Estadística Multivariada

ISBN: 978-956-306-074-4
Registro de Propiedad Intelectual: 200.545
Colección: Herramientas para la formación de profesores de matemáticas.
Diseño: Jessica Jure de la Cerda
Diseño de Ilustraciones: Cristina Felmer Plominsky, Catalina Frávega Thomas, Aurora Muñoz Lacourly
Diagramación: Pedro Montealegre Barba, Francisco Santibáñez Palma
Financiamiento: Proyecto Fondef D05I-10211

Datos de contacto para la adquisición de los libros:
Para Chile:
1. En librerías para clientes directos.
2. Instituciones privadas directamente con:
   Juan Carlos Sáez C.
   Director Gerente
   Comunicaciones Noreste Ltda.
   J.C. Sáez Editor
   jcsaezc@vtr.net
   www.jcsaezeditor.blogspot.com
   Oficina: (56 2) 3260104 - (56 2) 3253148
3. Instituciones públicas o fiscales: www.chilecompra.cl

Desde el extranjero:
1. Liberalia Ediciones: www.liberalia.cl
2. Librería Antártica: www.antartica.cl
3. Argentina: Ediciones Manantial: www.emanantial.com.ar
4. Colombia: Editorial Siglo del Hombre
   Fono: (571) 3377700
5. España: Tarahumara, tarahumara@tarahumaralibros.com
   Fono: (34 91) 3656221
6. México: Alejandría Distribución Bibliográfica, alejandria@alejandrialibros.com.mx
   Fono: (52 5) 556161319 - (52 5) 6167509
7. Perú: Librería La Familia, Avenida República de Chile # 661
8. Uruguay: Dolmen Ediciones del Uruguay
   Fono: 00-598-2-7124857

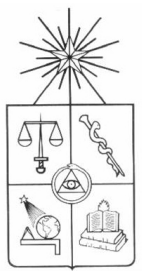
Estadística Multivariada | Nancy Lacourly
Facultad de Ciencias Físicas y Matemáticas, Universidad de Chile
nlacourl@dim.uchile.cl

ESTA PRIMERA EDICIÓN DE 2.000 EJEMPLARES
Se terminó de imprimir en febrero de 2011 en WORLDCOLOR CHILE S.A.

ESTADÍSTICA MULTIVARIADA

Nancy Lacourly

Universidad de Chile



Dirección del Proyecto Fondef D05I-10211
Herramientas para la Formación de Profesores de Matemática

Patricio Felmer, Director del Proyecto
Universidad de Chile.

Leonor Varas, Directora Adjunta del Proyecto
Universidad de Chile.

Salomé Martínez, Subdirectora de Monografías
Universidad de Chile.

Cristián Reyes, Subdirector de Estudio de Casos
Universidad de Chile.

Presentación de la Colección

La colección de monografías que presentamos es el resultado del generoso esfuerzo de los autores, quienes han dedicado su tiempo y conocimiento a la tarea de escribir un texto de matemática. Pero este esfuerzo y generosidad no se encuentra plenamente representado en esta labor, sino que también en la enorme capacidad de aprendizaje que debieron mostrar, para entender y comprender las motivaciones y necesidades de los lectores: Futuros profesores de matemática.

Los autores, encantados una y otra vez por la matemática, sus abstracciones y aplicaciones, enfrentaron la tarea de buscar la mejor manera de traspasar ese encanto a un futuro profesor de matemática. Éste también se encanta y vibra con la matemática, pero además se apasiona con la posibilidad de explicarla, enseñarla y entregarla a los jóvenes estudiantes secundarios. Si la tarea parecía fácil en un comienzo, esta segunda dimensión puso al autor, matemático de profesión, un tremendo desafío. Tuvo que salir de su oficina a escuchar a los estudiantes de pedagogía, a los profesores, a los formadores de profesores y a sus pares. Tuvo que recibir críticas, someterse a la opinión de otros y reescribir una y otra vez su texto. Capítulos enteros resultaban inadecuados, el orden de los contenidos y de los ejemplos era inapropiado, se hacía necesario escribir una nueva versión y otra más. Conversaron con otros autores, escucharon sus opiniones, sostuvieron reuniones con los editores. Escuchar a los estudiantes de pedagogía significó, en muchos casos, realizar eventos de acercamiento, desarrollar cursos en base a la monografía, o formar parte de cursos ya establecidos. Es así que estas monografías recogen la experiencia de los autores y del equipo del proyecto, y también de formadores de profesores y estudiantes de pedagogía. Ellas son el fruto de un esfuerzo consciente y deliberado de acercamiento, de apertura de caminos, de despliegue de puentes entre mundos, muchas veces, separados por falta de comunicación y cuya unión es vital para el progreso de nuestra educación.

La colección de monografías que presentamos comprende una porción importante de los temas que usualmente encontramos en los currículos de formación de profesores de matemática de enseñanza media, pero en ningún caso pretende ser exhaustiva. Del mismo modo, se incorporan temas que sugieren nuevas formas de abordar los contenidos, con énfasis en una matemática más pertinente para el futuro profesor, la que difiere en su enfoque de la matemática para un ingeniero o para un licenciado en matemática, por ejemplo. El formato de monografía, que aborda temas específicos

con extensión moderada, les da flexibilidad para que sean usadas de muy diversas maneras, ya sea como texto de un curso, material complementario, documento básico de un seminario, tema de memoria y también como lectura personal. Su utilidad ciertamente va más allá de las aulas universitarias, pues esta colección puede convertirse en la base de una biblioteca personal del futuro profesor o profesora, puede ser usada como material de consulta por profesores en ejercicio y como texto en cursos de especialización y post-títulos. Esta colección de monografías puede ser usada en concepciones curriculares muy distintas. Es, en suma, una herramienta nueva y valiosa, que a partir de ahora estará a disposición de estudiantes de pedagogía en matemática, formadores de profesores y profesores en ejercicio.

El momento en que esta colección de monografías fue concebida, hace cuatro años, no es casual. Nuestro interés por la creación de herramientas que contribuyan a la formación de profesores de matemática coincide con un acercamiento entre matemáticos y formadores de profesores que ha estado ocurriendo en Chile y en otros lugares del mundo. Nuestra motivación nace a partir de una creciente preocupación en todos los niveles de la sociedad, que ha ido abriendo paso a una demanda social y a un interés nacional por la calidad de la educación, expresada de muy diversas formas. Esta preocupación y nuestro interés encontró eco inmediato en un grupo de matemáticos, inicialmente de la Universidad de Chile, pero que muy rápidamente fue involucrando a matemáticos de la Pontificia Universidad Católica de Chile, de la Universidad de Concepción, de la Universidad Andrés Bello, de la Universidad Federico Santa María, de la Universidad Adolfo Ibáñez, de la Universidad de La Serena y también de la Universidad de la República de Uruguay y de la Universidad de Colorado de Estados Unidos.

La matemática ha adquirido un rol central en la sociedad actual, siendo un pilar fundamental que sustenta el desarrollo en sus diversas expresiones. Constituye el cimiento creciente de todas las disciplinas científicas, de sus aplicaciones en la tecnología y es clave en las habilidades básicas para la vida. Es así que la matemática actualmente se encuentra en el corazón del currículo escolar en el mundo y en particular en Chile. No es posible que un país que pretenda lograr un desarrollo que involucre a toda la sociedad, descuide el cultivo de la matemática o la formación de quienes tienen la misión de traspasar de generación en generación los conocimientos que la sociedad ha acumulado a lo largo de su historia.

Nuestro país vive cambios importantes en educación. Se ha llegado a la convicción que la formación de profesores es la base que nos permitirá generar los cambios cualitativos en calidad que nuestra sociedad ha impuesto. Conscientes de que la tarea formativa de los profesores de matemática y de las futuras generaciones de jóvenes es extremadamente compleja, debido a que confluyen un sinnúmero de factores y disciplinas, a través de esta colección de monografías, sus editores, autores y todos los que han participado del proyecto en cada una de sus etapas, contribuyen a esta tarea, poniendo a disposición una herramienta adicional que ahora debe tomar vida propia en los formadores, estudiantes, futuros profesores y jóvenes de nuestro país.

Patricio Felmer y Salomé Martínez
Editores

Agradecimientos

Agradecemos a todos quienes han hecho posible la realización de este proyecto Fondef: “Herramientas para la formación de Profesores de Matemáticas". A Cristián Cox, quien apoyó con decisión la idea original y contribuyó de manera crucial para obtener la participación del Ministerio de Educación como institución asociada. Agradecemos a Carlos Eugenio Beca por su apoyo durante toda la realización del proyecto. A Rafael Correa, Edgar Kausel y Juan Carlos Sáez, miembros del Comité Directivo. Agradecemos a Rafael Benguria, Servet Martínez y Fidel Oteiza, miembros del Comité Editorial de la colección, quienes realizaron valiosos aportes a los textos. A Guillermo Marshall, Decano de la Facultad de Matemáticas de la Pontificia Universidad Católica de Chile y José Sánchez, entonces Decano de la Facultad de Ciencias Físicas y Matemáticas de la Universidad de Concepción, quienes contribuyeron de manera decisiva a lograr la integridad de la colección de 15 monografías. A Jaime San Martín, director del Centro de Modelamiento Matemático por su apoyo durante toda la realización del proyecto. Agradecemos a Víctor Campos, Ejecutivo de Proyectos de Fondef, por su colaboración y ayuda en las distintas etapas del proyecto.

Agradecemos también a Bárbara Ossandón de la Universidad de Santiago, a Jorge Ávila de la Universidad Católica Silva Henríquez, a Víctor Díaz de la Universidad de Magallanes, a Patricio Canelo de la Universidad de Playa Ancha en San Felipe y a Osvaldo Venegas y Silvia Vidal de la Universidad Católica de Temuco, quienes hicieron posible las visitas que realizamos a las carreras de pedagogía en matemática. Agradecemos a todos los evaluadores, alumnos, académicos y profesores -cuyos nombres no incluimos por ser más de una centena- quienes entregaron sugerencias, críticas y comentarios a los autores, que ayudaron a enriquecer cada uno de los textos.

Agradecemos a Marcela Lizana por su impecable aporte en todas las labores administrativas del proyecto, a Aldo Muzio por su colaboración en la etapa de evaluación, y también a Anyel Alfaro por sus contribuciones en la etapa final del proyecto y en la difusión de los logros alcanzados.

Dirección del Proyecto

Índice General

Las cifras no mienten, pero los mentirosos también usan cifras
Anónimo

Prefacio

Introducir el estudio de las Probabilidades y Estadística en la Enseñanza Media no ha sido fácil para los profesores de Matemática. Es posible que la dificultad emane de una formación basada en el aprendizaje de la matemática como una "ciencia exacta", reducible en último término a la aplicación de algoritmos.

La necesidad de una interdisciplinariedad en la formación del profesor de Enseñanza Media surgió hace algunos años, dejando atrás el carácter teórico y descontextualizado para dar lugar a un conocimiento práctico y contextualizado. La estadística puede permitir el encuentro de las matemáticas con otras disciplinas, como la biología o las ciencias sociales.

En la monografía "Introducción a la Estadística" [1] el lector descubrió el pensamiento estadístico y, a través de muchas ilustraciones y ejemplos, sus conceptos básicos. En esta monografía se refuerzan los conceptos de la teoría de tests de hipótesis y se entregan más justificaciones matemáticas y nuevas distribuciones de probabilidad. Se centra en métodos para datos multivariados.

Si bien esta monografía es más avanzada que la Introducción a la Estadística, tiene el mismo espíritu, prefiriendo explicar los conceptos de la estadística y la interpretación de los resultados sobre las demostraciones matemáticas de teoremas, que aun si no son ausentes, pueden saltarse en una primera lectura. En el Capítulo 1 presentamos el análisis en componentes principales, el método más simple y más importante del análisis descriptivo multivariado, que se funda en resultados del algebra lineal. El Capítulo 2 contiene la teoría de tests estadísticos para medias y proporciones basada en el modelo Normal. Se entregan varias aplicaciones, en particular para comparar más de dos poblaciones (ANOVA). En el Capítulo 3 se presenta la regresión lineal múltiple y la discusión de un caso. En el Capítulo 4, se describe un método de predicción alternativo a la regresión lineal y al ANOVA: los árboles de clasificación y regresión (CART). Es un método no lineal, que usa criterios presentados en los dos capítulos anteriores y permite jerarquizar la importancia de las variables en el modelo, cuya visualización lo hace muy interesante.

[1] N. Lacourly, *Introducción a la Estadística*, Editorial JC. Sáez, Santiago, 2011.

Hemos intercalado referencias históricas cuando eso pareció relevante, y agregado ejercicios de autoevaluación para ayudar a la comprensión del texto.

Por su ayuda en las varias fases de este libro, me gustaría agradecer a Lorena Cerda.

Quiero agradecer especialmente a los editores, Patricio Felmer y Salomé Martínez, no sólo por haber desarrollado el proyecto Fondef, que hizo posible esta monografía, sino además por la enorme dedicación que aplicaron a su corrección y producción.

Agradezco a Francisco Santibañez por todos los aportes que hizo en la diagramación del texto.

Agradezco a Juan Muñoz, mi esposo, quien siempre me prestó apoyo y sabe lo importante que ha sido para mí escribir este texto.

Finalmente, con una inmensa alegría, dedico este trabajo especialmente a mis queridos hijos.

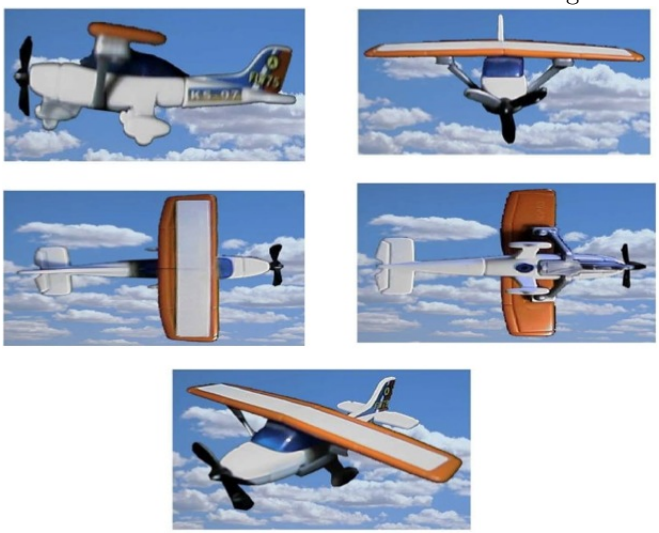
Nancy Lacourly 2010

Capítulo 1: Análisis en componentes principales

1.1 Introducción

Nuestro mundo está lleno de información que no siempre podemos analizar por lo complejo que resulta entenderla. Nos hemos acostumbrado a visualizar los objetos y los datos para describirlos, pues así es más fácil presentarlos e interpretarlos. Ilustremos con un ejemplo: en la Figura 1.1 se presentan fotos de un avión, tomadas desde diferentes ángulos. El avión es un objeto de tres dimensiones, y cada foto muestra un aspecto diferente de ese objeto, impresa sobre la superficie de un papel, que sólo tiene dos dimensiones. Los ángulos y la cantidad de fotos que necesitamos tomar para describir razonablemente el avión de esta manera dependen de varios factores, y particularmente, de la finalidad de esa descripción: las exigencias de una galería de arte son ciertamente distintas que las de una agencia de espionaje. Naturalmente, la complejidad intrínseca del objeto también es un factor importante.

FIGURA 1.1. Un avión visto de diferentes ángulos



Aunque vivimos en un espacio de 3 dimensiones, usamos permanentemente fotos y documentos, que muestran textos y gráficos bidimensionales. Al registrar las notas

de los alumnos en los cuatro años de Enseñanza Media, las notas de cada alumno constituyen un vector en $\mathbb{R}^p$, donde p es el número total de notas. Calcular el promedio de notas de cada año y después promediar los cuatros años es una manera de reducir esta realidad *multivariada* a una dimensión que facilita su interpretación y uso. Esto ha sido posible porque las notas de un mismo alumno se relacionan entre sí de alguna forma. Pero, el costo de esta simplificación es una disminución de la *precisión*. Después de promediar ya no sabemos en qué asignaturas o qué años le fue mejor a un alumno, o si ha tenido muchos altibajos. Cuando el obstetra aplica el test de Apgar a un recién nacido, está resumiendo varios aspectos relacionados, como el color de la piel, la frecuencia cardiaca, los reflejos, el tono muscular y la respiración, en un solo número (que no es un promedio). Al analizar estos indicadores sintéticos, es importante saber cuál fue el criterio que se usó para simplificar una realidad, que generalmente es más compleja.

Se dice que los datos son multivariados cuando se dispone de varias variables para cada objeto o individuo estudiado, como en la Tabla 1.1. Cuando las variables tienen un cierto grado de relación entre sí, los datos contienen ciertas redundancias que se pueden aprovechar para reducir la dimensión del espacio necesario para representarlos. El problema es cómo hacerlo perdiendo el mínimo posible de información.

La reducción de la dimensión debe satisfacer dos principios básicos: (i) Explicar lo más posible con la menor cantidad de elementos. Se habla de parsimonia. (ii) Debe facilitar la interpretación de los datos.

El **análisis en componentes principales** (ACP) intenta transformar los datos multivariados (generalmente complejos), con el fin de simplificarlos para facilitar su interpretación. Las transformaciones del ACP lo logran creando nuevas variables, llamadas **componentes principales**, que **no están correlacionadas entre sí**. No se trata de eliminar variables, sino de generar nuevas variables a partir de las variables originales, que contengan la información más relevante contenida en los datos, dejando de lado lo que es poco relevante. La relevancia se refiere a la variabilidad de los datos, pues es claro que una variable que varía poco (o nada) aporta poca (o ninguna) información para un estudio al interior de una población.

Las componentes principales se ordenan, entonces, según su capacidad para reproducir la variabilidad contenida en los datos. Además, las componentes principales se construyen de manera que, cada componente entrega información complementaria a la información producida por las otras componentes, vale decir que las componentes principales no están correlacionadas entre sí.

El ACP fue creado hace más de un siglo (1901) por Karl Pearson (Pearson[**13**]) quién desarrolló una intensa investigación sobre la aplicación de los métodos estadísticos en la biología y fue el fundador de la bioestadística. Sin embargo, el uso práctico de esta técnica tuvo que esperar la llegada de los computadores, pues involucra cálculos complejos (en particular el cálculo de valores y vectores propios) y representaciones gráficas elaboradas.

Actualmente, el ACP se usa en muchas disciplinas tales como la ingeniería, la biología y la educación, pues permite ampliar el campo de la Estadística Descriptiva (llamada también Estadística Exploratoria) presentada en Lacourly[**7**], ayudando a interpretar los datos y apoyando el desarrollo de los modelos inferenciales y predictivos.

El ACP es el método más simple de análisis exploratorio multivariado, que agrupa métodos descriptivos para datos multivariados. El ACP se relaciona con el análisis factorial clásico, utilizado en psicometría, que fue introducido por Spearman (1904) en su trabajo sobre la inteligencia. [1]

1.2 Concepto de índice

Los estudios estadísticos actuales se basan en grandes volúmenes de datos, con muchas variables y muchos individuos, lo que dificulta sus análisis. Se intenta, entonces, sintetizar en unos pocos valores los múltiples aspectos del fenómeno estudiado. Es lo que se hace cuando se toma el promedio de las notas de un alumno para medir su rendimiento general. Las universidades utilizan los puntajes de ingreso, que son promedios ponderados de las diferentes PSU y del promedio de notas de la Enseñanza Media (NEM). Las ponderaciones cambian según la universidad y la carrera. El Programa de las Naciones Unidas para el Desarrollo (PNUD) construye periódicamente un "índice de desarrollo humano" (IDH) que concentra los distintos aspectos demográficos y socioeconómicos de cada país en un solo número, que intenta medir el "desarrollo global" de cada país. El Índice de Precios al Consumidor (IPC) pretende reducir mensualmente una realidad numerosa de precios a un solo número, representativo del momento. En el problema de evaluación de proyectos o de cargos en una institución, se trata de tomar en cuenta varios aspectos: la formación, la experiencia, y varios aspectos sicológicos en un número único que permita ordenar los proyectos o los candidatos, para poder seleccionarlos.

Estos números se llaman **índices**. Un índice, que pretende representar varios aspectos en un solo número, tiene la gran ventaja de permitir ordenar los objetos que se miden, pero tiene el inconveniente de perder información,. Además no da necesariamente la misma importancia a todos los aspectos. En general, no es posible reconstruir los valores iniciales a partir de los valores del índice. Además, dos individuos que toman valores diferentes sobre las variables iniciales, pueden obtener el mismo valor sobre el índice; por ejemplo, dos alumnos con notas diferentes pueden tener el mismo promedio. A pesar de ello, los índices son de amplio uso cotidiano, pues permiten simplificar estudios complejos. La simplificación no es trivial pues debe permitir conservar la *mayor parte de la información* contenida en los datos.

[1] El Análisis Factorial busca factores que expliquen la mayor parte de la varianza común a las variables, mientras que el ACP busca factores que expliquen la máxima varianza contenida en los datos. La varianza común es la parte de la variación de una variable que es compartida con las otras variables. Para saber más ver Cuesta y Herrero[**5**].

Muchos índices son "lineales". Supongamos por ejemplo, los valores de 5 variables, x_1, x_2, x_3, x_4, x_5 medidas sobre un individuo, un índice Y es una combinación lineal de los 5 valores, o un promedio ponderado:

$$Y = (a_1x_1 + a_2x_2 + a_3x_3 + a_4x_4 + a_5x_5)/(a_1 + a_2 + a_3 + a_4 + a_5)$$

donde a_1, a_2, a_3, a_4, a_5 son las ponderaciones que definen el índice Y.

Considerando, por ejemplo, 3 individuos y las ponderaciones $a_1 = 2$, $a_2 = 1$, $a_3 = 1,5$, $a_4 = 0,5$, $a_5 = 2$, se obtiene los valores del índice 1 de la Tabla adjunta:

	x_1	x_2	x_3	x_4	x_5	Índice 1	Índice 2
Individuo 1	1	4	2	10	3	2,857	4,737
Individuo 2	1	4	5	3	2	2,714	2,789
Individuo 3	4	6	4	1	4	4,071	3,263

Observamos que el tercer individuo obtiene el mayor valor para el índice 1. Ahora, en el índice 2, usamos otra combinación lineal: $a_1 = 2$, $a_2 = 1$, $a_3 = 1,5$, $a_4 = 3$, $a_5 = 2$, que da el mayor valor al primer individuo. Lo anterior se debe esencialmente a la ponderación de la variable x_4.

El valor del índice depende de las ponderaciones utilizadas. Las ponderaciones pueden determinarse a priori, como se hace en los puntajes de ingreso para la universidad (Promedio ponderado de las PSU y de la NEM), o bien pueden obtenerse de un criterio que se optimiza. El ACP es un método que permite obtener ponderaciones, basándose en un criterio objetivo. Su ventaja está en su poder de simplificar los datos para mostrar, de manera aproximada, las distancias entre los individuos y las relaciones entre las variables y, de esta manera, permite explicar las diferencias entre los individuos.

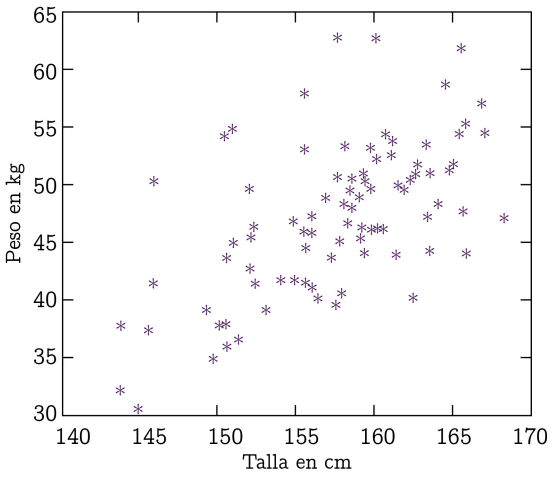
Para introducir el ACP, que permite determinar las ponderaciones de los índices, usaremos un ejemplo muy sencillo, donde se reducen datos bidimensionales en un índice (datos unidimencionales).

1.3 Ejemplo con dos variables

Queremos comparar 90 niñas entre 13 y 15 años según su peso y su talla. Cada niña puede representarse como un punto de coordenadas $(x, y) \in \mathbb{R}^2$, donde x es la talla e y el peso. Podemos entonces visualizar los 90 puntos en un gráfico de dispersión (Figura 2.1). Se puede apreciar una "nube de puntos" que muestra que existe una cierta relación entre el peso y la talla de las niñas: Las niñas altas tienden a tener un mayor peso que las niñas bajas. Es entonces natural buscar indicadores analíticos (o índices) que permiten cuantificar la forma y el grado de esta relación.

Si se trata de predecir el peso de una niña entre 13 y 15 años a partir de su talla, se podría utilizar la regresión lineal (Lacourly[**7**]). Pero aquí la idea no es predecir una de las variables a partir de la otra sino construir un índice de "corpulencia", que representa en parte el peso y en parte la talla en un solo valor. Reflexionemos sobre las condiciones que requiere este índice.

FIGURA 1.2. Peso y Talla niñas entre 13 y 15 años



1.3.1 Construcción de un índice de corpulencia

La forma más usual de obtener un índice c, que refleje a la vez el peso y la talla es un promedio ponderado (una combinación lineal) de las dos variables:[2]

$$c = \alpha x + \beta y.$$

El problema está en elegir los coeficientes α y β. Si se toma $\alpha = 3$ y $\beta = 2$, una niña de $50,5$ kg y $158,5$ cm obtiene un índice de $3 \times 158,5 + 2 \times 50,5 = 468,5$, mientras que otra que pese $37,8$ kg y mide $150,5$ obtiene $3 \times 150,5 + 2 \times 37,8 = 527,1$. Si en lugar de usar los coeficientes anteriores, tomamos $\alpha = 2$ y $\beta = 3$, la primera niña obtiene $2 \times 50,5 + 3 \times 158,5 = 576,5$, y la segunda $2 \times 37,8 + 3 \times 150,5 = 414,4$.

Notemos que tomando $\alpha = 20$ y $\beta = 30$, el índice difiere del segundo tan sólo por un factor de escala de 10, lo que no cambian las corpulencias relativas entre las niñas. En otras palabras, lo importante es la relación entre α y β. No se pierde entonces generalidad al imponer una condición de normalización. Se podría usar $\alpha + \beta = 1$, sin embargo en ACP se usa $\alpha^2 + \beta^2 = 1$, que se justifica a continuación. En el caso $\alpha = 3$ y $\beta = 2$, al normalizar obtenemos $\alpha = 0,8321$ y $\beta = 0,5547$. Geométricamente, podemos ver que los coeficientes α y β, cuando están normalizados, pueden considerarse como las coordenadas de un vector unitario $u = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$ de $\mathbb{R}^2$, o sea $||u||^2 = \alpha^2 + \beta^2 = 1$, donde $||v||$ es la norma del vector v.

[2]El índice de masa corporal, que es el cociente del peso en kilogramo y del cuadrado de talla en metro, es otra alternativa. Es un índice no lineal.

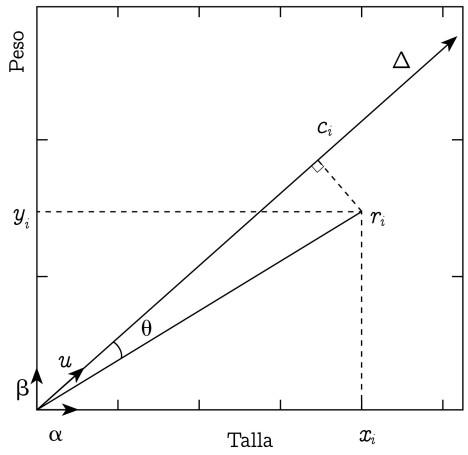
Entonces, si x_i e y_i son el peso y la talla de la niña i, su índice de corpulencia c_i es el producto escalar entre el vector $r_i = \begin{pmatrix} x_i \\ y_i \end{pmatrix}$ y el vector u:

$$c_i = \alpha x_i + \beta y_i =< r_i, u >= ||r_i|| \times ||u|| \times cos\theta = ||r_i|| cos\theta.$$

El valor c_i corresponde entonces a la norma de la proyección ortogonal P_{r_i} del vector r_i sobre el eje Δ definido por el vector u, $c_i = ||r_i|| cos\theta = ||P_{r_i}||$ (Figura 1.3).

A cada vector u corresponde un índice de corpulencia, que se obtiene calculando el producto escalar entre u y r_i.

FIGURA 1.3. Proyecciones ortogonales sobre una recta



1.3.2 Criterio para un índice óptimo

¿Como elegir un índice de corpulencia óptimo? El criterio de optimalidad dependerá de lo que se busca. En la Figura 1.4 vemos que los puntos r_i y r_k, que son diferentes, se proyectan sobre la recta Δ en puntos cercanos. Observamos que el hecho de proyectar ortogonalmente sobre una recta, no puede agrandar las distancias, sino achicarlas o dejarlas iguales. ¿Cuando quedan iguales? Cuando los puntos se encuentran sobre la recta Δ o sobre rectas paralelas a Δ.

Una condición deseable es, entonces, que el índice entregue valores de corpulencia muy diferentes a niñas con pesos y tallas muy diferentes, dado que niñas con pesos y tallas parecidos, siempre tomarán valores parecidos del índice.

Una forma de lograrlo sería elegir el vector $u = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$, que evita que niñas con pesos y tallas muy diferentes, tengan valores parecidos del índice. La manera de

lograrlo es buscando que las corpulencias estén tan dispersas como sea posible; en otras palabras, se busca el vector unitario u tal que la varianza de las proyecciones ortogonales de los puntos r_i sobre la recta Δ generada por el vector u sea máxima. Se trata, entonces, de maximizar la cantidad:

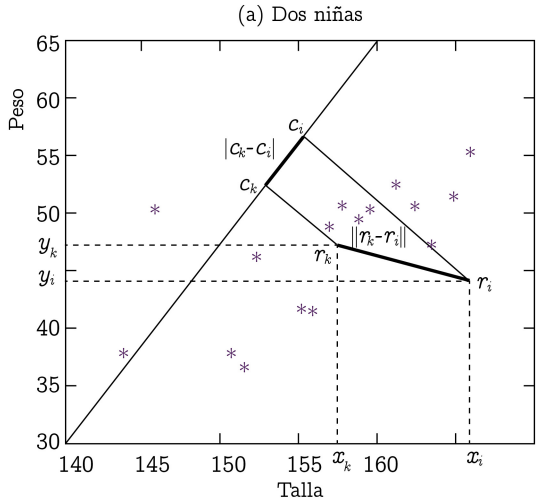
$$Var(c) = \frac{1}{90}\sum_{i=1}^{90}(c_i - \bar{c})^2, \tag{1.1}$$

donde

$$c_i = \alpha x_i + \beta y_i \quad \bar{c} = \tfrac{1}{90}\textstyle\sum_{i=1}^{90} c_i = \alpha\bar{x} + \beta\bar{y} \tag{1.2}$$

es la media de los c_i, con $\bar{x} = \frac{1}{90}\sum_{i=1}^{90} x_i$ y $\bar{y} = \frac{1}{90}\sum_{i=1}^{90} y_i$ las medias de los x_i e y_i, respectivamente.

FIGURA 1.4. Proyecciones ortogonales



Reemplazando c_i y $\bar{c}$ en la ecuación (1.1), obtenemos

$$\frac{1}{90}\sum_i(\alpha x_i + \beta y_i - \alpha\bar{x} - \beta\bar{y})^2 = \frac{1}{90}\sum_i[\alpha(x_i - \bar{x}) + \beta(y_i - \bar{y})]^2$$

$$= \frac{1}{90}[\alpha^2\sum_i(x_i - \bar{x})^2 + \beta^2\sum_i(y_i - \bar{y})^2 + 2\alpha\beta\sum_i(x_i - \bar{x})(y_i - \bar{y})],$$

o sea,

$$Var(c) = \alpha^2\,\mathrm{Var}(x) + 2\alpha\beta Cov(x,y) + \beta^2 Var(y), \tag{1.3}$$

donde $Var(x)$ y $Var(y)$ son las varianzas de los x_i e y_i, respectivamente, y $Cov(x, y)$ es la covarianza entre los x_i e y_i:

$$Var(x) = \frac{1}{90}\sum_{i=1}^{90}(x_i - \bar{x})^2;\ Var(y) = \frac{1}{90}\sum_{i=1}^{90}(y_i - \bar{y})^2;\ Cov(x, y) = \frac{1}{90}\sum_{i=1}^{90}(x_i - \bar{x})(y_i - \bar{y}).$$

Escribamos matricialmente la expresión de $Var(c)$. Definamos la matriz de varianzas-covarianzas V asociada a las dos variables:

$$V = \begin{pmatrix} Var(x) & Cov(x, y) \\ Cov(x, y) & Var(y) \end{pmatrix}.$$

Como $u = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$, obtenemos $Var(c) = u^t V u$.

Tenemos entonces que encontrar el vector u no nulo que maximiza la expresión $u^t V u$. La solución se basa en resultados del álgebra lineal.

Proposición 1.1. *El vector u no nulo, que maximiza la expresión $Var(c) = u^t V u$, es vector propio de la matriz V asociado al mayor valor propio.*

Demostración. Vemos que la solución involucra a la diagonalización de la matriz V. Recordamos entonces previamente algunas propiedades de diagonalización matricial (ver Gil[**6**] Sección 2.3). Además, presentaremos dos demostraciones alternativas.

- La matriz V es simétrica y, por lo tanto, diagonalizable en $\mathbb{R}$, es decir, existen dos valores propios reales distintos, λ_1 y λ_2, o un valor propio real doble ($\lambda_1 = \lambda_2$).
- Vimos que la expresión $u^t V u$ es una varianza, entonces es positiva o nula. La matriz V tiene sus valores propios no negativos (V es semidefinida positiva).
- Si los dos valores propios λ_1 y λ_2 de V son distintos, los dos subespacios propios asociados son rectas ortogonales de $\mathbb{R}^2$. Si u_1 y u_2 son los vectores propios unitarios (de norma igual a 1) asociados respectivamente a λ_1 y λ_2, forman una base ortonormal del plano: $Vu_1 = \lambda_1 u_1$ y $Vu_2 = \lambda_2 u_2$ con $||u_1|| = ||u_2|| = 1$ y $u_1 \perp u_2$, donde $u_1 \perp u_2$ designa la ortogonalidad de u_1 y u_2.
- Si V tiene un valor propio doble, el subespacio propio asociado es todo el plano $\mathbb{R}^2$. Se puede, entonces, elegir cualquiera de los vectores ortogonales u_1 y u_2 de norma 1 de $\mathbb{R}^2$. En este caso no hay un índice mejor que los otro, con el criterio de maximizar la varianza.

Sean $\{u_1, u_2\}$ vectores propios unitarios de V, que forman una base ortonormal del plano. Se puede escribir cualquier vector u del plano en esta base: $u = au_1 + bu_2$, donde a y b son las coordenadas de u sobre u_1 y u_2, respectivamente. Si $||u|| = 1$, $||u||^2 = a^2||u_1||^2 + b^2||u_2||^2 = a^2 + b^2$, dado que $u_1 \perp u_2$ y $||u_1|| = ||u_2|| = 1$.

Expresamos entonces $Var(c)$ a partir de u_1 y u_2:

$$Var(c) = u^t V u = (au_1 + bu_2)^t V (au_1 + bu_2) = a^2 u_1^t V u_1 + abu_1^t V u_2 + abu_2^t V u_1 + b^2 u_2^t V u_2,$$

donde u_1^t designa la traspuesta de u_1.

Como u_1 y u_2 son ortogonales, $u_1^t u_2 = 0$. Ahora bien, $Vu_2 = \lambda_2 u_2$, luego $u_1^t V u_2 = \lambda_2 u_1^t u_2 = 0$. Además, $||u_1|| = 1$, o sea, $||u_1||^2 = u_1^t u_1 = 1$, de lo cual se deduce que $u_1^t V u_1 = \lambda_1 u_1^t u_1 = \lambda_1$. De la misma manera obtenemos que $u_2^t V u_2 = \lambda_2$. Finalmente,

$$Var(c) = \lambda_1 a^2 + \lambda_2 b^2.$$

Vimos que $||u||^2 = a^2 + b^2$, luego tomando u unitario, $a^2 + b^2 = 1$. Si $\lambda_1 \geq \lambda_2$, $Var(c)$ toma su valor máximo cuando $a = 1$ y $b = 0$, o sea, $u = u_1$. □

Se presenta a continuación otra demostración de la proposición 1.1, que usa las derivaciones parciales con multiplicadores de Lagrange. Esta demostración no es indispensable para la comprensión del ACP, pero tiene un interés matemático. Retomamos el problema de optimización de $Var(c)$ (expresión 1.3), bajo la restricción $\alpha^2 + \beta^2 = 1$. Usamos un multiplicador de Lagrange λ y derivamos, con respecto a α y β, la expresión $Q = \alpha^2 Var(x) + 2\alpha\beta Cov(x, y) + \beta^2 Var(y) - \lambda(\alpha^2 + \beta^2 - 1)$,

$$\begin{cases} \frac{\partial Q}{\partial \alpha} = 2\alpha Var(x) + 2\beta Cov(x, y) - 2\lambda\alpha \\ \frac{\partial Q}{\partial \beta} = 2\beta Var(y) + 2\alpha Cov(x, y) - 2\lambda\beta. \end{cases}$$

Anulando las derivadas parciales, obtenemos un sistema de dos ecuaciones lineales con dos incógnitas:

$$\begin{cases} Var(x)\alpha + Cov(x, y)\beta = \lambda\alpha \\ Cov(x, y)\alpha + Var(y)\beta = \lambda\beta. \end{cases} \tag{1.4}$$

Recordando que $u = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$, el sistema de ecuaciones (3.4) puede escribirse matricialmente como:

$$Vu = \lambda u. \tag{1.5}$$

El vector u solución de la ecuación (1.5) es entonces un vector propio de la matriz V. Como $u^t V u = \lambda$, el vector propio solución corresponde al mayor valor propio λ_1.

Para los datos de las 90 niñas, $Var(x) = 41,61$, $Var(y) = 32,51$ y $Cov(x, y) = 21,78$, luego

$$V = \begin{pmatrix} 32,51 & 21,78 \\ 21,78 & 41,61 \end{pmatrix}.$$

Los valores propios son $\lambda_1 = 59,32$ y $\lambda_2 = 14,81$ y los vectores propios asociados son:

$$u_1 = \begin{pmatrix} 0,631 \\ 0,776 \end{pmatrix} \quad y \quad u_2 = \begin{pmatrix} -0,776 \\ 0,631 \end{pmatrix}.$$

Verifiquen que u_1 y u_2 son ortogonales y de norma 1.

Para el índice de corpulencia, basado en el criterio de máxima varianza, se obtienen las ponderaciones $\alpha = 0,631$ y $\beta = 0,776$. Para cada niña i se calcula su corpulencia $c_i = 0,631x_i + 0,776y_i$. La varianza de la corpulencia es $Var(c) = \alpha^2 \,\mathrm{Var}(x) + 2\alpha\beta Cov(x, y) + \beta^2 Var(y) = \lambda_1 = 59,32$.

Tenemos la propiedad algebraica siguiente: $\lambda_1 + \lambda_2 = Traza(V) = Var(x) + Var(y) = 74,12$.

El índice obtenido a partir del vector propio u_1 se llama "primera componente principal". Históricamente, el astrónomo y matemático belga Adolphe Quételet (1796-1874) realizó los primeros intentos de aplicar la estadística a las Ciencias Sociales. Es poco conocido que una de sus contribuciones es el famoso Índice de Masa Corporal (IMC), que mide la asociación entre el peso y la talla de un individuo:

$$IMC = \frac{Peso\ \ en\ \ kg}{(Talla\ \ en\ \ m)^2}.$$

En la Figura 1.5 se compara el IMC con el índice de corpulencia proporcionado por la primera componente principal. Si bien, observamos una gran concordancia entre los dos índices, aparecen algunas diferencias importantes. Para entender estas diferencias se identificaron algunas niñas. Por ejemplo, los valores de la primera componente principal de las niñas 10 y 18 son parecidos, pero sus IMC son bastante diferentes. En el gráfico de dispersión de la talla con el peso, vemos que la niña 18 es la más alta, siendo más bien delgada, y que pasa lo contrario con la niña 10 (Figura 1.6). El IMC refleja mejor esta situación, mientras que la primera componente principal trata de dar un índice más global, tomando en cuenta la magnitud tanto de la talla como del peso. En resumen, podemos decir que el IMC muestra las formas, o sea, la relación entre el peso y la talla de las niñas, mientras que la primera componente principal muestra los tamaños globales. Una niña delgada, pero alta, toma una primera componente principal relativamente elevada, pero un IMC bajo. Ahora bien, si tomamos como índice la variable asociada al valor propio λ_2, llamada "segunda componente principal", vemos que es parecido al IMC (Figura 1.7).

FIGURA 1.5. Primera componente principal y IMC

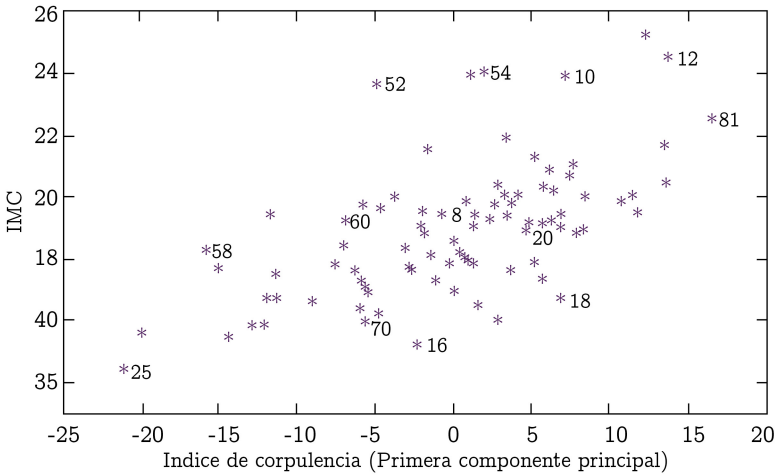


FIGURA 1.6. Peso y talla de niñas de entre 13 y 15 años

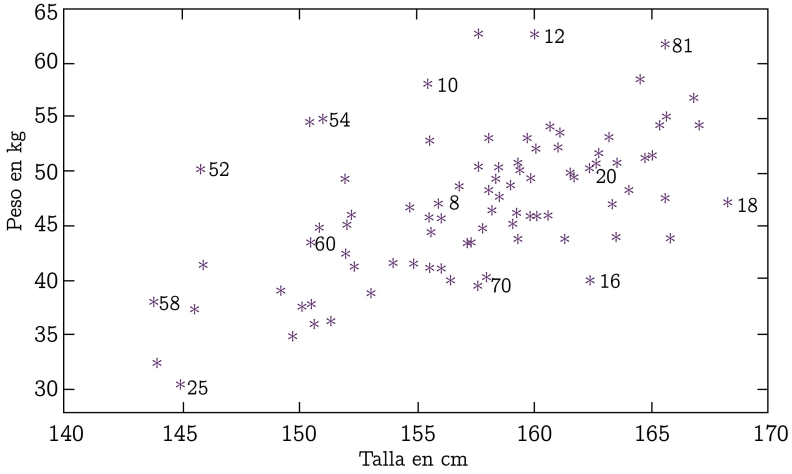
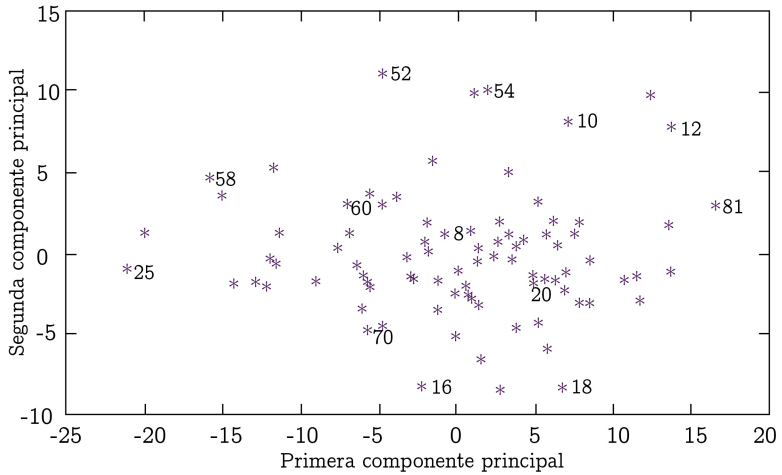


FIGURA 1.7. Primera y segunda componentes principales



1.3.3 Medida de calidad del índice

El índice de corpulencia reduce la información inicial (peso y talla) a una sola dimensión. Hay una "pérdida de información". Es importante evaluar cuánto se pierde o cuánto se recupera de la información del peso y de la talla con el índice de corpulencia. Es lo que se llama *calidad del índice*.

La información inicial se refiere al peso y a la talla de las niñas. Cuando reducimos las dos variables a un solo índice, el criterio utilizado trata de maximizar la "varianza" o dispersión del índice. Tendremos entonces un "buen" índice si la dispersión de este es similar a la dispersión de la nube de puntos inicial en $\mathbb{R}^2$. La dispersión del índice se mide con su varianza:

$$Var(c) = \frac{1}{90}\sum_{i=1}^{90}(c_i - \bar{c})^2,$$

que vale λ_1, y es el promedio de los cuadrados de las distancias entre el índice y su media.

De la misma manera, podemos calcular la dispersión de la nube en $\mathbb{R}^2$ como el promedio de los cuadrados de las distancias entre los puntos r_i y el punto promedio de la talla y de peso $\bar{r} = \begin{pmatrix} \bar{x} \\ \bar{y} \end{pmatrix}$:

$$I_{\bar{r}} = \frac{1}{90}\sum_{i=1}^{90}||r_i - \bar{r}||^2 = \frac{1}{90}\sum_{i=1}^{90}[(x_i - \bar{x})^2 + (y_i - \bar{y})^2] = Var(x) + Var(y).$$

Observamos que $Var(x) + Var(y)$ es la traza de la matriz V (suma de los términos de la diagonal), luego

$$Var(x) + Var(y) = traza(V) = \lambda_1 + \lambda_2.$$

La operación de proyección ortogonal reduce las distancias (Figura 1.8), por lo tanto,

$$Var(c) = \sum_{i=1}^{90}(c_i - \bar{c})^2 \leq \sum_{i=1}^{90}||r_i - \bar{r}||^2.$$

De aquí se deduce que una medida natural de calidad del índice sería

$$100 \times \frac{Var(c)}{Var(x) + Var(y)} = 100 \times \frac{\lambda_1}{Var(x) + Var(y)} = 100 \times \frac{\lambda_1}{\lambda_1 + \lambda_2},$$

que es el porcentaje de la varianza total $I_{\bar{r}}$ conservada por el índice, el que, en nuestro caso, vale

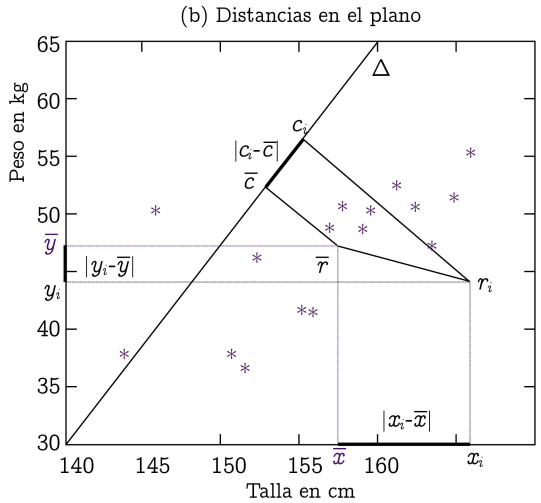
$$100 \times \frac{59,32}{59,32 + 14,81} = 80\,\%.$$

Si el porcentaje fuera 100 %, la nube de puntos pertenecería a la recta definida por el índice. Si el porcentaje fuera 0 %, la recta definida por el índice sería ortogonal al subespacio que contiene a la nube de puntos, lo que es imposible en este caso.

1.3.4 Efecto de la posición y de las unidades de medida: estandarización de los datos

¿Que pasaría si cambiamos las unidades de medición? Por ejemplo, si usamos la talla en dm (se divide por 10), la varianza valdría $0,325$, en vez de $32,51$. Verifiquen cómo cambian la matriz V y las ponderaciones del índice. Encontrarán $\alpha = 0,052$ y

FIGURA 1.8. Proyecciones ortogonales



$\beta = 0,999$. Este cambio de unidades de medición, que otorga a la talla una varianza mucho más pequeña que la varianza del peso, hace que en el índice se tome en cuenta esencialmente el peso.

Cuando el índice es una combinación lineal de las variables, conviene, entonces, construir un índice que no dependa de las unidades de medición. Una manera de evitar este problema consiste en transformar las variables para que tengan la misma varianza. Esto se obtiene dividiendo cada variable por su desviación estándar (raíz de la varianza), de manera que las varianzas normalizadas sean iguales a 1.

Las distancias entre los individuos no dependen de las medias de las variables. Se centran entonces las variables, es decir, a los valores de cada variable se les resta su media. De esta manera, todas las variables tienen su media nula. Así, cuando se centran y reducen los datos se habla de "estandarización", lo que supone que:

- Los datos no dependen de la unidad o escala de medida escogida
- Las variables tienen la misma media (media nula) y la misma dispersión (varianza 1).

En general, conviene estandarizar las variables, ya que esto evita que las unidades de mediciones afecten las distancias entre los individuos.

En este caso, la covarianza es igual al coeficiente de correlación lineal, que mide el grado de relación lineal que existe entre las dos variables. Calculamos entonces la matriz de correlaciones:

$$R = \begin{pmatrix} 1,0 & 0,59 \\ 0,59 & 1,0 \end{pmatrix},$$

cuyos valores propios son $\lambda_1 = 1,59$ y $\lambda_2 = 0,41$, y el vector propio asociado a λ_1 es $u = \begin{pmatrix} 0,71 \\ 0,71 \end{pmatrix}$, que define la "primera bisectriz", recta cuyos puntos toman los mismos valores en la abscisa y la ordenada. Entonces, el índice es igual a $0,71 \times x + 0,71 \times y$, que equivale a tomar el promedio entre las dos variables.

Observen que la varianza total en este caso es igual a 2, la traza de la matriz R, que es el número de variables. El porcentaje de varianza total conservada por la primera componente principal es $100 \times \frac{1,59}{2} = 79,5\,\%$.

1.4 Generalización a más de dos variables

1.4.1 Búsqueda de índices

Si queremos determinar un índice de corpulencia no solamente a partir del peso y de la talla, sino también incorporando el perímetro del brazo y el largo del tronco, podemos generalizar lo anterior. La dificultad está en el número de variables que caracterizan a los individuos u objetos de interés. En efecto, los gráficos de dispersión proporcionan un apoyo visual en el caso de dos variables y el análisis en componentes principales se muestra especialmente útil cuando hay muchas variables y muchas observaciones.

Consideremos un ejemplo con siete variables. Se trata de variables demográficas relativas a 20 países de América Latina, datos obtenidos del PNUD del año 2006 (Tabla 1.1) [3]. Para comparar los 20 países, podemos analizar nuevamente los gráficos de dispersión. La Figura 1.9(a) muestra los 20 países considerando como coordenadas la tasa de alfabetización y la esperanza de vida. Este gráfico permite comparar los países entre sí, y el en qué difieren o en qué se parecen respecto de estas dos variables y concluir de qué manera se relacionan las dos variables. Es así que Haití aparece como aislado, con una tasa de alfabetización y una esperanza de vida bajas, mientras que un grupo de países - Cuba, Chile, Costa Rica, Uruguay y Argentina - tienen los valores más altos en ambas variables, pero no encontramos ningún país con tasa de alfabetización alta y esperanza de vida baja. En efecto, el coeficiente de correlación lineal, que vale $0,79$, es relativamente alto y positivo, lo que permite concluir que existe un cierto grado de relación lineal entre estas dos variables. Si la tasa de alfabetización es alta, la esperanza de vida lo es también, y recíprocamente.

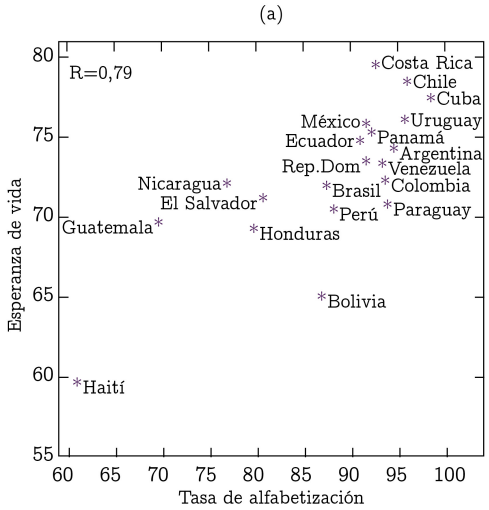
[3] En la tabla ‰ significa "por mil".

