sum_{i=1}^n (x_i - m)^2$	1827.764	5716.799

Pregunta 5.14. ¿Cuál es IMC medio de los dos grupos a comparar?

Pregunta 5.15. ¿Cuál es el valor de la t de Student?

Pregunta 5.16. ¿Qué conclusión sacas del resultado obtenido?

PROBLEMA 5.5

Se cree que la edad se relaciona al desarrollo de angina de pecho. Compruebe esta hipótesis. Para comprobarlo realice un test de hipótesis sobre si la media de la edad era la misma en los con angina de pecho que en los que no tenían angina de pecho.

	Edad	
	Angina de pecho	No angina de pecho
n	161	839
$\sum_{i=1}^n x_i$	8364	41974
$\sum_{i=1}^n (x_i - m)^2$	12841.602	62227.313

Pregunta 5.17. ¿En qué grupo la edad media es mayor?

Pregunta 5.18. ¿Cuál es el valor del test estadístico empleado?

Pregunta 5.19. ¿Cuál es el valor de la p?

Pregunta 5.20. ¿Qué conclusión sacas del resultado obtenido?

PROBLEMA 5.6

Se cree que el nivel de colesterol aumenta el riesgo de ACV. Para comprobarlo se comparan los niveles de colesterol en personas con y sin ACV. Compruebe esta hipótesis.

	Colesterol	
	ACV	No ACV
n	68	932
$\sum_{i=1}^n x_i$	18250	219624
$\sum_{i=1}^n (x_i - m)^2$	700000.059	1861438.567

Pregunta 5.21. ¿Cuál es el valor del test estadístico empleado?

Pregunta 5.22. ¿Cuál es el valor de la p?

Pregunta 5.23. ¿Calcula el intervalo de confianza al 95%?

Pregunta 5.24. ¿Qué conclusión sacas del resultado obtenido?

5.6

PROBLEMAS - CON EMPAREJAMIENTO

PROBLEMA 5.7

En este estudio se pretende comprobar si en solo un mes de estatinas el nivel de colesterol disminuye, para eso se mide el colesterol antes de empezar el tratamiento (COLESTEROL) y después de un mes de tratamiento con estatinas (COLESTEROL2).

Tto. 1 mes de Estatinas			
	Colesterol antes	Colesterol después	Diferencia
n	1000	1000	1000
Media (m)	237.874	237.330	-0.544
Desviación estándar (s)	45.544	44.736	53.856

Pregunta 5.25. ¿Podemos afirmar que después de un mes de tratamiento disminuye el nivel de colesterol de los sujetos de nuestro estudio?

Pregunta 5.26. Calcula el intervalo de confianza al 95 % de la diferencia de medias.

Pregunta 5.27. ¿Qué conclusión sacas del resultado obtenido?

PROBLEMA 5.8

A todos los participantes en el estudio se les recomendó realizar una actividad física diaria que superase los 10000 pasos. Al mes de esta recomendación se les volvió a mirar la glucemia en sangre (GLUCEMIA2).

Actividad física > 10000 pasos (1 mes)			
	Glucemia antes	Glucemia después	Diferencia
n	1000	1000	1000
Media (m)	81.252	136.352	-55.100
Desviación estándar (s)	27.567	23.424	33.972

Pregunta 5.28. ¿Podemos afirmar que después de un mes de actividad física el nivel de glucemia en sangre a disminuido?

Pregunta 5.29. Calcula el intervalo de confianza al 95 % de la diferencia de medias.

Pregunta 5.30. ¿Qué conclusión sacas del resultado obtenido?

PROBLEMA 5.9

Una vez que se han visto los resultados de la actividad física se pretende añadir en estos participantes una segunda recomendación, disminuir el nivel de sal en la dieta para ver cómo influye en la tensión arterial.

	Disminución sal en dieta (1 mes)		
	TAS antes	TAS después	Diferencia
n	1000	1000	1000
Media (m)	130.980	130.647	0.334
Desviación estándar (s)	21.026	20.887	26.351

Pregunta 5.31. ¿Podemos afirmar que después de un mes con un menor consumo de sal la tensión arterial sistólica al mes (TAS2) es menor que la tensión arterial sistólica antes de la recomendación (TAS)?

Pregunta 5.32. ¿Qué conclusión sacas de estos resultados?

6

Comparación de tres o más medias

6.1 Introducción

En este capítulo vamos a utilizar el ANOVA de una vía para comprobar hasta qué punto hay relación entre una variable independiente o de agrupación con más de dos categorías y una variable continua. Por tanto, es un procedimiento adecuado para comparar 3 o más medias.

Con estos ejercicios se pretende que practiques:

- El procedimiento de cálculo
- Las comparaciones múltiples

Los problemas se realizarán sobre el mismo estudio que relaciona la enfermedad cardiovascular y diabetes con varios factores del capítulo 5.

6.2 Procedimiento

Pasos para el cálculo del análisis de la varianza (ANOVA)

1. Datos necesarios.

Variable categórica	Número de pacientes	Media	Varianza
A	n_A	m_A	s_A^2
B	n_B	m_B	s_B^2
C	n_C	m_C	s_C^2
Total	n	m	s^2

2. Crear tabla del ANOVA.

Fuente de variación	Suma de cuadrados	Grados de libertad	Varianza	F	P
Entre grupos	$SC_{eg} = \sum_{i=1}^k n_i (m_i - m)^2$	$k-1$	$V_e = \frac{SC_{eg}}{(k-1)}$	$F = \frac{V_e}{V_r}$	$p > 0.05$ $p < 0.05$
Residual	$SCR = SCT - SC_{eg}$	$n - k$	$V_r = \frac{SCR}{(n-k)}$		
Total	$SCT = (n-1)s^2$	$n - 1$			

Donde:

- SC_{eg} : suma de cuadrados entre grupos
- SCR : suma de cuadrados residual
- SCT : suma de cuadrados total
- V_e : varianza entre grupos
- V_r : varianza residual
- g.l.: grados de libertad
- k : número de categorías
- F : valor del estadístico F
- p : significación estadística. Buscar el valor p en la tabla F con $k-1$, $n-k$ grados de libertad
- m_i es la media con el tratamiento i y m es la media total
- s_i^2 es la varianza con el tratamiento i y s^2 es la varianza total.
- n_i es el tamaño de muestra con tratamiento i y n es la muestra total

Vamos a considerar que las varianzas son homogéneas en la t de Student si la desviación típica mayor no supera el doble de la menor.

6.3 Descripción de la base de datos

La siguiente tabla presenta la descripción de todas las variables incluidas en la hoja **estudioEfCardiovascularDiabetes** del fichero **datosProblemasBest.xlsx**:

Nombre	Descripción	Codificación
ID	Número de identificación	Numérica
SEXO	Sexo	0: mujer, 1: varón
EDAD	Edad	Numérica (años cumplidos)

Nombre	Descripción	Codificación
GRUP_EDAD	Grupo de edad	Categórica 1: =<50 2: 50-65 3: >=65
IMC	Índice de masa corporal	Numérica
IMC_CAT	IMC categorizado	Categórica 1: Normopeso 2: Sobrepeso 3: Obesidad
COLESTEROL	Nivel de colesterol (mg/dl)	Numérica
COLESTEROL_CAT	Colesterol categorizado	Categórica 0: <200 1: 200-239 2: >=240 mg/dl
GLUCEMIA	Glucemia	Numérica
TAS	tensión arterial sistólica	Numérica (mm de Hg)

6.4 Descripción del estudio

Objetivo: Valorar los factores de riesgo de enfermedad cardiovascular y diabetes se realiza un estudio transversal con personas mayores de 30 años que acuden a un centro de salud.

Recogida de información: Los pacientes fueron reclutados entre el 1 de enero de 2018 y el 31 de diciembre de 2019. La información demográfica y clínica se obtuvo revisando la historia clínica.

Utilizando esta base de datos en Excel, responde a los siguientes problemas.

Las respuestas se darán en dos formatos Excel y Stata.

Los resultados se darán con una precisión de 3 decimales

6.5

PROBLEMAS

PROBLEMA 6.1

Se cree que el colesterol se relaciona con el Índice de masa corporal (IMC_cat). Para comprobarlo realice un test de hipótesis sobre si la media de colesterol era la misma en los pacientes con normopeso, sobrepeso y los que tenían obesidad.

IMC	Número de pacientes	Media	Desviación estandar
Normopeso	456	234.509	46.834
Sobrepeso	417	241.000	42.245
Obesidad	127	236.780	44.267
Total	1000	237.874	

Fuente de variación	Suma de cuadrados	Grados de libertad	Varianza	F	P
Entre grupos					
Residual	1987315.494				
Total	1999348.124				

Pregunta 6.1. ¿Cuál es el colesterol medio en los que tienen sobrepeso?

Pregunta 6.2. ¿Cuál es el valor de F?

Pregunta 6.3. ¿Cuál es el valor de la p?

Pregunta 6.4. ¿Qué conclusión sacas del resultado obtenido?

PROBLEMA 6.2

Se cree que la glucemia varía según el grupo de edad (GRUP_EDAD) del paciente. Para comprobarlo realice un test de hipótesis sobre si la media de glucemia era la misma en los 3 grupos de edad: pacientes hasta 50 años (Edad1), entre 50-64 años (Edad2) o iguales o mayores a 65 años (Edad3).

Compruebe esta hipótesis. Para ello, en primer lugar, complete esta tabla.

GLUCEMIA	Número de pacientes	Media	Desviación estandar
Edad-1	492	128.784	18.616
Edad-2	454	147.629	25.162
Edad-3	54	152.537	24.181
Total	1000	136.352	

Fuente de variación	Suma de cuadrados	Grados de libertad	Varianza	F	P
Entre grupos					
Residual	487942.835				
Total	548158.096				

Pregunta 6.5. ¿Cuál es la glucemia basal media en los tres grupos a comparar?

Pregunta 6.6. ¿Cuál es el valor de F?

Pregunta 6.7. ¿Cuál es el valor de la p?

Pregunta 6.8. ¿Qué conclusión sacas del resultado obtenido?

PROBLEMA 6.3

Se cree que la tensión arterial sistólica se relaciona con el Índice de masa corporal (IMC_CAT: normopeso/sobrepeso/obesidad)). Para comprobarlo realice un test de hipótesis sobre si la media de la tensión arterial sistólica era la misma en los pacientes con normopeso, sobrepeso y los que tenían obesidad.

En primer lugar, complete la siguiente tabla:

TAS	Número de pacientes	Media	Desviación estandar
Normopeso	456	128.498	21.630
Sobrepeso	417	132.332	19.806
Obesidad	127	132.831	21.056
Total	1000	130.647	

Fuente de variación	Suma de cuadrados	Grados de libertad	Varianza	F	P
Entre grupos					
Residual	431923.357				
Total	435819.391				

Pregunta 6.9. ¿Cuál es la tensión arterial sistólica en las personas con obesidad?

Pregunta 6.10. ¿Cuál es el valor de F?

Pregunta 6.11. ¿Cuál es el valor de la p?

Pregunta 6.12. ¿Qué conclusión sacas del resultado obtenido?

PROBLEMA 6.4

Se cree que la edad es diferente en función del IMC del paciente (IMC_CAT). Para comprobarlo realice un test de hipótesis sobre si la media de la edad en los pacientes con normopeso, sobrepeso y los que tenían obesidad.

Para ello, complete la siguiente tabla:

EDAD	Número de pacientes	Media	Desviación estandar
Normopeso	456	49.614	8.888
Sobrepeso	417	51.153	8.407
Obesidad	127	50.260	8.776
Total	1000	50.338	

Fuente de variación	Suma de cuadrados	Grados de libertad	Varianza	F	P
Entre grupos					
Residual	75050.673				
Total	75567.756				

Pregunta 6.13. ¿En qué grupo la edad media es mayor?

Pregunta 6.14. ¿Cuál es el valor de F?

Pregunta 6.15. ¿Cuál es el valor de la p?

Pregunta 6.16. ¿Qué conclusión sacas del resultado obtenido?

PROBLEMA 6.5

Se cree que la glucemia varía según el IMC (IMC_CAT) del paciente. Compruebe esta hipótesis.

GLUCEMIA	Número de pacientes	Media	Desviación estandar
Normopeso	456	132.508	24.371
Sobrepeso	417	137.747	21.663
Obesidad	127	145.575	22.596
Total	1000	136.352	

Fuente de variación	Suma de cuadrados	Grados de libertad	Varianza	F	P
Entre grupos					
Residual	529804.821				
Total	548158.096				

Pregunta 6.17. ¿Cuál es el valor del test empleado?

Pregunta 6.18. ¿Qué conclusión sacas del resultado obtenido?

7

Estimación de una proporción

7.1 Introducción

En este capítulo se abordará la estimación de una proporción utilizando la aproximación normal. Los problemas se realizarán sobre un estudio que relaciona el cáncer de mama con distintas exposiciones potencialmente de riesgo.

7.2 Fórmulas

- Estadístico z

$$z = \frac{p - \pi_0}{\sqrt{\frac{\pi_0 \times (1 - \pi_0)}{n}}}$$

- Intervalo de confianza de una proporción

$$p \pm z_{1-\alpha/2} \sqrt{\frac{p(1-p)}{n}}$$

7.3 Descripción de la base de datos

La siguiente tabla presenta la descripción de todas las variables incluidas en la hoja **estudioCaMama** del fichero **datosProblemasBest.xlsx**:

Nombre	Descripción	Codificación
id	Número de identificación	Numérica
caso	Caso de cáncer de mama	0: No cáncer de mama 1: Cáncer de mama
EdadMenar	Edad de la menarquia	Numérica
THS	Toma de terapia hormonal sustitutiva	0: No 1: Si

Nombre	Descripción	Codificación
AFamCancer	Antecedentes familiares de cáncer de mama	0: No 1: Familiar 1er grado 2: Familiar 2o grado 3: Otro familiar
Menopausia	Menopausia	0: No 1: Si

7.4 Descripción del estudio

Objetivo: Evaluar si distintas exposiciones como el consumo de anticonceptivos orales (AO), la terapia hormonal sustitutiva (THS) o los antecedentes familiares de cáncer de mama (AF), se asocian al desarrollo de cáncer de mama.

Diseño: Se recogió información sobre dos grupos de mujeres: 1) mujeres con cáncer de mama diagnosticado (por anatomía patológica) en cualquier centro sanitario de las 5 provincias participantes en el estudio entre el 1 de enero de 2013 y el 31 de diciembre de 2015 y 2) mujeres sin cáncer de mama seleccionadas al azar de la lista de mujeres adscritas a los centros de salud en las 5 provincias participantes, residentes en la provincia durante al menos 2 años.

Todos los problemas pueden resolverse utilizando **Excel** o **Stata**.



7.5



PROBLEMAS

PROBLEMA 7.1

Calcule la proporción muestral de mujeres participantes en el estudio cuya menarquia hubiera sucedido antes de los 13 años, y responda a las siguientes preguntas.

Pregunta 7.1. Construya el intervalo de confianza al 95% para π , la probabilidad de que la menarquia haya sucedido antes de los 13 años, usando la aproximación normal.

Pregunta 7.2. De acuerdo con el intervalo de confianza construido, ¿puede asegurarse, con una confianza del 95%, que es más probable que la menarquia sucediera antes, y no después, de los 13 años?

Pregunta 7.3. ¿Puede asumirse que la muestra procede de una población con $\pi=0.65$? Plantee las hipótesis del estudio.

PROBLEMA 7.2

Calcule la proporción muestral de mujeres participantes en el estudio cuyo primer parto hubiera sucedido después de los 30 años, y responda a las siguientes preguntas.

Pregunta 7.4. Construya el intervalo de confianza al 90% para π , la probabilidad de que el primer parto haya sucedido después de los 30 años, usando la aproximación normal.

Pregunta 7.5. De acuerdo con el intervalo de confianza construido, ¿puede asegurarse, con una confianza del 90%, que es más probable que el primer parto sucediera antes, y no después, de los 30 años?

Pregunta 7.6. ¿Puede asumirse que la muestra procede de una población con $\pi = 0.15$?

PROBLEMA 7.3

En este estudio se recogió información sobre el consumo de terapia hormonal sustitutiva (THS) en 1139 mujeres, de ellas se observó que 364 había consumido THS. Calcule la proporción muestral de mujeres participantes en el estudio consumidoras de THS, y responda a las siguientes preguntas.

Pregunta 7.7. Construya el intervalo de confianza al 99% para π , la probabilidad de haber consumido THS, usando la aproximación normal.

Pregunta 7.8. De acuerdo con el intervalo de confianza construido, ¿puede asegurarse, con una confianza del 99%, que es más probable haber consumido THS?

Pregunta 7.9. ¿Puede asumirse que la muestra procede de una población con $\pi = 0.35$?

PROBLEMA 7.4

En el mismo estudio, se define ahora un subgrupo de 41 mujeres con las siguientes características: (1) todas habían recibido terapia hormonal sustitutiva; y (2) tenían antecedentes familiares de cáncer de mama de primer grado. En este grupo, 6 mujeres habían recibido terapia hormonal sustitutiva durante más de dos años. Sea π la proporción de mujeres que, con las características (1) y (2), reciben terapia hormonal sustitutiva durante más de dos años.

Pregunta 7.10. Construya el intervalo de confianza al 95 % para π .

Pregunta 7.11. Construya el intervalo de confianza al 95 % exacto (usando la distribución binomial) para π .

Pregunta 7.12. ¿Qué conclusión con respecto a π puede extraerse de ellos?

Comparación de dos proporciones



8.1 Introducción

En este capítulo se abordará la comparación de dos proporciones utilizando la aproximación normal.

Los problemas se realizarán sobre el estudio que relaciona el cáncer de mama con distintas exposiciones potencialmente de riesgo presentado en el capítulo anterior.

8.2 Notación

N : Número total de mujeres del estudio.

n_1 : Total mujeres con cáncer de mama.

n_0 : Total de mujeres sin cáncer de mama.

$p_1 = P(\bullet \mid C = 1)$: Proporción de exposición en mujeres con cáncer de mama.

$p_0 = P(\bullet \mid C = 0)$: Proporción de exposición en mujeres sin cáncer de mama.

$P(\cdot)$: Proporción de exposición en toda la población de estudio.

8.3 Fórmulas

- Estadístico z

$$z = \frac{p_0 - p_1}{EE(p_0 - p_1)}$$

$$EE(p_0 - p_1) = \sqrt{P(1 - P) \left(\frac{1}{n_1} + \frac{1}{n_0} \right)}$$

- Intervalo de confianza de la diferencia entre dos proporciones:

$$p_1 - p_0 \pm z_{\alpha/2} \sqrt{\frac{P \times (1 - P)}{n_1} + \frac{P \times (1 - P)}{n_0}}$$

8.4 Descripción de la base de datos

La siguiente tabla presenta la descripción de todas las variables incluidas en la hoja **estudioCaMama** del fichero **datosProblemasBest.xlsx**:

Nombre	Descripción	Codificación
id	Número de identificación	Numérica
caso	Caso de cáncer de mama	0: No cáncer de mama 1: Cáncer de mama
AnovOrales	Toma de anovulatorios orales (AO)	0: No 1: Si
THS	Toma de terapia hormonal substitutiva	0: No 1: Si
AFamCancer	Antecedentes familiares de cáncer de mama	0: No 1: Familiar 1er grado 2: Familiar 2o grado 3: Otro familiar
Menopausia	Menopausia	0: No 1: Si
Menar11	Menarquia antes de los 11 años	0: No 1: Si

Todos los problemas pueden resolverse utilizando **Excel** o **Stata**.



8.5



PROBLEMAS

PROBLEMA 8.1

Al describir los datos obtenidos en el estudio se observó que, de las 1517 mujeres incluidas, 1158 habían tomado anovulatorios orales durante más de 1 año a lo largo de su vida fértil. Al evaluar esta exposición en mujeres con y sin cáncer de mama, 548 de las 711 mujeres con cáncer de mama y 610 de las 806 mujeres sin cáncer habían tomado AO durante más de 1 año.

Pregunta 8.1. Con los datos proporcionados sobre el estudio construya la tabla de contingencia correspondiente.

Pregunta 8.2. Calcule la proporción de exposición a AO en toda la población de estudio.

Pregunta 8.3. Calcule la proporción de exposición a AO en mujeres con y sin cáncer de mama.

Pregunta 8.4. ¿Hay diferencias en el consumo de AO en mujeres con y sin cáncer de mama?

Pregunta 8.5. Calcule el IC al 95 % de la diferencia de proporciones entre ambos grupos.

PROBLEMA 8.2

Los investigadores se plantean si los resultados del problema 8.1 pueden deberse a que la proporción de menopáusicas es inferior en el grupo de mujeres sin cáncer de mama, ya que es posible que las mujeres premenopáusicas recuerden mejor la exposición a AO por ser más reciente. Utilizando el fichero de Excel del capítulo 8 responda a las siguientes preguntas.

Pregunta 8.6. Construya la tabla de contingencia a partir de la base de datos.

Pregunta 8.7. Plantee las hipótesis de este estudio.

Pregunta 8.8. Realice el contraste de hipótesis aplicando la prueba z.

Pregunta 8.9. A la vista de los resultados, ¿se asocia el estado menopáusico a la condición de ser mujer con cáncer de mama?

Pregunta 8.10. Calcule el IC al 95 % de la diferencia de proporciones de menopáusicas en ambos grupos.

PROBLEMA 8.3

Se sabe que la paridad es un factor protector frente al desarrollo de cáncer de mama, ya que las mujeres que han tenido embarazos han estado expuestas a niveles menores de estrógenos circulantes. Se quiere comprobar si, en este estudio, las mujeres con y sin cáncer se diferencian entre sí en cuanto a haber tenido hijos.

Pregunta 8.11. Construya con Excel la tabla de contingencia a partir de la base de datos.

Pregunta 8.12. Plantee las hipótesis de este estudio.

Pregunta 8.13. Realice el contraste de hipótesis aplicando la prueba z.

Pregunta 8.14. A la vista de los resultados, ¿se asocia haber tenido hijos a la condición de ser mujer con cáncer de mama?

Pregunta 8.15. Calcule el IC al 95 % de la diferencia de proporciones de mujeres que han tenido hijos en ambos grupos.

PROBLEMA 8.4

Se quiere comprobar ahora si las mujeres con y sin cáncer se diferencian entre sí en cuanto al consumo de THS.

Pregunta 8.16. Construya con Excel la tabla de contingencia a partir de la base de datos.

Pregunta 8.17. Plantee las hipótesis de este estudio.

Pregunta 8.18. Realice el contraste de hipótesis aplicando la prueba z .

Pregunta 8.19. A la vista de los resultados, ¿se asocia el consumo de THS a la condición de ser mujer con cáncer de mama?

Pregunta 8.20. Calcule el IC al 95 % de la diferencia de proporciones de consumo de THS en ambos grupos.

PROBLEMA 8.5

Se sabe que la menarquia precoz (antes de los 11 años) puede aumentar el riesgo de cáncer de mama como consecuencia a una exposición más prolongada a estrógenos a lo largo de la vida fértil. Se quiere comprobar si las mujeres con cáncer presentan menarquia precoz en mayor proporción que las mujeres sin cáncer. La tabla siguiente resume los resultados encontrados en este estudio:

	Mujer con cáncer	Mujer sin cáncer	Total
Menarquia precoz	191	170	361
No Menarquia precoz	518	640	1158
Total	709	810	1519

Pregunta 8.21. Plantee las hipótesis de este estudio.

Pregunta 8.22. Realice el contraste de hipótesis aplicando la prueba z.

Pregunta 8.23. A la vista de los resultados, ¿se asocia la menarquia precoz al cáncer de mama?

Pregunta 8.24. Calcule el IC al 95% de la diferencia de proporciones de menarquia precoz en los dos grupos de comparación.

9

Tabla de contingencia

9.1 Introducción

En este capítulo se abordará el uso de tablas de contingencia para el cálculo de la prueba χ^2 aplicada a la comparación de dos o más proporciones.

Los problemas se realizarán sobre la misma base de datos del estudio que presentamos en el capítulo anterior que relaciona el cáncer de mama con diferentes factores de riesgo (consumo de anticonceptivos orales, terapia hormonal sustitutiva, etc.). En este capítulo, se responderán a algunas de las preguntas planteadas en el capítulo anterior utilizando ahora la prueba χ^2 .

9.2 Notación

Las tablas de contingencia son tablas que reflejan las frecuencias (observadas o esperadas) de las posibles combinaciones de dos variables cualitativas.

En la tabla de contingencia la notación utilizada es la siguiente:

- El valor O_{ij} es el valor **observado** en la fila i y columna j .
- El valor E_{ij} es el valor **esperado** en la fila i y columna j .
- $n_{i.}$ es el total en la fila i .
- $n_{.j}$ es el total en la columna j .
- n es total de efectivos.

Con esta misma terminología se pueden construir tablas de **m**x**n** (es decir, tablas con “m” filas y “n” columnas).

		Variable 2		
Variable 1		Categoría 1	Categoría 2	Total
	Categoría 1	O_{11}	O_{12}	$n_{1.}$
	Categoría 2	O_{21}	O_{22}	$n_{2.}$
	Total	$n_{.1}$	$n_{.2}$	n

Tabla 9.1: Tabla de valores observados

		Variable 2		
Variable 1		Categoría 1	Categoría 2	Total
	Categoría 1	E_{11}	E_{12}	$n_{1.}$
	Categoría 2	E_{21}	E_{22}	$n_{2.}$
	Total	$n_{.1}$	$n_{.2}$	n

Tabla 9.2: Tabla de valores esperados

f : número de filas

c : número de columnas

9.3 Fórmulas

- Tabla esperada

$$E_{i,j} = \frac{n_{i.} \times n_{.j}}{n}$$

- χ^2 (ji cuadrado)

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

9.4 Descripción de la base de datos

La siguiente tabla presenta la descripción de todas las variables incluidas en la hoja **estudioCaMama** del fichero **datosProblemasBest.xlsx**:

Nombre	Descripción	Codificación
id	Número de identificación	Numérica
caso	Caso de cáncer de mama	0: No cáncer de mama 1: Cáncer de mama
AnovOrales	Toma de anovulatorios orales (AO)	0: No 1: Si

Nombre	Descripción	Codificación
THS	Toma de terapia hormonal substitutiva	0: No 1: Si
AFamCancer	Antecedentes familiares de cáncer de mama	0: No 1: Familiar 1er grado 2: Familiar 2o grado 3: Otro familiar
Menopausia	Menopausia	0: No 1: Si
Menar11	Menarquia antes de los 11 años	0: No 1: Si
Hijos_cat	Número de hijos en 3 categorías	0: Sin hijos 1: 1 hijo 2: 2 o más hijos

Todos los problemas que planteamos a continuación pueden resolverse utilizando **Excel** o **Stata**.

9.5

PROBLEMAS

PROBLEMA 9.1

Se quiere valorar si el consumo de anticonceptivos orales (AnovOrales) se asocia al desarrollo de cáncer de mama.

Pregunta 9.1. Utilizando la base de datos en Excel, complete la tabla de valores observados (tabla reftab:tabla91) que muestra las proporciones observadas en el estudio de consumo de AO en casos (mujeres con cáncer de mama) y controles (mujeres sin cáncer de mama).

	Mujer con cáncer	Mujer sin cáncer	Total
Consumo AO > 1 año			
Consumo AO < 1 año			
Total			

Tabla 9.4: Relación entre consumo de AO y desarrollo de cáncer de mama

Una vez que tenga la tabla observada, complete la tabla 9.5 (de valores esperados) que muestra las proporciones esperadas de consumo de AO en los casos (mujeres con cáncer de mama) y en los controles (mujeres sin cáncer de mama) si el consumo de AO y el cáncer de mama fueran variables independientes.

	Mujer con cáncer	Mujer sin cáncer	Total
Consumo AO > 1 año			
Consumo AO < 1 año			
Total			

Tabla 9.5: Relación entre consumo de AO y desarrollo de cáncer de mama en caso de que ambas variables fueran independientes (tabla esperada)

A partir de las dos tablas anteriores responda las siguientes preguntas.

Pregunta 9.2. ¿Cuál es el valor del estadístico de contraste χ^2 ? Compare el resultado obtenido con el del problema 8.1.

Pregunta 9.3. ¿Cuál es la significación obtenida en el contraste de hipótesis?

Pregunta 9.4. ¿Cuál es el valor del estadístico a partir del cual se obtendría una asociación estadísticamente significativa entre el consumo de anovulatorios orales y el desarrollo de cáncer de mama?

Pregunta 9.5. Una vez realizado el contraste de hipótesis, ¿cuál es la conclusión del estudio?

PROBLEMA 9.2

Imagine ahora que lo que nos interesa es evaluar si el consumo de terapia hormonal sustitutiva (THS) se asocia al desarrollo de cáncer de mama. Para ello responda a las siguientes preguntas con los datos de la base de datos.

Pregunta 9.6. Calcule la proporción de consumidoras de THS que hay entre las mujeres con y sin cáncer de mama.

Pregunta 9.7. Compruebe si dicha proporción (consumidoras de THS) es distinta en las mujeres con cáncer de mama que en las controles.

Pregunta 9.8. A la vista de los resultados, ¿se asocia el consumo de THS al desarrollo de cáncer de mama?

PROBLEMA 9.3

Se quiere valorar si los antecedentes familiares de cáncer de mama se asocian al desarrollo de cáncer de mama. Para ello se ha recogido una variable en la base de datos que identifica a las mujeres sin antecedentes (0) y diferencia a las que tienen antecedentes en 3 grupos según el grado de parentesco (1: familiar de 1^{er} grado, 2: de 2^o grado y 3: otro familiar).

Pregunta 9.9. Construya a partir de la base de datos las tablas esperada y observada de la relación antecedentes familiares y cáncer de mama.

Pregunta 9.10. A la vista de las tablas, ¿se cumplen los criterios para evaluar si ambas variables se asocian mediante una prueba χ^2 ? Razone su respuesta.

Pregunta 9.11. ¿Cuántos grados de libertad tiene la prueba χ^2 ?

Pregunta 9.12. ¿Cuál es el valor del estadístico de contraste?

Pregunta 9.13. ¿Cuál es el valor de p?

Pregunta 9.14. ¿Cuál es la conclusión que se saca de este estudio?

PROBLEMA 9.4

En el capítulo anterior, con esta misma base de datos, se comprobó si las mujeres con cáncer y las del grupo control se diferencian en cuanto a haber tenido hijos. Para avanzar un poco más en este análisis se pretende ahora comprobar si hay diferencias en cuanto al número de hijos, categorizando esta variable en 3 grupos: 0: mujeres sin hijos, 1: mujeres con 1 hijo y 2: mujeres con 2 o más hijos.

Pregunta 9.15. ¿Cuál es el valor del estadístico χ^2 en este contraste de hipótesis?

Pregunta 9.16. ¿Cuál es el valor p?

Pregunta 9.17. ¿Cuál es el valor del estadístico a partir del cual se obtendría una asociación estadísticamente significativa entre el número de hijos y el desarrollo de cáncer de mama?

Pregunta 9.18. ¿Cuál es la conclusión del análisis?

Cálculo del tamaño de muestra

10

10.1 Introducción

Muchas veces el estudiante se puede plantear por qué en una investigación se han usado 150 personas en lugar de 200 o qué cosas influyen en que el tamaño de una muestra sea mayor o menor. A través de los ejercicios de este capítulo el estudiante practicará el cálculo del tamaño muestral de:

- Estimación de una proporción
- Estimación de una media
- Comparación de dos proporciones
- Comparación de dos medias
- Cálculo de la potencia del estudio

Para calcular el número necesario de sujetos que deben incluirse en un estudio se debe asumir básicamente 4 supuestos:

1. Probabilidad de cometer un error tipo 1 (riesgo alfa)
2. Probabilidad de cometer un error tipo 2 (riesgo beta)
3. Desviación estándar que se espera que tenga la variable que se comparará o raíz de $p(1-p)$ en el caso de proporciones
4. Diferencia mínima que se desea detectar en la comparación “magnitud del efecto”. Este valor es igual a la mitad de la amplitud del intervalo de confianza.

Procedimiento

FÓRMULAS cálculo del tamaño muestral

Concepto	Proporciones	Medias
Estimación de una proporción o media	$n = \frac{(z_{\frac{\alpha}{2}} + z_{\beta})^2 p(1-p)}{d^2}$	$n = \frac{(z_{\alpha/2} + z_{\beta})^2 s^2}{d^2}$
Comparación de dos proporciones o dos medias	$n = \frac{2(z_{\frac{\alpha}{2}} + z_{\beta})^2 p(1-p)}{d^2}$	$n = \frac{2(z_{\alpha/2} + z_{\beta})^2 s^2}{d^2}$
Cálculo de la potencia de un estudio	$z_{\beta} = \sqrt{\frac{nd^2}{2p(1-p)}} - z_{\alpha/2}$	$z_{\beta} = \sqrt{\frac{nd^2}{2s^2}} - z_{\alpha/2}$

* d es la precisión con la que quiero estimar

Valores utilizados frecuentemente en la distribución normal para $z_{\frac{\alpha}{2}}$ (2colas) o para z_{β} (1 cola)

α	$z_{\frac{\alpha}{2}}$	β	z_{β}
0.1	1.64	0.2	0.85
0.05	1.96	0.1	1.28
0.01	2.58	0.05	1.65



10.2



PROBLEMAS

PROBLEMA 10.1

Se va a realizar un estudio para conocer el porcentaje de escolares obesos en Cantabria. Por estudios previos, se espera que el resultado sea próximo al 15 %. Se desea que la nueva estimación tenga un intervalo de confianza al 90 % y una precisión de $\pm 2\%$.

Pregunta 10.1. ¿Cuántos escolares deberán incluirse en la muestra?

PROBLEMA 10.2

Se quiere conocer con una precisión de 5 años e intervalo de confianza del 95 %, la edad media a la que se produce un infarto de miocardio. Un estudio previo indica que la varianza puede ser de 200.

Pregunta 10.2. ¿Cuántos pacientes con infarto se deberían incluir en el estudio?

PROBLEMA 10.3

Aparece un nuevo medicamento contra el SIDA, la tishavirina. Se decide comparar su eficacia con un medicamento ya aceptado: la ribavirina. Se selecciona un grupo de pacientes; a la mitad se les dará tishavirina y a la otra mitad ribavirina. Se sabe que en el 70 % de los pacientes tratados con ribavirina la evolución de la enfermedad se detiene. Se considerará que la tishavirina es mejor que la ribavirina si consigue al menos un 3 % más de detenciones.

Pregunta 10.3. ¿Cuántos pacientes habrá que incluir en el estudio para realizarlo con un intervalo de confianza del 95 % y con una potencia del 90 %?

PROBLEMA 10.4

El tratamiento del enanismo hipofisario con hormona del crecimiento humana (hGH) permite ganar 15 centímetros de altura con una desviación estándar de 10. Aparece una nueva hormona modificada (mGH) y se realiza un estudio comparando las dos. Se asume que mGH será mejor si consigue al menos 5 centímetros más que hGH.

Pregunta 10.4. ¿Cuántos pacientes habrá que incluir para realizar el estudio con un intervalo de confianza al 90 % y una potencia del 90 %?

PROBLEMA 10.5

En el problema 10.4, los investigadores sólo pudieron reclutar 120 pacientes en total.

Pregunta 10.5. ¿Qué potencia estadística tuvo el estudio?

PROBLEMA 10.6

Con el siguiente problema se pretende que reflexione sobre cómo cambia el tamaño muestral al modificar los diferentes parámetros que integran la fórmula. Para ello previamente debe calcular el número de partos a incluir en nuestro estudio para conocer la frecuencia de síndrome de bajo peso al nacer.

Pregunta 10.6. ¿Cuál es el tamaño si la frecuencia que obtendremos es 0.2, el error α es del 5% y la precisión o diferencia detectable será de un 4%

Pregunta 10.7. ¿En cuánto aumentaría este tamaño si la frecuencia esperada fuese de 0.8?

Pregunta 10.8. ¿En cuánto aumentaría este tamaño si la frecuencia esperada fuese de 0.5?

Pregunta 10.9. ¿En cuánto aumentaría este tamaño si el error $\alpha=1\%$? (Error a $Z_{\frac{\alpha}{2}} = 2.6$)

Pregunta 10.10. ¿En cuánto aumentaría este tamaño si la precisión fuese de un 2%?

Pregunta 10.11. Con los resultados obtenidos en este problema cumplimente el número de partos necesario en las distintas situaciones planteadas. A continuación, saca una conclusión sobre cómo cambian las estimaciones del tamaño muestral en función de la frecuencia esperada, el error α y la precisión.

Tamaño muestral (partos)	—	—	—	—	—
Frecuencia esperada	0.2	0.8	0.5	0.2	0.2
Error α	5%	5%	5%	1%	5%
Precisión	4%	4%	4%	4%	2%

11

Regresión y correlación

11.1 Introducción

En este capítulo se abordarán las técnicas de correlación y de regresión lineal simple como método para evaluar la existencia de relación lineal entre dos variables continuas. Se utilizarán los datos de un estudio ecológico llevado a cabo con el objetivo de valorar la influencia de distintas variables sobre la puntuación obtenida en el informe PISA. El estudio está inspirado en el artículo del *New England Journal of Medicine* (Messerli FH. Chocolate consumption, cognitive function, and Nobel laureates. *N Engl J Med.* 2012; 367: 1562–1564), que recomendamos fuertemente leer y que está disponible en: biostat.jhsph.edu

11.2 Descripción del estudio

Para alcanzar los objetivos planteados se diseñó un estudio en el que se recogió información sobre el producto interior bruto de 30 países y distintas variables potencialmente asociadas al PIB. La información se obtuvo visitando varias páginas web el 19 de abril de 2013. El producto interior bruto a valores de paridad de poder adquisitivo per cápita se ha obtenido de las estimaciones del Fondo Monetario Internacional en: es.wikipedia.org-PIB, los resultados del informe PISA de 2009 se obtuvieron en es.wikipedia.org-PISA, y el consumo per cápita de azúcar refinado se consiguió en caobisco.eu, theobroma-cacao.de, y chocosuisse.ch.

Una vez obtenida la información se calculó la variable PISA con el valor medio de los tres informes PISA (en ciencias, matemáticas y lectura).

11.3 Notación

- r : coeficiente de correlación lineal
- R^2 : coeficiente de determinación
- $\text{Cov}(x, y)$: covarianza entre las variables x , y

- β_0 : intersección en el origen
- $\widehat{\beta}_0$: estimador de la intersección en el origen
- β_1 : pendiente de la recta
- $\widehat{\beta}_1$: estimador de la pendiente de la recta
- m_x : media muestral de x
- m_y : media muestral de y
- s_x : desviación típica muestral de x
- s_y : desviación típica muestral de y

11.4 Fórmulas

- Covarianza (x, y):

$$\text{Cov}(x, y) = \frac{\sum_{i=1}^n (x_i - m_x)(y_i - m_y)}{n - 1}$$

- Coeficiente de correlación lineal:

$$r = \frac{\text{Cov}(x, y)}{s_x \times s_y}$$

- Coeficiente de determinación: $R^2 = r^2$

- Pendiente de la recta:

$$\widehat{\beta}_1 = r \times \frac{s_y}{s_x}$$

- Intersección en el origen:

$$\widehat{\beta}_0 = m_y - \widehat{\beta}_1 \times m_x$$

- Recta de regresión:

$$y = \widehat{\beta}_0 + \widehat{\beta}_1 x$$

11.5 Descripción de la base de datos

La siguiente tabla presenta la descripción de todas las variables incluidas en la hoja **estudioPIB** del fichero **datosProblemasBest.xlsx**:

Nombre	Descripción	Codificación
país	País	Cadena
poblacion	Número de habitantes en 2012	Numérica

Nombre	Descripción	Codificación
pib	Producto interior bruto per cápita (en miles de dólares)	Numérica
consumochocolate	Consumo anual de chocolate per cápita (en Kg)	Numérica
consumoazucar	Consumo anual de azúcar per cápita (en Kg)	Numérica
pisamate	Puntuación en el informe PISA 2009 en matemáticas	Numérica
pisaciencia	Puntuación en el informe PISA 2009 en ciencias	Numérica
psalectura	Puntuación en el informe PISA 2009 en habilidades de lectura	Numérica
nobel	Número de premios nobel a lo largo de la historia	Numérica

Todos los problemas que se plantean a continuación pueden resolverse utilizando Excel o Stata.

11.6

PROBLEMAS

PROBLEMA 11.1

Se quiere valorar la influencia del PIB sobre los resultados de un país en el PISA. La media del PIB de los países participantes fue de 20.72 mil dólares ($s = 12.09$), mientras que la puntuación PISA obtuvo una media de 493.12 puntos ($s = 33.69$). Se sabe que la covarianza entre ambas variables es $\text{cov}_{(x,y)} = 144.42$ puntos · mil dólares.

Pregunta 11.1. ¿Cuánto vale el coeficiente de correlación lineal de la relación entre ambas variables? ¿Cómo se interpreta?

Pregunta 11.2. ¿Qué porcentaje de la nota obtenida en el PISA se puede explicar por PIB del país?

Pregunta 11.3. ¿Cuánto cambia la nota media del PISA por cada 1000 dólares más de PIB?

Pregunta 11.4. Es imposible que un país tenga PIB = 0 dólares. Pero si fuera posible, ¿cuál sería su valor esperado del PISA?

Pregunta 11.5. La siguiente tabla muestra las puntuaciones medias y las desviaciones estándar de cada una de las tres áreas de conocimiento evaluadas en el PISA (matemáticas, ciencias y lectura):

	Media	Desviación estándar
PISA ciencias	498.1	34.4
PISA matemáticas	492.4	38.6
PISA lectura	488.9	29.8

Al calcular las covarianzas (x,y) de cada una de estas áreas con el PIB obtenemos que:

$$\text{Cov}(\text{PIB}, \text{pisaciencias}) = 139$$

$$\text{Cov}(\text{PIB}, \text{pisamates}) = 150.63$$

$$\text{Cov}(\text{pisalectura}, \text{PIB}) = 143.62$$

Calcule de nuevo los parámetros solicitados en las preguntas 11.2 a 11.4 para los resultados de estas tres áreas de conocimiento evaluadas en el pisa.

PROBLEMA 11.2

Se quiere conocer en qué medida los resultados en las tres áreas de conocimiento evaluadas presentan correlación entre sí. Sabiendo que las covarianzas entre estas áreas son:

$$\text{Cov}(\text{pisamate}, \text{pisaciencias}) = 1278.82$$

$$\text{Cov}(\text{pisamate}, \text{pisalectura}) = 1075.23$$

$$\text{Cov}(\text{pisalectura}, \text{pisaciencia}) = 975.32$$

y utilizando los datos de la pregunta 11.5 responda a las siguientes cuestiones.

Pregunta 11.6. ¿Cuál es el valor del coeficiente de correlación lineal entre la puntuación PISA en matemáticas y en ciencia? Interprete el resultado.

Pregunta 11.7. ¿Cuál es el valor del coeficiente de correlación lineal entre la puntuación PISA en matemáticas y en lectura? Interprete el resultado.

Pregunta 11.8. ¿Cuál es el valor del coeficiente de correlación lineal entre la puntuación PISA en ciencias y en lectura? Interprete el resultado.

PROBLEMA 11.3

En el mismo estudio planteado al inicio del capítulo, se quiso evaluar la influencia del PIB del país sobre el consumo de azúcares refinados. La siguiente tabla resume la media y desviación de ambas variables y la covarianza entre ellas:

	Media	Desviación estándar
PIB	29.71	12.09
Consumo azúcar	5.82	7.86
<i>Cov(PIB, consumo azúcar)</i>	-2.23	

Pregunta 11.9. ¿Hay una relación lineal entre el PIB del país y el consumo de azúcares refinados?

Pregunta 11.10. Dibuje la recta de regresión entre ambas variables.

Pregunta 11.11. ¿Qué porcentaje del consumo de azúcar se puede explicar por el PIB del país?

Pregunta 11.12. ¿Cuánto cambia el consumo medio de azúcar por cada 1000 dólares de PIB?

PROBLEMA 11.4

Se sospecha que, para suplir malas puntuaciones PISA en lectura, ciertos países incentivan el consumo de chocolate entre sus ciudadanos con la esperanza de producir más premios Nobel. Investigue la relación entre consumo de chocolate y número de premios Nobel en los 15 países con menor puntuación PISA en lectura.

Pregunta 11.13. Sabiendo que, en los 15 países con menor puntuación, la covarianza entre el consumo de chocolate y el número de premios nobel es 50.74 ¿qué porcentaje de la variabilidad del número de premios Nobel es atribuible al consumo de chocolate en países con baja puntuación PISA en lectura?

La siguiente tabla resume la media y desviación del consumo de chocolate y del número de premios nobeles en estos 15 países:

	Media	Desviación estándar
Número de premios Nobel	12.60	30.18
Consumo chocolate	3.65	2.30

Pregunta 11.14. ¿Cuál sería el número de premios Nobel esperado si el consumo de chocolate se correspondiera con el consumo medio de los 15 países con mayor puntuación PISA en lectura?

Pregunta 11.15. A la luz de los resultados, uno de los países decide regalar una tableta de chocolate de 125 g a cada habitante a principios de año y así tratar de mejorar la puntuación PISA en lectura. ¿Cuántos premios Nobel adicionales se espera encontrar con respecto al escenario en que no hubiera invertido en chocolate?

PROBLEMA 11.5

La siguiente tabla resume la media y desviación de consumo de chocolate y número de premios nobel en los 30 países del estudio:

	Media	Desviación estándar
Número de premios Nobel	28.23	64.99
Consumo chocolate	4.64	2.82
$Cov(\text{consumo azúcar}, \text{nobel})$	39.32	

Investigue ahora la relación entre consumo de chocolate y el número de premios Nobel considerando los datos de todos los países disponibles.

Pregunta 11.16. Con el consumo de chocolate español, ¿cuántos premios Nobel debería haber conseguido España?

Pregunta 11.17. ¿Cuál es el error que comete el modelo de regresión lineal con los datos españoles?

Pregunta 11.18. Entre dos países hay una diferencia de 3 Kg de consumo anual de chocolate per cápita. ¿Qué deberían tener en el número de premios Nobel?

12

Regresión lineal múltiple

12.1 Introducción

En este capítulo no se utilizará ninguna base de datos. Se harán problemas sobre:

- Interpretación de los coeficientes de un modelo de regresión lineal múltiple
 - Para variables continuas
 - Para variables categóricas
- Cálculo de intervalos de confianza de los coeficientes.
- Cálculo de valores de p.
- Interpretación de los términos de interacción.

12.2 Notación

- Modelo general:

$$y = \sum_{i=0}^k \beta_i x_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

, donde y es el efecto; $x_1, x_2, \dots, x_k$ son las variables de exposición. β_0 es la constante, $\beta_1, \beta_2, \dots, \beta_k$ son los parámetros de cada exposición. Los parámetros β no son conocidos y no es posible llegar a conocerlos. El modelo de regresión hace una estimación de estos parámetros.

- Los parámetros estimados se representarán como $\widehat{\beta}$. De esta forma, el modelo estimado será:

$$y = \sum_{i=0}^k \widehat{\beta}_i x_i + \varepsilon_j = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + \cdots + \widehat{\beta}_k x_k + \varepsilon_j$$

, donde ε_j es el error que el modelo comete para el paciente (o la observación) j .

- *Error estándar de $\widehat{\beta}_i$* : $EE(\widehat{\beta}_i)$
- Intervalo de confianza al 95 %: IC 95 %

Fórmulas fundamentales:

- Test de Wald para el parámetro β_i :

$$t = \frac{\widehat{\beta}_i}{EE(\widehat{\beta}_i)}$$

- Intervalo de confianza al 95 % de β_i :

$$\widehat{\beta}_i \mp 1.96 \times EE(\widehat{\beta}_i)$$



12.3



PROBLEMAS

PROBLEMA 12.1

Se siguió a 1242 pacientes para conocer la influencia de diferentes nutrientes sobre el nivel de colesterol LDL (medido en mg/dL). Se obtienen los siguientes resultados:

Nutriente	Unidades	$\widehat{\beta}$	$EE(\widehat{\beta})$
Hidratos de carbono	g/día	1.10	0.31
Ácidos grasos saturados	g/día	1.54	0.23
Ácidos grasos poliinsaturados	g/día	-0.83	0.43
Ácidos grasos monoinsaturados	g/día	-2.20	0.45
Constante	-	12.50	1.57

Pregunta 12.1. Complete la siguiente tabla.

Nutriente	$\widehat{\beta}$	IC 95 %	p
Hidratos de carbono	1.10		
Ácidos grasos saturados	1.54		
Ácidos grasos poliinsaturados	-0.83		
Ácidos grasos monoinsaturados	-2.20		
Constante	12.50		

Pregunta 12.2. ¿Qué unidades tiene el coeficiente $\widehat{\beta}$ de grasas saturadas?

Pregunta 12.3. Un paciente tiene los siguientes consumos diarios: 180 gramos de hidratos de carbono, 40 gramos de ácidos grasos saturados, 20 gramos de ácidos grasos poliinsaturados y 18 gramos de ácidos grasos monoinsaturados. ¿Qué nivel de colesterol LDL espera que tenga?

Pregunta 12.4. Dos pacientes A y B comen todos los días las mismas cantidades, salvo que A añade 20 g de aceite de oliva (ácidos grasos monoinsaturados) y quita 20 g de grasa de origen animal (ácidos grasos saturados). ¿Qué diferencia espera que tengan en sus concentraciones de colesterol LDL?

PROBLEMA 12.2

Se quiere conocer cómo influyen la edad (en años) y el nivel de estudios en el índice de masa corporal (IMC, en kg/m^2). El nivel de estudios se ha clasificado en cuatro categorías (sin estudios / primarios / secundarios / universitarios).

Pregunta 12.5. ¿Cuántas variables será necesario incluir en el modelo?

Pregunta 12.6. En la siguiente tabla se presenta en filas el nivel de estudios y en columnas las variables ficticias creadas para incluirlo en el modelo. Rellene las casillas restantes con los códigos 1 y 0, usando el nivel de estudios primarios como referencia.

Variables ficticias			
Nivel de estudios	estudio_1	estudio_2	estudio_3
Sin estudios			
Primarios			
Secundarios			
Universitarios			

Pregunta 12.7. Tras obtener el modelo de regresión lineal en el que se han introducido la edad y el nivel de estudios como variables x , decide hacer una representación gráfica de los resultados, poniendo edad en el eje X e índice de masa corporal en el eje Y. ¿Qué aspecto tendrá la figura?

Pregunta 12.8. ¿Qué unidades tendrá el coeficiente $\widehat{\beta}$ de la edad?

Pregunta 12.9. ¿Qué unidades tendrá el coeficiente $\widehat{\beta}$ de estudio_2?

PROBLEMA 12.3

Con la codificación que se indica en la solución de la pregunta 12.5, el modelo obtenido es:

Variable	$\widehat{\beta}$	EE(b)
Edad	0.21	0.08
estudios_1	1.37	0.41
estudios_2	-0.22	0.38
estudios_3	-0.93	0.40
Constante	15.21	0.30

Pregunta 12.10. Complete la siguiente tabla.

Variable	$\widehat{\beta}$	IC 95 %	p
Edad	0.21		
estudios_1	1.37		
estudios_2	-0.22		
estudios_3	-0.93		
Constante	15.21		

Pregunta 12.11. Interprete el coeficiente $\widehat{\beta}$ de la edad y su intervalo de confianza.

Pregunta 12.12. Interprete los coeficientes $\widehat{\beta}$ de las variables estudios_1, estudios_2 y estudios_3 y sus intervalos de confianza.

Pregunta 12.13. ¿Qué índice de masa corporal espera que tengan las siguientes personas?

- (A) Persona con 33 años y estudios universitarios
- (B) Persona con 45 años y estudios primarios
- (C) Persona con 28 años sin estudios
- (D) Persona con 70 años y estudios secundarios

PROBLEMA 12.4

Se obtiene un modelo de regresión lineal entre la tensión arterial sistólica (TAS) en mmHg (variable Y) y la edad en años y el consumo diario de sal en gramos (variables X); se utilizan solo pacientes con al menos 30 años. El resultado se presenta en la siguiente tabla.

Variable	$\widehat{\beta}$ (IC 95%)	p
Edad	2.5 (2.2, 2.8)	<0.001
Consumo de sal	8.0 (7.8, 8.2)	<0.001
Edad x consumo de sal	-0.10 (-0.14, -0.06)	<0.001
Constante	5.0 (4.7, 5.3)	<0.001

Pregunta 12.14. ¿Qué unidades tienen los coeficientes b de cada variable?

Variable	Unidades de $\widehat{\beta}$
Edad	
Consumo de sal	
Edad x consumo de sal	
Constante	

Pregunta 12.15. Considere un paciente de 40 años cuyo consumo de sal diario es 7 gramos. Si consigue que el consumo de sal baje a 5 gramos, ¿en cuánto bajará su tensión arterial sistólica?