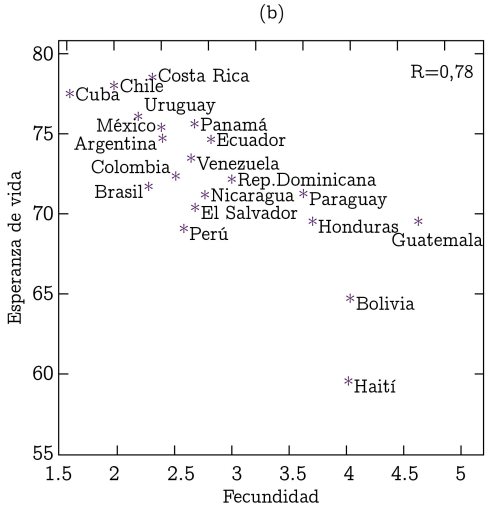
TABLA 1.1. Datos de 20 países de América Latina (PNUD 2006)

i	1	2	3	4	5	6	7
PAÍS	Porcentaje población urbana	Tasa de alfabe-tización (%)	Esperanza de vida (años)	Fecundidad (Nacimientos vivos por mujer)	Mortalidad infantil (‰ de nacimientos vivos)	Educación (indicador)	Usuarios Internet (‰ de habitantes)
ARGENTINA	90,1	97,2	74,8	2,4	15	0,947	177
BOLIVIA	64,2	86,7	64,7	4,0	52	0,865	52
BRASIL	84,2	88,6	71,7	2,3	31	0,883	195
CHILE	87,6	95,7	78,3	2,0	8	0,914	172
COLOMBIA	72,7	92,8	72,3	2,5	17	0,869	104
COSTA RICA	61,7	94,9	78,5	2,3	11	0,876	254
CUBA	75,5	99,8	77,7	1,6	6	0,952	17
ECUADOR	62,8	91,0	74,7	2,8	22	0,858	47
EL SALVADOR	59,8	80,6	71,3	2,9	23	0,772	93
GUATEMALA	47,2	69,1	69,7	4,6	32	0,685	79
HAITÍ	38,8	61,0	59,5	4,0	84	0,542	70
HONDURAS	46,5	80,0	69,4	3,7	31	0,771	36
MEXICO	76,0	91,6	75,6	2,4	22	0,863	181
NICARAGUA	59,0	76,7	71,9	3,0	30	0,747	27
PANAMÁ	70,8	91,9	75,1	2,7	19	0,878	64
PARAGUAY	58,5	93,5	71,3	3,5	20	0,853	34
PERÚ	72,6	87,9	70,7	2,7	23	0,872	164
R. DOMINICANA	66,8	87,0	71,5	3,0	26	0,827	169
URUGUAY	92,0	96,8	75,9	2,2	14	0,942	193
VENEZUELA	93,4	93,0	73,2	2,7	18	0,872	125
Media	69,0	84,7	72,4	2,87	25,2	0,84	112,6
Desviación estándar	15,5	21,3	4,5	0,75	17,2	0,10	70,3

Examinen y comenten el gráfico de dispersión de la fecundidad con la esperanza de vida (Figura 1.9(b)), que muestra las semejanzas y diferencias entre los países y cómo se relacionan las dos variables.

FIGURA 1.9. Ejemplos de los países de América Latina





Examinar los 21 gráficos de dispersión que se pueden construir con las siete variables es muy fastidioso y sería difícil hacer una síntesis de las interpretaciones obtenidas de ellos. Podemos entonces construir un índice que simplifique las siete variables, tal como se hizo en el caso de las dos variables, peso y talla en la Sección 1.3.

La tabla de los datos de los 20 países con las 7 variables define una matriz de datos $X = (x_{ij})$, donde x_{ij} es el valor que toma el país i sobre la variable j, $i = 1, 2, ..., 20$ y $j = 1, 2, ..., 7$. Para cada país i consideramos el vector $\underline{x}_i$ de $\mathbb{R}^7$:

$$\underline{x}_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{i7} \end{pmatrix}.$$

La primera coordenada corresponde al valor que toma el país i sobre una primera variable (Porcentaje de población urbana), la segunda coordenada corresponde al valor que toma el país i sobre una segunda variable (Tasa de alfabetización), etc. Así tenemos 20 puntos en $\mathbb{R}^7$, espacio que llamaremos "espacio de los países" o más generalmente "espacio de los individuos".

De la misma manera, podemos definir el "espacio de las variables", en el cual cada variable es un vector de $\mathbb{R}^{20}$ con una coordenada para cada país:

$$\underline{x}^k = \begin{pmatrix} x_{1k} \\ x_{2k} \\ \vdots \\ x_{20,k} \end{pmatrix}.$$

Buscamos ahora un vector $u = (\alpha_1, \alpha_2, ..., \alpha_7)$ de $\mathbb{R}^7$, tal que $c_i = \sum_{k=1}^{7} \alpha_k x_{ik}$ es el valor del índice para el país i. El índice óptimo se obtiene, como lo hicimos en el caso de dos variables, minimizando la varianza del índice c:

$$Var(c) = \frac{1}{20}\sum_{i=1}^{20}(c_i - \bar{c})^2,$$

donde $\bar{c} = \frac{1}{20}\sum_{i=1}^{20} c_i = \sum_{k=1}^{7} \alpha_k \underline{x}_k$, con $\underline{x}_k = \frac{1}{20}\sum_{i=1}^{20} x_{ik}$.

$$Var(c) = \frac{1}{20}\sum_{i=1}^{20}[\sum_{k=1}^{7} \alpha_k(x_{ik} - \bar{x}_k)]^2 = \frac{1}{20}\sum_{i=1}^{20}\sum_{k=1}^{7} \alpha_k^2(x_{ik} - \bar{x}_k)^2$$
$$+\frac{2}{20}\sum_{i=1}^{20}\sum_{k=1}^{7}\sum_{j=k+1}^{6} \alpha_k\alpha_j(x_{ik} - \bar{x}_k)(x_{ij} - \bar{x}_j).$$

Se deduce que

$$Var(c) = \frac{1}{20}\sum_{k=1}^{20}(c_i - \bar{c})^2 = \sum_{k=1}^{7} \alpha_k^2 Var(\underline{x}^k) + 2\sum_{k<j} \alpha_k\alpha_j Cov(\underline{x}^k, \underline{x}^j). \tag{1.6}$$

Ahora bien, conviene escribir matricialmente la expresión 1.6: $Var(c) = u^t V u$, donde $V = (v_{kj})$ es la matriz de varianzas y covarianzas de las siete variables.

$$v_{kj} = \begin{cases} Var(\underline{x}^k) & si \;\; k = j \\ Cov(\underline{x}^k, \underline{x}^j) & si \;\; k \neq j. \end{cases}$$

Como vimos anteriormente, vamos a "estandarizar " las variables. La matriz V se transforma en la matriz de correlaciones $R = (r_{kj})$, que usaremos de aquí en adelante, la que está dada por:

$$r_{kj} = \begin{cases} 1 & si \;\; k = j \\ \frac{Cov(\underline{x}^k, \underline{x}^j)}{\sqrt{Var(\underline{x}^k)}\sqrt{Var(\underline{x}^j)}} & si \;\; k \neq j. \end{cases}$$

El índice $c_i = u^t \underline{x}_i$ tiene como varianza: $Var(c) = u^t R u$. Como $Var(c)$ es no negativa, se deduce que $u^t R u \geq 0$. La matriz R, además de ser simétrica, es semi-definida positiva (Ver Gil[**6**]), por lo cual la matriz R es diagonalizable, sus valores propios son positivos o nulos y, más aún, existe una base ortonormal de $\mathbb{R}^7$ formada de vectores propios de R.

Sean $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_7 \geq 0$ los siete valores propios de R ordenados de mayor a menor.

Volvamos al problema de optimización:

($\mathcal{P}_1$): Buscar $u \in \mathbb{R}^7$ que maximiza $u^t R u$ sujeto a $\alpha_1^2 + \alpha_2^2 ... + \alpha_7^2 = u^t u = 1$.

Se deja como ejercicio a los lectores probar que la solución es $u = u_1$, el vector propio de R asociado al mayor valor propio λ_1. El vector propio u_1 es el "primer eje principal". Las coordenadas de u_1 nos proporcionan las ponderaciones que se deben aplicar para construir el índice c^1 "primera componente principal" o "primer factor principal". Para el país i, el índice c^1 se calcula como $c_i^1 = \sum_{k=1}^{7} u_{k1} x_{ik} = u_1^t \underline{x}_i$, el que está dado por,

$$c^1 = 0,37 P.urb. + 0,40 Alf. + 0,41 Esp.vida - 0,38 Fecun. - 0,41 M.inf. + 0,42 Educ. + 0,23 Inter. \quad (1.7)$$

donde los coeficientes $(0,37,\ 0,41,\ ...)$ son las coordenadas del vector u_1. La varianza del índice c^1 es igual a λ_1 y la calidad del índice es igual a $100 \times \frac{\lambda_1}{7}$, dado que $\sum_{k=1}^{7} \lambda_k = traza(R) = 7$.

En el ejemplo, el mayor valor propio encontrado es igual a 5,03, y la primera componente principal conserva $100 \times 5,03/7 = 71,8\,\%$ de la variabilidad total.

Observamos que algunas ponderaciones son positivas y otras negativas. Esto nos permite interpretar el índice construido c^1. Por ejemplo, el índice aumenta cuando aumenta el Porcentaje de población urbana, la Tasa de alfabetización, la Esperanza de Vida, el Gasto en educación y el Número de usuarios de Internet, pero disminuye cuando aumentan la Fecundidad y la Tasa de mortalidad infantil. Es razonable, entonces, decir que el índice se comporta como un "indicador de desarrollo". En la matriz de correlaciones (Tabla 1.2), la Fecundidad tiene una correlación positiva con la Mortalidad infantil y estas dos variables son correlacionadas negativamente con las otras cinco variables y estas cinco variables son correlacionadas positivamente entre sí. Vemos, entonces, una cierta consistencia con los signos del índice.

Tabla 1.2. Matriz de correlaciones (Datos PNUD)

	Urbana	Alfabet.	Esp. vida	Fecund.	M. inf.	Educ.	Inter.
Pob. urbana	1,00	0,65	0,62	-0,76	-0,61	0,82	0,53
Alfabetización	0,65	1,00	0,80	-0,59	-0,89	0,91	0,25
Esp. vida	0,62	0,80	1,00	-0,80	-0,94	0,77	0,39
Fecundidad	-0,76	-0,59	-0,80	1,00	0,71	-0,75	-0,46
Mort. infantil	-0,61	-0,89	-0,94	0,71	1,00	-0,80	-0,30
Educación	0,82	0,91	0,77	-0,75	-0,80	1,00	0,36
Internet	0,53	0,25	0,39	-0,46	-0,30	0,36	1,00

Es interesante calcular también las correlaciones del índice con cada una de las siete variables iniciales. De hecho,

Proposición 1.2. *El coeficiente de correlación entre la primera componente principal* c^1 *y la variable inicial* $\underline{x}^k$ *es:*

$$Cor(c^1, \underline{x}^k) = \sqrt{\lambda_1} u_{k1}.$$

Demostración. [4] Para la variable $\underline{x}^k$, $Cor(c^1, \underline{x}^k) = \frac{Cov(c^1, \underline{x}^k)}{\sqrt{Var(c^1) Var(\underline{x}^k)}}$. Tomando en cuenta que las variables $\underline{x}^k$ son de media nula, se deduce que c^1 es de media nula. Además, las variables $\underline{x}^k$ tienen varianza igual a 1 y la varianza de c^1 es igual a λ_1. Se deduce entonces:

$$Cor(c^1, \underline{x}^k) = \frac{Cov(c^1, \underline{x}^k)}{\sqrt{Var(c^1)}\sqrt{Var(\underline{x}^k)}} = \frac{1}{20} \frac{\sum_{i=1}^{20} c_i^1 x_{ik}}{\sqrt{\lambda_1}}$$

$$Cor(c^1, \underline{x}^k) = \frac{\sum_{i=1}^{20} \frac{1}{20} \sum_{j=1}^{7} \alpha_j x_{ij} x_{ik}}{\sqrt{\lambda_1}} = \frac{\sum_{j=1}^{7} \alpha_j \frac{1}{20} \sum_{i=1}^{20} x_{ij} x_{ik}}{\sqrt{\lambda_1}} = \frac{\sum_{j=1}^{7} \alpha_j r_{jk}}{\sqrt{\lambda_1}}, \tag{1.8}$$

donde $r_{jk} = \frac{1}{20} \sum_{i=1}^{20} x_{ij} x_{ik}$ es el coeficiente de correlación entre las variables $\underline{x}^j$ y $\underline{x}^k$ estandarizadas. Finalmente, recordando que $Ru_1 = \lambda_1 u_1$, obtenemos que el numerador de la ecuación (1.8) es el producto de la fila k de R y del vector u_1 y es igual a $\lambda_1 u_{k1}$. Se obtiene entonces

$$Cor(c^1, \underline{x}^k) = \sqrt{\lambda_1} u_{k1}.$$

Este resultado simplifica el cálculo de los coeficientes de correlación para esta situación particular. □

Los coeficientes de correlación entre las variables y la primera componente principal son parecidos, algunos siendo positivos y otros negativos, (Tabla 1.3), a excepción de la variable Internet que tiene una correlación de 0,50.

TABLA 1.3. Correlaciones entre las variables iniciales y la primera componente principal

Porcentaje pob. Urb.	Alfabetización	Esp. vida	Fecundidad	Mortalidad inf.	Educación	Internet
0,84	0,89	0,91	-0,86	-0,91	0,93	0,50

Podemos ordenar los 20 países según la primera componente principal (segunda columna de la Tabla 1.4(a)), que conserva 71,8 % de la varianza. Si encontramos que 71,8 % es poco, podemos buscar completar la primera componente principal con otro índice. Ahora bien, un segundo índice parecido al primero no nos aportará nada nuevo. Conviene, entonces, buscar un nuevo índice, lo más diferente posible de la primera componente principal c^1. Para esto, se impone que el nuevo índice no esté

[4] La demostración es interesante, pero no es indispensable para entender lo que sigue. En una primera lectura se puede obviar.

correlacionado al primero, lo que equivale a que el vector unitario u, que define el nuevo índice, sea ortogonal al eje principal u_1. El problema de optimización para el segundo índice es similar al problema $(\mathcal{P}_1)$ al cual se agrega la condición de ortogonalidad $< u, u_1 >= u^t u_1 = 0$:

$(\mathcal{P}_2)$: Buscar $u \in \mathbb{R}^7$ que maximiza $u^t R u$ sujeto a $u^t u = 1$ y $u^t u_1 = 0$.

La solución del problema $(\mathcal{P}_2)$ se obtiene de manera parecida a la del problema $(\mathcal{P}_1)$, salvo que tenemos dos restricciones en vez de una. El vector propio u_2 de la matriz R asociado al segundo mayor valor propio $\lambda_2 = 0,95$ es solución del problema $(\mathcal{P}_2)$: $Ru_2 = \lambda_2 u_2$. El vector u_2 es el segundo eje principal y el índice asociado es la segunda componente principal c^2 (tercera columna de la Tabla 1.4(a)), que conserva $100 \times \lambda_2/7 = 100 \times 0,95/7 = 13,6\,\%$ de la variabilidad total. Obtenemos así dos nuevas variables, c^1 y c^2, no correlacionadas entre sí, cada una mostrando aspectos diferentes de los datos.

1.4.2 Representación en el espacio de los individuos

Sea P el plano generado por los vectores unitarios u_1 y u_2 en el espacio $\mathbb{R}^7$. El punto $\underline{c}_i = \begin{pmatrix} c_{i1} \\ c_{i2} \end{pmatrix}$ es la proyección ortogonal del punto $\underline{x}_i$ de $\mathbb{R}^7$ sobre el plano P. El gráfico de dispersión de las dos primeras componentes principales c^1 y c^2 representa entonces las proyecciones ortogonales de los países sobre el plano P que conserva la mayor varianza (Figura 1.10(a)). Se observará que el origen $\begin{pmatrix} 0 \\ 0 \end{pmatrix}$ corresponde al punto promedio de la nube de puntos, pues, para construir las dos componentes principales, se centraron las variables.

Como los ejes principales son ortogonales, la varianza conservada en el plano P es la suma de las varianzas sobre cada eje. Se deduce del Teorema de Pitágoras:

$$\frac{1}{90}\sum_{i=1}^{90} ||\underline{x}_i||^2 = \frac{1}{90}\sum_{i=1}^{90}((c_i^1)^2 + (c_i^2)^2) = \frac{1}{90}\sum_{i=1}^{90}(c_i^1)^2 + \frac{1}{90}\sum_{i=1}^{90}(c_i^2)^2 = \lambda_1 + \lambda_2.$$

Se puede calcular, entonces, el porcentaje de la varianza conservada en el plano P:

$$100 \times \frac{\lambda_1 + \lambda_2}{7} = 100 \times \frac{5,03 + 0,95}{7} = 71,8 + 13,6 = 85,4\,\%.$$

Este alto porcentaje permite decir que las distancias entre los países que tenemos en $\mathbb{R}^7$ se deformaron poco cuando se proyectaron sobre el plan P. En consecuencia, el gráfico de dispersión de la Figura 1.10(a) es una foto bastante fiel de la representación de los países en $\mathbb{R}^7$. Es así que, por ejemplo, vemos que Haití está separado del resto de los países; que hay un grupo de países vecinos, como Chile, Argentina y Uruguay, y que Cuba se destaca del resto, pero del lado opuesto de Haití.

Es difícil presentar aquí la tabla de todas las distancias entre los países (es una tabla de 20×20), pero a título indicativo se muestran las distancias entre los países y Chile en $\mathbb{R}^7$ y en P (Tabla 1.4(b)). Observamos que todas las distancias en P son

más pequeñas que las distancias en R^7, pero se conservan bastante bien, salvo en el caso de Costa Rica y Argentina.

Tabla 1.4. Componentes principales y distancias

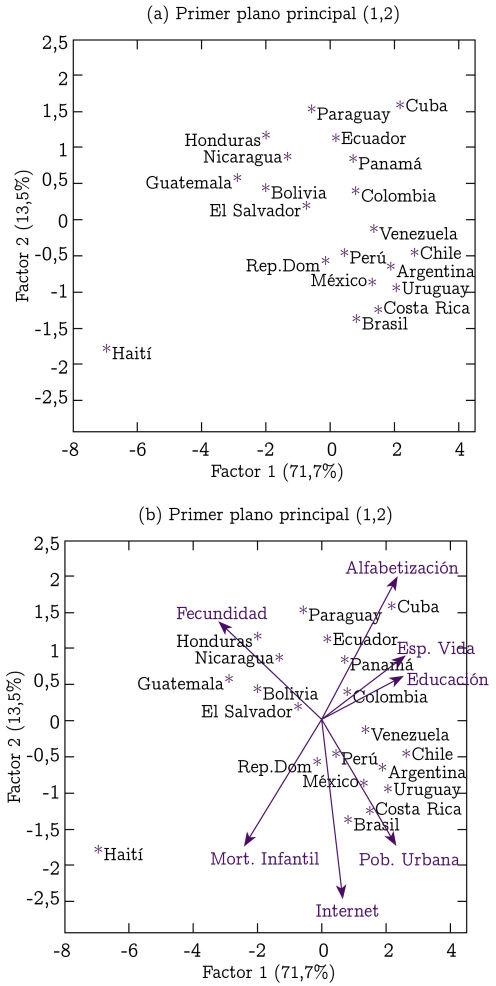
	(a) Componentes principales		(b) Distancias	
País	c^1	$c2$	En espacio $\mathbb{R}^7$	En plano P
Haití	-7,013	-1,797	9,76	9,71
Guatemala	-2,945	0,561	5,90	5,65
Bolivia	-2,121	0,417	5,45	4,81
Honduras	-2,052	1,167	5,05	4,95
Nicaragua	-1,320	0,907	4,20	4,17
El Salvador	-0,732	0,273	3,50	3,43
Paraguay	-0,599	1,485	3,92	3,77
R. Dominicana	-0,053	-0,623	2,88	2,66
Ecuador	0,156	1,135	2,99	2,94
Perú	0,444	-0,645	2,45	2,17
Panamá	0,677	0,864	2,40	2,36
Colombia	0,727	0,255	2,19	2,02
Brasil	1,003	-1,401	2,16	1,84
México	1,248	-0,755	1,52	1,38
Venezuela	1,281	-0,315	1,85	1,34
Costa Rica	1,840	-1,036	2,17	0,93
Argentina	2,156	-0,687	1,12	0,49
Cuba	2,243	1,613	2,51	2,14
Uruguay	2,456	-0,918	0,88	0,44
Chile	2,606	-0,499	0,00	0,0

Ahora, cuando dos países se ven diferentes en el plano P, sería interesante saber en qué difieren. Por ejemplo, Cuba y Uruguay lo hacen de manera notoria, en la segunda componente principal, mientras que Cuba y Haití en ambas componentes. Como las componentes principales son función de las variables iniciales, podemos explicar estas diferencias a partir de ellas usando sus correlaciones con las dos primeras componentes principales. Además de los 20 países, podemos representar las variables iniciales en el plano P también. Cada eje canónico de $\mathbb{R}^7$ corresponde a una variable. Por ejemplo, el primero corresponde al porcentaje de población urbana. Proyectamos, entonces, este eje sobre el plan principal P. El vector unitario de la dirección de esta proyección está dado por la primera coordenada de u_1 y u_2, respectivamente. Se proyectan de esta manera los 7 ejes canónicos. La proyección del eje k tiene como vector director $\begin{pmatrix} u_{k1} \\ u_{k2} \end{pmatrix}$.

Si las distancias entre los países cambian poco, estas direcciones proyectadas (ejes de la Figura 1.10(b)) nos permiten interpretar las similitudes y diferencias entre los países. El gráfico se llama "biplot". Por ejemplo, lo que nos dice el gráfico, es que

Brasil, Costa Rica y Uruguay tienen posiblemente el más alto porcentaje de población urbana, y Paraguay, Haití y Honduras el más bajo. Verifiquen con la Tabla 1.1 que esto es correcto. El costo que se debe pagar por simplificar la información reduciendo los datos de los países en $\mathbb{R}^7$ a 2 dimensiones, se observa claramente en el caso de Venezuela, el que posee el mayor porcentaje de población urbana pero que en el gráfico no queda de manifiesto.

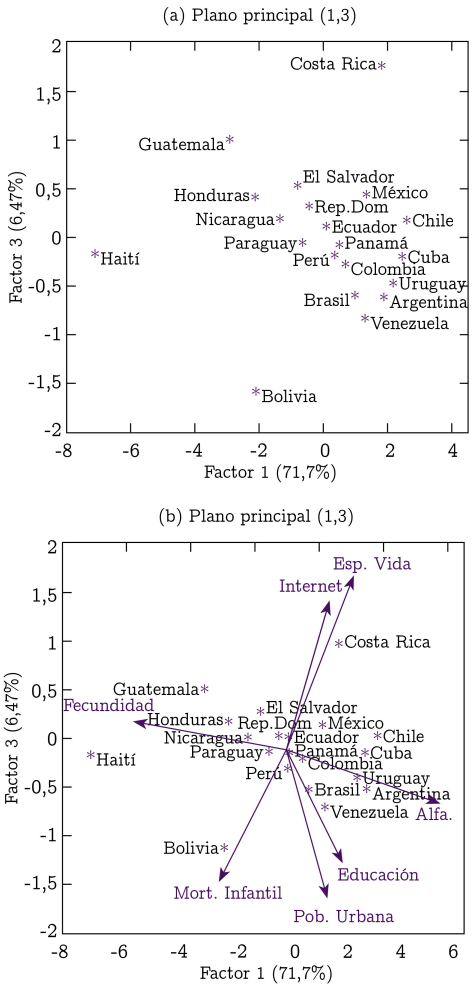
FIGURA 1.10. Representación en el primer plano principal



Podemos buscar una tercera componente principal tomando el vector propio asociado al tercer valor propio $\lambda_3 = 0,45$. El gráfico de dispersión de la primera y de

la tercera componentes (Figuras 1.11(a) y (b)) muestra otros aspectos de los datos. Interpreten este gráfico.

FIGURA 1.11. Primera y tercera componentes principales



1.4.3 Representación en el espacio de las variables

Nos interesamos aquí por las relaciones entre las variables. Vimos anteriormente la matriz de correlaciones (Tabla 1.2). Los coeficientes de correlación lineal nos proporcionan relaciones entre dos variables, pero es difícil ver qué pasa entre tres o más

variables. Por ejemplo, la correlación entre Alfabetización y Mortalidad infantil es negativa (-0,89) y la correlación entre Mortalidad infantil y Esperanza de vida es negativa también (-0,94). ¿Podemos deducir la correlación entre Esperanza de vida y Alfabetización? Es positiva (0,80). La correlación entre Alfabetización y Porcentaje de población urbana es igual a 0,65 y entre Porcentaje de población urbana e Internet es igual a 0,53; sin embargo, entre Alfabetización e Internet es pequeña (0,25). Vamos a tratar de mostrar estas relaciones gráficamente.

Vimos cómo estas variables se relacionan con las componentes principales, mediante los coeficientes de correlación entre variables antiguas y las nuevas. En la Tabla 1.5 presentamos las correlaciones con las tres primeras componentes principales.

TABLA 1.5. Correlaciones entre las variables iniciales y c^1 y c^2

	Pob. urbana	Alfabetización	Esp. vida.	Fecundidad	Mortalidad inf.	Educación	Internet
c^1	0,84	0,89	0,91	-0,86	-0,91	0,93	0,50
c^2	-0,27	0,32	0,14	0,16	-0,28	0,11	-0,80
c^3	0,41	0,04	-032	-0,02	0,24	0,25	-0,24

Podemos interpretar las componentes principales:

- La Fecundidad y la Mortalidad están altamente correlacionadas negativamente con la primera componente principal.
- La Alfabetización, Esperanza de vida, Educación y Porcentaje de población urbana están altamente correlacionadas positivamente con la primera componente principal.
- El número de usuarios de Internet es la única variable altamente correlacionada (negativamente) con la segunda componente principal.

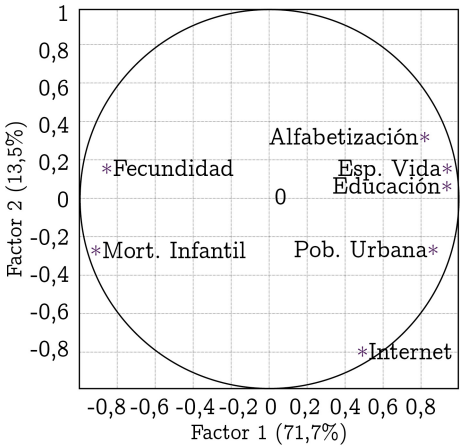
En resumen, se podría decir que la primera componente principal es un indicador demográfico y la segunda, un indicador de desarrollo tecnológico.

Representamos, ahora, en un gráfico de dispersión las variables iniciales, tomando como abscisa la primera componente principal y como ordenada la segunda componente principal. Se usan entonces como coordenadas los coeficientes de correlación lineal de la Tabla 1.5 obteniendo la Figura 1.12(a). Este gráfico permite interpretar las componentes principales en términos de las variables iniciales y representar de manera aproximada las correlaciones entre las variables iniciales. Además se dibujó el círculo $\mathcal{C}$ centrado en el origen de radio 1.

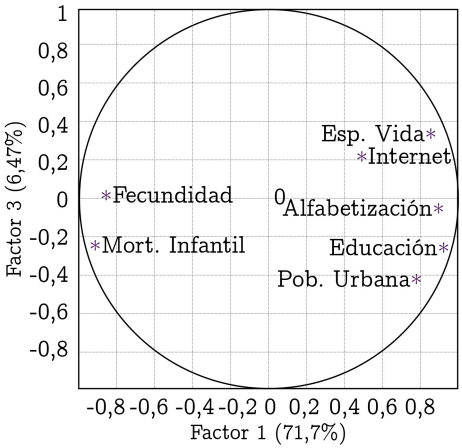
Observamos que los puntos-variables cercanos tienen correlaciones altas y positivas. Los opuestos, tales como Mortalidad infantil y Esperanza de vida, tienen una correlación alta negativa. Finalmente, Internet y Alfabetización, que tienen una correlación pequeña, forman un ángulo recto. Además, las siete variables se encuentran cercanas a la circunferencia del círculo $\mathcal{C}$. En la Figura 1.12(b), las variables son representadas sobre la primera y la tercera componentes principales. Veamos ahora cómo interpretar estos gráficos.

FIGURA 1.12. Círculos de correlaciones

(a) Círculo de correlación (1,2)



(b) Círculo de correlación (1,3)



La nube de los siete puntos $\underline{x}^1, \underline{x}^2, ..., \underline{x}^7$, cuyas coordenadas son los valores que toman los países sobre las siete variables, puede representarse en el espacio R^{20}. Veamos, en primer lugar, una interpretación geométrica del coeficiente de correlación lineal. Como las variables son centradas y reducidas (media 0 y varianza 1), los vectores que representan las variables $\underline{x}^k$ son de norma 1: $||\underline{x}^k|| = 1$[5], lo que ubica

[5]Se calculan las normas y los productos escalares usando ponderaciones $\frac{1}{20}$ para cada país: $||\underline{x}^k||^2 = \frac{1}{20}\sum_i x_{ik}^2$ y el producto escalar entre $\underline{x}^k$ y $\underline{x}^j$: $< \underline{x}^k, \underline{x}^j >= \frac{1}{20}\sum_i x_{ik}x_{ij}$.

los siete puntos sobre la hiperesfera de $\mathbb{R}^{20}$ de radio 1 centrada en el origen (Figura 1.13(a)). Además, el coeficiente de correlación lineal entre dos variables $\underline{x}^j$ y $\underline{x}^k$ es $r_{jk} = \frac{1}{20}\sum_{i=1}^{20} x_{ij}x_{ik} = \frac{1}{20}(\underline{x}^j)^t\underline{x}^k$, que es igual al producto escalar entre $\underline{x}^k$ y $\underline{x}^j$, que a su vez es igual al coseno del ángulo θ formado por los dos vectores.

Se considera ahora P_c el plano formado por los dos vectores componentes principales c^1 y c^2 de $\mathbb{R}^{20}$ (Figura 1.13(a)) y proyectamos los vectores $\underline{x}^k$ sobre este plano. La proyección $\underline{x}_o^k$ de $\underline{x}^k$ está necesariamente al interior del círculo $\mathcal{C}$ de radio 1 centrado en el origen del plano P_c, y $\underline{x}_o^k$ estará sobre la circunferencia del círculo $\mathcal{C}$ si y solo si $\underline{x}^k$ está contenido en P_c, o sea, $\underline{x}_o^k = \underline{x}^k$. Tenemos varios comentarios sobre este gráfico llamado "círculo de correlación" (Figura 1.13(b)):

- Sean a_k y b_k las coordenadas de la proyección $\underline{x}_o^k$ de $\underline{x}^k$ sobre P_c, donde a_k es la coordenada sobre c^1 y b_k la coordenada sobre c^2. La coordenada a_k es el producto escalar de $\underline{x}^k$ con c^1 dividido por la norma de c^1. Luego, es igual al coeficiente de correlación lineal entre $\underline{x}^k$ con c^1. De manera similar, encontramos que b_k es igual al coeficiente de correlación lineal entre $\underline{x}^k$ y c^2. Además, vimos anteriormente que $Cor(\underline{x}^k, c^1) = \sqrt{\lambda_1}u_{k1}$ y $Cor(\underline{x}^k, c^2) = \sqrt{\lambda_2}u_{k2}$. Luego, $a_k = \sqrt{\lambda_1}u_{k1}$ y $b_k = \sqrt{\lambda_2}u_{k2}$.
- Si la proyección $\underline{x}_o^k$ del punto $\underline{x}^k$ aparece cerca de la circunferencia del círculo $\mathcal{C}$, significa que $\underline{x}^k$ es cercano del plano P_c y, por lo tanto, que la variable correspondiente a $\underline{x}^k$ está bien "representada" por las dos componentes principales c^1 y c^2.
- Las proyecciones $\underline{x}_o^k$ y $\underline{x}_o^j$ de los puntos $\underline{x}^k$ y $\underline{x}^j$ sobre P_c forman un ángulo θ_o que es distinto del ángulo θ formado por los puntos $\underline{x}^k$ y $\underline{x}^j$, salvo si estos pertenecen al plano P_c. Sin embargo, si los puntos $\underline{x}^k$ y $\underline{x}^j$ no están en el plano P_c pero son cercanos a la circunferencia del círculo $\mathcal{C}$, el ángulo θ_o diferirá poco del ángulo θ. En este caso, el coseno del ángulo θ^o es una buena estimación del coeficiente de correlación entre $\underline{x}^k$ y $\underline{x}^j$.

Vimos que la representatividad global de las variables a partir de las dos componentes principales es de 85,4 %, que podríamos considerar como buena. En realidad, decir que es bueno o no es relativo y quizás subjetivo. Depende de la complejidad inicial de los datos y de cuánto pudimos descubrir de esta complejidad en el plano de las dos primeras componentes principales.

Las siete variables no necesariamente están todas bien representadas en el plano. La calidad de representación de la variable $\underline{x}^k$ a partir de las dos componentes principales es:

$$\sqrt{cor(\underline{x}^k, c^1)^2 + cor(\underline{x}^k, c^2)^2},$$

que es el largo de la proyección de la variable sobre el plano (Tabla 1.6). Podemos decir entonces que si la proyección de una variable es cercana a la circunferencia $\mathcal{C}$, entonces está bien representada para las dos componentes principales.

FIGURA 1.13. Espacio de las variables

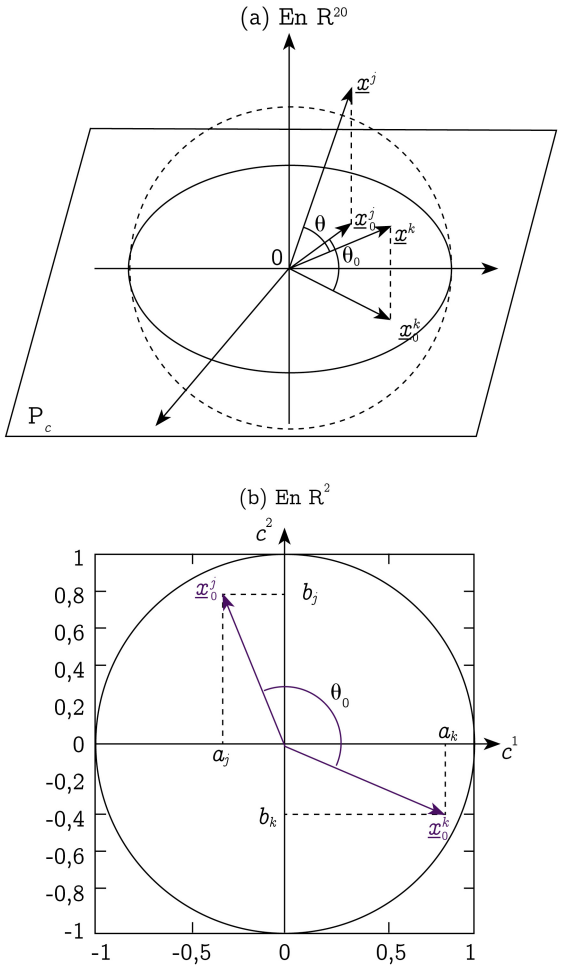


TABLA 1.6. Representatividad de las variables en el plano

Porcentaje pob. urbana	Alfabetización	Esp. vida	Fecundidad	Mortalidad inf.	Educación	Internet
0,88	0,94	0,92	0,88	0,95	0,94	0,94

El círculo de correlaciones permite interpretar las componentes principales en términos de las variables iniciales y representar geométricamente la matriz de correlaciones R:

- La Fecundidad y la Mortalidad infantil forman un ángulo pequeño; vale decir, tienen una alta correlación positiva entre sí.
- La Fecundidad y el Porcentaje de población urbana forman un ángulo cercano a π; tienen una alta correlación negativa entre sí;
- La Mortalidad Infantil y la Alfabetización forman un ángulo cercano a π; tienen una alta correlación negativa entre sí.
- Internet es la única variable altamente correlacionada con la segunda componente principal.
- La Fecundidad y la Mortalidad tienen una alta correlación negativa con la primera componente principal.
- La Alfabetización, Esperanza de vida, Educación y Porcentaje de población urbana tienen una alta correlación positiva con la primera componente principal.

1.5 Puntos suplementarios

A veces es interesante agregar en el estudio individuos o variables, que no formen parte de la construcción de las componentes principales.

1.5.1 Individuos

Queremos comparar países, como Túnez o Francia con los de América Latina. Son dos nuevos países que no forman parte de América Latina, entonces, es conveniente no considerarlos para construir el plano principal, sino proyectarlos a posteriori en el plano correspondientes a los 20 países de América Latina para que no afecten los ejes principales. Las proyecciones se calculan de la siguiente manera (Tabla 1.7):

(a) Se calculan las medias y desviaciones estándares de las siete variables sobre los 20 países de América Latina.

(b) Se estandarizan los valores de las siete variables de los dos nuevos países. Por ejemplo, para la Esperanza de vida de Túnez es igual a 73,5 y la media y la desviación estándar de la misma variable para los 20 países de América Latina son 72,4 y 4,5 respectivamente. Entonces el valor de la Esperanza de vida estandarizado para Túnez será $\frac{73,5-72,4}{4,5} = 2,4$. Note que los valores negativos de los datos estandarizados son los que están debajo del promedio.

(c) Se calculan los productos escalares de los valores estandarizados con los vectores propios u_1 y u_2, respectivamente (ver la ecuación (1.7)). Para Túnez, la coordenada sobre el primer eje principal es: $-0,37 \times 3,80 - 0,40 \times 0,49 + 0,41 \times 0,24 + 0,38 \times 1,15 + 0,41 \times 0,30 - 0,42 \times 0,92 - 0,23 \times 0,25 = -1,39$. Sobre el segundo eje: $0,28 \times 3,80 - 0,33 \times 0,49 + 0,14 \times 0,24 - 0,16 \times 1,15 + 0,29 \times 0,30 - 0,11 \times 0,92 + 0,82 \times 0,25 = 0,94$.

Podemos ahora comparar los nuevos países con los de América Latina (Figura 1.14(a)). Túnez es parecido a Nicaragua, y Francia se diferencia mucho de los otros países. Analicen.

1.5.2 Variables

Ahora queremos relacionar el Producto Nacional Bruto (PNB) con las siete variables demográficas. Podemos calcular la correlación del PNB con cada una de las siete variables e interpretarlas. Podemos también representar el PNB en el círculo de correlaciones. Basta entonces calcular el coeficiente de correlación entre el PNB con cada una de las dos componentes principales (Tabla 1.8). En el círculo de correlaciones, el PNB tiene como coordenadas estos dos coeficientes de correlación (Figura 1.14(b)). Se puede hacer lo mismo con los resultados de la prueba TIMSS (Trends in International Mathematics and Science Study). Interpreten el gráfico.

Tabla 1.7. Datos de los nuevos países

	Porcentaje pob. urbana	Alfabetización	Esp. de vida	Fecundidad	Mortal. infantil	Educación	Internet
Media	69,0	84,7	72,4	2,87	25,2	0,84	112,6
Desv. estándar	15,5	21,3	4,5	0,75	17,2	0,10	70,3
Túnez	10,1	74,3	73,5	2,0	20,0	0,75	95,0
Francia	61,0	100,0	80,2	1,9	4,0	0,98	430,0
Túnez estandarizado	-3,80	-0,49	0,24	-1,15	-0,30	-0,92	-0,25
Francia estandarizado	-0,52	0,71	1,72	-1,28	-1,23	1,47	4,52

Tabla 1.8. Correlaciones de las nuevas variables

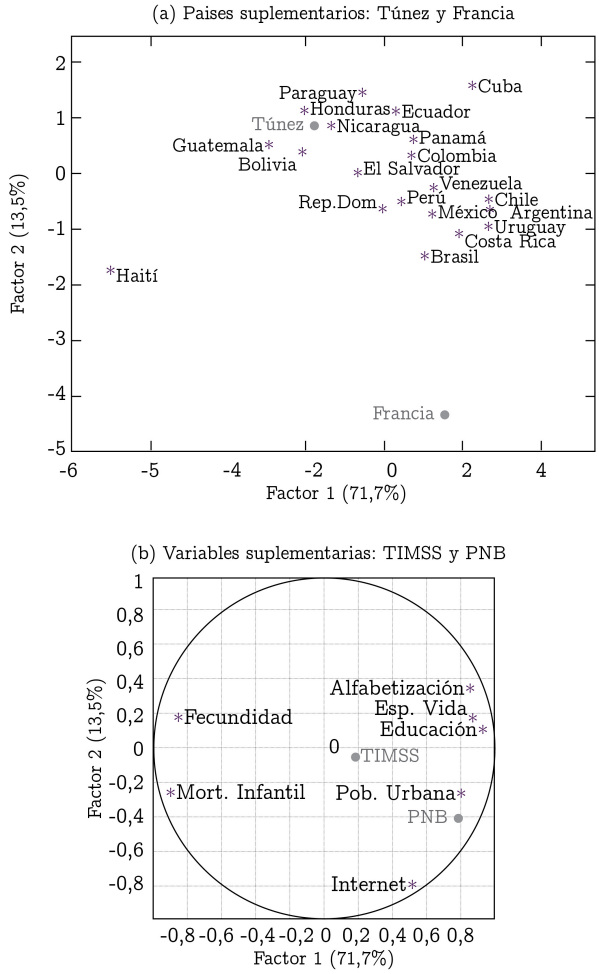
Variable	c^1	c^2
PNB	0,78	-0,43
TIMSS	0,17	-0,07

1.6 Análisis de la PSU con componentes principales

El ingreso a las universidades del Consejo de rectores toma en cuenta la Prueba de Selección Universitaria (PSU) y el promedio de las nota de Enseñanza Media (NEM). La PSU esta compuesta por cuatro tipos de prueba: matemática, lenguaje, historia y ciencia, que tienen diferente número de preguntas. Para que las cuatro pruebas sean comparables, los puntajes de cada una son transformados para obtener una distribución Normal de media 500 y desviación estándar 110 ($\mathcal{N}(500, 110)$). Se aplica el mismo tipo de transformación a la NEM. De esta manera, los puntajes de las cuatro pruebas y de la NEM tienen la misma escala de medición.

Todos los postulantes no rinden necesariamente las cuatro pruebas. Por ejemplo, en las carreras de ingeniería y medicina, la prueba de historia no es obligatoria, y en

FIGURA 1.14. Puntos suplementarios



derecho, la prueba de ciencia tampoco. Sin embargo, los postulantes pueden rendir las cuatro pruebas. En este estudio, para tener un conjunto de postulantes más homogéneos, nos limitamos a tomar a los que rindieron las cuatro pruebas, que provienen de colegios Humanistas-Cientificos y que egresaron de la Enseñanza Media el mismo año que rinden la PSU. De la PSU 2009, que analizamos aquí, se consideraron alrededor de 40.000 postulantes.

Tenemos solamente cinco variables, pero muchos individuos. Resultará, entonces, difícil representar a los postulantes en los gráficos de dispersión, además de que esta

información individual no es muy interesante. Ahora bien, datos complementarios, tales como el género, la dependencia y la región del colegio, son variables cualitativas que también pueden ser utilizadas. Buscaremos no solamente ver como se relacionan las cinco notas, sino también cómo las tres variables cualitativas se relacionan con las notas entre sí, lo que es más interesante que analizar a los postulantes individualmente.

Aplicamos en primer lugar un ACP sobre los cinco puntajes de los 40.000 postulantes. Podremos mostrar el círculo de correlaciones de las dos primeras componentes principales, pero como no sería realista tratar de representar a todos los postulantes en un gráfico, será mucho más práctico y, sobre todo, interesante hacerlo con grupos de postulantes definidos por el género, la dependencia y la región del colegio, para detectar si estas variables cualitativas influyen sobre los puntajes de la PSU.

Los resultados del ACP sobre los cinco puntajes se muestran en las Tablas 1.9 y 1.10. Estudien las dos tablas y, en particular, observen:

- Se tiene cinco valores y vectores propios.
- Los vectores propios de la Tabla 1.9 son de norma 1.
- Los vectores-columnas de correlaciones de la Tabla 1.10 son de norma igual a la raíz del valor propio correspondiente.
- Los vectores-filas de correlaciones de la Tabla 1.10 son de norma 1.

Tabla 1.9. Valores y vectores propios

Variable	1	2	3	4	5
Valor propio	3,747	0,546	0,366	0,186	0,156
Porcentaje	74,9	10,9	7,3	3,7	3,1
Porcentaje acumulado	74,9	85,8	93,2	96,9	100,0
Vector propio	1	2	3	4	5
NEM	0.3842	0,8702	0,2974	-0,0812	0,0131
Matemática	0,4694	0,0043	-0,4773	0,4212	-0,6119
Lenguaje	0,4765	-0,2097	0,1180	0,5447	0,6468
Historia	0,4350	-0,4339	0,6517	-0,2628	-0,3587
Ciencias	0,4645	-0,1028	-0,4950	-0,6710	0,2799
NORMA del vector propio	1	1	1	1	1

Se tomó una muestra aleatoria de 150 postulantes para realizar el gráfico del primer plano principal (Figura 1.15). Podemos decir que los postulantes a la derecha tienen buenos puntajes en todas las pruebas, pero que los que se encuentran en la parte superior tienen una buena NEM comparado con los cuatro puntajes de la PSU. Las correlaciones sobre las dos primeras C.P. (Figura 1.16(a)) y las C.P. 1 y 3 (Figura 1.16(b)) permiten entender mejor las relaciones entre las cinco notas dadas en la Tabla

Tabla 1.10. Correlación entre variables y componentes principales

Variable	C.P. 1	C.P. 2	C.P. 3	C.P. 4	C.P. 5	Norma
NEM	0,743	0,643	0,179	-0,035	0,005	1
Matemática	0,908	0,003	-0,289	0,181	-0,241	1
Lenguaje	0,922	-0,155	0,071	0,235	0,254	1
Historia	0,842	-0,320	0.,394	-0,113	-0,141	1
Ciencias	0,899	-0,076	-0,299	-0,289	0,110	1
Valor propio	3,747	0,546	0,366	0,186	0,156	
Norma del vector de correlaciones	1,94	0,73	0,60	0,43	0,39	

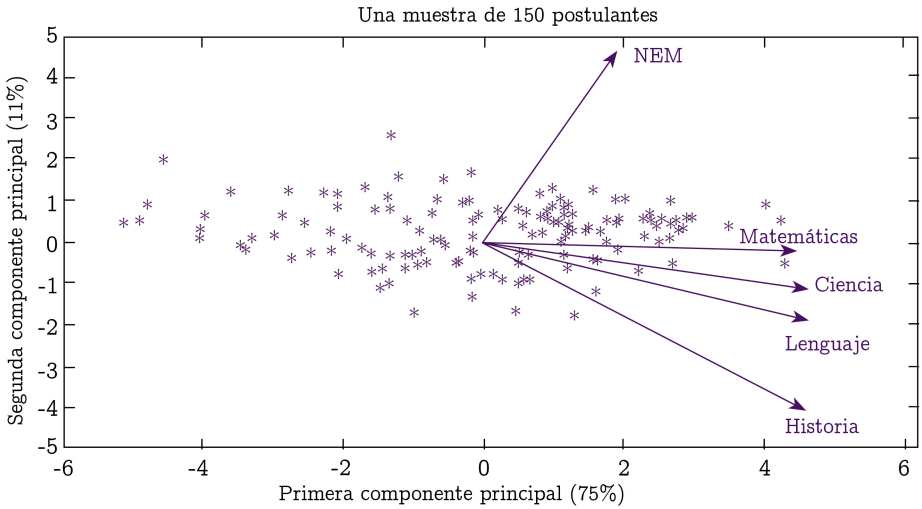
1.10. La primera C.P., que conserva el 75 % de la varianza total, podemos interpretarla como un índice de buenos resultados en general; la segunda C.P., que conserva de 11 % de la varianza total, opone la NEM a las pruebas PSU, especialmente la prueba de Historia y la tercera C.P. opone las pruebas científicas (Matemática y Ciencia) a los otros puntajes. Observen que la distancia al origen de los puntos de la Figura 1.16(a) es igual a la raíz de la suma de los cuadrados de las correlaciones con las dos primeras C.P. (Ver Tabla 1.11). Verifiquen con la Tabla 1.10, que la distancia al origen de los puntos de la Figura 1.16(b) es igual a la raíz de la suma de los cuadrados de las correlaciones con las C.P. 1 y 3.

En general, los puntajes de Matemática y Ciencia están muy correlacionados, lo que significa que si un postulante tiene un alto puntaje en Matemática, también lo tiene en Ciencia. Lo mismo para Lenguaje e Historia. La NEM, que es la nota más correlacionada con la segunda C.P., tiene un comportamiento un poco distinto. Trataremos de explicar esto más adelante.

Tabla 1.11. Correlación con las dos primeras componentes principales

VARIABLE	C.P. 1	C.P. 2	Norma
NEM	0,743	0,643	0,983
Matemática	0,908	0,003	0,908
Lenguaje	0,922	-0,155	0,935
Historia	0,842	-0,320	0,900
Ciencias	0,899	-0,076	0,902

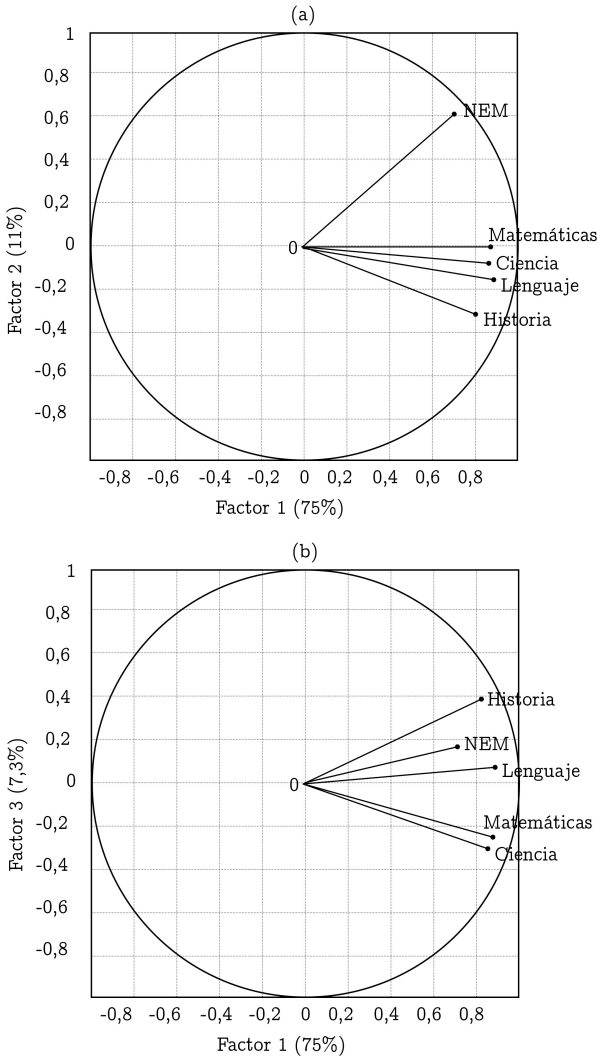
FIGURA 1.15. Plano principal de los cinco puntajes



Vimos que el gráfico de la Figura 1.15 no presenta especial interés, dado que muestra solamente 150 postulantes y no aporta más información que el círculo de correlaciones. Es ahí donde podemos usar las variables que caracterizan a los postulantes: género, dependencia y región del colegio.

Calculamos los promedios de las cuatro pruebas de la PSU y de la NEM para los grupos definidos por las tres variables cualitativas (Tabla 1.12). Obtenemos 20 grupos. ¡OJO!, los grupos no son disjuntos. Entonces, en vez de mostrar postulantes en el plano principal, podemos representar estos promedios como puntos-individuos suplementarios sobre los ejes principales. Tenemos dos puntos para el genero (H para Hombre y M para Mujer), tres para la dependencia del colegio (PP para particulares pagados, PS para particulares subvencionados y Mu para municipales) y 15 para la región (Figura 1.17). Podemos ver que en promedio les va mejor a los colegios particulares pagados. Los colegios particulares subvencionados tienen resultados regulares, ya que el promedio se encuentra cerca del origen, y los colegios municipales tienen resultados bajos. Entre los géneros hay menos diferencia, pero a los hombres les va mejor. Finalmente, se destaca la región RM (13). La región 15 tiene alto puntaje en la NEM y bajo en la PSU de Historia, lo que explica su posición. Completen la interpretación de la tabla y del gráfico.

FIGURA 1.16. Círculos de correlaciones de los cinco puntajes



Finalmente, calculamos los promedios de los cinco puntajes formando grupos disjuntos al cruzar las tres variables cualitativas. Obtenemos 86 grupos, siendo 4 grupos, de los 90, vacíos. Es así, por ejemplo, que tenemos el grupo "MuM3" formado de las mujeres de la tercera región de los colegios municipales, o el grupo "PPH9", formados por los hombres de la novena región de colegios particulares pagados. En la

Tabla 1.12. Promedios por grupo

Puntaje	NEM	Matemática	Lenguaje	Historia	Ciencia
Mu	566	553	525	519	496
PS	576	592	556	543	524
PP	627	688	639	611	608
H	577	608	568	559	543
M	606	593	563	535	519
1	586	606	549	529	510
2	573	594	557	537	526
3	579	567	526	507	490
4	573	566	528	525	498
5	587	597	562	543	525
6	576	571	539	534	507
7	587	588	553	539	518
8	561	565	529	521	504
9	565	589	551	545	523
10	603	617	574	565	549
11	594	611	582	594	554
12	616	616	571	564	513
13	593	636	599	579	569
14	590	554	518	526	505
15	619	582	528	497	513

Figura 1.18 se encuentran los 86 grupos en el primer plano principal, y en la Figura 1.19, los grupos en el plano principal (1,3). Se omitieron los ejes que representaban a los cinco puntajes, ya que son los mismos que en los otros gráficos. De estos dos gráficos pueden sacar conclusiones. Por ejemplo, que una alta nota NEM no asegura puntajes altos en la PSU (obviamente da cierta ventaja en el puntaje de ingreso); que hay diferencias en los resultados entre las regiones; que los colegios PP obtienen mejores puntajes (¡lo que no es ninguna novedad!).

FIGURA 1.17. ACP PSU: promedios de los tres tipos de grupos

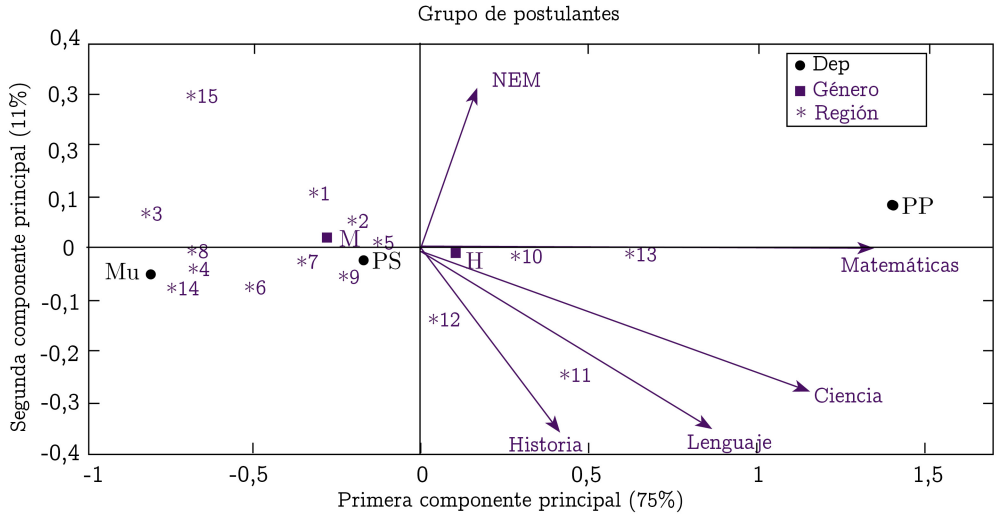


FIGURA 1.18. PSU: Factores 1 y 2 de los 86 grupos

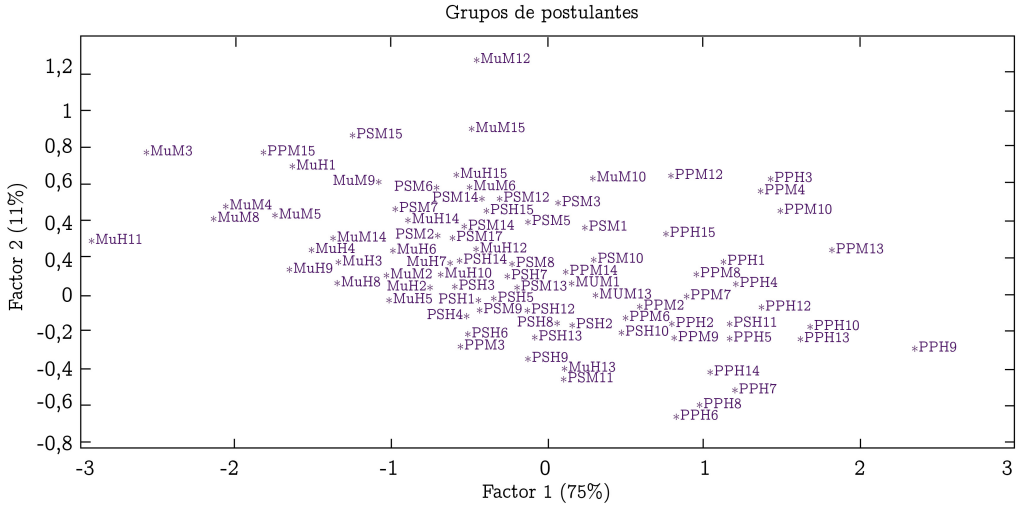
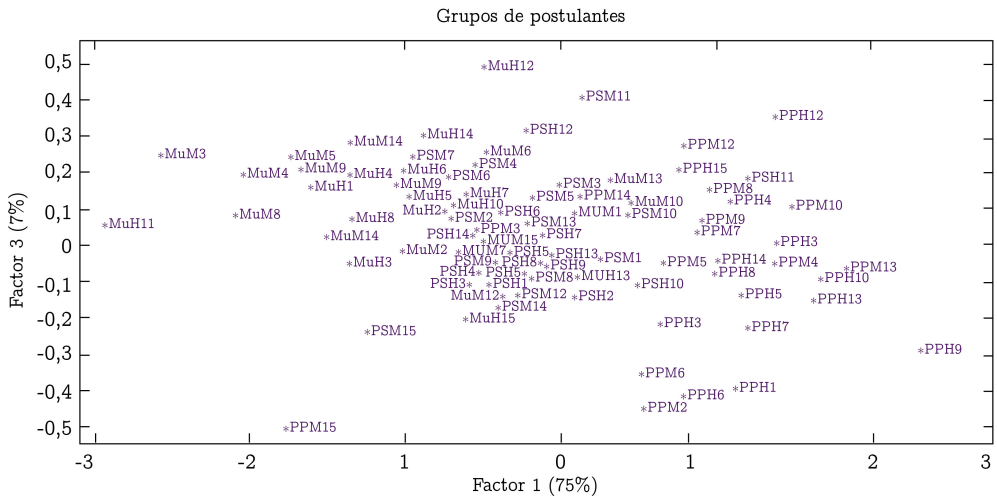


FIGURA 1.19. PSU: Factores 1 y 3 de los 86 grupos



1.7 Resumen de la terminología

Índice: Reduce a un valor único un conjunto de valores.
a^t: Designa la traspuesta de a.
Gráfico de dispersión: Gráfico de puntos en $\mathbb{R}^2$.
Matriz de correlaciones: Matriz simétrica construida a partir de las correlaciones entre variables medidas sobre los mismos individuos.
Espacio de los individuos: Espacio vectorial en el que los puntos son los individuos y los ejes representan a las variables.
Espacio de las variables: El espacio vectorial en el que los puntos son las variables y los ejes representan a los individuos.

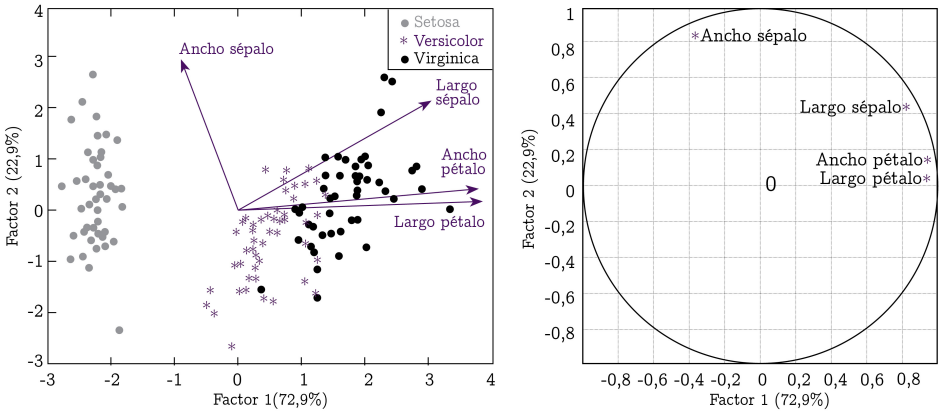
Inercia: Suma de los cuadrados de las distancias de una nube de puntos a un punto dado.
Eje principal: El eje definido por un vector propio de la matriz de covarianza o de correlaciones, en el caso de la estadística y de la matriz de inercia, en el caso de la mecánica sólida.
Componente principal: Variable definida por un eje principal.
Factor principal: Componente principal.
Contribución a la varianza de una variable: Porcentaje de varianza conservada por la variable.
Círculo de correlaciones: Representaciones de las variables iniciales sobre las componentes principales.

1.8 Ejercicios

Ejercicio 1.1. Se efectua el ACP de los datos de 150 flores con cuatro mediciones: ancho y largo del sépalo y ancho y largo del pétalo (Ejemplo debido a R. FIsher). Se busca saber cuáles de las variables permiten distinguir las tres especies de flores: Setosa, Versicolor y Virginica, observando la Figura 1.20.

(a) ¿Cuál es el porcentaje de conservación de la varianza en el plano principal?
(b) ¿Se distinguen las especies en el primer plano principal? ¿Cuáles son las variables que las diferencian?
(c) ¿Qué variables explican la primera componente principal? ¿Qué variables explican la segunda componente principal?
(d) ¿Cuáles son las mediciones más correlacionadas? ¿Cuáles son las menos correlacionadas?

FIGURA 1.20. ACP y círculo de correlación de las flores



Ejercicio 1.2. Consideren la matriz de correlación $R = \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}$.

(a) Encuentren la expresión de los valores propios de R.
(b) Calculen los vectores propios de norma 1.
(c) Construyan el círculo de correlaciones $\mathcal{C}$.
(d) Den el porcentaje de conservación de la varianza sobre la primera componente principal.
(e) Estudien el caso $r = 0$.

Ejercicio 1.3. En la tabla siguiente se muestran los valores de cinco variables obtenidas en 20 alumnos que quieren entrar a alguna universidad del consejo de rectores. Las variables en estudio son la distancia en kilómetros al lugar del colegio en el que estudiaban (DIST), el promedio de horas que hacían actividad física a la semana

(EF), índice de masa corporal (IMC), IQ (coeficiente intelectual) y NEM (promedio de notas con el cual postulan a las universidades). Se quiere determinar las relaciones existentes entre dichas variables intentando reducir la dimensionalidad del problema vía un análisis en componentes principales.

Individuo	Distancia $[km]$	EF $[hrs]$	IMC	IQ	NEM
1	0,1413	0,8773	25,7906	88,1222	5,2000
2	3,1442	0,4168	31,6282	77,9768	5,1000
3	0,3070	0,1311	25,0718	109,8634	6,5000
4	0,7752	2,1935	35,5118	94,8136	6,3000
5	0,1473	0,2609	24,7085	103,2737	6,8000
6	0,5500	0,0106	22,1104	102,3406	5,8000
7	0,2746	1,5676	15,9921	100,2147	5,5000
8	0,6772	1,7244	31,1364	89,9606	6,0000
9	0,9510	0,1203	23,0603	90,5285	5,3000
10	1,3300	0,7332	20,5239	96,2557	4,8000
11	1,3143	0,1800	24,2057	88,1411	5,6000
12	0,3059	0,2488	36,0617	89,4410	5,7000
13	0,9558	0,0251	26,2138	114,7248	7,0000
14	0,4905	0,0793	25,7133	100,5574	6,5000
15	1,5131	1,6264	29,9394	87,8268	6,0000
16	0,2877	0,0668	22,2821	99,5877	5,2000
17	0,7775	0,1744	26,7577	88,7166	5,5000
18	0,1206	6,5881	22,7429	86,5072	5,7000
19	0,1266	1,2532	20,9371	97,3890	5,9000
20	0,4173	0,0742	35,5664	109,5347	6,3000

Tabla 1.13. Matriz de correlaciones

	Distancia	EF	IMC	IQ	NEM
Distancia	1,0000	-0,1736	0,2433	-0,4705	-0,3113
EF	-0,1736	1,0000	-0,0822	-0,3397	-0,0691
IMC	0,2433	-0,0822	1,0000	-0,1683	0,2695
IQ	-0,4705	-0,3397	-0,1683	1,0000	0,6494
NEM	-0,3113	-0,0691	0,2695	0,6494	1,0000

(a) Analicen la matriz de correlaciones.

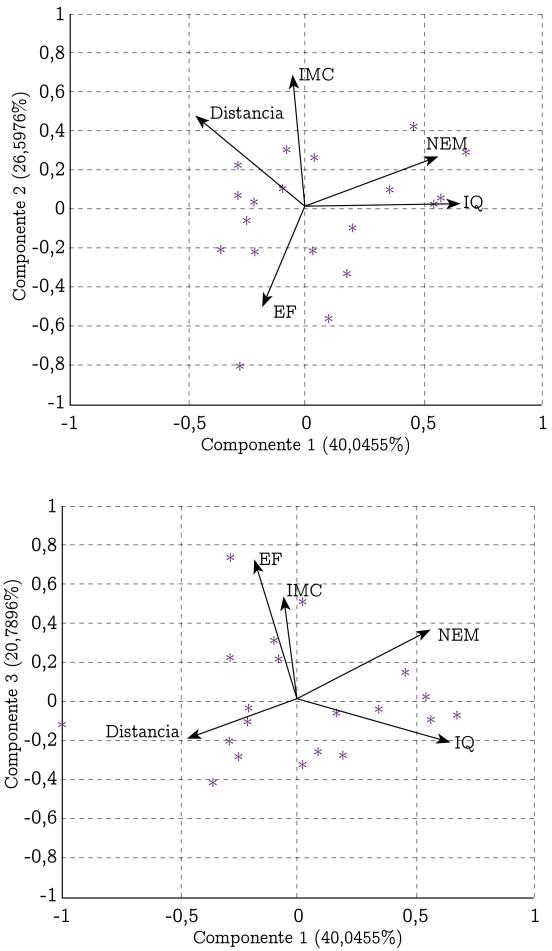
(b) Los valores propios de la matriz de correlaciones son de mayor a menor:

$$2,0023 \ - \ 1,3299 \ - \ 1,0494 \ - \ 0,4583 \ - \ 0,1601$$

Determinen el porcentaje acumulado de varianza, explicado por las tres primeras componentes principales.

(c) Los resultados del ACP se muestran en la figura (4.3). Interpreten los resultados.

FIGURA 1.21. ACP de postulantes a las universidades



Ejercicio 1.4. En la tabla siguiente se adjuntan los valores de 6 variables relacionadas con el estado nutricional de 22 alumnos de entre los 14 y 16 años, en un determinado colegio de Santiago. Las variables medidas son peso en kg, talla en cm, índice de masa corporal (IMC), perímetro de cintura, perímetro de cadera y porcentaje de grasa corporal. Se busca reducir la dimensionalidad del problema realizando un análisis en componentes principales.

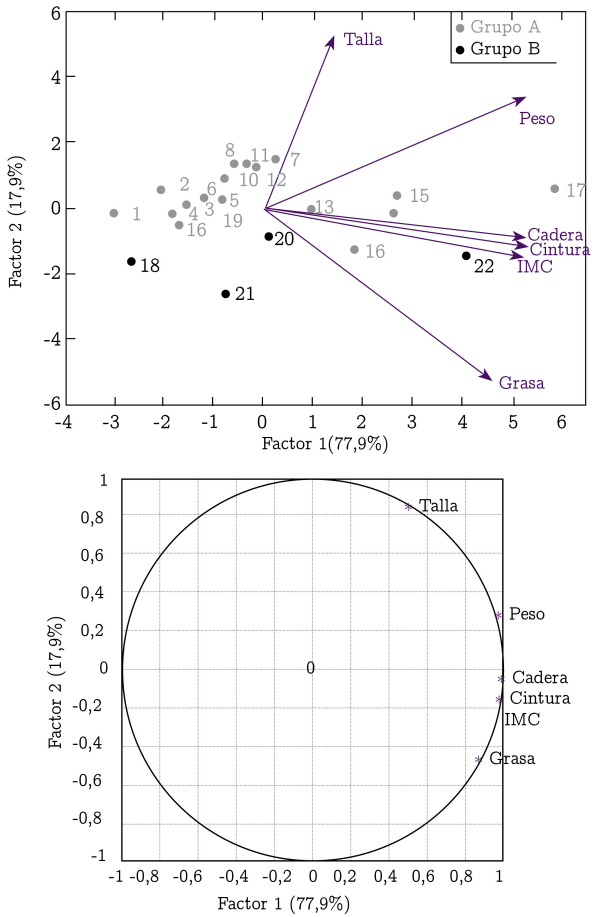
(a) Analicen la matriz de correlaciones(Tabla 1.14).
(b) Los valores propios de la matriz de correlaciones son, de mayor a menor: $4,675$, $1,080$, $0,129$, $0,068$, $0,046$ y $0,002$. Determinen el porcentaje acumulado de varianza explicada por las dos primeras componentes principales.
(c) El círculo de correlaciones se muestra en la figura (1.22(b)). Interpreten los resultados.
(d) Los resultados del ACP se muestran en la figura (1.22(a)). Interpreten los resultados. ¿Qué pueden decir de los 2 grupos presentes en la muestra?

Alumno	Peso [kg]	Talla [cm]	IMC	P. Cintura [cm]	P. Cadera [cm]	% Grasa	GRUPO
1	41,8	154,6	17,5	66,0	76,0	15,65	B
2	48,1	162,0	18,3	69,5	79,5	16,29	B
3	49,5	159,0	19,6	72,2	81,8	17,52	B
4	48,2	156,5	19,7	69,5	81,0	17,85	B
5	55,0	162,0	21,0	72,5	83,0	19,72	B
6	52,3	160,6	20,3	73,0	83,8	17,32	B
7	60,5	175,0	19,8	78,0	89,0	21,01	B
8	58,5	168,5	20,6	73,5	87,5	15,46	B
9	48,3	154,5	20,2	73,7	78,2	18,69	B
10	57,2	167,6	20,4	70,7	83,0	19,59	B
11	60,7	171,3	20,7	73,0	85,5	19,01	B
12	63,7	171,0	21,8	72,5	82,0	20,74	B
13	61,0	164,0	22,7	81,5	90,7	25,02	B
14	71,4	165,0	26,2	88,5	96,5	26,64	B
15	75,3	168,5	26,5	84,0	98,0	25,46	B
16	62,7	159,0	24,8	84,0	91,5	31,90	B
17	94,8	174,6	31,1	102,5	105,0	30,37	B
18	38,8	144,3	18,6	65,6	81,3	19,47	A
19	53,0	163,0	19,9	72,5	85,5	20,58	A
20	52,8	158,5	21,0	79,5	87,0	26,80	A
21	44,3	142,6	21,8	76,0	86,5	26,41	A
22	73,7	160,0	28,8	92,8	104,0	33,50	A

Tabla 1.14. Matriz de correlaciones

	Peso	Talla	IMC	P. Cintura	P. Cadera	% Grasa
Peso	1,000	0,700	0,905	0,892	0,882	0,641
Talla	0,700	1,000	0,338	0,383	0,393	0,045
IMC	0,905	0,338	1,000	0,956	0,943	0,844
P. Cintura	0,892	0,383	0,956	1,000	0,944	0,848
P. Cadera	0,882	0,393	0,943	0,944	1,000	0,827
% Grasa	0,641	0,045	0,844	0,848	0,827	1,000

FIGURA 1.22. ACP de nutrición



Ejercicio 1.5. En la tabla adjunta se encuentran cinco mediciones obtenidas sobre 23 peces. Dos mediciones son de radiactividad (escamas y ojos) y tres mediciones de tamaño (peso, largo y diámetro de los ojos). Los peces están repartidos en tres acuarios. Se busca saber cómo se relacionan las dos variables de radiactividad con las variables de tamaño, y si hay alguna diferencia entre los acuarios. Se adjunta la tabla de datos y la matriz de correlaciones de las cinco variables.

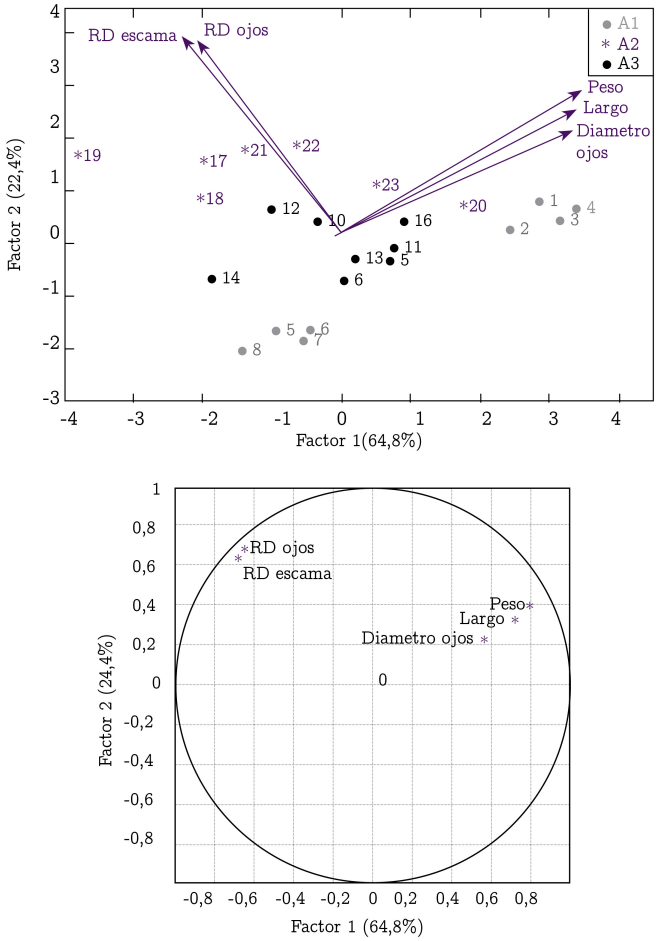
(a) Analicen la matriz de correlación de las cinco variables(Tabla 1.15).
(b) Se efectuó un ACP sobre las cinco variables. Interpreten el plano principal (Figura 1.23). Busquen si existen diferencias entre los acuarios.
(c) Comenten el círculo de correlaciones (Figura 1.23).

Pez	Radio escama	Radio ojos	Diámetro ojos	Largo	Peso	Acuario
1	142	10	11	214	132	A1
2	99	9	10	220	122	A1
3	121	6	11	220	129	A1
4	90	7	11	225	133	A1
5	244	8	9	168	57	A1
6	153	8	9	178	59	A1
7	162	7	9	176	59	A1
8	141	11	8	176	47	A1
9	169	13	10	182	72	A2
10	233	21	9	200	79	A2
11	220	12	11	185	80	A2
12	617	14	10	175	72	A2
13	211	14	10	189	75	A2
14	197	23	9	164	52	A2
15	191	13	10	195	86	A2
16	248	14	10	210	87	A2
17	461	32	9	181	72	A3
18	590	22	9	175	63	A3
19	809	31	8	170	49	A3
20	157	15	11	204	107	A3
21	690	22	9	190	83	A3
22	558	24	10	194	82	A3
23	345	19	11	190	91	A3

Tabla 1.15. Matriz de correlaciones

	Rad. escama	Rad. ojos	Diámetro ojos	Largo	Peso
Rad. escama	1,000	0,743	-0,395	-0,422	-0,381
Rad. ojos	0,743	1,000	-0,411	-0,379	-0,385
Diámetro ojo	-0,395	-0,411	1,000	0,677	0,803
Largo	-0,422	-0,379	0,677	1,00	0,938
Peso	-0,381	-0,385	0,803	0,938	1,000

FIGURA 1.23. ACP de los peces



Ejercicio 1.6. Sea la matriz de correlaciones de tres variables V_1, V_2 y V_3:

$$R = \begin{pmatrix} 1{,}00 & 0{,}70 & -0{,}80 \\ 0{,}70 & 1{,}00 & -0{,}21 \\ -0{,}80 & -0{,}21 & 1{,}00 \end{pmatrix}.$$

(a) Obtengan los valores y vectores propios de R.
(b) Representen las variables V_1, V_2 y V_3 en el círculo de correlaciones $\mathcal{C}$ de las dos primeras componentes principales.
(c) Den el porcentaje de conservación de la varianza en C.

Ejercicio 1.7. Se efectúa un ACP sobre los datos de 35 jugadores de tenis que fueron evaluados según 15 cualidades: Derecho, revés, servicio, volea, vuelta de servicio, smash, juego de pierna, lob, pelota corta, passing shot, regularidad, toque de pelota, estado psíquico, estado físico, juego en doble.

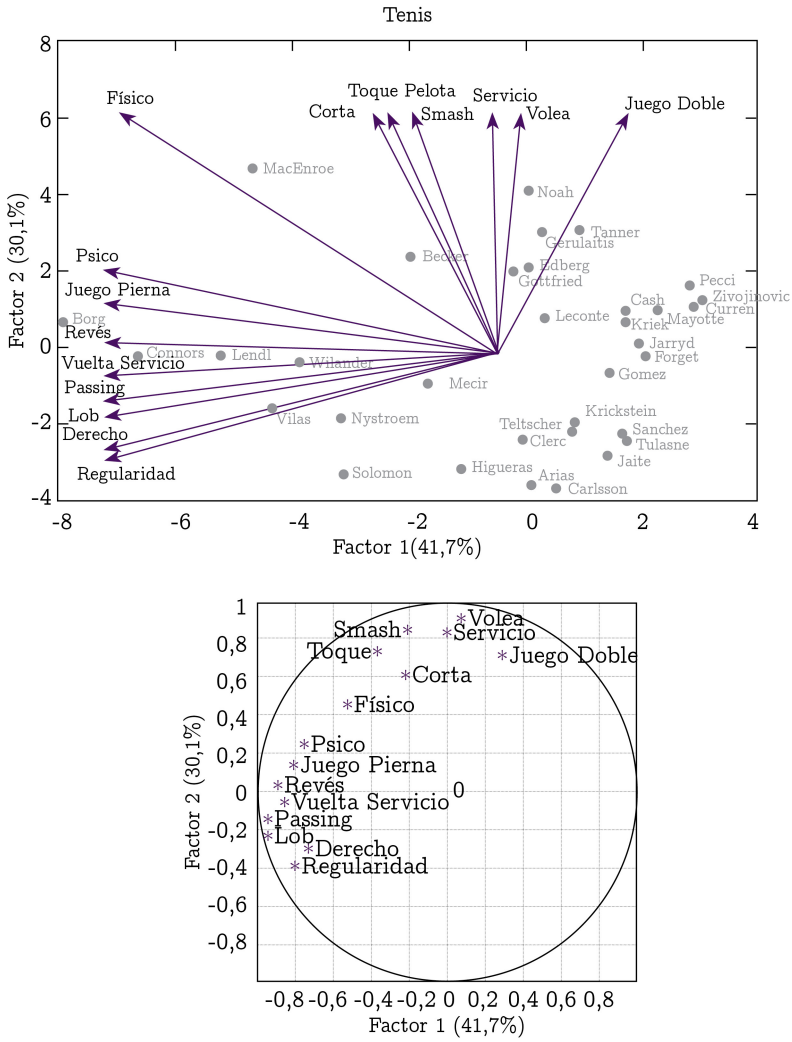
(a) Interpreten el círculo de correlaciones de la Figura 1.24. ¿Cómo se agrupan las cualidades de los jugadores?

(b) Interprete el plano principal de la Figura 1.24. ¿Como se interpreta cada componente principal?

(c) Ahora consideremos solamente la primera componente principal. Si queremos usarla como índice de 'ranking" de los jugadores, ¿que convendría hacer? Comparen la posición de los jugadores según la primera componente principal (C.P.) y el promedio de sus 15 cualidades dado en la tabla adjunta. Expliquen las diferencias entre la primera C.P. y el promedio de las cualidades.

Posición	1era C.P.	Jugador	Promedio
1	-6,763	BORG	7,8
2	-5,610	CONNORS	7,1
3	-4,273	LENDL	6,9
4	-3,822	MACENROE	8,4
5	-3,496	VILAS	6,0
6	-3,060	WILANDER	6,5
7	-2,441	NYSTROEM	5,9
8	-2,386	SOLOMON	5,1
9	-1,361	BECKER	6,7
10	-1,108	PERNFORS	5,5
11	-0,554	HIGUERAS	4,6
12	-0,434	MECIR	5,6
13	0,235	GOTTFRIED	6,2
14	0,371	CLERC	4,6
15	0,448	NOAH	6,7
16	0,472	EDBERG	6,0
17	0,489	ARIAS	4,1
18	0,674	GERULAITIS	6,2
19	0,747	LECONTE	5,5
20	0,925	CARLSSON	3,9
21	1,150	TELTSCHER	4,3
22	1,168	KRICKSTEIN	4,3
23	1,279	TANNER	5,9
24	1,680	JAITE	3,9
25	1,705	GOMEZ	4,9
26	1,894	SANCHEZ	4,3
27	1,970	KRIEK	4,7
28	1,989	CASH	5,1
29	1,996	TULASNE	3,9
30	2,175	JARRYD	4,9
31	2,273	FORGET	4,9
32	2,466	MAYOTTE	5,0
33	2,985	PECCI	4,9
34	3,040	CURREN	4,9
35	3,178	ZIVOJINOVIC	4,8

FIGURA 1.24. ACP de los jugadores de tenis



Capítulo 2: Test de hipótesis, teoría y aplicaciones

En el primer capítulo vimos el análisis en componentes principales, que permite describir datos multivariados. Mediante visualizaciones que pueden obtenerse del análisis, podemos sacar conclusiones, tales como relaciones entre variables o comparaciones de grupos de observaciones. Cuando los datos provienen de muestras, se buscará, en general, confirmar estas conclusiones. Queremos saber, por ejemplo, si las relaciones entre variables o los grupos de observaciones obtenidos se deben a la muestra particular analizada o son resultados que se pueden inducir a toda la población. Los test de hipótesis, objeto de este capítulo, permiten confirmar ciertas conclusiones; por ejemplo, que una proporción toma un determinado valor o las medias de dos grupos difieren. Por otra parte, el modelo de regresión (Capítulo 3) y el modelo CART (Capítulo 4) permiten confirmar ciertos tipos de relaciones. La calidad de los modelos obtenidos en estos dos capítulos se evalúan, en general, con un test de hipótesis, que son una herramienta fundamental de la Estadística.

En este capítulo, se profundiza la teoría de test estadísticos ya presentada en Lacourly [**7**]. Se asume que los conceptos básicos de probabilidad (Romagnoli[**14**] y Lladser [**8**]) y de muestreo aleatorio e inferencia (Lacourly[**7**]) son conocidos. Sin embargo, presentamos, a continuación, un repaso de los conceptos básicos necesarios para la comprensión del tema.

2.1 Conceptos básicos de inferencia estadística

Se llama "población" al conjunto de objetos o personas que nos interesa estudiar. Cuando no se puede tomar mediciones sobre todos los elementos de la población, se usa solamente una "muestra", que es un subconjunto de la población. Las mediciones obtenidas sobre una muestra, que se llaman "valores muestrales", proporcionarán, en general, resultados distintos de los que se deberían obtener de las mediciones de la población total. Por ejemplo, el consumo promedio de pan semanal de todos los hogares chilenos será distinto del consumo promedio de pan de una muestra de hogares. Las muestras más adecuadas son las obtenidas de manera aleatoria porque es un procedimiento de selección objetivo y permite hacer cálculos de errores en los resultados de inferencia. Por ejemplo, en una encuesta de opinión sobre la elección de un candidato se puede predecir el resultado de la votación de toda la población electoral con una cierta tasa de error a partir de las opiniones de solamente una muestra aleatoria. Lo fundamental es tener presente que estos resultados tienen un grado de incertidumbre que depende, en particular, del tamaño de la muestra.

Cuando utilizamos una muestra, se supone que estamos interesados en ciertas características de la población a nivel global y no individual. Por ejemplo, la media de la talla de los recién nacidos, la talla máxima o la desviación estándar de la talla. En la población, estas características se llaman "parámetros", y "estadísticos" en la muestra. Los parámetros son constantes y se deben determinar. Los estadísticos, que son funciones de los valores muestrales, varían, ya que dependen de la muestra. Cuando la muestra es aleatoria, los estadísticos son variables aleatorias (v.a.). Se trata entonces de decir algo sobre el valor de un parámetro de la población a partir del valor de un estadístico calculado sobre una muestra.

La teoría de la estimación permite atribuir un valor a cada parámetro de interés. Este valor, que se llama "estimación del parámetro", depende de la muestra aleatoria utilizada. A pesar de tener una sola muestra, conociendo el procedimiento de extracción, se puede calcular la varianza de las estimaciones del parámetro tal como si hubiéramos sacado todas las muestras posibles. La varianza de las estimaciones del parámetro depende del tamaño de la muestra. Mientras este es más grande, más pequeña es la varianza de las estimaciones. Pero la varianza de las estimaciones no dice mucho por sí sola. Sin embargo, a partir de ella podemos construir un intervalo de confianza, que permite señalar en qué rango de valores podría encontrarse el parámetro con una cierta probabilidad llamada "nivel de confianza". Un intervalo ancho indica una estimación poco precisa. Pero esto es muy relativo. El ancho del intervalo depende del nivel de confianza, que generalmente se elige elevado (95 % o 99 %), y de la varianza de las estimaciones, que depende a su vez del tamaño de la muestra. En efecto, cuando el tamaño de la muestra crece, el ancho del intervalo decrece y, por lo tanto, la precisión aumenta. Por ejemplo, si en 200 lanzamientos se obtuvieron 104 caras, la estimación de la probabilidad real p de sacar "cara" con esta moneda es $0,52$. El intervalo de p con un nivel de confianza de 95 % es $[0,45\ ;\ 0,59]$. Pero para un nivel de confianza de 99 % obtenemos el intervalo $[0,43\ ;\ 0,61]$ o si tenemos el 52 % de caras con 400 lanzamientos, el intervalo de 95 % de confianza es $[0,47\ ;\ 0,57]$.

Ahora bien, lo que nos interesa aquí es saber si la moneda está cargada. Esto se contesta con un test de hipótesis, que busca responder a una pregunta que involucra un parámetro. Se trata de determinar si un cierto valor o rango de valores dado a un parámetro es compatible con mediciones obtenidas en una muestra aleatoria. Con 52 % de caras en los 200 lanzamientos ¿podemos afirmar que la moneda está cargada para que salga "cara"? ¿O bien la moneda es equilibrada y el 52 % se debe al hecho de obtener este porcentaje sobre una muestra aleatoria? La teoría de tests de hipótesis, que da elementos para tomar decisiones a partir de datos empíricos, permitirá dar respuesta a esta pregunta.

Estimación y Tests de hipótesis son los temas fundamentales de la Inferencia estadística. Otro tema importante se refiere a métodos para extraer una muestra. No entraremos en detalle en este tema, pero en la Sección 2.5 discutiremos brevemente de la distinción entre el diseño muestral y el diseño experimental.

Para más detalles, pueden consultar la monografía de Lacourly[**7**] u otros textos, tales cómo Aliaga[**1**], Brook[**4**], Moore[**9**], Naiman[**10**] y Newman[**11**] y en documentos en Internet de Ycart[**15**] y Batanero[**2**][**3**].

2.2 Concepto de test de hipótesis

En muchos estudios, se quiere probar un resultado determinado o un impacto, que se traduce en una "hipótesis de trabajo". Consideramos las siguientes hipótesis de trabajo:

(a) El nuevo método de enseñanza de la matemática tiene un impacto positivo sobre el aprendizaje de los alumnos de 1°Medio.
(b) El precio del kilo de marraquetas en supermercados es más caro que en panaderías.
(c) Las niñas de 2°Medio de la Sra. Badilla tienen un promedio en matemática mayor que los niños.
(d) Esta moneda de 500 pesos está cargada a "cara".
(e) Existe una brecha en los resultados de la PSU de Matemática en el 2009 entre los diferentes tipos de colegios.
(f) El aumento de venta de 2 % de la empresa MEDOC es significativo.
(g) El precio unitario de la electricidad en Providencia es mayor que en Coyhaique.
(h) La vacuna TAPSA es eficaz para prevenir la gripe.

En cada uno de estos casos, por lo general se recogen datos que permiten comprobar si la afirmación es cierta. Se distinguen los casos (c), (e) y (g) de los demás. En efecto, si el promedio en matemática de todas las niñas del curso es 5,8 y los niños es 5,6, podremos decir con certeza que las niñas de 2°Medio de la Sra. Badilla tienen un promedio de matemática mayor que los niños. Para el caso (e), se puede comprobar que en la PSU de Matemática del 2009 los promedios por dependencia del colegio son distintos. Para el caso (g), basta mirar los cobros unitarios de las compañías de electricidad de Providencia y de Coyhaique para sacar una conclusión. En estos tres casos, la comprobación se basa sobre datos de toda la población (son censos). Para los demás es difícil y costoso recoger toda la información para comprobar con exactitud la afirmación. Se usa entonces una muestra aleatoria, que proporciona datos que conducen a una conclusión, pero ésta no es tan tajante como en los tres casos basados en un censo, ya que presenta una cierta incertidumbre. Los investigadores científicos, los médicos y nosotros mismos en la vida cotidiana nos enfrentamos frecuentemente a problemas similares.

Se plantean dos hipótesis, una llamada **Hipótesis nula**, que se contrasta con otra llamada **Hipótesis alternativa** y se establece una regla para elegir cuál de las dos es más compatible con los datos empíricos obtenidos de una muestra aleatoria.

Vemos que en el ejemplo (a) podemos tomar como hipótesis nula denotada H_o: "El nuevo método de enseñanza de la matemática tiene un impacto positivo sobre el aprendizaje de los alumnos de 1er Medio" y como hipótesis alternativa denotada H_1:

"El nuevo método de enseñanza de la matemática no tiene un impacto positivo sobre el aprendizaje de los alumnos de 1er Medio". Los roles de las dos hipótesis podrían parecer simétricos. Sin embargo, en las situaciones reales, no lo son, como veremos más adelante.

Tenemos entonces dos tipos de error de decisión posibles, llamados "error de tipo I" y "error de tipo II":

- Error de tipo I: declarar H_1 cierta cuando H_o es cierta, o sea, afirmar que el nuevo método de enseñanza de la matemática no tiene un impacto positivo sobre el aprendizaje de los alumnos de 1er Medio, cuando en realidad lo tiene.
- Error de tipo II: declarar H_o cierta cuando H_1 es cierta, o sea, afirmar que el nuevo método de enseñanza de la matemática tiene un impacto positivo sobre el aprendizaje de los alumnos de 1er Medio, cuando en realidad no lo tiene.

Las probabilidades de los errores Tipo I y Tipo II se designan convencionalmente con las letras griegas α y β, respectivamente. Veremos que no es posible tener ambas probabilidades de equivocarse tan pequeñas como uno quisiera al mismo tiempo. Mostraremos más adelante que si disminuimos la probabilidad α, la probabilidad β aumenta, y viceversa. En general, los dos errores no tienen las mismas consecuencias. Entonces, se tiene que elegir el error que consideramos más grave y "controlarlo". Si suponemos que el error controlado tiene la probabilidad α, vemos que controlamos el error de tipo I: Afirmar que el nuevo método de enseñanza de la matemática no tiene un impacto positivo sobre el aprendizaje de los alumnos de 1er Medio, cuando en realidad lo tiene. En este caso, el error de tipo II puede tomar un valor mayor o menor que α; no lo podemos controlar. Pero en el ejemplo, parece que el error de tipo II es el error más grave y es el que deberíamos controlar. Conviene en este caso tomar como hipótesis H_o: "El nuevo método de enseñanza de la matemática no tiene un impacto positivo sobre el aprendizaje de los alumnos de Primero Medio" y como hipótesis alternativa, denotada H_1: "El nuevo método de enseñanza de la matemática tiene un impacto positivo sobre el aprendizaje de los alumnos de 1er Medio" y controlar el error de tipo I.

Como ejercicio, escriban las hipótesis nula y alternativa en cada uno de los otros casos presentados anteriormente cuando el error controlado es el de tipo I y justifiquen su elección.

Vemos entonces que, en toda toma de decisión, es importante evaluar las consecuencias de equivocarse. En el caso de la moneda (d), si vamos a declarar que ella está cargada a "cara", tenemos que estar casi seguros de que efectivamente lo está, o sea, debemos controlar el error de tipo I, con H_o: la moneda no está cargada y H_1: la moneda esta cargada a "cara". En la mayoría de los casos, los roles de las hipótesis no son los mismos y las consecuencias de equivocarse en un sentido o en otro no tienen la misma importancia. Habitualmente, la hipótesis nula es la más conservadora, la que representa la tradición, el consenso; mientras que la hipótesis alternativa representa algo interesante, innovador, que queremos probar. Convencionalmente, se controla

entonces la probabilidad de cometer un error de Tipo I fijando un valor α_o máximo para α y se busca minimizar la probabilidad β, error de Tipo II, respetando el valor de α_o máximo preestablecido.

La situación es análoga a la de un juez que debe decidir si un inculpado es culpable o inocente sobre la base de las evidencias presentadas por la fiscalía y la defensa.

El juez tiene dos formas de equivocarse: dejar libre a un culpable o condenar a un inocente. En la mayoría de los países, el derecho no les da a los dos errores la misma importancia: se considera que es más grave condenar a un inocente que dejar libre a un culpable. Basado en este principio, el juez va a tratar de controlar el error de "condenar a un sospechoso que podría ser inocente" y, si las evidencias no le permiten establecer en forma fehaciente la culpabilidad del sospechoso, va a preferir dejarlo libre. La hipótesis nula será "el sospechoso es inocente". Es interesante recordar que la jurisprudencia no ha sido siempre la misma. En 1209, el enviado del papa Inocencio III recomendó matar a todos los habitantes de la ciudad francesa de Béziers, pues sabía que muchos eran herejes. Cuando le recordaron que entre los ciudadanos también había fieles, dijo "No importa, Dios los reconocerá".

Las dos hipótesis de los casos presentados anteriormente pueden escribirse mediante un parámetro de la población. Para decidir entre las dos hipótesis cuál está la más de acuerdo con los datos, se usa, entonces, un estadístico que se relaciona con el parámetro involucrado en las hipótesis. Por ejemplo, en el caso (d) se podrá tomar $H_o : p = \frac{1}{2}$ y $H_1 : p > \frac{1}{2}$, donde p es la verdadera probabilidad de sacar "cara" con esta moneda. Si en 200 lanzamientos obtuvimos 104 "caras", el estadístico es la proporción $\hat{p} = 0,52$ de "caras" obtenida en la muestra. Entonces, debemos verificar si la diferencia entre el valor observado y el valor de la hipótesis nula $\hat{p} - 0,50 = 0,02$ se debe al hecho de obtener $\hat{p}$ a partir de una muestra y la muestra es demasiado chica o si la moneda es realmente cargada a "cara".

Mientras más grande es la diferencia $\hat{p} - 0,5$, es más probable que la hipótesis H_1 sea cierta. Esta reflexión nos lleva a buscar un umbral c_α tal que si el valor del estadístico $\hat{p}$ evaluado sobre una muestra aleatoria lo sobrepasa, se rechaza la hipótesis nula con una probabilidad de equivocarse α. El umbral c_α define una **regla de decisión**: si $\hat{p} \geq c_\alpha$, se concluye que H_1 es cierta, y si $\hat{p} < c_\alpha$, se concluye que H_o es cierta. Pero en este último caso el error de decisión es β, que es un error no controlado, lo que no permite necesariamente aceptar H_1. Se dice, en general, que si $\hat{p} \geq c_\alpha$, se rechaza H_o, pero, que si $\hat{p} < c_\alpha$, **no se rechaza** H_o.

El umbral depende de la probabilidad α y de la hipótesis alternativa. Por ejemplo, si consideramos $H_o : p = \frac{1}{2}$ y $H_1 : p < \frac{1}{2}$, siendo H_1 la moneda está cargada a "sello", buscaremos un umbral tal que si el valor del estadístico es menor que el umbral, se rechaza la hipótesis nula. Una regla de decisión buscará entonces minimizar la probabilidad de error β dado un valor máximo α_0 de α.

2.3 Construcción de una regla de decisión en un caso simple

Mostramos a continuación las bases de la construcción de una regla de decisión mediante un ejemplo muy simplificado, que podría parecer poco realista, pero que permite introducir el procedimiento. Se deducirán después los casos más realistas encontrados en la práctica.

Los vecinos de un aeropuerto aseguran que el ruido promedio emitido por cierto tipo de aviones sobrepasa 80 decibeles, pero la dirección del aeropuerto asegura que es a lo más de 78 decibeles. Expertos consultados deciden registrar la intensidad del ruido de estos aviones. Como no pueden registrar datos en permanencia, toman una muestra aleatoria de 160 aviones y obtienen una intensidad media de 79,1 decibeles con una desviación estándar de 7 decibeles. La media de la muestra no es 78 y tampoco 80. Es ligeramente más cercano a 80, pero puede ser casual, dado que este resultado proviene de una muestra. Los expertos enfrentan, entonces, el dilema de decidir bajo incertidumbre. Se les presentan dos hipótesis para contrastar y tienen que optar por una de ellas considerando los resultados obtenidos de la muestra.

Planteamos las hipótesis del conflicto considerando $H_o : \mu = 80$ y $H_1 : \mu = 78$, donde μ es la media real del ruido (la media del ruido en la población). Los expertos saben que pueden tomar una decisión equivocada, pues la muestra es aleatoria, y que otra muestra daría posiblemente una media empírica distinta de 79,1. El error α es la probabilidad de decidir que el ruido promedio emitido es 78 decibeles, cuando en realidad el ruido es de 80 decibeles, o sea, α es el riesgo de equivocarse, si se da la razón al aeropuerto. El otro error, β, es la probabilidad de decidir que el ruido promedio emitido es 80 decibeles, cuando en realidad el ruido es de 78 decibeles, o sea, β es el riesgo de equivocarse, si se da la razón a los vecinos. Se escribe:

$$\alpha = \mathbb{P}(decidir H_1 | H_o \; es \; cierta), \quad \beta = \mathbb{P}(decidir H_o | H_1 \; es \; cierta).$$

Sabemos que el real valor μ puede ser un valor distinto a 78 u 80. Aquí, en el litigio entre el aeropuerto y sus vecinos no se consideran otros valores. Resolvemos entonces, en primer lugar, el test de hipótesis $H_o : \mu = 80$ y $H_1 : \mu = 78$ y más adelante incluiremos más valores de μ en las hipótesis.

Los expertos buscan minimizar los errores de decisión. Si deciden que el ruido promedio es 78 decibeles, quieren que sea con α lo más pequeño posible; y si deciden que vale 80, que sea con β lo más pequeño posible. Una "buena" **regla de decisión** debería minimizar los errores α y β. Pero una regla de decisión se determina a partir de los valores muestrales, o sea, un estadístico, que aquí es $\bar{x}$ la media de la muestra. Por ejemplo, una regla de decisión intuitiva es

$$(\mathcal{D}_1): \; Si \;\; \bar{x} > 79, \;\; se \;\; decide \;\; que \;\; H_o : \mu = 80 \;\; es \;\; cierta \;\; y$$
$$si \;\; \bar{x} \leq 79, \;\; se \;\; decide \;\; que \;\; H_1 : \mu = 78 \;\; es \;\; cierta.$$

Calculamos, entonces, los errores α y β asociados a esta regla de decisión. Se requiere una distribución de probabilidad para la media muestral

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n},$$

donde $x_1, x_2, ...x_n$ son los valores muestrales.

Si la distribución de población es Normal: $x_i \sim \mathcal{N}(\mu, \sigma)\ \forall i$, $\bar{x} \sim \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$ (Ver Lacouly[**7**]), ya que la suma de v.a. Normales sigue una distribución Normal. Nos falta determinar los parámetros de la distribución Normal, que son la media (o esperanza) y la desviación estándar de $\bar{x}$.

Calculemos estos dos parámetros. La esperanza

$$\mathbb{E}(\bar{x}) = \frac{\mathbb{E}(x_1) + \mathbb{E}(x_2) + ... + \mathbb{E}(x_n)}{n} = \frac{n\mu}{n} = \mu.$$

La desviación estándar es la raíz de la varianza, que es

$$V(\bar{x}) = \frac{Var(x_1) + Var(x_2) + ... + Var(x_n)}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n},$$

cuando x_1, x_2,..., x_n no dependen unos de los otros, lo que suponemos aquí. Luego, se deduce que $\bar{x}$ tiene una distribución Normal de media μ y desviación estándar $\tau = \frac{\sigma}{\sqrt{n}}$, denotada $\mathcal{N}(\mu, \tau)$. Ahora bien, si la distribución de población no es Normal, del Teorema Central del Límite se sabe que el resultado anterior es asintóticamente cierto, lo que significa que la media de la muestra tiende a una distribución Normal. Aproximamos, entonces, la distribución de $\bar{x}$ a la distribución $\mathcal{N}(\mu, \tau)$, donde $\tau = \frac{\sigma}{\sqrt{n}}$ (Ver Lacouly[**7**] y Lladser[**8**]).

Una vez determinada la distribución de $\bar{x}$, el cálculo de las probabilidades se hace mediante el computador o tablas de distribuciones. En Excel se usan las funciones "NORMDIST" y "NORMINV", en las cuales hay que especificar los valores numéricos de μ y τ. Se puede también usar la Tabla de distribución $\mathcal{N}(0, 1)$ dada en Anexo. Para esto, hay que transformar $\bar{x} \sim \mathcal{N}(\mu, \tau)$ para poder usar la distribución $\mathcal{N}(0, 1)$. La transformación se deduce del siguiente resultado:

> *Si a y b son dos constantes y* $X \sim \mathcal{N}(0, 1)$*, entonces* $Y = bX + a \sim \mathcal{N}(a, b)$*. Recíprocamente, si* $Y \sim \mathcal{N}(a, b)$*, entonces* $\frac{Y-a}{b} \sim \mathcal{N}(0, 1)$.

Se deduce de este resultado que si $\bar{x} \sim \mathcal{N}(\mu, \tau)$, entonces $Z = \frac{\bar{x}-\mu}{\tau} = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$.

Ahora, si tomamos la regla de decisión $(\mathcal{D}_1)$, tendremos

$$\alpha = \mathbb{P}(\bar{x} \leq 79 | \mu = 80)\ y\ \ \beta = \mathbb{P}(\bar{x} > 79 | \mu = 78).$$

Conociendo σ, podríamos calcular α y β, dado que μ vale 80 cuando H_o es cierta y 78 cuando H_1 es cierta.

Para mostrar cómo se calculan α y β, suponemos que σ es conocida con un valor igual a $7,2$[1]. Obtenemos los errores:

$$\alpha = \mathbb{P}\left(\frac{\bar{x}-80}{7,2/\sqrt{160}} \leq \frac{79-80}{7,2/\sqrt{160}}\right) = \mathbb{P}(Z \leq -1,757) \simeq 0,04,$$

$$\beta = \mathbb{P}\left(\frac{\bar{x}-78}{7,2/\sqrt{160}} > \frac{79-78}{7,2/\sqrt{160}}\right) = \mathbb{P}(Z \geq 1,757) \simeq 0,04,$$

donde $\frac{79-80}{7,2/\sqrt{160}} = -1,757$ y $\frac{79-78}{7,2/\sqrt{160}} = 1,757$. Se obtienen estas probabilidades aplicando el resultado,

Si a y b son constantes y X una variable aleatoria, entonces

$$\mathbb{P}(X \leq u) = \mathbb{P}(X - a \leq u - a) = \mathbb{P}\left(\frac{X-a}{b} \leq \frac{u-a}{b}\right).$$

Se usa entonces la Tabla de la distribución $\mathcal{N}(0,1)$ en Anexo después de redondear $1,757$ a $1,76$:

$$\mathbb{P}(Z \geq 1,76) = \mathbb{P}(Z \leq -1,76) = 0,0392.$$

En Excel se puede usar la función NORMDIST,

$$NORMDIST(79; 80; 7,2/SQRT(160); TRUE) = 0,03947...$$

$$1 - NORMDIST(79; 78; 7,2/SQRT(160); TRUE) = 0,03947...$$

Si la distribución Normal es continua, la precisión del cálculo depende del número de dígitos utilizados. Con la Tabla $N(0,1)$ tenemos que redondear a dos decimales (tomar $1,76$). Los cálculos con Excel son más precisos. El valor de la probabilidad es $0,03947...$, que redondeamos a $0,04$.

Con la regla de decisión $\mathcal{D}_1$, dado que $\bar{x} = 79,1$ es mayor que 79, aceptamos H_o con un error de $4\,\%$. Vale decir, que daríamos la razón al aeropuerto si estamos conforme con el error de $4\,\%$.

Definición 2.1. Se llama *región crítica* $\mathcal{R}$ al conjunto de valores del estadístico utilizado en la regla de decisión, con los cuales se rechaza la hipótesis nula H_o.

Con la regla de decisión $\mathcal{D}_1$, el estadístico de la regla de decisión es $\bar{x}$ y la región crítica es $\mathcal{R} = \{\bar{x}|\bar{x} \leq 79\}$. Vemos ilustrados los errores α y β en la Figura 2.1(a).