Pregunta 12.16. Considere ahora un paciente de 70 años cuyo consumo de sal diario es 7 gramos. Si consigue que el consumo de sal baje a 5 gramos, ¿en cuánto bajará su tensión arterial sistólica?

Pregunta 12.17. Tiene que organizar un programa para reducir la tensión arterial y quiere centrarse en los grupos en los que las medidas que tome sean más efectivas. ¿En qué edades será más importante recomendar la disminución del consumo de sal?

PROBLEMA 12.5

Se elabora un modelo de regresión lineal entre el nivel de hemoglobina en sangre (variable Y, medida en g/dL) y la edad (en años) y el sexo (codificado como mujer = 1, varón = 0). Se obtiene el siguiente resultado, al que llamaremos modelo 1.

Variable	$\widehat{\beta}$ (IC 95 %)
Edad	-0.010 (-0.015, -0.005)
Sexo	-0.90 (-1.05, - 0.75)
Constante	14.1 (12.0, 12.2)

Posteriormente, los investigadores pensaron que podía haber una interacción entre edad y sexo, por lo que introdujeron un término más. El resultado -que llamaremos modelo 2- aparece en la tabla siguiente.

Variable	$\widehat{\beta}$ (IC 95 %)
Edad	-0.009 (-0.13, -0.005)
Sexo	-1.02 (-1.20, -0.84)
Edad x sexo	0.006 (0.004, 0.008)
Constante	13.9 (13.7, 14.1)

Pregunta 12.18. Si tuviera que representar el modelo 1 en un gráfico con Edad en el eje X y hemoglobina en el eje Y, ¿qué forma tendría el gráfico? (No se pide hacer el gráfico, sino saber qué forma tendrá).

Pregunta 12.19. Si tuviera que representar el modelo 2 en un gráfico con Edad en el eje X y hemoglobina en el eje Y, ¿qué forma tendría el gráfico? (No se pide hacer el gráfico, sino saber qué forma tendrá).

Pregunta 12.20. Con el modelo 1, indique la diferencia de hemoglobina que habría entre los pacientes A y B de la tabla:

Paciente A	Paciente B	Hemoglobina _A - Hemoglobina _B
Varón de 38 años	Mujer de 38 años	
Varón de 62 años	Mujer de 62 años	
Varón de 38 años	Varón de 62 años	
Mujer de 38 años	Mujer de 62 años	

Pregunta 12.21. Con el modelo 2, indique la diferencia de hemoglobina que habría



entre los pacientes A y B de la tabla:

Paciente A	Paciente B	Hemoglobina _A - Hemoglobina _B
Varón de 38 años	Mujer de 38 años	
Varón de 62 años	Mujer de 62 años	
Varón de 38 años	Varón de 62 años	
Mujer de 38 años	Mujer de 62 años	

PROBLEMA 12.6

El nivel de hemoglobina en sangre (g/dL) suele ser más bajo en pacientes con enfermedades crónicas, pero en mujeres suele ser más alto tras la menopausia. Se elabora un modelo para estudiar el efecto conjunto de estas variables sobre la hemoglobina. Los resultados aparecen en esta tabla.

Variable	$\widehat{\beta}$ (IC 95 %)
Posmenopausia (1=sí, 0=no)	0.8 (0.5, 1.1)
Enfermedad crónica (1=sí, 0=no)	-0.5 (-0.8, -0.2)
Posmenopausia x enfermedad crónica	0.2 (0.0, 0.4)
Constante	11.5 (11.3, 11.7)

Pregunta 12.22. En las mujeres premenopáusicas, ¿en cuánto disminuye la hemoglobina si tienen enfermedad crónica?

Pregunta 12.23. Y en las mujeres posmenopáusicas, ¿en cuánto disminuye la hemoglobina si tienen enfermedad crónica?

Soluciones

1

Estadística descriptiva

SOLUCIÓN 1.1

Solución con calculadora**Media:**

$$m = \frac{\sum_{i=1}^n x_i}{n}$$

$$m = \frac{246 + 259 + 236 + 278 + 270}{5} = 257.8$$

Desviación estándar:

$$S = \sqrt{\frac{\sum (x_i - m)^2}{n - 1}}$$

$$S = \sqrt{\frac{(246 - 257.8)^2 + (259 - 257.8)^2 + (236 - 257.8)^2 + (278 - 257.8)^2 + (270 - 257.8)^2}{5 - 1}}$$

$$= \sqrt{\frac{1172.8}{4}} = 17.12$$

Mediana:

- Ordenar de menor a mayor los valores de la variable DuraGest:

236 246 259 270 278

- La mediana es el valor que se encuentra en el medio de un conjunto de datos:
259

Solución con Excel

- **Media:** PROMEDIO(F2:F544)
Resultado: 264.88 días
- **Desviación estándar:** DESVEST.M(F2:F544)
Resultado: 19.39
- **Mediana:** MEDIANA(F2:F544)
Resultado: 270

Nota: F hace referencia a la columna con los datos de interés.

Solución con Stata

```
. tabstat DuraGest,stat(mean sd p50)
```

Variable	Mean	Sd	p50
DuraGest	264.8785	19.39243	270

SOLUCIÓN 1.2

Solución con calculadora**Media:**

$$m = \frac{\sum_{i=1}^n x_i}{n}$$

- **Media del edificio A** (Building = 1):

$$m_1 = \frac{28 + 44 + 28 + 29 + 42}{5} = 34.2$$

- **Media del edificio B** (Building = 2):

$$m_2 = \frac{28 + 31 + 40 + 28 + 40}{5} = 33.4$$

Desviación estándar (S):

$$S = \sqrt{\frac{\sum (x_i - m)^2}{n - 1}}$$

- **Desviación estándar del edificio A:**

$$\begin{aligned} S_1 &= \sqrt{\frac{(28 - 34.2)^2 + (44 - 34.2)^2 + (28 - 34.2)^2 + (29 - 34.2)^2 + (42 - 34.2)^2}{5 - 1}} \\ &= \sqrt{\frac{206.8}{4}} = 8.07 \end{aligned}$$

- **Desviación estándar del edificio B:**

$$\begin{aligned} S_2 &= \sqrt{\frac{(28 - 33.4)^2 + (31 - 33.4)^2 + (40 - 33.4)^2 + (28 - 33.4)^2 + (40 - 33.4)^2}{5 - 1}} \\ &= \sqrt{\frac{151.2}{4}} = 6.15 \end{aligned}$$

Mediana:

- Ordena de menor a mayor los valores de la variable AgeMother en cada uno de los grupos:
 - Edificio A
28 28 29 42 44
 - Edificio B
28 28 31 40 40
- Resultado:
 - Mediana edificio A = 29
 - Mediana edificio B = 31

Solución con Excel

Definir etiquetas *AgeMother*. El material suplementario incluye un vídeo mostrando el proceso completo para “Etiquetar variables en Excel”.

- **Media:**
 - PROMEDIO(AgeMother_A)
Resultado Edificio A: 34.8 años
 - PROMEDIO(AgeMother_B)
Resultado Edificio B: 31.86 años
- **Desviación estándar:**
 - DESVEST.M(AgeMother_A)
Resultado edificio A: 5.85
 - DESVEST.M(AgeMother_B)
Resultado edificio B: 5.56
- **Mediana:**
 - MEDIANA(AgeMother_A)
Resultado edificio A: 36 días
 - MEDIANA(AgeMother_B)
Resultado edificio B: 33 días

Solución con Stata■ **Edificio A:**

```
. tabstat AgeMother if Building == 1, stat(mean sd p50)
```

Variable	Mean	Sd	p50
AgeMother	34.7972	5.854444	36

■ **Edificio B:**

```
. tabstat AgeMother if Building == 2, stat(mean sd p50)
```

Variable	Mean	Sd	p50
AgeMother	31.85992	5.556161	33

SOLUCIÓN 1.3

Solución con calculadora

Calcular la duración de la gestación en semanas:

DuraGest (semanas)
246/7 = 35.14
259/7 = 37
236/7 = 33.71
278/7 = 39.71
270/7 = 38.57

$$\text{Prematuros (\%)} = \frac{N_{\text{prematuros}}}{N} \times 100 = \frac{2}{5} \times 100 = 40\%$$

Solución con Excel

1. Generar variable DuraGest_semana: F2/7
Siendo F la columna donde se encuentra la variable de interés.
2. Autorrellenar. El material suplementario incluye un vídeo mostrando el proceso completo para “Autorrellenar celdas”
3. Porcentaje de niños prematuros:

$$(\text{CONTAR.SI}(L2:L544; "<37") / \text{CONTAR}(L2:L544)) * 100$$

La columna L contiene los datos de la variable *DuraGest_semanas*.
Resultado: 29.83

Solución con Stata

```
. gen prematuro=DuraGest/7 <37 if DuraGest!=.
. label define sino 0 "no" 1 "si"
. label values prematuro sino
. tab prematuro
```

prematurado	Freq.	Percent	Cum.
no	381	70.17	70.17
si	162	29.83	100.00

SOLUCIÓN 1.4

Solución con calculadora
Media:

$$m = \frac{\sum_{i=1}^n x_i}{n}$$

- **Media general:**

$$m = \frac{35.14 + 37 + 33.71 + 39.71 + 38.57 + 32.14 + 39.29 + 35.14 + 33.86 + 35.86}{10} = 36.04$$

- **Media del edificio A (Building = 1):**

$$m_1 = \frac{35.14 + 32.14 + 35.14 + 33.86 + 35.86}{5} = 34.43$$

- **Media del edificio B (Building = 2):**

$$m_2 = \frac{37 + 33.71 + 39.71 + 38.57 + 39.29}{5} = 37.66$$

Desviación estándar:

$$S = \sqrt{\frac{\sum (x_i - m)^2}{n - 1}}$$

- **Desviación estándar general:**

$$S = \sqrt{\frac{(35.1 - 36)^2 + (37 - 36)^2 + (33.7 - 36)^2 + (39.7 - 36)^2 + (38.6 - 36)^2 + (32.1 - 36)^2 + (39.3 - 36)^2 + (35.1 - 36)^2 + (33.9 - 36)^2 + (35.9 - 36)^2}{10 - 1}} = \sqrt{\frac{58.88}{9}} = 2.55$$

- **Desviación estándar del edificio A:**

$$S_1 = \sqrt{\frac{(35.14 - 34.43)^2 + (32.14 - 34.43)^2 + (35.14 - 34.43)^2 + (33.86 - 34.43)^2 + (35.86 - 34.43)^2}{5 - 1}} = \sqrt{\frac{8.62}{4}} = 1.47$$

- **Desviación estándar del edificio B:**

$$S_2 = \sqrt{\frac{(37 - 37.66)^2 + (33.71 - 37.66)^2 + (39.71 - 37.66)^2 + (38.57 - 37.66)^2 + (39.29 - 37.66)^2}{5 - 1}} = \sqrt{\frac{23.726}{4}} = 2.44$$

Mediana:

- Ordenar de menor a mayor los valores de la variable *DuraGest_semana*:
 - General:
32.14 33.71 33.86 35.14 35.14 35.86 37 38.57 39.29 39.71
 - Edificio A:
32.17 33.86 35.14 35.14 35.86
 - Edificio B:
33.71 37 38.57 39.29 39.71
- Resultado:
 - Mediana general = $\frac{35.14+35.86}{2} = 35.5$
 - Mediana Edificio A = 35.14
 - Mediana Edificio B = 38.57

Solución con Excel

Generar tres etiquetas sobre la variable *DuraGest_semana*. El material suplementario incluye un vídeo mostrando el proceso completo para “Etiquetar variables en Excel”.

- Cohorte general (*DuraGest*)
- Mujeres del edificio A (*DuraGest_A*)
- Mujeres del edificio B (*DuraGest_B*)

Medias:

- `PROMEDIO(DuraGest)`
Resultado media general: 37.84 semanas
- `PROMEDIO(DuraGest_A)`
Resultado edificio A: 36.43 semanas
- `PROMEDIO(DuraGest_B)`
Resultado edificio B: 39.41 semanas

Desviaciones estándar:

- `DESVEST.M(DuraGest)`
Resultado desviación estándar general: 2.77

- DESVEST.M(DuraGest_A)

Resultado edificio A: 2.78

- DESVEST.M(DuraGest_B)

Resultado edificio B: 1.71

Medianas:

- MEDIANA(DuraGest)

Resultado general: 38.57 semanas

- MEDIANA(DuraGest_A)

Resultado edificio A: 36.71 semanas

- MEDIANA(DuraGest_B)

Resultado edificio B: 39.86 semanas

Solución con Stata

- **Cohorte completa:**

```
. tabstat DuraGest_semana, stat(mean sd p50)
```

variable	mean	sd	p50
DuraGest_semana	37.83978	2.770347	38.57143

- **Edificio A:**

```
. tabstat DuraGest_semana if Building == 1, stat(mean sd p50)
```

variable	mean	sd	p50
DuraGest_semana	36.42707	2.783602	36.71429

- **Edificio B:**

```
. tabstat DuraGest_semana if Building == 2, stat(mean sd p50)
```

variable	mean	sd	p50
DuraGest_semana	39.4119	1.706467	39.85714

SOLUCIÓN 1.5

Solución con calculadora

- **n:** 10

- **Suma:**

$$\sum_{i=1}^n x_i = 11.3 + 10.8 + 12.9 + 12 + 10.4 + 11.6 + 13.2 + 11.3 + 12.3 + 10.8$$

$$= 116.6$$

- **Media:**

$$m = \frac{\sum_{i=1}^n x_i}{n}$$

$$= \frac{11.3 + 10.8 + 12.9 + 12 + 10.4 + 11.6 + 13.2 + 11.3 + 12.3 + 10.8}{10}$$

$$= 11.66$$

- **Mediana:**

$$\text{Mediana} = \frac{11.3 + 11.6}{2} = 11.45$$

- **Moda:** 10.8 y 11.3

- **Desviación estándar (S):**

$$S = \sqrt{\frac{\sum (x_i - m)^2}{n - 1}}$$

$$s = \sqrt{\frac{(11.3 - 11.6)^2 + (10.8 - 11.6)^2 + (12.9 - 11.6)^2 + (12 - 11.6)^2 + (10.4 - 11.6)^2 + (11.6 - 11.6)^2 + (13.2 - 11.6)^2 + (11.3 - 11.6)^2 + (12.3 - 11.6)^2 + (10.8 - 11.6)^2}{10 - 1}}$$

$$= \sqrt{\frac{7.7604}{9}} = 0.92$$

- **Varianza (S^2):** $S^2 = 0.92^2 = 0.86$

- **Valor mínimo:** 10.4

- **Valor máximo:** 13.2

- **Rango:** $13.2 - 10.4 = 2.8$

Solución con Excel

- **N:** CONTAR(K2:K544)
Resultado: 543
- **Suma:** SUMA(K2:K544)
Resultado: 6623.3
- **Media:** PROMEDIO(K2:K544)
Resultado: 12.20
- **Mediana:** MEDIANA(K2:K544)
Resultado: 12.3
- **Moda:** MODA.UNO(K2:K544)
Resultado: 13.2
- **Desviación estándar:** DESVEST.M(K2:K544)
Resultado: 1.06
- **Varianza:** VAR.S(K2:K544)
Resultado: 1.13
- **Valor mínimo:** MIN(K2:K544)
Resultado: 9.8
- **Valor máximo:** MAX(K2:K544)
Resultado: 13.8
- **Rango:** MAX(K2:K544) - MIN(K2:K544)
Resultado: 4

Nota: La columna K contiene los datos de interés.

Solución con Stata

```
. tabstat Hb, statistics (n sum mean p50 sd variance min max range)
```

variable	N	sum	mean	p50	sd	variance	min	max	range
Hb	543	6623.3	12.19761	12.3	1.063	1.130123	9.8	13.8	4

SOLUCIÓN 1.6

Solución con calculadora

$$\text{Peso_Bajo (\%)} = \frac{N \text{ neonatos con bajo peso}}{N} \times 100 = \frac{1}{10} \times 100 = 10\%$$

$$\text{Peso_Normal (\%)} = \frac{N \text{ neonatos con peso normal}}{N} \times 100 = \frac{8}{10} \times 100 = 80\%$$

$$\text{Peso_Alto (\%)} = \frac{N \text{ neonatos con peso alto}}{N} \times 100 = \frac{1}{10} \times 100 = 10\%$$

Solución con Excel■ **Peso menor de 2500g:**

`(CONTAR.SI(J2:J544;"<=2500")/CONTAR(J2:J544))*100`

Resultado: 15.10%

■ **Peso normal (>2500g y <4000g):**

`((CONTAR.SI(J2:J544;">2500")-CONTAR.SI(J2:J544;">4000"))/CONTAR(J2:J544))*100`

Resultado: 79.74%

■ **Peso mayor de 4000g:**

`(CONTAR.SI(J2:J544;">=4000")/CONTAR(J2:J544))*100`

Resultado: 5.16%

Nota: La columna J tiene los datos de interés.

Solución con Stata

```
. recode WeighNB 4000/max=3 min/2500=1 2500/4000=2,generate(peso_cat)
. label define peso_cat 1 "bajo" 2 "normal" 3 "alto"
. label values peso_cat peso_cat
```

```
. tab peso_cat
```

peso_cat	Freq.	Percent	Cum.
bajo	82	15.10	15.10
normal	433	79.74	94.84
alto	28	5.16	100.00
Total	543	100.00	

SOLUCIÓN 1.7

Solución con calculadora**■ Porcentaje (%) Mujeres ≤ 34 años**

$$\begin{aligned}\% \text{ Cohorte} &= \frac{N \text{ Mujeres } \leq 34 \text{ aos en cohorte}}{\text{Total}} \times 100 \\ &= \frac{6}{10} \times 100 = 60\%\end{aligned}$$

$$\begin{aligned}\% \text{ Edificio A} &= \frac{N \text{ Mujeres } \leq 34 \text{ aos en edificio A}}{\text{Total}} \times 100 \\ &= \frac{3}{5} \times 100 = 60\%\end{aligned}$$

$$\begin{aligned}\% \text{ Edificio B} &= \frac{N \text{ Mujeres } \leq 34 \text{ aos en edificio B}}{\text{Total}} \times 100 \\ &= \frac{3}{5} \times 100 = 60\%\end{aligned}$$

■ Porcentaje (%) Mujeres > 39 años

$$\begin{aligned}\% \text{ Cohorte} &= \frac{N \text{ Mujeres } \geq 40 \text{ aos en la cohorte}}{\text{Total}} \times 100 \\ &= \frac{4}{10} \times 100 = 40\%\end{aligned}$$

$$\begin{aligned}\% \text{ Edificio A} &= \frac{N \text{ Mujeres } \geq 40 \text{ aos en edificio A}}{\text{Total}} \times 100 \\ &= \frac{2}{5} \times 100 = 40\%\end{aligned}$$

$$\begin{aligned}\% \text{ Edificio B} &= \frac{N \text{ Mujeres } \geq 40 \text{ aos en edificio B}}{\text{Total}} \times 100 \\ &= \frac{2}{5} \times 100 = 40\%\end{aligned}$$

Solución con Excel

■ Mujeres ≤ 34 años:

- Cohorte:

$$(\text{CONTAR.SI}(\text{C2:C544}; \leq 34) / \text{CONTAR}(\text{C2:C544})) * 100$$

Resultado: 54.7 %

- Edificio A:

$$(\text{CONTAR.SI.CONJUNTO}(\text{C2:C544}; \leq 34; \text{B2:B544}; 1) / \text{CONTAR.SI}(\text{B2:B544}; 1)) * 100$$

Resultado: 46.85 %

- Edificio B:

$$\text{CONTAR.SI.CONJUNTO}(\text{C2:C544}; \leq 34; \text{B2:B544}; 2) / \text{CONTAR.SI}(\text{B2:B544}; 2) * 100$$

Resultado: 63.42 %

■ Mujeres ≥ 40 años

- Cohorte:

$$(\text{CONTAR.SI}(\text{C2:C544}; \geq 40) / \text{CONTAR}(\text{C2:C544})) * 100$$

Resultado: 17.5 %

- Edificio A:

$$(\text{CONTAR.SI.CONJUNTO}(\text{C2:C544}; \geq 40; \text{B2:B544}; 1) / \text{CONTAR.SI}(\text{B2:B544}; 1)) * 100$$

Resultado: 25.17 %

- Edificio B:

$$(\text{CONTAR.SI.CONJUNTO}(\text{C2:C544}; \geq 40; \text{B2:B544}; 2) / \text{CONTAR.SI}(\text{B2:B544}; 2)) * 100$$

Resultado: 8.95 %

Nota: C y B hacen referencia a las columnas con los datos de interés.

Solución con Stata

```
. recode AgeMother 40/max=3 min/34=1 34/40=2, generate(AgeMother_cat)
. label define edad_cat 1 "<=34" 2 "35-39" 3 ">40"
```

```
. label values AgeMother_Cat edad_cat
. tab Building AgeMother_Cat, row
```

Building	AgeMoter_Cat			Total
	<=34	35-39	>=40	
B	163	71	23	257
	63.42	27.63	8.95	100.00
A	134	80	72	286
	46.85	27.97	25.17	100.00
Total	297	151	95	543
	54.70	27.81	17.50	100.00

SOLUCIÓN 1.8

Solución con calculadora

Porcentajes

- Anemia en el Edificio A:

$$\begin{aligned} Anemia_{EdificioA} (\%) &= \frac{N \text{ mujeres con anemia edificio A}}{n \text{ edificio A}} \times 100 \\ &= \frac{3}{5} \times 100 = 60\% \end{aligned}$$

- Anemia en el Edificio B:

$$\begin{aligned} Anemia_{EdificioB} (\%) &= \frac{N \text{ mujeres con anemia edificio B}}{n \text{ edificio B}} \times 100 \\ &= \frac{2}{5} \times 100 = 40\% \end{aligned}$$

Medias

$$m = \frac{\sum_{i=1}^n x_i}{n}$$

- Edificio A:

- Con anemia

$$m_{A_anemia} = \frac{10.4 + 10.3 + 10.4}{3} = 10.4 \frac{mg}{dL}$$

- Sin anemia

$$m_{A_no_anemia} = \frac{11.7 + 11.7}{2} = 11.7 \frac{mg}{dL}$$

- Edificio B:

- Con anemia

$$m_{B_anemia} = \frac{10.6 + 10.7}{2} = 10.65 \frac{mg}{dL}$$

- Sin anemia

$$m_{B_no_anemia} = \frac{13.3 + 12.6 + 13.6}{3} = 13.17 \frac{mg}{dL}$$

Desviaciones estándar

$$S = \sqrt{\frac{\sum (x_i - m)^2}{n - 1}}$$

■ Edificio A:

- Con anemia

$$S_{A_anemia} = \sqrt{\frac{(10.4 - 10.4)^2 + (10.3 - 10.4)^2 + (10.4 - 10.4)^2}{3 - 1}} \\ = 0.06$$

- Sin anemia

$$S_{A_no_anemia} = \sqrt{\frac{(11.7 - 11.7)^2 + (11.7 - 11.7)^2}{2 - 1}} = 0$$

■ Edificio B:

- Con anemia

$$S_{B_anemia} = \sqrt{\frac{(10.6 - 10.65)^2 + (10.7 - 10.65)^2}{2 - 1}} = 0.07$$

- Sin anemia

$$S_{B_no_anemia} = \sqrt{\frac{(13.3 - 13.17)^2 + (12.6 - 13.17)^2 + (13.6 - 13.17)^2}{3 - 1}} \\ = 0.51$$

Solución con Excel

1. Generar variable: SI(K2>=11;0;1)
La columna K contiene los datos de la variable Hb.
2. Autorrellenar el resto de las filas.
3. Generar cuatro etiquetas sobre la variable Hb. El material suplementario incluye un vídeo mostrando el proceso completo para “Etiquetar variables en Excel”.

Porcentajes Insertar una tabla dinámica.

	1		2	
Etiquetas de fila	Cuenta de Anemia	Cuenta de Anemia2	Cuenta de Anemia	Cuenta de Anemia2
0	237	82.87 %	215	83.66 %
1	49	17.13 %	42	16.34 %
Total	286	100.00 %	257	100.00 %

Tabla 1.2: Tabla dinámica. Distribución de las mujeres en función del edificio en el que se encuentran y de sus niveles de hemoglobina. En columnas está el edificio (1=A;2=B) y en filas la Anemia (0=no;1=si)

Medias

- Mujeres del edificio A sin anemia (Hb_A0): PROMEDIO(Hb_A0)
Resultado: 12.45 mg/dL mujeres del edificio A sin anemia
- Mujeres del edificio A con anemia (Hb_A1): PROMEDIO(Hb_A1)
Resultado: 10.42 mg/dL mujeres del edificio A con anemia
- Mujeres del edificio B sin anemia (Hb_B0): PROMEDIO(Hb_B0)
Resultado: 12.62 mg/dL mujeres del edificio B sin anemia
- Mujeres del edificio B con anemia (Hb_B1): PROMEDIO(Hb_B1)
Resultado: 10.70 mg/dL mujeres del edificio B con anemia

Desviaciones estándar

- Mujeres del edificio A sin anemia (Hb_A0): DESVEST.M(Hb_A0)
Resultado: 0.84 mujeres del edificio A sin anemia
- Mujeres del edificio A con anemia (Hb_A1): DESVEST.M(Hb_A1)
Resultado: 0.26 mujeres del edificio A con anemia
- Mujeres del edificio B sin anemia (Hb_B0): DESVEST.M(Hb_B0)
Resultado: 0.81 mujeres del edificio B sin anemia
- Mujeres del edificio B con anemia (Hb_B1): DESVEST.M(Hb_B1)
Resultado: 0.20 mujeres del edificio B con anemia

Solución con Stata

Crear categorías y asignar nombres

```
. recode Hb 11/max=0 min/11=1, generate(anemia2)
```

```
. label define sino 0 "no" 1 "si"
. label values anemia sino
```

Porcentajes

```
. tab anemia Building, col
```

anemia	Building		Total
	A	B	
no	237 82.87	215 83.66	452 83.24
si	49 17.13	42 16.34	91 16.76
Total	286 100.00	257 100.00	543 100.00

Media y desviación típica

- Mujeres edificio A sin anemia:

```
. sum Hb if anemia==0&Building==1
```

Variable	Obs	Mean	Std. Dev.	Min	Max
Hb	237	12.44515	.8371838	11	13.8

- Mujeres edificio A con anemia:

```
. sum Hb if anemia==1&Building==1
```

Variable	Obs	Mean	Std. Dev.	Min	Max
Hb	49	10.42449	.2586267	9.8	10.9

- Mujeres edificio B sin anemia:

```
. sum Hb if anemia==0&Building==2
```

Variable	Obs	Mean	Std. Dev.	Min	Max
Hb	215	12.62047	.8137543	11	13.8

- Mujeres edificio B con anemia:

```
. sum Hb if anemia==1&Building==2
```

Variable	Obs	Mean	Std. Dev.	Min	Max
Hb	42	10.70476	.1962482	10.3	10.9

Probabilidad condicionada e independencia de sucesos

SOLUCIÓN 2.1

♣ Pregunta 2.1**Solución con calculadora**

$$P(\text{asbesto}) = \frac{\text{Personas expuestas al asbesto}}{\text{Total personas en el estudio}} = \frac{415}{844} = 0.4917$$

♣ Pregunta 2.2**Solución con calculadora**

$$P(\text{mesotelioma}) = \frac{\text{Personas con mesotelioma}}{\text{Total personas en el estudio}} = \frac{417}{844} = 0.4941$$

♣ Pregunta 2.3**Solución con calculadora**

$$\begin{aligned} P(\text{asbesto} \cap \text{mesotelioma}) &= \frac{\text{Personas expuestas al asbesto y que tienen mesotelioma}}{\text{Total personas en el estudio}} \\ &= \frac{226}{844} = 0.2678 \end{aligned}$$

♣ Pregunta 2.4**Solución con calculadora**

$$\begin{aligned} P(\text{no asbesto} \cap \text{no mesotelioma}) &= \frac{\text{Personas no expuestas al asbesto y que no tienen mesotelioma}}{\text{Total personas en el estudio}} \\ &= \frac{238}{844} = 0.2820 \end{aligned}$$

♣ Pregunta 2.5

Solución con calculadora

$$\begin{aligned}
 P(\text{asbesto} \mid \text{mesotelioma}) &= \frac{P(\text{asbesto} \cap \text{mesotelioma})}{P(\text{mesotelioma})} \\
 &= \frac{\text{Personas expuestas al asbesto y que tienen mesotelioma}}{\text{Personas con mesotelioma}} \\
 &= \frac{226}{417} = 0.5420
 \end{aligned}$$

Nota: El desarrollo completo sería:

$$\begin{aligned}
 P(\text{asbesto} \mid \text{mesotelioma}) &= \frac{P(\text{asbesto} \cap \text{mesotelioma})}{P(\text{mesotelioma})} \\
 &= \frac{\frac{\text{Personas expuestas al asbesto y que tienen mesotelioma}}{\text{Total personas en el estudio}}}{\frac{\text{Personas con mesotelioma}}{\text{Total personas en el estudio}}} \\
 &= \frac{\text{Personas expuestas al asbesto y que tienen mesotelioma}}{\text{Personas con mesotelioma}}
 \end{aligned}$$

En todos los ejemplos siguientes, se suprime la segunda igualdad y se llega directamente a la fórmula simplificada.

♣ Pregunta 2.6

Solución con calculadora

$$\begin{aligned}
 P(\text{no asbesto} \mid \text{mesotelioma}) &= \frac{P(\text{no asbesto} \cap \text{mesotelioma})}{P(\text{mesotelioma})} \\
 &= \frac{\text{Personas no expuestas al asbesto y que tienen mesotelioma}}{\text{Personas con mesotelioma}} \\
 &= \frac{191}{417} = 0.4580
 \end{aligned}$$

♣ **Pregunta 2.7****Solución con calculadora**

$$\begin{aligned} P(\text{mesotelioma} \mid \text{asbesto}) &= \frac{P(\text{asbesto} \cap \text{mesotelioma})}{P(\text{asbesto})} \\ &= \frac{\text{Personas expuestas al asbesto y que tienen mesotelioma}}{\text{Personas expuestas al asbesto}} \\ &= \frac{226}{415} = 0.5446 \end{aligned}$$

♣ **Pregunta 2.8****Solución con calculadora**

$$\begin{aligned} P(\text{no mesotelioma} \mid \text{asbesto}) &= \frac{P(\text{no mesotelioma} \cap \text{asbesto})}{P(\text{asbesto})} \\ &= \frac{\text{Personas expuestas al asbesto y que no tienen mesotelioma}}{\text{Personas expuestas al asbesto}} \\ &= \frac{189}{415} = 0.4554 \end{aligned}$$

SOLUCIÓN 2.2

♣ Pregunta 2.9
Solución con calculadora

$$P(TT) = \frac{\text{Pacientes con } TT}{\text{Total de pacientes}} = \frac{299}{844} = 0.3543$$

$$P(TC) = \frac{\text{Pacientes con } TC}{\text{Total de pacientes}} = \frac{412}{844} = 0.4882$$

$$P(CC) = \frac{\text{Pacientes con } CC}{\text{Total de pacientes}} = \frac{133}{844} = 0.1576$$

♣ Pregunta 2.10
Solución con calculadora

$$\begin{aligned} P(TT \mid \text{mesotelioma}) &= \frac{P(TT \cap \text{mesotelioma})}{P(\text{mesotelioma})} \\ &= \frac{\text{Pacientes con } TT \text{ y mesotelioma}}{\text{Pacientes con mesotelioma}} = \frac{180}{417} = 0.4317 \end{aligned}$$

$$\begin{aligned} P(TC \mid \text{mesotelioma}) &= \frac{P(TC \cap \text{mesotelioma})}{P(\text{mesotelioma})} \\ &= \frac{\text{Pacientes con } TC \text{ y mesotelioma}}{\text{Pacientes con mesotelioma}} = \frac{189}{417} = 0.4532 \end{aligned}$$

$$\begin{aligned} P(CC \mid \text{mesotelioma}) &= \frac{P(CC \cap \text{mesotelioma})}{P(\text{mesotelioma})} \\ &= \frac{\text{Pacientes con } CC \text{ y mesotelioma}}{\text{Pacientes con mesotelioma}} = \frac{48}{417} = 0.1151 \end{aligned}$$

♣ Pregunta 2.11

Solución con calculadora

$$\begin{aligned}P(TT \mid \text{no mesotelioma}) &= \frac{P(TT \cap \text{no mesotelioma})}{P(\text{no mesotelioma})} \\&= \frac{\text{Pacientes con TT y sin mesotelioma}}{\text{Pacientes sin mesotelioma}} = \frac{119}{427} = 0.2787\end{aligned}$$

$$\begin{aligned}P(TC \mid \text{no mesotelioma}) &= \frac{P(TC \cap \text{no mesotelioma})}{P(\text{no mesotelioma})} \\&= \frac{\text{Pacientes con TC y sin mesotelioma}}{\text{Pacientes sin mesotelioma}} = \frac{223}{427} = 0.5222\end{aligned}$$

$$\begin{aligned}P(CC \mid \text{no mesotelioma}) &= \frac{P(CC \cap \text{no mesotelioma})}{P(\text{no mesotelioma})} \\&= \frac{\text{Pacientes con CC y sin mesotelioma}}{\text{Pacientes sin mesotelioma}} = \frac{85}{427} = 0.1991\end{aligned}$$

SOLUCIÓN 2.3

♣ Pregunta 2.12
Solución con calculadora

$$P(CC) = \frac{\text{Pacientes con CC}}{\text{Total de pacientes}} = \frac{102}{844} = 0.1209$$

$$P(CG) = \frac{\text{Pacientes con CG}}{\text{Total de pacientes}} = \frac{420}{844} = 0.4976$$

$$P(GG) = \frac{\text{Pacientes con GG}}{\text{Total de pacientes}} = \frac{322}{844} = 0.3815$$

♣ Pregunta 2.13
Solución con calculadora

$$\begin{aligned} P(CC | \text{mesotelioma}) &= \frac{P(CC \cap \text{mesotelioma})}{P(\text{mesotelioma})} \\ &= \frac{\text{Pacientes con CC y mesotelioma}}{\text{Pacientes con mesotelioma}} = \frac{56}{417} = 0.1343 \end{aligned}$$

$$\begin{aligned} P(CG | \text{mesotelioma}) &= \frac{P(CG \cap \text{mesotelioma})}{P(\text{mesotelioma})} \\ &= \frac{\text{Pacientes con CG y mesotelioma}}{\text{Pacientes con mesotelioma}} = \frac{223}{417} = 0.5348 \end{aligned}$$

$$\begin{aligned} P(GG | \text{mesotelioma}) &= \frac{P(GG \cap \text{mesotelioma})}{P(\text{mesotelioma})} \\ &= \frac{\text{Pacientes con GG y mesotelioma}}{\text{Pacientes con mesotelioma}} = \frac{138}{417} = 0.3309 \end{aligned}$$

♣ Pregunta 2.14

Solución con calculadora

$$\begin{aligned} P(CC \mid \text{no mesotelioma}) &= \frac{P(CC \cap \text{no mesotelioma})}{P(\text{no mesotelioma})} \\ &= \frac{\text{Pacientes con CC y sin mesotelioma}}{\text{Pacientes sin mesotelioma}} = \frac{46}{427} = 0.1077 \end{aligned}$$

$$\begin{aligned} P(CG \mid \text{no mesotelioma}) &= \frac{P(CG \cap \text{no mesotelioma})}{P(\text{no mesotelioma})} \\ &= \frac{\text{Pacientes con CG y sin mesotelioma}}{\text{Pacientes sin mesotelioma}} = \frac{197}{427} = 0.4614 \end{aligned}$$

$$\begin{aligned} P(GG \mid \text{no mesotelioma}) &= \frac{P(GG \cap \text{no mesotelioma})}{P(\text{no mesotelioma})} \\ &= \frac{\text{Pacientes con GG y sin mesotelioma}}{\text{Pacientes sin mesotelioma}} = \frac{184}{427} = 0.4309 \end{aligned}$$

SOLUCIÓN 2.4

♣ Pregunta 2.15

Solución con calculadora

Si A y B son independientes, entonces $P(A \cap B) = P(A) \times P(B)$, por lo tanto:

$$\begin{aligned} P(TT \cap \text{mesotelioma}) &= P(TT) \times P(\text{mesotelioma}) \\ &= \frac{\text{Pacientes con } TT}{\text{Total de pacientes}} \times \frac{\text{Pacientes con mesotelioma}}{\text{Total de pacientes}} \\ &= \frac{299}{844} \times \frac{417}{844} = 0.1750 \end{aligned}$$

$$\begin{aligned} P(TC \cap \text{mesotelioma}) &= P(TC) \times P(\text{mesotelioma}) \\ &= \frac{\text{Pacientes con } TC}{\text{Total de pacientes}} \times \frac{\text{Pacientes con mesotelioma}}{\text{Total de pacientes}} \\ &= \frac{412}{844} \times \frac{417}{844} = 0.2412 \end{aligned}$$

$$\begin{aligned} P(CC \cap \text{mesotelioma}) &= P(CC) \times P(\text{mesotelioma}) \\ &= \frac{\text{Pacientes con } CC}{\text{Total de pacientes}} \times \frac{\text{Pacientes con mesotelioma}}{\text{Total de pacientes}} \\ &= \frac{133}{844} \times \frac{417}{844} = 0.0779 \end{aligned}$$

$$\begin{aligned} P(TT \cap \text{no mesotelioma}) &= P(TT) \times P(\text{no mesotelioma}) \\ &= \frac{\text{Pacientes con } TT}{\text{Total de pacientes}} \times \frac{\text{Pacientes sin mesotelioma}}{\text{Total de pacientes}} \\ &= \frac{299}{844} \times \frac{427}{844} = 0.1792 \end{aligned}$$

$$\begin{aligned} P(TC \cap \text{no mesotelioma}) &= P(TC) \times P(\text{no mesotelioma}) \\ &= \frac{\text{Pacientes con } TC}{\text{Total de pacientes}} \times \frac{\text{Pacientes sin mesotelioma}}{\text{Total de pacientes}} \\ &= \frac{412}{844} \times \frac{427}{844} = 0.2470 \end{aligned}$$

$$\begin{aligned} P(CC \cap \text{no mesotelioma}) &= P(CC) \times P(\text{no mesotelioma}) \\ &= \frac{\text{Pacientes con } CC}{\text{Total de pacientes}} \times \frac{\text{Pacientes sin mesotelioma}}{\text{Total de pacientes}} \\ &= \frac{133}{844} \times \frac{427}{844} = 0.0797 \end{aligned}$$

♣ Pregunta 2.16

Solución con calculadora

Cada casilla se completa multiplicando el número de participantes (844) por la probabilidad obtenida en la pregunta anterior. Por ejemplo, el número esperado de pacientes con TT y mesotelioma es: $844 \times 0.1750 = 147.7$. Observe que este mismo resultado podría haberse obtenido multiplicando el total de pacientes con mesotelioma (417) por el total de pacientes con TT (299) dividido por el total de pacientes en el estudio (844): $417 \times 299/844 = 147.7$. Generalizando, la regla general más sencilla para calcular el número de pacientes esperado en cada casilla es: *total de la columna $\times$ total de la fila / total de la tabla*.

Completando la tabla con este procedimiento, quedaría así:

Genotipo del snp1	Mesotelioma	No mesotelioma	Total
TT	147.7	151.3	299
TC	203.6	208.4	412
CC	65.7	67.3	133
Total	417	427	844

Tabla 2.1: Relación entre genotipos del snp1 y mesotelioma

Si las diferencias entre la 2.2 (tabla observada en el estudio) con la 2.1 (tabla esperada si el genotipo y el mesotelioma fueran independientes) fueran grandes, se deduciría que en el estudio el mesotelioma y el genotipo no son independientes; biológicamente eso podría significar que el genotipo influye en que aparezca mesotelioma. Si las diferencias entre las dos tablas no fueran grandes, se deduciría que sí son independientes y, por tanto, el genotipo no influye en que aparezca el mesotelioma. En el capítulo 9 (tablas de contingencia) se aprenderá cómo hacer un test estadístico para esta comparación entre tabla observada y tabla esperada.

SOLUCIÓN 2.5

♣ Pregunta 2.17

Solución con calculadora

El teorema de Bayes permite calcular $P(A | B)$ conociendo $P(A)$ y $P(B | A)$. Su fórmula general es:

$$P(A_i | B) = \frac{P(B | A_i) \times P(A_i)}{\sum_{k=1}^n P(B | A_k) \times P(A_k)}$$

A continuación, se desarrolla con detalle su aplicación a las dos primeras probabilidades. en el resto se presentarán solo las fórmulas y los resultados.

Para calcular $P(\text{mesotelioma} | TT)$ hay que aplicar la fórmula del teorema de Bayes dándose cuenta de que:

- A es tener o no mesotelioma.
- A1 es mesotelioma.
- A2 es no mesotelioma.
- B es el genotipo TT.

Por lo tanto, la fórmula del teorema de Bayes aplicado a esta probabilidad es:

$$\begin{aligned} P(\text{mesotelioma} | TT) \\ = \frac{P(TT | \text{mesotelioma}) \times P(\text{mesotelioma})}{P(TT | \text{mesotelioma}) \times P(\text{mesotelioma}) + P(TT | \text{no mesotelioma}) \times P(\text{no mesotelioma})} \end{aligned}$$

Por el enunciado de este problema, sabemos que $P(\text{mesotelioma}) = 0.001$; por lo tanto, $P(\text{no mesotelioma}) = 0.999$. Por los resultados de la pregunta 10, sabemos que $P(TT | \text{mesotelioma}) = 0.4317$ y por los resultados de la pregunta 2.11 sabemos que $P(TT | \text{no mesotelioma}) = 0.2787$. Por lo tanto:

$$\begin{aligned} P(\text{mesotelioma} | TT) \\ = \frac{P(TT | \text{mesotelioma}) \times P(\text{mesotelioma})}{P(TT | \text{mesotelioma}) \times P(\text{mesotelioma}) + P(TT | \text{no mesotelioma}) \times P(\text{no mesotelioma})} \\ = \frac{0.4317 \times 0.001}{0.4317 \times 0.001 + 0.2787 \times 0.999} = 0.0015 \end{aligned}$$

Es decir, solo 15 de cada 10000 personas que tengan el genotipo TT tendrán mesotelioma.

De la misma forma, Para calcular $P(\text{mesotelioma} | TC)$ hay que aplicar la fórmula del teorema de Bayes dándose cuenta de que:

- A es tener o no mesotelioma.
- A1 es mesotelioma.
- A2 es no mesotelioma.
- B es el genotipo TC.

Por lo tanto, la fórmula del teorema de Bayes aplicado a esta probabilidad es:

$$P(\text{mesotelioma} | TC) = \frac{P(TC | \text{mesotelioma}) \times P(\text{mesotelioma})}{P(TC | \text{mesotelioma}) \times P(\text{mesotelioma}) + P(TC | \text{no mesotelioma}) \times P(\text{no mesotelioma})}$$

Por el enunciado de este problema, sabemos que $P(\text{mesotelioma}) = 0.001$; por lo tanto, $P(\text{no mesotelioma}) = 0.999$. Por los resultados de la pregunta 10, sabemos que $P(TC | \text{mesotelioma}) = 0.4532$ y por los resultados de la pregunta 2.11 sabemos que $P(TC | \text{no mesotelioma}) = 0.5222$. Por lo tanto:

$$\begin{aligned} P(\text{mesotelioma} | TC) &= \frac{P(TC | \text{mesotelioma}) \times P(\text{mesotelioma})}{P(TC | \text{mesotelioma}) \times P(\text{mesotelioma}) + P(TC | \text{no mesotelioma}) \times P(\text{no mesotelioma})} \\ &= \frac{0.4532 \times 0.001}{0.4532 \times 0.001 + 0.5222 \times 0.999} = 0.0009 \end{aligned}$$

Es decir, solo 9 de cada 10000 personas que tengan el genotipo TC tendrán mesotelioma.

El resto de las probabilidades que se piden en el problema se dan sin más explicaciones:

$$\begin{aligned} P(\text{mesotelioma} | CC) &= \frac{P(CC | \text{mesotelioma}) \times P(\text{mesotelioma})}{P(CC | \text{mesotelioma}) \times P(\text{mesotelioma}) + P(CC | \text{no mesotelioma}) \times P(\text{no mesotelioma})} \\ &= \frac{0.1151 \times 0.001}{0.1151 \times 0.001 + 0.1991 \times 0.999} = 0.0006 \end{aligned}$$

$$\begin{aligned} P(\text{no mesotelioma} | TT) &= \frac{P(TT | \text{no mesotelioma}) \times P(\text{no mesotelioma})}{P(TT | \text{no mesotelioma}) \times P(\text{no mesotelioma}) + P(TT | \text{mesotelioma}) \times P(\text{mesotelioma})} \\ &= \frac{0.2787 \times 0.999}{0.2787 \times 0.999 + 0.4317 \times 0.001} = 0.9985 \end{aligned}$$

$$P(\text{no mesotelioma} | TC)$$

$$\begin{aligned} &= \frac{P(TC | \text{no mesotelioma}) \times P(\text{no mesotelioma})}{P(TC | \text{no mesotelioma}) \times P(\text{no mesotelioma}) + P(TC | \text{mesotelioma}) \times P(\text{mesotelioma})} \\ &= \frac{0.5222 \times 0.999}{0.5222 \times 0.999 + 0.4532 \times 0.001} = 0.9991 \end{aligned}$$

$$P(\text{no mesotelioma} | CC)$$

$$\begin{aligned} &= \frac{P(CC | \text{no mesotelioma}) \times P(\text{no mesotelioma})}{P(CC | \text{no mesotelioma}) \times P(\text{no mesotelioma}) + P(CC | \text{mesotelioma}) \times P(\text{mesotelioma})} \\ &= \frac{0.1991 \times 0.999}{0.1991 \times 0.999 + 0.1151 \times 0.001} = 0.9994 \end{aligned}$$

SOLUCIÓN 2.6

♣ Pregunta 2.18**Solución con calculadora**

Lo que se pregunta es $P(\text{COVID-19} \mid +)$. Para ello hay que utilizar el teorema de Bayes sabiendo que:

- A1 es COVID-19
- A2 es no COVID-19
- B es positivo
- $P(\text{COVID-19}) = 0.05$; por lo tanto: $P(\text{no COVID-19}) = 0.95$
- $P(+ \mid \text{COVID-19}) = 0.70$ (la sensibilidad)
- $P(- \mid \text{no COVID-19}) = 0.99$ (la especificidad), y por tanto $P(+ \mid \text{no COVID-19}) = 0.01$ (1 – especificidad).

Aplicando Bayes:

$$\begin{aligned}
 &P(\text{COVID-19} \mid +) \\
 &= \frac{P(+ \mid \text{COVID-19}) \times P(\text{COVID-19})}{P(+ \mid \text{COVID-19}) \times P(\text{COVID-19}) + P(+ \mid \text{no COVID-19}) \times P(\text{no COVID-19})} \\
 &= \frac{0.70 \times 0.05}{0.70 \times 0.05 + 0.01 \times 0.95} = 0.787
 \end{aligned}$$

Es decir, de cada 1000 personas que den positivo, solo 787 han tenido COVID-19.

♣ Pregunta 2.19**Solución con calculadora**

Se pregunta lo mismo que en la 2.18, solo que ahora $P(\text{COVID-19}) = 0.20$ y $P(\text{no COVID-19}) = 0.80$.

Aplicando Bayes:

$$\begin{aligned}
 &P(\text{COVID-19} | +) \\
 &= \frac{P(+ | \text{COVID-19}) \times P(\text{COVID-19})}{P(+ | \text{COVID-19}) \times P(\text{COVID-19}) + P(+ | \text{no COVID-19}) \times P(\text{no COVID-19})} \\
 &= \frac{0.70 \times 0.20}{0.70 \times 0.20 + 0.01 \times 0.80} = 0.946
 \end{aligned}$$

Es decir, un resultado positivo en Nueva York es mucho más fiable (946 enfermos de cada 1000 positivos) que en España (787 enfermos de cada 1000 positivos).

♣ Pregunta 2.20

Solución con calculadora

Ahora, lo que se pregunta es $P(\text{no COVID-19} | -)$.

Hay que aplicar Bayes sabiendo que:

- A1 es no COVID-19
- A2 es COVID-19
- B es resultado negativo
- $P(\text{COVID-19}) = 0.05$; por lo tanto: $P(\text{no COVID-19}) = 0.95$
- $P(+ | \text{COVID-19}) = 0.70$ (la sensibilidad); por tanto: $P(- | \text{COVID-19}) = 0.30$
- $P(- | \text{no COVID-19}) = 0.99$ (la especificidad).

$$\begin{aligned}
 &P(\text{no COVID-19} | -) \\
 &= \frac{P(- | \text{no COVID-19}) \times P(\text{no COVID-19})}{P(- | \text{no COVID-19}) \times P(\text{no COVID-19}) + P(- | \text{COVID-19}) \times P(\text{COVID-19})} \\
 &= \frac{0.99 \times 0.95}{0.99 \times 0.95 + 0.30 \times 0.05} = 0.984
 \end{aligned}$$

♣ **Pregunta 2.21****Solución con calculadora**

Se pide lo mismo que en la 2.20, pero con $P(\text{COVID-19}) = 0.20$.

Aplicando el teorema de Bayes:

$$\begin{aligned} & P(\text{no COVID-19} \mid -) \\ &= \frac{P(- \mid \text{no COVID-19}) \times P(\text{no COVID-19})}{P(- \mid \text{no COVID-19}) \times P(\text{no COVID-19}) + P(- \mid \text{COVID-19}) \times P(\text{COVID-19})} \\ &= \frac{0.99 \times 0.80}{0.99 \times 0.80 + 0.30 \times 0.20} = 0.930 \end{aligned}$$

Es decir, con la misma prueba diagnóstica, un resultado negativo falla en 70 pacientes de cada 1000 en Nueva York, pero solo falla en 16 pacientes de cada 1000 en España.

Distribuciones de probabilidad

SOLUCIÓN 3.1

Las condiciones del problema son: distribución normal con media $\mu = 38$, varianza $\sigma^2 = 25$, desviación típica $\sigma = 5$.

♣ Pregunta 3.1

Solución con calculadora

Se pregunta la probabilidad de $x > 38$. Pero 38 es la media. En la distribución normal, la media deja por encima al 50 % de la población.

♣ Pregunta 3.2

Solución con calculadora

Se pregunta la probabilidad de $x > 43$ con media 38 y desviación típica 5. Es decir: $P(x > 43 | \mu = 38, \sigma = 5)$.

- Calcular z :

$$z = \frac{x - \mu}{\sigma} = \frac{43 - 38}{5} = 1.00$$

¿Por qué es importante calcular z ? Porque solo tenemos la tabla $N(0,1)$ y porque con esta transformación sabemos que

$$P(x > 43 | \mu = 38, \sigma = 5) = P(x > 1 | \mu = 0, \sigma = 1)$$

, que es valor que sí podemos obtener de la tabla $N(0,1)$.

- A continuación, se busca en la tabla de la normal estandarizada $[N(0,1)]$ el valor $z = 1.00$ y se encuentra $P(x < 1 | \mu = 0, \sigma = 1) = 0.8413$. Por lo tanto,

$$\begin{aligned} P(x > 43 | \mu = 38, \sigma = 5) &= P(x > 1 | \mu = 0, \sigma = 1) \\ &= 1 - 0.8413 = 0.1587 = 15.87\% \end{aligned}$$

Solución con Excel

1-DISTR.NORM.N(43;38;5;VERDADERO)

Resultado: 0.15865525

Solución con Stata

```
di 1-normal((43-38)/5)
```

Resultado: 0.15865525

♣ Pregunta 3.3

Para saber qué porcentaje de la población tiene entre 33 y 43 mg/dL hay que calcular $P(x > 33) - P(x > 43)$. El segundo ya lo conocemos por la pregunta anterior: $P(x > 43) = 0.1587$.

Solución con calculadora

Para conocer $P(x > 33 | \mu = 38, \sigma = 5)$ se sigue el mismo orden que en la pregunta anterior.

- Cálculo de z :

$$z = \frac{x - \mu}{\sigma} = \frac{33 - 38}{5} = -1.00$$

- Consultar la tabla de la normal (0,1): En la tabla solo vienen los valores de $z > 0$, pero nuestro resultado es negativo: $z = -1$. Como la tabla es simétrica respecto al cero, eso quiere decir que

$$P(x > -1 | \mu = 0, \sigma = 1) = P(x < 1 | \mu = 0, \sigma = 1)$$

Pero la $P(x < 1 | \mu = 0, \sigma = 1)$ la encontramos en la pregunta anterior: 0.8413.