Si tomamos como región crítica $\mathcal{R} = \{\bar{x}|\bar{x} \leq 79,2\}$ en vez de $\mathcal{R} = \{\bar{x}|\bar{x} \leq 79\}$, el valor de α disminuye y el de β aumenta. Si tomamos como región crítica $\mathcal{R} = \{\bar{x}|\bar{x} \leq 78,5\}$ en vez de $\mathcal{R} = \{\bar{x}|\bar{x} \leq 79\}$, el valor de α aumenta y el de β disminuye. No podemos disminuir simultáneamente α y β a valores menores que $4\,\%$.

Ahora bien, obtuvimos, por ejemplo, los valores de α y β de $4\,\%$ fijando el umbral en 79. Podríamos proceder al reverso, especialmente si no encontramos el valor de $4\,\%$ satisfactorio para aceptar H_o. Podemos fijar el valor de α o el de β. En este caso,

[1]Eliminaremos este supuesto en la Sección 2.6.

el umbral c_α no será 79. Calculamos por ejemplo el umbral para un error α de 2 %. Utilizando la Tabla $\mathcal{N}(0,1)$, obtenemos,

$$\mathbb{P}(\bar{x} \leq u | \mu = 80) = 0,02 \Longrightarrow \mathbb{P}\left(\frac{\bar{x} - 80}{7,2/\sqrt{160}} \leq \frac{u - 80}{7,2/\sqrt{160}}\right) = 0,02,$$

donde $\frac{\bar{x}-80}{7,2/\sqrt{160}} \sim \mathcal{N}(0,1)$.

De la Tabla de la distribución Normal $\mathcal{N}(0,1)$ se deduce que $\frac{u-80}{7,2/\sqrt{160}} = -2,05$ y $u = 80 - 2,05 \times 7,2/\sqrt{160} = 78,83$. La región crítica para $\alpha = 2\,\%$ es entonces $\mathcal{R} = \bar{x} \leq 78,83\}$.

Calculamos el umbral c_α para el error $\alpha = 8\,\%$.

$$\mathbb{P}(\bar{x} \leq c_\alpha | \mu = 80) = 0,08 \Longrightarrow \mathbb{P}\left(\frac{\bar{x} - 80}{7,2/\sqrt{160}} \leq \frac{c_\alpha - 80}{7,2/\sqrt{160}}\right) = 0,08,$$

donde $\frac{\bar{x}-80}{7,2/\sqrt{160}} \sim \mathcal{N}(0,1)$.

De la Tabla de distribución Normal $\mathcal{N}(0,1)$ se deduce que $\frac{c_\alpha-80}{7,2/\sqrt{160}} = -1,41$ y $c_\alpha = 80 - 1,41 \times 7,2/\sqrt{160} = 79,2$.

La región crítica para $\alpha = 8\,\%$ es $\mathcal{R} = \{\bar{x} \leq 79,2\}$. Se concluye que si $\bar{x} \leq 79,2$, se rechaza la hipótesis nula $\mu = 80$ con un error de 8 %, pero que si $\bar{x} \leq 78,83$, se rechaza la hipótesis nula $\mu = 80$ con un error de 2 %. Con una media muestral de 79,1, observamos que tenemos diferentes conclusiones según el error α. Se da la razón al aeropuerto si $\alpha = 2\,\%$ y se da la razón a los vecinos si $\alpha = 8\,\%$. Mientras más pequeño es el error de tipo I, es más difícil dar la razón a los vecinos. Cuando se quiere controlar mejor la posibilidad de equivocarse rechazando la hipótesis nula, se toma un error de tipo I pequeño, pero vemos que puede haber una intención deliberada en la elección de la hipótesis nula. ¡Las hipótesis nula y alternativa elegidas aquí favorecen al aeropuerto!

En el caso de $\alpha = 8\,\%$, rechazamos H_o con una probabilidad de equivocarse de 8 %. Pero en el caso de $\alpha = 2\,\%$, no rechazamos H_o, por lo cual la probabilidad de equivocarse no es α, sino el error β. Calculamos entonces el error β correspondiente a la región crítica $\mathcal{R}_{0,02} = \{\bar{x} \leq 78,83\}$. En este caso, rechazamos H_1 cuando $\bar{x} > 78,83$ con $\mu = 78$:

$$\beta = \mathbb{P}(\bar{x} > 78,83 | \mu = 78) = \mathbb{P}\left(\frac{\bar{x} - 78}{7,2/\sqrt{160}} > \frac{78,83 - 78}{7,2/\sqrt{160}}\right) = \mathbb{P}(Z > 1,458) = 0,072,$$

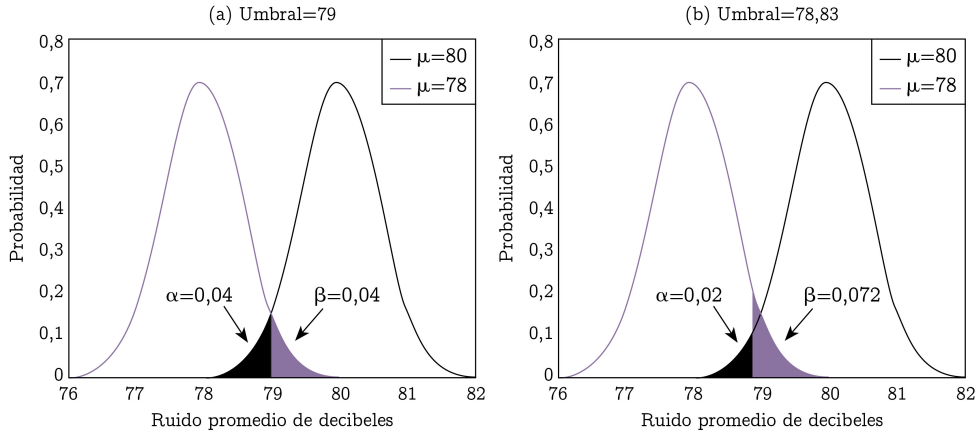
donde $Z \sim \mathcal{N}(0,1)$. Con la región crítica $\mathcal{R}_{0,02} = \{\bar{x} | \bar{x} \leq 78,83\}$, encontramos los errores α y β distintos (Figura 2.1(b)).

Ahora, si tomamos como valor umbral el valor obtenido en la muestra, la región crítica será $\mathcal{R}_\alpha = \{\bar{x} \leq 79,1\}$. ¿Qué error de tipo I tenemos en este caso?

$$\mathbb{P}(\bar{x} \leq 79,1 | \mu = 80) = \mathbb{P}\left(\frac{\bar{x} - 80}{7,2/\sqrt{160}} \leq \frac{79,1 - 80}{7,2/\sqrt{160}}\right) = \mathbb{P}(Z \leq -1,58) = 0,057,$$

donde $Z \sim \mathcal{N}(0,1)$.

FIGURA 2.1. Región crítica y errores tipo I y II



Se llama **p-valor** al error de tipo I definido por la región crítica basada en el umbral de 79,1, el valor numérico de $\bar{x}$ obtenido en la muestra. Se puede en este caso rechazar H_o con un error igual o mayor que 5,7 %. Dicho en otras palabras, **el p-valor es la probabilidad mínima con la cual se puede rechazar** H_o.

En resumen, una vez determinado cuál estadístico usar y la distribución que este sigue bajo la hipótesis nula, considerando las hipótesis nula y alternativa relativas al parámetro de la distribución de población, se busca entonces en qué parte de la distribución ubicar la región crítica. Esta depende de las hipótesis y de los errores α y β.

La regla de decisión ($\mathcal{D}_1$) tiene una cierta lógica. Como el valor de μ de la hipótesis alternativa H_1 es menor que el valor de la hipótesis nula H_o, es natural decidirse por H_1 cuando el valor de la media muestral $\bar{x}$ es menor que un valor umbral a determinar a partir de α. ¿Cómo cambia la regla de decisión si se intercambian las hipótesis nula y alternativa: $H_o^{'}: \mu = 78$ y $H_1^{'}: \mu = 80$? En este caso, deberíamos usar una regla que decide por $H_1^{'}$ si el valor de la media muestral $\bar{x}$ es mayor que un valor umbral.

Para construir la región crítica hicimos el supuesto que la desviación estándar σ en la población es conocida. Pero, en la práctica, es desconocida y se tiene que usar, en vez de la desviación estándar σ de la población, la desviación estándar de la muestra $s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$, donde $x_1, ..., x_n$ son los valores muestrales. En este caso $\frac{\bar{x}-\mu}{s/\sqrt{n}}$ no sigue una distribución Normal. Presentaremos tres funciones de distribución derivadas de la distribución Normal y algunas aplicaciones.

2.4 Tres distribuciones derivadas de la distribución Normal

En la Sección 2.3 acabamos de ver que un test de hipótesis sobre la media de la población requiere una distribución de probabilidad para la media muestral. Además, cuando la distribución de población es Normal, la media muestral también tiene una distribución Normal, lo que permite construir una regla de decisión cuando se conoce también la varianza σ^2 de la población. Pero, en general, no se conoce la varianza y ésta se estima a partir de la varianza de la muestra. Para determinar un estadístico para un test de hipótesis sobre una media, necesitaremos su distribución.

Buscaremos, en primer lugar, la distribución de la varianza muestral $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ y la distribución del estadístico $\frac{\bar{x}-\mu}{s/\sqrt{n-1}}$, que es el que se usa en los tests de hipótesis sobre una media de la población cuando no se conoce la varianza σ^2 y cuando los x_i siguen una distribución Normal.

Si $Z \sim \mathcal{N}(0,1)$, entonces Z^2 sigue una distribución llamada **chi cuadrado con** 1 **grado de libertad** y se denota $Z^2 \sim \chi^2_1$. Se generaliza esta distribución con la suma de cuadrados de variables $Z_i \sim \mathcal{N}(0,1)$ independientes entre sí.

Definición 2.2. La distribución de $U = \sum_{i=1}^{r} Z_i^2$, donde Z_i $(i = 1, 2, ..., r)$ son variables $\mathcal{N}(0,1)$ independientes entre sí, se llama distribución **chi cuadrado con** r **grados de libertad** y se denota χ^2_r.

La distribución χ^2_r depende solamente del parámetro r, que es el número de variables Z_i independientes que intervienen en la suma de cuadrados. Se encuentra en Anexo la tabla que proporciona valores de la distribución en función de r. Por ejemplo, $\mathbb{P}(\chi^2_{20} \geq 23,8) = 0,749$. También pueden usar las funciones CHIDIST y CHIINV de Excel. Si $U \sim \chi^2_{20}$:

$$1 - CHIDIST(23,8;20) = 0,749 \;\; y \;\; CHIINV(0,251;20) = 23,8.$$

Proposición 2.1. *Si* $U_1 \sim \chi^2_r$ *y* $U_2 \sim \chi^2_p$ *y* U_1 *y* U_2 *son variables aleatorias independientes, entonces* $U_1 + U_2 \sim \chi^2_{r+p}$.

Para demostrar la proposición basta utilizar la definición de la χ^2_r.

Tenemos dos aplicaciones inmediatas. Vimos que si $x_1, x_2, ..., x_n$ son los valores muestrales independientes entre sí y $x_i \sim \mathcal{N}(\mu, \sigma)$ para todo i, entonces $\frac{x_i-\mu}{\sigma} \sim \mathcal{N}(0,1)$ y $\frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$. Luego,

$$(a)\;\; m^2 = \left(\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\right)^2 \sim \chi^2_1, \;\; (b)\;\; v^2 = \sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2 \sim \chi^2_n.$$

De las dos distribuciones de m^2 y v^2 deducimos ahora la distribución de $w^2 = \sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{\sigma}\right)^2$. Observen que la expresión de w^2 se parece a la de v^2, salvo que se

reemplazó μ por $\bar{x}$. Los términos $\frac{x_i-\bar{x}}{\sigma}$ ya no son $\mathcal{N}(0,1)$. Sin embargo, tenemos la siguiente descomposición:

$$\sum_{i=1}^{n}\left(\frac{x_i-\mu}{\sigma}\right)^2=\left(\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\right)^2+\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{\sigma}\right)^2,$$

o sea, $v^2=m^2+w^2$.

Un resultado, que no demostraremos aquí, permite deducir la distribución de w^2.

Proposición 2.2. *En una población normal, la media muestral* $\bar{x}=\frac{1}{n}\sum x_i$ *y la varianza muestral* $s^2=\frac{1}{n}\sum(x_i-\bar{x})^2$ *son dos variables aleatorias independientes.*

Como los parámetros μ y σ son constantes, no cambia la condición de independencia de $\bar{x}$ y s^2. Se deduce que las v.a. $m=\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ y $w^2=\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{\sigma}\right)^2$ son independientes también, así como m^2 y w^2. Como $v^2=m^2+w^2$, $m^2\sim\chi_1^2$ y $v^2\sim\chi_n^2$, de la proposición 2.1 obtenemos que $w^2\sim\chi_{n-1}^2$.

Las distribuciones de m^2, v^2 y w^2 dependen de los parámetros de la población μ y σ^2. Sin embargo, observando que $w^2=n\frac{s^2}{\sigma^2}$, obtenemos que $\frac{m}{w}=\frac{\bar{x}-\mu}{s}$, donde $m\sim\mathcal{N}(0,1)$ y $w^2\sim\chi_{n-1}^2$ depende solamente de μ.

Definición 2.3. Si $Z\sim\mathcal{N}(0,1)$ y $U\sim\chi_r^2$ son dos variables aleatorias independientes, entonces la distribución de $\frac{Z}{\sqrt{U/r}}$ se denomina **t de Student con r grados de libertad**, lo cual denotamos por $T\sim t_r$.

La distribución de Student fue definida en 1908 por William Sealy Gosset, quien trabajaba en la fábrica de cerveza Guiness, que prohibía a sus empleados la publicación de artículos científicos debido a una difusión previa de secretos industriales. Esto llevó a William Gosset a publicar sus resultados bajo el seudónimo de Student.

En el Anexo encontrarán la tabla de la función de distribución t de Student t_r para diferentes valores de r. La distribución de Student es muy parecida a la distribución $\mathcal{N}(0,1)$, y, en particular, es simétrica con respecto al 0. Su esperanza es nula y su desviación estándar vale $\frac{r}{r-2}$ para $r>2$, que es mayor que 1 y depende del grado de libertad r. Su distribución depende solamente de r, y cuando r tiende al infinito, la distribución se acerca a la distribución $\mathcal{N}(0,1)$. Comparen los valores de $\mathbb{P}(Y\geq u)=0,05$ cuando $Y\sim t_{50}$, $Y\sim t_{120}$ y $Y\sim\mathcal{N}(0,1)$. Las funciones TDIST y TINV de Excel también permiten obtener las probabilidades de la distribución t de Student.

Aplicamos ahora la definición de la t de Student para encontrar la distribución del estadístico $\frac{\bar{x}-\mu}{s/\sqrt{n-1}}\sim t_{n-1}$.

Proposición 2.3. *Sea* $\{x_1,x_2,...,x_n\}$ *una muestra aleatoria donde* $x_i\sim\mathcal{N}(\mu,\sigma)$, $i=1,2...,n$. *Entonces, se obtiene la distribución*

$$T=\frac{\bar{x}-\mu}{s/\sqrt{n-1}}\sim t_{n-1},$$

donde $\bar{x}$ y s son la media y la desviación estándar muestrales.

En efecto $\bar{x} \sim \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$ y el estadístico T puede escribirse como $T = \frac{m}{\sqrt{\frac{w^2}{n-1}}}$, donde $m = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$ y $w^2 = \frac{ns^2}{\sigma^2} \sim \chi^2_{n-1}$. Lo interesante del estadístico T con una distribución de Student[2] es que no depende de σ.

Ejercicio: Mirando la Tabla de la t de Student, busque el tamaño n a partir del cual $T = \frac{\bar{x}-\mu}{s/\sqrt{n-1}} \sim t_{n-1}$ sigue aproximadamente una distribución $\mathcal{N}(0,1)$.

Presentamos una tercera distribución definida por Sir Ronald Fisher, que permite comparar varianzas muestrales.

Definición 2.4. Si las variables aleatorias $U \sim \chi^2_r$ y $V \sim \chi^2_p$ son independientes, entonces la distribución del cociente $F = \frac{U/r}{V/p}$ se denomina **F de Fisher** con r y p grados de libertad y se denota $F \sim F_{r,p}$.

La variable aleatoria F toma valores no negativos y su distribución depende de los dos parámetros r y p. En el anexo encuentran la tabla de la distribución F de Fisher. También puede usar las funciones FDIST y FINV de Excel. Note que si $F \sim F_{r,p}$, entonces $1/F \sim F_{p,r}$. Por ejemplo, $\mathbb{P}(F_{24,10} \geq 4,33) = \mathbb{P}(F_{p,r} \leq 0,231) = 0,01$.

Veamos una aplicación de la distribución de Fisher. Consideramos dos muestras aleatorias de alumnos de 1°Medio, una muestra en la población de niños y una muestra en la población de niñas. Las dos muestras son obtenidas de manera independiente una de la otra. Se considera, por ejemplo, la nota promedio en lenguaje, y queremos comparar las varianzas de la nota σ_1^2 y σ_2^2 en la población de los niños y la población de las niñas, respectivamente. Las hipótesis que se deben contrastar son:

$$H_o : \sigma_1^2 = \sigma_2^2 \;\; y \;\; H_1 : \sigma_1^2 \neq \sigma_2^2 \Longleftrightarrow H_o : \frac{\sigma_1^2}{\sigma_2^2} = 1 \;\; y \;\; H_1 : \frac{\sigma_1^2}{\sigma_2^2} \neq 1. \tag{2.1}$$

Si n_1 y n_2 son los tamaños muestrales respectivos y s_1^2 y s_2^2 las varianzas muestrales respectivas de las muestras de niños y niñas, entonces,

$$F_o = \frac{s_1^2 \frac{n_1}{n_1-1}}{s_2^2 \frac{n_2}{n_2-1}} \sim F_{n_1-1,n_2-1}.$$

En efecto, $U = n_1 \frac{s_1^2}{\sigma^2} \sim \chi_{n_1-1}$ y $V = n_2 \frac{s_2^2}{\sigma^2} \sim \chi_{n_2-1}$ y las dos variables U y V son independientes, dado que las muestras son independientes. Luego, $F_o = \frac{U/(n_1-1)}{V/(n_2-1)} \sim F_{n_1-1,n_2-1}$. Observen que el test no depende de las medias en la población.

Verifiquen que, si la χ^2 del numerador de la F tiene un grado de libertad (r=1), usar una $F_{1,p}$ de Fisher equivale a usar una t_p de Student, resultado dado en la siguiente proposición:

Proposición 2.4. *Si $G \sim F_{1,p}$, entonces $\sqrt{G} \sim t_p$.*

[2]En muchos textos de Estadística se usa como desviación estándar empírica de los valores muestrales a $s^* = \frac{1}{n-1}\sum_1^n (x_i - \bar{x})^2$ y entonces $T = \frac{\bar{x}-\mu}{s^*/\sqrt{n}} \sim t_{n-1}$.

2.5 Diseño experimental versus diseño muestral

En los ejemplos presentados en la primera sección, para tomar decisiones "bien informadas", se requieren datos, pero el método de recolección de datos no será el mismo para todos los problemas. Se tienen fundamentalmente dos clases de recolección de datos muestrales, que dependen del grado de control de ciertas variables.

(A) El diseño muestral. En el caso de la contaminación acústica por los aviones, los expertos requieren información obtenida de una muestra aleatoria de aviones despegando o aterrizando en el aeropuerto. En este caso, se pueden considerar, por ejemplo, los tipos de aviones en el diseño de la muestra para asegurar cubrir la variabilidad del ruido.

Si queremos hacer un seguimiento de los logros en matemática de los alumnos entre 1°y 4°Medio en los diferentes tipos de colegios del país, se puede aplicar una batería de pruebas a una muestra de alumnos durante su Enseñanza Media (EM). Se usa un diseño muestral que permite, en general, seleccionar una muestra de colegios y aplicar las pruebas a todos los alumnos de los colegios seleccionados. Los resultados obtenidos en la muestra deberían reflejar los logros de los alumnos de EM de todos los colegios del país.

Con un diseño muestral se busca una muestra de una determinada población, que permite medir ciertos aspectos sobre los elementos de la muestra con el objetivo de presentar lo mejor posible los aspectos de la población total.

(B) El diseño de experimentos. Si se busca medir el impacto del nuevo método de enseñanza en 1°Medio (caso (a)), se pueden tomar dos muestras aleatorias de alumnos de 1°Medio y aplicar el nuevo método de enseñanza solamente a los alumnos de una de las muestras. El grupo de alumnos que no tuvieron el nuevo método se llama "grupo control". Se comparan entonces los resultados de los dos grupos de alumnos. Es importante, aquí, que un alumno esté asignado al grupo control o al otro grupo de manera aleatoria, de manera que los dos grupos sean comparables antes de aplicar el nuevo método de enseñanza. Este caso es muy distinto del seguimiento de los logros de los alumnos de EM, ya que podemos producir, por ejemplo, estadísticas globales que reflejan el nivel de los estudiantes chilenos. Pero, en el caso (a), vamos a producir estadísticas al interior de cada grupo y compararlas, pues así podremos comprobar si el nuevo método de enseñanza tiene algún impacto.

En el caso (h) de la vacuna, se quiere medir el efecto o el impacto de la vacuna TAPSA para prevenir la gripe, pero no se busca recolectar estadísticas de la prevalencia[3] de la gripe. Aquí se requieren dos muestras también. Una de personas a las que se pone la vacuna y otra de personas que se quedarán sin vacuna. Este último es el "grupo control". Si la vacuna es eficaz, se espera que

[3]La prevalencia es el número total de casos de una enfermedad dada que existe en una población en un momento específico.

la prevalencia del primer grupo sea muy inferior a la prevalencia del grupo control.

El diseño experimental es una estrategia estadística que permite identificar y cuantificar las causas eventuales del efecto de "una intervención" (el nuevo método de enseñanza) sobre una variable de interés llamada "variable respuesta" (el rendimiento o logro del alumno). Es el diseño que se usa en las evaluación de impactos.

La 'intervención" corresponde a una o más variables controladas en el sentido que se sabe quién tuvo el nuevo método de enseñanza o quién tuvo la vacuna. Estas variables están vinculadas a las causas para medir el efecto que tienen en otra variable de interés. En el ejemplo del impacto de un nuevo método de enseñanza, el rendimiento de los alumnos es la variable respuesta y el cambio de método de enseñanza es la variable que podría influir sobre el rendimiento. En el ejemplo de la vacuna, "enfermarse de la gripe" es la variable respuesta y la "vacuna" es la variable que podría tener efecto sobre la gripe.

Veamos en qué difieren los dos tipos de diseños utilizando el cambio de método de enseñanza.

1. Diseño muestral: Se toma una muestra aleatoria de 1200 alumnos de 1°Medio. Se mide su rendimiento en matemática y, además, se recopila la información del método de enseñanza que recibió el alumno. Se pueden entonces comparar los rendimientos de los alumnos según el método de enseñanza.
2. Diseño experimental: Se toma una muestra aleatoria de 1200 alumnos de 1°Medio. Se reparten los 1200 alumnos de manera aleatoria en dos grupos de 600 alumnos. En un grupo se aplica el método antiguo de enseñanza y en el otro el nuevo método. La aleatoriedad de la elección de los grupos es un aspecto fundamental en el diseño experimental. Para evaluar el efecto del método nuevo de enseñanza no sería adecuado aplicar el método antiguo a 600 niños y el método nuevo a 600 niñas. La elección de los dos grupos no debe agregar el efecto de otra variable como aquí el género. La aleatorización requiere en general ser "controlada". Por ejemplo, se construyen los dos grupos de manera aleatoria, pero tomando la misma cantidad de niños y niñas y de diferentes tipos de colegios, de manera de controlar el efecto del género y de la dependencia del colegio (Tabla 2.1).

En el diseño muestral, se puede observar a posteriori el método de enseñanza recibido por el alumnos, mientras que en el segundo caso, el método de enseñanza está impuesto deliberadamente a un grupo de alumnos.

En ambos casos podemos comparar los rendimientos en matemática. Sin embargo, si buscamos medir el efecto del método de enseñanza sobre el rendimiento de los alumnos, el diseño experimental es una mejor estrategia. No solamente se asegura que haya el mismo número de alumnos por método de enseñanza, lo que proporciona precisión similar para los promedios, sino que permite controlar ciertas variables que

Tabla 2.1. Construcción controlada de los grupos

Grupo con método antiguo de enseñanza				
Género	Colegio municipal	Particular subvencionado	Particular pagado	Total
Niños	100	100	100	300
Niñas	100	100	100	300
Total	200	200	200	600

Grupo con método nuevo de enseñanza				
Género	Colegio municipal	Particular subvencionado	Particular pagado	Total
Niños	100	100	100	300
Niñas	100	100	100	300
Total	200	200	200	600

pueden influir en el rendimiento: tipo de colegio, género, región, etc., de manera que al inicio los grupos de alumnos sean comparables y que, en lo posible, las diferencias pueden atribuirse solamente al método de enseñanza.

2.6 Test en una población

Si consideramos el coeficiente intelectual (CI) de los niños chilenos, podemos interesarnos en diferentes características de la distribución de población del CI, que trataremos de especificar lo mejor posible mediante valores obtenidos de una muestra aleatoria. Las características pueden referirse a la distribución que es Normal, simétrica, unimodal, bimodal, etc., o bien a un parámetro –media, mediana, máximo etc.– de la distribución. En el primer caso, se trata de la forma de la distribución de la variable. Buscamos, entonces, verificar si la distribución de los valores muestrales es Normal, de Poisson, etc.. En el segundo caso, suponemos conocido el tipo de distribución de la población de la variable y buscamos definir sus parámetros, tales como media, desviación estándar, mediana, máximo. Si suponemos una distribución de población Normal, los parámetros de interés son la media y la desviación estándar. Si suponemos una distribución de Bernoulli (variable binaria que toma los valores 0 ó 1), el parámetro de la distribución es la probabilidad de obtener 1. En esta monografía, trataremos solamente los tests de hipótesis para la media de una distribución Normal y el parámetro de probabilidad para una distribución de Bernoulli. Los test para una mediana son más complejos.

2.6.1 Test para la media de una población

En la Sección 2.3, mostramos la manera de construir una regla de decisión con dos hipótesis relativas a un solo valor del parámetro μ de una distribución Normal, lo que es poco realista. Ahora vamos a volver a la construcción de una regla de decisión para un test de hipótesis sobre la media en una población normal $\mathcal{N}(\mu, \sigma)$. Supondremos también que σ es desconocida y usaremos entonces la distribución t de Student definida en la Sección 2.4. Se distinguen tres casos, que difieren por el recorrido de valores de la hipótesis alternativa, y que conducen a diferentes tipos de reglas de decisión. Suponiendo μ_o un valor dado, los casos son:

Caso 1 $H_o : \mu \geq \mu_o$ contra $H_1 : \mu < \mu_o$.
Caso 2 $H_o : \mu \leq \mu_o$ contra $H_1 : \mu > \mu_o$.
Caso 3 $H_o : \mu = \mu_o$ contra $H_1 : \mu \neq \mu_o$.

Caso 1

El director de un colegio asegura que el coeficiente intelectual (CI) de los alumnos de su colegio es al menos 110. Para probarlo, aplica un test de inteligencia a una muestra aleatoria de 41 alumnos y obtiene en la muestra un promedio $\bar{x} = 108$ con una desviación estándar $s = 10$. ¿El director está equivocado?

Vamos a suponer, en una primera etapa, que el director dice que el CI promedio μ del colegio es 110 en vez de "al menos 110". Las hipótesis nula y alternativa son entonces: $H_o : \mu = 110$ contra $H_1 : \mu < 110$. Si fijamos el error α, ¿cuál error controlamos? Controlamos el error de declarar que el director se equivoca cuando está en lo correcto.

Se encontró, en la muestra aleatoria de 41 alumnos, una media $\bar{x}$ del CI 108. Se trata entonces de saber si el valor de 108, que es menor que 110, se debe al hecho de medir el promedio en una muestra, siendo el promedio μ igual a 110, o si en realidad el promedio μ de la población es efectivamente inferior a 110.

Vemos que la hipótesis nula se refiere a un solo valor de μ, mientras que la hipótesis alternativa considera todos los valores menores que 110. Por el momento, no consideramos los valores mayores que 110. Esto permite precisar la distribución de $\bar{x}$ cuando la hipótesis nula es cierta, o sea, cuando $\mu = 110$. Con una hipótesis nula $H_o : \mu \leq 110$, la distribución no es única, dado que hay una distribución para cada valor de μ menor o igual a 110. Veremos más adelante cómo resolver este problema.

Ahora bien, si tuviéramos, por ejemplo, como hipótesis alternativa $H_1 : \mu = 106$, podríamos proceder como hicimos anteriormente en la Sección 2.2. Buscamos un umbral c_α tal que

$$\mathbb{P}(\bar{x} \leq c_\alpha | \mu = 110) = \alpha, \tag{2.2}$$

donde α es la probabilidad de equivocarse declarando que μ vale 106, cuando, en realidad, vale 110. La región crítica es entonces de la forma $\mathcal{R}_\alpha = \{\bar{x} \leq c_\alpha\}$. Revise por qué esta región crítica es correcta y no $\mathcal{R}_\alpha = \{\bar{x} \geq c_\alpha\}$.

Observamos ahora que la región crítica no depende de un valor especifico de H_1, sino solamente del hecho que $\mu < 110$. En efecto, si en vez de 106 en la hipótesis alternativa tomamos $H_1 : \mu = 105$, encontramos el mismo umbral c_α, que depende del valor α y de $\mu = 110$ (ecuación 2.2) y no depende de los valores de la hipótesis alternativa, como $\mu = 105$ o $\mu = 106$. Este umbral es el mismo para todo valor μ de la hipótesis alternativa. Luego, la región crítica del test para $H_o : \mu = 110$ contra $H_1 : \mu < 110$ es $\mathcal{R}_\alpha = \{\bar{x} \leq c_\alpha\}$, donde $\mathbb{P}(\bar{x} \leq c_\alpha | \mu = 110) = \alpha$.

Se supone que el CI en el colegio sigue una distribución $\mathcal{N}(\mu, \sigma)$, donde μ y σ son desconocidos. Luego la media muestral $\bar{x} \sim \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$. Como σ es desconocida, no podemos usar la distribución Normal en el test, sino la distribución de Student, que definimos en la Sección 2.4:

$$T = \frac{\bar{x} - \mu}{s/\sqrt{n-1}} \sim t_{n-1},$$

que permite usar la desviación estándar s de la muestra en vez de σ.

Reemplazando n, μ y s por sus valores, bajo la hipótesis nula H_o, $T = \frac{\bar{x}-110}{10/\sqrt{40}} \sim t_{40}$. De la tabla de la distribución de Student, para $\alpha = 0,05$, obtenemos $\mathbb{P}(t_{40} \geq 1,684) = 0,05$, y, por simetría, $\mathbb{P}(t_{40} \leq -1,684) = 0,05$. La región crítica se determina, entonces, por

$$\mathbb{P}\left(\frac{\bar{x} - 110}{10/\sqrt{40}} \leq -1,684\right) = \mathbb{P}(\bar{x} \leq 110 - 1,684 \times 10/\sqrt{40}) = \mathbb{P}(\bar{x} \leq 107,34) = 0,05.$$

La región es entonces $\mathcal{R}_{0,05} = \{\bar{x} \leq 107,34\}$.

El valor de la muestra, que fue 108, es mayor que 107,34. No podemos rechazar la hipótesis nula con el error de 5 %. Si aceptamos que el promedio del colegio es de 110, es con un error de tipo II. Pero no podemos calcular el error β, siendo que hay muchos valores de μ en la hipótesis alternativa: $H_1 : \mu < 110$.

Veamos, entonces, si queremos rechazar la hipótesis $H_o = 110$, con qué error α podríamos hacerlo. Sale del concepto de "p-valor" visto anteriormente.

Calculamos el p-valor del test, que es igual al error de tipo I correspondiente a la región crítica $\mathcal{R}_{p-valor} = \{\bar{x} \leq 108\}$, donde 108 es el CI promedio en la muestra de 41 alumnos (Figura 2.2(a)).

$$p\text{–}valor = \mathbb{P}(\bar{x} \leq 108 | \mu = 110) = \mathbb{P}\left(T \leq \tfrac{108-110}{10/\sqrt{40}}\right) = \mathbb{P}(T \leq -1,265) \approx 0,10,$$

donde $T = \frac{\bar{x}-110}{10/\sqrt{40}} \sim t_{40}$. La Tabla de la distribución de Student no proporciona la probabilidad correspondiente a $-1,265$. En realidad $\mathbb{P}(t_{40} \leq -1,303) = 0,10$. De aquí la aproximación a 10 % para $-1,265$. En Excel se usa la función TDIST, que calcula $\mathbb{P}(X \geq u)$ y no acepta valores negativos de u. Pero sabemos que la distribución t de Student es simétrica, $\mathbb{P}(t_{40} \geq 1,265) = \mathbb{P}(t_{40} \leq -1,265)$, y obtenemos el resultado con $TDIST(1,265; 40; 1) = 0,1066$.

Podemos decir, entonces, que si rechazamos H_o con el valor de 108, será con un error de 10 %. **Es el menor error de tipo I que permite rechazar** H_o. Tenemos

que decidir si estamos dispuestos a asumir un error de tipo I de 10 % para rechazar H_o. Con 10 %, le daríamos al director el beneficio de la duda.

Recordemos ahora, que el director dijo que el CI promedio de los alumnos de su colegio es **al menos** de 110. En este caso, la hipótesis nula es $H_o : \mu \geq 110$, quedando la hipótesis alternativa sin cambio, $H_1 : \mu < 110$. Sabemos que la región crítica depende del μ, de la hipótesis nula y de α el error de tipo I. Por ejemplo, para $\alpha = 5\,\%$, tendremos regiones críticas diferentes para $\mu = 110$ y $\mu = 112$ si imponemos que cada región crítica tenga exactamente un error de 5 %. Para $\mu = 112$,

$$\mathbb{P}\left(\frac{\bar{x} - 112}{10/\sqrt{40}} \leq -1,684\right) = \mathbb{P}(\bar{x} \leq 112 - 1,684 \times 10/\sqrt{40}) = \mathbb{P}(\bar{x} \leq 109,34) = 0,05.$$

Para $\mu = 112$, $\mathcal{R}_{0,05} = \{\bar{x} \leq 109,34\}$.

Necesitamos una sola región crítica, cualquiera sea el valor de μ de la hipótesis nula. Veamos cuál es el error de tipo I para $\mu = 112$ tomando la región crítica $\mathcal{R}_{0,05} = \{\bar{x} \leq 107,34\}$ obtenida con $\mu = 110$. Obtenemos $\mathbb{P}(\bar{x} \leq 107,34|\mu = 112) = \mathbb{P}\left(\frac{\bar{x}-112}{10/\sqrt{40}} \leq -2,95\right) = \mathbb{P}(t_{40} \leq -2,95) \leq 0,005$. El error de tipo I para $\mu = 112$ es más pequeño que el error de tipo I para $\mu = 110$. Como ejercicio, calculen el error de tipo I para $\mu = 111$ con la región $\mathcal{R}_{0,05} = \{\bar{x} \leq 107,34\}$.

Para todo valor $\mu < 110$, la región crítica $\mathcal{R}_{0,05} = \{\bar{x} \leq 107,34\}$ tiene un error de tipo I menor que 5 %. Se concluye que la región crítica $\mathcal{R}_{0,05} = \{\bar{x} \leq 107,34\}$ es la solución. Verifiquen por qué. Esta región crítica tiene un error de tipo I que no sobrepasa 5 % para todo valor de la hipótesis nula $H_o : \mu \geq 110$ contra $H_1 : \mu < 110$.

Veamos como cambian los resultados si el director hubiera usado una muestra de 122 alumnos, obteniendo el mismo promedio y la misma desviación estándar. En primer lugar, vemos que podemos usar la distribución Normal que aproxima bien la distribución de Student para un grado de libertad mayor que 120. Busquemos entonces la región crítica para $\alpha = 5\,\%$:

$$\mathbb{P}(\bar{x} \leq c_{0,05}|\mu = 110) = \mathbb{P}\left(\frac{\bar{x} - 110}{10/\sqrt{121}} \leq -1,65\right) = 5\,\%.$$

Obtenemos $c_{0,05} = 110 - 1,65\frac{10}{121} = 108,5$.

La región crítica es $\mathcal{R}_{0,05} = \{\bar{x} \leq 108,5\}$. Con un error de tipo I de 5 %, rechazamos H_o. Con la muestra de 41 alumnos, no se rechazó H_o. Mientras más grande es la muestra, más cercano a 110 se encuentra el umbral c_α.

Calculemos ahora el p-valor para la muestra de 122 alumnos:

$$\mathbb{P}(\bar{x} \leq 108|\mu = 110) = \mathbb{P}(\frac{\bar{x} - 110}{10/\sqrt{121}} \leq \frac{108 - 110}{10/\sqrt{121}}) = \mathbb{P}(Z \leq -2,2) = 1,4\,\%,$$

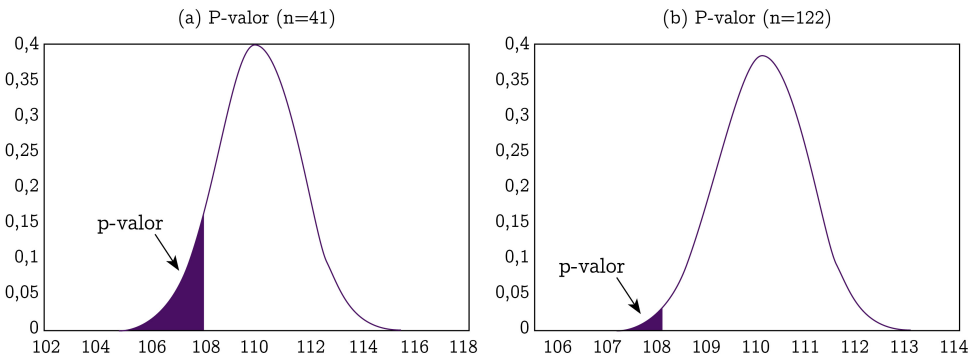
donde $Z \sim \mathcal{N}(0,1)$.

Observen que cuando el tamaño de la muestra es de 122 alumnos, no solamente rechazamos H_o con un error de 5 %, sino que también lo podemos hacer con un error de 1,4 % (Figura 2.2(b)). Verifiquen, por ejemplo, que no rechazamos H_o para un error

de 1 %. Vemos también que la distribución del gráfico (b) tiene menos dispersión que la del gráfico (a).

En resumen, con la muestra de 41 alumnos, se puede rechazar H_o con un error mínimo de 10 %, pero con 122 alumnos el error disminuye a 1,4 %. Se toma menos riesgo rechazando H_o con muestras grandes.

FIGURA 2.2. Efecto del tamaño de la muestra sobre el p-valor



Caso 2

Retomamos un ejemplo de la monografía [**7**], donde se estudia el proceso de fabricación de una bebida. Se sabe que la cadena de producción tiene una media diaria de 500 litros. La implementación de un nuevo proceso debería permitir aumentar la producción diaria. Se implementó el nuevo proceso sobre una de las cadenas, que en una muestra de 61 días dio una producción promedio de 520 litros con una desviación estándar de 75 litros. ¿Podemos decir que el nuevo proceso es eficaz?

Si μ es la media de producción diaria con el nuevo proceso, las hipótesis nula y alternativa son: $H_o : \mu \leq 500$ y $H_1 : \mu > 500$. En este caso, rechazaremos H_o cuando la media muestral $\bar{x}$ tome valores mayores que un umbral c_α: $\mathcal{R}_\alpha = \{\bar{x} \geq c_\alpha\}$. Procedemos como en el caso anterior, utilizando el estadístico $T = \frac{\bar{x}-500}{75/\sqrt{60}}$. Tenemos una muestra de tamaño 61; por lo tanto, $T \sim t_{60}$. Consideremos, por ejemplo, un error de tipo I de 1 %:

$$\mathbb{P}(\bar{x} \geq c_\alpha | \mu = 500) = \mathbb{P}\left(\frac{\bar{x} - 500}{75/\sqrt{60}} \geq \frac{c_\alpha - 500}{75/\sqrt{60}}\right) = \mathbb{P}(T \geq 2,39) = 1\,\%.$$

Se deduce el umbral $c_{0,01} = 500 + 2,39 \times 75/\sqrt{60} = 523,1$. Con una media muestral de 520, no podemos decir que el nuevo proceso sea eficaz con un error de decisión de 1 %.

Si aumentamos el error α a 5 %, $\mathbb{P}(T \geq 1,67) = 5\,\%$, por lo cual obtenemos un umbral

$$c_{0,05} = 500 + 1,67 \times 75/\sqrt{60} = 516,2.$$

Ahora, con un error de 5 %, podemos rechazar H_o y concluir que el nuevo proceso es eficaz. Para un error de tipo I de 1 %, no se rechaza H_o, que sí se rechaza para un error de tipo I de 5 %. Se concluye que el p-valor está entre 1 % y 5 %.

Para calcular el p-valor, el error más pequeño con el cual podemos rechazar H_o, tenemos que considerar la región crítica $\mathcal{R}_{p-valor} = \{\bar{x} \geq 520\}$:

$$p\text{–}valor = \mathbb{P}(\bar{x} \geq 520|\mu = 500) = \mathbb{P}(\frac{\bar{x} - 500}{75/\sqrt{60}} \geq \frac{520 - 500}{75/\sqrt{60}}) = \mathbb{P}(T \geq 2,06) \approx 0,02,$$

donde $T \sim t_{60}$. El error de 2 % es el más pequeño con el cual podemos equivocarnos y declarar que el nuevo proceso es eficaz (Figura 2.3(a)).

Vimos en el caso 1 cómo influye el tamaño de la muestra en la regla de decisión. Veamos ahora cómo influye la desviación estándar de los valores muestrales. Si la desviación estándar hubiera sido $s = 100$ litros en vez de 75 litros, dejando fijos el tamaño de la muestra $n = 61$ días y la media muestral de 520 litros, el p-valor sería mayor (Figura 2.3(b)):

$$p\text{–}valor = \mathbb{P}(\bar{x} \geq 520|\mu = 500) = \mathbb{P}\left(\frac{\bar{x} - 500}{100/\sqrt{60}} \geq \frac{520 - 500}{100/\sqrt{60}}\right) = \mathbb{P}(T \geq 1,55) > 5\,\%.$$

Nuevamente, la Tabla de la distribución de Student en Anexo no nos permite encontrar el valor exacto del p-valor. Pueden verificar en Excel que el p-valor exacto es igual a 6,3 %. Con esta desviación estándar mayor que la anterior, para decidir que el nuevo proceso es eficaz con un error menor al 6,3 % habría que tomar una muestra más grande. Es el costo que se pagará –tomar una muestra grande – cuando la varianza de la población sea grande.

Calculamos el tamaño de muestra mínimo requerido para poder bajar el p-valor de 6,3 % a 5 %. El estadístico T se escribe ahora como $T = \frac{\bar{x}-500}{100/\sqrt{n-1}}$, donde n es el tamaño muestral buscado. Esperamos un tamaño muestral mayor que $n = 60$, lo que permite aproximar la distribución del estadístico T a una Normal: $T \sim \mathcal{N}(0,1)$. Para un error de 5 %, $\mathbb{P}(T \geq 1,65) = 5\,\%$. Luego,

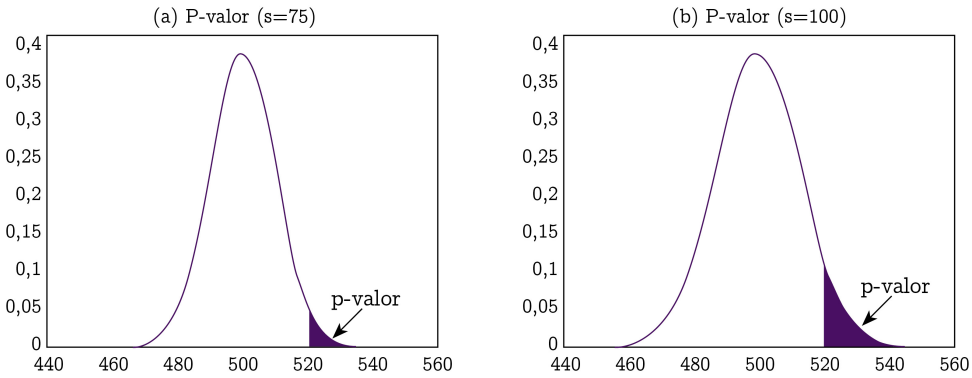
$$\mathbb{P}(\bar{x} \geq 520) = \mathbb{P}\left(\frac{\bar{x} - 500}{100/\sqrt{n-1}} \geq \frac{520 - 500}{100/\sqrt{n-1}}\right) = 5\,\%,$$

de donde se deduce que

$$\frac{520 - 500}{100/\sqrt{n-1}} = 1,65 \Longrightarrow \sqrt{n-1} = \frac{100 \times 1,65}{20} \Longrightarrow n = 69.$$

Tenemos que aumentar el tamaño muestral de 61 a 69. Como ejercicio, verifiquen que el tamaño muestral, que permite obtener un p-valor de 1 % en el caso de $s = 75$, es $n = 137$.

FIGURA 2.3. Efecto de la desviación estándar sobre el p-valor



Ahora, ¿qué pasa con β, el error de tipo II? La hipótesis H_1 no establece un valor específico de la media μ –sólo dice que es mayor que 500–, y hay infinitos valores de μ mayores que 500. Podemos calcular β para cada hipótesis alternativa posible H_1 para la región crítica obtenida a partir de un error α. Recordando que no se rechazó H_o con el error de tipo I de 1 %, cuando $s = 75$ y $n = 61$, tomaremos $\alpha = 1\,\%$.

En este caso, la región crítica es $\mathcal{R}_{0,01} = \{\bar{x} \geq 523, 1\}$ y β es la probabilidad de no rechazar H_o para un valor de μ mayor que 500. Por ejemplo, para $\mu = 505$:

$$\beta = \mathbb{P}(\bar{x} < 523, 1 | \mu = 505) = \mathbb{P}(\frac{\bar{x} - 505}{75/\sqrt{60}} < \frac{523, 1 - 505}{75/\sqrt{60}} = \mathbb{P}(T < 1, 87) = 0, 97,$$

donde $T \sim t_{60}$. Este error es bastante grande. La Tabla 2.2 y la Figura 2.4(a) muestran β para valores de μ entre 500 y 550 litros; β disminuye a medida que μ se aleja de 500, lo que no debería sorprendernos, pues es obviamente más fácil decidir entre valores muy distintos que entre valores parecidos.

TABLA 2.2. Error de Tipo II

μ	500	510	520	530	540	550
β	0,990	0,910	0,625	0,239	0,043	0,004

La probabilidad $1 - \beta$ de no equivocarse cuando H_1 es cierta se llama **potencia**. Minimizar β es equivalente a maximizar la potencia.

Si quisiéramos disminuir el error β, por ejemplo, para $\mu = 530$ (que vale 0,239 cuando α vale 1 %), habría que aumentar α. Para $\alpha = 5\,\%$, el umbral es de 516,2 y

$$\beta = \mathbb{P}(\bar{x} < 516,2|\mu = 530) = \mathbb{P}(\frac{\bar{x} - 530}{75/\sqrt{60}} < \frac{516,2 - 530}{75/\sqrt{60}} = \mathbb{P}(T < -1,43) = 0,079.$$

Ahora bien, si queremos disminuir β sin aumentar α, la única solución es aumentar el tamaño de la muestra. Por ejemplo, con $n = 150$, el umbral de $\alpha = 1\,\%$ es de 514,45, y aproximando la Student a la Normal $\mathcal{N}(0,1)$:

$$\beta = \mathbb{P}(\bar{x} < 514,45|\mu = 530) = \mathbb{P}(\frac{\bar{x} - 530}{75/\sqrt{149}} < \frac{514,45 - 530}{75/\sqrt{149}} \approx \mathbb{P}(Z < -2,53) = 0,007,$$

donde $Z \sim \mathcal{N}(0,1)$. Aun tomando $\alpha = 1\,\%$, logramos disminuir considerablemente β.

Caso 3

Se considera, en general, que la temperatura del cuerpo de un adulto con buena salud debería ser de 37°. Algunos médicos ponen en duda esta afirmación. Para verificar si ellos tienen la razón y no la creencia general, se obtiene una muestra aleatoria de 150 adultos sanos. De la muestra se obtiene una temperatura promedio $\bar{x} = 37,3°$ con un desviación estándar de 1,3°. Con este valor obtenido en la muestra, ¿podemos aceptar que la temperatura promedio de los adultos sanos es 37°?

Llamamos μ a la temperatura promedio en la población de los adultos sanos. Para plantear las hipótesis nula y alternativa tenemos que analizar lo que queremos probar. Nada en la objeción presentada por los médicos dice que μ podría ser mayor o menor que 37°. Se debe, entonces, contrastar el valor de $\mu = 37°$ contra valores de μ tanto menores como mayores que 37°. Además queremos controlar el error α de dar la razón a los médicos cuando en realidad $\mu = 37°$. Se toman entonces las hipótesis nula y alternativa: $H_o : \mu = 37^o$ contra $H_1 : \mu \neq 37^o$.

La región crítica no puede ser como en el Caso 1 o el Caso 2. En efecto, si $\mu > 37$, la región crítica tiene la forma del Caso 1 $\{\bar{x} \leq c_1\}$ y la forma del Caso 2 $\{\bar{x} \geq c_2\}$ en el caso contrario. Si tanto los valores menores como mayores que 37°son igualmente posible, es natural repartir el error α en dos partes iguales:

$$\mathbb{P}(\bar{x} \leq c_1|\mu = 37) = \frac{\alpha}{2}, \;\; \mathbb{P}(\bar{x} \geq c_2|\mu = 37) = \frac{\alpha}{2},$$

para obtener un error de tipo I total igual a α. La región crítica es entonces la unión de los dos subconjuntos:

$$\mathcal{R}_\alpha = \{\bar{x} \leq c_1\} \cup \{\bar{x} \geq c_2\}.$$

Se dice que la hipótesis $H_1 : \mu \neq 37$ es **bilateral**. En los dos casos anteriores, las hipótesis alternativas $H_1 : \mu \leq 110$ y $H_1 : \mu \geq 500$ se dicen **unilaterales**. Tenemos entonces aquí dos umbrales c_1 y c_2 por calcular. Si consideramos $\alpha = 5\,\%$:

$$\mathbb{P}(\bar{x} \leq c_1|\mu = 37) = 0,025 \Longrightarrow \mathbb{P}\left(\frac{\bar{x} - 37}{1,3/\sqrt{150}} \leq \frac{c_1 - 37}{1,3/\sqrt{150}}\right) = 0,025.$$

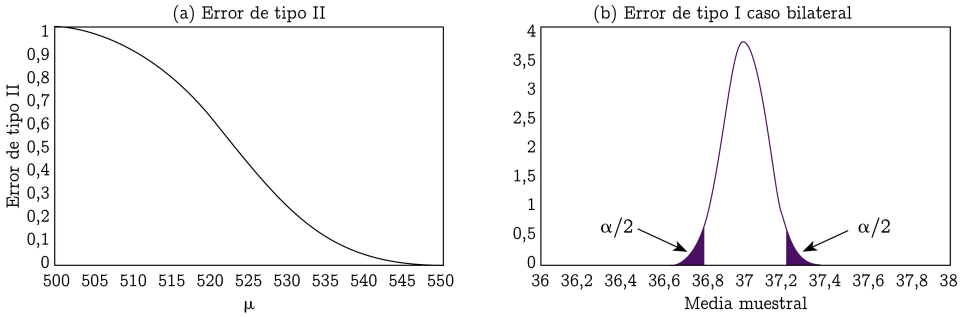
Aproximando la distribución de $Z = \frac{\bar{x}-37}{1,3/\sqrt{150}} \sim \mathcal{N}(0,1)$, y considerando que $\mathbb{P}(Z \leq -1,96) = 0,025$, se deduce el umbral $c_1 = 37 - 1,96 \times 1,3/\sqrt{150} = 36,8°$. El umbral c_2 se obtiene de la misma manera: $\mathbb{P}(Z \geq 1,96) = 0,025$. Se deduce $c_2 = 37 + 1,96 \times 1,3/\sqrt{150} = 37,2°$ (Figura 2.4(b)). Siendo que el valor de $\bar{x}$ en la muestra es igual a 37,3°, se rechaza H_o, dando la razón a los médicos con un error de 5 %.

Como ejercicio, comprueben que, para $\alpha = 2\,\%$, se sigue rechazando H_o, al obtener la región crítica $\mathcal{R}_{0,02} = \{\bar{x} \leq 36,75\} \cup \{\bar{x} \geq 37,25\}$. Calculemos ahora el p-valor considerando ambos lados de la distribución, como en la construcción de la región crítica:

$$p\text{–}valor = \mathbb{P}(|\bar{x}| \geq 37,3|\mu = 37) = 2 \times \mathbb{P}(\bar{x} \geq 37,3|\mu = 37) = 2 \times 0,0024 = 0,0048.$$

Podríamos rechazar H_o hasta un error $\alpha = 0,0048$.

FIGURA 2.4. Representación de los errores



Cuando el caso lo permite, es preferible usar una hipótesis alternativa unilateral. En efecto, el p-valor para una hipótesis alternativa bilateral toma el doble del valor del p-valor de la hipótesis unilateral.

Los conceptos de p-valor y región crítica son complementarios: el p-valor es la probabilidad más pequeña de equivocarnos si rechazamos la hipótesis H_o con el valor encontrado en la muestra; la región crítica es el conjunto de valores de $\bar{x}$ para los cuales se rechaza H_o con un error α controlado. En ambos casos, la decisión que se tomará depende del riesgo que estemos dispuestos a asumir. El p-valor se denomina también **nivel de significación**. Es común referirse a los errores de Tipo I como "falsos positivos" y a los de tipo II como "falsos negativos".

En resumen, para tomar una decisión entre dos alternativas relativas a una media μ a partir de datos muestrales, se formulan la hipótesis nula y la hipótesis alternativa sobre μ en función de lo que se quiere poner en evidencia. Se determina entonces el estadístico del test basado en $\bar{x}$ y su distribución –Normal (caso muestra grande) o t de Student (caso muestra pequeña)–, que se utilizará bajo la hipótesis nula. Se puede entonces usar una de las dos estrategias:

(a) Determinar la región crítica del test si se usa un error α fijado a priori. Si el valor de $\bar{x}$ observado en la muestra pertenece a la región crítica, se rechaza H_o.
(b) Calcular el p-valor a partir del valor de $\bar{x}$ observado en la muestra. Si el p-valor es muy pequeño, se rechaza la hipótesis nula. Si se sospechaba que la hipótesis nula era falsa, pero el p-valor no resultó tan pequeño como para rechazarla, tal vez habría que tomar una muestra más grande para confirmar la sospecha.

Los errores α más utilizados son 1 % y 5 %. ¿Por qué un error u otro? Depende obviamente del riesgo que se esté dispuesto a asumir. Este riesgo no es el mismo para todos. Es la ventaja del p-valor, que no requiere fijar a priori el valor de α. El p-valor permite a menudo tomar una decisión sin mucha dificultad. En general, cuando es menor que 5 %, se podrá rechazar H_o y cuando es mayor que 10 %, no se podrá rechazar H_o. El problema surge cuando el p-valor está entre 5 % y 10 %. Si vale 8 %, ¿estamos dispuestos a rechazar H_o con una probabilidad de equivocarse de 8 %? Depende de la gravedad de las consecuencias que puede tener un error de decisión.

2.6.2 Test para la proporción de una población

El resultado de un lanzamiento de una moneda es binario: "cara" o "sello". Los resultados aleatorios binarios pueden definirse mediante una variable aleatoria de Bernoulli X, que toma dos valores: $X = 1$ con probabilidad p y $X = 0$ con probabilidad $1 - p$ (denotada $X \sim \mathcal{B}(p)$). Definamos, por ejemplo, $X = 1$, si sale "cara" y $X = 0$, si sale "sello". El parámetro p es entonces la probabilidad de sacar "cara" en un lanzamiento. Si la moneda es equilibrada, $p = 0,5$. Si la moneda está cargada a "cara", $p > 0,5$. La distribución de Bernoulli se usa en situaciones en las cuales se tienen solo dos resultados posibles. Otro ejemplo: en una elección entre dos candidatos, Valverde y Rojo, un votante tiene dos alternativas. Se puede definir, por ejemplo, $X = 1$ si vota por Valverde y $X = 0$ si vota por Rojo. El parámetro p de la Bernoulli es, entonces, la probabilidad de que un elector vote por Valverde.

Volviendo al ejemplo de la moneda, si en 130 lanzamientos se obtuvieron 70 caras, ¿podemos concluir que la moneda está cargada a "cara"?

La hipótesis nula será $H_o : p = 0,5$ (moneda equilibrada) y la hipótesis alternativa será unilateral $H_1 : p > 0,5$ (moneda cargada hacia "cara"). La pregunta es si la moneda está cargada a "cara". No está considerado que $p < 0,5$. Ahora, si nos preguntamos si la moneda está cargada, sin especificar que esta cargada a "cara" o a "sello", la hipótesis alternativa sería $H_1 : p \neq 0,5$.

Construimos entonces la región crítica $\mathcal{R}$ para las hipótesis $H_o : p = 0,5$ contra $H_1 : p > 0,5$. Sea S el número de "caras" obtenido en los n=130 lanzamientos. Cuando la hipótesis H_1 es cierta, se espera un número S de "caras" mayor que el número de 'sellos", lo que lleva a construir una región crítica de la forma $\mathcal{R} = \{S \geq c\}$ de tal manera que $\mathbb{P}(S \geq c) = \alpha$.

Ahora bien, una suma de n variables $X_1, X_2, ..., X_n$ de Bernoulli de mismo parámetro p e independientes entre sí, sigue una distribución $\mathcal{B}inomial$ de parámetros n y p ($\mathcal{B}inomial(n,p)$). Bajo H_o, el estadístico S sigue, entonces, una $\mathcal{B}inomial(130;0,5)$.

Para calcular al valor c, se puede usar una Tabla de distribución Binomial (ver en Anexo), o bien EXCEL (función BINOMDIST para obtener el umbral c y CRITBINOM para obtener el p-valor). Cuando n es grande se puede aproximar la Binomial a una distribución Normal. Para usar la aproximación hay que calcular la esperanza y desviación estándar de la variable Binomial.

Para la variable de Bernoulli X,

$$\mathbb{E}(X) = p \times 1 + (1-p) \times 0 = p \;\; y \;\; Var(X) = p \times (1-p)^2 + (1-p) \times (0-p)^2 = p(1-p).$$

Se deduce la esperanza y varianza de la variable $S = \sum_{i=1}^{n} X_i \sim \mathcal{B}inomial(n,p)$, donde las X_i son independientes y $X_i \sim \mathcal{B}(p)$:

$$\mathbb{E}(S) = \sum_{i=1} n\mathbb{E}(X_i) = np \;\; y \;\; Var(S) = \sum_{i=1}^{n} Var(X_i) = np(1-p).$$

La aproximación es entonces $S \sim \mathcal{N}(np, \sqrt{np(1-p)})$. Bajo $H_o : \mu = 0,5$, $\mathbb{E}(S) = 130 \times 0,5 = 65$ y $Var(S) = 130 \times 0,5 \times 0,5 = 32,5$. Si usamos la aproximación a la distribución Normal, tenemos $S \sim \mathcal{N}(65; 5, 7)$.

Si tomamos $\alpha = 5\,\%$ y $S \sim \mathcal{B}inomial(130;0,5)$, con la función CRITBINOM de EXCEL:

$$CRITBINOM(130; 0,5; 0,95) = 74,$$

obtenemos $\mathbb{P}(S \geq 74) = 5\,\%$.

Si usamos la aproximación de la distribución $\mathcal{B}inomial(130;0{,}5)$ a la distribución $\mathcal{N}(65; 5, 7)$, donde $5,7 = Var(S)$, en EXCEL,

$$NORMINV(0,95; 65; 5, 7) = 74,37,$$

obtenemos el valor 74,37, que es cercano a 74, el valor exacto.

Tomamos entonces la región crítica $\mathcal{R}_\alpha = \{S \geq 74\}$ para un error de $5\,\%$. Encontramos en los 130 lanzamientos 70 "caras", que es menor que 74. No encontramos evidencia que la moneda esté cargada a "cara" para un error de $5\,\%$.

Calculamos ahora el $p{-}valor = \mathbb{P}(S \geq 70)$ del test, usando EXCEL. El valor exacto es:

$$1 - BINOMDIST(70; 130; 0,5; TRUE) = 0,167,$$

o la aproximación Normal

$$1 - NORMDIST(70; 65; 5, 7; TRUE) = 0,19.$$

Vimos que no podemos declarar la moneda cargada a "cara" cuando el error de tipo I es de $5\,\%$. Podríamos declararla cargada tomando un error α de al menos 0,167.

Como ejercicio, resuelvan los casos $H_o : p \geq 0,5$ contra $H_1 : p < 0,5$ (moneda cargada a "cara" contra cargada a "sello") y $H_o : p = 0,5$ contra $H_1 : p \neq 0,5$ (moneda no cargada contra cargada).

Veamos otro ejemplo. La Superintendencia de Telecomunicaciones (SUBTEL) encargó un estudio sobre la cobertura de una compañía de teléfonos celulares. El decreto estipula que la cobertura territorial de la telefonía celular de la compañía debería ser al menos 90 % del territorio $\mathcal{T}$ que pretende cubrir la compañía. En otras palabras, si una persona llama con un teléfono celular de la compañía de cualquier parte del territorio $\mathcal{T}$, debería poder comunicarse en al menos 90 % de los casos. En la duda, se realiza un experimento para comprobarlo. Se eligen al azar 800 puntos del territorio $\mathcal{T}$, de donde se intenta llamar y recibir llamadas. Se obtuvieron 710 llamadas con éxito (un porcentaje de 88,75 %). ¿Puede la SUBTEL multar a la compañía?

Como el resultado fue obtenido de una muestra, no se puede asegurar que la compañía no cumple con el decreto. La SUBTEL sólo debería multar a la compañía si está muy segura que ésta no cumple la cobertura ofrecida de 90 %, y necesita controlar el error de declarar que la compañía no la cumple cuando en realidad la cumple.

En este contexto se deducen las hipótesis por contrastar. Las hipótesis nula y alternativas son $H_o : p \geq 0,90$ contra $H_1 : p \leq 0,90$, pues se quiere controlar el error α de multar la compañía cuando no corresponde.

Buscamos ahora el estadístico del test. Tomamos como variable de interés a la variable binaria: $X = 1$, si la llamada es exitosa, y $X = 0$, en caso contrario. $X \sim \mathcal{B}ernoulli(p)$, donde p es la probabilidad de tener una llamada exitosa. Si la compañía cumple con la obertura, tenemos $p \geq 0,90$. El número de llamadas exitosas S sigue entonces una distribución $\mathcal{B}inomial(800, p)$.

Construyamos la región crítica para el número S de llamadas exitosas, con $\alpha =$ 5 % y las hipótesis $H_o : p \geq 0,90$ contra $H_1 : p \leq 0,90$. Sabemos que, bajo H_o, $S \sim \mathcal{B}inomial(800; 0{,}90)$. Buscamos entonces el umbral c_α tal que $\mathbb{P}(S \leq c_\alpha | p = 0,90) = 0,05$. En efecto, se rechaza H_o para valores menores de S. Ese valor resulta ser c_a=706. Como pudimos hacer 710 llamadas exitosas, con α=5 %, con la muestra de 800 llamadas, no se puede multar a la compañía.

Este cálculo también se puede hacer con una aproximación a la normal: $S \sim \mathcal{N}(720, \sqrt{72})$, donde $720 = 800 \times 0,9$ y $72 = 800 \times 0,9 \times (1 - 0,9)$. Se obtiene prácticamente la misma región crítica.

Calculamos el p-valor (BINOMDIST(710;800;0,90;TRUE)=0,132):

$$p\text{–}valor = \mathbb{P}(S \geq 710) = 0,132.$$

Tenemos un riesgo importante al multar a la compañía con tal p-valor.

2.7 Comparación de medias

Varias preguntas que planteamos al inicio del capítulo, no pueden ser respondidas con el test de media anterior. Por ejemplo, en evaluación de impacto o comparación de

tratamientos hay que comparar dos o más medias relativas a una misma medición. Se presentan dos situaciones:

(i) Se mide la variable de interés en dos grupos que provienen de dos poblaciones diferentes. Por ejemplo, se mide el sueldo de hombres y mujeres o bien se miden los logros de dos grupos de alumnos que tuvieron dos métodos de enseñanza diferentes. Se quiere comparar, entonces, las medias de los dos grupos. Otro ejemplo, se compara el precio del kilo de marraquetas en supermercados con el de panaderías.

(ii) Se mide la variable de interés en dos momentos distintos en una muestra que proviene de una sola población. Por ejemplo, para medir el impacto de un programa, se toman mediciones que se relacionan con el programa antes y después de la aplicación del mismo. Se comparan entonces las medias obtenidas antes y después del programa. Otro ejemplo: se comparan los sueldos del hombre y de la mujer en una pareja. Hay una sola población compuesta de parejas y dos mediciones de la variable sueldo, sueldo del cónyuge y sueldo de su pareja.

Como veremos a continuación, los test de hipótesis por aplicar en estos dos casos son distintos.

2.7.1 Comparación de dos medias en dos poblaciones

Una pregunta como *¿El precio del kilo de marraquetas en supermercados es más caro que en panaderías?* requiere comparar medias que provienen de dos poblaciones distintas. Tenemos la población de las panaderías y la población de los supermercados que venden pan. Si μ_1 es el precio promedio del kilo de marraquetas en panadería y μ_2 en supermercado, vamos a comparar los dos precios promedio. Se tiene que usar una muestra aleatoria de panaderías y una de supermercados para estimar los dos promedios. Además, la extracción de la primera muestra no debe depender de la extracción de la segunda.

Supongamos que a partir de una muestra aleatoria de 70 panaderías y una muestra aleatoria de 40 supermercados obtuvimos una media muestral $\bar{x}_1 = 980$ pesos en panadería y $\bar{x}_2 = 1100$ pesos en supermercado. ¿La diferencia entre las dos medias muestrales permite decir que el pan del supermercado es más caro o la diferencia es casual y se debe a que consideramos valores provenientes de muestras?

Queremos averiguar si efectivamente el precio del kilo de marraquetas en supermercados es más caro que en panaderías. Queremos controlar, entonces el error de declarar que el precio del kilo de marraquetas en supermercados es más caro que en panaderías, cuando no lo es. Nos conduce a las hipótesis nula y alternativa: $H_o : \mu_1 \leq \mu_2$ contra $H_1 : \mu_2 > \mu_1$. El estadístico natural que se usará es la diferencia de las medias muestrales $\bar{x}_2 - \bar{x}_1$ con una región crítica que será de la forma $\{\bar{x}_2 - \bar{x}_1 > c\}$. Lo difícil, en este caso, es encontrar un estadístico cuya distribución bajo H_o esté totalmente determinada y relacionado con la diferencia de las medias muestrales $\bar{x}_2 - \bar{x}_1$.

Suponiendo que los valores muestrales obtenidos de las panaderías siguen una distribución $\mathcal{N}(\mu_1, \sigma_1)$ y los valores muestrales obtenidos de los supermercados siguen una distribución $\mathcal{N}(\mu_2, \sigma_2)$, sabemos que $\bar{x}_1 \sim \mathcal{N}(\mu_1, \frac{\sigma_1}{\sqrt{70}})$ y $\bar{x}_2 \sim \mathcal{N}(\mu_2, \frac{\sigma_2}{\sqrt{40}})$. Como las dos muestras fueron extraídas de manera independiente, la varianza de la diferencia $\bar{x}_2 - \bar{x}_1$ es la suma de las varianzas: $Var(\bar{x}_2 - \bar{x}_1) = \frac{\sigma_1^2}{70} + \frac{\sigma_2^2}{40}$. Tenemos entonces

$$\bar{x}_2 - \bar{x}_1 \sim \mathcal{N}\left(\mu_2 - \mu_1, \sqrt{\frac{\sigma_1^2}{70} + \frac{\sigma_2^2}{40}}\right). \tag{2.3}$$

La desviación estándar $\sqrt{\frac{\sigma_1^2}{70} + \frac{\sigma_2^2}{40}}$ es desconocida. Usaremos entonces la distribución t-Student.

Proposición 2.5. *Si $\sigma_1^2 = \sigma_2^2$, el estadístico t de Student del test bajo la hipótesis $H_o : \mu_2 - \mu_1 = 0$ es*

$$T = \frac{(\bar{x}_2 - \bar{x}_1)/\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}{\sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}}} \sim t_{n_1+n_2-2}. \tag{2.4}$$

Demostración. Por una parte, de la ecuación 2.3, se tiene $Z = \frac{\bar{x}_2 - \bar{x}_1 - (\mu_2 - \mu_1)}{\sqrt{\frac{\sigma_1^2}{70} + \frac{\sigma_2^2}{40}}} \sim \mathcal{N}(0, 1)$. Por otra, si s_1^2 es la varianza muestral para las panaderías y s_2^2 para los supermercados, n_1 y n_2 los tamaños muestrales respectivos de panaderías y supermercados, entonces $U_1 = n_1 \frac{s_1^2}{\sigma_1^2} \sim \chi^2_{n_1-1}$ y $U_2 = n_2 \frac{s_2^2}{\sigma_2^2} \sim \chi^2_{n_2-1}$. Como las dos muestras son independientes, s_1^2 y s_2^2 son independientes. Además, $\bar{x}_1$ y s_1^2 son independientes entre sí y $\bar{x}_2$ y s_2^2 también (Proposición 2.2). Una primera consecuencia es que U_1 y U_2 son independientes y entonces $U_1 + U_2 \sim \chi^2_{n_1+n_2-2}$.

Construimos, entonces, la t-Student:

$$T = \frac{Z}{\sqrt{\frac{U_1 + U_2}{n_1 + n_2 - 2}}} \sim t_{n_1+n_2-2}.$$

Reemplazando Z, U_1 y U_2 por sus expresiones, obtenemos,

$$T = \frac{(\bar{x}_2 - \bar{x}_1 - (\mu_2 - \mu_1))/\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}{\sqrt{\frac{n_1 \frac{s_1^2}{\sigma_1^2} + n_2 \frac{s_2^2}{\sigma_2^2}}{n_1 + n_2 - 2}}} \sim t_{n_1+n_2-2}. \tag{2.5}$$

En el caso de una muestra, la ventaja de usar la t-Student está en que se elimina la desviación estándar σ del numerador y denominador (Proposición 2.3). No vemos que se produzca lo mismo en la expresión 2.5, salvo si $\sigma_1 = \sigma_2$ o $\sigma_1 = a\sigma_2$, donde a es conocido. Por lo general es difícil determinar el valor de a. Frecuentemente se toma $a = 1$. Se puede justificar comparando σ_1^2 y σ_2^2 mediante el test de las hipótesis dadas en 2.1 de la Sección 2.4 que involucra la distribución de Fisher. Lo que haremos

aquí, suponiendo que $\sigma_1 = \sigma_2$. Bajo la hipótesis nula $\mu_1 = \mu_2$, la expresión 2.5 se transforma entonces en:

$$T = \frac{(\bar{x}_2 - \bar{x}_1 - (\mu_2 - \mu_1))/\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}{\sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}}} \sim t_{n_1+n_2-2}. \tag{2.6}$$

□

Podemos proceder a construir la región crítica del test con una distribución t de Student con 108 grados de libertad. Bajo $H_o : \mu_1 = \mu_2$, para un error de tipo I de 5 %:

$$\mathbb{P}(t_{108} \geq 1,65) = 5\,\% \Longrightarrow \mathbb{P}\left(\frac{(\bar{x}_2 - \bar{x}_1)/\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}{\sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}}} \geq 1,65 \right) = 5\,\%$$

$$\Longrightarrow \mathbb{P}\left(\bar{x}_2 - \bar{x}_1 \geq 1,65 \times \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right) = 5\,\%$$

Reemplazando por los valores numéricos:

$$\mathbb{P}\left(\bar{x}_2 - \bar{x}_1 \geq 1,65 \times 0,198 \times 209,4\right) = 5\,\% \Longrightarrow \mathbb{P}\left(\bar{x}_2 - \bar{x}_1 \geq 68,5\right) = 5\,\%.$$

La región crítica es entonces $\mathcal{R} = \{\bar{x}_2 - \bar{x}_1 \geq 68,5\}$. Se rechaza H_o con un error de 5 %, puesto que encontramos que la diferencia entre las medias muestrales es $\bar{x}_2 - \bar{x}_1 = 120$ (Figura 2.5(a)).

Si queremos saber si podemos rechazar H_o para un error de tipo I menor que 5 %, calculamos el p-valor:

$$\mathbb{P}(\bar{x}_2 - \bar{x}_2 \geq 120) = \mathbb{P}\left(t_{108} \geq \frac{120/\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}{\sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}}} \right) = \mathbb{P}(t_{108} \geq 2,89) = 0,2\,\%.$$

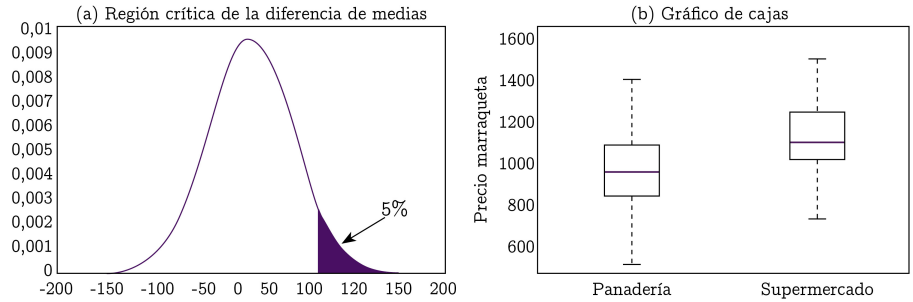
El p-valor, que es casi nulo, permite sentirse más seguro rechazando H_o. Se confirma que la diferencia es significativa con el gráfico de cajas de la Figura 2.5(b).

2.7.2 Comparación de dos medias en una población

Para evaluar el impacto de un laboratorio computacional para enseñar Estadística en 3°Medio se miden los logros de una muestra aleatoria de alumnos antes y después de su aplicación. Se comparan las medias de las notas obtenidas por los alumnos antes y después del laboratorio. Otro ejemplo: Para comprobar la eficacia de un tratamiento médico para bajar la presión arterial, se toma una muestra aleatoria de sujetos hipertensos y se mide su presión antes y después del tratamiento.

En ambos casos tenemos una sola muestra y una medición repetida para cada sujeto de la muestra. Entonces de las dos mediciones obtenidas de cada sujeto se

FIGURA 2.5. Comparación de los precios del pan



calculan dos medias, cuya comparación podría permitir comprobar la eficacia del laboratorio o del tratamiento.

Consideremos una muestra aleatoria de 150 alumnos de 3°Medio. A mitad de semestre, se les hace una prueba para medir sus logros antes de que se les aplique un laboratorio de Estadística. Al final del laboratorio, se les aplica la misma prueba. Cada uno de los 150 alumnos obtiene así dos notas: una antes y una después del laboratorio. Antes del laboratorio, la media muestral es $\bar{x}_1 = 4,85$ con una desviación estándar $s_1 = 0,94$, y después del laboratorio, la media muestral $\bar{x}_2 = 5,37$ y desviación estándar $s_2 = 1,14$. Parece que el laboratorio fue eficaz (Ver el gráfico de dispersión Figura 2.6(a) y gráfico de cajas 2.6(b)).

Si suponemos que la nota sigue una distribución $\mathcal{N}(\mu, \sigma)$, donde $\mu = \mu_1$ y $\sigma = \sigma_1$ son la media y desviación estándar antes del laboratorio, y $\mu = \mu_2$ y $\sigma = \sigma_2$ después, las hipótesis nula y alternativa son $H_o : \mu_2 = \mu_1$ contra $H_1 : \mu_2 > \mu_1$. No se espera que el laboratorio empeore los logros de los alumnos y H_o es la hipótesis que se espera rechazar.

Es importante constatar que las dos medias muestrales no son independientes, como es el caso con dos poblaciones (Sección 2.7.1). En efecto, se espera que las dos mediciones de un sujeto estén relacionadas. Si un sujeto tiene una buena nota en la primera prueba, se espera que su segunda nota sea en general parecida o mejor. De hecho, el coeficiente de correlación entre las dos notas es 0,73. Se recomienda describir los datos con un gráfico antes de aplicar un test de hipótesis. En el gráfico de dispersión (Figura 2.6(a)) de las dos notas, hay alumnos que no cambiaron su nota con el laboratorio y se encuentran sobre la recta Δ; alumnos que empeoraron después del laboratorio y se encuentran debajo de la recta, y finalmente alumnos que mejoraron después del laboratorio y están encima de la recta. Estos últimos son más numerosos que los otros.

Sabemos que el gráfico de cajas o boxplot (N. Lacourly [**7**]) permite visualizar características importantes de la distribución de una variable y mostrar las diferencias entre las dos notas también (Figura 2.6(b)). Estudiemos el gráfico, que muestra una

caja para la nota antes del laboratorio y una caja para la nota después de él. La línea gruesa horizontal dentro de una caja representa la mediana de la nota y permite ver dónde se posiciona la distribución. La caja contiene el 50 % de las distribución alrededor de la mediana y las otras dos líneas verticales posibilitan visualizar el recorrido de la distribución (extensión total de la distribución desde el mínimo hasta el máximo de los valores muestrales). Si las dos distribuciones son parecidas, las líneas medianas estarán más o menos al mismo nivel y los recorridos serán parecidos. En este caso, no habrá diferencia entre las notas antes y después del laboratorio. Lo importante es considerar las posiciones de las distribuciones en relación con los recorridos. La diferencia entre dos medianas no se interpreta de la misma manera si los recorridos son grandes o pequeños.

No podemos definir la desviación estándar de la diferencia de las dos medias muestrales de la misma manera que lo hicimos en 2.7.1. Sin embargo, vamos a ver que podemos calcularla directamente de las diferencias entre las dos notas.

Sean $\{x_{11}, x_{12}, ..., x_{1n}\}$ las notas antes del laboratorio y $\{x_{21}, x_{22}, ..., x_{2n}\}$ las notas después del laboratorio, donde $n = 150$. Definimos $\{d_1, d_2, ..., d_n\}$, las diferencias $d_i = x_{2i} - x_{1i}$.

Si $x_{1i} \sim \mathcal{N}(\mu_1, \sigma_1)$ y $x_{2i} \sim \mathcal{N}(\mu_2, \sigma_2)$, la diferencia $d_i \sim \mathcal{N}(\mu_1 - \mu_2, \sigma)$, donde σ depende de σ_1, σ_2 y de la relación entre las dos mediciones (covarianza). Si usamos las diferencias d_i, veremos que no necesitamos obtener σ de manera explícita a partir de σ_1 y σ_2. Basta usar el test de una media en una población de la Sección 2.7.2, donde las mediciones son las diferencias d_i y la media es la media de las diferencias. Las hipótesis nula y alternativa son entonces $H_o : \delta = 0$ contra $H_1 : \delta > 0$, donde $\delta = \mu_2 - \mu_1$.

La media muestral de las diferencias $\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i \sim \mathcal{N}(\mu_2 - \mu_1, \frac{\sigma}{\sqrt{n}})$.

Si $s^2 = \frac{1}{n}\sum_{i=1}^{n}(d_i - \bar{d})^2$ es la varianza muestral de las diferencias, $n\frac{s^2}{\sigma^2} \sim \chi^2_{n-1}$. Se deduce entonces que $\frac{\sqrt{n}(\bar{d}-(\mu_2-\mu_1))/\sigma}{\sqrt{n}\times s/(\sigma\times\sqrt{n-1})} = \frac{(\bar{d}-(\mu_2-\mu_1))}{s/\sqrt{n-1}} \sim t_{n-1}$.

Encontramos en la muestra $s = 0,79$. Deducimos la región crítica con $\alpha = 5\,\%$ y $\mu_1 - \mu_2 = 0$:

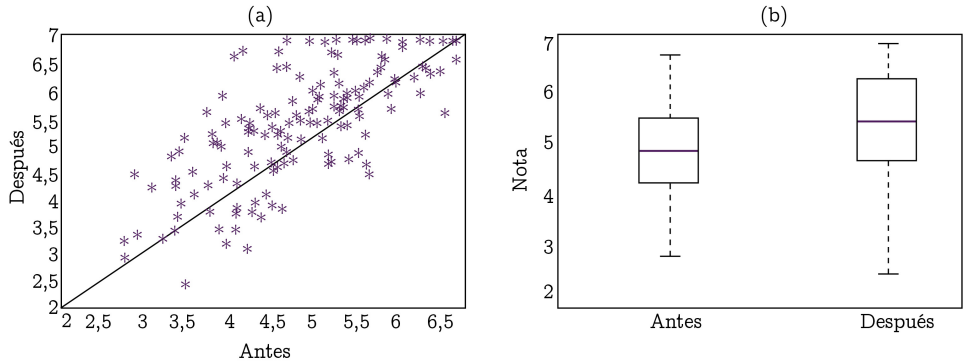
$$\mathbb{P}(t_{149} \geq 1,655) = 5\,\% \Longrightarrow \mathbb{P}\left(\frac{\bar{d}}{0,79/\sqrt{149}} \geq 1,655\right) = 5\,\%.$$

Se deduce la región crítica $\mathcal{R}_{0,05} = \{\bar{d} \geq 0,107\}$. La media $\bar{d}$ encontrada en la muestra fue $0,52$. Se rechaza H_o para $\alpha = 5\,\%$. Calculamos entonces el p-valor para saber si podemos rechazar con un error mucho más pequeño:

$$p\text{–}valor = \mathbb{P}(\bar{d} \geq 0,52) = \mathbb{P}(t_{149} \geq \frac{0,52}{0,79/\sqrt{149}}) = \mathbb{P}(t_{149} \geq 8,03) \approx 0,000.$$

Concluimos que el laboratorio fue muy eficaz.

FIGURA 2.6. Notas antes y después del laboratorio



Veamos otro ejemplo. Se aplica un tratamiento, que llamaremos ACME, a un grupo de 50 personas con hipertensión. A cada uno se le toma la presión antes y después del tratamiento. Comparamos las dos medias de la presión sistólica (La presión arterial tiene dos componentes: presión sistólica, que corresponde al valor máximo de la tensión arterial, cuando el corazón se contrae, y presión diastólica, que corresponde al valor mínimo de la tensión arterial entre latidos cardíacos). Se obtuvo antes del ACME una media muestral $\bar{x}_1 = 14,5$ y después del tratamiento, una media muestral $\bar{x}_2 = 10,45$, o sea, un promedio de las diferencias $\bar{d} = \bar{x}_1 - \bar{x}_2 = 4,06$. Sospechamos que el tratamiento es eficaz (Figuras 2.7(a) y (b)). Comprobemos si esta diferencia es suficiente para que el ACME sea considerado como un tratamiento eficaz mediante un test de hipótesis.

Se supone que la presión sistólica $x \sim \mathcal{N}(\mu, \sigma)$, donde $\mu = \mu_1$ es la media antes de tratamiento y $\mu = \mu_2$ después. Las hipótesis nula y alternativa son $H_o : \mu_1 = \mu_2$ contra $H_1 : \mu_2 < \mu_1$. En efecto, nos interesa solamente comprobar si disminuyó la presión con el tratamiento y queremos rechazar H_o a favor de H_1, que indica que el tratamiento es eficaz.

Para construir la región crítica, procedemos como en el caso del Laboratorio de Estadística. Se encontró en la muestra una desviación estándar de las diferencias de presión $s = 0,94$. Para $\alpha = 5\,\%$ y $\mu_1 - \mu_2 = 0$,

$$\mathbb{P}(t_{49} \geq 1,68) = 5\,\% \Longrightarrow \mathbb{P}\left(\frac{\bar{d}}{0,94/\sqrt{49}} \geq 1,68\right) = 5\,\%.$$

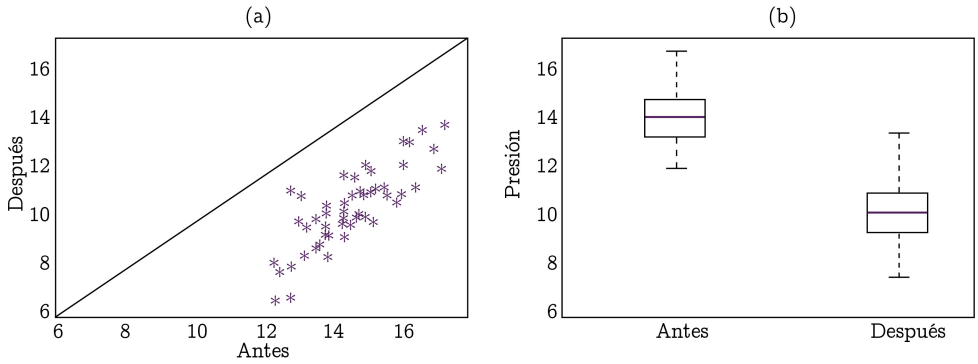
Se deduce la región crítica $\mathcal{R}_{0,05} = \{\bar{d} \geq 0,226\}$. La media $\bar{d}$ encontrada en la muestra fue $4,06$. Se rechaza H_o para $\alpha = 5\,\%$. Calculamos el p-valor para saber si podemos

rechazar con un error muy inferior:

$$p\text{–}valor = \mathbb{P}(\bar{d} \geq 4,06) = \mathbb{P}(t_{49} \geq \frac{4,06}{0,94/7}) = \mathbb{P}(t_{49} \geq 30,23) = 0,000$$

Concluimos que el ACME fue muy eficaz para bajar la presión.

FIGURA 2.7. Presión antes y después del tratamiento



2.8 Más de dos poblaciones: ANOVA

En la Sección 2.7.1 comparamos las medias de dos poblaciones distintas. En la práctica, es frecuente tener que compara más de dos poblaciones. Por ejemplo, queremos comparar tres métodos educativos distintos. En este caso, no podemos usar el test t-Student para dos poblaciones. Usamos en este caso el método de **Análisis de la varianza**, cuya abreviación es ANOVA[4]. La generalización del caso de dos medias en una población (Sección 2.7.2) a más de dos medias, como comparar varias notas sucesivas de un grupo de alumnos, no se aborda en esta monografía.

Consideremos entonces los tres métodos educativos $\mathcal{M}_1$, $\mathcal{M}_2$ y $\mathcal{M}_3$ para enseñar el Teorema de Pitágoras. Se eligen 450 alumnos, que se reparten al azar en tres grupos de 150. En un grupo se le enseña el teorema con el método $\mathcal{M}_1$, en otro grupo con el método $\mathcal{M}_2$ y en el último grupo con el método $\mathcal{M}_3$. Al final de la enseñanza del teorema de Pitágoras en los tres grupos, se aplica a los 450 alumnos una prueba sobre el tema.

Las distribuciones de las notas correspondientes a los tres métodos se presentan en el gráfico de cajas de la Figura 2.8 y los promedios y las varianzas de las notas por grupo se encuentran en la Tabla 2.3.

En el gráfico de cajas vemos que los tres recorridos son parecidos, pero que las medianas difieren. En la Tabla 2.3 las medias por método son diferentes y las varianzas

[4]ANOVA viene del inglés " Analysis Of Variance".

Tabla 2.3. Estadísticas de las notas por grupo

Método	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$	Total
Media	3,79	4,10	4,60	4,16
Varianza	0,24	0,28	0,25	0,37
Frecuencia	150	150	150	450

parecidas. En particular, observamos que el método $\mathcal{M}_3$ parece mejor que los otros dos y el método $\mathcal{M}_1$ parece más deficiente. Confirmemos ahora que los resultados son diferentes mediante un test estadístico.

Figura 2.8. Boxplot de los métodos

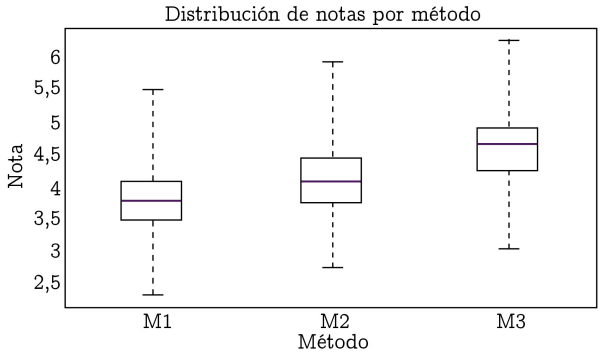


Tabla 2.4. Notaciones

Conjunto	Método	$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$	Total
Población	Media	μ_1	μ_2	μ_3	μ
	Varianza	σ_1^2	σ_2^2	σ_3^2	σ^2
Muestra	Media	$\bar{x}_1$	$\bar{x}_2$	$\bar{x}_3$	$\bar{x}$
	Varianza	s_1^2	s_2^2	s_3^2	s^2
	Tamaño	n_1	n_2	n_3	n

En vez de tomar las medianas, consideramos las medias; y en vez de tomar el recorrido, usaremos la varianza. Las notaciones utilizadas se encuentran en la Tabla 2.4.

Definamos ahora como hipótesis nula, las medias correspondientes a los tres métodos son iguales y como hipótesis alternativa, al menos una media difiere de las otras: $H_o : \mu_1 = \mu_2 = \mu_3 = \mu$ contra H_1: lo contrario.

En Tabla 2.3 vemos que al interior de los grupos las varianzas son parecidas (del orden de 0,25), pero que la varianza del total es mayor que la varianza de cada grupo (0,37), lo que hace pensar que en la varianza del total hay un efecto del cambio de posición de las distribuciones. Si, al contrario, las distribuciones tuvieran la misma media, la varianza del total sería parecida a las varianzas por grupo. Esta reflexión nos lleva a comparar estas varianzas. La siguiente proposición está en la base del ANOVA.

Proposición 2.6. *La varianza total s^2 se descompone en:*

$$s^2 = b^2 + w^2, \tag{2.7}$$

donde $b^2 = \frac{1}{n}\sum_{j=1}^{3} n_j(\bar{x}_j - \bar{x})^2$ es la varianza de las tres medias ponderando por el tamaño del grupo y $w^2 = \frac{1}{n}\sum_{j=1}^{3} n_j s_j^2$ es el promedio ponderado de las tres varianzas. Se llama **varianza intergrupos** *a b^2 y* **varianza intragrupos** *a w^2.*

Para la demostración, basta desarrollar la fórmula, que dejamos como ejercicio. Naturalmente ella sigue siendo válida cuando hay 2, 3 o más grupos.

Se presentan dos situaciones extremas:

- Si $b^2 = 0$, las tres medias x_j son iguales y $s^2 = w^2$. No hay diferencias entre los grupos.
- Si $w^2 = 0$, las tres varianzas s_j^2 son nulas y $s^2 = b^2$. Todos los valores del mismo grupo j son iguales a $\bar{x}_j$, la media del grupo. Los grupos difieren si las medias $\bar{x}_j$ difieren. En cada grupo, todos los valores son iguales a la media del grupo y los grupos difieren si sus medias difieren.

Comparando las varianzas b^2 y w^2, podemos determinar si los grupos son significativamente diferentes. Retomamos la hipótesis nula: $H_o : \mu_1 = \mu_2 = \mu_3 = \mu$. Si $\frac{b^2}{w^2}$ es grande, la varianza de las medias de los grupos es más importante que las varianzas al interior de los grupos. En este caso podemos decir que las medias de los grupos son diferentes y rechazar H_o. Utilizaremos entonces un estadístico basado en el cociente $\frac{b^2}{w^2}$ y que tenga una distribution F de Fisher.

Proposición 2.7. *Bajo la hipótesis nula $H_o : \mu_1 = \mu_2 = \mu_3 = \mu$ y suponiendo que $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma^2$, se tiene que*

$$\frac{b^2/2}{w^2/(n-3)} \sim F_{2,n-3}. \tag{2.8}$$

Demostración. [5] Sabemos que $n_1 \frac{s_1^2}{\sigma_1^2} \sim \chi^2_{n_1-1}$, $n_2 \frac{s_2^2}{\sigma_2^2} \sim \chi^2_{n_2-1}$ y $n_3 \frac{s_3^2}{\sigma_3^2} \sim \chi^2_{n_3-1}$, todos independientes entre sí. El supuesto $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma^2$ tiene sentido bajo la hipótesis nula, que pretende definir la igualdad de las distribuciones. Se tiene entonces

$$V = n\frac{w^2}{\sigma^2} = \frac{\sum_{j=1}^{3} n_j \; s_j^2}{\sigma^2} \sim \chi^2_{n-3}.$$

Por otra parte, $\bar{x}_j \sim \mathcal{N}(\mu_j, \frac{\sigma}{\sqrt{n_j}})$, $j = 1, 2, 3$, y $\bar{x} = \frac{1}{n}\sum_{j=1}^{3} n_j \; \bar{x}_j$. Se deduce que

$$U = n\frac{b^2}{\sigma^2} = \sum_{j=1}^{3} n_j \; \frac{(\bar{x}_j - \bar{x})^2}{\sigma^2} \sim \chi^2_2.$$

De la proposición 2.2, sabemos que las medias muestrales son independientes de las varianzas muestrales, de lo que se deduce que b^2 y w^2 son independientes. Usando la definición 2.4, se construye el estadístico con distribución de Fisher

$$F = \frac{U/2}{V/(n-3)} = \frac{b^2/2}{w^2/(n-3)} \sim F_{2,n-3}.$$

□

Obtenemos un estadístico cuya distribución es conocida y que permite comparar las dos varianzas intergrupos e intragrupos. Se rechaza H_o cuando b^2 es significativamente mayor que w^2. La región crítica del test es entonces de la forma $\mathcal{R}_\alpha = \{\frac{b^2/2}{w^2/(n-3)} \geq c_\alpha\}$.

Para $\alpha = 5\,\%$, $F \sim F_{2,447}$, y entonces encontramos $c_\alpha = 3,016$. Se puede usar la función FINV de Excel:

$$FINV(0,05; 2; 447) = 3,016.$$

En la muestra $nb^2 = 50,019$, $nw^2 = 115,331$ y el F observado $F_o = \frac{50,029/2}{115,331/447} = 96,93$, que es mayor que c_α. Se rechaza entonces que las tres medias son iguales con un error de $5\,\%$.

Calculamos ahora el p-valor, $\mathbb{P}(F_{2,447} \geq 96,93) \approx 0,00$. Se puede usar la función FDIST de Excel:

$$FDIST(96,93; 2; 447) = 0,0000.$$

Esto significa que se puede rechazar H_o con un error casi nulo. Se concluye que los métodos de enseñanza tienen efectos diferentes sobre el rendimiento de los alumnos.

Usualmente, se presentan los detalles de los cálculos del test en una tabla llamada "Tabla ANOVA" (Tabla 2.5), donde q es el número de grupos del estudio. La variable "método" representa lo que se llama **"factor"** y el método de comparación de varias

[5] En una primera lectura se puede omitir la demostración.

medias se llama **"ANOVA a un factor"**. Lo que no está explicado por el factor se llama "residuos", que es igual a la variabilidad intragrupos: nw^2. En efecto, si el factor influye sobre la variable X, la variabilidad de las medias de los grupos debe ser más importante que la variabilidad al interior de los grupos, de aquí que la variabilidad intragrupos es lo que se considera como "residuos" o "errores". Se puede agregar un factor al estudio, por ejemplo, la dependencia del colegio. En este caso se usa un ANOVA a dos factores, método que no se incluye en esta monografía.

Tabla 2.5. Tabla ANOVA

Fuente de varianza	Suma de cuadrados	Grados de libertad	Cuadrado medio	F-Fisher	p-valor
Factor método	$n\,b^2$	$q-1$	$\frac{n}{q-1}b^2$	F_o	$\mathbb{P}(F_{2,n-q} \geq F_o)$
Residuos	$n\,w^2$	$n-q$	$\frac{n}{n-q}w^2$		
Total	ns^2	$n-1$			

La Tabla ANOVA del caso (Tabla 2.6) muestra todos los cálculos numéricos que fueron desarrollados.

Tabla 2.6. Tabla ANOVA de los métodos de enseñanza

Fuente de varianza	Suma de cuadrados	Grados de libertad	Cuadrado medio	F	p-valor
Factor método	50,019	2	25,010	96,93	0,00
Residuos	115,331	447	0,258		
Total	164,35	449			

Un poco de historia. La foto de Sir Ronald Fisher en su escritorio a Whittinghome Lodge en 1952 fue obtenida del libro "R. A. Fisher, The Life of a Scientist", John Wiley & Sons 1978. Sir Ronald Fisher, matemático inglés, nació en Londres el 17 de febrero de 1890 y falleció en Adelaida, Australia, el 29 de julio de 1962. Los primeros resultados importantes de la estadística Matemática se deben al inglés Karl Pearson (1857-1936). Más recientemente, Sir Ronald Fisher, de la escuela biométrica inglesa, gran estadístico y gran genetista, tuvo mucha influencia en el campo de la genética y la agricultura. Desarrolló muchos métodos de la estadística inferencial, tales que el diseño de experimentos y el análisis de la varianza (ANOVA).

Consideramos ahora un experimento con 46 peces. Los peces fueron repartidos en tres acuarios (A1, A2 y A3), que tienen distintos niveles de radiactividad. Después de un tiempo se mide la radiactividad de los ojos y de los riñones de cada pez. Buscamos saber si las mediciones difieren de un acuario a otro.

Calculamos la media y varianza por acuario (Tabla 2.7) y construimos los boxplot (Figura 2.9(a) y (b)). Los boxplot muestran que los peces tienen un nivel de radioactividad de los ojos que difieren entre los acuarios, mientras que la radiactividad de los riñones no muestra diferencia. De los resultados de la Tabla y de los boxplot, se espera un efecto del factor "acuario" solo sobre la radiactividad de los ojos. Una explicación podría ser que el efecto del medio radiactivo de los acuarios es más importante sobre un órgano externo como los ojos, que sobre un órgano interno, como los riñones. Se confirman estas observaciones con un ANOVA para cada una de las dos variables con el factor "acuario" (Tablas 2.8 y 2.9). Para la radiactividad de lo ojos el p-valor es casi nulo, lo que muestra un efecto claro del acuario sobre esta variable. Ocurre lo contrario con la radiactividad de los riñones, cuyo p-valor es 0,905, que es demasiado alto para poder rechazar la igualdad de las medias.

Tabla 2.7. Resumen de los datos de los peces

Acuario	A1	A2	A3	Total
Frecuencia	16	16	14	46
	Radiactividad de los ojos			
Media	8,25	15,50	23,57	15,43
Varianza	2,44	14,75	32,22	53,90
	Radioactividad de los riñones			
Media	8,88	9,12	9,43	9,09
Varianza	13,72	3,05	18,42	10,96

FIGURA 2.9. Boxplot de los datos de los peces

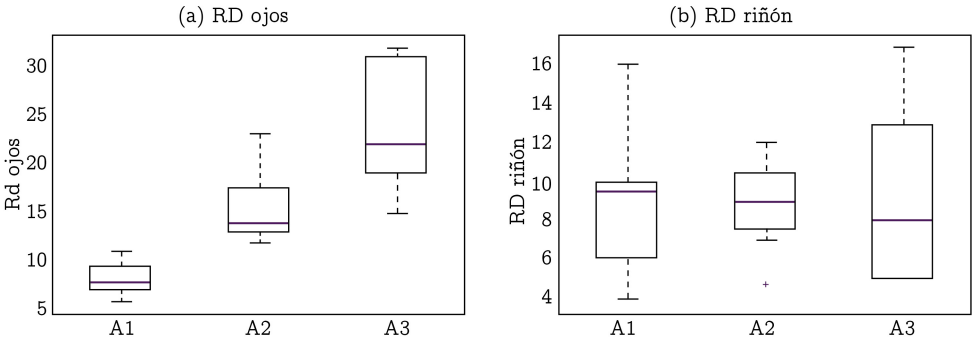


TABLA 2.8. Tabla ANOVA

Fuente de variabiliad	Suma de cuadrados	Grados de libertad	Cuadrado medio	F	p-valor
Radiactividad de los ojos					
Factor acuario	1752,88	2	876,44	55,88	0,000
Residuos	726,43	43	16,89		
Total	2479,30	45			
Radiactividad de los riñones					
Factor acuario	2,289	2	1,144	0,1	0,905
Residuos	490,929	43	11,417		
Total	493,217	45			

2.9 Resumen de la terminología

Población o universo: El conjunto de todos los objetos que se quiere estudiar.
Unidad estadística: Elemento sobre el cual se hacen mediciones.
Muestra: Subconjunto de la población.
Parámetro: Característica de la población que se busca estimar.
Valores muestrales: Valores de variables obtenidos en la muestra.

Muestreo: Método para obtener una muestra.
Muestreo aleatorio: Método de muestreo que selecciona la muestra de manera aleatoria.
Muestreo equiprobable: Muestreo que otorga la misma probabilidad de selección a todas las unidades de la población.

Muestreo aleatorio simple: Es un muestreo equiprobable en el cual se extraen las unidades una por una, de manera de que cada nueva unidad se obtiene del conjunto de las unidades no extraídas con equiprobabilidad.
Distribución en la población: Distribución de los valores de la variable de interés.
Distribución en el muestreo: Distribución de probabilidad de un estimador sobre todas las muestras posibles del mismo tamaño.
Error muestral: Error producido por el muestreo (método y tamaño de la muestra).
Error de Tipo I α: Rechazar la hipótesis nula cuando ésta es cierta.
Error de Tipo II β: Aceptar la hipótesis nula cuando ésta es falsa.
Región crítica o región de rechazo: Es el conjunto de valores de un test de hipótesis para los cuales la hipótesis nula es rechazada.
Potencia: Probabilidad que mide la habilidad de un test para rechazar la hipótesis nula cuando ésta es falsa. Es la probabilidad de tomar una decisión correcta. Vale $1 - \beta$.
Nivel de significación: Probabilidad del error de tipo I que está dispuesto a asumir.
p-valor: Probabilidad mínima de equivocarse con la cual se puede rechazar la hipótesis nula con el valor observado en la muestra del estadístico del test.
Varianza intragrupo: Promedio de las varianzas de una misma variable medida en varios grupos.
Varianza intergrupo: Varianza de los promedios de una misma variable medida en varios grupos.

2.10 Ejercicios

Ejercicio 2.1. ¿Cuáles de los casos siguientes requieren usar un test de hipótesis?

(a) El promedio de la PSU de Matemática de los alumnos de colegios particulares pagados es mayor que el de colegios municipales.
(b) El precio promedio de la bencina de 95 octanos es más caro en provincia que en la Región Metropolitana.
(c) El promedio en el curso de "biología de la célula", dictado por el profesor Juan Escobar, fue mayor que $4,5$.
(d) El kilo de pan es más barato en Concepción que en Santiago.

Ejercicio 2.2. ¿Los estudios siguientes requieren un estudio muestral o experimental? Justifique sus repuestas.

(a) El cambio en la malla curricular de la Enseñanza Media tiene un impacto positivo sobre el aprendizaje de los alumnos.
(b) Más del 25 % de los consumidores chilenos no miran los precios en el supermercado.
(c) Las piezas compuestas de plástico, fabricadas por la empresa Plásticos-Chile, son resistentes a altas temperaturas.
(d) La vacuna "oseltamivir" es eficaz para prevenir la gripe.

(e) Mientras más ácido es un lago, menos peces tiene.

Ejercicio 2.3. El peso medio de una muestra aleatoria de 81 personas de una determinada población es de 63,6 kg. Se sabe que la desviación estándar poblacional es de 6 kg. Con un nivel de significación del 0,05, ¿hay suficientes evidencias para rechazar la afirmación que la media poblacional del peso es de 65 kg?

Ejercicio 2.4. La vida media de una muestra de 100 tubos fluorescentes producidos por una empresa es de 1.570 horas, con una desviación estándar de 120 horas. ¿Se puede afirmar, con un nivel de significación de 5 %, que la duración media de los tubos es de 1.600 horas? Para esto:

(a) Determine las hipótesis nula y alternativa.
(b) Determine los errores de tipo I y II.
(c) Determine si el p-valor es mayor o menor que 5 %. Concluya.

Ejercicio 2.5. En el proceso de fabricación de una bebida se sabe que la producción por cadena tiene una media diaria de 500 litros con una desviación estándar de 100 litros. Se propone una modificación del proceso, con el objetivo de aumentar la producción diaria. Se implementó el nuevo proceso sobre una de las cadenas, que en una muestra de 60 días dio una producción promedio de 525 litros. ¿Podemos decir que el nuevo proceso es eficaz?

Ejercicio 2.6. Se desea estimar la proporción p de individuos daltónicos de una población a través del porcentaje observado en una muestra aleatoria de individuos de tamaño n. Si el porcentaje de individuos daltónicos en la muestra es igual al 30 %, calcule el valor de n para que, con un nivel de confianza del $0,95$, el error cometido en la estimación sea inferior al $3,1$ %. Ahora, si el tamaño de la muestra es de 64 individuos y el porcentaje de individuos daltónicos en la muestra es del 35 %, determine, usando un nivel de confianza del 1 %, el correspondiente intervalo de confianza para la proporción de daltónicos en la población. Deduzcan el error de estimación.

Ejercicio 2.7. El departamento de control de calidad de un laboratorio farmacéutico quiere revisar si los comprimidos de aspirina producidos por el laboratorio pesan 100 mg. En una muestra aleatoria de 350 comprimidos encuentran un promedio de 97 mg. con una desviación estándar de 78 mg. ¿El laboratorio tiene que revisar su proceso de producción?

Ejercicio 2.8. Se sabe que la renta anual de los individuos de una localidad sigue una distribución normal de media desconocida y con desviación estándar $0,24$ millones de pesos. Se ha observado además la renta anual de 16 individuos de esa localidad escogidos al azar, y se ha obtenido un valor medio de $1,6$ millones de pesos. Contraste, a un nivel de significación del 5 %, si la media de la distribución es de $1,45$ millones de pesos. Para esto:

(a) ¿Cuáles son las hipótesis nula y alternativa del contraste?