Por lo tanto,

$$P(x > 33 | \mu = 38, \sigma = 5) = P(x > -1 | \mu = 0, \sigma = 1) = 0.8413 = 84.13 \%$$

La pregunta pide el porcentaje entre 33 y 43 mg/dL:

$$\begin{aligned} P(33 < x < 43 | \mu = 38, \sigma = 5) \\ &= P(x > 33 | \mu = 38, \sigma = 5) - P(x > 43 | \mu = 38, \sigma = 5) \\ &= 84.13 - 15.87 = 68.26 \% \end{aligned}$$

Nota: nos podíamos haber ahorrado todos los cálculos si hubiéramos sabido que en cualquier distribución normal entre la $\mu - 1\sigma$ y la $\mu + 1\sigma$ está siempre el 68% de la población.

Solución con Excel

```
(1-DISTR.NORM.N(33;38;5;VERDADERO))-(1-DISTR.NORM.N(43;38;5;VERDADERO))
```

Resultado: 0.68268949

Solución con Stata

```
di normal((43-38)/5) - normal((33-38)/5)
```

Resultado: 0.68268949

♣ Pregunta 3.4**Solución con calculadora**

Se pide $P(x > 30 | \mu = 38, \sigma = 5)$.

- Calcular z :

$$z = \frac{x - \mu}{\sigma} = \frac{30 - 38}{5} = -1.60$$

- Consultar la tabla normal para $z = -1.60$. Por la simetría de la normal, sabemos que $P(x > -1.60 | \mu = 0, \sigma = 1) = P(x < 1.60 | \mu = 0, \sigma = 1)$.

Por lo tanto, consultamos la tabla para $z = 1.60$ y obtenemos que por debajo de este valor está el 94.52 % de la población. Por lo tanto:

$$\begin{aligned} P(x > 30 | \mu = 38, \sigma = 5) &= P(x > -1.60 | \mu = 0, \sigma = 1) \\ &= P(x < 1.60 | \mu = 0, \sigma = 1) \\ &= 94.52\% \end{aligned}$$

Solución con Excel

```
1-DISTR.NORM.N(30;38;5;VERDADERO)
```

Resultado: 0.94520071

Solución con Stata

```
di 1-normal((30-38)/5)
```

Resultado: 0.94520071

♣ Pregunta 3.5**Solución con calculadora**

Se pide $P(30 < x < 35 | \mu = 38, \sigma = 5)$ y esto es igual a:

$$P(x > 30 | \mu = 38, \sigma = 5) - P(x > 35 | \mu = 38, \sigma = 5)$$

Por la pregunta anterior, sabemos que el primer término es 94.52%. Falta calcular el segundo:

- Calcular z:

$$z = \frac{x - \mu}{\sigma} = \frac{35 - 38}{5} = -0.60$$

- Consultar la tabla normal para $z = -0.60$. Por la simetría de la normal, sabemos que $P(x > -0.60 | \mu = 0, \sigma = 1) = P(x < 0.60 | \mu = 0, \sigma = 1)$.

Por lo tanto, consultamos la tabla para $z = 0.60$ y obtenemos que por debajo de este valor está el 72.57% de la población. Por lo tanto:

$$\begin{aligned} P(x > 35 | \mu = 38, \sigma = 5) &= P(x > -0.60 | \mu = 0, \sigma = 1) \\ &= P(x < 0.60 | \mu = 0, \sigma = 1) \\ &= 72.57\% \end{aligned}$$

- Lo que se pide en la pregunta es $P(30 < x < 35 | \mu = 38, \sigma = 5)$:

$$\begin{aligned} P(30 < x < 35 | \mu = 38, \sigma = 5) &= P(x > 30 | \mu = 38, \sigma = 5) - P(x > 35 | \mu = 38, \sigma = 5) \\ &= 94.52 - 72.57 = 21.95\% \end{aligned}$$

Solución con Excel

```
(1-DISTR.NORM.N(30;38;5;VERDADERO))-(1-DISTR.NORM.N(35;38;5;VERDADERO))
```

Resultado: 0.21945383

Solución con Stata

```
di normal((35-38)/5) - normal((30-38)/5)
```

Resultado: 0.21945383

SOLUCIÓN 3.2

♣ Pregunta 3.6

Solución con calculadora

- Hay que buscar en el cuerpo de la tabla el valor más próximo al percentil 70 (es decir, a 0.7000). El valor más próximo es $p = 0.6985$ que corresponde a $z = 0.52$.
- Transformar la z en el valor x de la distribución de HDL. Se hace despejando x en la fórmula que hemos usado repetidamente en el ejercicio anterior:

$$z = \frac{x - \mu}{\sigma} \rightarrow x = z\sigma + \mu$$

Aplicado a este ejercicio:

$$x = z\sigma + \mu = 0.52 \times 5 + 38 = 40.6 \text{ mg/dL}$$

Solución con Excel

`INV.NORM(0.7;38;5)`

Resultado: 40.6220 mg/dL

Solución con Stata

`di invnormal(.7)*5+38`

Resultado: 40.62 mg/dL

♣ Pregunta 3.7

Solución con calculadora

Se siguen los mismos pasos que en la pregunta anterior.

- Hay que buscar en el cuerpo de la tabla el valor más próximo al percentil 40 (es decir, a 0.4000). En la tabla vienen solo los valores $p \geq 0.50$. Para resolver

el problema, tendremos que usar de nuevo la simetría de la tabla $N(0,1)$: el valor z que deje por debajo 0.40 será el negativo del valor z que deje por encima 0.40 (que es lo mismo que el valor z que deje por debajo 0.60). El valor más próximo a 0.60 es $p = 0.5987$ que corresponde a $z = 0.25$. Por lo tanto, el valor z que deja por debajo 0.40 será $z = -0.25$.

- Transformar la $z = -0.25$ en el valor x de la distribución de HDL.

Aplicado a este ejercicio:

$$x = z\sigma + \mu = -0.25 \times 5 + 38 = 36.75 \text{ mg/dL}$$

Solución con Excel

`INV.NORM(0.4;38;5)`

Resultado: 36.7332 mg/dL

Solución con Stata

`di invnormal(.4)*5+38`

Resultado: 36.73 mg/dL

♣ Pregunta 3.8

Solución con calculadora

Es muy parecida a la anterior.

- Hay que buscar en el cuerpo de la tabla el valor más próximo al percentil 10 (es decir, a 0.1000). En la tabla vienen solo los valores $p \geq 0.50$. Para resolver el problema, tendremos que usar de nuevo la simetría de la tabla $N(0,1)$: el valor z que deje por debajo 0.10 será el negativo del valor z que deje por encima 0.10 (que es lo mismo que el valor z que deje por debajo 0.90). El valor más próximo a 0.90 es $p = 0.8997$ que corresponde a $z = 1.28$. Por lo tanto, el valor z que deja por debajo 0.10 será $z = -1.28$.
- Transformar la $z = -1.28$ en el valor x de la distribución de HDL.

Aplicado a este ejercicio:

$$x = z\sigma + \mu = -1.28 \times 5 + 38 = 31.60 \text{ mg/dL}$$

Solución con Excel

`INV.NORM(0.1;38;5)`

Resultado: 31.5922 mg/dL

Solución con Stata

`di invnormal(.1)*5+38`

Resultado: 31.59 mg/dL

SOLUCIÓN 3.3

Todo este problema se hace igual que el 3.1. Es decir, tenemos valores en una distribución $N(0.17, 0.0009)$, tenemos que pasarlos a valores z de una distribución $N(0,1)$ y buscar la probabilidad correspondiente en la tabla de la normal. Observe que en $N(0.17, 0.0009)$, la desviación típica es 0.03 (raíz cuadrada de 0.0009).

♣ Pregunta 3.9

Solución con calculadora

Se pide $P(x > 0.23 | \mu = 0.17, \sigma = 0.03)$.

- Calcular z :

$$z = \frac{x - \mu}{\sigma} = \frac{0.23 - 0.17}{0.03} = 2.00$$

- Consultar la tabla normal para $z = 2.00$. Obtenemos que por debajo de este valor está el 97.72 % de la población. Por lo tanto:

$$P(x > 0.23 | \mu = 0.17, \sigma = 0.03) = 2.28\%.$$

Solución con Excel

`1-DISTR.NORM.N(0.23;0.17;0.03;VERDADERO)`

Resultado: 0.022750132

Solución con Stata

`di 1-normal((0.23-0.17)/0.03)`

Resultado: 0.02275013

♣ Pregunta 3.10

Solución con calculadora

Se pide $P(0.17 < x < 0.23 | \mu = 0.17, \sigma = 0.03)$.

Esto es igual a:

$$P(x > 0.17 | \mu = 0.17, \sigma = 0.03) - P(x > 0.23 | \mu = 0.17, \sigma = 0.03)$$

El segundo término se ha calculado en la pregunta anterior. El primero vale 0.50 porque 0.17 es la media. Por lo tanto:

$$P(0.17 < x < 0.23 | \mu = 0.17, \sigma = 0.03) = 0.50 - 0.0228 = 0.4772 = 47.72\%$$

Solución con Excel

```
(1-DISTR.NORM.N(0.17;0.17;0.03;VERDADERO))-(1-DISTR.NORM.N(0.23;0.17;0.03;VERDADERO))
```

Resultado: 0.47724987

Solución con Stata

```
di normal((0.23-0.17)/0.03)-normal((0.17-.17)/.03)
```

Resultado: 0.47724987

♣ Pregunta 3.11

Solución con calculadora

Se pide $P(0.12 < x < 0.20 | \mu = 0.17, \sigma = 0.03)$.

Esto es igual a: $P(x > 0.12 | \mu = 0.17, \sigma = 0.03) - P(x > 0.20 | \mu = 0.17, \sigma = 0.03)$. Calculemos cada término:

$$P(x > 0.12 | \mu = 0.17, \sigma = 0.03)$$

- Calcular z : $z = \frac{x - \mu}{\sigma} = \frac{0.12 - 0.17}{0.03} = -1.67$
- Consultar la tabla normal para $z = +1.67$. Obtenemos que por debajo de este valor está el 95.25% de la población; por lo tanto, por encima de $z = -1.67$ también estará el 95.25%. Por lo tanto:

$$P(x > 0.12 | \mu = 0.17, \sigma = 0.03) = 95.25\%$$

$$P(x > 0.20 | \mu = 0.17, \sigma = 0.03)$$

- Calcular z :

$$z = \frac{x - \mu}{\sigma} = \frac{0.20 - 0.17}{0.03} = 1.00$$

- Consultar la tabla normal para $z = +1.00$. Obtenemos que por debajo de este valor está el 84.13% de la población; por lo tanto, por encima estará el 15.87%. Por lo tanto:

$$P(x > 0.20 | \mu = 0.17, \sigma = 0.03) = 15.87\%$$

Finalmente,

$$P(0.12 < x < 0.20 | \mu = 0.17, \sigma = 0.03) = 95.25 - 15.87 = 79.38\%$$

Solución con Excel

```
(1-DISTR.NORM.N(0.12;0.17;0.03;VERDADERO))-(1-DISTR.NORM.N(0.2;0.17;0.03;VERDADERO))
```

Resultado: 0.79355439

Solución con Stata

```
di normal((0.20-0.17)/0.03)-normal((.12-.17)/.03)
```

Resultado: 0.79355439

♣ Pregunta 3.12

Solución con calculadora

Se pide $P(x < 0.14 | \mu = 0.17, \sigma = 0.03)$.

- Calcular z :

$$z = \frac{x - \mu}{\sigma} = \frac{0.14 - 0.17}{0.03} = -1.00$$

- Consultar la tabla normal para $z = +1.00$. Obtenemos que por debajo de este valor está el 84.13% de la población; por lo tanto, por encima de $z = +1.00$ está el 15.87%. Por simetría, este es el porcentaje que estará por debajo de $z = -1.00$. Por lo tanto:

$$P(x < 0.14 | \mu = 0.17, \sigma = 0.03) = 15.87\%$$

Solución con Excel

DISTR.NORM.N(0.14;0.17;0.03;VERDADERO)

Resultado: 0.15865525

Solución con Stata

di normal((.14-.17)/.03)

Resultado: 0.15865525

♣ Pregunta 3.13**Solución con calculadora**

Se pide $P(0.14 < x < 0.16 | \mu = 0.17, \sigma = 0.03)$.

Esto es igual a:

$$P(x < 0.16 | \mu = 0.17, \sigma = 0.03) - P(x < 0.14 | \mu = 0.17, \sigma = 0.03)$$

El segundo término se ha calculado en la pregunta anterior y vale 15.87%.

Calculemos el primer término: $P(x < 0.16 | \mu = 0.17, \sigma = 0.03)$

- Calcular z:

$$z = \frac{x - \mu}{\sigma} = \frac{0.16 - 0.17}{0.03} = -0.33$$

- Consultar la tabla normal para $z = +0.33$. Obtenemos que por debajo de este valor está el 62.93% de la población; por lo tanto, por encima de $z = +0.33$ también estará el 37.07%. Por simetría, este es también el porcentaje que estará por debajo de $z = -0.33$. Por lo tanto:

$$P(x < 0.16 | \mu = 0.17, \sigma = 0.03) = 37.07\%.$$

Finalmente,

$$P(0.14 < x < 0.16 | \mu = 0.17, \sigma = 0.03) = 37.07 - 15.87 = 21.20\%$$

Solución con Excel

```
DISTR.NORM.N(0.16;0.17;0.03;VERDADERO)-DISTR.NORM.N(0.14;0.17;0.03;VERDADERO)
```

Resultado: 0.21078609

Solución con Stata

```
di normal((0.16-0.17)/0.03)-normal((.14-.17)/.03)
```

Resultado: 0.21078609

SOLUCIÓN 3.4

Este problema se resuelve de forma análoga al problema 3.2. Primero se busca el valor p en el cuerpo de la tabla. Después se obtiene el valor z . Por último, se transforma z en x .

♣ Pregunta 3.14

Solución con calculadora

- Se busca en el cuerpo de la tabla el valor más próximo a $p = 0.4500$. Como en la tabla solo vienen los valores $p \geq 0.50$, habrá que buscar el complementario de 0.4500; esto es $1 - 0.4500 = 0.5500$.

El valor más próximo es $p = 0.5517$, que corresponde a $z = 0.13$. Por lo tanto, la z para $p = 0.4500$ es $z = -0.13$.

- Obtener x a partir de z :

$$x = z\sigma + \mu = -0.13 \times 0.03 + 0.17 = 0.1661 \text{ ng/mL}$$

Solución con Excel

`INV.NORM(0.45;0.17;0.03)`

Resultado: 0.16623016

Solución con Stata

`di invnormal(.45)*.03+.17`

Resultado: 0.16623016

♣ Pregunta 3.15

Solución con calculadora

Se pide el valor que deja por encima el 30%; por lo tanto, es el valor que deja por debajo el 70%.

- Se busca en el cuerpo de la tabla el valor más próximo a $p = 0.7000$. El valor más próximo es $p = 0.6985$, que corresponde a $z = 0.52$.

- Obtener x a partir de z :

$$x = z\sigma + \mu = +0.52 \times 0.03 + 0.17 = 0.1856 \text{ ng/mL}$$

Solución con Excel

INV.NORM(0.7;0.17;0.03)

Resultado: 0.185732015

Solución con Stata

di invnormal(.70)*.03+.17

Resultado: 0.18573202

♣ Pregunta 3.16

Solución con calculadora

Se pide el valor que deja por encima el 5%; por lo tanto, es el valor que deja por debajo el 95%.

- Se busca en el cuerpo de la tabla el valor más próximo a $p = 0.9500$. Hay dos valores a la misma distancia: $p = 0.9495$, que corresponde a $z = 1.64$, y $p = 0.9505$, que corresponde a $z = 1.65$. Cualquiera de los dos es correcto. En esta solución lo haremos con 1.64.

- Obtener x a partir de z :

$$x = z\sigma + \mu = +1.64 \times 0.03 + 0.17 = 0.2192 \text{ ng/mL}$$

Solución con Excel

INV.NORM(0.95;0.17;0.03)

Resultado: 0.21934561

Solución con Stata

```
di invnormal(.95)*.03+.17
```

Resultado: 0.21934561

♣ Pregunta 3.17**Solución con calculadora**

Se piden dos valores. Uno que deja por debajo el 10% y otro que deja por encima el 10%. Por lo tanto, los dos tendrán el mismo valor z con el signo cambiado.

Calculemos el valor que deja por encima el 10% (y, por tanto, deja por debajo el 90%):

- Se busca en el cuerpo de la tabla el valor más próximo a $p = 0.9000$.
El valor más próximo es: $p = 0.8997$, que corresponde a $z = 1.28$.
- Obtener x a partir de z :

$$x = z\sigma + \mu = +1.28 \times 0.03 + 0.17 = 0.2084 \text{ ng/mL}$$

- A continuación, repetimos la operación para el valor que deja por debajo el 10%. Sabemos que es el mismo z con el signo cambiado: $z = -1.28$. Solo falta obtener x :

$$x = z\sigma + \mu = -1.28 \times 0.03 + 0.17 = 0.1316 \text{ ng/mL}$$

Por lo tanto, los valores que dejan en medio el 80% de la distribución son 0.1316 y 0.2084.

Solución con Excel

- Valor que deja por debajo el 90%

```
INV.NORM(0.9;0.17;0.03)
```

Resultado: 0.20844655

- Valor que deja por debajo el 10%

```
INV.NORM(0.1;0.17;0.03)
```

Resultado: 0.13155345

Solución con Stata

- Valor que deja por debajo el 90 %

```
di invnormal(.90)*.03+.17
```

Resultado: 0.20844655

- Valor que deja por debajo el 10 %

```
di invnormal(.10)*.03+.17
```

Resultado: 0.13155345

SOLUCIÓN 3.5

♣ Pregunta 3.18

Solución con calculadora

Hay dos formas de resolver este problema. Una es utilizando la distribución binomial y la otra usando la definición de independencia que se aplicó en el capítulo 2.

- Usando la definición de independencia:

Los dos alelos se transmiten de forma independiente (cada uno procede de una planta distinta; vamos a llamarlas planta 1 y planta 2).

Por la definición de independencia sabemos que

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

, aplicado a este problema da:

$$P(AA) = P(A)P(A) = 0.6^2 = 0.36$$

$P(Av)$ es más complicado porque hay dos opciones: que A venga de la planta 1 y v de la planta 2 o al revés. Por eso, hay que multiplicar la fórmula por 2:

$$P(Av) = 2P(A)P(v) = 2 \times 0.6 \times 0.4 = 0.48$$

Por último,

$$P(vv) = P(v)P(v) = 0.4^2 = 0.16$$

Observe que la suma de estas tres posibilidades da 1. Si no se hubiera multiplicado por 2 el segundo cálculo, tendríamos un 24% de plantas sin genotipo.

- Usando la distribución binomial:

El número de copias de A que haya en dos alelos sigue una distribución binomial con $n = 2$, $p = 0.6$ [B(2,0.6)]. La fórmula general es:

$$P(A = k | n = 2, p = 0.6) = \binom{n}{k} p^k (1-p)^{n-k}$$

Por lo tanto:

$$\begin{aligned} P(AA) &= P(A = 2 | n = 2, p = 0.6) \\ &= \binom{2}{2} 0.6^2 (1 - 0.6)^{2-2} = 0.6^2 = 0.36 \end{aligned}$$

$$\begin{aligned} P(Av) &= P(A = 1 | n = 2, p = 0.6) \\ &= \binom{2}{1} 0.6^1 (1 - 0.6)^{2-1} = 2 \times 0.6 \times 0.4 = 0.48 \end{aligned}$$

$$\begin{aligned} P(vv) &= P(A = 0 | n = 2, p = 0.6) \\ &= \binom{2}{0} 0.6^0 (1 - 0.6)^{2-0} = (1 - 0.6)^2 = 0.16 \end{aligned}$$

Solución con Excel

Usando distribución binomial:

- P(AA): =DISTR.BINOM.N(2;2;0.6;FALSO)
Resultado: 0.36
- P(Av): =DISTR.BINOM.N(1;2;0.6;FALSO)
Resultado: 0.48
- P(vv): =DISTR.BINOM.N(0;2;0.6;FALSO)
Resultado: 0.16

Solución con Stata

Usando distribución binomial:

- P(AA): di binomialp(2,2,.6)
Resultado: 0.36
- P(Av): di binomialp(2,1,.6)
Resultado: 0.48
- P(vv): di binomialp(2,0,.6)
Resultado: 0.16

♣ **Pregunta 3.19**

Solución con calculadora

Se resuelve de la misma forma que el anterior.

$$P(LL) = P(L)P(L) = 0.8^2 = 0.64$$

$$P(Lr) = 2P(L)P(r) = 2 \times 0.8 \times 0.2 = 0.32$$

$$P(rr) = P(r)P(r) = 0.2^2 = 0.04$$

Solución con Excel

- P(LL): =DISTR.BINOM.N(2;2;0.8;FALSO)
Resultado: 0.64
- P(Lr): =DISTR.BINOM.N(1;2;0.8;FALSO)
Resultado: 0.32
- P(rr): =DISTR.BINOM.N(0;2;0.8;FALSO)
Resultado: 0.04

Solución con Stata

Usando distribución binomial:

- P(LL): di binomialp(2,2,.8)
Resultado: 0.64
- P(Lr): di binomialp(2,1,.8)
Resultado: 0.32
- P(rr): di binomialp(2,0,.2)
Resultado: 0.04

♣ **Pregunta 3.20****Solución con calculadora**

Como A y L son dominantes, serán plantas amarillas todas las que tengan genotipo AA o Av y serán plantas lisas todas las que tengan genotipo LL o Lr.

Las probabilidades son:

$$P(\text{amarilla}) = P(AA) + P(Av) = 0.84$$

$$P(\text{verde}) = P(vv) = 0.16$$

$$P(\text{lisa}) = P(LL) + P(Lr) = 0.96$$

$$P(\text{rugosa}) = P(rr) = 0.04$$

Vamos ahora con lo que se pregunta. Usando la propiedad de independencia:

$$P(\text{amarilla y lisa}) = P(\text{amarilla})P(\text{lisa}) = 0.84 \times 0.96 = 0.8064$$

$$P(\text{amarilla y rugosa}) = P(\text{amarilla})P(\text{rugosa}) = 0.84 \times 0.04 = 0.0336$$

$$P(\text{verde y lisa}) = P(\text{verde})P(\text{lisa}) = 0.16 \times 0.96 = 0.1536$$

$$P(\text{verde y rugosa}) = P(\text{verde})P(\text{rugosa}) = 0.16 \times 0.04 = 0.0064$$

SOLUCIÓN 3.6

♣ Pregunta 3.21

Solución con calculadora

Como la pregunta 3.18, esta puede resolverse usando independencia o usando la distribución binomial.

- Usando la noción de independencia de sucesos:

Hay que utilizar la propiedad de independencia que vimos en el problema anterior. Para el genotipo AB hay que multiplicar por 2 porque puede ocurrir que A venga del primer progenitor y B del segundo o al revés. Con eso es suficiente para:

$$P(AA) = P(A)P(A) = p^2$$

$$P(AB) = 2P(A)P(B) = 2p(1-p)$$

$$P(BB) = P(B)P(B) = (1-p)^2$$

- Usando la distribución binomial de parámetros $n = 2$ (porque hay dos loci) y p :

$$P(AA) = P(A = 2|n = 2, p) = \binom{2}{2} p^2 (1-p)^{2-2} = p^2$$

$$P(AB) = P(A = 1|n = 2, p) = \binom{2}{1} p^1 (1-p)^{2-1} = 2p(1-p)$$

$$P(BB) = P(A = 0|n = 2, p) = \binom{2}{0} p^0 (1-p)^{2-0} = (1-p)^2$$

A esta distribución de los genotipos en una población se le conoce en genética como Ley de Hardy-Weinberg en honor al matemático Godfrey Hardy y el ginecólogo Wilhelm Weinberg.

SOLUCIÓN 3.7

♣ Pregunta 3.22**Solución con calculadora**

La frecuencia de fibrosis quística será la probabilidad del genotipo aa:

$$P(aa) = P(a)P(a) = \left(\frac{1}{30}\right)^2 = \frac{1}{900} = 0.0011 = 0.11\%$$

♣ Pregunta 3.23**Solución con calculadora**

La frecuencia de portadores en los recién nacidos será la probabilidad del genotipo Aa. De acuerdo con lo que hemos visto en los problemas anteriores:

$$P(Aa) = 2P(A)P(a) = 2 \times \frac{29}{30} \times \frac{1}{30} = \frac{58}{900} = 0.0644 = 6.44\%$$

SOLUCIÓN 3.8

Nos encontramos con una distribución binomial en la que n es el número de trasplantes ($n = 10$) y p es la probabilidad de EICH ($p = 0.2$).

♣ Pregunta 3.24

Solución con calculadora

$$\begin{aligned} P(k=4|n=10, p=0.2) &= \binom{n}{k} p^k (1-p)^{n-k} \\ &= \binom{10}{4} 0.2^4 (1-0.2)^{10-4} \\ &= \frac{10!}{4!(10-4)!} \times 0.2^4 \times 0.8^6 = 0.0881 = 8.81\% \end{aligned}$$

Solución con Excel

DISTR.BINOM.N(4; 10; 0.2; FALSO)

Resultado: 0.8808038

Solución con Stata

di binomialp(10,4,.2)

Resultado: 0.8808038

♣ Pregunta 3.25

Solución con calculadora

$$\begin{aligned} P(k \geq 4|n=10, p=0.2) \\ &= 1 - P(k \leq 3|n=10, p=0.2) \\ &= 1 - (P(k=0|n=10, p=0.2) + P(k=1|n=10, p=0.2) \\ &\quad + P(k=2|n=10, p=0.2) + P(k=3|n=10, p=0.2)) \end{aligned}$$

Y ahora se calcula cada uno de los términos aplicando la fórmula que hemos visto en 3.24:

$$P(k=0|n=10, p=0.2) = \binom{10}{0} 0.2^0 (1-0.2)^{10-0}$$

$$= \frac{10!}{0!(10-0)!} \times 0.2^0 \times 0.8^{10} = 0.1074$$

$$P(k=1|n=10, p=0.2) = \binom{10}{1} 0.2^1 (1-0.2)^{10-1}$$

$$= \frac{10!}{1!(10-1)!} \times 0.2^1 \times 0.8^9 = 0.2684$$

$$P(k=2|n=10, p=0.2) = \binom{10}{2} 0.2^2 (1-0.2)^{10-2}$$

$$= \frac{10!}{2!(10-2)!} \times 0.2^2 \times 0.8^8 = 0.3020$$

$$P(k=3|n=10, p=0.2) = \binom{10}{3} 0.2^3 (1-0.2)^{10-3}$$

$$= \frac{10!}{3!(10-3)!} \times 0.2^3 \times 0.8^7 = 0.2013$$

Por lo tanto, la solución es:

$$P(k \geq 4|n=10, p=0.2) = 1 - P(k \leq 3|n=10, p=0.2)$$

$$= 1 - (0.1074 + 0.2684 + 0.3020 + 0.2013)$$

$$= 1 - 0.8791 = 0.1209 = 12.09\%$$

Solución con Excel

`(1-DISTR.BINOM.N(3;10;0.2;VERDADERO))`

Resultado: 0.120873882

Solución con Stata

`di binomialtail(10,4,.2)`

Resultado: 0.12087388

El mismo resultado se obtiene con la orden `di 1-binomial(10,3,.2)`

SOLUCIÓN 3.9

Tenemos una distribución binomial con $n = 15$ y $p = 0.07$.

♣ Pregunta 3.26

Solución con calculadora

$$\begin{aligned} P(k = 3|n = 15, p = 0.07) &= \binom{n}{k} p^k (1-p)^{n-k} \\ &= \binom{15}{3} 0.07^3 (1 - 0.07)^{15-3} \\ &= \frac{15!}{3!(15-3)!} \times 0.07^3 \times 0.93^{12} = 0.0653 = 6.53\% \end{aligned}$$

Solución con Excel

DISTR.BINOM.N(3; 15; 0.07; FALSO)

Resultado: 0.06532823

Solución con Stata

di binomialp(15,3,.07)

Resultado: 0.06532823

♣ Pregunta 3.27

Solución con calculadora

$$\begin{aligned} P(k \geq 3|n = 15, p = 0.07) &= 1 - P(k \leq 2|n = 15, p = 0.07) \\ &= 1 - (P(k = 0|n = 15, p = 0.07) \\ &\quad + P(k = 1|n = 15, p = 0.07) + P(k = 2|n = 15, p = 0.07)) \end{aligned}$$

Ahora se calcula cada término por separado:

$$\begin{aligned} P(k=0|n=15, p=0.07) &= \binom{15}{0} 0.07^0 (1-0.07)^{15-0} \\ &= \frac{15!}{0!(15-0)!} \times 0.07^0 \times 0.93^{15} = 0.3367 \end{aligned}$$

$$\begin{aligned} P(k=1|n=15, p=0.07) &= \binom{15}{1} 0.07^1 (1-0.07)^{15-1} \\ &= \frac{15!}{1!(15-1)!} \times 0.07^1 \times 0.93^{14} = 0.3801 \end{aligned}$$

$$\begin{aligned} P(k=2|n=15, p=0.07) &= \binom{15}{2} 0.07^2 (1-0.07)^{15-2} \\ &= \frac{15!}{2!(15-2)!} \times 0.07^2 \times 0.93^{13} = 0.2003 \end{aligned}$$

Y, por lo tanto:

$$\begin{aligned} P(k \geq 3|n=15, p=0.07) &= 1 - P(k \leq 2|n=15, p=0.07) \\ &= 1 - (0.3367 + 0.3801 + 0.2003) = 0.0829 = 8.29\% \end{aligned}$$

Solución con Excel

`1-DISTR.BINOM.N(2;15;0.07;VERDADERO)`

Resultado: 0.08286095

Solución con Stata

`di binomialtail(15,3,.07)`

Resultado: 0.08286095

El mismo resultado se obtiene con la orden `di 1-binomial(15,2,.07)`

SOLUCIÓN 3.10

Se trata de una distribución binomial con $n = 7$ pacientes, $p = 0.04$.

♣ Pregunta 3.28

Solución con calculadora

$$\begin{aligned} P(k = 0 | n = 7, p = 0.04) &= \binom{n}{k} p^k (1-p)^{n-k} = \binom{7}{0} 0.04^0 (1-0.04)^{7-0} \\ &= \frac{7!}{0!(7-0)!} \times 0.04^0 \times 0.96^7 = 0.7514 = 75.14\% \end{aligned}$$

Solución con Excel

DISTR.BINOM.N(0;7;0.04;FALSO)

Resultado: 0.75144748

Solución con Stata

di binomialp(7,0,.04)

Resultado: 0.75144748

♣ Pregunta 3.29

Solución con calculadora

$$\begin{aligned} P(k = 1 | n = 7, p = 0.04) &= \binom{n}{k} p^k (1-p)^{n-k} = \binom{7}{1} 0.04^1 (1-0.04)^{7-1} \\ &= \frac{7!}{1!(7-1)!} \times 0.04^1 \times 0.96^6 = 0.2192 = 21.92\% \end{aligned}$$

Solución con Excel

DISTR.BINOM.N(1;7;0.04;FALSO)

Resultado: 0.21917218

Solución con Stata

```
di binomialp(7,1,.04)
```

Resultado: 0.21917218

♣ Pregunta 3.30**Solución con calculadora**

$$\begin{aligned} P(k \geq 2 | n = 7, p = 0.04) &= 1 - P(k \leq 1 | n = 7, p = 0.04) \\ &= 1 - (P(k = 0 | n = 7, p = 0.04) \\ &\quad + P(k = 1 | n = 7, p = 0.04)) \end{aligned}$$

Pero los dos términos del interior del corchete están ya calculados de las preguntas 3.28 y 3.29. Por lo tanto:

$$\begin{aligned} P(k \geq 2 | n = 7, p = 0.04) &= 1 - P(k \leq 1 | n = 7, p = 0.04) \\ &= 1 - (0.7514 + 0.2192) = 0.0294 = 2.94\% \end{aligned}$$

Solución con Excel

```
1-DISTR.BINOM.N(1;7;0.04;VERDADERO)
```

Resultado: 0.0293034

Solución con Stata

```
di binomialtail(7,2,.04)
```

Resultado: 0.02938034

El mismo resultado se obtiene con la orden `di 1-binomial(7,1,.04)`

SOLUCIÓN 3.11

Sigue siendo una distribución binomial, pero ahora tiene $n = 2000$ y $p = 0.04$.

El problema es que manejar con la calculadora una binomial con $n = 2000$ es inviable, pero se puede utilizar la aproximación normal debido a que:

$$B(n, p) \approx N(\mu = np, \sigma^2 = np(1 - p))$$

♣ Pregunta 3.31

Solución con calculadora

$$\mu = np = 2000 \times 0.04 = 80 \text{ fallecidos}$$

$$\begin{aligned} \sigma^2 = np(1 - p) \rightarrow \sigma &= \sqrt{np(1 - p)} \\ &= \sqrt{2000 \times 0.04 \times (1 - 0.04)} = 8.76 \text{ fallecidos} \end{aligned}$$

♣ Pregunta 3.32

Solución con calculadora

Usando la aproximación normal, tenemos que calcular:

$$P(x < 60 | \mu = 80, \sigma = 8.76)$$

Para ello, primero tenemos que calcular el valor z :

$$z = \frac{x - \mu}{\sigma} = \frac{60 - 80}{8.76} = -2.28$$

A continuación, se busca en la tabla de la normal el valor p correspondiente. Si $z = +2.28$, $p = 0.9887$. Como en nuestro problema $z = -2.28$, su valor p será $1 - 0.9887 = 0.0113$: la probabilidad de que haya menos de 60 muertes en 2000 pacientes es del 1.13%.

Solución con Excel

Excel sí permite resolver el problema usando la distribución binomial, lo que permite ver hasta qué punto es buena la aproximación normal:

```
DISTR.BINOM.N(60;2000;0.04;VERDADERO)
```

Resultado: 0.01062072

Solución con Stata

Stata también permite resolver el problema usando la distribución binomial:

```
di binomial(2000,60,.04)
```

Resultado: 0.01062072

♣ Pregunta 3.33

Solución con calculadora

De nuevo, tendremos que utilizar la aproximación normal. Pero esta vez nos piden la probabilidad de que haya menos de 80 fallecimientos, cuando la media de la normal es 80. Como esta distribución es simétrica en torno a la media, la probabilidad de que haya menos de 80 fallecimientos es 50%.

Solución con Excel

```
DISTR.BINOM.N(80;2000;0.04;VERDADERO)
```

Resultado: 0.52969415

Solución con Stata

```
di binomial(2000,80,.04)
```

Resultado: 0.52969415

♣ **Pregunta 3.34****Solución con calculadora**

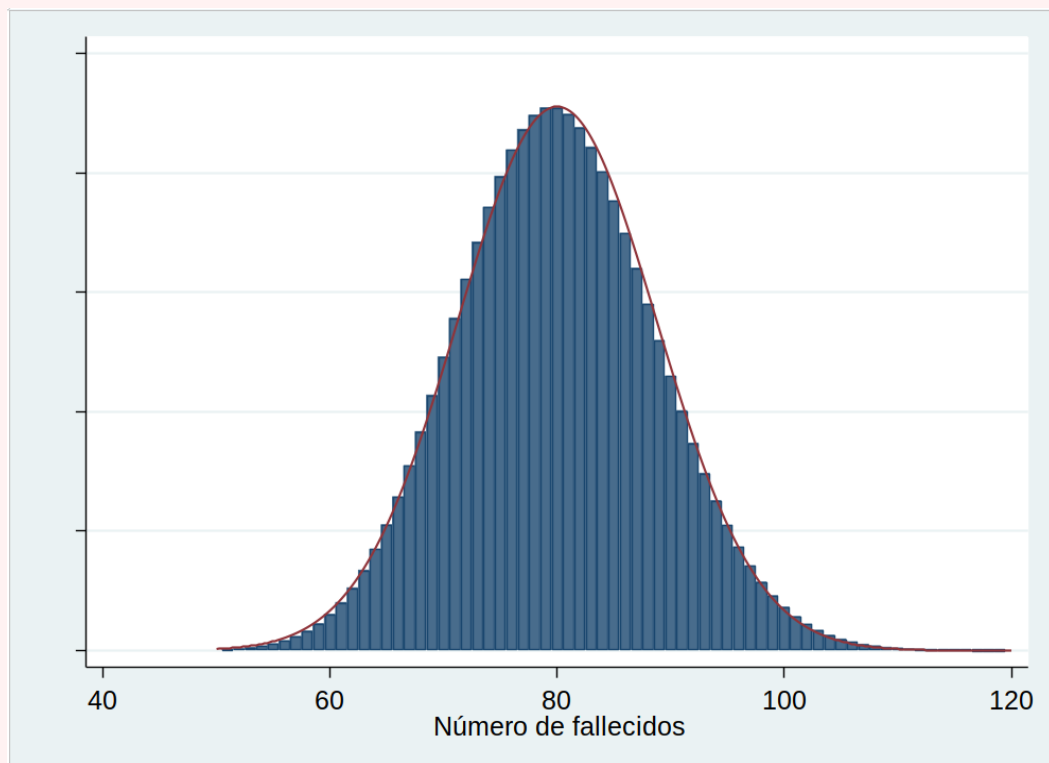
Seguimos necesitando la aproximación normal.

En primer lugar, calculemos z :

$$z = \frac{x - \mu}{\sigma} = \frac{90 - 80}{8.76} = 1.14$$

Consultando la tabla de la normal, encontramos que este valor z deja por debajo 0.8729. Por lo tanto, deja por encima 0.1271.

Resultado: la probabilidad de que se produzcan más de 90 fallecidos de 2000 pacientes es 12.71 %.



Esta figura muestra las dos distribuciones que hemos usado en este problema. En azul, las probabilidades de la distribución binomial $B(n = 2000, p = 0.04)$ y en rojo la función de densidad de la distribución normal con media $\mu = 80$ y desviación típica $\sigma = 8.76$. Puede verse que la aproximación es muy buena.

Solución con Excel

```
1-DISTR.BINOM.N(90;2000;0.04;VERDADERO)
```

Resultado: 0.116739396

Solución con Stata

```
di binomialtail(2000,91,.04)
```

Resultado: 0.1167394

El mismo resultado se obtiene con la orden `di 1-binomial(2000,90,.04)`

SOLUCIÓN 3.12

Solución con calculadora

En este problema seguimos teniendo una distribución binomial, pero con una n tan grande (125000) que no podemos usar la binomial y con una probabilidad tan pequeña ($5/125000 = 0.00004$) que la aproximación normal podría no funcionar. Para un fenómeno raro, la mejor aproximación es usar la distribución de Poisson cuya función de probabilidad es:

$$P(x = k | \mu) = \frac{\mu^k e^{-\mu}}{k!}$$

donde μ es la media.

♣ Pregunta 3.35

Solución con calculadora

Aplicado a este problema, tenemos $k = 8$ y $\mu = 5$:

$$P(x = 8 | \mu = 5) = \frac{\mu^k e^{-\mu}}{k!} = \frac{5^8 e^{-5}}{8!} = 0.0653 = 6.53\%$$

Si la media de muertes al año es 8, la probabilidad de que en un año tengamos 8 muertes es del 6.53 % (aproximadamente, un año de cada 16).

Solución con Excel

- Usando la distribución de Poisson:

`POISSON.DIST(8;5;FALSO)`

Resultado: 0.065278039

- Usando la distribución binomial:

`DISTR.BINOM.N(8; 125000; 0.00004; FALSO)`

Resultado: 0.06527778

Solución con Stata

- Usando la distribución de Poisson:

```
di poissonp(5,8)
```

Resultado: .06527804

- Usando la distribución binomial:

```
di binomialp(125000,8,.00004)
```

Resultado: 0.06527778

♣ Pregunta 3.36**Solución con calculadora**

Lo que se pide es la probabilidad de tener un año 8 muertes o más si la media anual es 5. Esto es igual a 1 menos la probabilidad de tener 7 muertes o menos:

$$\begin{aligned}
 P(x \geq 8 | \mu = 5) &= 1 - P(x \leq 7 | \mu = 5) \\
 &= 1 - (P(x = 0 | \mu = 5) + P(x = 1 | \mu = 5) + P(x = 2 | \mu = 5) \\
 &\quad + P(x = 3 | \mu = 5) + P(x = 4 | \mu = 5) + P(x = 5 | \mu = 5) \\
 &\quad + P(x = 6 | \mu = 5) + P(x = 7 | \mu = 5))
 \end{aligned}$$

Calculemos cada uno de los términos del corchete:

$$P(x = 0 | \mu = 5) = \frac{\mu^k e^{-\mu}}{k!} = \frac{5^0 e^{-5}}{0!} = 0.0067$$

$$P(x = 1 | \mu = 5) = \frac{\mu^k e^{-\mu}}{k!} = \frac{5^1 e^{-5}}{1!} = 0.0337$$

$$P(x = 2 | \mu = 5) = \frac{\mu^k e^{-\mu}}{k!} = \frac{5^2 e^{-5}}{2!} = 0.0842$$

$$P(x = 3 | \mu = 5) = \frac{\mu^k e^{-\mu}}{k!} = \frac{5^3 e^{-5}}{3!} = 0.1404$$

$$P(x = 4 | \mu = 5) = \frac{\mu^k e^{-\mu}}{k!} = \frac{5^4 e^{-5}}{4!} = 0.1755$$

$$P(x = 5 | \mu = 5) = \frac{\mu^k e^{-\mu}}{k!} = \frac{5^5 e^{-5}}{5!} = 0.1755$$

$$P(x = 6 | \mu = 5) = \frac{\mu^k e^{-\mu}}{k!} = \frac{5^6 e^{-5}}{6!} = 0.1462$$

$$P(x = 7 | \mu = 5) = \frac{\mu^k e^{-\mu}}{k!} = \frac{5^7 e^{-5}}{7!} = 0.1044$$

Finalmente:

$$\begin{aligned} P(x \geq 8 | \mu = 5) &= 1 - P(x \leq 7 | \mu = 5) \\ &= 1 - (0.0067 + 0.0337 + 0.0842 + 0.1404 \\ &\quad + 0.1755 + 0.1755 + 0.1462 + 0.1044) = 0.1344 = 13.44\% \end{aligned}$$

La probabilidad de que un año hubiera 8 muertes o más es 13.44%, aproximadamente, un año de cada 7.

Solución con Excel

- Usando la distribución de Poisson:

`1-POISSON.DIST(7;5;VERDADERO)`

Resultado: 0.133371674

- Usando la distribución binomial:

`1-DISTR.BINOM.N(7;125000;0.00004;VERDADERO)`

Resultado: 0.133367496

- Usando la aproximación normal:

Para usarla, ya conocemos la media (5), pero necesitamos calcular la desviación típica:

$$\sigma = \sqrt{np(1-p)} = \sqrt{125000 \times 0.00004 \times (1 - 0.00004)} = 2.236$$

`1-DISTR.NORM.N(8;5;2.236;VERDADERO)`

Resultado: 0.089849632

Solución con Stata

- Usando la distribución de Poisson:

```
di poissontail(5,8)
```

Resultado: 0.13337167

- Usando la distribución binomial:

```
di binomialtail(125000,8,0.00004)
```

Resultado: 0.1333675

El mismo resultado se obtiene con la orden:

```
di 1-binomial(125000,7,.00004)
```

- Usando la aproximación normal:

Para usarla, ya conocemos la media (5), pero necesitamos calcular la desviación típica:

$$\sigma = \sqrt{np(1-p)} = \sqrt{125000 \times 0.00004 \times (1 - 0.00004)} = 2.236$$

```
di 1-normal((8-5)/2.236)
```

Resultado: 0.08984963

Se comprueba que la distribución de Poisson aproxima muy bien a la binomial, pero la normal comete un error importante.

SOLUCIÓN 3.13

De nuevo tenemos una distribución binomial, pero esta vez ni siquiera nos dan la n , con lo que es imposible utilizar la binomial o la aproximación normal. La única opción es la distribución de Poisson.

Observe que, aunque no nos dan n , sí sabemos que la probabilidad es baja porque en cualquier provincia 9 / n será un número pequeño.

♣ Pregunta 3.37

Solución con calculadora

Aplicado a este problema, tenemos $k = 6$ y $\mu = 9$:

$$P(x = 6 | \mu = 9) = \frac{\mu^k e^{-\mu}}{k!} = \frac{9^6 e^{-9}}{6!} = 0.0911 = 9.11 \%$$

Si la media de casos al año es 9, la probabilidad de que en un año tengamos 6 casos es del 9.11 % (aproximadamente, un año de cada 9); no es una probabilidad baja.

Solución con Excel

`POISSON.DIST(6;9;FALSO)`

Resultado: 0.091090319

Solución con Stata

Usando la distribución de Poisson:

`di poissonp(9, 6)`

Resultado: 0.09109032

♣ Pregunta 3.38

Solución con calculadora

Lo que se pide es la probabilidad de tener un año 6 casos o menos si la media anual es 9.

$$\begin{aligned} P(x < 6 | \mu = 9) \\ &= P(x = 0 | \mu = 9) + P(x = 1 | \mu = 9) + P(x = 2 | \mu = 9) \\ &\quad + P(x = 3 | \mu = 9) + P(x = 4 | \mu = 9) \\ &\quad + P(x = 5 | \mu = 9) + P(x = 6 | \mu = 9) \end{aligned}$$

Si calculamos cada uno de los términos:

$$P(x = 0 | \mu = 9) = \frac{\mu^k e^{-\mu}}{k!} = \frac{9^0 e^{-9}}{0!} = 0.0001$$

$$P(x = 1 | \mu = 9) = \frac{\mu^k e^{-\mu}}{k!} = \frac{9^1 e^{-9}}{1!} = 0.0011$$

$$P(x = 2 | \mu = 9) = \frac{\mu^k e^{-\mu}}{k!} = \frac{9^2 e^{-9}}{2!} = 0.0050$$

$$P(x = 3 | \mu = 9) = \frac{\mu^k e^{-\mu}}{k!} = \frac{9^3 e^{-9}}{3!} = 0.0150$$

$$P(x = 4 | \mu = 9) = \frac{\mu^k e^{-\mu}}{k!} = \frac{9^4 e^{-9}}{4!} = 0.0337$$

$$P(x = 5 | \mu = 9) = \frac{\mu^k e^{-\mu}}{k!} = \frac{9^5 e^{-9}}{5!} = 0.0607$$

$$P(x = 6 | \mu = 9) = \frac{\mu^k e^{-\mu}}{k!} = \frac{9^6 e^{-9}}{6!} = 0.0911$$

Sumando, obtenemos:

$$\begin{aligned} P(x < 6 | \mu = 9) &= 0.0001 + 0.0011 + 0.0050 + 0.0150 + 0.0337 + 0.0607 + 0.0911 \\ &= 0.2076 = 20.76\% \end{aligned}$$

Si la media de casos anual es 9, en el 20 % de los años (un año de cada 5) se producirán 6 casos o menos. Esto indica que tener solo 6 casos no es una prueba de que la vacuna esté funcionando.

Solución con Excel

```
POISSON.DIST(6;9;VERDADERO)
```

Resultado: 0.20678084

Solución con Stata

```
di poisson(9,6)
```

Resultado: 0.20678084

SOLUCIÓN 3.14

♣ Pregunta 3.39**Solución con calculadora**

La diferencia con los problemas anteriores es que ahora la media es alta. Eso permite utilizar la aproximación normal sabiendo que:

$$P(\mu) \rightarrow N(\mu, \mu)$$

Es decir, una distribución de Poisson tiende a una normal en la que la media es igual a la varianza.

En este caso, como la media es 27, tendríamos:

$$\sigma^2 = 27$$

$$\sigma = \sqrt{27} = 5.20$$

Es decir, lo que tenemos que calcular para el problema es la probabilidad de obtener un valor 33 o mayor en una distribución normal de media 27 y desviación típica 5.20.

Primero, calculamos el valor z correspondiente a 33:

$$z = \frac{x - \mu}{\sigma} = \frac{33 - 27}{5.20} = 1.15$$

A continuación, consultamos esta z en la tabla normal. Encontramos que este valor deja por debajo el 87.49% de la distribución. Por lo tanto, lo que deja por encima es el 12.51%.

Resultado final: si la media es 27 casos al año, en 12.51% de los años se producirán 33 casos o más. Este no es un resultado infrecuente: ocurre en un caso de cada 8.

Solución con Stata

- Usando la aproximación normal: `di 1-normal((33-27)/5.2)`

Resultado: .12428162

- Usando la distribución de Poisson: `di poissontail(27,33)`

Resultado: .14537697

SOLUCIÓN 3.15

♣ Pregunta 3.40
Solución con calculadora

De acuerdo con el teorema de la probabilidad total,

$$P(I) = P(I \cap D) + P(I \cap D^C) = P(D)P(I|D) + P(D^C)P(I|D^C)$$

, donde I y D son los eventos “infectarse” y “mantener una distancia interpersonal de más de un metro”, respectivamente. Así,

$$P(I) = (1 - 0.2) \times 0.026 + 0.2 \times 0.128 = 0.0464$$

♣ Pregunta 3.41
Solución con calculadora

El número de infectados en las cenas de Nochebuena estaría representado por una variable aleatoria binomial; en concreto,

$$X \sim \text{Binomial}(n = 6, p = 0.128)$$

♣ Pregunta 3.42
Solución con calculadora

Sea

$$X \sim \text{Binomial}(n = 20, p = 0.128)$$

una variable aleatoria que representa el número de infectados. Así,

$$P(X \geq 2) = P(X \geq 1) = 1 - P(X \leq 1) = 1 - P(X = 0) - P(X = 1)$$

donde

$$P(X = 0) = \binom{20}{0} 0.128^0 (1 - 0.128)^{20-0} = 0.872^{20} = 0.0646$$

$$P(X = 1) = \binom{20}{1} 0.128^1 (1 - 0.128)^{20-1} = 20 \times 0.128 \times 0.872^{19} = 0.1897$$

Por tanto,

$$P(X \geq 2) = 1 - 0.0646 - 0.1897 = 0.7457$$

♣ Pregunta 3.43

Solución con calculadora

Sea

$$X \sim \text{Binomial}(n = 1000, p = 0.026)$$

una variable aleatoria que representa el número de infectados. Nótese que

$$P(X = 10) \approx P(Y = 10),$$

donde

$$Y \sim \text{Poisson}(\lambda = np = 26)$$

. Así,

$$P(Y = 10) = e^{-26} \frac{26^{10}}{10!} = 0.0002$$

4

Estimación de una media

SOLUCIÓN 4.1

♣ **Pregunta 4.1****Solución con calculadora**

$$m = \frac{\sum_{i=1}^n x_i}{n} = \frac{41808}{636} = 65.736$$

$$s^2 = \frac{\sum_{i=1}^n (x_i - m)^2}{n - 1} = \frac{133007.6}{636 - 1} = 209.461$$

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - m)^2}{n - 1}} = \sqrt{209.461} = 14.473$$

$$EEM = \frac{s}{\sqrt{n}} = \frac{14.473}{\sqrt{636}} = 0.574$$

Para IC 95%: $z_{\alpha/2} = 1.96$

$$IC \text{ superior} = m + z_{\alpha/2} \times \frac{s}{\sqrt{n}} = 65.736 + (1.96 \times 0.574) = 66.861$$

$$IC \text{ inferior} = m - z_{\alpha/2} \times \frac{s}{\sqrt{n}} = 65.736 - (1.96 \times 0.574) = 64.611$$

Interpretaciones correctas:

- Tenemos una confianza del 95% de que la media de la población está entre 64.6 y 66.8.
- En mi muestra he obtenido una media de 65.7, pero si repitiésemos el estudio 100 veces en 95 ocasiones los valores estarían comprendidos entre 64.6 y 66.9

Solución con Excel

Se genera una etiqueta con todos los valores de la edad denominada *edad*. Representa a toda la muestra y facilita el uso de fórmulas en **Excel**.

Para resolver el problema utilizaremos las siguientes funciones:

- Para obtener la n : CONTAR
- Para obtener la m : PROMEDIO

- Para obtener la s^2 : VAR.S
- Para obtener la s : RAIZ

	Valor	Función
n	636	CONTAR(edad)
m	65.736	PROMEDIO(edad)
S^2	209.461	VAR.S(edad)
s	14.473	RAIZ(S^2)
EEM	0.574	s/RAIZ(n)
IC 95 % superior	66.861	m+1.96*EEM
IC95 % inferior	64.611	m-1.96*EEM

Utilizando el analizador de datos: Estadística descriptiva .