(b) Determine la forma de la región crítica y el p-valor.
(c) ¿Se acepta la hipótesis nula con el nivel de significación indicado?

Ejercicio 2.9. La normativa relativa a la contaminación atmosférica estipula que los motores de los vehículos no deben emitir más de 5 ppm (partes por millón) de CO_2. Dentro del proceso de control de calidad, un fabricante midió la emisión de CO_2 de una muestra aleatoria de 64 motores. Obtuvo una media de $5,5$ ppm con una desviación estándar de $0,6$ ppm. ¿Podemos decir que el fabricante cumple con la normativa para un nivel de significación del 1 %?

Ejercicio 2.10. Suponga que una muestra aleatoria de 12 observaciones del pH de una cierta proteína (ver la Tabla 2.9), que supondremos normal $\mathcal{N}(\mu_1, \sigma^2)$, son obtenidas para un determinado medio de cultivo que denotaremos A. Por otro lado, 14 observaciones aleatorias del pH de la misma proteína (ver la Tabla 2.10) son obtenidas para otro medio de cultivo, que denotaremos B, las cuales supondremos normales $\mathcal{N}(\mu_2, \sigma^2)$. Tanto μ_1, μ_2 y σ son desconocidos. Explicite todos los supuestos apropiados y realice un test de comparación de medias para decidir el test: $H_0 : \mu_1 = \mu_2$ contra $H_1 : \mu_1 \leq \mu_2$, para un nivel de significación del 5 %.

Tabla 2.9. Datos del pH: Método de cultivo A

3,03	3,14	3,26	2,69	2,28	3,83
3,42	2,70	3,05	2,95	1,71	4,87

Tabla 2.10. Datos del pH: Método de cultivo B

3,15	3,47	3,59	3,55	3,34	3,35	3,66
2,03	5,02	3,45	3,84	3,65	3,25	3,73

Ejercicio 2.11. Un campus universitario tiene cuatro facultades: Arquitectura, Medicina, Derecho e Ingeniería Civil. Se quiere estudiar el tiempo que tarda un alumno en hacer una consulta en la base de datos de la biblioteca de su facultad y analizar la influencia del factor facultad en el tiempo de consulta. Complete la siguiente tabla ANOVA y dé el número involucrados en el estudio. Concluya si existe diferencia entre las facultades:

Ejercicio 2.12. Un agricultor desea analizar la influencia de un fertilizante sobre la producción de choclos. Para esto, aplica distintas dosis de fertilizante sobre 100 terrenos de control (Ver Tabla 2.12). Realice un test ANOVA y concluya.

Tabla 2.11. ANOVA de las Facultades

Fuente de variabilidad	Suma de cuadrados	Grados de libertad	Cuadrados medios	F	p-valor
Facultad	4125,7				0,000
Residuos					
Total	6673,6	49			

Tabla 2.12. Tabla de Datos

Dosis	Frecuencia	Media	Desviación estándar
D1	20	16,23	1,71
D2	20	13,38	1,94
D3	30	10,94	1,856
D4	30	8,97	1,394
Total	100	11,9	3,174

Capítulo 3: Regresión lineal múltiple

En este capítulo se generaliza el modelo de regresión simple, que fue introducido en la monografía *Introducción a la Estadística* [**7**]. Se usa nuevamente el criterio de los mínimos cuadrados para estimar los coeficientes del modelo y el coeficiente de correlación múltiple para evaluar su calidad. Aquí se introduce un modelo probabilístico de los errores para construir test de hipótesis e intervalos de confianza para reforzar la crítica de los modelos. Se termina el capítulo con un caso que describe en detalle cómo interpretar los resultados en presencia de autocorrelaciones entre las variables utilizadas en el modelo. Una situación como la descrita se encuentra frecuentemente en la práctica y desconcierta por sus aparentes contradicciones.

3.1 Un poco de historia

La primera documentación del método de mínimos cuadrados se atribuye al matemático francés Adrien-Marie Legendre (1752-1833), quien, sin consideraciones probabilísticas, lo aplica en el área de la Astronomía para resolver sistemas de ecuaciones donde las observaciones pueden contener errores y el número de las mismas supera el número de incógnitas. El matemático alemán Carl Friedrich Gauss (1777-1855), con una diferencia de 4 años, publica un método similar que incluye la suposición de que los errores se distribuyen como una distribución Normal ("ley de los errores"). De esta forma comienza el desarrollo de los modelos de regresión.

Posterior a Legendre y Gauss, el matemático inglés Sir Francis Galton (1822-1911) utilizó el término regresión para referirse a la relación hereditaria existente entre las características de los padres traspasada a los hijos. El origen de la palabra regresión es histórico. Cuando Galton estudió la relación de la talla del hijo a partir de la talla de su padre, descubrió que los hijos de padres muy altos o muy bajos no lo eran tanto como sus padres, "regresaban" a valores medios. La formalización de la relación hereditaria fue realizada posteriormente por Karl Pearson (1895-1980), quien tomando en consideración 1000 datos de familias de los registros de Galton, determinó que

$$\text{Altura hijo} = 85\text{cm} + 0{,}5 \ \ \text{altura del padre}.$$

Pearson, en su constante búsqueda de una teoría consistente con la teoría de la evolución de Darwin, ideó métodos de gran relevancia estadística, entre ellos, el coeficiente de correlación. Hoy en día el término regresión se utiliza para denominar la predicción del valor de una variable desconocida en función del valor de otras variables conocidas y su aplicación se encuentra en todas las áreas de la ciencia.

3.2 Desarrollo de un ejemplo

Consideremos los datos del Programa de las Naciones Unidas para el Desarrollo (PNUD) del capítulo 1, con el Producto Nacional Bruto (PNB) per cápita, la Tasa de alfabetización (%) y el Número de usuarios de Internet (por 1000 habitantes) de 20 países de América Latina (Tabla 3.1).

Tabla 3.1. Datos del PNUD 2006

País	PNB per cápita	Tasa de alfabetización	Nº de usuarios de Internet
Argentina	14280	97,2	177
Bolivia	2819	86,7	52
Brasil	8402	88,6	195
Chile	12027	95,7	172
Colombia	7304	92,8	104
Costa Rica	10180	94,9	254
Cuba	6000	99,8	17
Ecuador	4341	91,0	47
El Salvador	5255	80,6	93
Guatemala	4568	69,1	79
Haití	1663	63,0	70
Honduras	3430	80,0	36
México	10751	91,6	181
Nicaragua	3674	76,7	27
Panamá	7605	91,9	64
Paraguay	4642	93,5	34
Perú	6039	87,9	164
Rep. Dominicana	8217	87,0	169
Uruguay	9962	96,8	193
Venezuela	6632	93,0	125

3.2.1 Presentación del modelo

Nos proponemos expresar el PNB como función de las otras dos variables. La función más simple es una función lineal:

$$y = b_o + b_1 x_1 + b_2 x_2, \tag{3.1}$$

donde y es el PNB, x_1 la Tasa de alfabetización y x_2 el Número de usuarios de Internet. Los coeficientes b_o, b_1 y b_2 son los parámetros del modelo lineal a estimar a partir de los datos empíricos. Tenemos una ecuación por país que lleva a resolver el sistema de 20 ecuaciones lineales con tres incógnitas siguiente:

$$\begin{cases} 14280 = b_o + 97,2b_1 + 177b_2 \\ 2819 = b_o + 86,7b_1 + 52b_2 \\ 8402 = b_o + 88,6b_1 + 195b_2 \\ 12027 = b_o + 95,7b_1 + 172b_2 \\ 7304 = b_o + 92,8b_1 + 104b_2 \\ 10180 = b_o + 94,9b_1 + 254b_2 \\ 6000 = b_o + 99,8b_1 + 17b_2 \\ 4341 = b_o + 91,0b_1 + 47b_2 \\ 5255 = b_o + 80,6b_1 + 93x_2 \\ 4568 = b_o + 69,1b_1 + 79b_2 \\ 1663 = b_o + 63,0b_1 + 70b_2 \\ 3430 = b_o + 80,0b_1 + 36b_2 \\ 10751 = b_o + 91,6b_1 + 181b_2 \\ 3674 = b_o + 76,7b_1 + 27b_2 \\ 7605 = b_o + 91,9b_1 + 64b_2 \\ 4642 = b_o + 93,5b_1 + 34b_2 \\ 6039 = b_o + 87,9b_1 + 164b_2 \\ 8217 = b_o + 87,0b_1 + 169b_2 \\ 9962 = b_o + 96,8b_1 + 193b_2 \\ 6632 = b_o + 93,0b_1 + 125b_2 \end{cases} \tag{3.2}$$

Este sistema de ecuaciones, en general, no tiene solución. Podría tenerla si fueran sólo tres ecuaciones o si el sistema fuera de rango 3, es decir, si sólo tres de las ecuaciones fueran linealmente independientes.

Antes de continuar, vamos a escribir el sistema de ecuaciones en forma matricial. Esta notación es más cómoda para el análisis que hacemos después. Y es un vector que contiene los valores del PNB per cápita y X es una matriz que contiene en columnas los valores de la Tasa de alfabetización, los Números de usuarios de Internet y un vector de "1", que corresponde a la constante.

$$Y = \begin{pmatrix} 14280 \\ 2819 \\ 8402 \\ 12027 \\ 7304 \\ 10180 \\ 6000 \\ 4341 \\ 5255 \\ 4568 \\ 1663 \\ 3430 \\ 10751 \\ 3674 \\ 7605 \\ 4642 \\ 6039 \\ 8217 \\ 9962 \\ 6632 \end{pmatrix} \quad y \quad X = \begin{pmatrix} 97,2 & 177 & 1 \\ 86,7 & 52 & 1 \\ 88,6 & 195 & 1 \\ 95,7 & 172 & 1 \\ 92,8 & 104 & 1 \\ 94,9 & 254 & 1 \\ 99,8 & 17 & 1 \\ 91 & 47 & 1 \\ 80,6 & 93 & 1 \\ 69,1 & 79 & 1 \\ 63 & 70 & 1 \\ 80 & 36 & 1 \\ 91,6 & 181 & 1 \\ 76,7 & 27 & 1 \\ 91,9 & 64 & 1 \\ 393,5 & 34 & 1 \\ 87,9 & 164 & 1 \\ 87 & 169 & 1 \\ 96,8 & 193 & 1 \\ 93 & 125 & 1 \end{pmatrix}. \tag{3.3}$$

De esta manera lel sistema de ecuaciones (3.2) se escribe simplemente como:

$$Y = Xb, \tag{3.4}$$

donde $b = \begin{pmatrix} b_1 \\ b_2 \\ b_o \end{pmatrix}$. Para el país i, la ecuación se escribe $y_i = b_o + b_1 x_{i1} + b_2 x_{i2}$, donde x_{i1} y x_{i2} son respectivamente los valores de la Tasa de alfabetización y del Número de usuarios de Internet del país i.

Como el sistema de ecuaciones (3.4) en general no tiene solución en la incógnita b, buscaremos una solución aproximada, que sea la más aceptable.

Los modelos lineales se caracterizan por sumar sus términos. Sumamos entonces un término de "error" e_i para cada país i, que expresa la diferencia entre el valor real y_i y la expresión del modelo $b_o + b_1 x_{i1} + b_2 x_{i2}$. La ecuación del país i es entonces:

$$y_i = b_o + b_1 x_{i1} + b_2 x_{i2} + e_i, \quad i = 1, 2, ..., 20, \tag{3.5}$$

y la ecuación matricial

$$Y = Xb + e,$$

donde e es el vector formado de los errores e_i.

El PNB tiene un rol diferente al de las otras variables en el modelo, así que se denominan de manera distinta. Se llama **variable a explicar, dependiente o**

respuesta al PNB, y **variables independientes o explicativas** a la Tasa de alfabetización y Número de usuarios de Internet. Definir las variables que tratan de explicar la variable respuesta PNB con "variables independientes" no significa que lo son y que es estrictamente necesario que lo sean. Sin embargo, es una propiedad deseada, como lo veremos en el caso presentado al final del capítulo.

3.2.2 Solución de los mínimos cuadrados

Nos encontramos ahora con más incógnitas que ecuaciones. Tenemos 20 ecuaciones y 23 incógnitas: los 20 errores $e_1,\ e_2,\ ...\ e_{20}$ y los tres parámetros del modelo b_o, b_1 y b_2. No podemos obtener solución única para las 23 incógnitas resolviendo el sistema de ecuaciones lineales (3.5). Queremos que todos los términos de errores sean lo más pequeños posible. Lamentablemente, si tomamos un error e_i muy pequeño o nulo, es muy probable que nos lleve a que los otros errores sean grandes. Por ejemplo, si $e_1 = e_2 = e_3 = 0$, obtenemos un sistema de ecuaciones para los tres primeros países de tres ecuaciones con tres incógnitas cuya solución es única:

$$b_o = -63245 \qquad b_1 = 744,5 \qquad b_2 = 29,15.$$

Podemos entonces deducir los otros 17 errores calculando $e_i = y_i - (-63245 + 744,5x_{i1} + 29,15x_{i2})$ (Tabla 3.2). Los 17 errores no nulos parecen grandes, algunos positivos y otros negativos. Con este procedimiento no podemos llegar fácilmente a una solución satisfactoria. Como no podemos anular o hacer pequeños de manera fácil todos los errores al mismo tiempo, ¿por qué no buscar los tres parámetros minimizando una función de los errores? Es así que Gauss propuso el **criterio de los mínimos cuadrados** (Ver Sección 3.1 y Osses[**12**]):

$$\min_{b_o,b_1,b_2} \sum_{i=1}^{20} e_i^2 = \min_{b_o,b_1,b_2} \sum_{i=1}^{20} (y_i - b_o - b_1 x_{i1} - b_2 x_{i2})^2. \tag{3.6}$$

Se podrían usar otras funciones de los errores, tales que

$$\min_{b_o,b_1,b_2} \sum_{i=1}^{20} |e_i| \quad o \quad \min_{b_o,b_1,b_2} (\max(e_1, e_2, ..., e_{20})).$$

Pero encontrar la solución con estos dos últimos criterios es numéricamente más complejo que utilizando el criterio de los mínimos cuadrados. Lo que hace interesante el criterio de los mínimos cuadrados es su simplicidad y que proporciona una solución explícita.

Buscamos entonces la solución del problema de minimización 3.6. Basta derivar la expresión:

$$Q = \sum_{i=1}^{20} (y_i - b_o - b_1 x_{i1} - b_2 x_{i2})^2,$$

Tabla 3.2

País	Error del modelo e_i
Argentina	0
Bolivia	0
Brasil	0
Chile	-990
Colombia	-1572
Costa Rica	-4632
Cuba	-5552
Ecuador	-1534
El Salvador	5782
Guatemala	14065
Haití	15964
Honduras	6066
México	524
Nicaragua	9029
Panamá	565
Paraguay	-2715
Perú	-938
Rep. Dominicana	1764
Uruguay	-4487
Venezuela	3005

con respecto a b_o, b_1 y b_2.

$$\begin{cases} \frac{\partial Q}{\partial b_o} = -2\sum_{i=1}^{20}(y_i - b_o - b_1 x_{i1} - b_2 x_{i2}) \\ \frac{\partial Q}{\partial b_1} = -2\sum_{i=1}^{20}(y_i - b_o - b_1 x_{i1} - b_2 x_{i2}) x_{i1} \\ \frac{\partial Q}{\partial b_2} = -2\sum_{i=1}^{20}(y_i - b_o - b_1 x_{i1} - b_2 x_{i2}) x_{i2}. \end{cases}$$

Otra alternativa consiste en escribir el criterio matricialmente:

$$Q = \sum_{i=1}^{20}(y_i - b_o - b_1 x_{i1} - b_2 x_{i2})^2 = (Y - Xb)^t (Y - Xb) = Y^t Y - 2b^t X^t Y + b^t X^t X b,$$

y usar la derivación matricial:

$$\frac{\partial Q}{\partial b} = -2X^t Y + 2X^t X b.$$

Anulando la derivada, obtenemos la expresión llamada *"ecuaciones normales"*

$$X^t X b = X^t y. \tag{3.7}$$

La expresión (3.7) es un sistema de tres ecuaciones lineales con tres incógnitas que podemos escribir como:

$$Au = c, \text{ con } A = X^tX \text{ y } c = X^tY, \tag{3.8}$$

donde $A = X^tX$ es una matriz cuadrada simétrica de orden 3. Un sistema de ecuaciones lineales $Au = c$, donde A es cuadrada, puede no tener solución. Sin embargo, el sistema de ecuaciones (3.7) es siempre determinado, es decir, tiene al menos una solución. Esto se debe a que la matriz A y el vector c están relacionados a través la misma matriz X, como veremos a continuación. Presentamos ahora la demostración de la existencia de solución del sistema (3.7) mediante el análisis algebraico, pero veremos, en la Sección 3.2.2, un resultado geométrico que permite justificar la existencia de una solución de manera mucho más simple.

Proposición 3.1. *El sistema de ecuaciones lineales $X^tXb = X^tY$ siempre tiene solución. Además, si las columnas de X son linealmente independientes, el sistema tiene una solución única igual a $(X^tX)^{-1}X^tY$.*

Demostración. Sea $\mathcal{H} = \{Au | u \in \mathbb{R}^3\}$ el conjunto llamado "conjunto imagen" de A y $\mathcal{K} = \{X^tv | v \in \mathbb{R}^n\}$ el conjunto imagen de X^t. $\mathcal{H}$ y $\mathcal{K}$ son ambos subconjuntos de $\mathbb{R}^3$, además $\mathcal{H} \subset \mathcal{K}$, dado que $A = X^tX$.

Por un lado, sabemos que el sistema de ecuaciones lineales $Au = c$ tiene solución si y solo si $c \in \mathcal{H}$. Por otro lado, dada una matriz B, se llama "rango" de B (denotado $r(B)$) al número máximo de columnas linealmente independientes de la matriz B, que es igual también al número máximo de filas linealmente independientes de la matriz B, y la dimensión del conjunto imagen de B es igual a $r(B)$. Por otra parte, se llama "núcleo" o "Kernel" de B al subconjunto $\mathcal{K}er(B) = \{v | Bv = 0\}$ y "nulidad"de B (denotada $n(B)$) a la dimensión de $\mathcal{K}er(B)$. Si B tiene p filas, obtenemos entonces el siguiente resultado:

$$n(B) + r(B) = p.$$

Se deduce que $r(X) = r(X^t)$. Mostremos entonces que $r(X^tX) = r(X^t)$. En efecto, si $X^ta = 0 \Longrightarrow X^tXa = 0$, o sea $r(X^tX) \subset \mathcal{K}er(X^t)$. Recíprocamente, si $X^tXa = 0 \Longrightarrow aX^tXa = 0 \Longrightarrow X^ta = 0$ y $\mathcal{K}er(X^t) \subset \mathcal{K}er(X^tX)$. Se deduce que $\mathcal{K}er(X^t) = \mathcal{K}er(X^tX)$; por lo tanto, $n(X^tX)(X^t)$ y $r(X^tX) = r(X^t)$ y entonces $\mathcal{H} = \mathcal{K}$. Finalmente obtenemos que $X^tY \in \mathcal{H}$. Por lo tanto, el sistema de ecuaciones lineales $X^tXb = X^tY$ siempre tiene solución.

Tenemos una solución única cuando la matriz X^tX es invertible, y ésta es invertible cuando la matriz X es de rango 3, es decir, cuando sus tres columnas son linealmente independientes. La solución es en este caso: $(X^tX)^{-1}X^tY$. □

Si la matriz X es de rango inferior a 3, se tiene una infinidad de soluciones. Para obtener una solución única, basta detectar cuál o cuáles de las columnas son linealmente dependientes de las otras. Entonces, se reduce el modelo, eliminando las columnas linealmente dependientes, o sea, las columnas redundantes. La solución del

modelo reducido tendrá una solución única. Obviamente, el modelo reducido no es único.

En este capítulo supondremos que la matriz $X'X$ es invertible y entonces que el modelo tiene una solución única denotada $\hat{b}$:

$$\hat{b} = (X^t X)^{-1} X^t Y.$$

Para el ejemplo de los países de América Latina, obtenemos la solución: $\hat{b}_o = -9789$, $\hat{b}_1 = 152,5$ y $\hat{b}_2 = 29,1$. Para el país i, se denota $\hat{y}_i$ el PNB estimado por el modelo y $\hat{e}_i$ el error resultante (Tabla 3.3), es decir:

$$\hat{y}_i = -9789 + 152,5x_{i1} + 29,1x_{i2}; \qquad \hat{e}_i = y_i - \hat{y}_i.$$

Hay que distinguir el modelo teórico (3.5) expresado matricialmente como $Y = Xb+e$, donde b es el vector de los parámetros por estimar y e es el vector de los errores teóricos del modelo estimado $Y = X\hat{b} + \hat{e}$. Los errores estimados $\hat{e}_i$ se llaman **residuos**. Observaciones:

(i) Es importante notar que la magnitud de un coeficiente $\hat{b}_j$, que depende de la escala de medición, no indica la magnitud del efecto de la variable j. La magnitud está dada por las cantidades $\hat{b}_j x_{ij}$, que no dependen de la escala de medición de la variable j. Veremos más adelante cómo evaluar los efectos de las variables.

(ii) La interpretación del signo de los coeficientes es importante. Por ejemplo, el coeficiente de la tasa de alfabetización es positivo, entonces, en el modelo, cuando aumenta la tasa de alfabetización, incrementa el PNB. Los términos de residuos, si son elevados, pueden cambiar este efecto. Ocurre lo mismo con los usuarios de Internet. ¿Que pasaría si en vez de tomar la tasa de alfabetización en el modelo, utilizamos la tasa de analfabetismo? El coeficiente asociado sería negativo.

Representaciones geométricas de la solución de los mínimos cuadrados

Tenemos dos representaciones geométricas posibles, en $\mathbb{R}^3$ o en $\mathbb{R}^{20}$, según se representen como vectores los países o las variables.

Consideramos en primer lugar las variables. Sean $\tilde{y}_i = b_o + b_1 x_{i1} + b_2 x_{i2}$ e $\tilde{Y} = Xb$ el vector de $\mathbb{R}^{20}$ formado por los componentes $\tilde{y}_i$, $i = 1, 2, ..., 20$. Sea W el conjunto imagen de X, que es el subespacio vectorial de $\mathbb{R}^{20}$ generado por las columnas de la matriz X. El criterio de los mínimos cuadrados, que consiste en minimizar $\sum_i e_i^2 = \|e\|^2 = \|Y - \tilde{Y}\|^2$, busca el vector $\tilde{Y}$ de W que minimiza la distancia entre el vector Y y el subespacio W. La solución es entonces la proyección ortogonal de Y sobre W, que llamaremos $\hat{Y}$ (Figura 3.1(a)). Se deduce que el vector de residuos $\hat{e}$ es ortogonal al vector estimado $\hat{Y}$. Una consecuencia es que $\hat{e}$ es ortogonal a todo vector del subespacio W y en particular al vector $\mathbb{1}$ formado de coordenadas constantes iguales a 1. De esto

Tabla 3.3

País	y_i	$\hat{y}_i$	$\hat{e}_i$
Argentina	14280	10179,7	4100,3
Bolivia	2819	4945,3	-2126,3
Brasil	8402	9391,2	-989,2
Chile	12027	9805,6	2221,4
Colombia	7304	7387,0	-83,0
Costa Rica	10180	12066,8	-1886,8
Cuba	6000	5926,0	74,0
Ecuador	4341 5	455,8	-1114,8
El Salvador	5255	5206,6	48,4
Guatemala	4568	3045,8	1522,2
Haití	1663	1853,9 -	190,9
Honduras	3430	3458,5	-28,5
s México	10751	9441,9	1309,1
Nicaragua	3674	2693,6	980,4
Panamá	7605	6087,2	1517,8
Paraguay	4642	5459,3	-817,3
Perú	6039	8383,5	-2344,5
Rep. Dominicana	8217	8391,6	-174,6
Uruguay	9962	10583,7	-621,7
Venezuela	6632	8027,8	-1395,8

concluimos que los residuos tienen su media nula:

$$\hat{e} \perp \mathbb{1} \Longrightarrow \mathbb{1}^t \hat{e} = 0 \Longrightarrow \sum_i \hat{e}_i = 0. \tag{3.9}$$

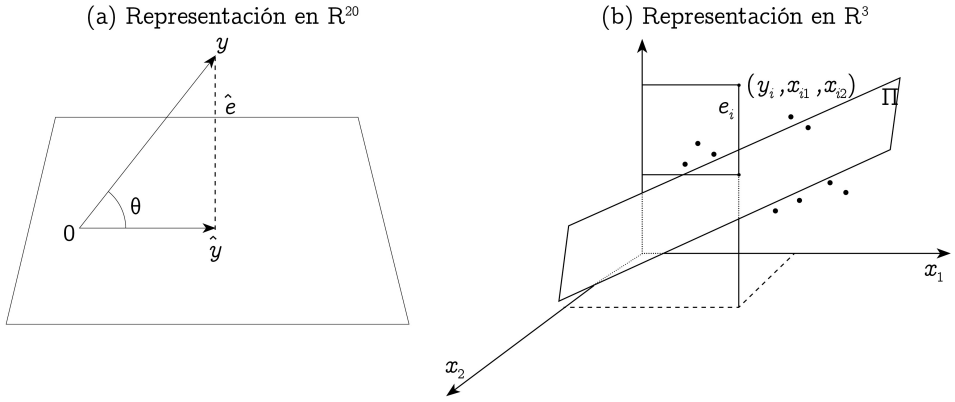
La representación geométrica en $\mathbb{R}^{20}$ del problema de los mínimos cuadrados permite justificar la existencia de solución de la ecuación (3.7), que mostramos algebraicamente en la Proposición 3.1. En efecto, la proyección ortogonal $\hat{Y}$ sobre el subespacio W siempre existe, luego $\hat{b}$ siempre existe, pues corresponde a los coeficientes que permiten expresar $\hat{Y}$ como combinación lineal de las columnas de la matriz X: $\hat{Y} = X\hat{b}$.

Si bien $\hat{Y}$ es siempre único, $\hat{b}$ no siempre lo es. Si las columnas de X son linealmente independientes, estas constituyen una base de W y entonces $\hat{b}$ es único. Cuando las columnas de X son linealmente dependientes, forman un conjunto generador de W, pero no constituyen una base. En este caso, hay una infinidad de maneras de escribir $\hat{Y}$ a partir del conjunto generador.

Ahora consideremos los países, que pueden representarse en $\mathbb{R}^3$. Cada país i es un vector z_i con tres coordenadas $z_i = (y_i, x_{i1}, x_{i2})$. Sea Π el plano de ecuación $y = \hat{b}_o + \hat{b}_1 x_1 + \hat{b}_2 x_2$. La estimación $\hat{y}_i$ de y_i se obtiene entonces proyectando el vector

z_i, que representa al país i sobre el plano Π paralelamente al eje de los y_i (Figura 3.1(b)). El residuo $\hat{e}_i$ es entonces el vector obtenido de la diferencia entre los vectores z_i y $\hat{y}_i$. El criterio de los mínimos cuadrados es entonces la suma de los cuadrados de las distancias entre los vectores z_i y el plano Π paralelamente al ejes de los y_i.

FIGURA 3.1. Representación geométrica del modelo



Evaluación de la solución de los mínimos cuadrados

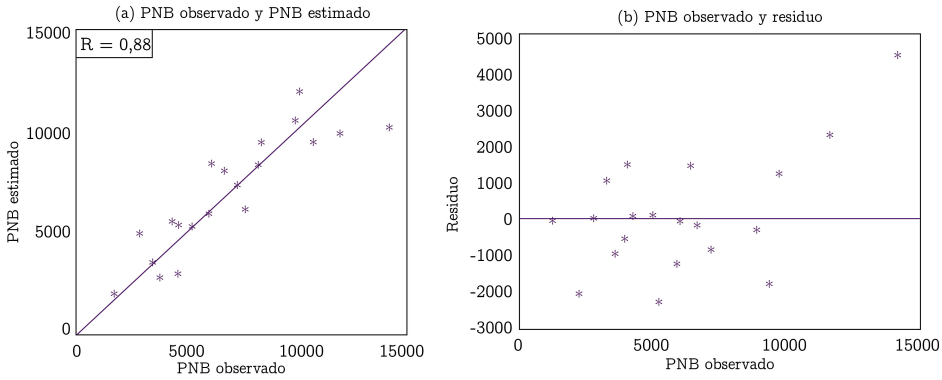
Una vez obtenida la solución de los mínimos cuadrados, tenemos que interpretar y evaluar el modelo estimado. A partir de las estimaciones $\hat{y}_i$ y los residuos $\hat{e}_i$, podemos construir un índice de calidad del modelo y dos gráficos que permiten comparar los valores observados y los valores estimados por el modelo.

Un modelo perfecto sería aquel que tiene todos los residuos $\hat{e}_i$ nulos. Al otro extremo, un modelo pésimo sería cuando los residuos son todos o casi todos iguales a los y_i. Obviamente, ninguno de estos dos casos ocurre en general. Un modelo con residuos no nulos puede ser "bueno". El problema es cómo definir un "buen" modelo ''. Salvo cuando el modelo es perfecto o pésimo, es un concepto subjetivo. Un modelo puede ser "bueno para una persona, pues le es muy útil, y el mismo modelo puede no ser tan "bueno" para otra.

Notemos que se usó el criterio de los mínimos cuadrados, por lo cual se espera de un buen modelo que los residuos sean pequeños, pero ¿qué tanto? Si ellos son pequeños, los valores de los PNB observados y_i y los PNB estimados $\hat{y}_i$ deberían ser parecidos. Además, si el modelo contiene todas las variables explicativas relevantes para explicar al PNB, los residuos no deberían depender del PNB, pues si fuera lo contrario, se pensaría que faltan una o más variables explicativas en el modelo o que el modelo no es lineal. En el gráfico de dispersión de los PNB observados y PNB

estimados, los 20 puntos deberían ubicarse cerca de la recta cuyos puntos tienen sus dos coordenadas iguales, recta que llamaremos "primera bisectriz" (Figura 3.2(a)). El gráfico muestra que los puntos son en general alineados con la primera bisectriz, salvo el punto con el PNB más alto. En el gráfico de dispersión de los PNB observados y_i y los residuos $\hat{e}$, los puntos se distribuyen alrededor del 0 sin mostrar alguna tendencia (Figura 3.2(b)).

FIGURA 3.2. Visualización de la calidad del modelo



Un índice que mide el grado de relación entre los y_i e $\hat{y}_i$ es el coeficiente de correlación lineal. Recordemos que el coeficiente de correlación lineal entre las variables X eY se calcula a partir de la covarianza $Cov(X,Y)$ y de las desviaciones estándar s_X y s_Y de las variables:

$$R = \frac{Cov(X,Y)}{s_X s_Y} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}}.$$

El coeficiente R varía entre -1 y $+1$. Cuando vale $+1$, los puntos (x_i, y_i) están alineados sobre una recta de pendiente positiva y cuando vale -1, los puntos están alineados sobre una recta de pendiente negativa. Cuando R es nulo, la distribución de los puntos no muestran ninguna tendencia lineal. Cuando R esta cercano a $+1$ o -1, los puntos no están alineados sobre una recta, pero existe, en general, una recta de la cual los puntos están cercanos. Para interpretar un coeficiente de correlación y el cuidado que se debe tener para interpretarlo vea Lacourly[**7**].

Como dijimos anteriormente, para aceptar el modelo tenemos que comprobar que los puntos $(y_i, \hat{y}_i)$ se encuentran cercanos a la primera bisectriz. El grado de cercanía

lo dará el coeficiente de correlación entre los y_i e $\hat{y}_i$:

$$R = \frac{\sum_i (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_i (y_i - \bar{y})^2}\sqrt{\sum_i (\hat{y}_i - \bar{\hat{y}})^2}} = \frac{\sum_i (y_i - \bar{y})(\hat{y}_i - \bar{y})}{\sqrt{\sum_i (y_i - \bar{y})^2}\sqrt{\sum_i (\hat{y}_i - \bar{y})^2}},$$

donde $\bar{y} = \frac{1}{20}\sum_i y_i$ y $\bar{\hat{y}} = \frac{1}{20}\sum_i \hat{y}_i = \frac{1}{20}\sum_i (y_i - \hat{e}_i) = \bar{y} - \frac{1}{20}\sum_i \hat{e}_i = \bar{y}$, dado que $\sum_i \hat{e}_i = 0$.

El coeficiente de correlación lineal entre los y_i e $\hat{y}_i$ se llama **coeficiente de correlación múltiple**.

Encontramos aquí que $R = 0,88$, lo que confirma una alto grado de relación lineal entre los y_i y los $\hat{y}_i$.

Usaremos el hecho de que los $\hat{y}_i$ se obtienen de los y_i para escribir el coeficiente R como el cociente de dos normas y un coseno. Se deducirá también que es siempre no negativo.

Proposición 3.2. *El coeficiente de correlación múltiple puede escribirse como*

$$R = \frac{\sqrt{\sum_i (\hat{y}_i - \bar{y})^2}}{\sqrt{\sum_i (y_i - \bar{y})^2}} = Coseno(Y, \hat{Y}).$$

Demostración. El numerador de la expresión de R puede escribirse en función de los $\hat{y}_i$. En efecto,

$$\sum_i (y_i - \bar{y})(\hat{y}_i - \bar{y}) = \sum_i y_i\hat{y}_i - \bar{y}\sum_i y_i - \bar{y}\sum_i \hat{y}_i + \sum_i \bar{y}_i^2 = \sum_i y_i\hat{y}_i - 20\bar{y}^2$$

$$\sum_i (y_i - \bar{y})(\hat{y}_i - \bar{y}) = \sum_i (\hat{y}_i + \hat{e}_i)\hat{y}_i - 20\bar{y}^2 = \sum_i \hat{y}_i^2 - 20\bar{y}^2 + \sum_i \hat{e}_i\hat{y}_i = \sum_i (\hat{y}_i - \bar{y})^2.$$

Se deduce que

$$R = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sqrt{\sum_i (y_i - \bar{y})^2}\sqrt{\sum_i (\hat{y}_i - \bar{y})^2}} = \frac{\sqrt{\sum_i (\hat{y}_i - \bar{y})^2}}{\sqrt{\sum_i (y_i - \bar{y})^2}}.$$

Geométricamente, en $\mathbb{R}^{20}$, R es el coseno del ángulo entre Y e $\hat{Y}$ (Figura 3.1(a)). □

Observe que $R \geq 0$. Se llama **coeficiente de determinación** a $R^2 = \dfrac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}$, que es el cociente de la varianza de los $\hat{y}_i$ con la varianza de los y_i. Podemos ver que el modelo produce una varianza de las estimaciones $\hat{y}_i$ menor que la varianza original de los y_i.

3.3 Estudio de la validez del modelo mediante tests de hipótesis

Vimos en el párrafo anterior el coeficiente de correlación múltiple y gráficos de dispersión, que son útiles para una primera crítica de la estimación del modelo, pero es importante confirmar los resultados con tests de hipótesis. Se requiere entonces un modelo probabilístico.

Se supone que los errores provienen de una distribución normal, $e_i \sim \mathcal{N}(0, \sigma)$, $i = 1, 2, ..., 20$, y que los e_i son independientes entre sí. La varianza σ^2 común a las 20 observaciones es desconocida. Si el modelo es correcto, este supuesto de distribución de los errores es bastante natural.

Previo a presentar los tests de hipótesis utilizados en regresión lineal, determinamos las distribuciones de los estadísticos $\hat{b}_j$ y $\sum_i \hat{e}_i^2$ obtenidos de la estimación de los mínimos cuadrados a partir de los supuestos definidos sobre los errores. Observen que cada error e_i se considera aleatorio, lo que hace que los y_i también lo sean. Aquí, supondremos que los x_{1i} y x_{2i} no son aleatorios.

3.3.1 Distribución de los $\hat{b}_j$

Se supone que los errores provienen de una distribución normal, $e_i \sim \mathcal{N}(0, \sigma)$, $i = 1, 2, ..., n$, y que los e_i son independientes entre sí con la varianza σ^2 desconocida y que los x_{1i} y x_{2i} no son aleatorios.

Consideremos la matriz $C = (X^t X)^{-1}$ y $C = (c_{ij})$, donde los c_{ij} son los elementos de la matriz C.[1]

Proposición 3.3. *El estimador $\hat{b}_j$ de b_j sigue una distribución $\mathcal{N}(b_j, \sigma_j)$, donde $\sigma_j = \sigma\sqrt{c_{jj}}$. Además la covarianza entre $\hat{b}_j$ y $\hat{b}_k$ es igual a $\sigma^2 c_{ik}$.*

Demostración. El estimador $\hat{b}_j$ de b_j es una combinación lineal de los y_i; por lo tanto, sigue una distribución Normal. Tenemos entonces que calcular su esperanza y varianza.

Los cálculos algebraicos son menos engorrosos si calculamos directamente la esperanza del vector $\hat{b}$ usando la linealidad del operador "esperanza" y el supuesto de

[1] Las demostraciones de las proposiciones 3.3 y 3.4 pueden omitirse en una primera lectura.

que la matriz X no es aleatoria,

$$\mathbb{E}(\hat{b}) = \mathbb{E}((X^tX)^{-1}X^tY) = (X^tX)^{-1}X^t\mathbb{E}(Y) = \\ (X^tX)^{-1}X^t\mathbb{E}(Xb+e) = (X^tX)^{-1}X^t\mathbb{E}(Xb) = b.$$

Se deduce que $\mathbb{E}(\hat{b}_j) = b_j$. Concluimos que $\hat{b}_j$ es un estimador insesgado de b_j.

$$Var(\hat{b}) = E[(\hat{b} - E(\hat{b}))(\hat{b} - E(\hat{b}))^t] = E[(\hat{b} - b)(\hat{b} - b)^t]\,. \tag{3.10}$$

Escribiendo $\hat{b} - b = (X^tX)^{-1}X^te$, obtenemos

$$\begin{aligned} Var(\hat{b}) &= E[(X^tX)^{-1}X^tee^tX(X^tX)^{-1}] \\ &= \left(X^tX\right)^{-1}X^tE\left(ee^t\right)X\left(X^tX\right)^{-1} \\ &= \sigma^2\left(X^tX\right)^{-1}X^tX\left(X^tX\right)^{-1} \\ &= \sigma^2\left(X^tX\right)^{-1}, \end{aligned}$$

dado que los errores son independientes entre sí y, por lo tanto, $\mathbb{E}\left(ee^t\right) = \sigma^2 I$. Lo que demuestra la proposición. □

Además, se puede mostrar que entre los estimadores insesgados de b_j, $\hat{b}_j$ es un estimador de mínima varianza (Teorema de Gauss-Markov). Lo que significa que si los supuestos sobre los errores son correctos, $\hat{b}_j$ es el estimador insesgado más preciso.

3.3.2 Distribución de los $\sum_i \hat{e}_i^2$

La distribución de los errores $e_i \sim \mathcal{N}(0,\sigma)$, $i = 1, 2, ..., n$, supone que los e_i son independientes entre sí y que la varianza σ^2 es desconocida. Para estimar la varianza σ^2 de los errores, podemos usar la varianza de los residuos. Vimos en la Sección 3.9, que la media de los residuos es nula: $\frac{1}{n}\sum \hat{e}_i = 0$. Luego $\frac{1}{n}\sum \hat{e}_i^2$ es la varianza de los residuos, y el estimador de σ^2 es $\hat{\sigma}^2 = \frac{1}{n}\sum_i \hat{e}_i^2$.

Proposición 3.4. *El estadístico* $n\frac{\hat{\sigma^2}}{\sigma^2} = \frac{\sum_i^n \hat{e}_i^2}{\sigma^2}$ *sigue una distribución* χ^2_{n-3}.

Demostración. Esta proposición no es fácil de demostrar con rigor en el marco de esta monografía. Sabemos que los $e_i \sim \mathcal{N}(0,\sigma)$ y que son independientes entre sí. Luego, $\sum_i^n \left(\frac{e_i}{\sigma}\right)^2 \sim \chi^2_n$.

Los residuos $\hat{e}_i$ son también variables normales, pero no son independientes entre sí, en particular $\sum_i^n \hat{e}_i = 0$. Sin embargo, se puede notar que el vector $\hat{e}$ pertenece a $W^\perp$, el subespacio ortogonal al subespacio vectorial W generado por las columnas de la matriz X (Sección 3.2.2). El subespacio W tiene dimensión 3, luego $W^\perp$ tiene dimensión n–3. El vector de residuos $\hat{e}$ puede, entonces, escribirse en una base de n–3 vectores de $W^\perp$, o sea que $\sum_i^n \hat{e}_i^2$ puede escribirse como la suma de $n-3$ variables independientes entre sí. De aquí podemos deducir que es posible escribir $\frac{\sum_i^n \hat{e}_i^2}{\sigma^2}$ como la suma de los cuadrados de $n-3$ variables $\mathcal{N}(0,1)$ independientes entre sí. □

Más generalmente, si el modelo de regresión tiene $p-1$ variables explicativas, hay p parámetros por estimar (la constante y los parámetros asociados a las $p-1$ variables explicativas). Entonces, $\frac{n\hat{\sigma^2}}{\sigma^2} = \frac{\sum_i^n \hat{e}_i^2}{\sigma^2} \sim \chi^2_{n-p}$.

3.3.3 Test de validez global

¿Qué pasaría si no tuviéramos ninguna variable explicativa en el modelo? Tendríamos un "modelo constante": $y_i = b_o$ para todo i. Nos preguntamos entonces si el aporte de las variables explicativas al modelo constante es significativo para explicar la variable respuesta. Para responder a la pregunta, comparamos dos modelos: uno sin las variables explicativas y el otro con las variables explicativas. Ponemos el primer modelo como hipótesis nula, pues buscaremos rechazarlo para probar que las variables explicativas aportan a la explicación de la variable PNB.

$$H_o : b_1 = b_2 = 0 \;\; \textit{contra} \;\; H_1 : b_1 \neq 0 \; y/o \; b_2 \neq 0.$$

$$\Updownarrow$$

$$H_o : \mathbb{E}(y_i) = b_o \;\; \forall \;\; i \;\; \textit{contra} \;\; H_1 : \mathbb{E}(y_i) = b_o + b_1 x_{i1} + b_2 x_{i2} \;\; \forall \;\; i.$$

Si se rechaza H_o, podemos decir que por lo menos una variable explicativa aporta información que permitirá estimar la variable respuesta PNB. Si no se rechaza la hipótesis H_o, entonces ninguna de las variables explicativas influye sobre la variable respuesta. En este caso el modelo es constante o bien las variables explicativas son otras.

El modelo bajo H_1 puede ser mejor o igual que bajo H_o, pero nunca peor. En efecto, el modelo de H_1 contiene al modelo de H_o. Hay una explicación geométrica también en $\mathbb{R}^{20}$ (Figura 3.1(a)). En el caso de H_o, el subespacio W es solamente una recta Δ (es la recta de los vectores con todas sus componentes iguales), que está contenida en el subespacio vectorial W de H_1. Se deduce que si $\hat{e}^o$ es el vector de los residuos bajo H_o, $\|\hat{e}^o\| \geq \|\hat{e}\|$. Además, si la diferencia $\|\hat{e}^o\|^2 - \|\hat{e}\|^2$ es pequeña, los modelos de H_o y H_1 son parecidos, y si, al contrario, la diferencia es grande, el modelo de H_1 es significativamente mejor que el modelo de H_o. Ahora bien, no está claro qué significa "pequeño" o "grande", siendo que la diferencia depende de la unidad de medición de los y_i, aquí el PNB. Vamos a construir una diferencia relativa, que permita evitar el problema de la escala de medición.

Consideremos en primer lugar la diferencia $SR = \|\hat{e}^o\|^2 - \|\hat{e}\|^2$, llamada *suma de cuadrados debida a la regresión*. Es fácil verificar que $\hat{e}_i^o = \bar{y} \; \forall i$. Se deduce entonces que

$$SR = \sum_i (y_i - \hat{e}_i)^2 - \sum_i (y_i - \hat{b}_i)^2 = \sum_i (y_i - \bar{y})^2 - \sum_i (y_i - \hat{b}_i)^2 = \sum_i (\hat{y}_i - \bar{y})^2.$$

Llamamos *suma de cuadrados debida a los residuos del modelo* a $SE = \sum_i \hat{e}_i^2$ y consideramos el cociente $\frac{SR}{SE}$, que es una expresión que no depende de la escala de medición de los y_i. Falta encontrar una distribución en relación con este cociente, que permita calcular la región crítica y el p-valor del test. Se puede demostrar que, bajo

H_o, $\frac{SR}{\sigma^2} \sim \chi^2_2$ y $\frac{SE}{\sigma^2} \sim \chi^2_{n-p}$, donde $n = 20$ es el número de observaciones y $p = 3$ es el número de parámetros del modelo. Además se puede demostrar que SR y SE son independientes. A partir de las dos variables χ^2 se puede construir una variable con una distribución de Fisher (Ver la definición 2.4 de la Sección 2.4 del Capítulo anterior). Se deduce entonces que bajo H_o,

$$F = \frac{SR/p}{SE/(n-p-1)} \sim F_{p,n-p-1}.$$

Rechazaremos H_o si el valor F_o del cociente calculado con los datos empíricos es grande. La región crítica del test es entonces de la forma $F \geq c$, donde $\mathbb{P}(F_{p,n-p-1} > F) = \alpha$. Rechazamos entonces H_o con un error α si $F_o \geq c$. El p-valor se obtiene del valor observado F_o:

$$\mathbb{P}(F_{p,n-p-1} \geq F_o).$$

En los paquetes computaciones de estadística se presentan usualmente los resultados de este test en una tabla llamada ANOVA (Tabla 3.4). El numerador SR/p y el denominador $SE/(n-p-1)$ de F se llaman "cuadrados medios". Observen que similarmente al método ANOVA visto en el Capítulo 2, la variabilidad total se descompone en variabilidad debida a la regresión y variabilidad debida a los residuos:

$$ST = \sum_i (y_i - \bar{y})^2 = SR + SE.$$

Tabla 3.4. Presentación de la tabla ANOVA

Fuente	Suma de cuadrados	Grados de libertad	Cuadrado medio	F	p-valor
Regresión	SR	p	SR/p	$\frac{SR/p}{SE/(n-p-1}$	$\mathbb{P}(F_{p,n-p-1} \geq \frac{SR/p}{SE/(n-p-1})$
Residuos	SE	$n-p-1$	$SE/(n-p-1)$		
Total	$SR+SE$	$n-1$			

Calculamos los elementos de la tabla ANOVA para el ejemplo de los países (Tabla 3.5). El p-valor es nulo, lo que permite concluir que el modelo tiene una cierta validez.

Tabla 3.5. Tabla ANOVA del ejemplo

Fuente	Suma de cuadrados	Grados de libertad	Cuadrado medio	F	p-valor
Regresión	160.380.000	2	80.190.000	28,44	0,000
Residuos	47.928.000	17	2.819.300		
Total	208.308.000	19			

Ejercicio: Muestren que el cociente F puede expresarse a partir del coeficiente de determinación R^2:

$$F = \frac{SR/p}{SE/(n-p-1)} = \frac{R^2/p}{(1-R^2)/(n-p-1)}.$$

3.3.4 Test local e intervalo de confianza para los parámetros b_j

Si se concluye que el modelo es válido globalmente debemos preguntarnos si todas las variables tienen la misma influencia sobre la variable respuesta PNB. La respuesta viene dada por los tests de validez local, que compara los modelos con o sin una de las variables explicativas. Si se trata de la variable j, las hipótesis nula y alternativa se escriben: $H_o : b_j = 0$ y $H_1 : b_j \neq 0$.

Vimos anteriormente que $\hat{b}_j \sim \mathcal{N}(b_j, \sigma_j)$, donde $\sigma_j = \sigma\sqrt{c_{jj}}$ (Sección 3.3.1) y $\frac{\sum_i \hat{e}_i^2}{\sigma^2}\chi^2_{n-p}$ (Sección 3.3.3).

Se puede entonces construir un estadístico t-Student con $Z = \frac{\hat{b}_j - b_j}{\sigma_j} \sim \mathcal{N}(0,1)$ y $U = \frac{\sum_i \hat{e}_i^2}{\sigma^2} \sim \chi^2_{n-p}$:

$$T_j = \frac{(\hat{b}_j - b_j)/\sigma_j}{\sqrt{U/(n-p-1)}}.$$

Esta expresión se simplifica definiendo $\hat{\sigma}_j^2 = \sqrt{c_{jj}}\frac{\sum_i \hat{e}_i^2}{n-p-1}$ como estimador de σ_j:

$$T_j = \frac{(\hat{b}_j - b_j)/\sigma_j}{\hat{\sigma}_j/\sigma_j} = \frac{\hat{b}_j - b_j}{\hat{\sigma}_j}.$$

Se puede demostrar que $\hat{b}_j$ y $\hat{\sigma}_j^2$ son independientes y se puede deducir que $T_j \sim t_{n-p-1}$, una distribución de Student con $n-p-1$ grados de libertad (Ver la definición 2.3 de la Sección 2.4 del Capítulo 2).

Tenemos aquí una hipótesis alternativa bilateral, lo que lleva a tomar una región crítica compuesta de dos intervalos situados en ambos extremos de la distribución de T_j (Ver el caso 3 de la Sección 2.7.2 del Capítulo 2). Bajo la hipótesis nula $H_o : b_j = 0$ contra $H_1 : b_j \neq 0$, la región crítica del test es de la forma

$$\mathcal{R} = \{\hat{b}_j \geq c_2\} \bigcap \{\hat{b}_j \leq c_1\}.$$

En los paquetes computaciones de Estadística se prefiere en general usar la estrategia del p-valor, que proporciona la probabilidad de equivocarse rechazando la hipótesis nula $H_o : b_j = 0$ cuando ésta es cierta. Los resultados de los test locales se presentan de la siguiente manera:

Con los datos de los 20 países, los tres coeficientes aparecen con p-valores muy pequeños, lo que permite concluir que las tres variables son significativas (Tabla 3.7).

Por otra parte, se pueden construir intervalos de confianza para los parámetros b_j utilizando la misma distribución de Student. Para un nivel de confianza del 95 %

TABLA 3.6. Resultados para los coeficientes

Variable	Coeficiente	Desv. estándar	T- Student	p-valor
Constante	$\hat{b}_o$	$\hat{\sigma}_o$	$T_o = \hat{b}_o/\hat{\sigma}_o$	$\mathbb{P}(\lvert t_{n-p}\rvert \geq T_o)$
Alfabetización	$\hat{b}_1$	$\hat{\sigma}_1$	$T_1 = h\beta_1/\hat{\sigma}_1$	$\mathbb{P}(\lvert t_{n-p}\rvert \geq T_1)$
Usuarios Internet	$\hat{b}_2$	$\hat{\sigma}_2$	$T_2 = \hat{b}_2/\hat{\sigma}_2$	$\mathbb{P}(\lvert t_{n-p}\rvert \geq T_2)$

$\mathbb{P}(t_{17} \geq 2,11) = \mathbb{P}(t_{17} \leq -2,11) = 0,025$. El intervalo es entonces

$$[\hat{b}_j - 2,11\hat{\sigma}_j;\ \hat{b}_j + 2,11\hat{\sigma}_j],$$

que aparece en la Tabla 3.7 con el título "IC".

TABLA 3.7. Resultados para los coeficientes

Variable	Coeficiente	Desv. estándar	T- Student	p-valor	IC (95 %)
Constante	-9789	3606	-2,715	0,007	[-17.397; -2.181]
Alfabetización	152,5	42,86	3,559	0,001	[62,09; 242,93]
Usuarios Internet	29,1	5,86	4,958	0,000	[16,70; 41,43]

3.4 Predicción

Una vez construido el modelo y obtenidas las estimaciones de los parámetros, podemos utilizarlo para hacer estimaciones de nuevos datos, en los cuales se conocen solamente las variables explicativas. Los nuevos datos deben ser parte de la misma población. Por ejemplo, podríamos aplicar el modelo a Surinam, que es de la misma región que los 20 países del modelo. Su alfabetización es de 89,6 % y usuarios de Internet 71 por mil habitantes. La estimación del PNB de Surinam es entonces:

$$\widehat{PNB}_S = -9789 + 152,5 \times 89,6 + 29,1 \times 71 = 5941,1.$$

Un intervalo de confianza para el PNB de Surinam se obtiene a partir de las varianzas y covarianzas de $\hat{b}$ dadas en la matriz $V = \hat{\sigma}^2(X^tX)^{-1}$, donde $\hat{\sigma}^2 = 2819300$. Si el vector de los nuevos datos es $x_S = (1 \ \ 89,6 \ \ 71)^t$, entonces el intervalo de nivel 95 % se calcula

$$[\widehat{PNB}_S - 2,11u;\ \widehat{PNB}_S + 2,11u]$$

con $u = \sqrt{x_S V x_S^t} = 1679,1 \times 0,2785 = 467,6$. Se obtiene el intervalo del PNB para un nivel de confianza de 95 %:

$$IC_S = [5941,1 - 2,11 \times 467,6;\ 5941,1 + 2,11 \times 467,6] = [4954,5;\ 6927,8],$$

donde $\mathbb{P}(t_{17} \leq -2,11) = \mathbb{P}(t_{17} \geq 2,11) = 0,05$. Si revisan en Internet, el valor real del PNB de Surinam en 2008 fue 7722, que queda fuera del intervalo. Surinam es un

país de la misma región que los 20 países de América Latina, pero es muy distinto políticamente, lo que podría explicar que el modelo subestima el PNB de Surinam.