<i>Edad Muestra</i>	
Media	65.73585
Error típico	0.573882
Mediana	68
Moda	79
Desviación estándar	14.47276
Varianza de la muestra	209.4608
Curtosis	0.090198
Coefficiente de asimetría	-0.76964
Rango	76
Mínimo	17
Máximo	93
Suma	41808
Cuenta	636
Nivel de confianza(95.0%)	1.126937

En la salida del Excel, el intervalo de confianza, en realidad es la precisión del intervalo, es decir, la mitad del intervalo. Por tanto, es necesario sumar y restar esta precisión para obtener el IC.

Media	65.736
Precisión	1.127
LIC	64.609
LSC	66.862

Nota: La resolución mediante el analizador de datos solo se muestra en el primer ejemplo.

Solución con Stata

Para obtener la media y el intervalo de confianza utilizamos el comando `mean` con la opción `level` :

```
. mean Edad, level(95)
```

Mean estimation

Number of obs = 636

	Mean	Std. Err.	[95% Conf. Interval]	
Edad	65.73585	.5738823	64.60891	66.86279

♣ Pregunta 4.2**Solución con calculadora**

$$m = \frac{\sum_{i=1}^n x_i}{n} = \frac{20423}{318} = 64.223$$

$$s^2 = \frac{\sum_{i=1}^n (x_i - m)^2}{n - 1} = \frac{63989.15}{318 - 1} = 201.859$$

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - m)^2}{n - 1}} = \sqrt{201.859} = 14.208$$

$$EEM = \frac{s}{\sqrt{n}} = \frac{14.208}{\sqrt{318}} = 0.797$$

Para IC 95%: $z_{\alpha/2} = 1.96$

$$IC \text{ superior} = m + z_{\alpha/2} \times \frac{s}{\sqrt{n}} = 64.223 + (1.96 \times 0.797) = 65.785$$

$$IC \text{ inferior} = m - z_{\alpha/2} \times \frac{s}{\sqrt{n}} = 64.223 - (1.96 \times 0.797) = 62.662$$

Interpretaciones correctas:

- Tenemos una confianza del 95% de que la media de los posibles controles de la población está entre 62.7 y 65.8.
- En mi muestra he obtenido una media de 64.2. pero si repitiésemos el estudio 100 veces en 95 ocasiones los valores estarían comprendidos entre 62.7 y 65.8.

Solución con Excel

Se genera una etiqueta con todos los valores de la edad para los controles (caso=0) denominada *edad_0*. Representa la muestra de controles y facilita el uso de fórmulas en Excel.

	Valor	Función
n	318	CONTAR(edad_0)
m	64.223	PROMEDIO(edad_0)
S ²	201.859	VAR.S(edad_0)
s	14.208	RAIZ(S ²)
EEM	0.797	s/RAIZ(n)
IC 95 % superior	65.785	m+1.96*EEM
IC95 % inferior	62.662	m-1.96*EEM

Solución con Stata

```
. mean Edad if Caso==0, level(95)
```

Mean estimation

Number of obs = 318

	Mean	Std. Err.	[95% Conf. Interval]	
Edad	64.22327	.7967278	62.65573	65.79081

♣ Pregunta 4.3

Solución con calculadora

$$m = \frac{\sum_{i=1}^n x_i}{n} = \frac{21385}{318} = 67.248$$

$$s^2 = \frac{\sum_{i=1}^n (x_i - m)^2}{n - 1} = \frac{67563.37}{318 - 1} = 213.134$$

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - m)^2}{n - 1}} = \sqrt{213.134} = 14.599$$

$$EEM = \frac{s}{\sqrt{n}} = \frac{14.599}{\sqrt{318}} = 0.819$$

Para IC 95%: $z_{\alpha/2} = 1.96$

$$IC\ superior = m + z_{\alpha/2} \times \frac{s}{\sqrt{n}} = 67.248 + (1.96 \times 0.819) = 68.853$$

$$IC\ inferior = m - z_{\alpha/2} \times \frac{s}{\sqrt{n}} = 67.248 - (1.96 \times 0.819) = 65.644$$

Interpretaciones correctas:

- Tenemos una confianza del 95% de que la media de los casos de la población está entre 65.6 y 68.9.
- En mi muestra he obtenido una media de 67.2. pero si repitiésemos el estudio 100 veces en 95 ocasiones los valores estarían comprendidos entre 65.6 y 68.9.

Solución con Excel

Se genera una etiqueta con todos los valores de la edad para los casos (caso=1) denominada *edad_1*. Representa la muestra de casos y facilita el uso de fórmulas en Excel.

	Valor	Función
n	318	CONTAR(edad_1)
m	67.248	PROMEDIO(edad_1)
S ²	213.134	VAR.S(edad_1)
s	14.599	RAIZ(S ²)
EEM	0.819	s/RAIZ(n)
IC 95% superior	68.853	m+1.96*EEM
IC95% inferior	65.644	m-1.96*EEM

Solución con Stata

```
. mean Edad if Caso==1, level(95)
```

Mean estimation

Number of obs = 318

	Mean	Std. Err.	[95% Conf. Interval]
Edad	67.24843	.8186768	65.6377 68.85915

SOLUCIÓN 4.2

♣ Pregunta 4.4

Solución con calculadora

$$m = \frac{\sum_{i=1}^n x_i}{n} = \frac{15970}{634} = 25.189$$

$$s^2 = \frac{\sum_{i=1}^n (x_i - m)^2}{n - 1} = \frac{28781.29}{634 - 1} = 45.468$$

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - m)^2}{n - 1}} = \sqrt{45.468} = 6.743$$

$$EEM = \frac{s}{\sqrt{n}} = \frac{6.743}{\sqrt{634}} = 0.268$$

Para IC 90%: $z_{\alpha/2} = 1.64$

$$IC \text{ superior} = m + z_{\alpha/2} \times \frac{s}{\sqrt{n}} = 25.189 + (1.64 \times 0.268) = 25.628$$

$$IC \text{ inferior} = m - z_{\alpha/2} \times \frac{s}{\sqrt{n}} = 25.189 - (1.64 \times 0.268) = 24.750$$

Interpretaciones correctas:

- Tenemos una confianza del 90% de que la media de Apache II de la población está entre 24.8 y 25.6.
- En mi muestra he obtenido una media de Apache II de 25.19 pero si repitiésemos el estudio 100 veces en 90 ocasiones los valores estarían comprendidos entre 24.8 y 25.6.

Solución con Excel

Se genera una etiqueta con todos los valores de Apache II denominada *apache*. Representa a toda la muestra y facilita el uso de fórmulas en Excel.

	Valor	Función
n	634	CONTAR (apache)
m	25.189	PROMEDIO (apache)
S ²	45.468	VAR.S (apache)
s	6.743	RAIZ (S ²)
EEM	0.268	s/RAIZ (n)
IC 90% superior	25.628	m+1.64*EEM
IC 90% inferior	24.750	m-1.64*EEM

Solución con Stata

```
. mean Apache, level(90)
```

Mean estimation

Number of obs = 634

	Mean	Std. Err.	[90% Conf. Interval]	
Apache	25.18927	.2677988	24.74814	25.63041

♣ Pregunta 4.5

Solución con calculadora

$$m = \frac{\sum_{i=1}^n x_i}{n} = \frac{7901}{318} = 24.846$$

$$s^2 = \frac{\sum_{i=1}^n (x_i - m)^2}{n-1} = \frac{12221.45}{318-1} = 38.553$$

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - m)^2}{n-1}} = \sqrt{38.553} = 6.209$$

$$EEM = \frac{s}{\sqrt{n}} = \frac{6.209}{\sqrt{318}} = 0.348$$

Para IC 90%: $z_{\alpha/2} = 1.64$

$$IC \text{ superior} = m + z_{\alpha/2} \times \frac{s}{\sqrt{n}} = 24.846 + (1.64 \times 0.348) = 25.417$$

$$IC inferior = m - z_{\alpha/2} \times \frac{s}{\sqrt{n}} = 24.846 - (1.64 \times 0.348) = 24.275$$

Interpretaciones correctas:

- Tenemos una confianza del 90% de que la media de Apache II de la población de controles está entre 24.3 y 25.4.
- En mi muestra he obtenido una media de Apache II de 24.8 pero si repitiésemos el estudio 100 veces en 90 ocasiones los valores estarían comprendidos entre 24.3 y 25.4.

Solución con Excel

Se genera una etiqueta con todos los valores de Apache II para los controles (caso=0) denominada *apache_0*. Representa la muestra de controles y facilita el uso de fórmulas en **Excel**.

	Valor	Función
n	318	CONTAR(<i>apache_0</i>)
m	24.846	PROMEDIO(<i>apache_0</i>)
S ²	38.553	VAR.S(<i>apache_0</i>)
s	6.209	RAIZ(S ²)
EEM	0.348	s/RAIZ(n)
IC 90% superior	25.417	m+1.64*EEM
IC 90% inferior	24.275	m-1.64*EEM

Solución con Stata

```
. mean Apache if Caso==0, level(90)
```

Mean estimation Number of obs = 318

	Mean	Std. Err.	[90% Conf. Interval]	
Apache	24.84591	.3481915	24.27151	25.42031

♣ Pregunta 4.6

Solución con calculadora

$$m = \frac{\sum_{i=1}^n x_i}{n} = \frac{8069}{316} = 25.535$$

$$s^2 = \frac{\sum_{i=1}^n (x_i - m)^2}{n - 1} = \frac{16483.62}{316 - 1} = 52.332$$

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - m)^2}{n - 1}} = \sqrt{52.332} = 7.234$$

$$EEM = \frac{s}{\sqrt{n}} = \frac{7.234}{\sqrt{316}} = 0.407$$

Para IC 90%: $z_{\alpha/2} = 1.64$

$$IC \text{ superior} = m + z_{\alpha/2} \times \frac{s}{\sqrt{n}} = 25.535 + (1.64 \times 0.407) = 26.202$$

$$IC \text{ inferior} = m - z_{\alpha/2} \times \frac{s}{\sqrt{n}} = 25.535 - (1.64 \times 0.407) = 24.867$$

Interpretaciones correctas:

- Tenemos una confianza del 90% de que la media de Apache II de la población susceptible de ser caso está entre 24.9 y 26.2.
- En mi muestra he obtenido una media de Apache II en los casos de 35.5 pero si repitiésemos el estudio 100 veces en 90 ocasiones los valores estarían comprendidos entre 24.9 y 26.2.

Solución con Excel

Se genera una etiqueta con todos los valores de Apache II para los casos (caso=1) denominada *apache_1*. Representa la muestra de casos y facilita el uso de fórmulas en Excel.

	Valor	Función
n	316	CONTAR(<i>apache_1</i>)
m	25.535	PROMEDIO(<i>apache_1</i>)
S ²	52.332	VAR.S(<i>apache_1</i>)
s	7.234	RAIZ(S ²)
EEM	0.407	s/RAIZ(n)
IC 90% superior	26.202	m+1.64*EEM
IC 90% inferior	24.867	m-1.64*EEM

Solución con Stata

```
. mean Apache if Caso==1, level(90)
```

Mean estimation

Number of obs = 316

	Mean	Std. Err.	[90% Conf. Interval]	
Apache	25.53481	.4069496	24.86346	26.20616

SOLUCIÓN 4.3

♣ **Pregunta 4.7****Solución con calculadora**

$$m = \frac{\sum_{i=1}^n x_i}{n} = \frac{1522.71}{627} = 2.429$$

$$s^2 = \frac{\sum_{i=1}^n (x_i - m)^2}{n - 1} = \frac{2621.29}{627 - 1} = 4.187$$

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - m)^2}{n - 1}} = \sqrt{4.187} = 2.046$$

$$EEM = \frac{s}{\sqrt{n}} = \frac{2.046}{\sqrt{627}} = 0.082$$

Para IC 80%: $z_{\alpha/2} = 1.28$

$$IC \text{ superior} = m + z_{\alpha/2} \times \frac{s}{\sqrt{n}} = 2.429 + (1.28 \times 0.082) = 2.533$$

$$IC \text{ inferior} = m - z_{\alpha/2} \times \frac{s}{\sqrt{n}} = 2.429 - (1.28 \times 0.082) = 2.324$$

Solución con Excel

Se genera una etiqueta con todos los valores de creatinina en sangre denominada *creatinina*. Representa a toda la muestra y facilita el uso de fórmulas en **Excel**.

	Valor	Función
n	627	CONTAR(creatinina)
m	2.429	PROMEDIO(creatinina)
S ²	4.187	VAR.S(creatinina)
s	2.046	RAIZ(S ²)
EEM	0.082	s/RAIZ(n)
IC 80% superior	2.533	m+1.28*EEM
IC 80% inferior	2.324	m-1.28*EEM

Solución con Stata

```
. mean CreatIng, level(80)
```

Mean estimation

Number of obs = 627

	Mean	Std. Err.	[80% Conf. Interval]	
CreatIng	2.428565	.0817216	2.323724	2.533406

♣ Pregunta 4.8**Solución con calculadora**

$$m = \frac{\sum_{i=1}^n x_i}{n} = \frac{571.32}{313} = 1.825$$

$$s^2 = \frac{\sum_{i=1}^n (x_i - m)^2}{n - 1} = \frac{597.065}{313 - 1} = 1.914$$

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - m)^2}{n - 1}} = \sqrt{1.914} = 1.383$$

$$EEM = \frac{s}{\sqrt{n}} = \frac{1.383}{\sqrt{313}} = 0.078$$

Para IC 80%: $z_{\alpha/2} = 1.28$

$$IC \text{ superior} = m + z_{\alpha/2} \times \frac{s}{\sqrt{n}} = 1.825 + (1.28 \times 0.078) = 1.925$$

$$IC \text{ inferior} = m - z_{\alpha/2} \times \frac{s}{\sqrt{n}} = 1.825 - (1.28 \times 0.078) = 1.725$$

Solución con Excel

Se genera una etiqueta con todos los valores de creatinina en sangre para los controles (caso=0) denominada *creatinina_0*. Representa la muestra de los controles y facilita el uso de fórmulas en Excel.

	Valor	Función
n	313	CONTAR(creatinina_0)
m	1.825	PROMEDIO(creatinina_0)
S ²	1.914	VAR.S(creatinina_0)
s	1.383	RAIZ(S ²)
EEM	0.078	s/RAIZ(n)
IC 80% superior	1.925	m+1.28*EEM
IC 80% inferior	1.725	m-1.28*EEM

Solución con Stata

```
. mean CreatIng if Caso==0, level(80)
```

Mean estimation

Number of obs = 313

	Mean	Std. Err.	[80% Conf. Interval]	
CreatIng	1.825304	.0781918	1.724884	1.925723

♣ Pregunta 4.9

Solución con calculadora

$$m = \frac{\sum_{i=1}^n x_i}{n} = \frac{951.39}{314} = 3.030$$

$$s^2 = \frac{\sum_{i=1}^n (x_i - m)^2}{n-1} = \frac{951.39}{314-1} = 5.740$$

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - m)^2}{n-1}} = \sqrt{5.740} = 2.396$$

$$EEM = \frac{s}{\sqrt{n}} = \frac{2.396}{\sqrt{314}} = 0.135$$

Para IC 80%: $z_{\alpha/2} = 1.28$

$$IC \text{ superior} = m + z_{\alpha/2} \times \frac{s}{\sqrt{n}} = 3.030 + (1.28 \times 0.135) = 3.203$$

$$IC\ inferior = m - z_{\alpha/2} \times \frac{s}{\sqrt{n}} = 3.030 - (1.28 \times 0.135) = 2.857$$

Solución con Excel

Se genera una etiqueta con todos los valores de creatinina en sangre para los casos (caso=1) denominada *creatinina_1*. Representa la muestra de los casos y facilita el uso de fórmulas en Excel.

	Valor	Función
n	314	CONTAR(creatinina_1)
m	3.030	PROMEDIO(creatinina_1)
S ²	5.740	VAR.S(creatinina_1)
s	2.396	RAIZ(S ²)
EEM	0.135	s/RAIZ(n)
IC 80% superior	3.203	m+1.28*EEM
IC 80% inferior	2.857	m-1.28*EEM

Solución con Stata

```
. mean CreatIng if Caso==1, level(80)
```

Mean estimation

Number of obs = 314

	Mean	Std. Err.	[95% Conf. Interval]	
CreatIng	3.029904	.1352103	2.856259	3.20355

SOLUCIÓN 4.4

♣ Pregunta 4.10

Solución con calculadora

$$m = \frac{\sum_{i=1}^n x_i}{n} = \frac{54358.59}{625} = 86.974$$

$$s^2 = \frac{\sum_{i=1}^n (x_i - m)^2}{n - 1} = \frac{3082005}{625 - 1} = 4939.111$$

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - m)^2}{n - 1}} = \sqrt{4939.111} = 70.279$$

$$EEM = \frac{s}{\sqrt{n}} = \frac{70.279}{\sqrt{625}} = 2.811$$

Para IC 99%: $z_{\alpha/2} = 2.58$

$$IC \text{ superior} = m + z_{\alpha/2} \times \frac{s}{\sqrt{n}} = 86.974 + (2.58 \times 2.811) = 94.227$$

$$IC \text{ inferior} = m - z_{\alpha/2} \times \frac{s}{\sqrt{n}} = 86.974 - (2.58 \times 2.811) = 79.721$$

Solución con Excel

Se genera una etiqueta con todos los valores de urea en sangre denominada *urea*. Representa a toda la muestra y facilita el uso de fórmulas en Excel.

	Valor	Función
n	625	CONTAR(urea)
m	86.974	PROMEDIO(urea)
S ²	4939.111	VAR.S(urea)
s	70.279	RAIZ(S ²)
EEM	2.811	s/RAIZ(n)
IC 99% superior	94.227	m+2.58*EEM
IC 99% inferior	79.721	m-2.58*EEM

Solución con Stata

```
. mean UreaIng, level(99)
```

Mean estimation

Number of obs = 625

	Mean	Std. Err.	[99% Conf. Interval]	
UreaIng	86.97374	2.811152	79.71048	94.237

♣ Pregunta 4.11**Solución con calculadora**

$$m = \frac{\sum_{i=1}^n x_i}{n} = \frac{21868.83}{311} = 70.318$$

$$s^2 = \frac{\sum_{i=1}^n (x_i - m)^2}{n - 1} = \frac{901446.9}{311 - 1} = 2907.893$$

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - m)^2}{n - 1}} = \sqrt{2907.893} = 53.925$$

$$EEM = \frac{s}{\sqrt{n}} = \frac{53.925}{\sqrt{311}} = 3.058$$

Para IC 99%: $z_{\alpha/2} = 2.58$

$$IC \text{ superior} = m + z_{\alpha/2} \times \frac{s}{\sqrt{n}} = 70.318 + (2.58 \times 3.058) = 78.207$$

$$IC \text{ inferior} = m - z_{\alpha/2} \times \frac{s}{\sqrt{n}} = 70.318 - (2.58 \times 3.058) = 62.429$$

Solución con Excel

Se genera una etiqueta con todos los valores de urea en sangre para los controles (caso=0) denominada *urea_0*. Representa la muestra de los controles y facilita el uso de fórmulas en Excel.

	Valor	Función
n	311	CONTAR(urea_0)
m	70.318	PROMEDIO(urea_0)
S ²	2907.893	VAR.S(urea_0)
s	53.925	RAIZ(S ²)
EEM	3.058	s/RAIZ(n)
IC 99% superior	78.207	m+2.58*EEM
IC 99% inferior	62.429	m-2.58*EEM

Solución con Stata

```
. mean UreaIng if Caso==0, level(99)
```

Mean estimation

Number of obs = 311

	Mean	Std. Err.	[99% Conf. Interval]
UreaIng	70.31777	3.0578	62.39262 78.24292

♣ Pregunta 4.12

Solución con calculadora

$$m = \frac{\sum_{i=1}^n x_i}{n} = \frac{32489.76}{314} = 103.471$$

$$s^2 = \frac{\sum_{i=1}^n (x_i - m)^2}{n-1} = \frac{2008827}{314-1} = 6417.976$$

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - m)^2}{n-1}} = \sqrt{6417.976} = 80.112$$

$$EEM = \frac{s}{\sqrt{n}} = \frac{80.112}{\sqrt{314}} = 4.521$$

Para IC 99%: $z_{\alpha/2} = 2.58$

$$IC \text{ superior} = m + z_{\alpha/2} \times \frac{s}{\sqrt{n}} = 103.471 + (2.58 \times 4.521) = 115.135$$

$$IC inferior = m - z_{\alpha/2} \times \frac{s}{\sqrt{n}} = 103.471 - (2.58 \times 4.521) = 91.806$$

Solución con Excel

Se genera una etiqueta con todos los valores de urea en sangre para los casos (caso=1) denominada *urea_1*. Representa la muestra de los casos y facilita el uso de fórmulas en Excel.

	Valor	Función
n	314	CONTAR(urea_1)
m	103.471	PROMEDIO(urea_1)
S ²	6417.976	VAR.S(urea_1)
s	80.112	RAIZ(S ²)
EEM	4.521	s/RAIZ(n)
IC 99 % superior	115.135	m+2.58×EEM
IC 99 % inferior	91.806	m-2.58×EEM

Solución con Stata

```
. mean UreaIng if Caso==1, level(99)
```

Mean estimation

Number of obs = 314

	Mean	Std. Err.	[99% Conf. Interval]	
UreaIng	103.4706	4.520997	91.75383	115.1873

SOLUCIÓN 4.5

♣ **Pregunta 4.13****Solución con calculadora**

$$m = \frac{\sum_{i=1}^n x_i}{n} = \frac{48425.4}{458} = 105.732$$

$$s^2 = \frac{\sum_{i=1}^n (x_i - m)^2}{n - 1} = \frac{15626.22}{458 - 1} = 34.193$$

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - m)^2}{n - 1}} = \sqrt{34.193} = 5.847$$

$$EEM = \frac{s}{\sqrt{n}} = \frac{5.847}{\sqrt{458}} = 0.273$$

Para IC 95%: $z_{\alpha/2} = 1.96$

$$IC \text{ superior} = m + z_{\alpha/2} \times \frac{s}{\sqrt{n}} = 105.732 + (1.96 \times 0.273) = 106.268$$

$$IC \text{ inferior} = m - z_{\alpha/2} \times \frac{s}{\sqrt{n}} = 105.732 - (1.96 \times 0.273) = 105.197$$

Solución con Excel

Se genera una etiqueta con todos los valores de cloro en sangre denominada *cloro*. Representa a toda la muestra y facilita el uso de fórmulas en **Excel**.

	Valor	Función
n	458	CONTAR(<i>cloro</i>)
m	105.732	PROMEDIO(<i>cloro</i>)
S ²	34.193	VAR.S(<i>cloro</i>)
s	5.847	RAIZ(S ²)
EEM	0.273	s/RAIZ(n)
IC 95 % superior	106.268	m+1.96*EEM
IC 95 % inferior	105.197	m-1.96*EEM

Solución con Stata

```
. mean C124h, level(95)
```

Mean estimation

Number of obs = 458

	Mean	Std. Err.	[95% Conf. Interval]	
C124h	105.7323	.2732349	105.1954	106.2693

♣ Pregunta 4.14**Solución con calculadora**

$$m = \frac{\sum_{i=1}^n x_i}{n} = \frac{23952}{225} = 106.453$$

$$s^2 = \frac{\sum_{i=1}^n (x_i - m)^2}{n - 1} = \frac{6977.76}{225 - 1} = 31.151$$

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - m)^2}{n - 1}} = \sqrt{31.151} = 5.581$$

$$EEM = \frac{s}{\sqrt{n}} = \frac{5.581}{\sqrt{225}} = 0.372$$

Para IC 95%: $z_{\alpha/2} = 1.96$

$$IC \text{ superior} = m + z_{\alpha/2} \times \frac{s}{\sqrt{n}} = 106.453 + (1.96 \times 0.372) = 107.183$$

$$IC \text{ inferior} = m - z_{\alpha/2} \times \frac{s}{\sqrt{n}} = 106.453 - (1.96 \times 0.372) = 105.724$$

Solución con Excel

Se genera una etiqueta con todos los valores de cloro en sangre para los controles (caso=0) denominada *cloro_0*. Representa la muestra de los controles y facilita el uso de fórmulas en Excel.

	Valor	Función
n	225	CONTAR(cloro_0)
m	106.453	PROMEDIO(cloro_0)
S ²	31.151	VAR.S(cloro_0)
s	5.581	RAIZ(S ²)
EEM	0.372	s/RAIZ(n)
IC 95% superior	107.183	m+1.96*EEM
IC 95% inferior	105.724	m-1.96*EEM

Solución con Stata

```
. mean C124h if Caso==0, level(95)
```

Mean estimation Number of obs = 225

	Mean	Std. Err.	[95% Conf. Interval]	
C124h	106.4533	.3720855	105.7201	107.1866

♣ Pregunta 4.15

Solución con calculadora

$$m = \frac{\sum_{i=1}^n x_i}{n} = \frac{24473.4}{223} = 105.036$$

$$s^2 = \frac{\sum_{i=1}^n (x_i - m)^2}{n - 1} = \frac{8418.537}{223 - 1} = 36.287$$

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - m)^2}{n - 1}} = \sqrt{36.287} = 6.024$$

$$EEM = \frac{s}{\sqrt{n}} = \frac{6.024}{\sqrt{233}} = 0.395$$

Para IC 95%: $z_{\alpha/2} = 1.96$

$$IC \text{ superior} = m + z_{\alpha/2} \times \frac{s}{\sqrt{n}} = 103.471 + (1.96 \times 0.395) = 105.810$$

$$IC inferior = m - z_{\alpha/2} \times \frac{s}{\sqrt{n}} = 103.471 - (1.96 \times 0.395) = 104.263$$

Solución con Excel

Se genera una etiqueta con todos los valores de cloro en sangre para los casos (caso=1) denominada *cloro_1*. Representa la muestra de los casos y facilita el uso de fórmulas en Excel.

	Valor	Función
n	233	CONTAR(cloro_1)
m	105.036	PROMEDIO(cloro_1)
S ²	36.287	VAR.S(cloro_1)
s	6.024	RAIZ(S ²)
EEM	0.395	s/RAIZ(n)
IC 95 % superior	105.810	m+1.96*EEM
IC 95 % inferior	104.263	m-1.96*EEM

Solución con Stata

```
. mean C124h if Caso==1, level(95)
```

Mean estimation Number of obs = 233

	Mean	Std. Err.	[95% Conf. Interval]	
C124h	105.0361	.3946357	104.2585	105.8136

SOLUCIÓN 4.6

Solución con calculadora

	Muestra		Controles		Casos	
	media	IC 95 %	media	IC 95 %	media	IC 95 %
Edad	65.736	64.611-66.861	64.223	62.662-65.785	67.248	65.644-68.853
	media	IC 90 %	media	IC 90 %	media	IC 90 %
Puntuación Apache II	25.189	24.750-25.628	24.846	24.275-25.417	25.535	24.867-26.202
	media	IC 80 %	media	IC 80 %	media	IC 80 %
Creatinina en sangre	2.429	2.324-2.533	1.825	1.725-1.925	3.030	2.857-3.203
	media	IC 99 %	media	IC 99 %	media	IC 99 %
Urea sangre	86.974	79.721-94.227	70.318	62.429-78.207	103.471	91.806-115.135
	media	IC 95 %	media	IC 95 %	media	IC 95 %
Cloro sangre	105.732	105.197-106.268	106.453	105.724-107.183	105.036	104.263-105.810

♣ Pregunta 4.16

Solución con calculadora

- Tenemos una confianza del 95 % de que la media de edad en los controles está entre 62.662 y 65.785.
- En mi muestra he obtenido una media de 64.223. pero si repitiésemos el estudio 100 veces en 95 ocasiones las edades de los controles estarían comprendidos entre 62.662 y 65.785.

♣ Pregunta 4.17

Solución con calculadora

- Tenemos una confianza del 95 % de que la media de edad en los casos está entre 65.644 y 68.853.
- En mi muestra he obtenido una media de 67.248. pero si repitiésemos el estudio 100 veces en 95 ocasiones las edades de los casos estarían comprendidos entre 65.644 y 68.853.

♣ Pregunta 4.18**Solución con calculadora**

- Tenemos una confianza del 95 % de que la puntuación Apache II de los controles está entre 24.275 y 25.417.
- En mi muestra he obtenido una media de 24.846. pero si repitiésemos el estudio 100 veces en 95 ocasiones los valores de Apache II de los controles estarían comprendidos entre 24.275 y 25.417.

♣ Pregunta 4.19**Solución con calculadora**

- Tenemos una confianza del 95 % de que la puntuación Apache II de los casos está entre 24.867 y 26.202.
- En mi muestra he obtenido una media de 25.535. pero si repitiésemos el estudio 100 veces en 95 ocasiones los valores de Apache II de los casos estarían comprendidos entre 24.867 y 26.202.

♣ Pregunta 4.20**Solución con calculadora**

- Tenemos una confianza del 95 % de que la media de urea en sangre en los controles está entre 62.429 y 78.207.
- En mi muestra he obtenido una media de 70.318. pero Si repitiésemos el estudio 100 veces en 95 ocasiones los valores de los controles estarían comprendidos entre 62.429 y 78.207.

♣ Pregunta 4.21**Solución con calculadora**

- Tenemos una confianza del 95 % de que la media de urea en sangre en los casos está entre 91.806 y 115.135.
- En mi muestra he obtenido una media de 103.471. pero Si repitiésemos el estudio 100 veces en 95 ocasiones los valores de los controles estarían comprendidos entre 91.806 y 115.135.

♣ Pregunta 4.22**Solución con calculadora**

- Tenemos una confianza del 95 % de que la media de cloro en sangre en los controles está entre 105.724 y 107.183.
- En mi muestra he obtenido una media de 106.453. pero Si repitiésemos el estudio 100 veces en 95 ocasiones los valores de los controles estarían comprendidos entre 105.724 y 107.183.

♣ Pregunta 4.23**Solución con calculadora**

- Tenemos una confianza del 95 % de que la media de urea en sangre en los casos está entre 104.263 y 105.810.
- En mi muestra he obtenido una media de 105.036. pero Si repitiésemos el estudio 100 veces en 95 ocasiones los valores de los controles estarían comprendidos entre 104.263 y 105.810.

SOLUCIÓN 4.7

♣ Pregunta 4.24

Solución con calculadora

$$m = \frac{\sum_{i=1}^n x_i}{n} = \frac{41808}{636} = 65.736$$

$$s^2 = \frac{\sum_{i=1}^n (x_i - m)^2}{n - 1} = \frac{133007.6}{636 - 1} = 209.461$$

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - m)^2}{n - 1}} = \sqrt{209.461} = 14.473$$

$$EEM = \frac{s}{\sqrt{n}} = \frac{14.473}{\sqrt{636}} = 0.574$$

Solución con Excel

Se genera una etiqueta con todos los valores de la edad denominada *edad*. Representa a toda la muestra y facilita el uso de fórmulas en Excel.

	Valor	Función
n	636	CONTAR(edad)
m	65.736	PROMEDIO(edad)
S ²	209.461	VAR.S(edad)
s	14.473	RAIZ(S ²)
EEM	0.574	s/RAIZ(n)

Solución con Stata

Para obtener la media, la desviación típica y el error estándar `tabstat` con la opción `stat`:

```
. tabstat Edad, stat(n mean sd semean)
```

Variable	N	Mean	sd	se(mean)
Edad	636	65.73585	14.47276	.5738823

♣ Pregunta 4.25

Solución con calculadora

$$m = \frac{\sum_{i=1}^n x_i}{n} = \frac{29127}{447} = 65.161$$

$$s^2 = \frac{\sum_{i=1}^n (x_i - m)^2}{n - 1} = \frac{94274.4}{447 - 1} = 211.378$$

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - m)^2}{n - 1}} = \sqrt{211.378} = 14.539$$

$$EEM = \frac{s}{\sqrt{n}} = \frac{14.539}{\sqrt{447}} = 0.688$$

Solución con Excel

Se genera una etiqueta con todos los valores de la edad en la submuestra 1 (submuestra1=1) denominada *edad_submuestra_1*. Representa a toda la muestra y facilita el uso de fórmulas en Excel.

	Valor	Función
n	447	CONTAR(edad_submuestra_1)
m	65.161	PROMEDIO(edad_submuestra_1)
S ²	211.378	VAR.S(edad_submuestra_1)
s	14.539	RAIZ(S ²)
EEM	0.688	s/RAIZ(n)

Solución con Stata

```
. tabstat Edad if submuestra1==1, stat(n mean sd semean)
```

Variable	N	Mean	sd	se(mean)
Edad	447	65.16107	14.53883	.6876631

♣ Pregunta 4.26

Solución con calculadora

$$m = \frac{\sum_{i=1}^n x_i}{n} = \frac{12143}{184} = 65.994$$

$$s^2 = \frac{\sum_{i=1}^n (x_i - m)^2}{n - 1} = \frac{40176.99}{184 - 1} = 219.546$$

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - m)^2}{n - 1}} = \sqrt{219.546} = 14.817$$

$$EEM = \frac{s}{\sqrt{n}} = \frac{14.817}{\sqrt{184}} = 1.092$$

Solución con Excel

Se genera una etiqueta con todos los valores de la edad en la submuestra 2 (submuestra2=1) denominada *edad_submuestra_2*. Representa a toda la muestra y facilita el uso de fórmulas en Excel.

	Valor	Función
n	184	CONTAR(edad_submuestra_2)
m	65.995	PROMEDIO(edad_submuestra_2)
S ²	219.546	VAR.S(edad_submuestra_2)
s	14.817	RAIZ(S ²)
EEM	1.092	s/RAIZ(n)

Solución con Stata

```
. tabstat Edad if submuestra2==1, stat(n mean sd semean)
```

Variable	N	Mean	sd	se(mean)
Edad	184	65.99457	14.8171	1.092331

♣ Pregunta 4.27

Solución con calculadora

$$m = \frac{\sum_{i=1}^n x_i}{n} = \frac{54358.59}{625} = 86.974$$

$$s^2 = \frac{\sum_{i=1}^n (x_i - m)^2}{n - 1} = \frac{3082005}{625 - 1} = 4939.111$$

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - m)^2}{n - 1}} = \sqrt{4939.111} = 70.279$$

$$EEM = \frac{s}{\sqrt{n}} = \frac{70.279}{\sqrt{625}} = 2.811$$

Solución con Excel

Se genera una etiqueta con todos los valores de urea en sangre denominada *urea*. Representa a toda la muestra y facilita el uso de fórmulas en Excel.

	Valor	Función
n	625	CONTAR(urea)
m	86.974	PROMEDIO(urea)
S ²	4939.111	VAR.S(urea)
s	70.279	RAIZ(S ²)
EEM	2.811	s/RAIZ(n)

Solución con Stata

```
. tabstat UreaIng, stat(n mean sd semean)
```

Variable	N	Mean	sd	se(mean)
UreaIng	625	86.97374	70.27881	2.811152

♣ Pregunta 4.28

Solución con calculadora

$$m = \frac{\sum_{i=1}^n x_i}{n} = \frac{37015.16}{437} = 84.703$$

$$s^2 = \frac{\sum_{i=1}^n (x_i - m)^2}{n - 1} = \frac{2168226}{437 - 1} = 4972.995$$

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - m)^2}{n - 1}} = \sqrt{4972.995} = 70.519$$

$$EEM = \frac{s}{\sqrt{n}} = \frac{70.519}{\sqrt{437}} = 3.373$$

Solución con Excel

Se genera una etiqueta con todos los valores de urea en sangre para la submuestra 1 (submuestra1=1) denominada *urea_submuestra_1*. Representa a toda la submuestra 1 y facilita el uso de fórmulas en Excel.

	Valor	Función
n	437	CONTAR(urea_submuestra_1)
m	84.703	PROMEDIO(urea_submuestra_1)
S ²	4972.995	VAR.S(urea_submuestra_1)
s	70.519	RAIZ(S ²)
EEM	3.373	s/RAIZ(n)

Solución con Stata

```
. tabstat UreaIng if submuestra1==1, stat(n mean sd semean)
```

Variable	N	Mean	sd	se(mean)
UreaIng	437	84.70289	70.51946	3.373403

♣ Pregunta 4.29

Solución con calculadora

$$m = \frac{\sum_{i=1}^n x_i}{n} = \frac{16062.29}{181} = 88.742$$

$$s^2 = \frac{\sum_{i=1}^n (x_i - m)^2}{n - 1} = \frac{1115155}{181 - 1} = 6195.307$$

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - m)^2}{n - 1}} = \sqrt{6195.307} = 78.710$$

$$EEM = \frac{s}{\sqrt{n}} = \frac{78.710}{\sqrt{181}} = 5.850$$

Solución con Excel

Se genera una etiqueta con todos los valores de urea en sangre para la submuestra 2 (submuestra2=2) denominada *urea_submuestra_2*. Representa a toda la submuestra 1 y facilita el uso de fórmulas en **Excel**.

	Valor	Función
n	181	CONTAR(urea_submuestra_2)
m	88.742	PROMEDIO(urea_submuestra_2)
S ²	6195.307	VAR.S(urea_submuestra_2)
s	78.710	RAIZ(S ²)
EEM	5.850	s/RAIZ(n)

Solución con Stata

```
. tabstat UreaIng if submuestra2==1, stat(n mean sd semean)
```

Variable	N	Mean	sd	se(mean)
UreaIng	181	88.74196	78.71027	5.850488

♣ **Pregunta 4.30**

Solución con calculadora

Edad	Todos	Submuestra1=1	Submuestra2=1
N	636	447	184
Media	65.74	65.16	65.99
Desviación típica	14.47	14.53	14.81
Error estándar de la media	0.57	0.69	1.09

Urea en sangre	Todos	Submuestra1=1	Submuestra2=1
N	625	437	181
Media	86.97	84.70	88.74
Desviación típica	70.28	70.52	78.71
Error estándar de la media	2.81	3.37	5.85

Las tres medias son parecidas, las tres desviaciones típicas son parecidas, pero el error estándar de la media es mayor y por tanto el IC es más ancho cuanto menor sea el tamaño de la muestra.

SOLUCIÓN 4.8

♣ **Pregunta 4.31****Solución con calculadora**

1. Definir las hipótesis:

$$H_0 : \mu = 2.4$$

$$H_1 : \mu \neq 2.4$$

2. Calcular el estadístico t:

$$t = \frac{m - \mu_0}{s/\sqrt{n}} = \frac{2.6 - 2.4}{2.18/\sqrt{418}} = 1.744$$

3. Calcular los n-1 grados de libertad:

$$g.l. = n - 1 = 418 - 1 = 417$$

4. Buscar
- $t = 1.744$
- en la tabla t de Student con 417 grados de libertad, obtenemos
- $p = 0.082$
- .

Conclusión: puesto que el valor-p (0.082) calculado es mayor que el nivel de significación $\alpha = 0.05$, no se puede rechazar la hipótesis nula H_0 con un nivel de significación al 5%.

Solución con Stata

Para obtener el valor de p utilizamos el comando `ttest` :

```
. ttest CreatIng=2.4 if Sexo==1
```

One-sample t test

Variable	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.Interval]	
CreatIng	418	2.585981	.1066428	2.180318	2.376356	2.795605
mean = mean(CreatIng)				t = 1.7440		
Ho: mean = 2.4				degrees of freedom = 417		
Ha: mean <2.4		Ha: mean != 2.4		Ha: mean >2.4		
Pr(T <t) = 0.9590		Pr(T > t) = 0.0819		Pr(T >t) = 0.0410		

Interpretación de la prueba, según la hipótesis planteada:

- H_0 : La media es igual a 2.4.
- El valor-p cuando H_1 es diferente de 2.4 es igual a 0.0819.

Conclusión: puesto que el valor-p calculado es mayor que el nivel de significación $\alpha = 0.05$, no se puede rechazar la hipótesis nula H_0 .

♣ Pregunta 4.32

Solución con calculadora

1. Definir las hipótesis:

$$H_0 : \mu = 1.7$$

$$H_1 : \mu \neq 1.7$$

2. Calcular el estadístico t:

$$t = \frac{m - \mu_0}{s/\sqrt{n}} = \frac{2.1 - 1.7}{1.70/\sqrt{209}} = 3.499$$

3. Calcular los n-1 grados de libertad:

$$g.l. = n - 1209 - 1 = 208$$

4. Buscar $t = 3.499$ en la tabla t de Student con 208 grados de libertad, obtenemos $p = 0.001$.

Conclusión: puesto que el valor-p computado es menor que el nivel de significación $\alpha = 0.05$, se debe rechazar la hipótesis nula H_0 , y aceptar la hipótesis alternativa H_1 con un nivel de significación al 5%.

Solución con Stata

```
. ttest CreatIng=1.7 if Sexo==0
```

One-sample t test

Variable	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.Interval]	
CreatIng	209	2.113732	.1182517	1.709545	1.880606	2.346858
mean = mean(CreatIng)				t = 3.4987		
Ho: mean = 1.7				degrees of freedom = 208		
Ha: mean <1.7		Ha: mean != 1.7		Ha: mean >1.7		
Pr(T <t) = 0.9997		Pr(T > t) = 0.0006		Pr(T >t) = 0.0003		

Interpretación de la prueba, según la hipótesis planteada:

- H_0 : La media es igual a 1.7.
- El valor-p cuando H_1 es diferente de 1.7 es igual a 0.0006.

Conclusión: puesto que el valor-p computado es menor que el nivel de significación $\alpha = 0.05$, se debe rechazar la hipótesis nula H_0 , y aceptar la hipótesis alternativa H_a .

♣ Pregunta 4.33

Solución con calculadora

1. Definir las hipótesis:

$$H_0 : \mu = 69$$

$$H_1 : \mu \neq 69$$

2. Calcular el estadístico t:

$$t = \frac{m - \mu_0}{s/\sqrt{n}} = \frac{70.3 - 69}{53.9/\sqrt{311}} = 0.4310$$

3. Calcular los n-1 grados de libertad:

$$g.l. = n - 1 = 311 - 1 = 310$$

4. Buscar $t = 0.4310$ en la tabla t de Student con 310 grados de libertad, obtenemos $p = 0.6668$.

Conclusión: puesto que el valor-p calculado es mayor que el nivel de significación $\alpha = 0.05$, no se puede rechazar la hipótesis nula H_0 con un nivel de significación al 5%.

Solución con Stata

```
. ttest UreaIng=69 if Caso==0
```

One-sample t test

Variable	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.Interval]	
UreaIng	311	70.31777	3.0578	53.92488	64.3011	76.33444
mean = mean(UreaIng)					t = 0.4310	
Ho: mean = 69					degrees of freedom = 310	
Ha: mean <69			Ha: mean != 69		Ha: mean >69	
Pr(T <t) = 0.6666			Pr(T > t) = 0.6668		Pr(T >t) = 0.3334	

♣ Pregunta 4.34

Solución con calculadora

- Definir las hipótesis:

$$H_0 : \mu = 106$$

$$H_1 : \mu > 106$$

- Calcular el estadístico t:

$$t = \frac{m - \mu_0}{s/\sqrt{n}} = \frac{106.5 - 106}{5.58/\sqrt{225}} = 1.218$$

- Calcular los n-1 grados de libertad:

$$g.l. = n - 1225 - 1 = 224$$

- Buscar $t = 1.218$ en la tabla t de Student con 224 grados de libertad, obtenemos $p = 0.1122$.

Conclusión: puesto que el valor-p calculado es mayor que el nivel de significación $\alpha = 0.05$, no se puede rechazar la hipótesis nula H_0 .

Solución con Stata

```
. ttest C124h=106 if Caso==0
```

One-sample t test

Variable	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.Interval]	
C124h	225	106.4533	.3720855	5.581282	105.7201	107.1866
mean = mean(CreatIng)					t = 1.2184	
Ho: mean = 106					degrees of freedom = 224	
Ha: mean <106			Ha: mean != 106		Ha: mean >106	
Pr(T <t) = 0.8878			Pr(T > t) = 0.2244		Pr(T >t) = 0.1122	

Comparación de dos medias

SOLUCIÓN 5.1

♣ Pregunta 5.1**Solución con calculadora**

$$m = \frac{\sum_{i=1}^n x_i}{n} = \frac{87562}{329} = 266.146$$

El colesterol medio de los fallecidos es de 266.146.

Solución con Excel

Se generan dos etiquetas: una etiqueta con los valores de colesterol (COLESTEROL) en los vivos (MUERTE=0) denominada *colesterol_muerto_0* y otra con los valores de colesterol (COLESTEROL) en los fallecidos (MUERTE=1) denominada *colesterol_muerto_1*. Representa la muestra de los vivos y fallecidos y facilita el uso de fórmulas en Excel.

Para resolver el problema utilizaremos las siguientes funciones:

- Para obtener la *n*: CONTAR .
- Para obtener la *m*: PROMEDIO .

	Valor	Función
n	329	CONTAR(colesterol_muerte_1)
m	266.146	PROMEDIO(colesterol_muerte_1)

Utilizando el analizador de datos: Estadística descriptiva .

Estadística descriptiva

Entrada
Rango de entrada: ETIQUETA

Agrupado por:
☒ Columnas
☐ Filas

☐ Góttulos en la primera fila

Opciones de salida
☒ Rango de salida: SJS31

☐ En una hoja nueva:
☐ En un libro nuevo

☒ Resumen de estadísticas
☒ Nivel de confianza para la media 95 %
☐ K-ésimo mayor: 1
☐ K-ésimo menor: 1

Aceptar Cancelar Ayuda

Colesterol fallecidos

Media	266.145897
Error típico	2.41445877
Mediana	262
Moda	272
Desviación estándar	43.7943154
Varianza de la muestra	1917.94206
Curtosis	0.29383848
Coefficiente de asimetría	0.35817588
Rango	280
Mínimo	159
Máximo	439
Suma	87562
Cuenta	329
Nivel de confianza(95.0 %)	4.7497784

Nota: La resolución mediante el analizador de datos solo se muestra en el primer ejemplo.

Solución con Stata

Para obtener la media utilizamos el comando `mean` :

```
. mean COLESTEROL if MUERTE==1
```

Mean estimation

Number of obs = 329

	Mean	Std. Err.	[95% Conf. Interval]
COLESTEROL	266.1459	2.414459	261.3961 270.8957

♣ Pregunta 5.2

Solución con calculadora

$$m(\text{fallecidos}) = \frac{\sum_{i=1}^n x_i}{n} = \frac{87562}{329} = 266.146$$

$$m(\text{vivos}) = \frac{\sum_{i=1}^n x_i}{n} = \frac{150312}{671} = 224.012$$

$$s^2 = \frac{\sum_{i=1}^n (x_i - m)^2}{n - 1}$$

$$s^2(\text{fallecidos}) = \frac{629084.997}{329 - 1} = 1917.942$$

$$s^2(\text{vivos}) = \frac{9783555.9046}{671 - 1} = 1460.233$$

En estos problemas vamos a considerar la regla de Box, es decir, si una varianza no es cuatro veces mayor que la otra, se puede usar directamente la t. Solo si es al menos cuatro veces mayor, usaremos el test de Welch.

$$\text{Homogeneidad de Varianzas} = \frac{\text{varianza}_{\text{mayor}}}{\text{varianza}_{\text{menor}}} = \frac{1917.942}{1460.233} = 1.313$$

$$\text{Homogeneidad de Varianzas} = 1.313 < 4 \rightarrow \text{Varianzas Homogneas}$$

Estas varianzas son casi iguales. Al dividir la varianza mayor entre la menor observamos que el resultado es menor de 4, y, por tanto, trabajaremos con la t de Student para varianzas homogéneas.

Varianzas homogéneas

$$\begin{aligned} s_p^2 &= \frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{(n_A - 1) + (n_B - 1)} \\ &= \frac{(329 - 1) 1917.942 + (671 - 1) * 1460.233}{(329 - 1) + (671 - 1)} = 1610.66 \end{aligned}$$

$$s_p = \sqrt{s_p^2} = \sqrt{1610.66} = 40.133$$

$$EEDM = s_p \sqrt{\frac{1}{n_A} + \frac{1}{n_B}} = 40.133 \sqrt{\frac{1}{329} + \frac{1}{671}} = 2.701$$

$$g.l = n_A + n_B - 2 = 161 + 839 - 2 = 998$$

$$t = \frac{dm}{EEDM} = \frac{m_A - m_B}{s_p \sqrt{\frac{1}{n_A} + \frac{1}{n_B}}} = \frac{224.0119 - 266.1459}{2.701} = -15.5987$$

El valor de la t-Student es de 15.599.

Solución con Excel

	Valor		Función
	Fallecidos (muerte=1)	Vivos (muerte=0)	
MEDIA	266.146	224.012	PROMEDIO(etiqueta)
VAR	1917.942*	1460.233	VAR.S(etiqueta)
N	329	671	CONTAR(etiqueta)
gl	328	670	N - 1
Homogeneidad de varianzas	1.313	1917.942/1460.233	

* Varianza mayor

Estas varianzas son casi iguales. Al dividir la varianza mayor entre la menor observamos que el resultado es menor de 4.

	Valor	Función
Varianza conjunta (S^2)	1610.662	(VAR muertos*gl muertos + VAR vivos*gl vivos) / (gl muertos+gl vivos)
Desviación ponderada (S)	40.133	RAIZ(varianza conjunta)
Diferencia de medias (DM)	-42.134	media mayor-media menor
Error Estándar de la DM (EEDM)	2.701	S* RAIZ(1/N muertos + 1/N vivos)
gl t de Student	998	N muertoss + N vivos - 2
t de Student	-15.5987	dm/eedm
Sig (una cola)	1.48E-49	T.DIST.RT(t-student;gl t-student)
Valor critico de t (una cola)	1.646	T.INV(0.95;gl t-student)
Sig (bilateral)	2.97E-49	T.DIST.2T(t-student;gl t-student)
Valor critico de t (dos colas)	1.962	T.INV.2T(0.05;gl t-student)

Utilizando el analizador de datos:

Prueba para varianzas de dos muestras

Prueba F para varianzas de dos muestras

Entrada

Rango para la variable 1: colesterol_muerte_1

Rango para la variable 2: colesterol_muerte_2

☐ Bótilos

Alfa: 0.05

Opciones de salida

☒ Rango de salida: SAS4

☐ En una hoja nueva:

☐ En un libro nuevo

Aceptar Cancelar Ayuda

	Fallecidos	Vivos
Media	266.1459	224.0119
Varianza	1917.942	1460.233
Observaciones	329	671
Grados de libertad	328	670
F	1.31345	
P(F<=f) una cola	0.001799	
Valor crítico para F (una cola)	1.167073	

Prueba T para dos muestras suponiendo varianzas iguales

Prueba t para dos muestras suponiendo varianzas iguales

Entrada

Rango para la variable 1: colesterol_muerte_1

Rango para la variable 2: colesterol_muerte_2

Diferencia hipotética entre las medias:

☐ Bótilos

Alfa: 0.05

Opciones de salida

☒ Rango de salida: SX\$24

☐ En una hoja nueva: COLESTEROL-MUERTE

☐ En un libro nuevo

Aceptar Cancelar Ayuda

	Fallecidos	Vivos
Media	266.1459	224.0119
Varianza	1917.942	1460.233
Observaciones	329	671
Varianza agrupada	1610.662	
Diferencia hipotética de las medias	0	
Grados de libertad	998	
Estadístico t	15.59874	
P(T<=t) una cola	1.48E-49	
Valor crítico de t (una cola)	1.646382	
P(T<=t) dos colas	2.97E-49	
Valor crítico de t (dos colas)	1.962344	

Solución con Stata

```
. ttest COLESTEROL, by(MUERTE) level(95)
```

Two-sample t test with equal variances

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
0	671	224.0119	1.4752	38.2129	221.1154	226.9085
1	329	266.1459	2.4145	43.7943	261.3961	270.8957
combined	1000	237.874	1.4147	44.7364	235.0979	240.6501
diff		-42.134	2.7012		-47.4345	-36.83346

diff = mean(0) - mean(1) t = -15.5987

Ho: diff = 0

degrees of freedom = 998

Ha: diff < 0

Ha: diff != 0

Ha: diff > 0

Pr(T < t) = 0.0000

Pr(|T| > |t|) = 0.0000

Pr(T > t) = 1.0000

♣ Pregunta 5.3

Solución con calculadora

Las hipótesis planteadas en el ejercicio son:

$$H_0 : \mu(\text{Colesterol})_{\text{muertos}} \leq \mu(\text{Colesterol})_{\text{vivos}}$$

$$H_1 : \mu(\text{Colesterol})_{\text{muertos}} > \mu(\text{Colesterol})_{\text{vivos}}$$

$$\text{t-Student} \rightarrow t = -15.5987 \text{ gl} = 998 \rightarrow 2.97 \times 10^{-49}$$

El valor de p teniendo en cuenta una cola es 5.37×10^{-43} , es decir, $p < 0.05$.