3.5 Estudio de un caso

Una universidad quiere determinar cuáles son las variables que permiten hacer un pronóstico del rendimiento en el primer año de los alumnos de la carrera de Pedagogía en Matemática. Se han recopilado informaciones de 70 alumnos que ingresaron en esta carrera[2]. Los datos son:

x_1 : Promedio de Música de la Enseñanza Media (EM).
x_2 : Promedio de Gimnasia de la Enseñanza Media (EM).
x_3 : Promedio de Castellano de la Enseñanza Media (EM).
x_4 : PSU de Matemática.
y : Promedio final de notas del primer año de la carrera de Pedagogía Matemática.

Usamos un modelo de regresión:

$$(\mathcal{M}_1) \quad y_i = b_o + b_1 x_{i1} + b_2 x_{i2} + b_3 x_{i3} + b_4 x_{i4} + e_i, \;\; i = 1, 2, ..., 70.$$

Los resultados del modelo $\mathcal{M}_1$ son presentados en las Tablas 3.8 y 3.9. Encontramos un coeficiente de correlación múltiple de 0,92, lo que es relativamente alto. El p-valor de la F del ANOVA es casi nulo, por lo que podemos decir que el modelo $\mathcal{M}_1$ es significativo. Parece que podríamos inferir "bastante bien" el rendimiento de un alumno de primer año de la carrera de Pedagogía Matemática a partir de las cuatro notas. Hagamos un análisis más profundo de la validez del modelo, examinando el gráfico de dispersión de la nota promedio de primer año observada y de la nota estimada (Figura 3.3(a)) y del gráfico de dispersión de la nota promedio de primer año observada y de los residuos (Figura 3.3(b)). En el primer gráfico se espera que los puntos estén alineados sobre la primera bisectriz, y el segundo gráfico no debería mostrar relación entre los residuos y la notas finales de primer año. Estas dos condiciones no se cumplen. En el primer gráfico, los alumnos con notas bajas o altas son alejados de la primera bisectriz. Se ve como una relación logarítmica. Los residuos correspondientes a las notas bajas o altas son son entonces grandes. Esto deja pensar que faltan variables explicativas o que el modelo no es lineal. No tenemos más variables explicativas, así que probamos otros modelos con las mismas variables, pero aplicando algunas transformaciones a una o más variables del modelo inicial y manteniendo un modelo lineal. La transformación que resultó más adecuada fue la aplicación del logaritmo a la nota final de primer año. Dio como resultado el siguiente modelo:

$$(\mathcal{M}_2) \quad ln(y_i) = b_o + b_1 x_{i1} + b_2 x_{i2} + b_3 x_{i3} + b_4 x_{i4} + e_i \;\; (\forall i),$$

[2]Los datos utilizados son ficticios, pero suficientemente realistas para una buena comprensión del ejemplo.

El modelo mejoró claramente. No solo el coeficiente de correlación múltiple creció de 0,92 a 0,978, sino que los gráficos reflejan las dos condiciones que debería cumplir un modelo aceptable (Figuras 3.4(a) y (b)). Los p-valores del test global (Tabla 3.10) y de los test locales (Tablas 3.11) son significativos, salvo la PSU de Matemática, cuyo p-valor vale 0,471, lo que es demasiado alto para rechazar que el parámetro asociado es nulo. La PSU de Matemática parece no influir sobre la nota final de primer año.

Se percibe una contradicción. En efecto, si examinamos la matriz de correlaciones, observamos que el coeficiente de correlación del logaritmo de la nota final del primer año con la PSU de Matemática vale 0,741 y es la variable con mayor correlación (Tabla 3.12). Parece una contradicción con el hecho de no encontrar la PSU de Matemática con un efecto significativo en el modelo $\mathcal{M}_2$. De hecho, si eliminamos la PSU de Matemática del modelo $\mathcal{M}_2$, el coeficiente de correlación múltiple sigue igual a $0,988$. Sin embargo, si tuviéramos que poner una sola variable explicativa en el modelo, de los cuatro promedios sería la mejor variable. ¿Cómo explicar esta contradicción? Examinando con más cuidado la matriz de correlaciones, vemos que la PSU de Matemática esta correlacionada también con las notas de música, gimnasia y castellano, pero que estas tres últimas notas son poco correlacionadas entre sí. Cuando estas tres variables explicativas están en el modelo, la PSU de Matemática está ya representada a través de éstas. Es un fenómeno que puede presentarse cuando las variables explicativas no son independientes entre sí[3]. Se habla de "paradoja de Simpson", si la asociación entre dos variables cambia cuando se controla el efecto de una tercera variable.

Tabla 3.8. Tabla ANOVA del modelo ($\mathcal{M}_1$)

Fuente	Suma de cuadrados	Grados de libertad	Cuadrado medio	F	p-valor
Regresión	57,766	4	14,442	91,404	0,000
Residuos	10,263	65	0,158		
Total	68,029	69			

[3]De aquí que se llaman las variables explicativas "variables independientes".

TABLA 3.9. Coeficientes del modelo ($\mathcal{M}_1$)

Variable	Coeficiente	Desv. estándar	T- Student	p-valor
Constante	-10,232	0,72 5	-14,108	0,000
Música	1,035	0,113	9,183	0,000
Gimnasia	0,578	0,091	6,326	0,000
Castellano	0,624	0,064	9,705	0,000
PSU Matemática	0,0015	0,0012	1,259	0,212

FIGURA 3.3. Gráficos de validez modelo ($\mathcal{M}_1$)

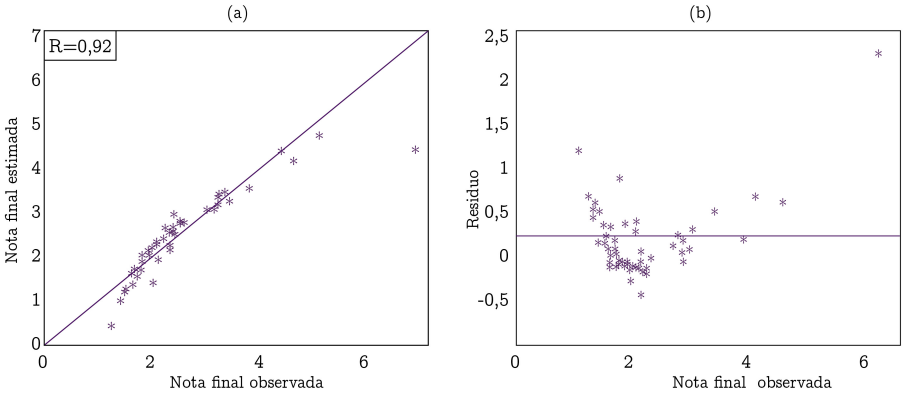


FIGURA 3.4. Gráficos de validez modelo ($\mathcal{M}_2$)

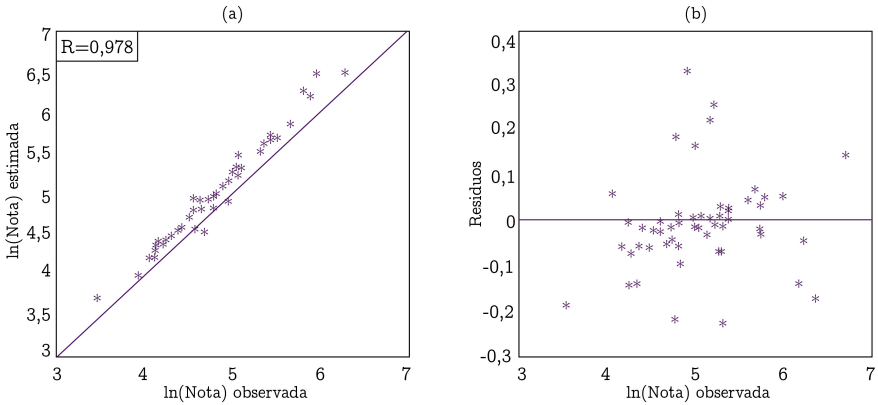


Tabla 3.10. Tabla ANOVA del modelo ($\mathcal{M}_2$)

Fuente	Suma de cuadrados	Grados de libertad	Cuadrado medio	F	p-valor
Regresión	7,967	4	1,992	355,7	0,000
Residuos	0,365	65	0,0056		
Total	8,332	69			

Tabla 3.11. Coeficientes del modelo ($\mathcal{M}_2$)

Variable	Coeficiente	Desv. estándar	T- Student	p-valor
Constante	-3,909	0,331	11,82	0,000
Música	0,407	0,0160	25,50	0,000
Gimnasia	0,204	0,0136	15,01	0,000
Castellano	0,240	0,0088	27,40	0,000
PSU Matemática	0,0004	0,0005	0,724	0,471

Tabla 3.12. Matriz de correlaciones para el modelo ($\mathcal{M}_2$)

	Música	Gimnasia	Castellano	PSU Matemática	ln(Nota)
Música	1,000	0,092	-0,055	0,390	0,606
Gimnasia	0,092	1,000	-0,228	0,449	0,303
Castellano	-0,055	-0,228	1,000	0,403	0,629
PSU Matemática	0,390	0,449	0,402	1,000	0,741
ln(Nota)	0,606	0,303	0,629	0,741	1,000

3.6 Resumen de la terminología

Coeficiente de correlación lineal: Índice estadístico que mide la relación lineal entre dos variables cuantitativas. No indica causalidad.
Regresión lineal simple: Es un método matemático que expresa una relación lineal de una variable numérica a partir de otra. El modelo no es simétrico entre las dos variables.

Variable a explicar o respuesta: Es una variable que depende del valor que toman otras variables.
Variable explicativa o independiente: Los cambios en los valores de una variable explicativa determinan cambios en los valores de otras (variable dependiente).
Error del modelo: Es la diferencia entre el valor observado de la variable a

explicar con el estimador obtenido del modelo.
Residuos: Son las estimaciones de los errores del modelo obtenidas de los datos observados.

Criterio de mínimos cuadrados: Es la función de los errores teóricos del modelo que se minimiza para estimar los parámetros del modelo.
Predicción: Estimaciones de la variable respuesta que se obtienen de un modelo.

3.7 Ejercicios

Ejercicio 3.1. Se realiza el estudio del rendimiento R del primer año de una muestra de alumnos de carreras de pedagogía en matemáticas en función de sus resultados en la PSU y la NEM. Se plantea un modelo de regresión:

$$R = b_0 + b_1 \text{Matematica} + b_2 \text{NEM} + b_3 \text{Lenguaje} + b_4 \text{Ciencia} + e. \tag{3.11}$$

Supongan que los errores del modelo (3.11) cumplen los supuestos usuales de normalidad y correlación nula, $e_i \sim \mathcal{N}(0, \sigma^2)\ \forall i = 1, \ldots, n$ y $Cov(e_i, e_j) = 0\ \forall i \neq j$.

(a) Los resultados de la regresión se encuentran en las tablas 3.13 y 3.14. Complete los resultados de las tablas.
(b) Deduzcan de la tabla 3.14 el número de observaciones de la muestra. Justifiquen su respuesta.
(c) Expliquen cómo se obtuvieron los p-valores de la tabla 3.13.
(d) Comente los resultados del modelo considerando que el coeficiente de correlación múltiple es igual a 0,40.

Tabla 3.13. Coeficientes del modelo 3.11

Variable	Estimación	Desviación estándar	t-student	p-valor
Constante	-53,98	12,2152	?	0,000
Matemática	0,0312	0,0106	2,958	0,003
NEM	0,0245	?	3,129	0,002
Lenguaje	0,0135	0,0077	1,744	0,082
Ciencia	?	0,0096	7,843	0,000

Ejercicio 3.2. Un Instituto de Estudios Ambientales quiere analizar la influencia de algunos factores sobre la contaminación en SO_2. Las variables consideradas son: SO_2, número de empresas, la población (en miles de habitantes) y la temperatura del ambiente (en grados Fahrenheit). Se plantea un modelo de la forma:

$$SO_2 = b_0 + b_1 \text{número empresas} + b_2 \text{temperatura} + b_3 \text{ población} + e. \tag{3.12}$$

Supongan que los errores del modelo (3.12) cumplen los supuestos usuales de normalidad y correlación nula, $e_i \sim \mathcal{N}(0, \sigma^2)\ \forall i = 1, \ldots, n$ y $Cov(e_i, e_j) = 0\ \forall i \neq j$.

Tabla 3.14. Tabla ANOVA del modelo 3.11

Fuente	Suma cuadrados	Grados libertad	Cuadrados medios	F	p-valor
Regresión	12540	4	3135	?	0,0000
Residuos	?	499	126,51		
Total	75671	?			

(a) Se registraron los valores de las variables sobre una muestra. Se precedió a plantear el modelo (3.12). Los resultados de la regresión se encuentran en las tablas 3.15, 3.16 y 3.17. Complete los resultados de las tablas 3.16 y 3.17.

(b) Expliquen cómo se obtuvieron los p-valores de la tabla 3.16. Deduzcan de la tabla 3.17 el número de observaciones de la muestra. Justifiquen su respuesta.

(c) Basándose en los resultados anteriores, uno de los investigadores propone un modelo alternativo:

$$SO_2 = b_0 + b_1 \text{número empresas} + b_2 \text{temperatura} + e, \tag{3.13}$$

cuyos resultados se encuentran en las tablas 3.18 y 3.19. Interpreten el modelo y compárenlo con el anterior. Decidan si el modelo (3.13) da resultados muy diferentes al modelo (3.12), indicando los supuestos y criterios en los cuales basan su decisión. Justifiquen su respuesta.

Tabla 3.15. Matriz de Correlaciones

Variable	SO_2	Número empresas	Temperatura	Población
SO_2	1,000	0,869	-0,463	0,794
Número empresas		1,000	-0,325	0,971
Temperatura			1,000	-0,259
Población				1,000

TABLA 3.16. Coeficientes del modelo 3.12

Variable	Estimación	Desviación estándar	t-student	p-valor
Constante	43,9534	19,505	?	0,0386
Número empresas	?	0,0156	3,229	0,0052
Temperatura	-0,4485	0,3355	-1,337	0,2000
Población	-0,0240	?	-1,572	0,1339
R=0,905				

TABLA 3.17. Tabla ANOVA del modelo 3.12

Fuente	Suma cuadrados	Grados libertad	Cuadrados medios	F	p-valor
Regresión	8.619,6	3	?	24,18	0,0000
Residuos	1.901,2	?	118,83		
Total	?	19			

TABLA 3.18. Coeficientes del modelo 3.13

Variable	Estimación	Desviación estándar	t-student	p-valor
Constante	47,373	20,8274	2,2745	0,0362
Número empresas	0,0266	0,0040	6,6509	0,000
Temperatura	-0,5826	0,3486	-1,6712	0,1130
R=0,89				

TABLA 3.19. Tabla ANOVA del modelo 3.13

Fuente	Suma cuadrados	Grados libertad	Cuadrados medios	F	p-valor
Regresión	8.325,8	2	4.162,9	32,24	0,0000
Residuos	2.195,0	17	129,12		
Total	10521	19			

Ejercicio 3.3. Consideramos las siguientes variables obtenidas en diferentes parcelas de $4ha$:

- x_1: la primera medición de volumen de madera;
- y: la segunda medición de volumen de madera;
- x_2: el número de pinos;

- x_3: la edad promedio de los pinos;
- x_4: el volumen promedio por pino ($x_4 = \frac{x_1}{x_2}$).

(a) Complete los resultados de la regresión de y sobre las 4 otras variables (Tabla 3.20).
(b) Complete la tabla de análisis de las varianzas (Tabla 3.21).
(c) ¿Cuántas parcelas tenemos?
(d) Dé el coeficiente de correlación múltiple R.
(e) Dé el resultado del test de la hipótesis: $H_0 : b_1 = b_2 = b_3 = b_4 = 0$. Concluye.
(f) Dé un intervalo de confianza a 95 % para b_2. Concluye
(g) Dé el resultado del test de hipótesis $H_0 : b_3 = 0$ contra $b_3 \neq 0$ al nivel de significación $\alpha = 0,05$.

Tabla 3.20. Coeficientes del ejercicio 3.3

Variable	Estimación	Desviación estándar	t-student	p-valor
Constante	23,45	14,90	?	0,122
X_1	0,9321	0,08602	?	0,000
X_2	?	0,4721	1,5554	?
X_3	-0,4982	0,1520	?	0,002
X_4	3,486	?	1,533	0,132

Tabla 3.21. Tabla ANOVA del ejercicio 3.3

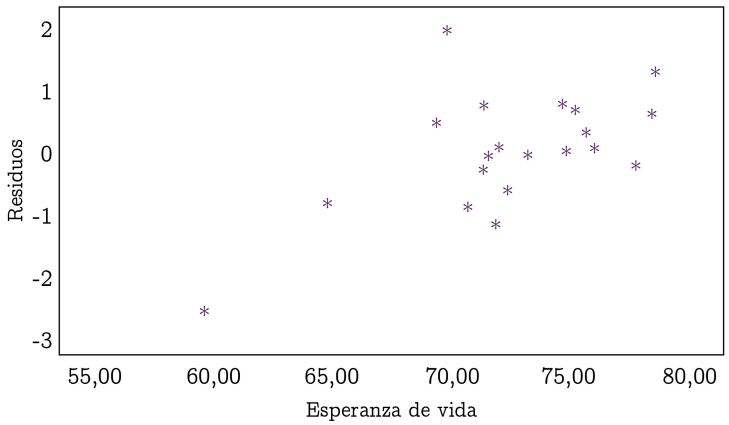
Fuente	Suma cuadrados	Grados libertad	Cuadrados medios	F	p-valor
Regresión	887994	4	222000	?	0,0000
Residuos	?	?	29558		
Total	902773	54			

Ejercicio 3.4. Consideramos tres indicadores demográficos para 20 países de América Latina: y la esperanza de vida; x la tasa de natalidad y z el porcentaje de población urbana y estudiamos la relación que existe entre ellas a partir del modelo: $y_i = b_0 + b_1 x_i + b_2 z_i + e_i$. Se obtuvieron los siguientes resultados:

Variable	Coeficiente	Desv. estándar	t-Student	P-valor
Constante	84,704	7,870	10,763	0,000
Tasa natalidad	-4,594	1,346	-3,414	0,003
% Pob. urbana	0,012	0,065	0,188	0,853
Coeficiente de correlación múltiple: $R = 0,796$			F de Fisher: 14,739	

(a) Interpretando el coeficiente de correlación múltiple R y el F de Fisher, concluya si las variables x y z son significativas en el modelo. Vea en particular si ambas variables x y z son realmente significativas.
(b) Dé intervalos de confianza al 95 % para los coeficientes b_1 y b_2.
(c) Si la varianza de la esperanza de vida es igual a 30,5, estime $Var(\hat{e}_i) = \sigma^2$.
(d) Interpretando la figura (3.5), concluya si el supuesto usual sobre los erores teóricos $\mathbb{E}[e_i] = 0$ se cumple. Comente qué otras propiedades de los estimadores puede deducir de esta figura.

FIGURA 3.5. Gráfico de y vs residuos



Ejercicio 3.5. Estudiamos si la nota de un control de Estadística está ligada al tiempo de trabajo personal, el porcentaje de asistencia a clase y el tiempo que toma el alumno en el trayecto entre su casa y la escuela. Se tienen los datos de una muestra de 100 alumnos. Se realizan tres regresiones lineales: La regresión lineal de la nota de un control de Estadística sobre las tres otras variables (Tabla 3.22); ($\mathcal{R}_2$) La misma

regresión lineal sin la variable Trayecto (Tabla 3.23) y ($\mathcal{R}_3$) La regresión sobre la variable Trabajo personal (Tabla 3.24).

(a) Concluya si el modelo de la Tabla 3.22 es globalmente significativo. Justifique su conclusión. Precise los grados de libertad, si corresponde.
(b) ¿Las tres variables tienen la misma importancia en el modelo? Justifique.
(c) Para el modelo de la Tabla 3.24 construya un intervalo de confianza para el coeficiente de la variable "Trabajo Personal" a un nivel de confianza igual al 95 %.
(d) ¿Con cuál de los tres modelos se quedaría?

TABLA 3.22. Resultados de la regresión $\mathcal{R}_1$

Variable	Coeficiente	Desv. estándar	t-Student	P-valor
Constante	1,4900	0,1592	9,358	0,0000
Trabajo personal	0,0153	0,0006	26,090	0,0000
% de asistencia	0,0119	0,0013	9,338	0,0000
Trayecto	0,0001	0,0017	0,079	0,9372
Coeficiente de correlación múltiple: $R = 0,941$			F de Fisher: 246,501	

TABLA 3.23. Resultados de la regresión $\mathcal{R}_2$

Variable	Coeficiente	Desv. estándar	t-Student	P-valor
Constante	1,4973	0,1296	11,552	0,0000
Trabajo personal	0,0153	0,0006	26,572	0,0000
% de asistencia	0,0119	0,0013	9,388	0,0000
Coeficiente de correlación múltiple: $R = 0,9325$			F de Fisher: 373,576	

TABLA 3.24. Resultados de la regresión $\mathcal{R}_3$

Variable	Coeficiente	Desv. típica	t-Student	P-valor
Constante	2,2063	0,1448	15,667	0,0000
Trabajo personal	0,0147	0,0008	18,677	0,0000
Coeficiente de correlación múltiple: $R = 0,8836$			F de Fisher: 348,83	

Ejercicio 3.6. En una muestra de $n = 20$ pacientes se han recogido los siguientes datos: nivel de colesterol en plasma sanguíneo (en mg/100 ml), edad (en años), consumo de grasas saturadas (en gr/semana) y nivel de ejercicio (cuantificado como 0: ningún ejercicio, 1: ejercicio moderado y 2: ejercicio intenso) y se realizo el ajuste a un modelo lineal entre el nivel de colesterol y las demás variables. A continuación se presentan los resultados:

Variable	Coeficiente	Desv. estándar	t-Student	P-valor
Constante	99,937	61,275	1,631	0,122
Edad	2,346	1,056	2,223	0,041
Grasas	2,306	0,720	3,201	0,006
Ejercicio	-6,248	19,831	-0,315	0,757

Coeficiente de correlación múltiple: $R = 0,831$

(a) ¿Cómo se calcula y qué representa el coeficiente de correlación múltiple?
(b) Para medir la calidad global del modelo se usa el estadístico F. Muestre que F se puede calcular a partir del coeficiente de correlación múltiple y que vale 11,90. Dé sus grados de libertad. Interprételo y concluya.
(c) Se decide eliminar la variable "Ejercicios" del modelo. ¿Le parece justificado? El coeficiente de correlación múltiple es igual en este caso a 0,832. ¿Qué concluye?
(d) ¿Cuáles son los otros resultados que se debería obtener para verificar los supuestos del modelo?

Ejercicio 3.7. Se desea explicar el consumo de bencina de diferentes vehículos en función de ciertas características a partir de un modelo lineal. Las variables explicativas consideradas son: la potencia, el peso, el tiempo de aceleración y el número de cilindros del vehículo. A partir de 391 observaciones, se efectúo la regresión del consumo de bencina (de media 23,48 y desviación típica de 7,78) sobre las cuatro otras variables; se obtuvieron los resultados siguientes:

Variable	Media	Desv. estándar	Coeficiente	Desv. estándar del coef.	t-Student	P-valor
Constante			46,244	2,510	18,424	0,000
Potencia	104,24	38,28	-0,045	0,016	-2,755	0,006
Peso	2.973, 10	845,83	-0,005	0,001	-7,052	0,000
Aceleración	15,53	2,76	-0,028	0,128	-0,215	0,830
Cilindros	5,47	1,70	-0,401	0,304	-1,320	0,188

Coeficiente de correlación múltiple: $R = 0,916$	F de Fisher: 230,735	p-valor: 0,000

(a) ¿Cómo se calculan los t-Student? Interprete los t-Student. Dé los grados de libertad.
(b) ¿Qué es la F de Fisher? Dé los grados de libertad de la F de Fisher. Concluya la bondad del modelo.

(c) Construya un intervalo de confianza a un nivel del 95 % para el parámetro $b_{aceleracin}$ asociado a la variable "Aceleración". Interprete para discutir la significación de la aceleración del vehículo para explicar el consumo de bencina en el modelo.

Capítulo 4: Árboles de clasificación y de regresión

Todos los días tomamos decisiones a partir de ciertas informaciones[1]. Por ejemplo, antes de salir de la casa en la mañana miramos por la ventana o escuchamos el informe meteorológico para decidir si vamos a llevar un paraguas. Después de dar la PSU, el estudiante debe elegir tres carreras en función de los puntajes obtenidos y de su preferencia vocacional.

Para implementar un plan de acción con el objeto de combatir el tabaquismo de los alumnos de Enseñanza Media (EM), el Ministerio de Educación recoge información sobre el tema mediante una encuesta en los colegios. Esta se realizó con una muestra aleatoria de 1500 alumnos[2]. Los resultados mostraron que 30 % de 1500 alumnos fuman. Además consideró la edad y el género de los alumnos. Estas informaciones permitirán al ministerio enfocar mejor su campaña preventiva. Tenemos entonces tres variables:

- El tabaquismo, que es una variable binaria (el alumno fuma o no fuma);
- El género, que es una variable binaria (hombre o mujer);
- La edad, que es una variable continua, pero considerando que en la Enseñanza Media el recorrido de edad es del orden de 5 años, se transformó en una variable binaria: el alumno tiene menos de 15 años o más.

El ministerio busca definir el perfil de alumnos fumadores a partir de las variables género y edad. Se calculan entonces los porcentajes de alumnos fumadores según su edad y género (Tabla 4.1). Otra manera de presentar la misma información es mediante un "árbol" (Figura 4.1). En la caja superior del árbol se muestra el número total de alumnos y los porcentajes de alumnos que fuman y no fuman. Vemos que en el total de los 1500 alumnos encuestados hay 30 % que fuman. Enseguida, el grupo de los 1500 alumnos se divide o segmenta en dos subgrupos según el género. Observamos que 33 % de las niñas fuman y 27 % de los niños. Dividamos ahora cada uno de esos dos últimos subgrupos según la edad: menor d 15 años o no. De esta manera clasificamos a los 1500 alumnos en cuatro grupos según la edad y el género. Observamos diferencias entre los cuatro grupos. En particular, el mayor porcentaje de fumadores corresponde a las niñas de al menos 15 años y el menor porcentaje a los niños de a lo más 15 años.

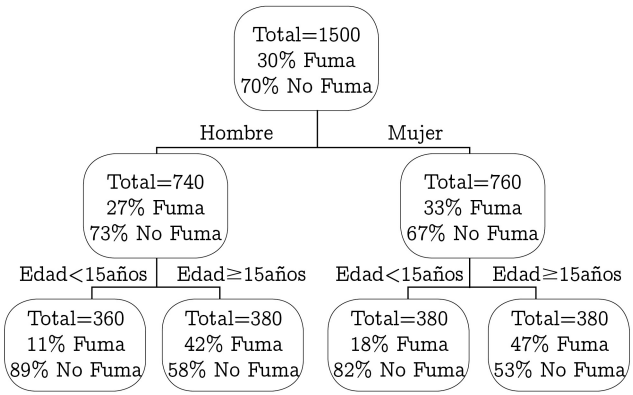
[1]Para una mejor comrensión de este capítulo es necesario conocer el método ANOVA del Capítulo 2 y la regresión múltiple del Capítulo 3.

[2]Los datos son ficticios, pero tratan de reflejar la realidad.

Tabla 4.1. Resultados de la encuesta

Género	< 15 años	≥ 15 años	Total
Mujer	380 18 %	380 47 %	740 27 %
Hombre	360 11 %	380 42 %	760 33 %
Total	740 15 %	760 45 %	1500 30 %

Figura 4.1. Árbol de los resultados de la encuesta



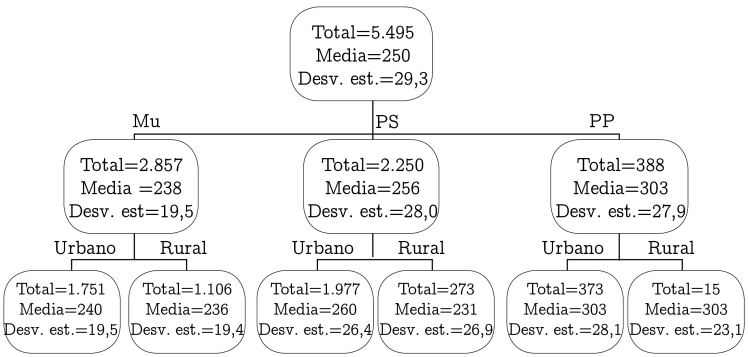
Consideramos ahora las medias por colegio del Sistema Nacional de Evaluación de Resultados de Aprendizaje del Ministerio de Educación de Chile (SIMCE) en Matemática del 8º Básico del año 2007. Buscamos si las medias se relacionan con la dependencia del colegio (municipal, particular subvencionado o particular pagado) y el sector (urbano o rural) (Tabla 4.2). Nuevamente construimos un árbol para representar la información (Figura 4.2). En cada caja aparece la frecuencia de colegios, la media y la desviación estándar del puntaje SIMCE correspondiente al subgrupo. Vemos que los promedios SIMCE difieren más entre los grupos definidos por la dependencia que los grupos definidos por el sector.

En el caso del tabaquismo, observamos más diferencias entre los dos subgrupos de edad. En el árbol de la Figura 4.3(a) la primera segmentación se realizó según la edad y después el género. Este árbol es más interesante que el de la Figura 4.1; las segmentaciones estan ordenadas desde arriba hacia abajo, de mayor a menor diferencia. En el ejemplo del SIMCE, se prefiere el orden del árbol de la Figura 4.2.

Tabla 4.2. Resultados SIMCE

	Municipal	Subvencionado	Part. Pagado	Total
Urbano	1751 243	1976 264	373 312	4100 259
Rural	1106 237	273 241	15 308	1394 239
Total	2857 242	2249 263	388 312	5494 256

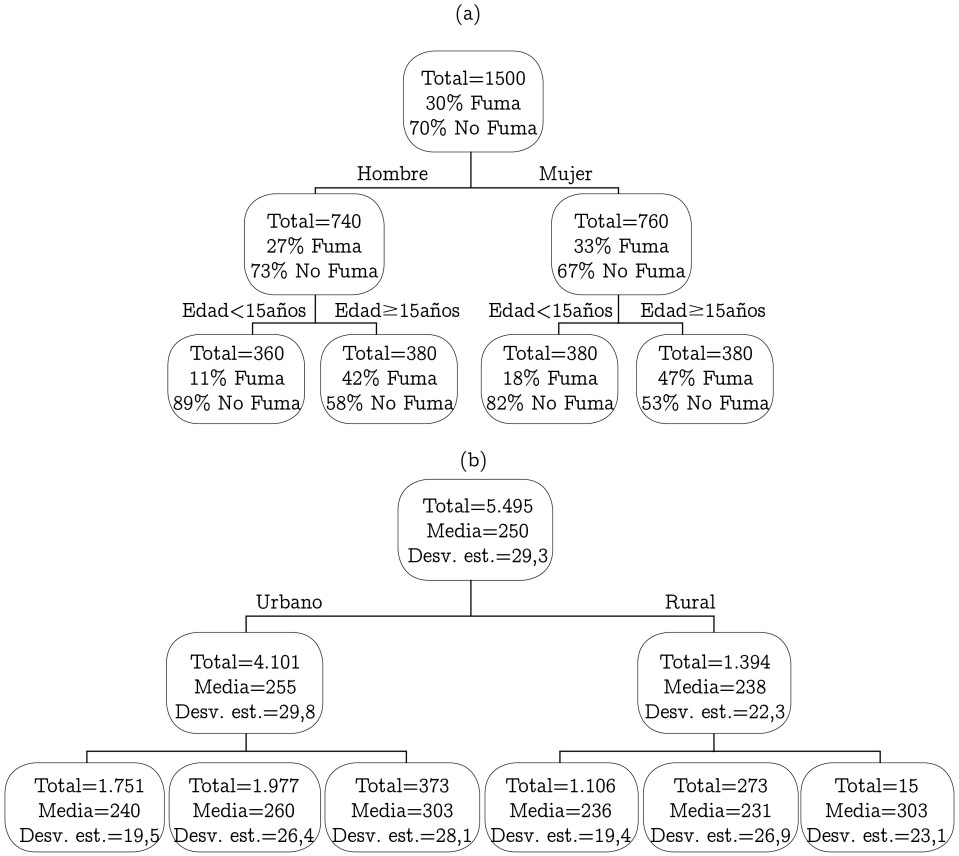
Figura 4.2. Árbol de los resultados SIMCE



En los dos ejemplos anteriores se usaron sólo dos variables para generar los subgrupos. Naturalmente, podemos tomar en cuenta más variables de manera de caracterizar, por ejemplo, los colegios si estas permiten tener una mejor homogeneidad del SIMCE en cada subgrupo. En este caso, las tablas son más complicadas construir e interpretar. Cuando se consideran más de dos variables es más fácil usar un árbol. Estos árboles se llaman "árboles de decisión". En efecto, un árbol del tipo 4.1 o 4.3(a) permite definir un perfil de alumnos que fuman, lo que podrá ser utilizado en una campaña preventiva contra el tabaquismo.

Las agrupaciones más finas, que son las obtenidas en la parte inferior del árbol, permiten eventualmente hacer predicciones. Por ejemplo, podemos decir que un alumno de EM de 14 años tiene una probabilidad de 11 % de ser fumador, mientras que una alumna de la misma edad tiene una probabilidad de 18 %. Con el ejemplo del SIMCE, para un colegio urbano municipal podemos predecir un puntaje SIMCE de 8º Básico de 240 con un intervalo de confianza de 90 % $[240-1,96\frac{19,5}{\sqrt{1751}};\ \ 240+1,96\frac{19,5}{\sqrt{1751}} = [239,09;\ \ 240,91]$.

FIGURA 4.3. Ejemplos de árboles con el orden cambiado



Los grupos de estos ejemplos fueron definidos a partir de variables cualitativas, incluso la edad del primer ejemplo fue transformada en una variable binaria (< 15 años o ≥ 15 años). La elección de estas dos últimas clases puede ser discutible.

Como vimos anteriormente, es mejor empezar a dividir por la edad y después por el género en el caso del tabaquismo (Figura 4.3(a)) y primero por la dependencia y después por el sector en el caso del SIMCE (Figura 4.2). Pero estos órdenes de aparición de las segmentaciones son subjetivos y no se basan en un criterio riguroso. Finalmente, observamos que entre el sector urbano y rural se nota poca diferencia, lo que podría llevar a decidir eliminar esta variable del árbol y quizás, introducir otras variables que caracterizan mejor los resultados SIMCE.

Hasta ahora se construyeron la jerarquía de los árboles con criterios intuitivos. El método de "árboles de clasificación y regresión" proporciona criterios para construirlos.

4.1 ¿Qué es un árbol de decisión?

Un árbol de decisión permite comunicar información mediante criterios fáciles de interpretar. Se utiliza como modelo de predicción, especialmente en el ámbito de la inteligencia artificial.

El método utiliza un enfoque visual de agrupaciones de datos mediante reglas fáciles de entender. Las variables utilizadas para definir los grupos son las "variables explicativas", a las que llamaremos también "variables de segmentación" y la variable que se mide dentro los subgrupos es la "variable respuesta". Los miembros de los grupos son definidos a partir de los valores de las variables explicativas, y en cada grupo, se estudian las características de la variable respuesta. En la parte baja del árbol se obtienen varios subgrupos. Si dentro cada uno de estos subgrupos los sujetos toman valores parecidos sobre la variable respuesta y, de un subgrupo a otro, los valores difieren bastante, podemos caracterizar perfiles de sujetos con las variables explicativas y relacionarlos con valores de la variable respuesta. En este caso, las variables explicativas, que identifican a los miembros de los grupos, permitirán hacer predicciones de la variable respuesta.

El método de los árboles de clasificación y de regresión (CART) permite construir de manera óptima tales árboles. CART tiene el mismo propósito que el ANOVA o la regresión lineal, pero difiere en varios aspectos:

- Las relaciones son no lineales,
- Visualiza las relaciones mediante un árbol,
- Puede usar cualquier tipo de variables, nominales o numéricas, tanto para la variable respuesta como para las variables de segmentación.

Con CART, los cálculos numéricos y los gráficos no son simples de realizar. Esta metodología requiere un software ad hoc llamado "R", el cual puede bajarse gratuitamente de Internet[3]. También sirve el Add-in "decision.xla" para Excel, que se encuentra en www.ormm.net/.

El uso de árboles de decisión tuvo su origen en las ciencias sociales con los trabajos de J. Sonquist y J. Morgan (1964), de la Universidad de Michigan, y el programa AID (Automatic Interaction Detection), que fue uno de los primeros métodos de ajuste de los datos basados en árboles de clasificación.

En estadística, Robert Kass (1980) introdujo un algoritmo recursivo de clasificación no binario, llamado CHAID (Chi-square automatic interaction detection). Más tarde, L. Breiman, J. Friedman, R. Olshen y C. Stone (1984) introdujeron un nuevo algoritmo para la construcción de árboles y los aplicaron a problemas de regresión y clasificación. El método es conocido como CART por la sigla en inglés de "Classification and regression trees" [4].

[3]R es un software gratuito especializado en métodos estadísticos. Permite también implementar la regresión múltiple o el análisis en componentes principales. Se encuentra en www.r-project.org/.

[4]Casi al mismo tiempo el proceso de inducción mediante árboles de decisión comenzó a ser usado en "Machine Learning" en ciencias de la computación y en "Pattern Recognition" en ingeniería eléctrica.

El método CART es parte de lo que se llama "Data Mining" o "Minería de Datos", que se puso de moda en muchos tipos de instituciones y empresas. Por ejemplo, los registros de un banco contienen muchas informaciones de sus clientes. Puede usarlas para determinar el perfil de los clientes morosos y decidir si conceder o no un crédito a un nuevo solicitante. El Servicio de Impuestos Internos puede tratar de caracterizar las empresas que hacen fraudes fiscales. El departamento de recursos humanos de una multitienda puede examinar los procesos de contrataciones pasadas y determinar reglas de decisión que harán más eficientes los procesos de contrataciones futuras.

A continuación, presentamos en primer lugar los elementos que componen los árboles de decisión mediante dos ejemplos simples y después definimos los criterios de construcción del método de CART.

4.1.1 Descripción de un árbol binario de regresión

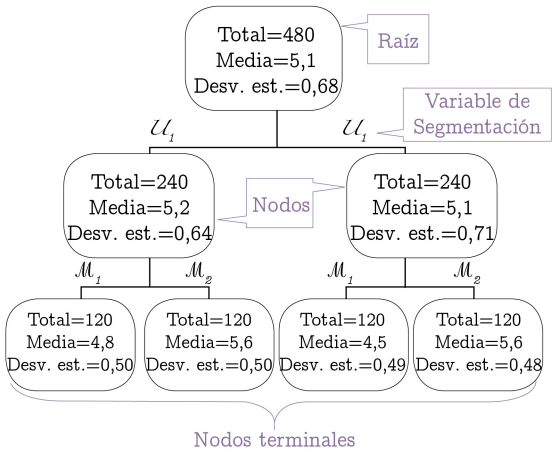
Un grupo de profesores de matemática de dos universidades, $\mathcal{U}_1$ y $\mathcal{U}_2$, quieren comparar dos métodos de enseñanzas, $\mathcal{M}_1$ y $\mathcal{M}_2$. Elaboran un experimento que consiste en elegir al azar dos grupos de estudiantes en cada universidad para aplicar el método $\mathcal{M}_1$ a un grupo y el método $\mathcal{M}_2$ al otro. Se obtienen así cuatro subgrupos definidos por el método de enseñanza y la universidad. Durante el semestre se aplican las mismas pruebas a los alumnos de los cuatro grupos. Al final del semestre se obtienen los rendimientos de los alumnos.

La variable respuesta, que es el rendimiento, es cuantitativa, y las dos variables de segmentación son binarias. Queremos ver si un método es mejor que el otro y si influye la universidad. Podemos llegar a los cuatro subgrupos en dos etapas. Se divide el total de los alumnos, en una primera, etapa según la universidad y después según el método de enseñanza (Figura 4.4).

El gráfico muestra un árbol jerárquico, que es un tipo de grafo. Si se lee de arriba hacia abajo, la "raíz" es el nodo superior, que contiene la totalidad de los 480 estudiantes. La raíz se divide en dos nodos, llamados "hijos", según una **regla de decisión**, que corresponde a valores que toma una variable de segmentación, que aquí es la universidad. El nodo que contiene a los dos hijos se llama naturalmente "padre" de estos. El nodo hijo de la izquierda contiene a los estudiantes de la universidad $\mathcal{U}_1$ y el hijo de la derecha contiene a los estudiantes de la universidad $\mathcal{U}_2$. Cada uno de estos dos nodos se divide a su vez en otros dos, uno con los estudiantes formados con el método de enseñanza $\mathcal{M}_1$ y el otro con los estudiantes formados con el método de enseñanza $\mathcal{M}_2$. Cada uno de los cuatro nodos obtenidos, que aquí son "nodos terminales", contiene estudiantes de la misma universidad y formados con el mismo método.

En general, los nodos se dividen en dos grupos según una pregunta o variable de segmentación, y los valores de la variable que usan en la segmentación para dividir un grupo en dos subgrupos corresponde a una **regla de decisión**. El método CHAID

FIGURA 4.4. Árbol de regresión



citado en la sección anterior permite realizar segmentaciones con más de dos subgrupos, pero a menudo el árbol no resulta fácil de interpretar. Se recomienda entonces usar los árboles binarios.

Para detectar si el método de enseñanza tiene efecto sobre el rendimiento y si hay diferencias entre las dos universidades se ponen en evidencia las características del rendimiento de cada nodo: la media y la desviación estándar del rendimiento y el tamaño del nodo, que es la frecuencia de estudiantes que pertenecen al nodo. Se observa poca diferencia entre las medias de los grupos de la primera segmentación (4,9 y 5,2). En los nodos terminales obtenidos de la segmentación con el método de enseñanza se observan diferencias mucho más importantes (4,8 y 5,6 para la Universidad $\mathcal{U}_1$ y 4,5 y 5,6 para la Universidad $\mathcal{U}_2$). Observamos también que las desviaciones estándar de los subgrupos son menores que la desviación estándar del total. Podemos decir, a primera vista, que hay un impacto del método de enseñanza sobre el rendimiento del estudiante. Para comprobarlo, podemos hacer un test de hipótesis de comparación de medias o también un ANOVA.

Se habla de árbol de regresión por la naturaleza de la variable respuesta, que es numérica. Más adelante vamos a "optimizar" el orden de las variables de segmentación de manera de jerarquizarlas en función de su impacto sobre la variable respuesta. Eliminaremos también las segmentaciones que no muestran una diferencia significativa sobre la variable respuesta entre los dos subgrupos producidos.

4.1.2 Ejemplo de un árbol binario de clasificación

La fábrica de chocolate Blagne quiere determinar el perfil de los consumidores de sus productos para dirigir mejor su campaña publicitaria. El departamento de estudios

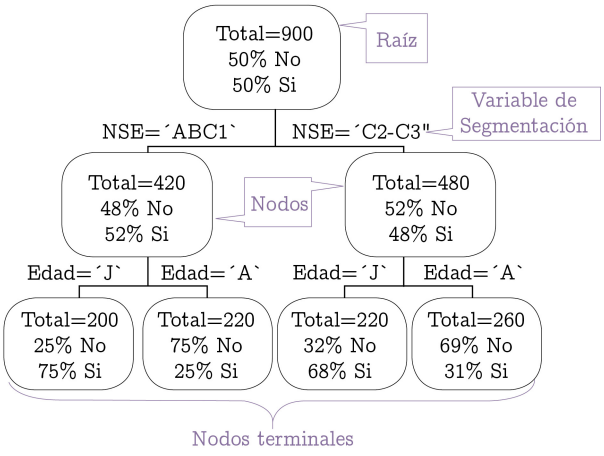
de mercados aplica entonces una encuesta a 900 personas de la cual se obtienen tres variables:

(a) A la pregunta ¿Consume el chocolate Blagne?, el encuestado responde “SI” o “NO”. Se obtiene la variable “consumo”.
(b) El encuestador define el “Nivel socio-economico” del encuestado. El encuestado está clasificado en “ABC1” o “C2-C3”. Se obtiene la variable “NSE”.
(c) El encuestador pregunta la edad y define la “clase de edad” del encuestado. El encuestado obtiene el código “J”, si es menor de 35 años, o “A”, si tiene al menos 35 años. Se obtiene así la variable “edad”.

Observamos que las tres variables son binarias, ya que toman solamente dos valores cada uno. Considerando el propósito del estudio, la variable “consumo ” es la variable respuesta que quisiéramos poder explicar a partir de la edad y del NSE. Nuevamente tenemos cuatro grupos posibles combinando las dos alternativas de la edad y las dos alternativas del NSE. Como primera variable de segmentación usamos el NSE. Los nodos se definen como en el ejemplo de la sección anterior. Sin embargo, las estadísticas del nodo por considerar son diferentes, pues la variable respuesta es binaria. Con los porcentajes de “SI” y de “NO” del “consumo” en los nodos, podemos determinar si existe un perfil de consumidores de Blagne (Figura 4.5). Parecería que el consumo se relaciona con la edad, pero poco con el NSE.

Se habla de árbol de clasificación por la variable respuesta, que permite clasificar a los encuestados en dos grupos: los consumidores del chocolate Blagne y los no consumidores.

FIGURA 4.5. Árbol de clasificación



Presentamos ahora la manera de "optimizar" el orden de las segmentaciones mediante criterios de segmentación para obtener homogeneidad de la variable respuesta dentro de los nodos y heterogeneidad entre los nodos.

En los dos ejemplos anteriores pudimos dividir fácilmente un grupo en dos subgrupos de manera natural, dado que las variables de segmentación que utilizamos (Universidad y método de enseñanza o edad y NSE) son binarias. Vemos también cómo dividir un nodo en dos subgrupos con una variable de segmentación numérica o nominal con más de dos categorías.

4.2 División binaria

Para obtener las divisiones binarias diferenciamos las variables de segmentación nominales de las numéricas.

4.2.1 Variable de segmentación nominal

Supongamos ahora que, en el ejemplo de la Sección 4.1.1, se tienen tres universidades ($\mathcal{U}_1$, $\mathcal{U}_2$ y $\mathcal{U}_3$) en vez de dos. Para dividir un grupo en dos subgrupos a partir de la variable universidad tenemos tres maneras de combinar las tres categorías:

Tabla 4.3

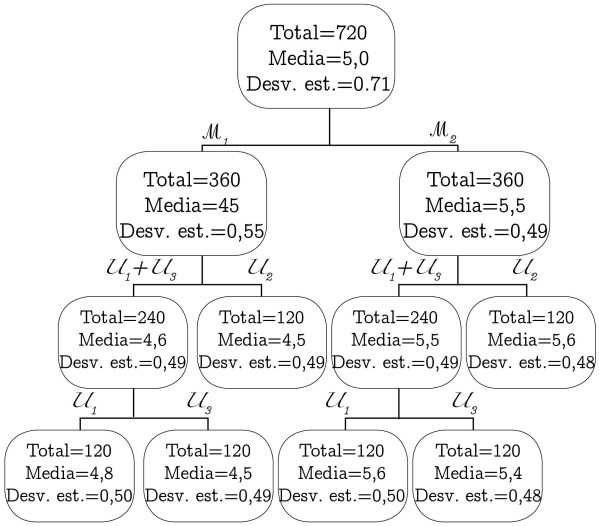
	Grupo1	Grupo 2
Caso 1	$\mathcal{U}_1$	$\mathcal{U}_2 + \mathcal{U}_3$
Caso 2	$\mathcal{U}_2$	$\mathcal{U}_1 + \mathcal{U}_3$
Caso 3	$\mathcal{U}_3$	$\mathcal{U}_1 + \mathcal{U}_2$

En el caso 2, por ejemplo, no se puede distinguir entre las universidades $\mathcal{U}_1$ y $\mathcal{U}_3$. Sin embargo, en una segmentación posterior se podrá separar el grupo "$\mathcal{U}_1$+$\mathcal{U}_3$" en dos un subgrupos, uno con $\mathcal{U}_1$ y el otro con $\mathcal{U}_3$ (Figura 4.6). En general, si la variable nominal tiene q categorías, se agrupan las categorías en dos subgrupos excluyentes, que pueden subdividirse, a su vez, en dos subgrupos, etc..

4.2.2 Variable de segmentación numérica

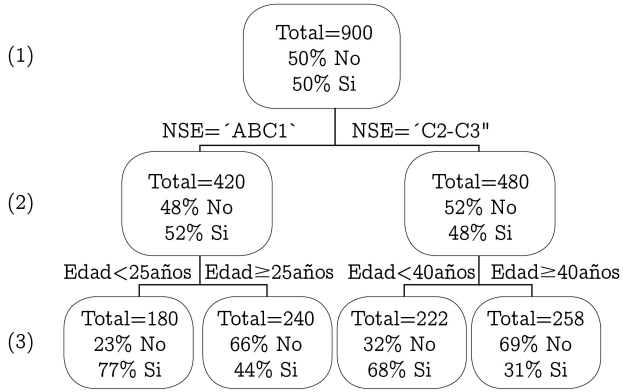
Supongamos que tomamos la edad de los consumidores en años en vez de las dos clases de edad "J: menor de 35 años" y "A: mayor o igual a 35 años" en el ejemplo de la Sección 4.1.2. Para dividir un grupo en dos subgrupos con la edad en años tenemos muchas posibilidades. De manera natural, nos limitaremos a considerar solamente las segmentación en dos grupos del tipo "menor que u" y "mayor que u" , donde u toma los valores de todas las distintas edades presentes en la muestra. Por ejemplo, si $u = 25$, un grupo será "menor que 25 años" y el otro "mayor o igual a 25 años". El valor de u no será necesariamente el mismo en las dos segmentaciones. Por ejemplo, el corte es

FIGURA 4.6. División con una variable nominal no binaria



en 25 años en vez de 35 para el NSE "ABC1" y 40 años para el NSE "C2-C3" (Paso del nivel (2) al nivel (3) en el árbol de la Figura 4.7).

FIGURA 4.7. División con una variable numérica



4.3 Construcción del árbol de regresión

En el ejemplo de la Sección 4.1.1 podríamos intercambiar el orden de las variables de segmentaciones, método de enseñanza y universidad. En la Figura 4.8(a) se divide primero por universidad y después por método de enseñanza. En la Figura 4.8(b) se divide primero por método de enseñanza y después por universidad. Si queremos jerarquizar el efecto de las variables de segmentación sobre el rendimiento, parecería que el segundo árbol es más adecuado. En efecto, se observa no solamente una mayor diferencia entre las medias de los dos subgrupos en la primera división, sino, también, una disminución importante de las varianzas al interior de los subgrupos en la figura de la derecha. ¿Cómo podemos definir un criterio que permita elegir de manera automática las segmentaciones que produzcan subgrupos diferentes entre sí y que dentro los subgrupos haya poco variabilidad?

4.3.1 Criterio de segmentación

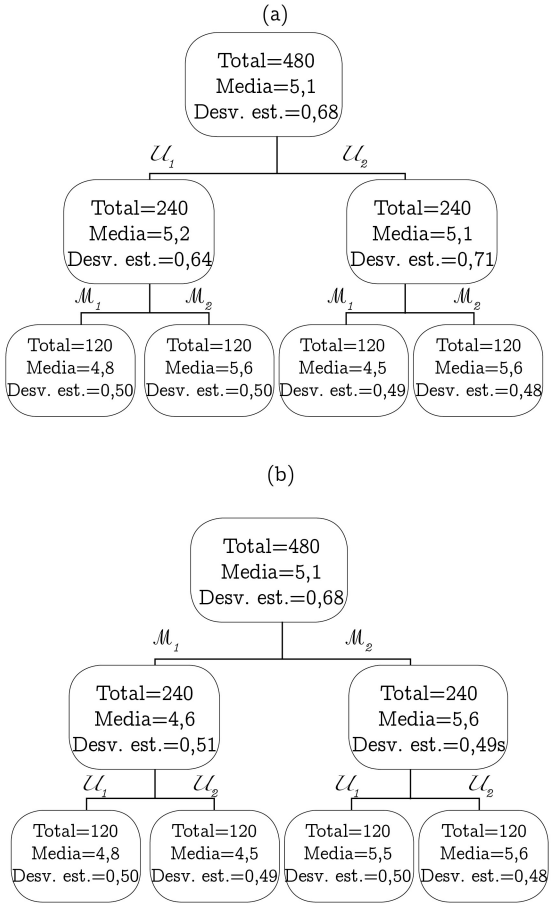
Recordamos que en el método ANOVA, presentado en la Sección 2.8 del capítulo 2, definimos w^2 como la varianza intragrupos, que es el promedio de las varianzas dentro de los grupos, y b^2 como la varianza intergrupos, que es la varianza de las medias de los grupos. Además, obtenemos la varianza total $v^2 = b^2 + w^2$. Vimos también, que el estadístico $F = \frac{b^2/(q-1)}{w^2/(n-q)}$, donde q es el número de grupos y n el número de sujetos (Ecuación 2.8 del capítulo 2), sigue una distribución de Fisher con $q-1$ y $n-q$ grados de libertad bajo la hipótesis nula de igualdad de las medias de los grupos. Aquí tenemos solamente dos grupos, por lo cual es equivalente utilizar el test F del ANOVA o el test t de Student de comparación de dos medias (Ver la Proposición 2.4 del Capítulo 2). Utilizaremos aquí el estadístico F e introducimos también otro criterio, que es la "razón de correlación",

$$\eta = \frac{Varianza\ \ intergrupo}{Varianza\ \ total} = \frac{b^2}{v^2}.$$

El coeficiente η varia entre 0 y 1. Cuando las medias de los grupos son iguales, $\eta = 0$; y cuando dentro los grupos las varianzas son nulas y las medias de los grupos son distintas, entonces $\eta = 1$. Una razón de correlación η alta indica que los grupos se diferencian. Seleccionaremos entonces la segmentación que tiene la mayor razón de correlación.