♣ Pregunta 5.4

Solución con calculadora

Utilizamos los datos calculados en el apartado anterior.

$$IC \ 95\% : dm \pm t_{0.025; n_A + n_B - 2} \times EEDM$$

$$IC \text{ Inferior} = -42.133 - 1.96 * 2.701 = -47.43$$

$$IC \text{ Superior} = -42.133 + 1.96 * 2.701 = -36.83$$

Utilizando los datos calculados en el apartado anterior obtenemos un IC al 95 % de -47.43 - -36.83.

Solución con Excel

	Valor	Función
IC Inferior	-47.43	DM-(T.INV.2T(0.05;gl t-student)*EEDM)
IC Superior	-36.83	DM+(T.INV.2T(0.05;gl t-student)*EEDM)

Solución con Stata

Con el comando `ttest` empleado en el apartado anterior obtenemos el IC al 95 %. Estos valores se muestran en la fila *diff* de la tabla de resultados correspondiente.

♣ **Pregunta 5.5****Solución con calculadora**

Como $p < 0.05$ rechazamos la hipótesis nula y podemos concluir que el nivel de colesterol aumenta el riesgo de muerte.

Con una confianza del 95% se puede afirmar que las personas que murieron presentaban de media niveles de colesterol entre 36.6 y 47.7 mg/dl más altos que la población que no se murió.

SOLUCIÓN 5.2

♣ Pregunta 5.6

Solución con calculadora

$$m = \frac{\sum_{i=1}^n x_i}{n}$$

$$m(\text{fallecidos}) = \frac{44054}{329} = 133.903$$

$$m(\text{vivos}) = \frac{86593}{671} = 129.051$$

La tensión arterial media en fallecidos es de 133.903 y en los vivos es de 129.051.

Solución con Excel

Se generan dos etiquetas: una etiqueta con los valores de la tensión arterial sistólica (TAS) en los vivos (MUERTE=0) denominada *TAS_muerto_0* y otra con los valores en los fallecidos (MUERTE=1) denominada *TAS_muerto_1*. Representa la muestra de los vivos y fallecidos y facilita el uso de fórmulas en Excel.

■ Fallecidos

	Valor	Función
n	329	CONTAR(TAS_muerto_1)
m	133.903	PROMEDIO(TAS_muerto_1)

■ Vivos

	Valor	Función
n	671	CONTAR(TAS_muerto_0)
m	129.051	PROMEDIO(TAS_muerto_0)

Solución con Stata

- Fallecidos . mean TAS if MUERTE==1

Mean estimation

Number of obs = 329

	Mean	Std. Err.	[95% Conf. Interval]	
TAS	133.9027	1.268884	131.4066	136.3989

- Vivos . mean TAS if MUERTE==0

Mean estimation

Number of obs = 671

	Mean	Std. Err.	[95% Conf. Interval]	
TAS	129.0507	.755899	127.5665	130.5349

♣ Pregunta 5.7**Solución con calculadora**

$$m = \frac{\sum_{i=1}^n x_i}{n}$$

$$m \text{ (fallecidos)} = \frac{44054}{329} = 133.903$$

$$m \text{ (vivos)} = \frac{86593}{671} = 129.051$$

$$s^2 = \frac{\sum_{i=1}^n (x_i - m)^2}{n - 1}$$

$$s^2(\text{fallecidos}) = \frac{173745.387}{329 - 1} = 529.712$$

$$s^2(\text{vivos}) = \frac{256876.778}{671 - 1} = 383.398$$

$$\text{Homogeneidad de Varianzas} = \frac{\text{varianza}_{\text{mayor}}}{\text{varianza}_{\text{menor}}} = \frac{529.712}{383.398} = 1.382$$

$$\text{Homogeneidad de Varianzas} = 1.382 < 4 \rightarrow \text{Varianzas Homogneas}$$

Estas varianzas son casi iguales. Al dividir la varianza mayor entre la menor observamos que el resultado es menor de 4, y, por tanto, trabajaremos con la t de

Student para varianzas homogéneas.

Varianzas homogéneas

$$s_p^2 = \frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{(n_A - 1) + (n_B - 1)}$$

$$= \frac{(329 - 1) 529.712 + (671 - 1) 383.398}{(329 - 1) + (671 - 1)} = 431.485$$

$$s_p = \sqrt{s_p^2} = \sqrt{431.485} = 20.772$$

$$EEDM = s_p \sqrt{\frac{1}{n_A} + \frac{1}{n_B}} = 20.772 \sqrt{\frac{1}{329} + \frac{1}{671}} = 1.398$$

$$g.l = n_A + n_B - 2 = 329 + 671 - 2 = 998$$

$$t = \frac{dm}{EEDM} = \frac{m_A - m_B}{s_p \sqrt{\frac{1}{n_A} + \frac{1}{n_B}}} = \frac{133.90 - 129.06}{1.398} = 3.471$$

El valor de la t-Student es de 3.471.

Solución con Excel

	Valor		Función
	Fallecidos (muerte=1)	Vivos (muerte=0)	
MEDIA	133.903	129.051	PROMEDIO(etiqueta)
VAR	529.712*	383.398	VAR.S(etiqueta)
N	329	671	CONTAR(etiqueta)
gl	328	670	N - 1
Homogeneidad de varianzas		1.382	529.712/383.398

* Varianza mayor

Estas varianzas son casi iguales. Al dividir la varianza mayor entre la menor observamos que el resultado es menor de 4.

Valor	Función
Varianza conjunta (S^2)	431.485 (VAR muertos*gl muertos + VAR vivos*gl vivos) / (gl muertos+gl vivos)
Desviación ponderada (S)	20.772 RAIZ(varianza conjunta)
Diferencia de medias (DM)	4.84 media mayor-media menor
Error Estándar de la DM (EEDM)	1.398 $S * \text{RAIZ}(1/N \text{ muertos} + 1/N \text{ vivos})$
gl t de Student	998 $N \text{ muertoss} + N \text{ vivos} - 2$
t de Student	3.471 dm/eedm
Sig (una cola)	0.003 T.DIST.RT(t-student;gl t-student)
Valor critico de t (una cola)	1.646 T.INV(0.95;gl t-student)
Sig (bilateral)	0.001 T.DIST.2T(t-student;gl t-student)
Valor critico de t (dos colas)	1.962 T.INV.2T(0.05;gl t-student)

Solución con Stata

```
. ttest TAS , by(MUERTE) level(95)
```

Two-sample t test with equal variances

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
0	671	129.0507	.755899	19.58056	127.5665	130.5349
1	329	133.9027	1.268884	23.01546	131.4066	136.3989
combined	1,000	130.647	.6604965	20.88673	129.3509	131.9431
diff		-4.852065	1.398053		-7.595526	-2.108604

diff = mean(0) - mean(1) t = -3.4706

Ho: diff = 0

degrees of freedom = 998

Ha: diff <0

Ha: diff != 0

Ha: diff >0

Pr(T <t) = 0.0003

Pr(|T| >|t|) = 0.0005

Pr(T >t) = 0.9997

♣ Pregunta 5.8

Solución con calculadora

Las hipótesis planteadas en el ejercicio son:

$$H_0 : \mu(TAS)_{muertos} \leq \mu(TAS)_{vivos}$$

$$H_1 : \mu(TAS)_{muertos} > \mu(TAS)_{vivos}$$

$$T - \text{Student} \rightarrow t = 3.471 \text{ gl} = 998 \rightarrow P = 0.003$$

El valor de p teniendo en cuenta una cola es 0.0003, es decir, $p < 0.05$.

♣ Pregunta 5.9

Solución con calculadora

$$IC\ 95\% : dm \pm t_{0.025; n_A + n_B - 2} \times EEDM$$

$$IC\ Inferior = 4.852 - 1.96 * 1.398 = 2.109$$

$$IC\ Superior = 4.852 + 1.96 * 1.398 = 7.597$$

Utilizando los datos calculados en el apartado anterior obtenemos un IC al 95% de 2.109 - 7.597.

Solución con Excel

	Valor	Función
IC Inferior	2.109	DM-(T.INV.2T(0.05;gl t-student)*EEDM)
IC Superior	7.597	DM+(T.INV.2T(0.05;gl t-student)*EEDM)

Solución con Stata

Con el comando `ttest` empleado en el apartado anterior obtenemos el IC al 95%. Estos valores se muestran en la fila *diff* de la tabla de resultados correspondiente.

♣ Pregunta 5.10**Solución con calculadora**

Como $p < 0.05$ rechazamos la hipótesis nula y podemos concluir que la tensión arterial sistólica aumenta el riesgo de muerte.

Con una confianza del 95% se puede afirmar que las personas que murieron presentaban de media tensión arterial sistólica entre 2.1 y 7.6 mm de Hg más altos que la población que no se murió.

SOLUCIÓN 5.3

♣ Pregunta 5.11

Solución con calculadora

Las hipótesis planteadas en el ejercicio son:

$$H_0 : \mu(TAS)_{fumador} \leq \mu(TAS)_{noFumador}$$

$$H_1 : \mu(TAS)_{fumador} > \mu(TAS)_{noFumador}$$

$$m = \frac{\sum_{i=1}^n x_i}{n}$$

$$m \text{ (fumadores)} = \frac{62506.5}{486} = 128.614$$

$$m \text{ (no fumadores)} = \frac{68140.5}{514} = 132.569$$

$$s^2 = \frac{\sum_{i=1}^n (x_i - m)^2}{n - 1}$$

$$s^2 \text{ (fumadores)} = \frac{219421.912}{486 - 1} = 452.416$$

$$s^2 \text{ (no fumadores)} = \frac{21490.298}{514 - 1} = 414.211$$

$$\text{Homogeneidad de Varianzas} = \frac{\text{varianza}_{mayor}}{\text{varianza}_{menor}} = \frac{452.416}{414.211} = 1.092$$

$$\text{Homogeneidad de Varianzas} = 1.092 < 4 \rightarrow \text{Varianzas Homogneas}$$

Estas varianzas son casi iguales. Al dividir la varianza mayor entre la menor observamos que el resultado es menor de 4, y, por tanto, trabajaremos con la t de Student para varianzas homogéneas.

Varianzas homogéneas

$$\begin{aligned} s_p^2 &= \frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{(n_A - 1) + (n_B - 1)} \\ &= \frac{(486 - 1) 452.416 + (514 - 1) * 414.211}{(486 - 1) + (514 - 1)} = 432.778 \end{aligned}$$

$$s_p = \sqrt{s_p^2} = \sqrt{432.778} = 20.803$$

$$EEDM = s_p \sqrt{\frac{1}{n_A} + \frac{1}{n_B}} = 20.803 \sqrt{\frac{1}{486} + \frac{1}{514}} = 1.316$$

$$g.l = n_A + n_B - 2 = 486 + 514 - 2 = 998$$

$$t = \frac{dm}{EEDM} = \frac{m_A - m_B}{s_p \sqrt{\frac{1}{n_A} + \frac{1}{n_B}}} = \frac{128.614 - 132.569}{1.316} = -3.0047$$

$$T - \text{Student} \rightarrow t = -3.0047 \quad g.l = 998 \rightarrow P = 0.9986$$

El valor de p teniendo en cuenta una cola es 0.9986, es decir, $p > 0.05$.

Como $p > 0.05$ **la hipótesis nula es verdadera y no hay evidencia estadística** para concluir que la tensión arterial sistólica es mayor en los fumadores.

Solución con Excel

	Valor		Función
	TAS (fumador=1)	TAS (fumador=0)	
MEDIA	128.614	132.569	PROMEDIO(etiqueta)
VAR	452.416*	414.211	VAR.S(etiqueta)
N	486	514	CONTAR(etiqueta)
gl	485	513	N - 1
Homogeneidad de varianzas	1.092	452.416/414.211	

* Varianza mayor

Estas varianzas son casi iguales. Al dividir la varianza mayor entre la menor observamos que el resultado es menor de 4.

Valor	Función
Varianza conjunta (S^2)	432.778
Desviación ponderada (S)	20.803
Diferencia de medias (DM)	3.955
Error Estándar de la DM ($EEDM$)	1.316
gl t de Student	998
t de Student	3.005
Sig (una cola)	0.0014
Valor critico de t (una cola)	1.646
Sig (bilateral)	0.003
Valor critico de t (dos colas)	1.962

Solución con Stata

```
. ttest TAS, by(FUMADOR) level(95)
```

Two-sample t test with equal variances

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
0	514	132.5691	.897696	20.35218	130.8055	134.3327
1	486	128.6142	.9648304	21.27008	126.7184	130.51
combined	1,000	130.647	.6604965	20.88673	129.3509	131.9431
diff	3.954869	1.316233	1.371967	6.53777		

diff = mean(0) - mean(1) t = 3.0047

Ho: diff = 0

degrees of freedom = 998

Ha: diff <0

Ha: diff != 0

Ha: diff >0

Pr(T < t) = 0.9986

Pr(|T| > |t|) = 0.0027

Pr(T > t) = 0.0014

♣ Pregunta 5.12

Solución con calculadora

$$IC\ 95\% : dm \pm t_{0.025; n_A + n_B - 2} \times EEDM$$

$$IC\ Inferior = 3.955 - 1.962 \times 1.316 = 1.372$$

$$IC\ Superior = 3.955 + 1.962 \times 1.316 = 6.538$$

Utilizando los datos calculados en el apartado anterior obtenemos un IC al 95 % de 1.372 - 6.538.

Solución con Excel

	Valor	Función
IC Inferior	1.372	DM-(T.INV.2T(0.05;g1 t-student)*EEDM)
IC Superior	6.538	DM+(T.INV.2T(0.05;g1 t-student)*EEDM)

Solución con Stata

Con el comando `ttest` empleado en el apartado anterior obtenemos el IC al 95 %. Estos valores se muestran en la fila *diff* de la tabla de resultados correspondiente.

♣ Pregunta 5.13

Solución con calculadora

Con una confianza del 95% se puede afirmar que la población de fumadores en media presenta niveles de la tensión arterial sistólica entre 1.37 y 6.54 mm de Hg más altos que la población de no fumadores.

SOLUCIÓN 5.4

♣ Pregunta 5.14

Solución con calculadora

$$m = \frac{\sum_{i=1}^n x_i}{n}$$

$$m(\text{fumadores}) = \frac{11014.28}{486} = 22.663$$

$$m(\text{no fumadores}) = \frac{14786.44}{514} = 28.767$$

El IMC medio en los fumadores es de 22.663 y en los no fumadores es de 28.767.

Solución con Excel

Se generan dos etiquetas: una etiqueta con los valores del IMC en los no fumadores (FUMADOR=0) denominada *IMC_fumador_0* y otra con los valores en los fumadores (FUMADOR=1) denominada *IMC_fumador_1*.

■ Fumador

	Valor	Función
n	586	CONTAR(IMC_fumador_1)
m	22.663	PROMEDIO(IMC_fumador_1)

■ No fumador

	Valor	Función
n	514	CONTAR(IMC_fumador_0)
m	28.767	PROMEDIO(IMC_fumador_0)

Solución con Stata

- Fumador . mean IMC if FUMADOR==1

Mean estimation

Number of obs = 486

	Mean	Std. Err.	[95% Conf. Interval]	
IMC	22.66313	.0880584	22.4901	22.83615

- No fumador . mean IMC if FUMADOR==0

Mean estimation

Number of obs = 514

	Mean	Std. Err.	[95% Conf. Interval]	
IMC	28.76739	.1472435	28.47812	29.05667

♣ Pregunta 5.15**Solución con calculadora**

$$m = \frac{\sum_{i=1}^n x_i}{n}$$

$$m \text{ (fumadores)} = \frac{11014.28}{486} = 22.663$$

$$m \text{ (no fumadores)} = \frac{14786.44}{514} = 28.767$$

$$s^2 = \frac{\sum_{i=1}^n (x_i - m)^2}{n - 1}$$

$$s^2 \text{ (fumadores)} = \frac{1827.767}{486 - 1} = 3.769$$

$$s^2 \text{ (no fumadores)} = \frac{5716.799}{514 - 1} = 11.144$$

$$\text{Homogeneidad de Varianzas} = \frac{\text{varianza}_{\text{mayor}}}{\text{varianza}_{\text{menor}}} = \frac{11.144}{3.769} = 2.957$$

$$\text{Homogeneidad de Varianzas} = 2.957 < 4 \rightarrow \text{Varianzas Homogneas}$$

Aunque estas varianzas son casi iguales vamos a **suponer que son heterogéneas** para ver un ejemplo con el test de Welch.

Varianzas heterogéneas

$$sd = \sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}} = \sqrt{\frac{3.769}{486} + \frac{11.14}{514}} = 0.172$$

$$gl^* = \frac{\left(\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}\right)^2}{\frac{\left(\frac{s_A^2}{n_A}\right)^2}{n_A-1} + \frac{\left(\frac{s_B^2}{n_B}\right)^2}{n_B-1}} = \frac{\left(\frac{3.769}{486} + \frac{11.14}{514}\right)^2}{\frac{\left(\frac{3.769}{486}\right)^2}{486-1} + \frac{\left(\frac{11.14}{514}\right)^2}{514-1}} = 832.887$$

$$t = \frac{m_A - m_B}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}} = \frac{22.663 - 28.767}{0.172} = -35.580$$

El valor de la t-Student es de -35.580.

Solución con Excel

	Valor	Función
	IMC (fumador=1)	IMC (fumador=0)
MEDIA	22.663	28.767
VAR	3.769	11.144*
N	486	514
gl	485	513
Homogeneidad de varianzas	2.957	11.144/3.769

* Varianza mayor

Solución con Stata

En este caso aplicamos la opción `unequal` al comando `ttest` para indicarle que resuelva el test con varianzas desiguales.

```
. ttest IMC, by(FUMADOR) level(95) unequal
```

Two-sample t test with unequal variances

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
0	514	28.76739	.1472435	3.338242	28.47812	29.05667
1	486	22.66313	.0880584	1.941284	22.4901	22.83615
combined	1,000	25.80072	.1298832	4.107267	25.54584	26.0556
diff	6.104265	.1715662	5.767513	6.441018		

diff = mean(0) - mean(1) t = 35.5797

Ho: diff = 0

Satterthwaite's degrees of freedom = 832.887

Ha: diff < 0

Ha: diff != 0

Ha: diff > 0

Pr(T < t) = 1.0000

Pr(|T| > |t|) = 0.0000

Pr(T > t) = 0.0000

♣ Pregunta 5.16

Solución con calculadora

Las hipótesis planteadas en el ejercicio son:

$$H_0 : \mu(IMC)_{noFumador} \leq \mu(IMC)_{fumador}$$

$$H_1 : \mu(IMC)_{noFumador} > \mu(IMC)_{fumador}$$

Partimos del supuesto en el cuál la hipótesis nula es cierta, y bajo esta premisa se calcula el valor de p. Una $p < 0.05$ indica que la hipótesis nula es falsa y una $p > 0.05$ que es verdadera.

Obtenemos el valor de p a partir de la t-Student calculada:

$$t - Student \rightarrow t = -35.580 \text{ gl} = 832.887 \rightarrow P = 1.2 \times 10^{-169}$$

Por lo tanto, y dado que la p calculada es de 1.2×10^{-169} , es decir, $p < 0.05$ podemos concluir que la hipótesis nula es falsa, es decir, el índice de masa corporal (IMC) es mayor en los no fumadores.

SOLUCIÓN 5.5

♣ Pregunta 5.17

Solución con calculadora

$$m = \frac{\sum_{i=1}^n x_i}{n}$$

$$m \text{ (angina)} = \frac{8364}{161} = 51.950$$

$$m \text{ (no angina)} = \frac{41974}{839} = 50.029$$

La media es mayor en el grupo de edad que desarrolla angina.

Solución con Excel

Se generan dos etiquetas: una etiqueta con los valores de la edad (EDAD) en la muestra sin angina de pecho (Angina=0) denominada *EDAD_angina_0* y otra con las edades de la muestra con angina (Angina=1) denominada *EDAD_angina_1*. Representa las dos muestras y facilita el uso de fórmulas en Excel.

■ Angina

	Valor	Función
n	161	CONTAR(EDAD_angina_1)
m	51.950	PROMEDIO(EDAD_angina_1)

■ No angina

	Valor	Función
n	839	CONTAR(EDAD_angina_0)
m	50.029	PROMEDIO(EDAD_angina_0)

Solución con Stata

- Angina . mean EDAD if ANGINA==1

Mean estimation

Number of obs = 161

	Mean	Std. Err.	[95% Conf. Interval]	
EDAD	51.95031	.706052	50.55593	53.34469

- No angina . mean EDAD if ANGINA==0

Mean estimation

Number of obs = 839

	Mean	Std. Err.	[95% Conf. Interval]	
EDAD	50.02861	.2975004	49.44467	50.61254

♣ Pregunta 5.18**Solución con calculadora**

$$m = \frac{\sum_{i=1}^n x_i}{n}$$

$$m(\text{angina}) = \frac{8364}{161} = 51.950$$

$$m(\text{no angina}) = \frac{41974}{839} = 50.029$$

$$s^2 = \frac{\sum_{i=1}^n (x_i - m)^2}{n - 1}$$

$$s^2(\text{fumadores}) = \frac{12841.602}{161 - 1} = 80.260$$

$$s^2(\text{no fumadores}) = \frac{62227.313}{839 - 1} = 74.257$$

$$\text{Homogeneidad de Varianzas} = \frac{\text{varianza}_{\text{mayor}}}{\text{varianza}_{\text{menor}}} = \frac{80.260}{74.257} = 1.081$$

$$\text{Homogeneidad de Varianzas} = 1.081 < 4 \rightarrow \text{Varianzas Homogneas}$$

Estas varianzas son casi iguales. Al dividir la varianza mayor entre la menor observamos que el resultado es menor de 4, y, por tanto, trabajaremos con la t de

Student para varianzas homogéneas.

Varianzas homogéneas

$$s_p^2 = \frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{(n_A - 1) + (n_B - 1)}$$

$$= \frac{(161 - 1) 80.260 + (839 - 1) * 74.257}{(161 - 1) + (839 - 1)} = 75.219$$

$$s_p = \sqrt{s_p^2} = \sqrt{75.219} = 8.673$$

$$EEDM = s_p \sqrt{\frac{1}{n_A} + \frac{1}{n_B}} = 8.673 \sqrt{\frac{1}{161} + \frac{1}{839}} = 0.746$$

$$g.l = n_A + n_B - 2 = 161 + 839 - 2 = 998$$

$$t = \frac{dm}{EEDM} = \frac{m_A - m_B}{s_p \sqrt{\frac{1}{n_A} + \frac{1}{n_B}}} = \frac{51.950 - 50.029}{0.746} = 2.575$$

El valor del test es de 2.575.

Solución con Excel

	Valor		Función
	EDAD (angina=1)	EDAD (angina=0)	
MEDIA	51.950	50.029	PROMEDIO(etiqueta)
VAR	80.260*	74.257	VAR.S(etiqueta)
N	161	839	CONTAR(etiqueta)
gl	160	838	N - 1
Homogeneidad de varianzas		1.081	80.260/74.257

* Varianza mayor

Estas varianzas son casi iguales. Al dividir la varianza mayor entre la menor observamos que el resultado es menor de 4.

Valor		Función
Varianza conjunta (S^2)	75.219	(VAR angina*gl angina + VAR noAngina*gl noAngina) / (gl angina+gl noAngina)
Desviación ponderada (S)	8.673	RAIZ(varianza conjunta)
Diferencia de medias (DM)	1.922	media mayor-media menor
Error Estándar de la DM (EEDM)	0.746	S* RAIZ(1/N angina + 1/N noAngina)
gl t de Student	998	N angina + N noAngina - 2
t de Student	2.5752	dm/eedm
Sig (una cola)	0.0051	T.DIST.RT(t-student;gl t-student)
Valor critico de t (una cola)	1.646	T.INV(0.95;gl t-student)
Sig (bilateral)	0.010	T.DIST.2T(t-student;gl t-student)
Valor critico de t (dos colas)	1.962	T.INV.2T(0.05;gl t-student)

Solución con Stata

```
. ttest EDAD, by(ANGINA) level(95)
```

Two-sample t test with equal variances

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
0	839	50.02861	.2975004	8.617247	49.44467	50.61254
1	161	51.95031	.706052	8.958795	50.55593	53.34469
combined	1,000	50.338	.2750335	8.697321	49.79829	50.87771
diff	-1.921705	.7462269	-3.386059	-.4573514		

diff = mean(0) - mean(1) t = -2.5752

Ho: diff = 0

degrees of freedom = 998

Ha: diff <0

Ha: diff != 0

Ha: diff >0

Pr(T <t) = 0.0051

Pr(|T| >|t|) = 0.0102

Pr(T >t) = 0.9949

♣ Pregunta 5.19**Solución con calculadora**

Las hipótesis planteadas en el ejercicio son:

$$H_0 : \mu(Edad)_{angina} = \mu(Edad)_{noAngina}$$

$$H_1 : \mu(Edad)_{angina} \neq \mu(Edad)_{noAngina}$$

$$t - \text{Student} \rightarrow t = 2.575 \text{ gl} = 998 \rightarrow P = 0.01016$$

El valor de p teniendo en cuenta las dos colas es 0.01016, es decir, $p < 0.05$.

♣ Pregunta 5.20**Solución con calculadora**

$p < 0.05$ Rechazamos la hipótesis nula y podemos concluir que la edad se relaciona al desarrollo de angina de pecho.

Con una confianza del 95 % se puede afirmar que las personas con angina de pecho presentaban una edad media entre 0.5 y 3 años más altos que la población que tenían angina de pecho.

SOLUCIÓN 5.6

♣ **Pregunta 5.21****Solución con calculadora**

$$m = \frac{\sum_{i=1}^n x_i}{n}$$

$$m \text{ (ACV)} = \frac{18250}{68} = 268.382$$

$$m \text{ (no ACV)} = \frac{235.648}{932} = 235.648$$

$$s^2 = \frac{\sum_{i=1}^n (x_i - m)^2}{n - 1}$$

$$s^2 \text{ (ACV)} = \frac{70000.059}{68 - 1} = 1044.777$$

$$s^2 \text{ (no ACV)} = \frac{1861438.567}{932 - 1} = 1999.397$$

$$\text{Homogeneidad de Varianzas} = \frac{\text{varianza}_{\text{mayor}}}{\text{varianza}_{\text{menor}}} = \frac{1999.397}{1044.777} = 1.914$$

$$\text{Homogeneidad de Varianzas} = 1.914 < 4 \rightarrow \text{Varianzas Homogéneas}$$

Estas varianzas son casi iguales. Al dividir la varianza mayor entre la menor observamos que el resultado es menor de 4, y, por tanto, trabajaremos con la t de Student para varianzas homogéneas.

Varianzas homogéneas

$$\begin{aligned} s_p^2 &= \frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{(n_A - 1) + (n_B - 1)} \\ &= \frac{(68 - 1) 1044.777 + (932 - 1) * 1999.397}{(68 - 1) + (932 - 1)} = 1935.309 \end{aligned}$$

$$s_p = \sqrt{s_p^2} = \sqrt{1935.305} = 43.992$$

$$EEDM = s_p \sqrt{\frac{1}{n_A} + \frac{1}{n_B}} = 43.992 \sqrt{\frac{1}{68} + \frac{1}{932}} = 5.526$$

$$g.l = n_A + n_B - 2 = 68 + 932 - 2 = 998$$

$$t = \frac{dm}{EEDM} = \frac{m_A - m_B}{s_p \sqrt{\frac{1}{n_A} + \frac{1}{n_B}}} = \frac{268.382 - 235.648}{5.526} = 5.937$$

El valor del test es de 5.937.

Solución con Excel

	Valor		Función
	ACV = 1	ACV = 0	
MEDIA	268.382	235.648	PROMEDIO(etiqueta)
VAR	1044.777	1999.397*	VAR.S(etiqueta)
N	68	932	CONTAR(etiqueta)
gl	67	931	N - 1
Homogeneidad de varianzas		1.914	199.397/1044.777

* Varianza mayor

Estas varianzas son casi iguales. Al dividir la varianza mayor entre la menor observamos que el resultado es menor de 4.

	Valor	Función
Varianza conjunta (S^2)	1935.309	(VAR ACV*gl ACV + VAR noACV*gl noACV) / (gl ACV+gl noACV)
Desviación ponderada (S)	43.992	RAIZ(varianza conjunta)
Diferencia de medias (DM)	32.734	media mayor-media menor
Error Estándar de la DM (EEDM)	5.526	S* RAIZ(1/N ACV + 1/N noACV)
gl t de Student	998	N ACV + N noACV - 2
t de Student	5.9237	dm/eedm
Sig (una cola)	0.0000	T.DIST.RT(t-student;gl t-student)
Valor critico de t (una cola)	1.646	T.INV(0.95;gl t-student)
Sig (bilateral)	0.000	T.DIST.2T(t-student;gl t-student)
Valor critico de t (dos colas)	1.962	T.INV.2T(0.05;gl t-student)

Solución con Stata

```
. ttest COLESTEROL, by(ACV) level(95)
```

Two-sample t test with equal variances

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
0	932	235.6481	1.464676	44.71462	232.7736	238.5225
1	68	268.3824	3.919741	32.32301	260.5585	276.2062
combined	1,000	237.874	1.414691	44.73644	235.0979	240.6501
diff	-32.73428	5.526024	-43.57824	-21.89032		

diff = mean(0) - mean(1) t = -5.9237

Ho: diff = 0

degrees of freedom = 998

Ha: diff < 0

Ha: diff != 0

Ha: diff > 0

Pr(T < t) = 0.0000

Pr(|T| > |t|) = 0.0000

Pr(T > t) = 1.0000

♣ Pregunta 5.22**Solución con calculadora**

Las hipótesis planteadas en el ejercicio son:

$$H_0 : \mu(\text{Colesterol})_{ACV} \leq \mu(\text{Colesterol})_{noACV}$$

$$H_1 : \mu(\text{Colesterol})_{ACV} > \mu(\text{Colesterol})_{noACV}$$

$$T - \text{Student} \rightarrow t = 5.937 \text{ gl} = 998 \rightarrow = 2.16 \times 10^{-09}$$

El valor de p teniendo en cuenta una cola es 5.16×10^{-09} , es decir, $p < 0.05$.

♣ Pregunta 5.23**Solución con calculadora**

$$IC\ 95\% : dm \pm t_{0.025; n_A + n_B - 2} \times EEDM$$

$$IC\ Inferior = 32.734 - 1.96 * 5.526 = 21.890$$

$$IC\ Superior = 32.734 + 1.96 * 5.526 = 43.578$$

Utilizando los datos calculados en el apartado anterior obtenemos un IC al 95 % de 21.890 - 43.578.



Solución con Excel

	Valor	Función
IC Inferior	21.890	$DM - (T.INV.2T(0.05; gl_{t-student}) * EEDM)$
IC Superior	43.578	$DM + (T.INV.2T(0.05; gl_{t-student}) * EEDM)$

Solución con Stata

Con el comando `ttest` empleado en el apartado anterior obtenemos el IC al 95 %. Estos valores se muestran en la fila *diff* de la tabla de resultados correspondiente.

♣ Pregunta 5.24

Solución con calculadora

Por lo tanto, y dado que la p calculada es menor a 0.05 podemos rechazar la hipótesis nula y concluir que el nivel de colesterol aumenta el riesgo de ACV.

SOLUCIÓN 5.7

♣ **Pregunta 5.25****Solución con calculadora**

Las hipótesis planteadas en el ejercicio son:

$$H_0 : \mu(\text{Colesterol})_{\text{despues}} \geq \mu(\text{Colesterol})_{\text{antes}}$$

$$H_1 : \mu(\text{Colesterol})_{\text{despues}} < \mu(\text{Colesterol})_{\text{antes}}$$

$$m = \frac{\sum_{i=1}^n x_i}{n}$$

$$m(\text{antes}) = \frac{237874}{1000} = 237.874$$

$$m(\text{despues}) = \frac{237330}{1000} = 237.330$$

$$dm = m_A - m_B = 237.330 - 237.874 = -0.544$$

$$s_d = \sqrt{s_d^2} = \sqrt{2990.430} = 53.856$$

$$EEDM = \frac{s_d}{\sqrt{n}} = \frac{53.856}{\sqrt{1000}} = 1.703$$

$$t = \frac{dm}{EEDM} = \frac{-0.544}{1.703} = -0.319$$

$$t - \text{Student} \rightarrow t = -0.319 \text{ gl} = 1000 \rightarrow P = 0.375$$

El valor de p teniendo en cuenta una cola es 0.375, es decir, $p > 0.05$. Por lo tanto, **la hipótesis nula es verdadera y no hay evidencia estadística** para afirmar que después de un mes de tratamiento se disminuye el nivel de colesterol de los sujetos de nuestro estudio.

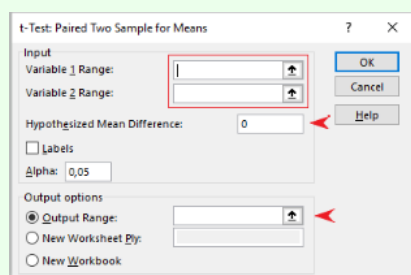
Solución con Excel

Se generan dos etiquetas: una etiqueta con los valores del colesterol antes (COLESTEROL) y otra con los valores de colesterol después (COLESTEROL2). Además, se crea una nueva variable que contiene la diferencia entre el nuevo colesterol y el antiguo ($COLESTEROL2 - COLESTEROL$) que denominada $d_{colesterol}$.

	Valor		Función	
	COLESTEROL	COLESTEROL2	Diff	
N	1000	1000	1000	CONTAR(etiqueta)
MEDIA	237.874	237.330	-0.544	PROMEDIO(etiqueta)
		Varianza - S2	2900.430	VAR.S(etiqueta)
		Desviación estándar - Sd	53.856	RAIZ(s2)
		Error estándar - EE(d)	1.703	sd/ RAIZ(n)

	Valor	Función
t de Student	-0.3194	(m-0)/ee(d)
Sig (una cola)	0.3747	T.DIST.RT(abs(t-student);gl t-student)
Valor crítico de t (una cola)	1.646	T.INV(0.95;gl t-student)

Utilizando el analizador de datos: prueba T para dos muestras emparejadas



	COLESTEROL	COLESTEROL2
Media	237.33	237.874
Varianza	2074.239339	2001.34947
Observaciones	1000	1000
Coefficiente de correlación de Pearson	0.288386871	
Diferencia hipotética de las medias	0	
Grados de libertad	999	
Estadístico t Stat	-0.319424092	
P(T<=t) una cola	0.374735863	
Valor crítico de t (una cola)	1.646380345	
P(T<=t) dos colas	0.749471727	
Valor crítico de t (dos cola)	1.962341461	

Solución con Stata

```
. ttest COLESTEROL2=COLESTEROL
```

Paired t test

Variable	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
COLESTEROL2	1,000	237.33	1.440222	45.54382	234.5038	240.1562
COLESTEROL	1,000	237.874	1.414691	44.73644	235.0979	240.6501
diff	1,000	-.544	1.703065	53.85564	-3.885995	2.797995

mean(diff) = mean(COLESTEROL2 - COLESTEROL) t = -0.3194

Ho: mean(diff) = 0 degrees of freedom = 999

Ha: mean(diff) < 0

Ha: mean(diff) != 0

Ha: mean(diff) > 0

Pr(T < t) = 0.3747

Pr(|T| > |t|) = 0.7495

Pr(T > t) = 0.6253

♣ Pregunta 5.26**Solución con calculadora**

$$IC\ 95\% : dm \pm 1.962 \times EE(d)$$

$$IC\ inferior : -0.544 - 1.962 \times 1.703 = -3.882$$

$$IC\ superior : -0.544 + 1.962 \times 1.703 = 2.7494$$

Utilizando los datos calculados en el apartado anterior obtenemos un IC al 95 % de -3.882 - 2.794.

Solución con Excel

	Valor	Función
IC Inferior	-3.882	DM-(T.INV.2T(0.05;g1 t-student)*EEDM)
IC Superior	2.794	DM+(T.INV.2T(0.05;g1 t-student)*EEDM)

♣ Pregunta 5.27

Solución con calculadora

El valor de p teniendo en cuenta una cola es 0.375, es decir, $p > 0.05$. Por lo tanto, **la hipótesis nula es verdadera y no hay evidencia estadística** para afirmar que después de un mes de tratamiento se disminuye el nivel de colesterol de los sujetos de nuestro estudio.

Con una confianza del 95 % se puede afirmar que las personas, después de 1 mes de tratamiento con estatinas, presentaban una reducción del nivel de colesterolde entre -3.882 y 2.794 .

SOLUCIÓN 5.8

♣ **Pregunta 5.28****Solución con calculadora**

Las hipótesis planteadas en el ejercicio son:

$$H_0 : \mu(\text{Glucemia})_{\text{despus}} \geq \mu(\text{Glucemia})_{\text{antes}}$$

$$H_1 : \mu(\text{Glucemia})_{\text{despus}} < \mu(\text{Glucemia})_{\text{antes}}$$

$$m = \frac{\sum_{i=1}^n x_i}{n}$$

$$m \text{ (antes)} = \frac{136352}{1000} = 136.352$$

$$m \text{ (despues)} = \frac{81252}{1000} = 81.252$$

$$dm = m_A - m_B = 81.252 - 136.352 = -55.100$$

$$s_d = \sqrt{s_d^2} = \sqrt{1154.084} = 33.972$$

$$\text{EEDM} = \frac{s_d}{\sqrt{n}} = \frac{33.972}{\sqrt{1000}} = 1.074$$

$$t = \frac{dm}{\text{EEDM}} = \frac{-55.100}{1.074} = -51.290$$

$$t - \text{Student} \rightarrow t = -51.290 \text{ gl} = 1000 \rightarrow P = 1.996 \times 10^{-282}$$

El valor de p teniendo en cuenta una cola es 1.996×10^{-282} , es decir, $p < 0.05$. Por lo tanto, **la hipótesis nula es falsa y hay evidencia estadística** para afirmar que después de un mes de actividad física se disminuye el nivel de glucemia en sangre de los sujetos de nuestro estudio.

Solución con Excel

Se generan dos etiquetas: una etiqueta con los valores de glucemia antes (GLUCEMIA) y otra con los valores de glucemia después (GLUCEMIA2). Además, se crea una nueva variable que contiene la diferencia entre la nueva glucemia y la antigua

(*GLUCEMIA2* – *GLUCEMIA*) que denominada *d_glucemia*.

	Valor			Función
	GLUCEMIA	GLUCEMIA2	D_GLUCEMIA	
N	1000	1000	1000	CONTAR(etiqueta)
MEDIA	136.352	81.252	-55.100	PROMEDIO(etiqueta)
		Varianza - S2	2900.430	VAR.S(etiqueta)
		Desviación estándar - Sd	53.856	RAIZ(s2)
		Error estándar - EE(d)	1.703	sd/ RAIZ(n)

	Valor	Función
t de Student	-51.2900	(m-0)/ee(d)
Sig (una cola)	0.0000	T.DIST.RT(abs(t-student);gl t-student)
Valor critico de t (una cola)	1.646	T.INV(0.95;gl t-student)

Solución con Stata

```
. ttest GLUCEMIA2=GLUCEMIA
```

Paired t test

Variable	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
GLUCEMIA2	1,000	81.252	.8717353	27.56669	79.54136	82.96264
GLUCEMIA	1,000	136.352	.7407475	23.42449	134.8984	137.8056
diff	1,000	-55.1	1.074283	33.97181	-57.20811	-52.99189

mean(diff) = mean(GLUCEMIA2 - GLUCEMIA) t = -51.2900

Ho: mean(diff) = 0

degrees of freedom = 999

Ha: mean(diff) <0

Ha: mean(diff) != 0

Ha: mean(diff) >0

Pr(T <t) = 0.0000

Pr(|T| >|t|) = 0.0000

Pr(T >t) = 1.0000

♣ Pregunta 5.29

Solución con calculadora

$$IC\ 95\% : dm \pm 1.962 \times EE(dm)$$

$$IC\ Inferior : 33.97 + 1.962 \times 1.074 = -57.206$$

$$IC\ Superior : 33.97 - 1.962 \times 1.074 = -52.994$$

Utilizando los datos calculados en el apartado anterior obtenemos un IC al 95 % de $-57.206 - -52.994$.

Solución con Excel

	Valor	Función
IC Inferior	-57.206	$DM - (T.INV.2T(0.05;gl \ t\text{-student}) * EEDM)$
IC Superior	-52.994	$DM + (T.INV.2T(0.05;gl \ t\text{-student}) * EEDM)$

♣ Pregunta 5.30

Solución con calculadora

El valor de p teniendo en cuenta una cola es 1.9967×10^{-282} , es decir, $p < 0.05$. Por lo tanto, **la hipótesis nula es falsa y hay evidencia estadística** para afirmar que después de un mes de actividad física el nivel de glucemia en sangre a disminuido en los sujetos de nuestro estudio.

Con una confianza del 95 % podemos afirmar que después de un mes de actividad física el nivel de glucemia de la población disminuye en promedio entre -57.21 y -52.99 mg/dl. Además, como el intervalo de confianza no incluye al 0, podemos estar seguros al 95 % de que los individuos que realizan actividad física durante un mes consiguen una reducción en los niveles de glucemia.

SOLUCIÓN 5.9

♣ Pregunta 5.31

Solución con calculadora

Las hipótesis planteadas en el ejercicio son:

$$H_0 : \mu(TAS)_{despues} \geq \mu(IMC)_{antes}$$

$$H_1 : \mu(TAS)_{despues} < \mu(IMC)_{antes}$$

$$m = \frac{\sum_{i=1}^n x_i}{n}$$

$$m \text{ (antes)} = \frac{130647}{1000} = 130.647$$

$$m \text{ (despues)} = \frac{130980.5}{1000} = 130.981$$

$$dm = m_A - m_B = 130.981 - 130.647 = 0.334$$

$$s_d = \sqrt{s_d^2} = \sqrt{694.37} = 26.341$$

$$= \frac{s_d}{\sqrt{n}} = \frac{26.341}{\sqrt{1000}} = 0.833$$

$$t = \frac{dm}{0.833} = \frac{0.334}{0.833} = 0.400$$

$$t - \text{Student} \rightarrow t = 0.400 \text{ gl} = 1000 \rightarrow P = 0.345$$

El valor de p teniendo en cuenta una cola es 0.345, es decir, $p > 0.05$. Por lo tanto, **la hipótesis nula es verdadera y no hay evidencia estadística** para afirmar que después de un mes de bajo consumo en sal se disminuye la tensión arterial sistólica de los sujetos de nuestro estudio.

Solución con Excel

Se generan dos etiquetas: una etiqueta con los valores de glucemia antes (GLUCEMIA) y otra con los valores de glucemia después (GLUCEMIA2). Además, se crea una nueva variable que contiene la diferencia entre la nueva glucemia y la antigua

$(GLUCEMIA2 - GLUCEMIA)$ que denominada $d_{glucemia}$.

	Valor			Función
	TAS	TAS2	D_TAS	
N	1000	1000	1000	CONTAR(etiqueta)
MEDIA	130.647	130.981	0.334	PROMEDIO(etiqueta)
	Varianza - S2		694.370	VAR.S(etiqueta)
	Desviación estándar - Sd		26.351	RAIZ(s2)
	Error estándar - EE(d)		0.833	sd/RAIZ(n)

	Valor	Función
t de Student	0.4002	$(m-0)/ee(d)$
Sig (una cola)	0.3445	T.DIST.RT(abs(t-student);gl t-student)
Valor crítico de t (una cola)	1.646	T.INV(0.95;gl t-student)

Solución con Stata

```
. ttest TAS2=TAS
```

Paired t test

Variable	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf. Interval]	
TAS2	1,000	130.9805	.6649131	21.0264	129.6757	132.2853
TAS	1,000	130.647	.6604965	20.88673	129.3509	131.9431
diff	1,000	.3335	.8332889	26.35091	-1.301697	1.968697

mean(diff) = mean(TAS2 - TAS) t = 0.4002

Ho: mean(diff) = 0

degrees of freedom = 999

Ha: mean(diff) < 0

Ha: mean(diff) != 0

Ha: mean(diff) > 0

Pr(T < t) = 0.6555

Pr(|T| > |t|) = 0.6891

Pr(T > t) = 0.3445

♣ Pregunta 5.32

Solución con calculadora

El valor de p teniendo en cuenta una cola es 0.3445, es decir, $p > 0.05$. Por lo tanto, **la hipótesis nula es verdadera y no hay evidencia estadística** para afirmar que después de un mes con un menor consumo de sal la tensión arterial sistólica es menor que la tensión arterial sistólica antes.

Comparación de tres o más medias

SOLUCIÓN 6.1

Nota: En este ejercicio se muestra el proceso completo de resolución. A partir del segundo, se omiten los pasos para realizar la comparación dos a dos que ya se ha trabajado en el capítulo anterior mostrando solo los resultados obtenidos.

♣ Pregunta 6.1

Solución con calculadora

$$m \text{ (sobrepeso)} = \frac{\sum_{i=1}^n x_i}{n} = \frac{100867}{417} = 241.887$$

La media de colesterol en la muestra con sobrepeso es de 241.887.

Solución con Excel

Se generan tres etiquetas: una etiqueta con los valores del colesterol (COLESTEROL) en la muestra con normopeso denominada *colesterol_imc_1*, otra con sobrepeso denominada *colesterol_imc_2* y, por último, otra con la muestra de obesidad denominada *imc_3*. Representa la muestra y facilita el uso de fórmulas en Excel.

COLESTEROL			
	Valor		
	Normopeso	Sobrepeso	Obesidad
n	456	417	127
m	234.509	241.887	236.780

CONTAR(etiqueta)

PROMEDIO(etiqueta)

Utilizando el analizador de datos: Estadística descriptiva

Estadística descriptiva ? X

Entrada
Rango de entrada: ETIQUETA

Agrupado por: ☒ Columnas ☐ Filas

☐ Rotulos en la primera fila

Opciones de salida
☒ Rango de salida: SJS31

☐ En una hoja nueva:

☐ En un libro nuevo

☒ Resumen de estadísticas

☒ Nivel de confianza para la media 95 %

☐ K-ésimo mayor: 1

☐ K-ésimo menor: 1

Aceptar Cancelar Ayuda

Colesterol sobrepeso

Media	241.8872902
Error típico	2.068723495
Mediana	239
Moda	225
Desviación estándar	42.24452919
Varianza de la muestra	1784.600247
Curtosis	-0.26892062
Coefficiente de asimetría	0.31798637
Rango	218
Mínimo	154
Máximo	372
Suma	100867
Cuenta	417

Nota: La resolución mediante el analizador de datos solo se muestra en el primer ejemplo.

Solución con Stata

```
. mean COLESTEROL if IMC_CAT==2
```

Mean estimation

Number of obs = 417

	Mean	Std. Err.	[95% Conf. Interval]	
COLESTEROL	241.8873	2.068723	237.8208	245.9537

♣ Pregunta 6.2

Solución con calculadora

■ Varianzas

$$s^2 = \frac{\sum_{i=1}^n (x_i - m)^2}{n - 1}$$

$$s^2(\text{normopeso}) = \frac{998012}{456 - 1} = 2193.433$$

$$s^2(\text{sobrepeso}) = \frac{742393.7}{417 - 1} = 1784.600$$

$$s^2(\text{obesidad}) = \frac{246909.8}{127 - 1} = 1959.602$$

$$s^2 = \frac{1999348.12}{1000 - 1} = 2001.349$$

■ Fuente de variación: **Total**

$$\begin{aligned} Total(SCT) &= (n - 1)s^2 \\ &= (1000 - 1)2001.349 = 1999348.124 \end{aligned}$$

■ Fuente de variación: **Entre grupos**

$$\begin{aligned} Entre\ grupos(SCeg) &= \sum_{i=1}^k n_i (m_i - m)^2 \\ &= 456 (234.509 - 237.874)^2 \\ &\quad + 417 (241.887 - 237.874)^2 \\ &\quad + 456 (236.780 - 237.874)^2 \\ &= 12032.630 \end{aligned}$$

$$Entre\ grupos(g.l) = k - 1 = 3 - 1 = 2$$

$$s_{entre\ grupos}^2 = \frac{SCeg}{g.l.\ entre\ grupos} = \frac{12032.630}{2} = 6016.315$$

■ Fuente de variación: **Residual**

$$\begin{aligned} Residual(SCR) &= SCT - SCeg \\ &= 1999348.124 - 12032.630 = 1987315.494 \end{aligned}$$

$$Residual(g.l) = n - k = 1000 - 3 = 997$$

$$s_{residual}^2 = \frac{SCR}{g.l.\ residuales} = \frac{1987315.494}{997} = 1993.295$$

■ **Valor F**

$$F = \frac{s_{entre\ grupos}^2}{s_{residual}^2} = \frac{6016.315}{1993.295} = 3.018$$

El valor de F es 3.018.

■ **Valor p**

$$F = 3.018 \quad g.l = 997; 2 \rightarrow P = 0.049$$

Como hay diferencias significativas entre los grupos $p(p=0.049)$ se realizan **comparaciones dos a dos** con la corrección de **Bonferroni** para determinar en qué grupos hay diferencias. Como se vio en el capítulo 5, estas comparaciones dos a dos se realizan con la t de Student. Para ello, todavía nos falta conocer las varianzas de cada grupo:

$$s^2(\text{normopeso}) = 2193.433$$

$$s^2(\text{sobrepeso}) = 1784.600$$

$$s^2(\text{obesidad}) = 1959.602$$

■ Primera comparación: **Normopeso - Sobrepeso**

$$\text{Homogeneidad de Varianzas} = \frac{\text{varianza}_{\text{mayor}}}{\text{varianza}_{\text{menor}}} = \frac{2193.433}{1784.600} = 1.229$$

$$\text{Homogeneidad de Varianzas} = 1.229 < 4 \rightarrow \text{Varianzas Homogéneas}$$

Varianzas homogéneas

$$\begin{aligned} s_p^2 &= \frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{(n_A - 1) + (n_B - 1)} \\ &= \frac{(456 - 1) 2193.433 + (417 - 1) 1784.600}{(456 - 1) + (417 - 1)} \\ &= 1998.170 \end{aligned}$$

$$s_p = \sqrt{s_p^2} = \sqrt{1998.170} = 44.701$$

$$EEDM = s_p \sqrt{\frac{1}{n_A} + \frac{1}{n_B}} = 44.701 \sqrt{\frac{1}{456} + \frac{1}{417}} = 3.029$$

$$g.l = n_A + n_B - 2 = 456 + 417 - 2 = 871$$

$$t = \frac{dm}{EEDM} = \frac{m_A - m_B}{s_p \sqrt{\frac{1}{n_A} + \frac{1}{n_B}}} = \frac{241.887 - 234.509}{3.029} = 2.4361$$

$$t - \text{Student} \rightarrow t = 2.4361 \text{ } g.l = 871 \rightarrow P = 0.015$$

Como se realizan 3 comparaciones dos a dos para obtener una **p penalizada** para un valor significativo de $p < 0.05$ debemos multiplicar el valor de la p de la t de Student por el número de comparaciones realizado, en este caso por 3:

$$P = 0.015 \times 3 \text{ comparaciones} = 0.045$$

■ Segunda comparación: **Normopeso - Obesidad**

$$\text{Homogeneidad de Varianzas} = \frac{\text{varianza}_{\text{mayor}}}{\text{varianza}_{\text{menor}}} = \frac{2193.433}{1959.602} = 1.119$$

$$\text{Homogeneidad de Varianzas} = 1.119 < 4 \rightarrow \text{Varianzas Homogéneas}$$

Varianzas homogéneas

$$\begin{aligned} s_p^2 &= \frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{(n_A - 1) + (n_B - 1)} \\ &= \frac{(456 - 1) 2193.433 + (417 - 1) 1959.602}{(456 - 1) + (417 - 1)} \\ &= 2142.723 \end{aligned}$$

$$s_p = \sqrt{s_p^2} = \sqrt{2142.723} = 46.290$$

$$EEDM = s_p \sqrt{\frac{1}{n_A} + \frac{1}{n_B}} = 46.290 \sqrt{\frac{1}{456} + \frac{1}{417}} = 2.271$$

$$g.l = n_A + n_B - 2 = 456 + 417 - 2 = 871$$

$$t = \frac{dm}{EEDM} = \frac{m_A - m_B}{s_p \sqrt{\frac{1}{n_A} + \frac{1}{n_B}}} = \frac{234.509 - 236.780}{2.271} = -0.989$$

$$t - \text{Student} \rightarrow t = -0.989 \text{ gl} = 871 \rightarrow P = 0.325$$

$$p \text{ penalizada} = p \times \text{número comparaciones} = 0.325 \times 3 = 0.975 \rightarrow 1$$

■ Tercera comparación: **Sobrepeso - Obesidad**

$$\text{Homogeneidad de Varianzas} = \frac{\text{varianza}_{\text{mayor}}}{\text{varianza}_{\text{menor}}} = \frac{1784.600}{1959.602} = 0.908$$

$$\text{Homogeneidad de Varianzas} = 0.908 < 4 \rightarrow \text{Varianzas Homogéneas}$$

Varianzas homogéneas

$$\begin{aligned} s_p^2 &= \frac{(n_A - 1)s_A^2 + (n_B - 1)s_B^2}{(n_A - 1) + (n_B - 1)} \\ &= \frac{(417 - 1) 1784.600 + (127 - 1) 1929.602}{(417 - 1) + (127 - 1)} \\ &= 1825.283 \end{aligned}$$

$$s_p = \sqrt{s_p^2} = \sqrt{1825.283} = 42.723$$

$$EEDM = s_p \sqrt{\frac{1}{n_A} + \frac{1}{n_B}} = 42.723 \sqrt{\frac{1}{417} + \frac{1}{127}} = 4.330$$

$$g.l = n_A + n_B - 2 = 417 + 127 - 2 = 542$$

$$t = \frac{dm}{EEDM} = \frac{m_A - m_B}{s_p \sqrt{\frac{1}{n_A} + \frac{1}{n_B}}} = \frac{241.887 - 236.780}{4.330} = 1.1796$$

$$t - Student \rightarrow t = 1.180 \quad g.l = 542 \rightarrow P = 0.239$$

$$p \text{ penalizada} = p \times \text{número comparaciones} = 0.239 \times 3 = 0.716$$

Solución con Excel

	COLESTEROL			
	Normopeso	Sobrepeso	Obesidad	Total
m	234.509	241.887	236.780	237.874
s2	2193.433	1784.600	1959.602	2001.349
n	456	417	127	1000
gl	455	416	126	999
k	1	1	1	3

Para rellenar la tabla con el excel utilizamos las siguientes funciones:

- Suma de cuadrados de SCT: `gl total * s2 total`
- Suma de cuadrados de SCR:
`gl normopeso * s2 normopeso + gl sobrepeso * s2 sobrepeso + gl obesidad * s2 obesidad`
- Suma de cuadrados de SC_{eg} : `SCT-SCR`
- Grados de libertad SCT: `n-1`
- Grados de libertad SCR: `n-k`
- Grados de libertad SC_{eg} : `k-1`

- Varianzas tanto de SCR como de SC_{eg} : $SC/g.l$
- Valor de F: $\text{Varianza SCE} / \text{Varianza SCR}$
- Valor de P: $F.DIST.RT(F; g.l. SC_{eg}; g.l. SCR)$

	SC	g.l	S^2	F	P
SC_{eg}	12032.630	2	6016.315	3.018	0.049
SCR	1987315.494	997	1993.295		
SCT	1999348.124	999			

Utilizando el analizador de datos: Anova: factor simple

A diferencia de los ejercicios planteados en los capítulos anteriores, la función Anova del analizador de datos no utiliza rangos de valores como entrada, sino que necesita una matriz de datos sobre los que realizar la operación. Por lo tanto, el primer paso será generar dicha matriz.

EDAD1	EDAD2	EDAD3
121	215	202
121	115	145
122,5	170	145
122,5	128	145
117,5	128	124
97,5	139,5	178,5
112	139,5	170
122,5	147,5	145
130	117	145
130	107	117
136	157	193
113	133	193
113	139	115
121	140	166
110,5	123	190
117	139	162,5
117	139	175
122,5	153,5	135
122,5	165	150
...		

Anova: Single Factor

Input Range: | ↑ Matriz de datos

Grouped By: ☒ Columns ☐ Rows

☒ Labels in first row

Alpha: 0,05

Output options: ☒ Output Range: ↑ ←

☐ New Worksheet Ply:

☐ New Workbook

OK Cancel Help

SUMMARY				
Grupo	Contar	Suma	Promedio	Varianza
IMC1	456	106936	234.5087719	2193.43289
IMC2	417	100867	241.8872902	1784.600247
IMC3	127	30071	236.7795276	1959.6018

ANOVA						
Fuente de variación	SS	df	MS	F	P-valor	F crit
Entre grupos	12032.629	2	6016.3148	3.0182756	0.049332	3.004752
Dentro grupos	1987315.494	997	1993.2953			
Total	1999348.124	999				

Como hay diferencias significativas entre los grupos ($p = 0.049 < 0.05$) se realiza **comparaciones dos a dos** con la corrección para determinar en qué grupos hay diferencias.

■ Primera comparación: **Normopeso - Sobrepeso**

Valor		Función	
	Normopeso	Sobrepeso	
MEDIA	234.509	241.887	PROMEDIO(etiqueta)
VAR	2193.433*	1784.600	VAR.S(etiqueta)
N	456	417	CONTAR(etiqueta)
gl	455	416	N - 1
Homogeneidad de varianzas		1.229	2193.433/ 1784.600

* Varianza mayor

Estas varianzas son casi iguales. Al dividir la varianza mayor entre la menor observamos que el resultado es menor de 4.

	Valor	Función
Varianza conjunta (S^2)	1998.170	$(VAR\ normopeso * gl\ normopeso + VAR\ sobrepeso * gl\ sobrepeso) / (gl\ normopeso + gl\ sobrepeso)$
Desviación ponderada (S)	44.701	$RAIZ(varianza\ conjunta)$
Diferencia de medias (DM)	7.379	$media\ mayor - media\ menor$
Error Estándar de la DM ($EEDM$)	3.029	$S * RAIZ(1/N\ normopeso + 1/N\ sobrepeso)$
$gl\ t$ de Student	871	$N\ normopeso + N\ sobrepeso - 2$
t de Student	2.4361	$dm / eedm$
Sig (una cola)	0.0075	$T.DIST.RT(t\text{-}student; gl\ t\text{-}student)$
Valor critico de t (una cola)	3.554	$T.INV(0.95; gl\ t\text{-}student)$
Sig (bilateral)	0.015	$T.DIST.2T(t\text{-}student; gl\ t\text{-}student)$
Valor critico de t (dos colas)	3.153	$T.INV.2T(0.05; gl\ t\text{-}student)$

▪ Segunda comparación: **Normopeso - Obesidad**

	Valor		Función
	Normopeso	Obesidad	
MEDIA	234.509	236.780	$PROMEDIO(etiqueta)$
VAR	2193.433*	1959.602	$VAR.S(etiqueta)$
N	456	127	$CONTAR(etiqueta)$
gl	455	126	$N - 1$
Homogeneidad de varianzas	1.119		$2193.433 / 1959.602$

* Varianza mayor

Estas varianzas son casi iguales. Al dividir la varianza mayor entre la menor observamos que el resultado es menor de 4.

Valor		Función
Varianza conjunta (S^2)	2142.723	(VAR normopeso*gl normopeso + VAR obesidad*gl obesidad) / (gl normopeso+gl obesidad)
Desviación ponderada (S)	46.290	RAIZ(varianza conjunta)
Diferencia de medias (DM)	2.271	media mayor-media menor
Error Estándar de la DM (EEDM)	4.644	S* RAIZ(1/N normopeso + 1/N obesidad)
gl t de Student	581	N normopeso + N obesidad - 2
t de Student	0.4889	dm/eedm
Sig (una cola)	0.3125	T.DIST.RT(t-student;gl t-student)
Valor critico de t (una cola)	3.562	T.INV(0.95;gl t-student)
Sig (bilateral)	0.625	T.DIST.2T(t-student;gl t-student)
Valor critico de t (dos colas)	2.401	T.INV.2T(0.05;gl t-student)

■ Tercera comparación: **Sobrepeso - Obesidad**

Valor		Función	
	Sobrepeso	Obesidad	
MEDIA	241.887	236.780	PROMEDIO(etiqueta)
VAR	1784.600	1959.602*	VAR.S(etiqueta)
N	417	127	CONTAR(etiqueta)
gl	416	126	N - 1
Homogeneidad de varianzas		1.098	1959.602/1784.600

* Varianza mayor

Estas varianzas son casi iguales. Al dividir la varianza mayor entre la menor observamos que el resultado es menor de 4.

	Valor	Función
Varianza conjunta (S^2)	1825.283	(VAR sobrepeso*gl sobrepeso + VAR obesidad*gl obesidad) / (gl sobrepeso+gl obesidad)
Desviación ponderada (S)	42.723	RAIZ(varianza conjunta)
Diferencia de medias (DM)	5.108	media mayor-media menor
Error Estándar de la DM ($EEDM$)	4.330	$S * \text{RAIZ}(1/N \text{ sobrepeso} + 1/N \text{ obesidad})$
gl t de Student	542	$N \text{ sobrepeso} + N \text{ obesidad} - 2$
t de Student	1.1796	dm/eedm
Sig (una cola)	0.1193	T.DIST.RT(t-student;gl t-student)
Valor critico de t (una cola)	3.562	T.INV(0.95;gl t-student)
Sig (bilateral)	0.239	T.DIST.2T(t-student;gl t-student)
Valor critico de t (dos colas)	2.401	T.INV.2T(0.05;gl t-student)

Para obtener la **p penalizada** para un valor significativo de $p < 0.05$ debemos multiplicar el valor de la p de la t de Student por el número de comparaciones realizado, en este caso.

	Grupos a comparar	p t de Student	p penalizada
Colesterol	Normopeso - Sobrepeso	0.015045798	0.045137394
	Normopeso - Obesidad	0.625083132	1.00
	Sobrepeso - Obesidad	0.238675502	0.716026507

Utilizando el analizador de datos:

Prueba T para dos muestras suponiendo varianzas iguales

Prueba t para dos muestras suponiendo varianzas iguales ? X

Entrada

Rango para la variable 1: colesterol_muerte_1

Rango para la variable 2: colesterol_muerte_0

Diferencia hipotética entre las medias:

☐ Bótilos

Alfa: 0.05

Opciones de salida

☒ Rango de salida: SX524

☐ En una hoja nueva: COLESTEROL-MUERTE

☐ En un libro nuevo

Aceptar Cancelar Ayuda

	Normopeso	Sobrepeso
Media	234.508	241.887
Varianza	2193.432	1784.600
Observaciones	456	417
Varianza agrupada	1998.169	
Diferencia hipotética de las medias	0	
Grados de libertad	871	
Estadístico t	-2.43610675	
P(T<=t) una cola	0.00752*	
Valor crítico de t (una cola)	1.6466	
P(T<=t) dos colas	0.015045*	
Valor crítico de t (dos colas)	1.9626	

	Normopeso	Obesidad	Sobrepeso	Obesidad
Media	234.5087719	236.7795276	241.8872902	236.7795276
Varianza	2193.43289	1959.6018	1784.600247	1959.6018
Observaciones	456	127	417	127
Varianza agrupada	2142.722533		1825.283265	
Diferencia hipotética de las medias	0		0	
Grados de libertad	581		542	
Estadístico t	-0.48891959		1.179603376	
P(T<=t) una cola	0.312541566*		0.119337751*	
Valor crítico de t (una cola)	1.647480506		1.64766985	
P(T<=t) dos colas	0.625083132*		0.238675502*	
Valor crítico de t (dos colas)	1.964055442		1.964350493	

Cuando hacemos la t de Student con el *anализador de datos*, la p obtenida no está corregida por el número de comparaciones. Como en este ejemplo tenemos 3 comparaciones dos a dos se considera significativa toda $p < 0.0167$, es decir, $p = 0.05$ dividida entre las 3 comparaciones. Lo mismo para las siguientes tablas.

Solución con Stata

En **stata** utilizamos el comando `oneway`. Además, indicando la opción de `bonferroni` después de la coma nos realiza el análisis completo con una sola llamada.

```
. oneway COLESTEROL IMC__CAT, tab bonferroni
```

IMC_CAT	Summary of COLESTEROL		
	Mean	Std. Dev.	Freq.
1	234.50877	46.8341	456
2	241.88729	42.244529	417
3	236.77953	44.26739	127
Total	237.874	44.736445	1,000

Analysis of Variance					
Source	SS	df	MS	F	Prob >F
Between groups	12032.6297	2	6016.31484	3.02	0.0493
Within groups	1987315.49	997	1993.29538		
Total	1999348.12	999	2001.34947		

Bartlett's test for equal variances: $\chi^2(2) = 4.6128$ Prob> $\chi^2 = 0.100$

Comparison of COLESTEROL by IMC_CAT (Bonferroni)			
Row Mean - Col Mean			
	1	2	
2	7.37852		
	0.045		
3	2.27076	-5.10776	
	1.000	0.778	

♣ Pregunta 6.3

Solución con calculadora

El valor de la p es de 0.049332.

♣ Pregunta 6.4

Solución con calculadora

Las hipótesis planteadas en el ejercicio son:

$$H_0 : \mu(\text{Colesterol})_{\text{normopeso}} = \mu(\text{Colesterol})_{\text{sobrepeso}} = \mu(\text{Colesterol})_{\text{obesidad}}$$

$$H_1 : \mu(\text{Colesterol})_{\text{normopeso}} \neq \mu(\text{Colesterol})_{\text{sobrepeso}} \neq \mu(\text{Colesterol})_{\text{obesidad}}$$

El valor de p es menor de 0.05 por lo tanto **tenemos evidencia estadística suficiente para rechazar la hipótesis nula** y concluir que el colesterol se relaciona con el índice de masa corporal.

Al hacer comparaciones dos a dos con la t de Student corregida por Bonferroni vemos que solo hay diferencias significativas entre la categoría 2 (sobrepeso) respecto de la categoría 1 (normopeso). Así, podemos observar que **la categoría 2 (media 241.9) es significativamente mayor el nivel de colesterol que la 1 (media 234.5)**. Sin embargo, la categoría 3 tiene resultados equivalentes a la 1 y a la 2 (no hay diferencias significativas entre estas categorías).

SOLUCIÓN 6.2

♣ Pregunta 6.5

Solución con calculadora

$$Media(m) = \frac{\sum_{i=1}^n x_i}{n}$$

$$m(edad1) = \frac{633651.5}{492} = 128.784$$

$$m(edad2) = \frac{64753.5}{454} = 142.629$$

$$m(edad3) = \frac{8237}{54} = 152.537$$

La media de glucemia para el grupo de edad 1 es de 128.784, para el grupo 2 de 142.629 y para el grupo 3 de 152.537.

Solución con Excel

Se generan tres etiquetas, una etiqueta con los valores de glucemia (GLUCEMIA) por cada grupo de edad (glucemia_edad_1, glucemia_edad_2 Y glucemia_edad_3). Representa la muestra y facilita el uso de fórmulas en Excel.

GLUCEMIA				
	Valor			Función
	EDAD1	EDAD2	EDAD3	
n	492	454	54	CONTAR(etiqueta)
m	128.784	142.629	152.537	PROMEDIO(etiqueta)

Solución con Stata

```
. mean GLUCEMIA if GRUP_EDAD==1
```

Mean estimation

Number of obs = 492

	Mean	Std. Err.	[95% Conf. Interval]	
GLUCEMIA	128.7835	.8392642	127.1345	130.4325



```
. mean GLUCEMIA if GRUP_EDAD==2
```

Mean estimation

Number of obs = 454

	Mean	Std. Err.	[95% Conf. Interval]	
GLUCEMIA	142.6289	1.180892	140.3081	144.9496

```
. mean GLUCEMIA if GRUP_EDAD==3
```

Mean estimation

Number of obs = 54

	Mean	Std. Err.	[95% Conf. Interval]	
GLUCEMIA	152.537	3.290682	145.9368	159.1373

♣ Pregunta 6.6

Solución con calculadora

■ Varianzas

$$s^2 = \frac{\sum_{i=1}^n (x_i - m)^2}{n - 1}$$

$$s^2(edad = 1) = \frac{170154.7}{492 - 1} = 346.547$$

$$s^2(edad = 2) = \frac{286796.7}{454 - 1} = 633.105$$

$$s^2(edad = 3) = \frac{30991.43}{54 - 1} = 584.744$$

$$s^2 = \frac{548158.096}{1000 - 1} = 548.707$$

■ Fuente de variación: **Total**

$$\begin{aligned} Total(SCT) &= (n - 1)s^2 \\ &= (1000 - 1)548.707 = 548158.096 \end{aligned}$$

- Fuente de variación: **Entre grupos**

$$\begin{aligned} \text{Entre grupos}(SCeg) &= \sum_{i=1}^k n_i (m_i - m)^2 \\ &= 492 (128.784 - 136.352)^2 \\ &\quad + 454 (142.629 - 136.352)^2 \\ &\quad + 54 (152.537 - 136.352)^2 \\ &= 60215.261 \end{aligned}$$

$$\text{Entre grupos}(g.l) = k - 1 = 3 - 1 = 2$$

$$s_{\text{entre grupos}}^2 = \frac{\text{SCE}}{\text{g.l. entre grupos}} = \frac{60215.261}{2} = 30107.631$$

- Fuente de variación: **Residual**

$$\begin{aligned} \text{Residual}(SCR) &= SCT - SCeg \\ &= 548158.096 - 60215.261 = 487942.835 \end{aligned}$$

$$\text{Residual}(g.l) = n - k = 1000 - 3 = 997$$

$$s_{\text{residual}}^2 = \frac{\text{SCR}}{\text{g.l. residuales}} = \frac{487942.835}{997} = 489.411$$

- **Valor F**

$$F = \frac{s_{\text{entre grupos}}^2}{s_{\text{residual}}^2} = \frac{30107.631}{489.411} = 61.518$$

El valor de F es 61.518.

- **Valor p**

$$F = 61.518 \quad g.l = 997; 2 \rightarrow P = 6.417 \times 10^{-26}$$

Como hay diferencias significativas entre los grupos $p(p=6.417 \times 10^{-26})$ se realiza **comparaciones dos a dos** con la corrección de **Bonferroni** para determinar en qué grupos hay diferencias.

En el problema 6.2 se muestra el desarrollo completo para obtener los resultados de las comparaciones dos a dos con la t de Student y la corrección de Bonferroni. A continuación, se muestran los resultados de las comparaciones correspondientes al presente ejercicio siguiendo el mismo desarrollo.

	Grupos a comparar	p t de Student	p penalizada
Glucemia	edad1 - edad2	3.7298×10^{-21}	1.11894×10^{-20}
	edad1 - edad3	7.48831×10^{-17}	2.24649×10^{-16}
	edad2 - edad3	0.006236475	0.018709425

Solución con Excel

GLUCEMIA				
	Edad1	Edad2	Edad3	Total
MEDIA	128.784	142.629	152.537	136.352
s2	346.547	633.105	584.744	548.707
n	492	454	54	1000
gl	491	453	53	999
k	1	1	1	3

	SC	g.l	S ²	F	P
SC_{eg}	60215.261	2	30107.631	61.518	0.000
SCR	487942.835	997	489.411		
SCT	548158.096	999			

Solución con Stata

```
. oneway GLUCEMIA GRUP_EDAD, tab bonferroni
```

GRUP_EDAD	Summary of GLUCEMIA		
	Mean	Std. Dev.	Freq.
1	128.78354	18.615779	492
2	142.62885	25.161584	454
3	152.53704	24.181478	54
Total	136.352	23.424492	1,000

Analysis of Variance					
Source	SS	df	MS	F	Prob >F
Between groups	60215.2614	2	30107.6307	61.52	0.0000
Within groups	487942.835	997	489.411068		
Total	548158.096	999	548.706803		

Bartlett's test for equal variances: $\chi^2(2) = 43.2756$ Prob> $\chi^2 = 0.000$

Comparison of GLUCEMIA by GRUP_EDAD (Bonferroni)				
Row Mean - Col Mean			1	2
2			13.8453	
			0.000	
3			23.7535	9.90818
			0.000	0.006

♣ Pregunta 6.7

Solución con calculadora

El valor de la p es de 6.417×10^{-26}

♣ Pregunta 6.8

Solución con calculadora

Las hipótesis planteadas en el ejercicio son:

$$H_0 : \mu(\text{Glucemia})_{edad1} = \mu(\text{Glucemia})_{edad2} = \mu(\text{Glucemia})_{edad3}$$

$$H_1 : \mu(\text{Glucemia})_{edad1} \neq \mu(\text{Glucemia})_{edad2} \neq \mu(\text{Glucemia})_{edad3}$$

El valor de p es menor de 0.05 por lo tanto tenemos evidencia estadística suficiente para rechazar la hipótesis nula y concluir que la glucemia varía según el grupo de edad.

Al hacer comparaciones dos a dos con la t de Student corregida por Bonferroni vemos que hay diferencias significativas entre las comparaciones de las tres cate-

gorías. Así, podemos observar que en el grupo de edad 1 (media 128.7) el nivel de glucemia es significativamente menor que grupo de edad 2 (media 142.63) y este, a su vez, presenta esta significación en el grupo de edad 3 (media 152.53).

SOLUCIÓN 6.3

♣ Pregunta 6.9

Solución con calculadora

$$m(\text{obesidad}) = \frac{\sum_{i=1}^n x_i}{n} = \frac{16869.5}{127} = 132.831$$

La tensión arterial sistólica media en la muestra con obesidad es de 132.831.

Solución con Excel

Se generan tres etiquetas, una etiqueta con los valores de tensión arterial sistólica (TAS) por cada grupo de edad (*tas_edad_1*, *tas_edad_2* Y *tas_edad_3*). Representa la muestra y facilita el uso de fórmulas en Excel.

		TAS		
		Valor		
		Normopeso	Sobrepeso	Obesidad
n	456	417	127	CONTAR(etiqueta)
m	128.498	132.332	132.831	PROMEDIO(etiqueta)

Solución con Stata

```
. mean TAS if IMC_CAT==3
```

Mean estimation

Number of obs = 127

	Mean	Std. Err.	[95% Conf. Interval]	
TAS	132.8307	1.868449	129.1331	136.5283

♣ Pregunta 6.10



Solución con calculadora

■ Varianzas

$$s^2 = \frac{\sum_{i=1}^n (x_i - m)^2}{n - 1}$$

$$s^2(\text{normopeso}) = \frac{212875}{456 - 1} = 467.857$$

$$s^2(\text{sobrepeso}) = \frac{163183.7}{417 - 1} = 392.269$$

$$s^2(\text{obesidad}) = \frac{55864.61}{127 - 1} = 443.370$$

$$s^2 = \frac{435819.391}{1000 - 1} = 436.256$$

■ Fuente de variación: **Total**

$$\begin{aligned} \text{Total}(SCT) &= (n - 1)s^2 \\ &= (1000 - 1)436.256 = 435819.391 \end{aligned}$$

■ Fuente de variación: **Entre grupos**

$$\begin{aligned} \text{Entre grupos}(SCeg) &= \sum_{i=1}^k n_i (m_i - m)^2 \\ &= 456 (128.498 - 130.647)^2 \\ &\quad + 417 (132.332 - 130.647)^2 \\ &\quad + 127 (132.831 - 130.647)^2 \\ &= 3896.034 \end{aligned}$$

$$\text{Entre grupos}(g.l) = k - 1 = 3 - 1 = 2$$

$$s^2_{\text{entre grupos}} = \frac{SCeg}{\text{g.l. entre grupos}} = \frac{3896.034}{2} = 1948.017$$

■ Fuente de variación: **Residual**

$$\begin{aligned} \text{Residual}(SCR) &= SCT - SCeg \\ &= 435819.391 - 3896.034 = 431923.357 \end{aligned}$$

$$\text{Residual}(g.l) = n - k = 1000 - 3 = 997$$

$$s^2_{\text{residual}} = \frac{SCR}{\text{g.l. residuales}} = \frac{431923.357}{997} = 433.223$$

■ **Valor F**

$$F = \frac{s_{\text{entre grupos}}^2}{s_{\text{residual}}^2} = \frac{1948.017}{433.223} = 4.497$$

El valor de F es 4.497.

■ **Valor p**

$$F = 4.497 \quad gl = 997; 2 \rightarrow P = 0.011$$

Como hay diferencias significativas entre los grupos $p(p=0.011)$ se realiza **comparaciones dos a dos** con la corrección de **Bonferroni** para determinar en qué grupos hay diferencias.

En el problema 6.2 se muestra el desarrollo completo para obtener los resultados de las comparaciones dos a dos con la t de Student y la corrección de Bonferroni. A continuación, se muestran los resultados de las comparaciones correspondientes al presente ejercicio siguiendo el mismo desarrollo.

	Grupos a comparar	p t de Student	p penalizada
TAS	Normopeso - Sobrepeso	0.006590371	0.019771113
	Normopeso - Obesidad	0.045111937	0.135335811
	Sobrepeso - Obesidad	0.806783058	2.420349175

Solución con Excel

	TAS			
	Normopeso	Sobrepeso	Obesidad	Total
MEDIA	128.498	132.332	132.831	130.647
s2	467.857	392.269	443.370	436.256
n	456	417	127	1000
gl	455	416	126	999
k	1	1	1	3

	SC	g.l	S ²	F	P
SC_{eg}	3896.034	2	1948.017	4.497	0.011
SCR	431923.357	997	433.223		
SCT	435819.391	999			

Solución con Stata

```
. oneway TAS IMC_CAT, tab bonferroni
```

IMC_CAT	Summary of TAS		
	Mean	Std. Dev.	Freq.
1	128.49781	21.630006	456
2	132.33213	19.805773	417
3	132.83071	21.056351	127
Total	130.647	20.886734	1,000

Analysis of Variance					
Source	SS	df	MS	F	Prob >F
Between groups	3896.03356	2	1948.01678	4.50	0.0114
Within groups	431923.357	997	433.223027		
Total	435819.391	999	436.255647		

Bartlett's test for equal variances: $\chi^2(2) = 3.3934$ Prob> $\chi^2 = 0.183$

Comparison of TAS by IMC_CAT (Bonferroni)		
Row Mean - Col Mean	1	2
2	3.83433	0.020
3	4.3329	.498574
	0.115	1.000

♣ Pregunta 6.11

Solución con calculadora

El valor de la p es de 0.011.

♣ Pregunta 6.12

Solución con calculadora

Las hipótesis planteadas en el ejercicio son:

$$H_0 : \mu(TAS)_{normopeso} = \mu(TAS)_{sobrepeso} = \mu(TAS)_{obesidad}$$

$$H_1 : \mu(TAS)_{normopeso} \neq \mu(TAS)_{sobrepeso} \neq \mu(TAS)_{obesidad}$$

El valor de p es menor de 0.05 por lo tanto **tenemos evidencia estadística suficiente para rechazar la hipótesis nula**, y por tanto, la tensión arterial sistólica se relaciona con el índice de masa corporal.

Al hacer comparaciones dos a dos con la t de Student corregida por Bonferroni vemos que solo hay diferencias significativas entre la categoría 2 (sobrepeso) respecto de la categoría 1 (normopeso). Así, podemos observar que en **la categoría 2 (media 132.33) la tensión arterial sistólica es significativamente mayor que en la 1 (media 128.5)**. Sin embargo, la categoría 3 tiene resultados equivalentes a la 1 y a la 2 (no hay diferencias significativas entre estas categorías).

SOLUCIÓN 6.4

♣ Pregunta 6.13

Solución con calculadora

$$Media(m) = \frac{\sum_{i=1}^n x_i}{n}$$

$$m \text{ (normopeso)} = \frac{22624}{456} = 49.614$$

$$m \text{ (sobrepeso)} = \frac{21331}{417} = 51.153$$

$$m \text{ (obesidad)} = \frac{6383}{127} = 50.260$$

La edad media mayor se encuentra en el grupo de sobrepeso (IMC_CAT = 2) con un valor de 51.153.

Solución con Excel

Se generan tres etiquetas, una etiqueta con los valores de edad (EDAD) por cada grupo de IMC (*edad_imc_1*, *edad_imc_2* y *edad_imc_3*). Representa la muestra y facilita el uso de fórmulas en Excel.

				EDAD
	Valor			Función
	Normopeso	Sobrepeso	Obesidad	
n	456	417	127	CONTAR(etiqueta)
m	49.614	51.153	50.260	PROMEDIO(etiqueta)

Solución con Stata

```
. mean EDAD if IMC_CAT==1
```

Mean estimation Number of obs = 127

	Mean	Std. Err.	[95% Conf. Interval]	
EDAD	49.61404	.4162225	48.79608	50.43199

```
. mean EDAD if IMC_CAT==2
```

Mean estimation

Number of obs = 417

	Mean	Std. Err.	[95% Conf. Interval]	
EDAD	51.15348	.4116945	50.34422	51.96274

```
. mean EDAD if IMC_CAT==3
```

Mean estimation

Number of obs = 127

	Mean	Std. Err.	[95% Conf. Interval]	
EDAD	50.25984	.7787495	48.71872	51.80096

♣ Pregunta 6.14

Solución con calculadora

■ Varianzas

$$s^2 = \frac{\sum_{i=1}^n (x_i - m)^2}{n - 1}$$

$$s^2(\text{normopeso}) = \frac{35944.07}{456 - 1} = 78.998$$

$$s^2(\text{sobrepeso}) = \frac{29402.18}{417 - 1} = 70.678$$

$$s^2(\text{obesidad}) = \frac{9704.425}{127 - 1} = 77.019$$

$$s^2 = \frac{75567.756}{1000 - 1} = 75.643$$

■ Fuente de variación: **Total**

$$\begin{aligned} \text{Total}(SCT) &= (n - 1)s^2 \\ &= (1000 - 1)75.643 = 75567.756 \end{aligned}$$

- Fuente de variación: **Entre grupos**

$$\begin{aligned} \text{Entre grupos}(SCeg) &= \sum_{i=1}^k n_i (m_i - m)^2 \\ &= 456 (49.614 - 50.338)^2 \\ &\quad + 417 (51.153 - 50.338)^2 \\ &\quad + 127 (50.260 - 50.338)^2 \\ &= 517.083 \end{aligned}$$

$$\text{Entre grupos}(g.l) = k - 1 = 3 - 1 = 2$$

$$s_{\text{entre grupos}}^2 = \frac{\text{SCE}}{\text{g.l. entre grupos}} = \frac{517.083}{2} = 258.542$$

- Fuente de variación: **Residual**

$$\begin{aligned} \text{Residual}(SCR) &= SCT - SCeg \\ &= 75567.756 - 517.083 = 75050.673 \end{aligned}$$

$$\text{Residual}(g.l) = n - k = 1000 - 3 = 997$$

$$s_{\text{residual}}^2 = \frac{SCR}{\text{g.l. residuales}} = \frac{75050.673}{997} = 75.277$$

- **Valor F**

$$F = \frac{s_{\text{entre grupos}}^2}{s_{\text{residual}}^2} = \frac{258.542}{75.277} = 3.435$$

El valor de F es 3.435.

- **Valor p**

$$F = 3.435 \quad gl = 997; 2 \rightarrow P = 0.033$$

Como hay diferencias significativas entre los grupos $p(p=0.033)$ se realiza **comparaciones dos a dos** con la corrección de **Bonferroni** para determinar en qué grupos hay diferencias.

En el problema 6.2 se muestra el desarrollo completo para obtener los resultados de las comparaciones dos a dos con la t de Student y la corrección de Bonferroni. A continuación, se muestran los resultados de las comparaciones correspondientes al presente ejercicio siguiendo el mismo desarrollo.

	Grupos a comparar	p t de Student	p penalizada
Edad	Normopeso - Sobrepeso	0.008867143	0.026601429
	Normopeso - Obesidad	0.468037669	1.404113006
	Sobrepeso - Obesidad	0.299724335	0.899173005

Solución con Excel

EDAD				
	Normopeso	Sobrepeso	Obesidad	Total
MEDIA	49.614	51.153	50.260	130.647
s2	78.998	70.678	77.019	436.256
n	456	417	127	1000
gl	455	416	126	999
k	1	1	1	3

	SC	g.l	S ²	F	P
SC_{eg}	517.083	2	258.542	3.435	0.033
SCR	75050.673	997	75.277		
SCT	75567.756	999			

Solución con Stata

```
. oneway EDAD IMC_CAT, tab bonferroni
```

IMC_CAT	Summary of EDAD		
	Mean	Std. Dev.	Freq.
1	49.614035	8.8880795	456
2	51.153477	8.4070394	417
3	50.259843	8.7760611	127
Total	50.338	8.6973214	1,000

Analysis of Variance					
Source	SS	df	MS	F	Prob >F
Between groups	517.08317	2	258.541585	3.43	0.0326
Within groups	75050.6728	997	75.2765023		
Total	75567.756	999	75.6433994		

Bartlett's test for equal variances: $\chi^2(2) = 1.3782$ Prob> $\chi^2 = 0.502$

Comparison of EDAD by IMC_CAT (Bonferroni)			
Row Mean - Col Mean		1	2
2		1.53944	
		0.027	
3		.645807	-.893635
		1.000	0.929

♣ Pregunta 6.15

Solución con calculadora

El valor de la p es de 0.033.

♣ Pregunta 6.16

Solución con calculadora

Las hipótesis planteadas en el ejercicio son:

$$H_0 : \mu(Edad)_{normopeso} = \mu(Edad)_{sobrepeso} = \mu(Edad)_{obesidad}$$

$$H_1 : \mu(Edad)_{normopeso} \neq \mu(Edad)_{sobrepeso} \neq \mu(Edad)_{obesidad}$$

El valor de p es menor de 0.05 por lo tanto **tenemos evidencia estadística suficiente para rechazar la hipótesis nula**. Y por tanto, concluir que la edad varía según el grupo de índice de masas corporal del paciente.

Al hacer comparaciones dos a dos con la t de Student corregida por Bonferroni vemos que solo hay diferencias significativas entre la categoría 2 (sobrepeso) respecto de la categoría 1 (normopeso). Así, podemos observar que **la categoría 2 (media 51.15) la edad media es significativamente mayor que la categoría 1 (media 49.61)**. Sin embargo, la categoría 3 tiene resultados equivalentes a la categoría 1 y a la categoría 2 (no hay diferencias significativas entre estas categorías).

SOLUCIÓN 6.5

♣ **Pregunta 6.17****Solución con calculadora**■ **Varianzas**

$$s^2 = \frac{\sum_{i=1}^n (x_i - m)^2}{n - 1}$$

$$s^2(\text{normopeso}) = \frac{270249.7}{456 - 1} = 593.955$$

$$s^2(\text{sobrepeso}) = \frac{195222.6}{417 - 1} = 469.285$$

$$s^2(\text{obesidad}) = \frac{64332.54}{127 - 1} = 510.576$$

$$s^2 = \frac{548158.096}{1000 - 1} = 548.707$$

■ **Fuente de variación: *Total***

$$\begin{aligned} \text{Total}(SCT) &= (n - 1)s^2 \\ &= (1000 - 1)548.707 = 548158.096 \end{aligned}$$

■ **Fuente de variación: *Entre grupos***

$$\begin{aligned} \text{Entre grupos}(SCeg) &= \sum_{i=1}^k n_i (m_i - m)^2 \\ &= 456 (132.508 - 136.352)^2 \\ &\quad + 417 (137.747 - 136.352)^2 \\ &\quad + 127 (145.575 - 136.352)^2 \\ &= 18353.275 \end{aligned}$$

$$\text{Entre grupos}(g.l) = k - 1 = 3 - 1 = 2$$

$$s_{\text{entre grupos}}^2 = \frac{SCeg}{\text{g.l. entre grupos}} = \frac{18353.275}{2} = 9176.637$$

■ **Fuente de variación: *Residual***

$$\begin{aligned} \text{Residual}(SCR) &= SCT - SCeg \\ &= 548158.096 - 18353.275 = 529804.821 \end{aligned}$$

$$Residual(g.l) = n - k = 1000 - 3 = 997$$

$$s_{residual}^2 = \frac{SCR}{g.l. \text{ residuales}} = \frac{529804.821}{997} = 531.399$$

■ **Valor F**

$$F = \frac{s_{entre \text{ grupos}}^2}{s_{residual}^2} = \frac{9176.637}{531.399} = 17.269$$

El valor de la F es 17.269.

■ **Valor p**

$$F = 17.269 \quad gl = 997; 2 \rightarrow P = 4.239 \times 10^{-8}$$

Como hay diferencias significativas entre los grupos $p(p=0.011)$ se realiza **comparaciones dos a dos** con la corrección de **Bonferroni** para determinar en qué grupos hay diferencias.

En el problema 6.2 se muestra el desarrollo completo para obtener los resultados de las comparaciones dos a dos con la t de Student y la corrección de Bonferroni. A continuación, se muestran los resultados de las comparaciones correspondientes al presente ejercicio siguiendo el mismo desarrollo.

	Grupos a comparar	p t de Student	p penalizada
Glucemia	Normopeso - Sobrepeso	0.000858346	0.002575039
	Normopeso - Obesidad	8.42244E-08	2.52673x10 ⁻⁰⁷
	Sobrepeso - Obesidad	0.000451985	0.001355955

Solución con Excel

	GLUCEMIA			
	Normopeso	Sobrepeso	Obesidad	Total
MEDIA	132.508	137.747	145.575	136.352
s2	593.955	469.285	510.576	548.707
n	456	417	127	1000
gl	455	416	126	999
k	1	1	1	3

	SC	g.l	S ²	F	P
SC _{eg}	18353.275	2	9176.637	17.269	0.000
SCR	529804.821	997	531.399		
SCT	548158.096	999			

Solución con Stata

```
. oneway GLUCEMIA IMC__CAT, tab bonferroni
```

GRUP_EDAD	Summary of GLUCEMIA		
	Mean	Std. Dev.	Freq.
1	132.50768	24.371201	456
2	137.747	21.662987	417
3	145.5748	22.595922	127
Total	136.352	23.424492	1,000

Analysis of Variance					
Source	SS	df	MS	F	Prob >F
Between groups	18353.2747	2	9176.63737	17.27	0.0000
Within groups	529804.821	997	531.399018		
Total	548158.096	999	548.706803		

Bartlett's test for equal variances: chi2(2) = 6.0976 Prob>chi2 = 0.047

Comparison of GLUCEMIA by IMC_CAT
(Bonferroni)

Row Mean - Col Mean	1	2
2	5.23933	
	0.002	
3	13.0671	7.8278
	0.000	0.003

♣ Pregunta 6.18

Solución con calculadora

Las hipótesis planteadas en el ejercicio son:

$$H_0 : \mu(\text{Glucemia})_{\text{normopeso}} = \mu(\text{Glucemia})_{\text{sobrepeso}} = \mu(\text{Glucemia})_{\text{obesidad}}$$



$$H_1 : \mu(\text{Glucemia})_{\text{normopeso}} \neq \mu(\text{Glucemia})_{\text{sobrepeso}} \neq \mu(\text{Glucemia})_{\text{obesidad}}$$

El valor de p es de 4.239×10^{-8} . Como p es menor de 0.05 por lo tanto **tenemos evidencia estadística suficiente para rechazar la hipótesis nula** y concluir que la glucemia varía según el IMC categorizado del paciente.

Al hacer comparaciones dos a dos con la t de Student corregida por Bonferroni vemos que hay diferencias significativas entre las comparaciones de las tres categorías. Así, podemos observar que en **la categoría 1 (normopeso; media 128.7) el nivel de glucemia es significativamente menor que en la categoría 2 (sobrepeso: media 142.63) y este, a su vez, presenta esta significación en la categoría 3 (obesidad; media 152.53).**

Estimación de una proporción



SOLUCIÓN 7.1

Solución con calculadora

La proporción de mujeres con menarquia antes de los 13 años es de un 61.55%

Solución con Excel

Primero, generamos la variable **Menar13**, que indica si la menarquia había sucedido antes (1) o después (0) de los 13 años:

En **Excel**, aplicaremos la función lógica **SI** :

```
=SI(ESBLANCO(H2); ;SI(H2<13;1;0))
```

donde H2 es el valor de la variable **EdadMenar** de la primera mujer de la base de datos.

Ahora, es necesario borrar aquellas celdas de la variable **Menar13** cuyo valor es , que, a pesar de que no son visibles, Excel interpreta como texto. Para ello, seleccionamos el rango de valores de la variable **Menar13** y pulsamos **F5**. Hacemos clic en «Especial...» y, activando *Celdas con fórmulas*, dejamos seleccionado únicamente «Texto». Después de aceptar, se suprimen las celdas seleccionadas.

Utilizando **Estadística descriptiva** del Analizador de datos con las opciones *Resumen de estadísticas* y *Nivel de de confianza para la media* marcadas, al tratarse de una variable binaria codificada como 0 y 1, la salida nos da la proporción de mujeres con menarquia antes de los 13 años bajo la etiqueta «Media».

Solución con Stata

En **Stata** utilizaremos los comandos **generate** y **replace** :

```
. generate Menar13 = EdadMenar <13  
  
. replace Menar13 = . if missing(EdadMenar)  
  
. tabulate Menar13
```

Menar13	Freq.	Percent	Cum.
0	584	38.45	38.45
1	935	61.55	100.00
Total	1,519	100.00	

Tabla 7.1: Solución Stata - Poblema 7.1

♣ Pregunta 7.1

Solución con calculadora

La proporción muestral es

$$p = \frac{935}{1519} = 0.6155 .$$

Empleando la aproximación normal, los extremos del intervalo de confianza al 95 % pueden calcularse mediante

$$p \pm z_{1-\alpha/2} \sqrt{\frac{p(1-p)}{n}} = 0.6155 \pm 1.96 \sqrt{\frac{0.6155 \times (1 - 0.6155)}{1519}} ,$$

es decir,

$$0.6155 \pm 1.96 \times 0.0125 \Rightarrow (0.5911; 0.6400) .$$

El intervalo de confianza al 95 % es de 0.5911 - 0.6400.

Solución con Excel

En la salida del Analizador de datos, el intervalo de confianza es, en realidad, la precisión del intervalo, es decir, la mitad de su amplitud. Por tanto, es necesario sumarla a y restarla de la proporción muestral para obtener el intervalo de confianza:

$$0.6155 \pm 0.0245 \Rightarrow (0.5911; 0.6400) .$$

Solución con Stata

En **Stata**, con el comando `ci` :

```
. ci proportions Menar13, wald
```

Variable	Obs	Proportion	Std. Err	Binomial Wald	
				[95% Conf. Interval]	
Menar13	1,519	.6155365	.0124818	.5910728	.6400003

Tabla 7.2: Solución Stata - Pregunta 7.1

♣ Pregunta 7.2

Solución con calculadora

Si se repitiera el estudio con muestras independientes un número elevado de veces, el 95% de los intervalos de confianza calculados contendrían el verdadero parámetro poblacional. Asumiendo que este intervalo de confianza pertenece a ese 95%, apoya la afirmación del enunciado.

♣ Pregunta 7.3

Solución con calculadora

$$H_0 : \pi = 0.65$$

$$H_1 : \pi \neq 0.65$$

Usando la aproximación normal, llevamos a cabo un contraste de hipótesis sobre la proporción muestral:

$$z = \frac{p - \pi_0}{\sqrt{\frac{\pi_0 \times (1 - \pi_0)}{n}}} = \frac{0.6155 - 0.65}{0.0125} = -2.8161$$

que sigue una distribución normal estándar.

Para calcular el valor p, en la tabla de la distribución normal $z = 2.82$, obtendremos $p = 0.0024$, como el contraste de hipótesis es a dos colas multiplicamos p por 2, $p = 0.0049$.

A la luz de los resultados obtenidos, se puede decir que existen diferencias estadísticamente significativas entre ambas proporciones por ser $p = 0.0049 < 0.05$, por lo que no puede asumirse que la muestra proceda de una población con $\pi = 0.65$.

Solución con Excel

En **Excel**, para calcular el valor p , teniendo en cuenta que se trata de un contraste de hipótesis a dos colas, escribiremos

```
=2*(1-DISTR.NORM.ESTAND.N(ABS(-2.8161);VERDADERO))
```

donde VERDADERO le indica a Excel que utilice la función de distribución en vez de la función de densidad de probabilidad.

Solución con Stata

En **Stata**, el comando `prtest` realiza la misma función:

```
. prtest Menar13 == 0.65
```

One-sample test of proportion Number of obs = 1519

Variable	Mean	Std. Err	[95% Conf. Interval]	
Menar13	.6155365	.0124818	.5910728	.6400003

$p = \text{proportion}(\text{Menar13})$ $z = -2.8161$

Ho: $p = 0.65$

Ha: $p < 0.65$

Ha: $p \neq 0.65$

Ha: $p > 0.65$

$\Pr(Z < z) = 0.0024$

$\Pr(|Z| > |z|) = 0.0049$

$\Pr(Z > z) = 0.9976$

Tabla 7.3: Solución Stata - Pregunta 7.3

Nota: en esta salida de Stata se obtienen 3 posibles valores de p . El que figura a la izquierda ($p = 0.0024$) corresponde a la hipótesis nula $H_0: \pi < 0.65$; el del centro ($p = 0.0049$) se refiere a la hipótesis nula $H_0: \pi = 0.65$ y el de la derecha ($p = 0.9976$) corresponde a la hipótesis nula $H_0: \pi > 0.65$. En este caso, puesto que la pregunta de investigación se plantea como igualdad ($H_0: \pi = 0.65$), nos quedaremos con el valor p del contraste bilateral ($p = 0.005$).

SOLUCIÓN 7.2

Solución con Excel

Primero, generamos la variable **Parto30**, que indica si el primer parto había sucedido antes (0) o después (1) de los 30 años.

En **Excel**, aplicaremos la función lógica **SI** :

```
=SI(K2=99; ;SI(K2>30;1;0))
```

donde K2 es el valor de la variable **EdadParto** de la primera mujer de la base de datos y 99 el código para los datos faltantes.

Ahora, es necesario borrar aquellas celdas de la variable **Parto30** cuyo valor es , que, a pesar de que no son visibles, Excel interpreta como texto. Para ello, seleccionamos el rango de valores de la variable **Parto30** y pulsamos **F5**. Hacemos clic en «Especial...» y, activando *Celdas con fórmulas*, dejamos seleccionado únicamente «Texto». Después de aceptar, se suprimen las celdas seleccionadas.

Utilizando **Estadística descriptiva** del Analizador de datos con las opciones *Resumen de estadísticas* y *Nivel de confianza para la media* marcadas, al tratarse de una variable binaria codificada como 0 y 1, la salida nos da la proporción de mujeres cuyo primer parto hubiera sucedido después de los 30 años bajo la etiqueta «Media»: 0.1171.

Solución con Stata

```
. generate Parto30 = EdadParto >30
```

```
. tabulate Parto30
```

Parto30	Freq.	Percent	Cum.
0	807	88.29	88.29
1	107	11.71	100.00
Total	914	100.00	

Tabla 7.4: Solución Stata - Poblema 7.2

♣ Pregunta 7.4

Solución con calculadora

La proporción muestral es

$$p = \frac{107}{914} = 0.1171$$

Empleando la aproximación normal, los extremos del intervalo de confianza al 90 % pueden calcularse mediante

$$p \pm z_{1-\alpha/2} \sqrt{\frac{p(1-p)}{n}} = 0.1171 \pm 1.645 \sqrt{\frac{0.1171 \times (1 - 0.1171)}{914}},$$

es decir,

$$0.1171 \pm 1.645 \times 0.0106 \Rightarrow (0.0996; 0.1346)$$

Nota: en este problema al tratarse de un IC al 90 % cambia el valor de $Z_{1-\alpha/2}$ (es de 1.645, en vez de 1.96).

Solución con Excel

En la salida del Analizador de datos, el intervalo de confianza es, en realidad, la precisión del intervalo, es decir, la mitad de su amplitud. Por tanto, es necesario sumarla a y restarla de la proporción muestral para obtener el intervalo de confianza:

$$0.1171 \pm 0.0175 \Rightarrow (0.0996; 0.1346)$$

Solución con Stata

En **Stata**, con el comando `ci`,

```
. ci proportions Parto30, wald level(90)
```

Variable	Obs	Proportion	Std. Err	Binomial Wald	
				[90% Conf. Interval]	
Parto30	914	.1170678	.0106343	.0995759	.1345597

Tabla 7.5: Solución Stata - Pregunta 7.4

♣ Pregunta 7.5

Solución con calculadora

Si se repitiera el estudio con muestras independientes un número elevado de veces, el 90 % de los intervalos de confianza calculados contendrían el verdadero parámetro poblacional. Asumiendo que este intervalo de confianza pertenece a ese 90 %, apoya la afirmación del enunciado.

♣ Pregunta 7.6

Solución con calculadora

Usando la aproximación normal, llevamos a cabo un contraste de hipótesis sobre la proporción muestral:

$$z = \frac{p - \pi_0}{\sqrt{\frac{\pi_0 \times (1 - \pi_0)}{n}}} = \frac{0.3196 - 0.35}{0.0138} = -2.1525 ,$$

que sigue una distribución normal estándar.

Para calcular el valor p, en la tabla de la distribución normal para $z=2.79$ encontramos $p=0.0026$, como el contraste de hipótesis es a dos colas, $p=0.0052$.

A la luz de los resultados obtenidos, se puede decir que existen diferencias estadísticamente significativas entre ambas proporciones por ser $p = 0.0053 < 0.10$, por lo que no puede asumirse que la muestra proceda de una población con $\pi = 0.15$.

Solución con Excel

En **Excel**, teniendo en cuenta que se trata de un contraste de hipótesis a dos colas, escribiremos

`=2*(1-DISTR.NORM.ESTAND.N(ABS(-2.7883);VERDADERO))`

donde VERDADERO le indica a Excel que utilice la función de distribución en vez de la función de densidad de probabilidad.

Solución con Stata

En **Stata**, el comando `prtest` realiza la misma función:

```
. prtest Parto30 == 0.15, level(90)
```

One-sample test of proportion Number of obs = 914

Variable	Mean	Std. Err	[90% Conf. Interval]	
Parto30	.1170678	.0106343	.0995759	.1345597

$p = \text{proportion}(\text{Parto30})$ $z = -2.7883$

Ho: $p = 0.15$

Ha: $p < 0.15$

Ha: $p \neq 0.15$

Ha: $p > 0.15$

$\Pr(Z < z) = 0.0026$

$\Pr(|Z| > |z|) = 0.0053$

$\Pr(Z > z) = 0.9974$

Tabla 7.6: Solución Stata - Pregunta 7.6

SOLUCIÓN 7.3

♣ Pregunta 7.7**Solución con calculadora**

La proporción muestral es

$$p = \frac{364}{1139} = 0.3196 \text{ .}$$

Empleando la aproximación normal, los extremos del intervalo de confianza al 90% pueden calcularse mediante

$$p \pm z_{1-\alpha/2} \sqrt{\frac{p(1-p)}{n}} = 0.3196 \pm 2.576 \sqrt{\frac{0.3196(1-0.3196)}{1139}} \text{ ,}$$

es decir,

$$0.3196 \pm 2.576 \times 0.0138 \Rightarrow (0.284; 0.355) \text{ .}$$

Solución con Excel

Usando el Analizador de datos, sumamos a y restamos de la proporción muestral («Media») la precisión del intervalo («Nivel de confianza (99.0%)»):

$$0.3196 \pm 0.0354 \Rightarrow (0.284; 0.355) \text{ .}$$

Solución con Stata

En **Stata**, con el comando `ci` ,

```
. ci proportions THS, wald level(99)
```

Variable	Obs	Proportion	Std. Err	Binomial Wald	
				[99% Conf. Interval]	
THS	1,139	.3195786	.0138171	.2839882	.355169

Tabla 7.7: Solución Stata - Pregunta 7.7

♣ Pregunta 7.8

Solución con calculadora

Si se repitiera el estudio con muestras independientes un número elevado de veces, el 99% de los intervalos de confianza calculados contendrían el verdadero parámetro poblacional. Asumiendo que este intervalo de confianza pertenece a ese 99%, apoya la afirmación del enunciado.

♣ Pregunta 7.9

Solución con calculadora

Usando la aproximación normal, llevamos a cabo un contraste de hipótesis sobre la proporción muestral:

$$z = \frac{p - \pi_0}{\sqrt{\frac{\pi_0 \times (1 - \pi_0)}{n}}} = \frac{0.3196 - 0.35}{0.0138} = -2.1525 ,$$

que sigue una distribución normal estándar.

Para calcular el valor p, en la tabla de la distribución normal para $z=2.15$ encontramos $p=0.0158$, como el contraste de hipótesis es a dos colas, $p=0.0316$.

A la luz de los resultados arrojados, se puede decir que no existen diferencias estadísticamente significativas entre ambas proporciones por ser $p = 0.0314 > 0.01$, por lo que no puede descartarse que la muestra proceda de una población con $\pi = 0.35$.

Solución con Excel

En **Excel** teniendo en cuenta que se trata de un contraste de hipótesis a dos colas, escribiremos

```
=2*(1-DISTR.NORM.ESTAND.N(ABS(-2.1525);VERDADERO))
```

donde VERDADERO le indica a Excel que utilice la función de distribución en vez de la función de densidad de probabilidad.

Solución con Stata

En **Stata**, el comando `prtest` realiza la misma función:

```
. prtest THS == 0.35, level(99)
```

One-sample test of proportion			Number of obs = 1139	
Variable	Mean	Std. Err	[99% Conf. Interval]	
THS	.3195786	.0138171	.2839882	.355169
p = proportion(THS)			z = -2.1525	
Ho: p = 0.35				
Ha: p <0.35		Ha: p != 0.35		Ha: p >0.35
Pr(Z <z) = 0.0157		Pr(Z > z) = 0.0314		Pr(Z >z) = 0.9843

Tabla 7.8: Solución Stata - Pregunta 7.9

SOLUCIÓN 7.4

♣ Pregunta 7.10

Solución con calculadora

La proporción muestral es

$$p = \frac{6}{41} = 0.1463 \text{ .}$$

Empleando la aproximación normal, los extremos del intervalo de confianza de π al 95 % pueden calcularse mediante

$$0.1463 \pm 1.96 \sqrt{\frac{0.1463(1 - 0.1463)}{41}} \text{ ,}$$

es decir,

$$0.1463 \pm 1.96 \times 0.0552 \Rightarrow (0.0382; 0.2545) \text{ .}$$

Solución con Stata

En **Stata**, con el comando **cii**,

```
. cii proportions 41 6, wald
```

Variable	Obs	Proportion	Std. Err	Binomial Wald	
				[95% Conf. Interval]	
	41	.1463415	.0551993	.0381528	.2545302

Tabla 7.9: Solución Stata - Pregunta 7.10

♣ Pregunta 7.11

Solución con calculadora

Se trata de hallar el intervalo (a, b) tal que

$$P(X \leq 5 \mid \pi = a) = 0.975 \quad y \quad P(X \leq 6 \mid \pi = b) = 0.025$$

Solución con Stata

En **Stata**, respectivamente, `invbinomial(41, 5, 0.975)` y `invbinomial(41, 6, 0.025)`, es decir, (0,0557;0,2917).

Con el comando `cii`,

```
. cii proportions 41 6, exact
```

Variable	Obs	Proportion	Std. Err	Binomial Exact	
				[95% Conf. Interval]	
	41	.1463415	.0551993	.0556574	.2917305

Tabla 7.10: Solución Stata - Pregunta 7.11

♣ Pregunta 7.12**Solución con calculadora**

Si se repitiera el estudio con muestras independientes un número elevado de veces, el 95 % de los intervalos de confianza calculados contendrían el verdadero parámetro poblacional. Asumiendo que este intervalo de confianza pertenece a ese 95 %, la proporción de mujeres que, con las características (1) y (2), recibe THS durante más de dos años es menor que la de aquellas que, con las mismas características, reciben THS durante menos de 2 años.

8

Comparación de dos proporciones

SOLUCIÓN 8.1

♣ Pregunta 8.1**Solución con calculadora**

	Mujer con cáncer	Mujer sin cáncer	
Consumo AO > 1 año	548	610	1158
Consumo AO < 1 año	163	196	359
	711	806	1517

♣ Pregunta 8.2**Solución con calculadora**

$$P(AO > 1 \text{ año}) = \frac{1158}{1517} = 0.763$$

♣ Pregunta 8.3**Solución con calculadora**

$$p_1 = P(AO > 1 \text{ año} \mid C = 1) = \frac{548}{711} = 0.771$$

$$p_0 = P(AO > 1 \text{ año} \mid C = 0) = \frac{610}{806} = 0.757$$

♣ Pregunta 8.4

Solución con calculadora

Para responder a esta pregunta, se calcula el valor del estadístico z de la distribución normal:

$$z = \frac{p_0 - p_1}{EE(p_0 - p_1)}$$

Así,

$$p_0 - p_1 = \frac{610}{806} - \frac{548}{711} = -0.0139$$

y

$$EE(p_0 - p_1) = \sqrt{\frac{1158}{1517} \left(1 - \frac{1158}{1517}\right) \left(\frac{1}{806} + \frac{1}{711}\right)} = 0.0219$$

Por lo tanto,

$$z = -\frac{0.0139}{0.0219} = -0.6366$$

que sigue una distribución normal estándar.

Para calcular el valor p , teniendo en cuenta que se trata de un contraste de hipótesis a dos colas:

En la tabla de la distribución normal estandarizada obtenemos para $z = -0.64$ un valor $p = 0.261$ que multiplicamos este valor por 2 (contraste de hipótesis a dos colas) obteniendo una $p = 0.522$.

A la luz de los resultados, se puede decir que no existen diferencias estadísticamente significativas entre ambas proporciones por ser $p = 0.5244 > 0.05$.

Solución con Excel

En **Excel**, escribiremos:

```
=2*(1-DISTR.NORM.ESTAND.N(ABS(-0.6366);VERDADERO))
```

donde VERDADERO le indica a Excel que utilice la función de distribución en vez de la función de densidad de probabilidad. Esto da un resultado de 0.5244

Solución con Stata

En **Stata**, el comando `prtest`, junto con su versión inmediata, realiza la misma función. Los parámetros que se le introducen son el número de observaciones de una de las muestras, su proporción, el número de observaciones de la otra muestra y su proporción:

```
. prtesti 806 `=610/806' 711 `=548/711'
```

Two-sample test of proportions					x: Number of obs = 806
					y: Number of obs = 711
	Mean	Std. Err.	z	P> z	[95% Conf. Interval]
x	.7568238	.0151109			.727207 .7864406
y	.7707454	.0157645			.7398476 .8016432
diff	-.0139216	.0218371			-.0567215 .0288783
	under Ho:	.0218679	-0.64	0.524	
diff = prop(x) - prop(y)					z = -0.6366
Ho: diff = 0					
Ha: diff <0		Ha: diff != 0		Ha: diff >0	
Pr(Z <z) = 0.2622		Pr(Z > z) = 0.5244		Pr(Z >z) = 0.7378	

♣ Pregunta 8.5**Solución con calculadora**

Calculados la diferencia de proporciones y su error estándar en la pregunta anterior, la construcción del intervalo de confianza al 95 % de la diferencia de proporciones es inmediata:

$$-0.0139 \pm 1.96 \times 0.0219 \Rightarrow (-0.0567; 0.0289)$$

Es decir, si se repitiera el mismo experimento un número elevado de veces, el 95 % de las veces, el verdadero valor de la diferencia de proporciones estaría incluido entre los extremos del intervalo de confianza que hemos calculado. Al incluir tanto valores positivos como negativos, se puede concluir que las diferencias no son estadísticamente significativas.

Solución con Excel

La comparación de una variable binaria en función de otra variable binaria es equivalente a una prueba t de Student. Así, ordenamos por la variable `Caso` y etiquetamos las submuestras de `Anov0rales` que en función de aquella se generan. En el analizador de datos, seleccionamos *Prueba t para dos muestras suponiendo varianzas iguales* e indicamos el nombre de las etiquetas que se refieren a las submuestras antes mencionadas. El intervalo de confianza al 95 % es $(-0.0567; 0.0289)$.

Solución con Stata

El comando usado para resolver la pregunta anterior, `prtesti`, retorna también el intervalo de confianza de la comparación de dos proporciones:

$(-0.0567; 0.0289)$

SOLUCIÓN 8.2

♣ **Pregunta 8.6****Solución con Excel**

Para obtener la tabla observada procederemos de la siguiente manera:

Aplicaremos la orden `CONTAR.SI.CONJUNTO(VAR1;valor;VAR2;valor)` a cada celda interior de la tabla. Esta orden cuenta el número de celdas de la base de datos que cumplen unos determinados criterios. Una vez etiquetadas las variables Hijos y Caso, la orden para la casilla 1,1 sería:

`=CONTAR.SI.CONJUNTO(Menopausia;1;Caso;1)`

	Mujer con cáncer	Mujer sin cáncer	
Menopáusicas	=CONTAR.SI.CONJUNTO(Menopausia;1;Caso;1)		
No menopáusicas			
Total			

Solución con Stata

En Stata,

`. tab Menopausia Caso`

	Mujer con cáncer	Mujer sin cáncer	Total
Menopáusicas	544	604	1148
No menopáusicas	167	211	378
Total	711	815	1526

♣ **Pregunta 8.7**

Solución con calculadora

Sean π_0 y π_1 las proporciones poblacionales de menopáusicas en mujeres sin o con cáncer de mama, respectivamente. Así,

$$H_0 : \pi_0 \geq \pi_1$$

$$H_1 : \pi_0 < \pi_1$$

Nota: en este caso el contraste de hipótesis será unilateral (o de 1 cola) ya que se plantea el sentido de la diferencia de proporciones.

♣ **Pregunta 8.8**

Solución con calculadora

$$p_1 = P(\text{Menopáusicas} \mid C = 1) = \frac{544}{711} = 0.765$$

$$p_0 = P(\text{Menopáusicas} \mid C = 0) = \frac{604}{815} = 0.741$$

Así,

$$p_0 - p_1 = \frac{604}{815} - \frac{544}{711} = -0.024$$

y

$$EE(p_0 - p_1) = \sqrt{\frac{1148}{1526} \left(1 - \frac{1148}{1526}\right) \left(\frac{1}{815} + \frac{1}{711}\right)} = 0.0222$$

Por lo tanto,

$$z = -\frac{0.024}{0.0222} = -1.081$$

que sigue una distribución normal estándar.

Para calcular el valor p , teniendo en cuenta que se trata de un contraste de hipótesis a dos colas,

En la tabla de la distribución normal estandarizada obtenemos para $z = 1.08$ un valor $p = 0.14$.

Solución con Excel

En **Excel** escribiremos

```
=DISTR.NORM.ESTAND.N(-1.081;VERDADERO)
```

donde VERDADERO le indica a Excel que utilice la función de distribución en vez de la función de densidad de probabilidad. Nótese que, en esta ocasión, al tratarse de un contraste de hipótesis a una cola, no hemos calculado el valor absoluto del estadístico de contraste ni hemos multiplicado por dos el valor p resultante. Tampoco se lo hemos restado a 1, ya que se trata de un contraste de cola izquierda.

Esto da un resultado de $p = 0.139$.

Solución con Stata

En **Stata**, el comando `prtest`, junto con su versión inmediata, realiza la misma función:

```
. prtesti 815 "=604/815" 711 "=544/711"
```

Two-sample test of proportions					x: Number of obs = 815
					y: Number of obs = 711
	Mean	Std. Err.	z	P> z	[95% Conf. Interval]
x	.7411043	.0153435			.7110317 .7711769
y	.7651195	.0158984			.7339593 .7962798
diff	-.0240153	.0220948			-.0673203 .0192898
	under Ho:	.0221527	-1.08	0.278	
diff = prop(x) - prop(y)					z = -1.0841
Ho: diff = 0					
Ha: diff <0		Ha: diff != 0		Ha: diff >0	
Pr(Z <z) = 0.1392		Pr(Z > z) = 0.2783		Pr(Z >z) = 0.8608	

♣ Pregunta 8.9**Solución con calculadora**

A la luz de los resultados obtenidos, se puede decir que no existen diferencias estadísticamente significativas entre ambas proporciones por ser $p = 0.1392 > 0.05$.



Por tanto el estado menopáusico no se asocia en nuestro estudio al cáncer de mama.

♣ Pregunta 8.10

proporciones de menopáusicas en ambos grupos.

Solución con calculadora

Calculados la diferencia de proporciones y su error estándar, la construcción del intervalo de confianza al 95 % de la diferencia de proporciones es inmediata:

$$-0.024 \pm 1.96 \times 0.0222 \Rightarrow (-0.0674; 0.01940)$$

Es decir, si se repitiera el mismo experimento un número elevado de veces, el 95 % de las veces, el verdadero valor de la diferencia de proporciones estaría incluido entre los extremos del intervalo de confianza que hemos calculado. Al incluir tanto valores positivos como negativos, se puede concluir que las diferencias no son estadísticamente significativas.

Solución con Excel

La comparación de una variable binaria en función de otra variable binaria es equivalente a una prueba *t* de Student. Así, ordenamos por la variable `Caso` y etiquetamos las submuestras de `Menopausia` que en función de aquella se generan. En el analizador de datos, seleccionamos *Prueba t para dos muestras suponiendo varianzas iguales* e indicamos el nombre de las etiquetas que se refieren a las submuestras antes mencionadas. El intervalo de confianza al 95 % es $(-0.0674; 0.01940)$.

Solución con Stata

El comando usado para resolver la pregunta anterior, `prtesti`, retorna también el intervalo de confianza de la comparación de dos proporciones:

$$(-0.0674; 0.01940)$$

SOLUCIÓN 8.3

♣ Pregunta 8.11**Solución con Excel**

Antes de nada, hay que crear la variable dicotómica Hijos y codificarla como $0 = No$ / $1 = S$. Para ello, contamos con la variable NumHijos (número de hijos), que podemos transformar de forma sencilla.

En **Excel**, aplicaremos la función lógica SI :

```
=SI ( J2=0;0;1 )
```

donde J2 es el valor de la variable *NumHijos* de la primera mujer de la base de datos.

Para obtener la tabla observada, aplicaremos la orden `CONTAR.SI.CONJUNTO(VAR1;valor;VAR2;valor)` a cada celda interior de la tabla. Esta orden cuenta el número de celdas de la base de datos que cumplen unos determinados criterios. Una vez etiquetadas las variables Hijos y Caso, la orden para la casilla 1,1 sería:

```
=CONTAR.SI.CONJUNTO(Hi jos;1;Caso;1)
```

Solución con Stata

Antes de nada, hay que crear la variable dicotómica Hijos y codificarla como $0 = No$ / $1 = S$. Para ello, contamos con la variable NumHijos (número de hijos), que podemos transformar de forma sencilla.

En **Stata**,

```
recode NumHi jos (0=0) (1/max=1), generate(Hi jos)
```

Para obtener la tabla observada utilizamos el comando `tab` ,

```
. tab Hi jos Caso
```


Solución con calculadora

Así, la tabla de contingencia queda como sigue:

	Mujer con cáncer	Mujer sin cáncer	
Con hijos	413	500	913
Sin hijos	306	318	624
	719	818	1537

♣ Pregunta 8.12**Solución con calculadora**

En este caso, el planteamiento de las hipótesis es bilateral, puesto que los investigadores pretenden evaluar si hay asociación entre haber tenido hijos y desarrollar cáncer de mama, pero no se plantea la dirección de dicha asociación. Así, sean π_0 y π_1 las proporciones poblacionales de mujeres con hijos en aquellas sin o con cáncer de mama, respectivamente. Así,

$$H_0 : \pi_0 = \pi_1$$

$$H_1 : \pi_0 \neq \pi_1$$

♣ Pregunta 8.13**Solución con calculadora**

$$p_1 = P(\text{Con hijos} \mid C = 1) = \frac{413}{719} = 0.574$$

$$p_0 = P(\text{Sin hijos} \mid C = 0) = \frac{500}{818} = 0.611$$

Así,

$$p_0 - p_1 = \frac{500}{818} - \frac{413}{719} = 0.0368$$

y

$$EE(p_0 - p_1) = \sqrt{\frac{913}{1537} \left(1 - \frac{913}{1537}\right) \left(\frac{1}{818} + \frac{1}{719}\right)} = 0.0251$$

Por lo tanto,

$$z = \frac{0.0368}{0.0251} = 1.4661$$

que sigue una distribución normal estándar.

Para calcular el valor p , teniendo en cuenta que se trata de un contraste de hipótesis a dos colas,

En la tabla de la distribución normal estandarizada obtenemos para $z = 1.47$ un valor $p = 0.070$ que multiplicado por 2 (al ser un contraste bilateral) nos da $p = 0.14$.

Solución con Excel

En **Excel** escribiremos

```
=2*(1-DISTR.NORM.ESTAND.N(ABS(1.4661);VERDADERO))
```

donde VERDADERO le indica a Excel que utilice la función de distribución en vez de la función de densidad de probabilidad.

La solución obtenida con Excel es $p = 0.0711$, como el contraste es bilateral habrá que multiplicar este resultados por 2, con lo que obtendremos un valor $p = 0.142$.

Solución con Stata

En **Stata**, el comando `prtest` realiza la misma función:

```
. prtest Hijos, by(Caso)
```

Two-sample test of proportions					0: Number of obs = 818
					1: Number of obs = 719
	Mean	Std. Err.	z	P> z	[95% Conf. Interval]
0	.6112469	.0170439			.5778415 .6446523
1	.5744089	.0184392			.5382687 .6105491
diff	.036838	.0251097			-.0123761 .0860522
	under Ho:	.0251044	1.47	0.142	
diff = prop(0) - prop(1)					z = 1.4674
Ho: diff = 0					
Ha: diff <0			Ha: diff != 0		Ha: diff >0
Pr(Z <z) = 0.9289			Pr(Z > z) = 0.1423		Pr(Z >z) = 0.0711

♣ Pregunta 8.14

Solución con calculadora

A la luz de los resultados arrojados, se puede decir que no existen diferencias estadísticamente significativas entre ambas proporciones por ser $p = 0.1423 > 0.05$.

♣ Pregunta 8.15

Solución con calculadora

Calculados la diferencia de proporciones y su error estándar, la construcción del intervalo de confianza al 95 % de la diferencia de proporciones es inmediata:

$$0.0368 \pm 1.96 \times 0.0251 \Rightarrow (-0.0124; 0.0861)$$

Es decir, si se repitiera el mismo experimento un número elevado de veces, el 95 % de las veces, el verdadero valor de la diferencia de proporciones estaría incluido entre los extremos del intervalo de confianza que hemos calculado. Al incluir el IC tanto valores positivos como negativos, se puede concluir que las diferencias no son estadísticamente significativas.

Solución con Excel

La comparación de una variable binaria en función de otra variable binaria es equivalente a una prueba t de Student. Así, ordenamos por la variable `Caso` y etiquetamos las submuestras de `Hi` y `jos` que en función de aquella se generan. En el analizador de datos, seleccionamos *Prueba t para dos muestras suponiendo varianzas iguales* e indicamos el nombre de las etiquetas que se refieren a las submuestras antes mencionadas. El intervalo de confianza al 95 % es $(-0.0124; 0.0861)$.

Solución con Stata

El comando usado para resolver la pregunta anterior, `prtest`, retorna también el intervalo de confianza de la comparación de dos proporciones:

$$(-0.0124; 0.0861)$$

SOLUCIÓN 8.4

♣ Pregunta 8.16

Solución con calculadora

	Mujer con cáncer	Mujer sin cáncer	
Con THS	168	196	364
Sin THS	375	400	775
	543	596	1139

Solución con Excel

Para obtener la tabla observada, aplicaremos la orden `CONTAR.SI.CONJUNTO(VAR1;valor;VAR2;valor)` a cada celda interior de la tabla. Esta orden cuenta el número de celdas de la base de datos que cumplen unos determinados criterios. Una vez etiquetadas las variables THS y Caso, la orden para la casilla 1,1 sería:

`=CONTAR.SI.CONJUNTO(THS;1;Caso;1)`

Solución con Stata

`. tab THS Caso`

♣ Pregunta 8.17

Solución con calculadora

Nuevamente, el planteamiento de las hipótesis es bilateral, puesto que los investigadores pretenden evaluar si hay asociación entre el consumo de THS y desarrollar cáncer de mama, pero no se plantea la dirección de dicha asociación. Así, sean π_0 y π_1 las proporciones poblacionales de mujeres con THS en aquellas sin o con cáncer de mama, respectivamente. Así,

$$H_0 : \pi_0 = \pi_1$$

$$H_1 : \pi_0 \neq \pi_1$$

♣ Pregunta 8.18

Solución con calculadora

$$p_1 = P(\text{Con THS} \mid C = 1) = \frac{168}{543} = 0.309$$

$$p_0 = P(\text{Sin THS} \mid C = 0) = \frac{196}{596} = 0.329$$

Así,

$$p_0 - p_1 = \frac{196}{596} - \frac{168}{543} = 0.0195$$

y

$$EE(p_0 - p_1) = \sqrt{\frac{364}{1139} \left(1 - \frac{364}{1139}\right) \left(\frac{1}{596} + \frac{1}{543}\right)} = 0.0277$$

Por lo tanto,

$$z = \frac{0.0195}{0.0277} = 0.70397$$

que sigue una distribución normal estándar.

Para calcular el valor p , teniendo en cuenta que se trata de un contraste de hipótesis a dos colas,

En la tabla de la distribución normal estandarizada obtenemos para $z = 0.70$ un valor $p = 0.242$, como es un contraste a dos colas, multiplicaremos este valor por 2, obteniendo $p = 0.484$.

Solución con Excel

En Excel escribiremos

$$=2*(1-\text{DISTR.NORM.ESTAND.N}(\text{ABS}(0.70397);\text{VERDADERO}))$$

donde VERDADERO le indica a Excel que utilice la función de distribución en vez de la función de densidad de probabilidad. Eso da un resultado de 0.4815.

Solución con Stata

En **Stata**, el comando `prtest` realiza la misma función:

```
. prtest THS if THS != 9, by(Caso)
```

Two-sample test of proportions					0: Number of obs = 596
					1: Number of obs = 543
	Mean	Std. Err.	z	P> z	[95% Conf. Interval]
0	.3288591	.0192437			.2911421 .366576
1	.3093923	.0198368			.2705129 .3482716
diff	.0194668	.0276372			-.0347012 .0736348
	under Ho:	.0276641	0.70	0.482	
diff = prop(0) - prop(1)					z = 0.7037
Ho: diff = 0					
Ha: diff <0		Ha: diff != 0		Ha: diff >0	
Pr(Z <z) = 0.7592		Pr(Z > z) = 0.4816		Pr(Z >z) = 0.2408	

♣ Pregunta 8.19

Solución con calculadora

A la luz de los resultados arrojados, se puede decir que no existen diferencias estadísticamente significativas entre ambas proporciones por ser $p = 0.4816 > 0.05$.

♣ Pregunta 8.20

Solución con calculadora

Calculados la diferencia de proporciones y su error estándar, la construcción del intervalo de confianza al 95% de la diferencia de proporciones es inmediata:

$$0.0195 \pm 1.96 \times 0.0277 \Rightarrow (-0.0347; 0.0737)$$

Es decir, si se repitiera el mismo experimento un número elevado de veces, el 95 %

de las veces, el verdadero valor de la diferencia de proporciones estaría incluido entre los extremos del intervalo de confianza que hemos calculado. Al incluir tanto valores positivos como negativos, se puede concluir que las diferencias no son estadísticamente significativas.

Solución con Excel

La comparación de una variable binaria en función de otra variable binaria es equivalente a una prueba t de Student. Así, ordenamos por la variable `Caso` y etiquetamos las submuestras de `THS` que en función de aquella se generan. En el analizador de datos, seleccionamos *Prueba t para dos muestras suponiendo varianzas iguales* e indicamos el nombre de las etiquetas que se refieren a las submuestras antes mencionadas. El intervalo de confianza al 95 % es $(-0.0347; 0.0737)$.

Solución con Stata

El comando usado para resolver la pregunta anterior, `prtest`, retorna también el intervalo de confianza de la comparación de dos proporciones:

$(-0.0347; 0.0737)$

SOLUCIÓN 8.5

♣ Pregunta 8.21**Solución con calculadora**

Sean π_0 y π_1 las proporciones poblacionales de menarquia precoz en mujeres sin o con cáncer de mama, respectivamente. Así,

$$H_0 : \pi_0 \geq \pi_1$$

$$H_1 : \pi_0 < \pi_1$$

Nota: en este caso el contraste de hipótesis será unilateral (o de 1 cola) ya que se plantea el sentido de la diferencia de proporciones.

♣ Pregunta 8.22**Solución con calculadora**

$$p_1 = P(\text{Menarquia precoz} \mid C = 1) = \frac{191}{709} = 0.269$$

$$p_0 = P(\text{Menarquia precoz} \mid C = 0) = \frac{170}{810} = 0.210$$

Así,

$$p_1 - p_0 = \frac{191}{709} - \frac{170}{810} = 0.06$$

y

$$EE(p_0 - p_1) = \sqrt{\frac{361}{1519} \left(1 - \frac{361}{1519}\right) \left(\frac{1}{709} + \frac{1}{810}\right)} = 0.0219$$

Por lo tanto,

$$z = -\frac{0.06}{0.0219} = -2.719$$

En la tabla de la distribución normal estandarizada obtenemos para $z = -2.72$ un valor $p = 0.0033$.

Solución con Excel

En **Excel**, escribiremos:

```
=1-DISTR.NORM.ESTAND.N(2.719;VERDADERO)
```

donde VERDADERO le indica a Excel que utilice la función de distribución en vez de la función de densidad de probabilidad. Esto da un resultado de 0.00327.

Solución con Stata

En **Stata**, el comando `prtest`, junto con su versión inmediata, realiza la misma función. Los parámetros que se le introducen son el número de observaciones de una de las muestras, su proporción, el número de observaciones de la otra muestra y su proporción:

```
. prtest Menar11 if Menar11 != 9, by(Caso)
```

Two-sample test of proportions

0: Number of obs = 810

1: Number of obs = 709

	Mean	Std. Err.	z	P> z	[95% Conf. Interval]
0	.2098765	.0143083			.1818329 .2379202
1	.2693935	.0166614			.2367377 .3020493
diff	-.059517	.021962			-.1025617 -.0164722
	under Ho:	.0218909	-2.72	0.007	

diff = prop(0) - prop(1)

z = -2.7188

Ho: diff = 0

Ha: diff <0

Ha: diff != 0

Ha: diff >0

Pr(Z <z) = 0.0033

Pr(|Z| >|z|) = 0.0066

Pr(Z >z) = 0.9967

♣ Pregunta 8.23

Solución con calculadora

Sí se asocian. La proporción de mujeres que tuvieron una menarquia precoz es significativamente mayor en las mujeres con cáncer de mama.

♣ Pregunta 8.24

Solución con calculadora

Calculados la diferencia de proporciones y su error estándar, la construcción del intervalo de confianza al 95 % de la diferencia de proporciones es inmediata:

$$0.0595 \pm 1.96 \times 0.0219 \Rightarrow (0.0166; 0.103)$$

Si se repitiera el estudio con muestras independientes un número elevado de veces, el 95% de los intervalos de confianza calculados contendrían el verdadero parámetro poblacional.

Solución con Excel

La comparación de una variable binaria en función de otra variable binaria es equivalente a una prueba *t* de Student. Así, ordenamos por la variable *Caso* y etiquetamos las submuestras de *Menar11* que en función de aquella se generan. En el analizador de datos, seleccionamos *Prueba t para dos muestras suponiendo varianzas iguales* e indicamos el nombre de las etiquetas que se refieren a las submuestras antes mencionadas. El intervalo de confianza al 95% es (0.0166;0.103).

Solución con Stata

El comando usado para resolver la pregunta anterior, `prtest`, retorna también el intervalo de confianza de la comparación de dos proporciones:

$$(0.0166; 0.103)$$

Tabla de contingencia

SOLUCIÓN 9.1

♣ Pregunta 9.1

Solución con calculadora

	Mujer con cáncer	Mujer sin cáncer	Total
Consumo AO > 1 año	548	610	1158
Consumo AO < 1 año	163	196	359
Total	711	806	1517

Tabla 9.1: Relación entre consumo de AO y desarrollo de cáncer de mama (**Tabla observada**)

Para completar la **tabla de valores esperados** utilizaremos la siguiente fórmula:

$$E_{i,j} = \frac{n_{i.} \times n_{.j}}{n}$$

Por ejemplo, el valor esperado en la casilla 1,1 es:

$$E_{1,j} = \frac{n_{1.} \times n_{.j}}{n} = \frac{1158 \times 711}{1517} = 542.74$$

	Mujer con cáncer	Mujer sin cáncer	Total
Consumo AO > 1 año	542.74 $\frac{1158 \times 711}{1517}$	615.26 $\frac{1158 \times 806}{1517}$	1158
Consumo AO < 1 año	168.26 $\frac{359 \times 711}{1517}$	190.74 $\frac{359 \times 806}{1517}$	359
Total	711	806	1517

Tabla 9.2: Relación entre consumo de AO y desarrollo de cáncer de mama (**Tabla esperada**)

Solución con Excel

Para preparar la tabla de datos observados en **Excel** puede utilizarse una *tabla dinámica* o la opción **CONTARSICONJUNTO**.

Solución con Stata

El comando `tabulate` con la opción `expected` produce la misma salida que la del cuadro «Solución»:

```
tabulate Anov0rales Caso, expected
```

♣ Pregunta 9.2

Solución con calculadora

$$\begin{aligned}\chi^2 &= \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \\ &= \frac{(548 - 542.7)^2}{542.7} + \frac{(610 - 615.3)^2}{615.3} + \frac{(163 - 168.3)^2}{168.3} + \frac{(196 - 190.7)^2}{190.7} \\ &= 0.41\end{aligned}$$

El valor de la ji cuadrado (0.41) es igual al cuadrado del valor obtenido aplicando la prueba z ($0.636^2 = 0.4053$). Ambos procedimientos son válidos para resolver la comparación de dos proporciones.

Solución con Excel

Una vez se tienen las tablas observada y esperada, el estadístico χ^2 se obtiene aplicando la fórmula mediante operaciones matriciales, es decir, se escribe en una celda:

```
(RANGO_OBSERVADO-RANGO_ESPERADO)^2/RANGO_ESPERADO
```

y se pulsa la combinación Ctrl + Mayús + Intro: 0.41.

Solución con Stata

El comando `tabulate` con la opción `chi2` produce la misma salida que la del cuadro «Solución»:

```
tabulate AnovOrales Caso, chi2:0.41
```

♣ Pregunta 9.3**Solución con calculadora**

En la tabla de la ji cuadrado, con 1 g.l. el valor obtenido de χ^2 (0.41) se sitúa entre 0.27 y 0.45, luego $0.60 > p > 0.50$

Solución con Excel

En **Excel** se obtiene el valor exacto de p con la orden:

```
DISTR.CHICUAD.CD (0.41;1)
```

, donde **0.41** es el valor del estadístico de contraste (es decir de χ^2 calculado arriba) y **1** son los grados de libertad (se obtienen multiplicado el número de filas -1 por el número de columnas -1: $(f - 1) \times (c - 1)$).

Se obtiene $p = 0.524$

Solución con Stata

El comando `tabulate` con la opción `chi2` produce la misma salida que la del cuadro «Solución»:

```
tabulate AnovOrales Caso, chi2:0.524.
```

♣ Pregunta 9.4**Solución con calculadora**

Valor crítico de χ^2 (para tablas de 1 grado de libertad y nivel de significación de 0.05)=**3.84**

Solución con Excel

En Excel se obtiene con la orden:

```
INVCHICUAD.CD (0.05;1)
```

, donde 0.05 es el nivel de significación y 1 los g.l. de la tabla de contingencia.

Solución con Stata

En Stata, la función se denomina `invchi2tail`, y se le suministran los grados de libertad y el nivel de significación:

```
display invchi2tail(1, 0.05) : 3.84.
```

♣ Pregunta 9.5**Solución con calculadora**

A la vista de los resultados podemos concluir:

- 1) El consumo de anovulatorios orales y el cáncer de mama son variables independientes.
- 2) La proporción de consumidoras de anovulatorios orales entre las mujeres con o sin cáncer de mama no presenta diferencias estadísticamente significativas.

Solución con Stata

En Stata todos los cálculos anteriores se realizarían con la siguiente orden:

```
. tab Anov0rales Caso, chi
```

Anov0rales	Caso		Total
	0	1	
0	196	163	359
1	610	548	1,158
Total	806	711	1,517

Pearson $\chi^2(1) = 0.4053$ Pr = 0.524

En general, para obtener el **valor crítico** del estadístico χ^2 en Stata se utiliza la orden:

```
. di invchi2tail(df, p)
```

, donde df son los grados de libertad y p la p.

En nuestro estudio, como trabajamos con una tabla de 2x2 (anovulatorios (Si/No) y cáncer (Si/No) sustituiremos los g.l. por 1:

```
. di invchi2tail(1, 0.05)
```

El resultado obtenido es: 3.8414588.

SOLUCIÓN 9.2

♣ Pregunta 9.6

Solución con calculadora

A partir de la tabla observada (la podemos hacer en Excel con la opción tabla dinámica)

	Mujer con cáncer	Mujer sin cáncer	Total
THS	168	196	364
No THS	375	400	775
Total	543	596	1139

Tabla 9.3: Relación entre THS y desarrollo de cáncer de mama (**Tabla observada**)

vemos que hay 168 consumidoras de THS entre las 543 mujeres con cáncer de mama luego la proporción de consumo de THS es

$$\frac{168}{543} = 0.3094 = 30.94\%$$

, mientras que de las 596 mujeres sin cáncer de mama, 196 consumieron THS,

$$\frac{196}{596} = 0.3289 = 32.89\%$$

♣ Pregunta 9.7

Solución con calculadora

Para ello calcularemos la tabla de efectivos esperados y aplicaremos la prueba χ^2 .

Para completar la tabla de valores esperados utilizaremos la siguiente fórmula:

$$E_{i,j} = \frac{n_{i.} \times n_{.j}}{n}$$

Por ejemplo, el valor esperado en la casilla 1,1 es:

$$E_{1,j} = \frac{n_{1.} \times n_{.j}}{n} = \frac{364 \times 543}{1139} = 173.53$$

	Mujer con cáncer	Mujer sin cáncer	Total
THS	173.53 $\frac{364 \times 543}{1139}$	190.47 $\frac{364 \times 596}{1139}$	364
No THS	369.47 $\frac{775 \times 543}{1139}$	405.53 $\frac{775 \times 596}{1139}$	775
Total	543	596	1139

Tabla 9.4: Relación entre THS y desarrollo de cáncer de mama (**Tabla esperada**)

Cálculo del estadístico de contraste χ^2 :

$$\begin{aligned} \chi^2 &= \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \\ &= \frac{(168 - 173.5)^2}{173.5} + \frac{(196 - 190.5)^2}{190.5} + \frac{(375 - 369.5)^2}{369.5} + \frac{(400 - 405.5)^2}{405.5} \\ &= 0.495 \end{aligned}$$

Nota: Este problema se ha resuelto en el capítulo anterior (problema 8.4) aplicando la distribución normal. Puede comprobarse que elevando al cuadrado el valor z se obtiene el estadístico ji cuadrado ($0.70397^2 = 0.495$)

Solución con Excel

Para preparar la tabla de datos observados en Excel puede utilizarse una *tabla dinámica* o la opción CONTARSICONJUNTO.

Una vez se tienen las tablas observada y esperada, el estadístico χ^2 se obtiene aplicando la fórmula mediante operaciones matriciales, es decir, se escribe en una celda:

`(RANGO_OBSERVADO-RANGO_ESPERADO)^2/RANGO_ESPERADO`

y se pulsa la combinación Ctrl + Mayús + Intro: 0.495.

Solución con Stata

El comando `tabulate` con las opciones `expected` y `chi2` produce la misma salida que la del cuadro «Solución»:

```
tabulate THS Caso, expected chi2:  $\chi^2 = 0.495$ .
```

♣ Pregunta 9.8

Solución con calculadora

En la tabla de la ji cuadrado, con 1 g.l. el valor obtenido de χ^2 (0.495) se sitúa entre 0.45 y 0.71, lo que corresponde a $0.5 > p > 0.40$

Conclusión: El consumo de THS no se asocia al cáncer de mama.

Solución con Excel

En **Excel** se obtiene el valor exacto de p con la orden:

```
DISTR.CHICUAD.CD (0.495; 1)
```

, donde **0.495** es el valor del estadístico de contraste (es decir de χ^2 calculado arriba) y **1** son los grados de libertad (se obtienen multiplicado el número de filas -1 por el número de columnas -1 : $(f - 1) \times (c - 1)$).

Se obtiene un valor $p = 0.482$, por lo tanto, el consumo de THS no se asocia al cáncer de mama.

Solución con Stata

Con un solo comando podemos obtener el porcentaje de expuestas a THS en los dos grupos (con y sin cáncer de mama) y el estadístico de contraste χ^2

```
tab THS Caso, col chi
```

THS	Caso		Total
	0	1	
0	400 67.11	375 69.06	775 68.04
1	196 32.89	168 30.94	364 31.96
Total	596 100.00	543 100.00	1,139 100.00

Pearson $\chi^2(1) = 0.4952$ Pr = 0.482

SOLUCIÓN 9.3

♣ Pregunta 9.9

Solución con calculadora

Observe que la variable *AFamCancer* (antecedentes familiares) tiene 4 categorías (0-3) mientras que la variable cáncer de mama tiene 2 (0,1). Por tanto, las tablas de contingencia de este problema serán tablas de 4 filas ($f=4$) y 2 columnas ($c=2$).

Antecedentes familiares (AFamCancer)	Mujer con cáncer	Mujer sin cáncer	Total
Sin antecedentes (0)	413	573	986
1: Familiar 1er grado	99	60	159
2: Familiar 2o grado	90	58	148
3: Otro familiar	117	128	245
Total	719	819	1538

Tabla 9.5: Relación entre Antecedentes familiares y desarrollo de cáncer de mama (Tabla observada)

Antecedentes familiares (AFamCancer)	Mujer con cáncer	Mujer sin cáncer	Total
Sin antecedentes (0)	460.95	525.05	986
1: Familiar 1er grado	74.33	84.67	159
2: Familiar 2o grado	69.19	78.81	148
3: Otro familiar	114.54	130.46	245
Total	719	819	1538

Tabla 9.6: Relación entre Antecedentes familiares y desarrollo de cáncer de mama (Tabla esperada)

Solución con Excel

Para preparar la tabla de datos observados en Excel puede utilizarse una *tabla dinámica* o la opción CONTARSICONJUNTO.

Solución con Stata

Para obtener la tabla observada con los porcentajes de mujeres con antecedentes familiares de cáncer de mama en cada grupo, añadiremos la opción `col` al comando `tab`:

```
. tab AFamCancer Caso, col
```

AFamCancer	Caso		Total
	0	1	
0	573 69.96	413 57.44	986 64.11
1	60 7.33	99 13.77	159 10.34
2	58 7.08	90 12.52	148 9.62
3	128 15.63	117 16.27	245 15.93
Total	819 100.00	719 100.00	1,538 100.00

♣ Pregunta 9.10**Solución con calculadora**

Sí, en este estudio puede aplicarse la prueba χ^2 ya que todos los efectivos de la tabla esperada son mayores de 5.

♣ Pregunta 9.11**Solución con calculadora**

Para el cálculo de los g.l. se aplica la fórmula: $g.l. = (filas - 1) \times (columnas - 1)$
 En este caso: $g.l. = (4 - 1) \times (2 - 1) = 3$

♣ Pregunta 9.12

Solución con calculadora

$$\begin{aligned}
 \chi^2 &= \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \\
 &= \frac{(413 - 460.9)^2}{460.9} + \frac{(573 - 525.1)^2}{525.1} + \frac{(99 - 74.3)^2}{74.3} + \frac{(60 - 84.7)^2}{84.7} \\
 &\quad + \frac{(90 - 69.2)^2}{69.2} + \frac{(58 - 78.8)^2}{78.8} + \frac{(117 - 114.5)^2}{114.5} + \frac{(128 - 130.5)^2}{130.5} \\
 &= 36.595
 \end{aligned}$$

Solución con Stata

Para el cálculo del estadístico χ^2 y su valor p, añadiremos la opción `chi` al comando `tab` :

```
. tab AFamCancer Caso, chi
```

AFamCancer	Caso		Total
	0	1	
0	573 69.96	413 57.44	986 64.11
1	60 7.33	99 13.77	159 10.34
2	58 7.08	90 12.52	148 9.62
3	128 15.63	117 16.27	245 15.93
Total	819 100.00	719 100.00	1,538 100.00

Pearson $\chi^2(3) = 36.5951$ Pr = 0.000

♣ **Pregunta 9.13****Solución con calculadora**

En la tabla de la ji cuadrado, con 3 g.l. el valor obtenido de χ^2 (36.595) se sitúa por encima de 16.27, luego $p < 0.001$

Solución con Excel

En **Excel** se obtiene el valor exacto de p con la orden:

`DISTR.CHICUAD.CD (36.595;3)`

, siendo 36.595 el valor de la χ^2 y 3 los grados de libertad.

De esta forma obtenemos una $p = 1.131.04 \times 10^{-8}$.

Para la mayor parte de las aplicaciones de estadística en medicina, no necesitamos expresar la p con tanta precisión, podríamos poner simplemente $p < 0.001$.

♣ **Pregunta 9.14****Solución con calculadora**

Los antecedentes familiares de cáncer de mama se asocian al desarrollo de cáncer de mama. La tabla de valores observados muestra que las mujeres con cáncer de mama presentan un porcentaje claramente superior de antecedentes familiares de primer (13.8% vs 7.3%) y 2º grado (12.5% vs 7.1 %) que las mujeres sin cáncer.

SOLUCIÓN 9.4

Nota: este problema plantea valorar la asociación entre tener hijos y desarrollar cáncer de mama, utilizando una información más detallada de la variable hijos que la expuesta en el problema 8.3 del capítulo anterior. Para valorar la relación entre tener o no tener hijos (variable dicotómica) y tener o no cáncer de mama podemos aplicar indistintamente la prueba z o la ji cuadrado. Sin embargo, si queremos evaluar si tener un solo hijo o más cambia la asociación respecto a no tenerlos, la prueba z no podrá aplicarse porque una de las variables tiene más de dos categorías.

♣ Pregunta 9.15

Solución con calculadora

Tabla observada

	Mujer con cáncer	Mujer sin cáncer	Total
2 o más hijos	222	282	504
1 hijo	191	218	409
NO hijos	306	319	625
Total	719	819	1538

Tabla esperada al azar

	Mujer con cáncer	Mujer sin cáncer
2 o más hijos	235.62	268.38
1 hijo	191.20	217.80
NO hijos	292.18	332.82

$$E_{i,j} = \frac{n_{i.} \times n_{.j}}{n}$$

Cálculo de la ji cuadrado

$$\begin{aligned}
 \chi^2 &= \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \\
 &= \frac{(222 - 235.6)^2}{235.6} + \frac{(282 - 268.4)^2}{268.4} + \frac{(191 - 191.2)^2}{191.2} \\
 &\quad + \frac{(218 - 217.8)^2}{217.8} + \frac{(306 - 292.2)^2}{292.2} + \frac{(319 - 332.8)^2}{332.8} \\
 &= 2.705
 \end{aligned}$$

Solución con Excel

Para preparar la tabla de datos observados en **Excel** puede utilizarse una *tabla dinámica* o la opción **CONTARSICONJUNTO**.

Una vez se tienen las tablas observada y esperada, el estadístico χ^2 se obtiene aplicando la fórmula mediante operaciones matriciales, es decir, se escribe en una celda:

`(RANGO_OBSERVADO-RANGO_ESPERADO)²RANGO_ESPERADO`

y se pulsa la combinación **Ctrl + Mayús + Intro**: 2.705.

Solución con Stata

En Stata todos los cálculos anteriores se realizarían con la siguiente orden:

```
. tab Hijos_cat Caso , chi
```

Hijos_cat	Caso		Total
	0	1	
0	319	306	625
1	218	191	409
2	282	222	504
Total	819	719	1,538

Pearson chi2(2) = 2.7051 Pr = 0.259

♣ **Pregunta 9.16**

Solución con calculadora

En la tabla de la ji cuadrado, con 2 g.l. el valor obtenido de χ^2 (2.705) se sitúa entre 2.41 y 3.22, luego $0.30 > p > 0.20$

Solución con Excel

En **Excel** se obtiene el valor exacto de p con la orden:

`DISTR.CHICUAD.CD (2.705;2)`

siendo 2.705 el valor de la χ^2 y 2 los grados de libertad $((3-1)X(2-1))$.

Se obtiene un valor $p = 0.259$

♣ **Pregunta 9.17**

Solución con Excel

En **Excel** se obtiene con la orden:

`INVCHICUAD.CD (0.05;2)`

, donde 0.05 es el nivel de significación y 2 los g.l. de la tabla de contingencia.

La solución es: $\chi^2 = 5.991$

Solución con Stata

En **Stata** utilizamos la función `invchi2tail`:

`. di invchi2tail(2,0.05)`

El resultado obtenido es: 5.9914645

♣ Pregunta 9.18**Solución con calculadora**

En nuestro estudio no se ha encontrado una relación estadísticamente significativa entre el número de hijos (categorizado como sin hijos, 1 hijo o más hijos) y el cáncer de mama.

10

Cálculo del tamaño de muestra

SOLUCIÓN 10.1

♣ Pregunta 10.1

Solución con calculadora

Comparación de una proporción.

- $IC = 90\%; \text{Error } \alpha = 10\% \rightarrow z_{\alpha/2} = 1.64$
- $\text{Precision} = 2\% \rightarrow d = 0.02$
- $\text{Frecuencia esperada}(p) = 15\% \rightarrow p = 0.15$

$$n = \frac{(z_{\frac{\alpha}{2}} + z_{\beta})^2 p(1-p)}{d^2}$$

$$= \frac{(1.64 + 0.854)^2 \times 0.15 \times 0.85}{0.02^2} = 1976.3 \rightarrow \mathbf{1977 \text{ escolares}}$$

Solución con Excel

Error (b)	0.2	zb	0.85
Error (a)	0.10	z	1.64
Precisión (d)	0.02		
Frecuencia esperada (p)	0.15		
	Valor	Función	
Tamaño muestra (n)	1976.282	$((z+zb)^2 \cdot p \cdot (1-p)) / (d^2)$	

Solución con Stata

Si utilizamos Stata como calculadora deberemos introducir la siguiente operación:

```
. di (1.64+.85)^2*.15*.85/.02^2
```

Resultado: 1976.2819

A través, de comando `power oneproportion`. En la orden `power oneproportion`, el primer número que se pone NO es la proporción que dé la muestra (0.15) sino la hipótesis nula que, en este caso, podría ser 0.13 o 0.17.

```
. power oneproportion .13, diff(.02) alpha(.10) test(wald)
```

Performing iteration ...

Estimated sample size for a one-sample proportion test

Wald z test

Ho: $p = p_0$ versus Ha: $p \neq p_0$

Study parameters:

alpha = 0.1000

power = 0.8000

delta = 0.0200

p0 = 0.1300

pa = 0.1500

diff = 0.0200

Estimated sample size:

N = 1,971

```
. power oneproportion .17, diff(-.02) alpha(.10) test(wald)
```

Performing iteration ...

Estimated sample size for a one-sample proportion test

Wald z test

Ho: $p = p_0$ versus Ha: $p \neq p_0$

Study parameters:

alpha = 0.1000

power = 0.8000

delta = 0.0200

p0 = 0.1700

pa = 0.1500

diff = -0.0200

Estimated sample size:

N = 1,971

SOLUCIÓN 10.2

♣ Pregunta 10.2**Solución con calculadora****Comparación de una media.**

- $IC = 95\%; \text{Error } \alpha = 5\% \rightarrow z_{\alpha/2} = 1.96$
- *Diferencia mínima que se desea detectar* $Precisin(d) = 5$
- $\text{Varianza}(s^2) = 200s = 14.14$

$$n = \frac{\left(z_{\frac{\alpha}{2}} + z_{\beta}\right)^2 s^2}{d^2}$$

$$= \frac{(1.64 + 0.854)^2 \times 200}{5^2} = 63.17 \rightarrow \mathbf{64 \text{ pacientes}}$$

Solución con Excel

Error (b)	0.2	z_{β}	0.85
Error (a)	0.05	z	1.96
Precisión (d)	5		
Varianza (s2)	200.00		
	Valor	Función	
Tamaño muestra (n)	63.1688	$(D16+D15)^2*B18/B17^2$	

Solución con Stata

```
. power onemean 55 60, sd(14.14) (55 y 60 son dos edades arbitrarias con diferencia 5)
```

```
Performing iteration ...
```

```
Estimated sample size for a one-sample mean test  
t test
```

```
Ho:  $m = m_0$  versus Ha:  $m \neq m_0$ 
```

```
Study parameters:
```

```
alpha = 0.0500  
power = 0.8000  
delta = 0.3536  
m0 = 0.5500  
ma = 0.6000  
sd = 14.1400
```

```
Estimated sample size:  
N = 65
```

SOLUCIÓN 10.3

♣ Pregunta 10.3**Solución con calculadora****Comparación de dos proporciones**

- $IC = 95\%; Error_{\alpha} = 5\% \rightarrow z_{\alpha/2} = 1.96$
- $Potencia = 90\%; b = 10\% \rightarrow z_b = 1.28 <$
- $Diferencia\ minima\ que\ se\ desea\ detectar = 3\% \rightarrow Precision(d) = 0.03$
- $Frecuencia(p) = 70\% \rightarrow p = 0.7$

$$Varianza(s^2) = p(1 - p) = 0.7 \times 0.3 = 0.21$$

$$n = \frac{2(z_{\alpha/2} + z_{\beta})^2 s^2}{d^2}$$

$$= \frac{2(1.96 + 1.28)^2 \times 0.21}{0.03^2} = 4899 \text{ pacientes por grupo}$$

Solución con Excel

Error (a)	0.05	z_{α}	1.96
Potencia (b)	0.10	z_b	1.28
Precision (d)	0.03		
Frecuencia esperada (p)	0.7		
	Valor	Función	
Desviación estándar (s)	0.458258	RAIZ(p*(1-p))	
Tamaño muestra (n)	4898.88	$(2*(z_{\alpha}+z_b)^2*(s^2))/(d^2)$	

Solución con Stata

```
. power twoproportions .7 .73,beta(.1)
```

```
Performing iteration ...
```

```
Estimated sample sizes for a two-sample proportions test
```

```
Pearson's chi-squared test
```

```
Ho:  $p_2 = p_1$  versus Ha:  $p_2 \neq p_1$ 
```

```
Study parameters:
```

```
alpha = 0.0500
```

```
beta = 0.1000
```

```
delta = 0.0300 (difference)
```

```
p1 = 0.7000
```

```
p2 = 0.7300
```

```
Estimated sample size:
```

```
N = 9514
```

```
N per group = 4757
```

SOLUCIÓN 10.4

♣ Pregunta 10.4**Solución con calculadora****Comparación de dos medias.**

- $IC = 90\%; Error = 10\% \rightarrow z_{\alpha/2} = 1.64$
- $Potencia = 90\%; b = 10\% \rightarrow z_b = 1.28$
- $Diferencia\ minima\ que\ se\ desea\ detectar = 5\% \rightarrow Precision(d) = 0.05$
- $Desviación\ estandar\ esperada(s) = 10$

$$n = \frac{2(z_{\alpha/2} + z_b)^2 s^2}{d^2}$$

$$= \frac{2(1.64 + 1.28)^2 \times 10^2}{5^2} = 68.2112 \rightarrow \mathbf{69\ pacientes}$$

Solución con Excel

Error (a)	0.10	z _α	1.64
Potencia (b)	0.10	z _β	1.28
Precision (d)	5.00		
Desviación estandar (s)	10		
	Valor	Función	
Tamaño muestra (n)	68.2112	$(2*(z_{\alpha} + z_{\beta})^2 * (s^2)) / (d^2)$	

Solución con Stata

```
. power twomeans 15 20, sd(10) power(.9) alpha(.1)
```

Performing iteration ...

Estimated sample sizes for a two-sample means test

t test assuming $sd1 = sd2 = sd$

Ho: $m2 = m1$ versus Ha: $m2 \neq m1$

Study parameters:

alpha = 0.1000

power = 0.9000

delta = 5.0000

m1 = 15.0000

m2 = 20.0000

sd = 10.0000

Estimated sample size:

N = 140

N per group = 70

SOLUCIÓN 10.5

♣ Pregunta 10.5**Solución con calculadora****Potencia estadística.**

- A partir de la fórmula empleada para calcular el tamaño muestral, se despeja z_β :

$$n = \frac{2(z_{\frac{\alpha}{2}} + z_\beta)^2 s^2}{d^2} \Rightarrow z_\beta = \sqrt{\frac{nd^2}{2s^2}} - z_{\frac{\alpha}{2}}$$

- Y se sustituye por los datos del problema: $d=5$, $s=10$, $\alpha=0.05$, y $n=60$ (los pacientes que hubo realmente en cada grupo):

$$z_\beta = \sqrt{\frac{nd^2}{2s^2}} - z_{\alpha/2} = \sqrt{\frac{60 \times 5^2}{2 \times 10^2}} - 1.64 = 1.09$$

- Mirando este valor de z en la tabla normal se obtiene la probabilidad bajo la curva ($=1-p$):

$$\beta = 0.14 \rightarrow potencia = 0.86$$

Solución con Excel

Error (a)	0.10	za	1.64
Precision (d)	5		
Desviación estandar (s)	10		
Tamaño muestra (n)	60		
Potencia estadística	Valor	Función	
z_β	1.099	RAIZ((n*(d^2))/(2*(s^2)))-za	
β	0.864	DISTR.NORM.ESTAND.N(zb;VERDADERO)	

Solución con Stata

```
. power twomeans 15 20, sd(10) n(120) alpha(.1)
```

```
Estimated power for a two-sample means test  
t test assuming sd1 = sd2 = sd  
Ho: m2 = m1 versus Ha: m2 != m1
```

```
Study parameters:
```

```
alpha = 0.1000  
N = 120  
N per group = 60  
delta = 5.0000  
m1 = 15.0000  
m2 = 20.0000  
sd = 10.0000
```

```
Estimated power:  
power = 0.8595
```

SOLUCIÓN 10.6

♣ Pregunta 10.6**Solución con calculadora**

$$n = \frac{\left(z_{\frac{\alpha}{2}} + z_{\beta}\right)^2 p(1-p)}{d^2}$$

$$= \frac{(0.85 + 1.96)^2 \times 0.2 \times 0.8}{0.04^2} = 789.6 \rightarrow \mathbf{790} \text{ escolares}$$

Solución con Excel

Error (b)	0.2	z_{β}	0.85
Error (a)	0.05	z	1.96
Precisión (d)	0.04		
Frecuencia esperada (p)	0.2		
	Valor	Función	
Tamaño muestra (n)	789.61	$((z+z_{\beta})^2 \cdot p \cdot (1-p)) / (d^2)$	

Solución con Stata

```
. power oneproportion .16, diff(.04) alpha(.05) test(wald)
```

Estimated sample size for a one-sample proportion test

Wald z test

Ho: p = p0 versus Ha: p != p0

Study parameters:

alpha = 0.0500

power = 0.8000

delta = 0.0400

p0 = 0.1600

pa = 0.2000

diff = 0.0400

Estimated sample size:

N = 785

♣ Pregunta 10.7

Solución con calculadora

$$n = \frac{\left(z_{\frac{\alpha}{2}} + z_{\beta}\right)^2 p(1-p)}{d^2}$$

$$= \frac{(0.85 + 1.96)^2 \times 0.8 \times 0.2}{0.04^2} = 789.6 \rightarrow \mathbf{790} \text{ escolares}$$

Solución con Excel

Error (b)	0.2	z_{β}	0.85
Error (a)	0.05	z	1.96
Precisión (d)	0.04		
Frecuencia esperada (p)	0.8		
	Valor	Función	
Tamaño muestra (n)	789.61	$((z+z_b)^2 \cdot p \cdot (1-p)) / (d^2)$	

Solución con Stata

```
. power oneproportion .76, diff(.04) alpha(.05) test(wald)
```

Estimated sample size for a one-sample proportion test

Wald z test

Ho: $p = p_0$ versus Ha: $p \neq p_0$

Study parameters:

alpha = 0.0500

power = 0.8000

delta = 0.0400

p0 = 0.7600

pa = 0.8000

diff = 0.0400

Estimated sample size:

N = 785

♣ Pregunta 10.8

Solución con calculadora

$$n = \frac{\left(z_{\frac{\alpha}{2}} + z_{\beta}\right)^2 p(1-p)}{d^2}$$

$$= \frac{(0.85 + 1.96)^2 \times 0.5 \times 0.5}{0.04^2} = 1233.8 \rightarrow \mathbf{1234 \text{ escolares}}$$

Solución con Excel

Error (b)	0.2	z_{β}	0.85
Error (a)	0.05	z	1.96
Precisión (d)	0.04		
Frecuencia esperada (p)	0.5		
	Valor	Función	
Tamaño muestra (n)	1233.766	$((z+z_b)^2 \cdot p \cdot (1-p)) / (d^2)$	

Solución con Stata

```
. power oneproportion .46, diff(.04) alpha(.05) test(wald)
```

Estimated sample size for a one-sample proportion test
Wald z test

Ho: $p = p_0$ versus Ha: $p \neq p_0$

Study parameters:

```
alpha = 0.0500
power = 0.8000
delta = 0.0400
p0 = 0.4600
pa = 0.5000
diff = 0.0400
```

Estimated sample size:

```
N = 1,227
```

♣ Pregunta 10.9

Solución con calculadora

$$n = \frac{\left(z_{\frac{\alpha}{2}} + z_{\beta}\right)^2 p(1-p)}{d^2}$$

$$= \frac{(0.85 + 2.58)^2 \times 0.2 \times 0.8}{0.04^2} = 1233.8 \rightarrow \mathbf{1234 \text{ escolares}}$$

Solución con Excel

Error (b)	0.2	z_{β}	0.85
Error (a)	0.01	z	2.6
Precisión (d)	0.04		
Frecuencia esperada (p)	0.2		
	Valor	Función	
Tamaño muestra (n)	1190.25	$((z+z_b)^2 \cdot p \cdot (1-p)) / (d^2)$	

Solución con Stata

```
. power oneproportion .16, diff(.04) alpha(.01) test(wald)
```

Estimated sample size for a one-sample proportion test
Wald z test

Ho: $p = p_0$ versus Ha: $p \neq p_0$

Study parameters:

```
alpha = 0.1000
power = 0.8000
delta = 0.0400
p0 = 0.1600
pa = 0.2000
diff = 0.0400
```

Estimated sample size:

```
N = 1,168
```

♣ Pregunta 10.10

Solución con calculadora

$$n = \frac{\left(z_{\frac{\alpha}{2}} + z_{\beta}\right)^2 p(1-p)}{d^2}$$

$$= \frac{(0.85 + 1.96)^2 \times 0.2 \times 0.8}{0.02^2} = 3158.4 \rightarrow \mathbf{3159} \text{ escolares}$$

Solución con Excel

Error (b)	0.2	z_{β}	0.85
Error (a)	0.05	z	1.96
Precisión (d)	0.02		
Frecuencia esperada (p)	0.2		
	Valor	Función	
Tamaño muestra (n)	3158.44	$((z+z_b)^2 \cdot p \cdot (1-p)) / (d^2)$	

Solución con Stata

```
. power oneproportion .18, diff(.02) alpha(.05) test(wald)
```

Estimated sample size for a one-sample proportion test

Wald z test

Ho: $p = p_0$ versus Ha: $p \neq p_0$

Study parameters:

alpha = 0.0500

power = 0.8000

delta = 0.0200

p0 = 0.1800

pa = 0.2000

diff = 0.0200

Estimated sample size:

N = 3,140

♣ **Pregunta 10.11**

Solución con calculadora

Tamaño muestral (partos)	790	790	1234	1191	3159
Frecuencia esperada	0.2	0.8	0.5	0.2	0.2
Error α	5%	5%	5%	1%	5%
Precisión	4%	4%	4%	4%	2%

Conclusiones:

1. Con una la frecuencia esperada es 50% se requiere el mayor tamaño muestral porque es cuando $p(1-p)$ es máximo. Es lo que se conoce como “máxima indeterminación”.
2. Si se reduce el error alpha al 1 %, somos más precisos, entonces se incrementa el tamaño muestral, ya que entonces alpha medios es 0.005, es decir deja por debajo el 99.5% de la distribución frente al error alpha del 5% donde alpha medios es 0.025, es decir, deja por debajo el 97.5% de la distribución.
3. Reducir la precisión a la mitad supone multiplicar por 4 el tamaño muestral ($790partos \times 4 = 3160partos$)

11

Regresión y correlación

SOLUCIÓN 11.1

♣ Pregunta 11.1**Solución con calculadora**

Con los datos del enunciado podemos calcular directamente el coeficiente de correlación aplicando la fórmula:

$$r = \frac{\text{cov}_{xy}}{s_x \times s_y} = \frac{144.42}{12.09 \times 33.69} = 0.35$$

Interpretación: El coeficiente de correlación lineal vale 0.354, por tanto se trata de una relación lineal moderada positiva.

Solución con Excel

En **Excel** se obtiene con la orden `PEARSON(Pisa;pib)`. Tras el paréntesis se pone el nombre de las variables dependiente (Pisa) e independiente (pib) si se han creado etiquetas para definir los datos.

Si no se han creado etiquetas, se definirá la columna que contiene esos campos de valor para toda la población de estudio en la base de datos. En este caso utilizaríamos la orden `PEARSON(L1:L31;C1:C31)`, donde la variable Pisa ocupa las celdas L1 a L31.

La solución que se obtiene es $r = 0.354$

Solución con Stata

En **Stata** se utilizará **correlate**: `. correlate pisa pib`

(obs=30)

	pisa	pib
pisa	1.0000	
pib	0.3545	1.0000

♣ Pregunta 11.2

Solución con calculadora

Esta pregunta se responde calculando el coeficiente de determinación.

$$r^2 = 0.35^2 = 0.13$$

Interpretación: El 13 % de la puntuación obtenida en el PISA puede explicarse por el PIB del país.

Solución con Excel

En Excel se obtiene también con la orden: =COEFICIENTE.R2(Pisa;pib)

Solución con Stata

Ver solución en la pregunta 11.4.

♣ Pregunta 11.3

Solución con calculadora

Esta pregunta se responde calculando la pendiente de la recta.

$$\widehat{\beta}_1 = r \times \frac{s_y}{s_x} = 0.35 \times \frac{33.69}{12.09} = 0.988$$

Interpretación: Por cada 1000 dólares de aumento del PIB aumenta 0.988 puntos la nota del PISA.

Solución con Excel

En Excel podemos utilizar las siguientes orden: =PENDIENTE(Pisa;pib)

Solución con Stata

Ver solución en la pregunta 11.14.

♣ **Pregunta 11.4****Solución con calculadora**

Para responder a esta pregunta calcularemos la intersección de la recta en el origen ($x=0$).

$$\widehat{\beta}_0 = m_{\text{pisa}} - \widehat{\beta}_1 \times m_{\text{pib}} = 439.12 - 0.988 \times 20.72 = \mathbf{463.77}$$

Solución con Excel

En **Excel** podemos utilizar las siguientes orden:
`=INTERSECCION.EJE(Pisa;pib)` `=INTERSECCION.EJE(Pisa;pib)`

Para obtener todos los parámetros anteriores (ejercicios 11.1 a 11.4) de una sola vez podemos utilizar el **analizador de datos de Excel** y seleccionar **Regresión**.

Con esta opción hay que introducir las etiquetas de las variables Y(Pisa) y X(pib).

Estadísticas de la regresión	
Coefficiente de correlación múltiple	0.35445414
Coefficiente de determinación R ²	0.12563773
R ² ajustado	0.09441051
Error típico	32.0615206
Observaciones	30

Coefficiente de correlación múltiple	0.35445414
Coefficiente de determinación R ²	0.12563773
R ² ajustado	0.09441051
Error típico	32.0615206
Observaciones	30

Análisis de varianza					
	Grados libertad	Suma cuadrados	Promedio cuadrados	F	Valor crítico F
Regresión	1	4135.756	4135.756	4.023	0.055
Residuos	28	28782.351	1027.941		
Total	29	32918.107			

	Grados libertad	Suma cuadrados	Promedio cuadrados	F	Valor crítico F
Regresión	1	4135.756	4135.756	4.023	0.055
Residuos	28	28782.351	1027.941		
Total	29	32918.107			

	Coef.	Error típico	Estadístico t	Probabilidad	Inferior 95 %	Superior 95 %
Intercepción	463.776	15.758	29.431	1.3×10^{-22}	431.5	496.055
PIB	0.988	0.492	2.006	0.0546	-0.021	1.996

Solución con Stata

Solución con **Stata** para los ejercicios 11.1 a 11.4.

En **Stata** con el comando **regress** obtendremos el análisis de regresión completo:

```
. regress pisa pib
```

				Number of obs	= 30
				F(1, 28)	= 4.02
Source	SS	df	MS	Prob>F	= 0.0546
Model	4135.75717	1	4135.75717	R-squared	= 0.1256
Residual	28782.3509	28	1027.9411	Adj R-squared	= 0.0944
Total	32918.1081	29	1135.10718	Root MSE	= 32.062

pisa	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
pib	.9875088	.4923201	2.01	0.055	-.0209632 1.995981
_cons	463.7763	15.75789	29.43	0.000	431.4977 496.0549

♣ Pregunta 11.5

Solución con calculadora

- Relación PIB-PISA matemáticas:

$$r = \frac{\text{cov}_{xy}}{s_x \times s_y} = \frac{150.6}{12.09 \times 38.6} = 0.32$$

$$\widehat{\beta}_1 = r \times \frac{s_y}{s_x} = 0.32 \times \frac{38.6}{12.09} = 1.03$$

$$\widehat{\beta}_0 = m_{\text{pisamate}} - \widehat{\beta}_1 \times m_{\text{pib}} = 492.4 - 1.03 \times 20.72 = 463.77$$

- Relación PIB-PISA ciencias:

$$r = \frac{\text{cov}_{xy}}{s_x \times s_y} = \frac{139}{12.09 \times 34.4} = 0.334$$

$$\widehat{\beta}_1 = r \times \frac{s_y}{s_x} = 0.33 \times \frac{34.4}{12.09} = 0.95$$

$$\widehat{\beta}_0 = m_{\text{pisaciencias}} - \widehat{\beta}_1 \times m_{\text{pib}} = 498.1 - 0.95 \times 20.72 = 469.86$$

- Relación PIB-PISA lectura:

$$r = \frac{\text{cov}_{xy}}{s_x \times s_y} = \frac{143.62}{12.09 \times 29.8} = 0.40$$

$$\widehat{\beta}_1 = r \times \frac{s_y}{s_x} = 0.40 \times \frac{29.8}{12.09} = 0.98$$

$$\widehat{\beta}_0 = m_{\text{pisalectura}} - \widehat{\beta}_1 \times m_{\text{pib}} = 488.9 - 0.98 \times 20.72 = 459.68$$

Solución con Stata

■ Relación PIB-PISA matemáticas: `. regress pisamate pib`

Source	SS	df	MS	Number of obs		
				F(1, 28)	= 3.26	
Model	4499.1172	1	4499.1172	Prob>F	= 0.0816	
Residual	38614.0828	28	1379.07439	R-squared	= 0.1044	
Total	43113.2	29	1486.66207	Adj R-squared	= 0.0724	
				Root MSE	= 37.136	

pisamate	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
pib	1.029976	.5702395	1.81	0.082	-.1381067	2.198059
_cons	461.7921	18.25189	25.30	0.000	424.4048	499.1794

■ Relación PIB-PISA ciencias: `. regress pisaciencia pib`

Source	SS	df	MS	Number of obs		
				F(1, 28)	= 3.52	
Model	3831.20888	1	3831.20888	Prob>F	= 0.0712	
Residual	30509.4911	28	1089.62468	R-squared	= 0.1116	
Total	34340.7	29	1184.16207	Adj R-squared	= 0.0798	
				Root MSE	= 33.009	

pisacienc.	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
pib	.9504546	.5068762	1.88	0.071	-.0878343	1.988743
_cons	469.8552	16.22379	28.96	0.000	436.6223	503.0882

■ Relación PIB-PISA lectura: `. regress pisalectura pib`

Source	SS	df	MS	Number of obs		
				F(1, 28)	= 5.30	
Model	4090.53912	1	4090.53912	Prob>F	= 0.0290	
Residual	21614.9275	28	771.961698	R-squared	= 0.1591	
Total	25705.4667	29	886.395402	Adj R-squared	= 0.1291	
				Root MSE	= 27.784	

pisalect.	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
pib	.9820955	.4266398	2.30	0.029	.1081636	1.856027
_cons	459.6816	13.65563	33.66	0.000	431.7093	487.6539

SOLUCIÓN 11.2

♣ Pregunta 11.6

Solución con calculadora

Aplicando la fórmula del coeficiente de correlación lineal obtendremos:

$$r = \frac{\text{cov}_{xy}}{s_x \times s_y} = \frac{1278.82}{34.4 \times 38.6} = 0.96$$

Interpretación: Ambas variables (resultados pisa en ciencias y en matemáticas) están fuertemente correlacionadas con un coeficiente de correlación ($r=0.96$) próximo a 1 (relación lineal perfecta positiva).

Solución con Excel

En **Excel** se obtiene con la orden: `PEARSON(pisamate;pisciencia)`

La solución que se obtiene es $r = 0.9638$

Solución con Stata

En **Stata** con el comando `correlate` obtendremos los coeficientes de correlación entre las variables: `. correlate pisaciencia pisamate`

(obs=30)

	pisaciencia	pisamate
pisaciencia	1.0000	
pisamate	0.9638	1.0000

♣ Pregunta 11.7

Solución con calculadora

Aplicando la fórmula del coeficiente de correlación lineal obtendremos:

$$r = \frac{\text{cov}_{xy}}{s_x \times s_y} = \frac{1075.2}{38.6 \times 29.8} = 0.94$$

Interpretación: Ambas variables están fuertemente correlacionadas con un coeficiente de correlación ($r=0.94$) próximo a 1 (relación lineal perfecta positiva).

Solución con Excel

En Excel se obtiene con la orden `PEARSON(pisamate;pisalectura)`

La solución que se obtiene es $r = 0.9367$

Solución con Stata

En Stata:

```
. correlate pisamate pisalectura
```

(obs=30)

	pisamate	pisalectura
pisamate	1.0000	
pisalectura	0.9367	1.0000

♣ Pregunta 11.8

Solución con calculadora

Aplicando la fórmula del coeficiente de correlación lineal obtendremos:

$$r = \frac{\text{cov}_{xy}}{s_x \times s_y} = \frac{975.32}{34.4 \times 29.8} = 0.95$$

Interpretación: Ambas variables están fuertemente correlacionadas con un coeficiente de correlación ($r=0.95$) próximo a 1 (relación lineal perfecta positiva).

Solución con Excel

En Excel se obtiene con la orden `PEARSON(pisalectura;pisaciencia)`

La solución que se obtiene es $r = 0.9520$

Solución con Stata

En Stata:

```
. correlate pisaciencia pisalectura
```

(obs=30)

	pisaciencia	pisalectura
pisaciencia	1.0000	
pisalectura	0.9520	1.0000

Estos tres resultados se pueden obtener con una sola orden:

```
. correlate pisaciencia pisamate pisalectura
```

(obs=30)

	pisaciencia	pisamate	pisalectura
pisaciencia	1.0000		
pisamate	0.9638	1.0000	
pisalectura	0.9520	0.9367	1.0000

SOLUCIÓN 11.3

♣ Pregunta 11.9**Solución con calculadora**

El coeficiente de correlación de Pearson indica la fuerza de la relación lineal entre dos variables continuas. Podemos calcularlo con los datos del enunciado aplicando la fórmula:

$$r = \frac{\text{cov}_{xy}}{s_x \times s_y} = \frac{-2.23}{12.09 \times 7.86} = -0.023$$

Interpretación: Se trata de una asociación débil ($|R| \approx 0$) e inversa ($R < 0$).

Solución con Excel

En Excel: `PEARSON(consumoazucar;pib)`

Solución con Stata

En Stata: `. correlate consumoazucar pib`

(obs=30)

	consumoazucar	pib
consumoazucar	1.0000	
pib	-0.0234	1.0000

♣ Pregunta 11.10**Solución con Stata**

Primero, ajustamos un modelo de regresión lineal simple en que el consumo de azúcares refinados depende del PIB del país:

`. regress consumoazucar pib`

Source	SS	df	MS	Number of obs	= 30
Model	.98213912	1	.98213912	F(1, 28)	= 0.02
Residual	1789.95453	28	63.9269475	Prob>F	= 0.9022
Total	1790.93667	29	61.7564369	R-squared	= 0.0005
				Adj R-squared	= -0.0351
				Root MSE	= 7.9954

consu r.	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
pib	-.0152177	.1227737	-0.12	0.902	-.2667083 .2362728
_cons	6.273227	3.929669	1.60	0.122	-1.776335 14.32279

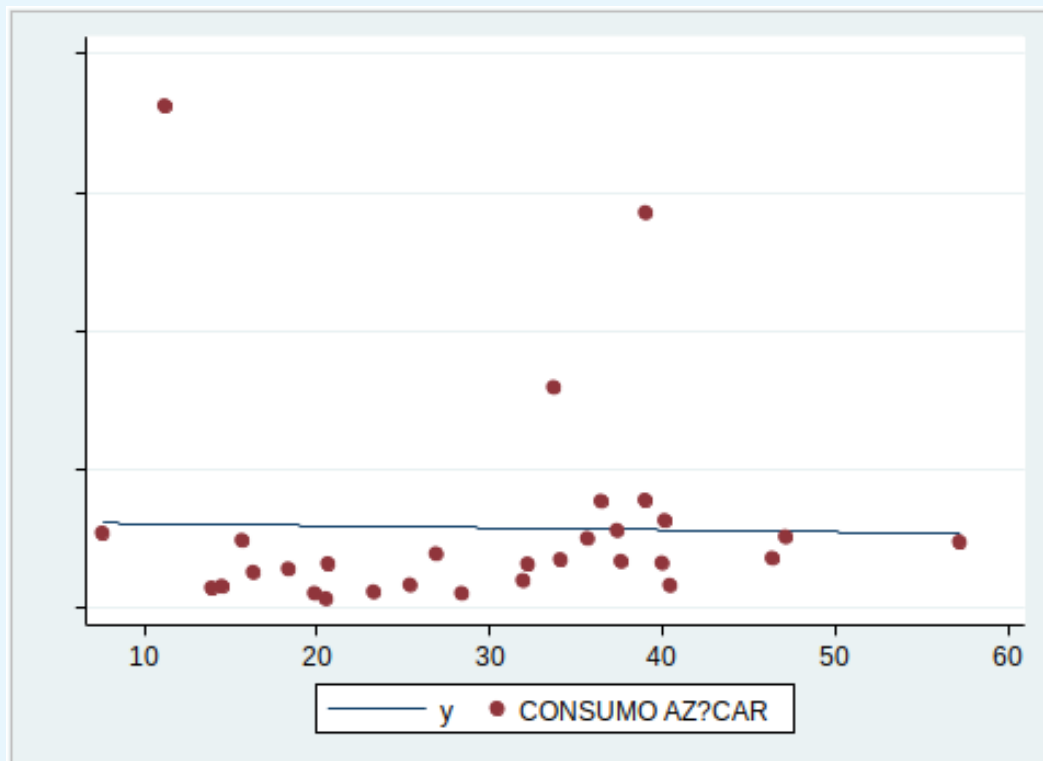
Por lo tanto, la ecuación del modelo es:

$$\text{Consumo de azcar} = 6.2732 - 0.0152 \times \text{pib}$$

Stata, después de ajustar un modelo de regresión, permite acceder a los coeficientes ajustados mediante `_b[nombre de la variable]`. Así,

```
graph twoway function y = _b[_cons] + _b[pib] * x, range(pib) || scatter
consumoazucar pib
```

, produce la siguiente figura



♣ Pregunta 11.11

Solución con calculadora

El coeficiente de determinación, R^2 , da cuenta de la variación en consumo de azúcar que es justificable por la variable que representa el PIB del país.

Se obtiene elevando al cuadrado el coeficiente de correlación lineal:

$$R^2 = (-0.023)^2 = 0.0005$$

Interpretación: El 0.055 % de la variabilidad del consumo de azúcar se puede explicar por el PIB del país.

Solución con Excel

En Excel, `COEFICIENTE.R2(consumoazucar; pib)`

Se obtiene $R^2 = 0.00054839$

Solución con Stata

En Stata: `. correlate consumoazucar pib`

(obs=30)		
	consumoazucar	pib
consumoazucar	1.0000	
pib	-0.0234	1.0000

Si utilizamos Stata como una calculadora `. di r(C)[2,1]^2`

Resultado: .00054839

♣ Pregunta 11.12

Solución con calculadora

Esta pregunta se responde calculando la pendiente de la recta.

$$\widehat{\beta}_1 = -0.023 \times \frac{7.86}{12.09} = -0.015$$

Solución con Excel

En Excel se obtiene con la orden: `PENDIENTE(consumoazucar;pib)`

Solución con Stata

La orden y la salida en Stata aparece en los resultados de la pregunta 11.10.

SOLUCIÓN 11.4

Solución con Stata

Al haber 30 países en la base de datos, los 15 con menor puntuación PISA en lectura pueden identificarse mediante la mediana (percentil 50) de puntuación PISA en lectura.

En **Stata**,

```
. tabstat pisalectura, stat(median)
```

variable	p50
pisalectura	494.5

Es decir, 494.5 es la puntuación PISA en lectura por debajo de la cual se sitúan los 15 países a que hace referencia el enunciado del problema.

♣ Pregunta 11.13

Solución con calculadora

El porcentaje de variabilidad explicada por la variable independiente corresponde al coeficiente de determinación. Para responder esta pregunta calcularemos primero r y luego lo elevaremos al cuadrado:

$$r = \frac{\text{cov}_{xy}}{s_x \times s_y} = \frac{50.74}{2.30 \times 30.18} = 0.732$$

Siendo $R^2 = 0.536$

El coeficiente de determinación, R^2 , da cuenta de la variación en número de premios Nobel que es justificable por la variable que representa el consumo de chocolate.

Solución con Excel

En **Excel**, habrá que seleccionar sólo los países que cumplan la condición de tener una puntuación en pisalectura < 494.5 , para ello se puede utilizar un filtro (datos $\rightarrow$ filtro) y crear las etiquetas de las variables de los 15 países seleccionados, que hemos llamado **nobel15** y **chocolate15** respectivamente.

La función a ejecutar es: `COEFICIENTE.R2(nobel15;chocolate15)`

Solución con Stata

En **Stata**, con la orden `if`, seleccionamos los registros con los que vamos a trabajar, en este caso `if pisalectura<494.5`

```
. correlate nobel consumochocolate if pisalectura < 494.5
```

(obs=15)

	nobel	consum e
nobel	1.0000	
consumocho e	0.7318	1.0000

```
. display r(C)[2,1]^2
```

Resultado: .53558471

Es decir, el 53.56% de la variabilidad en el número de premios Nobel se puede explicar por el consumo de chocolate en los países con puntuación Pisa en lectura inferior a la mediana.

♣ Pregunta 11.14

Solución con Excel

Primero, calculamos el consumo medio de chocolate de los 15 países con mayor puntuación PISA en lectura.

En **Excel**, filtrando puntuación superior a la mediana `=PROMEDIO(chocolate)`

Solución con Stata

En **Stata**, con la orden `if ( if pisalectura>494.5 )`.

```
. tabstat consumochocolate if pisalectura>494.5, stat(mean)
```

variable	mean
consumocho e	5.622667

Solución con calculadora

Es decir, 5.6227 kg. Ahora, ajustamos un modelo de regresión lineal simple en que el número de premios Nobel depende del consumo de chocolate:

$$\widehat{\beta}_1 = 0.73 \times \frac{30.18}{2.30} = 9.61$$

$$\widehat{\beta}_0 = 12.6 - 9.61 \times 3.65 = -22.48$$

Con estos datos podemos construir la ecuación del modelo

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x = -22.48 + 9.61x$$

Y, substituyendo x por el valor del consumo medio de chocolate (5.63), obtendremos:

$$\widehat{y} = -22.48 + 9.61 \times 5.63 = 31.56$$

Por tanto, para un país con un consumo medio de chocolate de 5.63, esperaremos una media de **31.56 premios nobel**

Solución con Excel

En Excel:

1º Calculamos la pendiente:

=PENDIENTE(nobel15;chocolate15)

Resultado: $\widehat{\beta}_1 = 9.6118$

2º Calculamos la intersección en el origen:

=INTERSECCION.EJE(nobel15;chocolate15)

Resultado: $\widehat{\beta}_0 = -22.4829$

La ecuación del modelo es

$$\widehat{y} = -22.4829 + 9.6118x$$

Solución con Stata

En Stata

```
. regress nobel consumochocolate if pisalectura < 494.5
```

Source	SS	df	MS	Number of obs	= 15
Model	6827.41961	1	6827.41961	F(1, 13)	= 14.99
Residual	5920.18039	13	455.398491	Prob>F	= 0.0019
Total	12747.6	14	910.542857	R-squared	= 0.5356
				Adj R-squared	= 0.4999
				Root MSE	= 21.34

nobel	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
consumochocolate	9.611752	2.48239	3.87	0.002	4.248874 14.97463
_cons	-22.48289	10.60456	-2.12	0.054	-45.39265 .4268575

La ecuación del modelo es

$$\widehat{y} = -22.4829 + 9.6118x,$$

por lo que, si $x = 5.6227 \Rightarrow \widehat{y} = 31.5614$ premios Nobel, en contraposición a los 12.6 que tienen de media los países con menor puntuación PISA en lectura.

```
. summarize nobel if pisalectura < 494.5
```

Variable	Obs	Mean	Std. Dev.	Min	Max
nobel	15	12.6	30.1752	0	119

♣ Pregunta 11.15

Solución con calculadora

Por cada incremento de 1kg en el consumo de chocolate, se espera encontrar 9.6118 premios Nobeles más. Si el país decide incrementar el consumo en 125 g, es decir, $1/8$ kg, se esperaría encontrar $9.6118/8 = 1.2015$ premios Nobel más.

SOLUCIÓN 11.5

♣ **Pregunta 11.16****Solución con Stata**

Primero, hallamos el consumo de chocolate y el número de premios Nobel españoles:

En **Stata**,

```
. list consumochocolate nobel if pais == "ESPAÑA"
```

	consumo	chocolate	nobel
17.	3		8

Solución con calculadora

Después, ajustamos un modelo de regresión lineal simple en que el número de premios Nobel dependa del consumo de chocolate, siguiendo los siguientes pasos: Primero, calculamos r :

$$r = \frac{\text{cov}_{xy}}{s_x \times s_y} = \frac{39.32}{2.82 \times 64.99} = 0.215$$

Segundo, calculamos los coeficientes:

$$\widehat{\beta}_1 = 0.215 \times \frac{64.99}{2.82} = 4.95$$

$$\widehat{\beta}_0 = 28.23 - 4.64 \times 4.95 = 5.27$$

Tercero, con estos datos ya tenemos la ecuación del modelo:

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x = 5.27 + 4.95x$$

Solución con Excel

En **Excel** se obtiene con las órdenes: `PENDIENTE(nobel;consumochocolate)` y `INTERSECCION.EJE(nobel;consumochocolate)` o directamente con el analizador de datos **Regresión**.

Solución con Stata

En Stata: `. regress nobel consumochocolate`

Source	SS	df	MS	Number of obs	= 30
Model	5646.46238	1	5646.4624	F(1, 28)	= 1.35
Residual	116828.904	28	4172.4609	Prob>F	= 0.2545
Total	12747.6	14	910.54286	R-squared	= 0.00461
				Adj R-squared	= 0.0120
				Root MSE	= 64.595

nobel	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
consumochocolate	4.951883	4.256752	1.16	0.255	-3.767677 13.67144
_cons	5.274752	22.99088	0.23	0.820	-41.81993 52.36944

La ecuación del modelo es

$$y = 5.2748 + 4.9519x ,$$

por lo que, si $x = 3 \Rightarrow y = 20.13$ premios Nobel, en contraposición a los 8 de España.

♣ Pregunta 11.17

Solución con calculadora

El error o residuo que se comete es, precisamente, la diferencia entre la predicción y el dato real, es decir, $20.13 - 8 = 12.13$ premios Nobel menos de los que nos corresponde por nuestro consumo de chocolate.

♣ Pregunta 11.18

Solución con calculadora

Algebraicamente, la diferencia en número de premios Nobel se corresponde con el coeficiente del consumo de chocolate multiplicado por la diferencia en consumo anual de chocolate, esto es, $4.9519 \times 3 = 14.8557$ premios Nobel más.

12

Regresión lineal múltiple

SOLUCIÓN 12.1

♣ Pregunta 12.1**Solución con calculadora**

Cada intervalo de confianza se obtiene con la fórmula:

$$\widehat{\beta}_i \mp 1.96 \times EE(\widehat{\beta}_i)$$

, por ejemplo, para hidratos de carbono:

$$1.10 \mp 1.96 \times 0.31 = 0.49, 1.71$$

El valor de p se obtiene con el test de Wald. Como el número de grados de libertad es muy alto ($n - 1 = 1241$), el resultado obtenido sirve para buscar la p en la tabla normal. Aquí se han utilizado siempre valores p de dos colas.

Por ejemplo, para hidratos de carbono:

$$t = \frac{\widehat{\beta}_i}{EE(\widehat{\beta}_i)} = \frac{1.10}{0.31} = 3.55$$

Consultando este valor en la tabla Normal, se observa que deja por debajo el 0.9998. Por lo tanto, lo que queda por encima es 0.0002. Como se trata de obtener un valor de dos colas, hay que multiplicar este resultado por 2: 0.0004

Nutriente	$\widehat{\beta}$	IC 95 %	p
Hidratos de carbono	1.10	0.49, 1.71	0.0004
Ácidos grasos saturados	1.54	1.09, 1.99	<0.001
Ácidos grasos poliinsaturados	-0.83	-1.67, 0.01	0.054
Ácidos grasos monoinsaturados	-2.20	-3.08, -1.32	<0.001
Constante	12.50	9.42, 15.58	<0.001

♣ Pregunta 12.2

Solución con calculadora

La respuesta es mg/dL/(g/día). Para comprenderlo hay que tener en cuenta:

1. Las unidades de la variable Y son mg/dL
2. Las unidades de ácidos grasos saturados son g/día
3. El resultado de multiplicar el coeficiente por los ácidos grasos tiene que ser mg/dL. Es decir: *unidades de LDL = unidades de $\beta \times$ unidades de ácidos grasos saturados*

4. Por lo tanto:

$$\frac{\text{mg}}{\text{dL}} = \text{unidades de } \beta \times \frac{\text{g}}{\text{da}}$$

5. Despejando:

$$\text{unidades de } \beta = \frac{\text{mg}}{\text{dL}} / \left(\frac{\text{g}}{\text{da}} \right)$$

♣ Pregunta 12.3

Solución con calculadora

El modelo que sale de la tabla del enunciado es:

$$\begin{aligned} LDL &= 12.50 + 1.10 \times \text{hidratos de carbono} \\ &\quad + 1.54 \times \text{ácidos g. saturados} \\ &\quad - 0.83 \times \text{ácidos g. poliinsaturados} \\ &\quad - 2.20 \times \text{ácidos g. monoinsaturados} \end{aligned}$$

Aplicado a este paciente sería:

$$LDL = 12.50 + 1.10 \times 180 + 1.54 \times 40 - 0.83 \times 20 - 2.20 \times 18 = 215.9$$

♣ Pregunta 12.4

Solución con calculadora

En la tabla del enunciado, el coeficiente para ácidos grasos monoinsaturados es -2.20. Si el paciente A añade 20 g, tendrá $2.20 \times 20 = 44.00 \frac{\text{mg}}{\text{dL}}$ menos de LDL.

Pero además ha suprimido 20 g de ácidos grasos saturados. A cada gramo que quite le corresponde 1.54 -según indica la tabla del enunciado. Por lo tanto, lo que pierde de LDL por bajar 20 g los ácidos grasos saturados es: $1.54 \times 20 = 30.80 \frac{\text{mg}}{\text{dL}}$

En total, habrá perdido $44.00 + 30.80 = 74.80 \text{ mg/dL}$.

SOLUCIÓN 12.2

♣ Pregunta 12.5
Solución con calculadora

Hay que introducir cuatro variables: una para la edad y tres para el nivel de estudios. La regla para una variable categórica es que hay que incluir tantas variables ficticias como categorías - 1. Por ejemplo, para estudiar el sexo solo hace falta una variable (2 categorías - 1). Para el nivel de estudios tenemos cuatro categorías; por lo tanto, hacen falta tres variables.

♣ Pregunta 12.6
Solución con calculadora

	Variables ficticias		
Nivel de estudios	estudio_1	estudio_2	estudio_3
Sin estudios	1	0	0
Primarios	0	0	0
Secundarios	0	1	0
Universitarios	0	0	1

La categoría de referencia (estudios primarios) se pone a 0 en las tres variables ficticias. Cada una de las otras categorías debe tener un 1 en una variable y 0 en el resto. De esta forma se identifica bien el nivel de estudios de cada paciente: si tiene 1 en estudio_1 es una persona sin estudios; si tiene 1 en estudio_2, tiene estudios secundarios. Si tiene 1 en estudio_3, tiene estudios universitarios. Y si tiene 0 en todas, tiene estudios primarios.

♣ Pregunta 12.7

Solución con calculadora

La figura tendrá cuatro rectas paralelas, una para cada nivel de estudios.

Si solo se tuviera en cuenta la edad, la figura sería una recta. Al añadir las tres variables ficticias, se añaden tres rectas más. Son todas paralelas porque el coeficiente de cada variable ficticia lo único que hace es añadir una cantidad fija (es decir, desplazar la recta hacia arriba o hacia abajo).

♣ Pregunta 12.8

Solución con calculadora

Las unidades son $\frac{\text{kg}}{\text{m}^2 \times \text{años}}$

Hay dos formas de comprenderlo:

1. La interpretación del coeficiente $\widehat{\beta}$ es en cuánto aumenta Y (índice de masa corporal) por cada unidad que aumente X (edad). Por lo tanto, sus unidades son las del índice de masa corporal divididas entre las unidades de edad.
2. El resultado final de la regresión es el índice de masa corporal medido en $\frac{\text{kg}}{\text{m}^2}$. Como el $\widehat{\beta}$ de la edad va multiplicado por el número de años, necesitamos que las unidades de b multiplicadas por el número de años nos den como resultado $\frac{\text{kg}}{\text{m}^2}$.

♣ Pregunta 12.9

Solución con calculadora

Las unidades son $\frac{\text{kg}}{\text{m}^2}$. El $\widehat{\beta}$ de estudio_2 va multiplicado por un 1 o por un 0, pero estudio_2 no tiene unidades. Por lo tanto, las unidades de $\widehat{\beta}$ tienen que ser las mismas que las del índice de masa corporal.

SOLUCIÓN 12.3

♣ Pregunta 12.10

Solución con calculadora

Variable	$\widehat{\beta}$	IC 95 %	p
Edad	0.21	0.05, 0.37	0.009
estudios_1	1.37	0.57, 2.17	<0.001
estudios_2	-0.22	-0.96, 0.52	0.563
estudios_3	-0.93	-1.71, -0.15	0.020
Constante	15.21	14.62, 15.80	<0.001

En cada fila, el intervalo de confianza se obtiene como $\widehat{\beta} \pm 1.96 \times EE(\widehat{\beta})$ y el valor p se obtiene buscando en la tabla normal el valor $z = \left| \frac{\widehat{\beta}}{EE(\widehat{\beta})} \right|$; el valor p obtenido se debe multiplicar por dos para tener las dos colas.

Es importante tener presente el valor absoluto en el cálculo de z.

♣ Pregunta 12.11

Solución con calculadora

El coeficiente de edad (0.21) indica que por cada año más que se cumple, el índice de masa corporal aumenta en 0.21 kg/m^2 . El intervalo de confianza indica que puede haber un error al azar que haga que este aumento sea tan bajo como 0.05 kg/m^2 o tan alto como 0.37 kg/m^2 .

♣ Pregunta 12.12

Solución con calculadora

Cada coeficiente del nivel de estudios hay que interpretarlo respecto al nivel de estudios primarios (que se puso como categoría de referencia). Por lo tanto:

- Las personas sin estudios (variable estudios_1) tienen de media 1.37 kg/m^2 más que los que tienen estudios primarios. Al azar, este valor podría cambiar entre 0.57 kg/m^2 y 2.17 kg/m^2 .
- Las personas con estudios secundarios (variable estudios_2) tiene un índice de masa corporal muy parecido al de quienes tiene estudios primarios: 0.22 kg/m^2 menos. Al azar, el resultado podría estar entre que su índice de masa corporal es menor en 0.97 kg/m^2 o mayor que el de estudios primarios en 0.52 kg/m^2 .
- Las personas con estudios universitarios (variable estudios_3) tienen menor índice de masa corporal que las que tienen estudios primarios: 0.93 kg/m^2 menos. La variación al azar está siempre en el lado de menor índice: puede estar entre 1.71 kg/m^2 menos y 0.15 kg/m^2 menos que los de estudios primarios.

♣ Pregunta 12.13

Solución con calculadora

El modelo que aparece en la tabla de enunciado se puede expresar con esta fórmula:

$$IMC = 15.21 + 0.21 \times edad + 1.37 \times estudios_1 - 0.22 \times estudios_2 - 0.93 \times estudios_3$$

Aplicado a la persona **A**, sabemos que $edad = 33$, $estudios_1=0$, $estudios_2=0$ y $estudios_3=1$. Por lo tanto:

$$IMC = 15.21 + 0.21 \times 33 + 1.37 \times 0 - 0.22 \times 0 - 0.93 \times 1 = 21.21 \text{ kg/m}^2$$

La persona **B** tiene 45 años y, como sus estudios son primarios, $estudios_1 = estudios_2 = estudios_3 = 0$. Por lo tanto:

$$IMC = 15.21 + 0.21 \times 45 + 1.37 \times 0 - 0.22 \times 0 - 0.93 \times 0 = 24.66 \text{ kg/m}^2$$

Para **C**, tenemos 28 años, $estudio_1=1$, $estudio_2=0$ y $estudio_3=0$. Su IMC esperado será:

$$IMC = 15.21 + 0.21 \times 28 + 1.37 \times 1 - 0.22 \times 0 - 0.93 \times 0 = 22.46 \text{ kg/m}^2$$

Por último, **D** tiene 70 años, $estudio_1=0$, $estudio_2=1$ y $estudio_3=0$. Esperamos que su IMC sea:

$$IMC = 15.21 + 0.21 \times 70 + 1.37 \times 0 - 0.22 \times 1 - 0.93 \times 0 = 29.69 \text{ kg/m}^2$$

SOLUCIÓN 12.4

♣ Pregunta 12.14

Solución con calculadora

Las unidades de cada coeficiente son el resultado de dividir las unidades de la variable Y (mmHg) entre las unidades de la variable X. Por lo tanto:

Variable	Unidades de $\widehat{\beta}$
Edad	$\frac{\text{mmHg}}{\text{años}}$
Consumo de sal	$\frac{\text{mmHg}}{\text{gramos}}$
Edad x consumo de sal	$\frac{\text{mmHg}}{\text{años} \times \text{gramos}}$
Constante	mmHg

♣ Pregunta 12.15

Solución con calculadora

El modelo que indica el enunciado corresponde a esta fórmula:

$$TAS = 5.0 + 2.5 \times edad + 8.0 \times sal - 0.10 \times edad \times sal$$

Aplicaremos la fórmula a las preguntas 12.15 y 12.16.

Para la pregunta 12.15, encontraremos la solución de dos formas. Para la 12.16 nos limitaremos a la segunda forma.

- Primera forma: El paciente 12.15 tiene 40 años y consumo 7 gramos. Su nivel esperado de TAS es:

$$TAS = 5.0 + 2.5 \times 40 + 8.0 \times 7 - 0.10 \times 40 \times 7 = 133 \text{ mmHg}$$

Si baja su consumo de sal a 5 g, la nueva TAS que esperamos es:

$$TAS = 5.0 + 2.5 \times 40 + 8.0 \times 5 - 0.10 \times 40 \times 5 = 125 \text{ mmHg}$$

Por lo tanto, el descenso es de 8 mmHg.

■ Segunda forma:

Como el consumo de sal ha bajado en 2 g, pero la edad no cambia, el descenso de TAS será:

$$\begin{aligned}\text{Cambio de TAS} &= 2.5 \times \text{cambio en edad} \\ &\quad + 8.0 \times \text{cambio en sal} \\ &\quad - 0.10 \times \text{edad} \times \text{cambio en sal} \\ &= 8.0 \times 2 - 0.10 \times 40 \times 2 = 8 \text{ mmHg}\end{aligned}$$

♣ Pregunta 12.16

Solución con calculadora

Ahora tenemos un paciente con 70 años que baja su consumo de sal en 2 g. El cambio de TAS que esperamos es:

$$\begin{aligned}\text{Cambio de TAS} &= 2.5 \times \text{cambio en edad} \\ &\quad + 8.0 \times \text{cambio en sal} \\ &\quad - 0.10 \times \text{edad} \times \text{cambio en sal} \\ &= 8.0 \times 2 - 0.10 \times 70 \times 2 = 2 \text{ mmHg}\end{aligned}$$

♣ Pregunta 12.17

Solución con calculadora

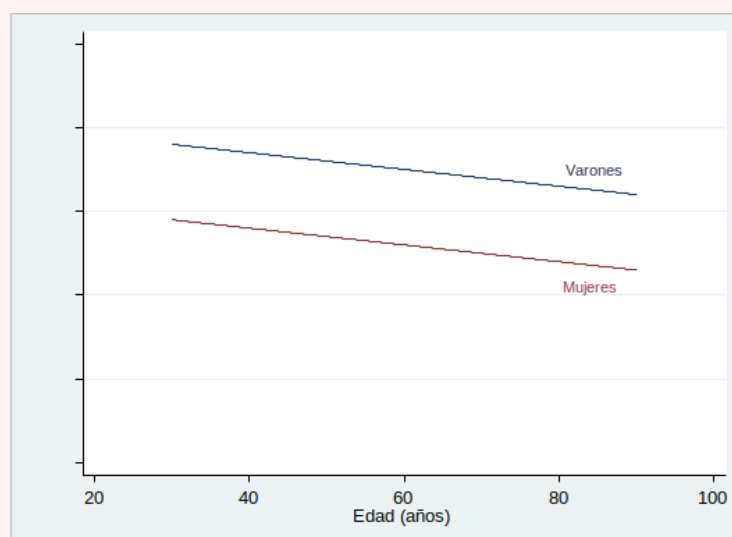
De acuerdo con las preguntas 12.15 y 12.16, la misma reducción en consumo de sal es más efectiva en personas más jóvenes. Por lo tanto, la recomendación preferente debería ir a los jóvenes.

SOLUCIÓN 12.5

♣ Pregunta 12.18**Solución con calculadora**

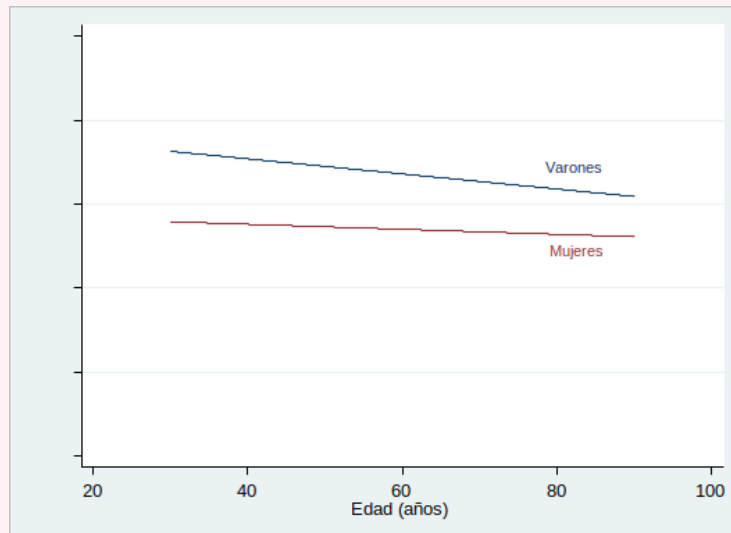
Un modelo que tuviera solo la edad como variable X sería una recta. Al tener también el sexo, se obtienen dos rectas paralelas: una que muestra la relación entre edad y hemoglobina en mujeres y otra en varones. Las dos rectas descienden al aumentar la edad porque su coeficiente es negativo. Pero la recta de mujeres está por debajo de la de varones porque el coeficiente de sexo es también negativo.

El siguiente gráfico muestra el resultado. La escala Y empieza en 10 g/dL para que se aprecie mejor que las líneas son paralelas.

**♣ Pregunta 12.19****Solución con calculadora**

Siguen saliendo dos rectas, pero con el modelo 2 no son paralelas: el término de interacción edad x sexo hace que la recta de mujeres no descienda tan rápido con la edad. El modelo 2 es un ejemplo de interacción entre una variable continua (edad) y otra dicotómica (sexo). El efecto que tiene la edad sobre el nivel de hemoglobina es distinto en varones que en mujeres.

El siguiente gráfico muestra el resultado. La escala Y se inicia en 10 g/dL para que se aprecie mejor que las rectas no son paralelas.



♣ Pregunta 12.20

Solución con calculadora

En el modelo 1, un varón y una mujer de la misma edad tienen una diferencia de 0.90 g/dL (da igual qué edad tengan mientras sea la misma para los dos). Dos varones de distinta edad tienen una diferencia de $0.01 \times \text{diferencia de edad}$, y lo mismo ocurre con dos mujeres de la misma edad. Por lo tanto:

Paciente A	Paciente B	Hemoglobina _A - Hemoglobina _B
Varón de 38 años	Mujer de 38 años	0.90 g/dL
Varón de 62 años	Mujer de 62 años	0.90 g/dL
Varón de 38 años	Varón de 62 años	0.34 g/dL
Mujer de 38 años	Mujer de 62 años	0.34 g/dL

♣ **Pregunta 12.21**

Solución con calculadora

En el modelo 2 hay una interacción entre edad y sexo. Esto quiere decir que la diferencia entre un varón y una mujer de la misma edad ya no será constante: depende de la edad concreta que tengan ambos. Y la diferencia entre dos varones de distinta edad ya no será la misma que entre dos mujeres de las mismas edades. Aplicando el modelo se obtiene:

Paciente A	Paciente B	Hemoglobina _A - Hemoglobina _B
Varón de 38 años	Mujer de 38 años	0.904 g/dL
Varón de 62 años	Mujer de 62 años	0.834 g/dL
Varón de 38 años	Varón de 62 años	0.216 g/dL
Mujer de 38 años	Mujer de 62 años	0.144 g/dL

SOLUCIÓN 12.6

♣ Pregunta 12.22**Solución con calculadora**

En una mujer premenopáusica sin enfermedad crónica, la media esperada de hemoglobina es 11.5g/dL (es decir, de todo el modelo solo interviene la constante: la variable posmenopausia vale 0, la variable enfermedad crónica también vale 0 y su interacción posmenopausia x enf. crónica también es 0).

En cambio, en una mujer premenopáusica con enfermedad crónica, la variable enfermedad crónica vale 1; por lo tanto, la media de hemoglobina será $11.5 - 0.5 = 11.0\text{g/dL}$.

¿Cuánto baja la hemoglobina si una mujer premenopáusica tiene enfermedad crónica? 0.5g/dL .

♣ Pregunta 12.23**Solución con calculadora**

En una mujer posmenopáusica sin enfermedad crónica, la media esperada de hemoglobina es $11.5 + 0.8 = 12.3\text{g/dL}$ (es decir, de todo el modelo solo intervienen la constante y la variable posmenopausia: la variable enfermedad crónica vale 0 y su interacción posmenopausia x enf. crónica también es 0).

En cambio, en una mujer posmenopáusica con enfermedad crónica, la variable enfermedad crónica vale 1 y la interacción posmenopausia x enf. crónica también vale 1. Por lo tanto, para conocer la media de hemoglobina hay que sumar todos los coeficientes: $11.5 + 0.8 - 0.5 + 0.2 = 12.0\text{g/dL}$.

¿Cuánto baja la hemoglobina si una mujer premenopáusica tiene enfermedad crónica? $12.3 - 12.0 = 0.3\text{g/dL}$.

Este es un ejemplo de interacción entre dos variables dicotómicas. El efecto que tiene la enfermedad crónica sobre el nivel de hemoglobina es distinto en mujeres premenopáusicas que en mujeres posmenopáusicas.



Septiembre, 2021

Hay muchos buenos libros de bioestadística. La mayoría de ellos desarrollan los mismos conceptos con enfoques ligeramente distintos. Lo que les diferencia fundamentalmente es el estilo. Algunos son más teóricos que otros, algunos hacen más hincapié en la realización de problemas o en el uso de paquetes estadísticos. Lo que no se encuentra con facilidad es un libro dedicado exclusivamente a hacer problemas. Eso es lo que presentamos aquí.

El libro es repetitivo. En algunos capítulos se propone una y otra vez el mismo tipo de problema con diferentes datos. Puede que la primera vez un alumno quiera ir directamente a la solución, pero hacer lo mismo en los siguientes problemas no le aportará nada: tiene que enfrentarse al papel en blanco para intentar encontrar la solución por sí mismo.

Nuestro enfoque es que saberse los comandos de un paquete estadístico no es lo mismo que saber estadística por el mismo motivo que aprender el funcionamiento de un ecocardiógrafo no le convierte a uno en cardiólogo. Por eso creemos que los problemas hay que hacerlos al menos una vez con calculadora para comprender bien qué es lo que uno se trae entre manos. Pero a lo largo del libro aportamos también soluciones con Excel o con Stata. La mayor parte de los capítulos parten de una base de datos en Excel accesible mediante los enlaces indicados en el texto. En algunos problemas, el primer paso es usar esa base de datos para generar una tabla a partir de la cual se puede realizar el resto del problema con calculadora. El alumno que no esté interesado en el manejo de Excel o Stata, puede saltarse ese paso, consultar directamente la tabla en las soluciones y seguir con el problema «a mano».

No queremos que nadie se llame a engaño: este no es un libro sobre cómo resolver problemas de bioestadística con Excel o con Stata. El eje del libro es la solución con calculadora. Las soluciones con Excel y Stata se aportan como complemento. El mensaje es: peléese usted con las fórmulas hasta que las entienda. Solo después de eso busque un programa con el que hacer las cuentas.