(a) Dados los valores del rendimiento de los 480 alumnos, calculamos para cada variable de segmentación el coeficiente η. En este ejemplo, tenemos solamente dos casos por considerar, dado que las dos variables de segmentación son binarias (Tabla 4.4). La varianza intergrupos para los métodos de enseñanza $\mathcal{M}_1$ y $\mathcal{M}_2$ es netamente mayor que la varianza intergrupos para las universidades $\mathcal{U}_1$ y $\mathcal{U}_2$ y más aún en relación con la varianza total, que es lo que muestra el coeficiente η de 0,46 contra 0,007 para las universidades. Según este criterio, elegimos dividir la raíz según el método de enseñanza. Además se calculó el valor de la F de Fisher del test ANOVA de comparación de

FIGURA 4.8. Intercambio del orden de las segmentaciones



medias. El p-valor de la F del método de enseñanza es nulo, lo que permite rechazar la igualdad de las medias del rendimiento de los grupos de métodos de enseñanza. Para la universidad, el p-valor es igual a 0,071, lo que es más elevado que para el método de enseñanza. Además no podemos rechazar la igualdad de las medias de los grupos definidos por la universidad con un error de tipo I de 5 %. Nos quedamos entonces con el método de enseñanza como primera segmentación.

(b) Seguimos entonces con el árbol 4.8(b), que tiene como primera segmentación el método de enseñanza. Ahora tenemos que aplicar el criterio η en cada uno de los dos subgrupos definidos por el método de enseñanza. Aquí no hay

muchas alternativas. Nos queda solamente la universidad como variable de segmentación. Sin embargo, nos preguntamos si tiene sentido segmentar con la universidad para ambos métodos. Vemos que los coeficientes η y los p-valores de la F de Fisher de los dos métodos son muy diferentes (Tabla 4.5). Para el método $\mathcal{M}_1$, $\eta = 0,0527$ y el p-valor de la F es nulo, y para el método $\mathcal{M}_2$, $\eta = 0,000$ y el p-valor es 0,884, que es muy alto. Las dos segmentaciones no tienen la misma importancia y no está claro si aplicar ambas. A continuación, estudiamos criterios para no seguir dividiendo en subgrupos cuando éstos no son muy diferentes.

Tabla 4.4

Variable	Tamaño	Varianza intergrupo	Varianza total	η	F	p-valor
Universidad	480	0,0031	0,464	0,007	3,26	0,071
Método de enseñanza	480	0,214	0,464	0,46	408,3	0,000

Tabla 4.5

Variable	Tamaño	Varianza intergrupo	Varianza total	η	F	p-valor
Método $\mathcal{M}_1$						
Universidad	240	0,0137	0,2589	0,0527	13,25	0,000
Método $\mathcal{M}_2$						
Universidad	240	0,0000	0,2418	0,000	0,021	0,884

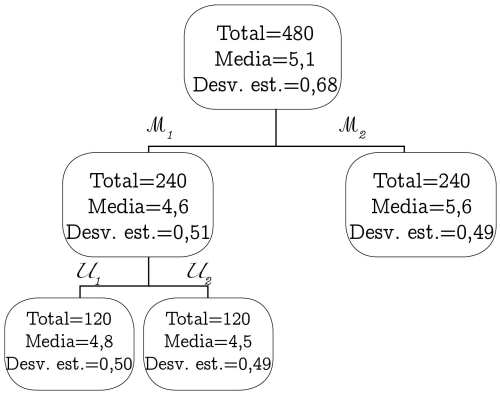
4.3.2 Criterio de poda

¿Cuándo detener la segmentación de un nodo en la construcción del árbol de decisión? Obviamente, cuando no existan más variables explicativas que permitan segmentar. Cuando hay muchas variables de segmentación y varias divisiones binarias para cada una, el árbol puede crecer bastante. Cuando el árbol es muy grande, su interpretación se pone difícil. Es inútil recargarlo con muchos nodos y ramas, si algunos de estos no aportan a la explicación de la variable respuesta. En este caso hay que buscar una manera de parar de segmentar. Usualmente se construye un árbol más largo que lo necesario y luego se van eliminando los nodos de poca utilidad. Se habla de "poda" del árbol.

Acabamos de ver en la tabla anterior que en el segundo nivel del árbol 4.8(b) no hay diferencia en el rendimiento de los alumnos de las dos universidades en el grupo del método $\mathcal{M}_2$, pues el p-valor de la F del ANOVA es igual a $0,884$ (Tabla 4.5). Para el grupo del método $\mathcal{M}_1$, el p-valor de la F es nulo. Se puede entonces eliminar, o sea, podar los dos nodos colgando del método $\mathcal{M}_2$.

El criterio natural para podar el árbol es el p-valor de la F, que se usa en cada nodo para decidir si seguir la segmentación debajo del nodo. Se puede decidir, por ejemplo, aceptar la segmentación con un p-valor menor que 5 % y podar el nodo en el caso contrario (Figura 4.9).

FIGURA 4.9. Árbol podado (2 universidades)



Construyamos el árbol del ejemplo de la Sección 4.2.1 con tres universidades. La Tabla 4.6 entrega el detalle del orden de las segmentaciones con el coeficiente η y el criterio de poda con un p-valor de la F de 5 %.

A partir de la tabla, las decisiones se toman de la siguiente manera:

(a) Raíz $\longrightarrow$ Nivel (1): En esta primera segmentación, tenemos cuatro segmentaciones posibles: el método, que tiene un coeficiente η igual a 0,47; la segmentación entre la Universidad $\mathcal{U}_1$ y $\mathcal{U}_2 + \mathcal{U}_3$ con $\eta = 0,017$; la segmentación entre la Universidad $\mathcal{U}_2$ y $\mathcal{U}_1 + \mathcal{U}_3$ con $\eta = 0,000$; la segmentación entre la Universidad $\mathcal{U}_3$ y $\mathcal{U}_1 + \mathcal{U}_2$ con $\eta = 0,021$. Se elige dividir con el método, que tiene el coeficiente η más grande.

(b) Nivel (1) $\longrightarrow$ (2): Se tienen dos grupos a dividir: el grupo del método $\mathcal{M}_1$ y el grupo $\mathcal{M}_2$. Para ambos grupos tenemos tres segmentaciones posibles definidas por las universidades. Para el método $\mathcal{M}_1$ la segmentación de las universidades $\mathcal{U}_1$ y $\mathcal{U}_2 + \mathcal{U}_3$ tiene el mayor coeficiente η (0,078) y su p-valor es nulo. Se elige utilizar esta segmentación. Para el método $\mathcal{M}_2$, la segmentación de las universidades $\mathcal{U}_3$ y $\mathcal{U}_1 + \mathcal{U}_2$ tiene el mayor coeficiente η (0,019) y su p-valor es menor que 5 %. Se elige utilizar esta segmentación.

(c) Nivel (2) $\longrightarrow$ (3): Queda por ver si los cuatros nodos que obtuvimos en el nivel (2) pueden seguir segmentándose. Los nodos $(\mathcal{M}_1, \mathcal{U}_1)$ y $(\mathcal{M}_2, \mathcal{U}_3)$ no pueden segmentarse más. Son nodos terminales (nodos sombreados). Sólo los otros dos nodos pueden segmentarse. En el nodo $(\mathcal{M}_1, \mathcal{U}_2 + \mathcal{U}_3)$ podemos tratar de separar las universidades $\mathcal{U}_2$ y $\mathcal{U}_3$, y en el nodo $(\mathcal{M}_2, \mathcal{U}_1 + \mathcal{U}_2)$ podemos

tratar de separar las universidades $\mathcal{U}_1$ y $\mathcal{U}_2$. Para el primer nodo tenemos un p-valor de 0,008, que es menor que 5 %. Conservamos esta segmentación. En el segundo nodo, el p-valor de 0,884 es muy alto, lo que nos lleva a podar esta segmentación. Se agregó una cruz en el lugar de la poda. Se borran entonces los nodos por debajo de esta cruz.

Finalmente, nos quedamos con los 5 nodos terminales sombreados en la Figura 4.10.

Tabla 4.6

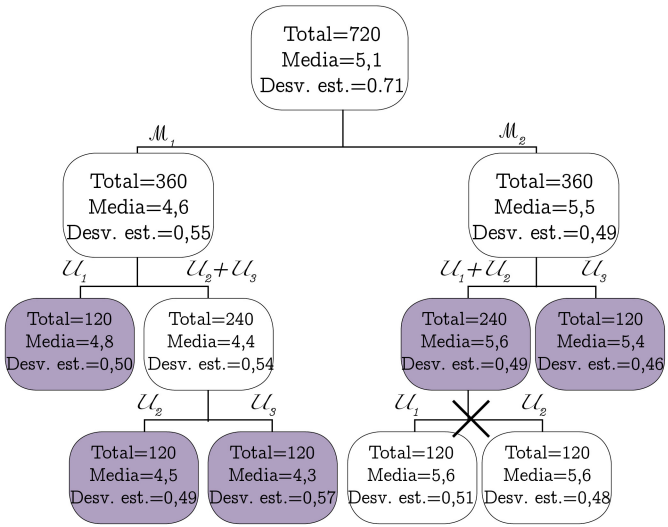
Segmentación	Tamaño	Varianza intergrupos	Varianza total	η	F	p-valor
Raíz ->Nivel (1) para el Nodo RAÍZ						
Método: $\mathcal{M}_1$ y $\mathcal{M}_2$	**720**	**0,239**	**0,509**	**0,47**	**637,6**	**0,000**
Universidad: $\mathcal{U}_1$ y $\mathcal{U}_2 + \mathcal{U}_3$	720	0,009	0,509	0,017	12,14	0,000
Universidad: $\mathcal{U}_2$ y $\mathcal{U}_1 + \mathcal{U}_3$	720	0,0000	0,509	0,000	0,022	0,636
Universidad: $\mathcal{U}_3$ y $\mathcal{U}_1 + \mathcal{U}_2$	720	0,011	0,509	0,021	15,77	0,000
Nivel (1) ->Nivel (2) para el Nodo MÉTODO $\mathcal{M}_1$						
Universidad: $\mathcal{U}_1$ y $\mathcal{U}_2 + \mathcal{U}_3$	**360**	**0,024**	**0,302**	**0,078**	**30,41**	**0,000**
Universidad: $\mathcal{U}_2$ y $\mathcal{U}_1 + \mathcal{U}_3$	360	0,000	0,302	0,000	0,1558	0,691
Universidad: $\mathcal{U}_3$ y $\mathcal{U}_1 + \mathcal{U}_2$	360	0,020	0,302	0,067	25,70	0,0030
Nivel (1) ->Nivel (2) para el Nodo MÉTODO $\mathcal{M}_2$						
Universidad: $\mathcal{U}_1$ y $\mathcal{U}_2 + \mathcal{U}_3$	360	0,000	0,237	0,004	1,380	0,241
Universidad: $\mathcal{U}_2$ y $\mathcal{U}_1 + \mathcal{U}_3$	360	0,0014	0,237	0,006	2,053	0,153
Universidad: $\mathcal{U}_3$ y $\mathcal{U}_1 + \mathcal{U}_2$	**360**	**0,0045**	**0,237**	**0,019**	**6,897**	**0,009**
Nivel (2) ->Nivel(3) para el Nodo MÉTODO $\mathcal{M}_1$ y UNIVERSIDADES $\mathcal{U}_2 + \mathcal{U}_3$						
Universidad: U_2 y U_3	**240**	**0,009**	**0,291**	**0,029**	**7,20**	**0,008**
Nivel (2) ->Nivel(3) para el Nodo MÉTODO $\mathcal{M}_2$ y UNIVERSIDADES $\mathcal{U}_1 + \mathcal{U}_2$						
Universidad: U_1 y U_2	240	0,000	0,242	0,000	0,0213	0,884

4.3.3 Predicción

Un vez podado el árbol, podemos hacer una partición de los 720 estudiantes en cinco grupos perfectamente identificados por su universidad y método de enseñanza recibido. Los cinco grupos tienen diferentes rendimientos entre sí, y, al interior de cada grupo, los rendimientos difieren poco.

Podemos usar estos resultados para hacer predicciones del rendimiento de nuevos alumnos en condiciones similares utilizando las medias y desviaciones estándar del grupo al cual pertenecen. Por ejemplo, se espera que un alumno de la universidad $\mathcal{U}_3$

FIGURA 4.10. Árbol podado (tres universidades)



con el método $\mathcal{M}_1$ tendrá más o menos un rendimiento de 4,3. Como toda predicción está sujeta a errores, podemos construir un intervalo de confianza (Lacourly[**7**]) para el rendimiento esperado del alumno. Para un nivel de confianza de 95 % obtenemos el intervalo:

$$IC_{95\,\%} = [4,3 - 1,96 \times \frac{0,71}{\sqrt{120}},\ 4,3 + 1,96 \times \frac{0,71}{\sqrt{120}}] = [4,15;\ 4,45],$$

donde 120 es el número de alumnos del grupo y 0,71 la desviación estándar del rendimiento.

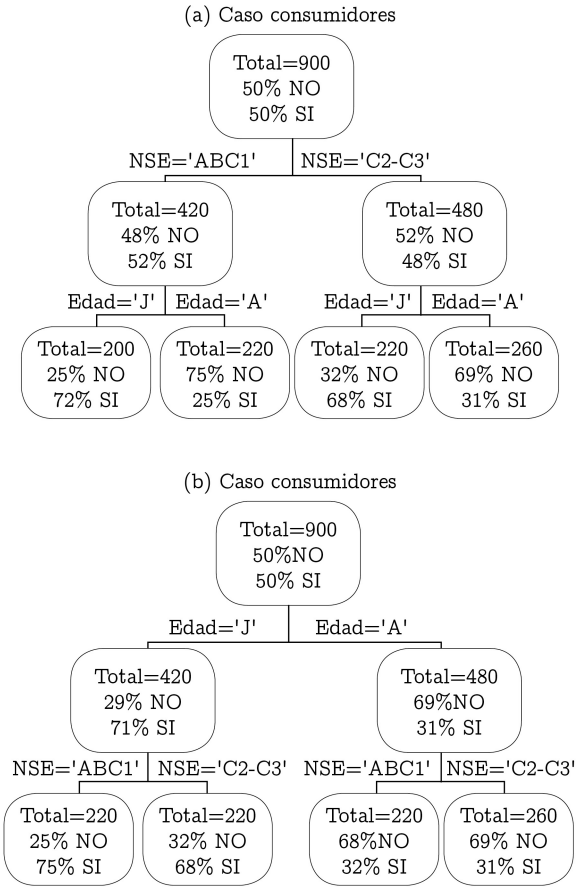
Ejercicio: Entregue la predicción del rendimiento con un intervalo de confianza de 95 % para un alumno de la Universidad $\mathcal{U}_2$ que tuvo el método de enseñanza $\mathcal{M}_2$.

4.4 Construcción del árbol de clasificación

En el ejemplo de la Sección 4.1.2, la variable respuesta "consumo" es binaria. No podemos usar el criterio η para elegir las segmentaciones del árbol como lo hicimos anteriormente. Pero examinamos las estadísticas de los nodos cuando se divide la raíz con el NSE (Figura 4.11(a)) o con la edad (Figura 4.11(b)). Observemos que cuando se divide la raíz con el NSE, hay poca diferencia entre los dos grupos socio-económicos. Ambos tienen valores cercanos al 50 % de "SI" y de "NO" como en la raíz. Si dividimos la raíz con los dos grupos de edad, los resultados son distintos. En el grupo "J" hay mucho más "SI" que "NO", y en el grupo "A" ocurre lo contrario. Esta segmentación

en que se utiliza la edad es más interesante para el estudio de mercado de la fábrica de chocolate.

FIGURA 4.11. Árboles del ejemplo de los consumidores

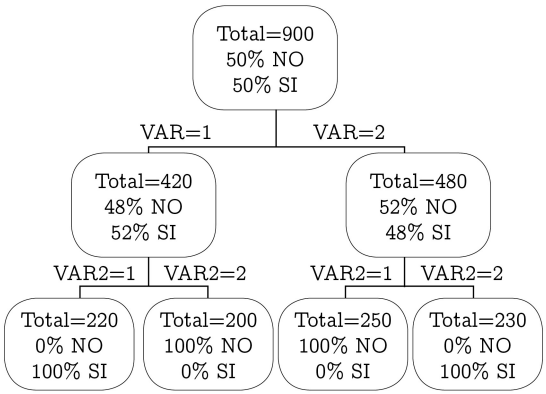


4.4.1 Criterio de segmentación

Para introducir el criterio de segmentación que utilizaremos, consideramos un caso ideal, cuyas variables de segmentación, llamadas VAR1 y VAR2, son binarias (Figura 4.12). En los nodos terminales aparece una situación extrema. Los nodos tienen 100 % de "NO" o 100 % de "SI". En este caso, las variables VAR1 y VAR2 determinan perfectamente el perfil de los consumidores del chocolate Blagne. Para estos nodos,

se habla de **"pureza"**. Cuando hay algunos "NO'y "SI" en un nodo, se dice que es impuro. Buscaremos entonces segmentaciones para obtener los grupos con la **máxima pureza**, o sea, que los porcentajes de "SI" y "NO" en un nodo sean lo más diferentes posible.

FIGURA 4.12. Situación ideal



Se pueden construir varios índices de impureza. El más utilizado, en economía, por ejemplo, es el índice de Gini, que es una medida definida por el estadístico italiano Corrado Gini para medir la desigualdad de los ingresos en una población. Es un número entre 0 y 1, en donde 0 se corresponde con la perfecta igualdad (todos tienen los mismos ingresos) y 1 se corresponde con la perfecta desigualdad (una persona tiene todos los ingresos y los demás ninguno). Este índice puede utilizarse para medir cualquier forma de distribución desigual. El índice de impureza de Gini se basa en la misma idea.

Retomamos el árbol del ejemplo 4.1.2 (Figura 4.11(a)). Sean $p_N(t_1)$ y $p_S(t_1)$ las proporciones de "SI" y "NO", respectivamente, en el nodo t_1. Observamos que $p_S(t_1) = 1-p_N(t_1)$, $p_S(t_1)\times p_N(t_1) = p_S(t_1)(1-p_S(t_1)) = p_N(t_1)(1-p_N(t_1))$. Además $p_S(t_1) \times p_N(t_1)$ toma el valor 0 cuando $p_S(t_1)$ vale 0 o 1, y toma el valor máximo 0,25 cuando $p_S(t_1) = p_N(t_1) = 0,5$. Se define entonces el **índice de impureza de Gini** del nodo t_1 como:

$$\gamma(t_1) = p_N(t_1)(1 - p_N(t_1)) + p_S(t_1)(1 - p_S(t_1)) = 2\, p_N(t_1)p_S(t_1).$$

Mientras más pequeño es el índice de Gini, más puro es el nodo. Por ejemplo, el nodo NSE="ABC1" tiene un índice de Gini igual a: $2 \times 0,48 \times 0,52 = 0,50$ y su nodo hijo NSE="ABC1" y EDAD="J" tiene un índice de Gini igual a: $2 \times 0,25 \times 0,75 = 0,375$. El índice de Gini del otro nodo hijo NSE="ABC1" y EDAD="A" vale $2\times 0,68\times 0,32 = 0,435$. El nodo padre es más impuro que sus nodos hijos.

En un nodo t dado y dos nodos hijos t_1 y t_2 de t, se define el índice

$$G(t) = \frac{n_1}{n}\gamma(t_1) + \frac{n_2}{n}\gamma(t_2)$$

donde $\gamma(t_1)$ y $\gamma(t_2)$ son los índices de Gini de t_1 y t_2, respectivamente, y n, n_1 y n_2 son los tamaños de los nodos t, t_1 y t_2, respectivamente. $G(t)$ es un promedio ponderado de los índices de Gini de los dos nodos t_1 y t_2. En el nodo t se elige, entonces, entre las posibles segmentaciones (t_1, t_2), aquella que produce la mayor reducción de impureza promedio $G(t)$.

En la tabla 4.7 se presentan los índices de Gini γ de los nodos y los índices G obtenidos, y se indican los tamaños de los nodos entre paréntesis.

Tabla 4.7. Índices de Gini

Raíz	NSE="ABC1"	NSE="C2-C3"	NSE
0,50 (900)	0,499 (420)	0,499 (480)	0,499
Raíz	Edad="J"	Edad="A"	Edad
0,50 (900)	0,408 (420)	0,430 (480)	0,420
Edad="J"	NSE="ABC1"	NSE="C2-C3"	NSE con Edad="J"
0,408 (420)	0,375 (200)	0,434 (220)	0,406
Edad="A'	NSE="ABC1"	NSE="C2-C3"	NSE con Edad="A"
0,430 (480)	0,434 (220)	0,426 (260)	0,430

Desde la raíz, los índices G de la edad y el NSE son respectivamente 0,420 y 0,499. Desde la raíz, la edad permite la mayor reducción de impureza. En el nivel siguiente, se calcula el índice G en cada categoría de la edad. En el caso de la categoría "J", la impureza se reduce a 0,406; y en la categoría "A" la impureza G es de 0,430.

4.4.2 Clasificación y error de clasificación

En el caso de una variable respuesta nominal la predicción consiste en clasificar un sujeto en una de las categorías de la variable. En el caso del árbol 4.11(b) tenemos que clasificar un nuevo sujeto como consumidor o no consumidor del chocolate Blagne. La regla de clasificación natural y simple consiste en clasificar un sujeto como consumidor del chocolate si el porcentaje de respuestas "SI" de su grupo es mayor que el porcentaje de respuestas "NO". Por ejemplo, el nodo un nodo terminal ("J", "ABC1") tiene un 75 % de sujetos que respondieron "SI'. Clasificaremos entonces todos los sujetos de este grupo como consumidores del chocolate. El nodo terminal ("A" ,"C2-C3") tiene un 31 % de sujetos que respondieron "SI'. Clasificaremos entonces todos los sujetos de este grupo como no consumidor del chocolate.

Cuando los nodos no son puros, estas clasificaciones tienen errores. En el primer caso tenemos un 25 % de errores de clasificación y 31 % en el segundo caso.

Si aplicamos este criterio de clasificación (respuesta "SI " o "NO") a un nodo, podemos calcular la tasa de errores de clasificación para las distintas segmentaciones posibles, pues conocemos sus respuestas reales. Por ejemplo, desde la raíz, con el NSE clasificaremos 200 de los 420 encuestados con respuesta "SI", en circunstancias que respondieron "NO" (árbol 4.11(a)); y con la edad clasificaremos 120 de los 420 encuestados con respuesta "SI", en circunstancias que respondieron "NO" (árbol 4.11(b)). El NSE tiene una tasa de error de 48 % y la edad una tasa de error de 30 %. Podemos usar este criterio de minimizar los errores de clasificación para elegir la segmentación. Aquí se usa la edad como primera variable de segmentación desde la raíz, como en el caso del índice de Gini.

Lo ideal entonces es no tener errores de clasificación de las observaciones de los nodos. Calculemos las tasas de errores de clasificación del árbol 4.11(b). En la Tabla 4.8, para cada nodo se presenta el número de errores de clasificación y entre paréntesis el tamaño del nodo. Como ya vimos, la edad es la mejor elección de segmentación desde la raíz (30 % contra 48 % del NSE). Seguimos la tabla usando como primera segmentación la edad. En el nodo ("J", "ABC1") tenemos 25 % (50/200) de errores de clasificación y en el nodo ("J", "C2-C3") 32 % (70/220). La tasa total de errores de clasificación de esta segmentación es entonces 28,6 % (120/420). Notemos el decrecimiento de la tasa de errores cuando bajamos en el árbol.

Tabla 4.8. Tasas de errores de clasificación

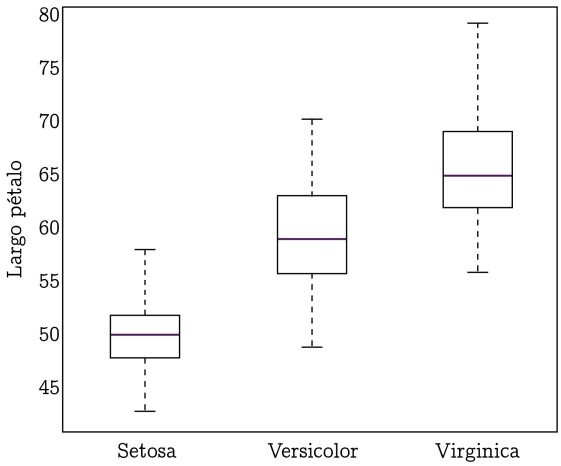
Raíz	NSE="ABC1"	NSE="C2-C3"	Total	Tasa NSE
450 (900)	200 (420)	230 (480)	430 (900)	48 %
Raíz	Edad="J"	Edad="A"	Total	Tasa Edad
450 (900)	120 (420)	150 (480)	270 (900)	30 %
Edad="J"	NSE="ABC1"	NSE="C2-C3"	Total	Tasa NSE en edad "J"
120 (420)	50 (200)	70 (220)	120 (420)	28,6 %
Edad="A	NSE="ABC1"	NSE="C2-C3"	Total	Tasa NSE en edad "A"
150 (480)	70 (220)	80 (260)	150 (480)	31,25 %

4.4.3 Caso de variable respuesta nominal no binaria

La variable respuesta del Ejemplo 4.1.2 es binaria. Veamos otro ejemplo, donde la variable respuesta tiene más de dos categorías.

Tomamos los datos famosos de R. Fisher, citados en sus escritos y presentados en el ejercicio 1.1 del primer capítulo. Se consideran tres especies de iris (flores): Setosa, Versicolor y Virginica, y cuatro mediciones: Largo del pétalo (LP), ancho del pétalo (AP), largo del sépalo (LS) y ancho del sépalo (AS). Se busca detectar cuáles de las cuatro mediciones discriminan mejor las tres especies. Tenemos entonces cuatro variables de segmentación numéricas y una variable respuesta nominal con tres categorías. En primer lugar, podemos visualizar las especies con un boxplot para cada una de las cuatro mediciones. En la Figura 4.13 se muestra los boxplot del largo del pétalo, donde podemos observar que las especies se diferencian. La especie Setosa tiene un longitud de pétalo más pequeña, mientras que la especie Virginica tiene el pétalo más largo.

FIGURA 4.13. Árbol de los iris



Para dividir un nodo se buscan los cortes de las mediciones que producen nodos hijos lo más puros posible. Tenemos que definir la impureza para el caso de una

variable respuesta con tres categorías, que se puede generalizar con un número de cualquier categoría. Los dos criterios definidos para una variable respuesta binaria se generalizan fácilmente.

Si $p_1(t)$, $p_2(t)$ y $p_3(t)$ son las proporciones de las tres especies en el nodo t, y $p_1(t) + p_2(t) + p_3(t) = 1$, el índice de Gini del nodo t se define como:

$$\gamma(t) = p_1(t)(1 - p_1(t)) + p_2(t)(1 - p_2(t)) + p_3(t)(1 - p_3(t)).$$

En un nodo t dado y dos nodos hijos t_1 y t_2, se calcula el índice de impureza como la media ponderada de los índices de Gini de t_1 y t_2:

$$G(t) = \frac{n_1}{n}\gamma(t_1) + \frac{n_2}{n}\gamma(t_2),$$

donde n, n_1 y n_2 son los tamaños de los nodos t, t_1 y t_2, respectivamente.

Se elige, entonces, entre las posibles segmentaciones del nodo t a aquella que produce la mayor reducción de impureza $G(t)$.

Aquí las variables de segmentación son numéricas. Para cada variable de segmentación tenemos que calcular los índices de Gini G para cada corte u posible y nos quedamos con el corte que minimiza la impureza.

Por ejemplo, en la raíz tenemos 50 flores de cada especie. El índice de Gini de la raíz vale entonces:

$$\gamma(raiz) = \frac{1}{3}(1 - \frac{1}{3}) + \frac{1}{3}(1 - \frac{1}{3}) + \frac{1}{3}(1 - \frac{1}{3}) = 0,667.$$

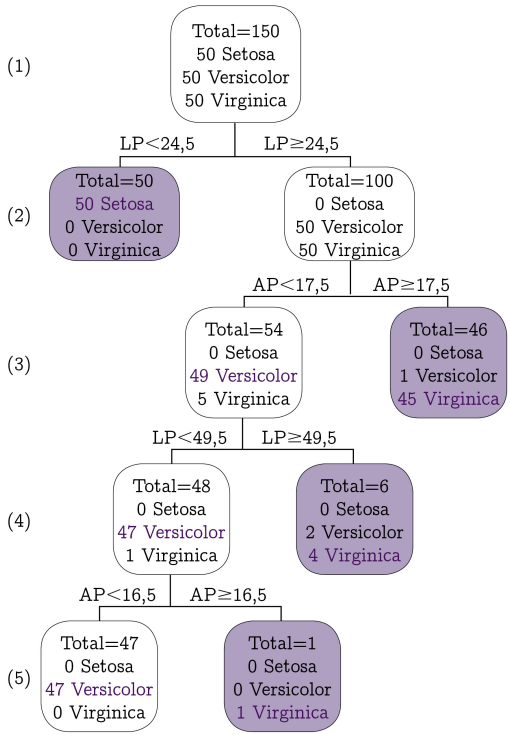
Consideremos, por ejemplo, la segmentación con el ancho del pétalo (AP) y un corte en el valor 12. Obtenemos una disminución de la impureza de la raíz de $0,667$ a $G = 0,408$. Considerando el largo del pétalo con un corte en 24,5, obtenemos una mayor disminución: $G = 0,333$ (Tabla 4.9). No podemos mostrar todas las etapas del cálculo aquí. Presentamos solamente el árbol resultante (Figura 4.14). Vemos que la mejor segmentación es en el corte 25,5 del largo del pétalo. No hubo índice G menor que 0,333 cuando se segmenta desde la raíz. Observen que el nodo $LP < 24,5$ es puro. Es un nodo terminal y todo iris con un largo de pétalo menor que 24,5 será clasificado en la especie Setosa y sin error de clasificación. La otra rama $LP \geq 24,5$ puede segmentarse. Esta vez la mejor división es en 17,5 del ancho del pétalo. Mostramos el árbol con cinco niveles, donde tres de los nodos terminales son puros. Pero uno de los nodos tiene una sola flor. Posiblemente habría que eliminar el último nivel, imponiendo, por ejemplo, un número mínimo de seis flores por nodo. Observen que las variables de segmentación largo y ancho del pétalo aparecen más de una vez. Por ejemplo, en el nivel (4) en el nodo izquierdo está definido por un largo de pétalo entre 24,5 y 49,5 y un ancho de pétalo menor que 17,5. Las otras dos variables de segmentación no intervienen. No significa que no permiten distinguir las tres especies, sino que el largo y ancho del pétalo son más eficaces para discriminar las especies.

Podemos calcular las tasas de errores de clasificación asociadas al árbol 4.14, que tiene una tasa de error total de 2 % (3/150). Si podamos el último nivel, la tasa sube

TABLA 4.9. Índices de Gini de los iris

Nodo	Frecuencia Nodo	Frecuencia Setosa	Frecuencia Versicolor	Frecuencia Virginica	Índice de Gini
$AP < 12$	60	50	10	0	0,278
$AP \geq 12$	90	0	40	50	0,494
G: $(0,278 \times 60 + 0,494 \times 90)/150 = 0,408$					
$LP < 24,5$	50	50	0	0	0,000
$LP \geq 24,5$	100	0	50	50	0,50
G: $(0,00 \times 50 + 0,50 \times 100)/150 = 0,333$					

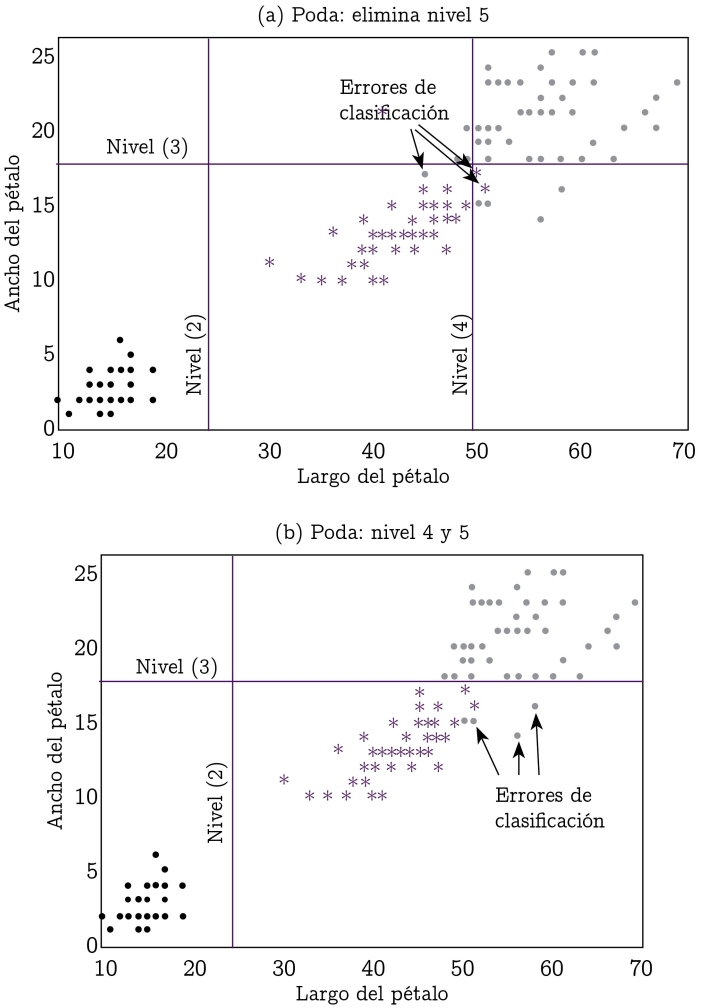
FIGURA 4.14. Árbol de los iris



a 2,7 % (4/150), que es pequeño también. La última segmentación parece forzada, ya que tiene un solo iris en uno de los nodos.

Mostramos gráficos de dispersión del largo y del ancho del pétalo (Figuras 4.15). Las especies fueron marcadas con diferentes símbolos. Las líneas corresponden a las diferentes segmentaciones utilizadas. El gráfico (a) corresponde al árbol con tres errores de clasificación cuando se conservan los cinco niveles, y el gráfico (b) corresponde al árbol con cuatro errores de clasificación cuando se podan los niveles 4 y 5.

FIGURA 4.15. Gráficos de dispersión de los iris



4.5 Resumen de la terminología

Variable respuesta: Variable que se busca explicar a partir de otras variables.
Variable explicativa: Variable que influye sobre una variable respuesta.
Árbol de regresión: Árbol de decisión cuya variable respuesta es numérica.
Árbol de clasificación: Árbol de decisión cuya variable respuesta es nominal.
Raíz del árbol: El nivel más alto del árbol que contiene todas las observaciones.
Nodo: Subconjunto de los datos definidos por una o más variables explicativas.
Nodo terminal: Nodo que no se ha dividido.
Regla de decisión: Conjunto de valores de una o más variables explicativas que se eligen para definir un subconjunto de datos.
Varianza intragrupo: Promedio de las varianzas de una misma variable medida en varios grupos.
Varianza intergrupo: Varianza de los promedios de una misma variable medida en varios grupos.
Índice de Gini: Criterio de segmentación basado en la impureza de los nodos.
Tasa de errores de clasificación: Tasa de errores obtenida clasificando observaciones de un árbol.

4.6 Ejercicios

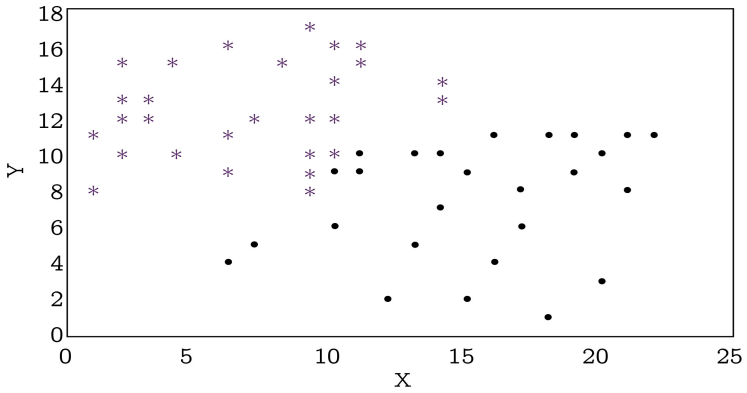
Ejercicio 4.1. Clasifica los siguientes casos en árbol de regresión o árbol de clasificación y especifica el tipo de variables de segmentación que se utilizan.

(a) Un cardiólogo estudia la posibilidad que sobrevivan más de 30 días pacientes que ingresan a un hospital con un ataque al corazón en un hospital considerando la presión arterial, el pulso, la edad y si es su primer ataque.
(b) El Banco Central hace un estudio para predecir la bancarrota de una empresa en función de indicadores económicos.
(c) Un nutricionista quiere modelar con CART la relación entre el índice de masa corporal (IMC), la edad y el género.
(d) La Unidad Técnico-Pedagógica de un colegio hace un estudio sobre la enseñanza del teorema de Pitágoras. En el estudio se considera el rendimiento de los alumnos en una prueba sobre el teorema, el género, el profesor y el método de enseñanza del profesor.
(e) El fisco busca detectar patrones de contribuyentes que permiten distinguir entre las declaraciones de impuestos legítimas y las fraudulentas con el objetivo de desarrollar así mecanismos para tomar medidas rápidas frente a ellas.
(f) Con el objetivo de detectar cuanto antes aquellos clientes que puedan estar pensando en rescindir sus contratos para, posiblemente, pasarse a la competencia, un banco encarga un estudio de patrones de comportamiento de clientes actuales y pasados. Estos patrones serán una ayuda para determinar el perfil de los clientes más proclives a darse de baja. El banco podrá hacer promociones especiales, etc., a los clientes con este perfil con el fin de retenerlos.

(g) El departamento de recursos humanos de un colegio recopila información sobre sus profesores para identificar las características de aquellos de mayor éxito. Los datos considerados se relacionan con los esfuerzos de sus profesores, su participación en diversas actividades y grupos de trabajo y los resultados obtenidos por sus alumnos. La información obtenida puede ayudar a la contratación de profesores a futuro.

Ejercicio 4.2. Se aplica un modelo CART a los datos de la figura adjunta, que tiene dos variables de segmentación, X_1 y X_2, y una variable respuesta binaria ("Rojo" y "Azul").

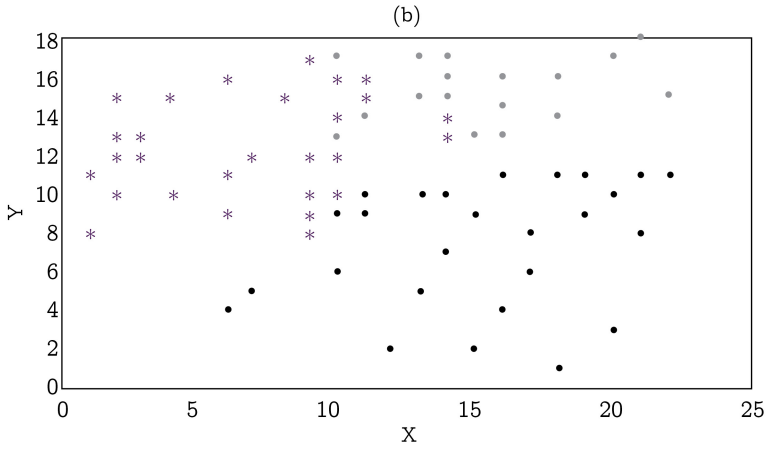
(a) Separa con tres líneas horizontales y/o verticales los dos grupos, "Rojo" y "Azul", de manera de minimizar la tasa de errores de clasificación.
(b) Construye el árbol de clasificación asociado.
(c) Calcula los coeficientes de Gini del árbol obtenido en (b).
(d) Clasifica una nueva observación con $X_1 = 8$ y $X_2 = 14$. Encuentra la probabilidad de equivocarse.



Ejercicio 4.3. Se quiere construir un modelo CART a los datos de la figura adjunta, que tiene dos variables de segmentación, X_1 y X_2, y una variable respuesta con tres categorías ("Rojo", "Azul" y "Verde") (Figura adjunta).

(a) Separa con tres líneas horizontales y/o verticales los dos grupos, "Rojo" y "Azul", de manera de formar grupos de tal manera que se minimice la tasa de errores de clasificación.
(b) Construye el árbol de clasificación asociado.
(c) Clasifica una nueva observación con $X_1 = 16$ y $X_2 = 16$. Encuentra la probabilidad de equivocarse.
(d) Clasifica una nueva observación con $X_1 = 5$ y $X_2 = 5$. Encuentra la probabilidad de equivocarse.

(e) Se poda el árbol de nivel. Clasifica nuevamente la observación con $X_1 = 5$ y $X_2 = 5$. Encuentra la nueva probabilidad de equivocarse.



Ejercicio 4.4. En un estudio de la PSU de Matemática de 2009, se obtienen las estadísticas por dependencia y género de la Región Metropolitana (Tablas 4.10 y 4.11).

(a) ¿Cómo se calcula el coeficiente η de las tablas?
(b) ¿Cuál es la primera segmentación de la raíz que optimiza η (Tabla 4.10)? Justifica.
(c) En la Tabla 4.11 están los resultados de las segmentaciones que podrían seguir. Construye el árbol correspondiente.
(d) Dé el árbol final usando un criterio de poda de 5 %.
(e) Utilizando la Tabla 4.12 y el árbol (d), estima la PSU de Matemática de un alumno de un colegio particular pagado. Da un intervalo de confianza de 95 %.
(f) Utilizando la Tabla 4.12 y el árbol (d), estima la PSU de Matemática de una alumna de un colegio municipal. Da un intervalo de confianza de 95 %.

Tabla 4.10. Primera segmentación

Variable	Tamaño	Varianza intergrupo	Varianza total	η	F	p-valor
Nivel (1) ->(2)	Raíz					
Género	97.228	0,833	12.865	0,0000	6,29	0,012
Dependencia (Mu)-(PS+PP)	97.228	107,14	12.865	0,008	816,52	0,000
Dependencia (PS)-(Mu+PP)	97.228	51,2	12.865	0,004	388,52	0,000
Dependencia (PP)-(Mu+PS)	97.228	532,5	12.865	0,040	4.198	0,000

Tabla 4.11. Segundas segmentaciones

Variable	Tamaño	Varianza intergrupo	Varianza total	η	F	p-valor
Nivel (2) ->(3)	Nodo PP					
Género	14.866	5,177	13.500	0,0004	5,7	0,017
Nivel (2) ->(3)	Nodo Mu+PS					
Género	82.362	0,144	12.122	0,00001	0,98	0,32
Mu, PS	82.362		12.122	0,002	134,7	0,000
Nivel (3) ->(4)	Nodo MU					
Género	29.162	0,0162	12.026	0,000	0,04	0,84
Nivel (3) ->(4)	Nodo PS					
Género	53.200	0,78	12.143	0,000	13,44	0,064

Tabla 4.12. Primera segmentación

Género		PP	PS	MU	Total
H	Frecuencia	7720	23582	14537	45839
	Media	559,9	502,3	492,1	508,7
	Desv. Estándar	116,3	110,8	109,5	113,8
M	Frecuencia	7146	29618	14625	51389
	Media	564,5	500,5	491,9	506,9
	Desv. Estándar	116,0	109,7	109,8	113,1
Total	Frecuencia	14866	53200	29162	97228
	Media	462,1	501,3	492,0	507,8
	Desv. Estándar	116,2	110,2	109,7	113,4

Anexo 1: Solución de los ejercicios

Capítulo 1

Ejercicio 1.1

(a) El porcentaje de conservación de la varianza en el plano principal es $72,9\,\% + 22,9\,\% = 95,8\,\%$, lo que es bastante elevado e indica correlaciones fuertes entre las variables.

(b) Las especies se distinguen en el primer plano principal, especialmente la Setosa. Se diferencian según la primera C.P., que está muy correlacionada con el ancho y el largo del pétalo.

(c) El ancho y el largo del pétalo, y en menor grado el largo del sépalo. La segunda componente principal se explica sobre todo con el ancho del sépalo.

(d) El ancho y el largo del pétalo son altamente correlacionados positivamente. El ancho del sépalo es poco correlacionado a las tres otras variables.

Ejercicio 1.2 Consideramos la matriz de correlación $R = \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}$.

(a) Los valores propios son solución de la ecuación: $det(R - \lambda I_2) = (1-\lambda)^2 - r^2 = 0$. Se obtienen dos valores propios: $1 + r$ y $1 - r$.

(b) Se tiene que resolver: $Ru_1 = (1 + r)u_1$ y $Ru_2 = (1 - r)u_2$. Si $u_1 = \begin{pmatrix} u_{11} \\ u_{21} \end{pmatrix}$. Se obtiene $u_{11} = u_{21}$. Con la normalización, $u_1 = \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix}$.

Si $u_2 = \begin{pmatrix} u_{12} \\ u_{22} \end{pmatrix}$, obtenemos $u_{12} = -u_{22} = \frac{1}{\sqrt{2}}$. O sea, $u_2 = \begin{pmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{pmatrix}$.

(c) En el círculo de correlaciones $\mathcal{C}$, las coordenadas de las variables son iguales a $\sqrt{1+r}u_1$ sobre la primera C.P. y $\sqrt{1-r}u_2$ sobre la segunda C.P. Obtenemos para la primera variable $\begin{pmatrix} \sqrt{1+r}/\sqrt{2} \\ \sqrt{1-r}/\sqrt{2} \end{pmatrix}$ y para la segunda $\begin{pmatrix} \sqrt{1+r}/\sqrt{2} \\ -\sqrt{1-r}/\sqrt{2} \end{pmatrix}$.

Dibuja el círculo de correlaciones y observa que los dos puntos son sobre la circunferencia del círculo $((\sqrt{1+r})^2/2 + (\sqrt{1-r})^2/2 = 1)$. Como partimos de una representación en $\mathbb{R}^2$, era de esperar que la representación en el plano conservará toda la información.

(d) El porcentaje de conservación de la varianza sobre la primera componente principal es $100 \times (1 + r)/2$, si $r > 0$, si no es $100 \times (1 - r)/2$.

(e) En el caso $r = 0$, hay un valor propio doble que es igual a 1. Cualquier vector unitario del plano es vector propio.

Ejercicio 1.5

(a) Se observa que las dos variables de radiactividad tienen una correlación alta (0,743) y las mediciones de tamaño, también. Pero las correlaciones entre las variables de radiactividad y las variables de tamaño son bajas.

(b) En el plano principal se conserva 64,8 %+22,4 %= 87,2 % de la varianza total. Se pueden ver cuáles son los peces grandes (por ejemplo, 1,2,3,4) y los pequeños (por ejemplo, 5,6,7,8) y cuáles tienen radiactividad alta (por ejemplo, 21 y 22) y cuáles tienen radiactividad baja (por ejemplo, 1 y 2). Cada acuario tiene peces de diferentes tamaños, pero en el acuario A1 los peces tienen radiactividades bajas; en A2, medianas y en A3, altas. Concluimos que los acuarios se diferencian por la radiactividad.

(c) La primera C.P. está correlacionada positivamente con las variables de tamaño y negativamente con las variables de radiactividad. Las proyecciones de las variables se ubican cerca de la circunferencia del círculo, lo que permiten decir que el coseno del ángulo formado por dos variables es una buena aproximación del coeficiente de correlación entre ellas. De esta manera, tenemos una buena visualización de la matriz de correlación y la confirmación de las conclusiones del punto (a). En particular, las variables de radiactividad hacen ángulos cercanos al 90^o con las variables de tamaño.

Ejercicio 1.6

(a) Los valores propios son: $2,17$; $0,79$ y $0,036$. Los vectores propios son:

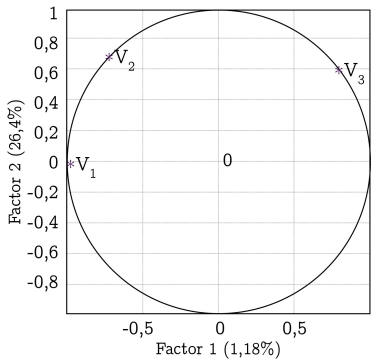
$$u_1 \begin{pmatrix} -0,67 \\ -0,50 \\ 0,55 \end{pmatrix} \quad u_2 \begin{pmatrix} -0,028 \\ 0,76 \\ 0,65 \end{pmatrix} \quad u_3 \begin{pmatrix} 0,84 \\ -0,42 \\ 0,52 \end{pmatrix}.$$

(b) Las coordenadas de las tres variables en el círculo son: $V_1 = (-0,989;\ -0,025)$, $V_2 = (-0,736;\ 0,673)$ y $V_3 = (0,807;\ 0;582)$ (Ver la figura adjunta).

(c) El porcentaje de conservación de la varianza en C es casi 100 %: 72,4 %+26,4 %=98,8 %, lo que es casi 100 %. Era de esperar que las variables fueran bastante correlacionadas.

Ejercicio 1.7

(a) En el círculo de correlaciones aparecen dos grupos de cualidades: regularidad y técnica del revés, etc., que están altamente correlacionadas negativamente con la primera C.P., y los golpes especiales, tales como smash o servicios que están altamente correlacionadas positivamente con la segunda C.P. Estos dos grupos de cualidades no son correlacionados. Las cualidades del físico y síquico se encuentran entremedio.



(b) Un jugador arriba del cuadrante de la izquierda tiene todas las cualidades. Encontramos jugadores regulares con técnica en los golpes en general, como Borg, y otros con golpes potentes y especiales, como McEnroe. A la izquierda se encuentran los jugadores arriba del promedio.

(c) Si queremos usar la primera C.P. como índice de 'ranking" de los jugadores, convendría cambiarle el signo, de manera que los mejores jugadores tomen los valores mayores. El orden de los jugadores según la primera C.P. y el promedio de las 15 cualidades dadas son diferentes y a veces bastante diferentes. Se debe a que la primera C.P. no toma mucho en cuenta de las cualidades del segundo grupo. Es así que McEnroe se ve desfavorecido con la primera C.P. El problema aquí es que las cualidades del primer grupo son más numerosas que las del segundo grupo, lo que favorece el primer grupo en la primera C.P.

Capítulo 2

Ejercicio 2.1

(a) Tenemos los datos de censo. No debería necesitar un test de hipótesis.
(b) Los datos provienen de una muestra. Se requiere un test de hipótesis.
(c) Tenemos los datos de censo. No debería necesitar un test de hipótesis.
(d) Los datos provienen de una muestra. Se requiere un test de hipótesis.

Ejercicio 2.2

(a) Se requiere un diseño experimental con dos grupos de alumnos elegidos al azar.
(b) Se requiere un diseño muestral.
(c) Se requiere diseño experimental en el cual se someta a las piezas a diversas temperaturas.
(d) Se requiere diseño experimental.

(e) Se requiere un diseño muestral. Se recolectan datos en diversos lagos, en los cuales se miden la acidez y la cantidad de peces.

Ejercicio 2.3

La región crítica para $\alpha = 5\,\%$ es $\mathcal{R} = \{\bar{x} \leq 65 - 1,66\frac{6}{80} = 63,89\}$. Se rechaza la hipótesis nula $H_o : \mu = 65$ contra $H_1 : \mu < 65$ dado que 63,5 es menor al 63,89.

Ejercicio 2.4

(a) $H_o : \mu = 1.600$ contra $H_1 : \mu < 1600$.

(b) $\alpha = \mathbb{P}(\bar{x} \leq c|\mu = 1600) = \mathbb{P}(\frac{\bar{x}-1600}{120/\sqrt{99}} < \frac{c-1600}{120/\sqrt{99}}) = 0,05$.
$\mathbb{P}(t_{99} < -1,66) = 0,05 \Longrightarrow c = 1600 - 1,66 \times 120/sqrt99 = 1580$. Se rechaza la hipótesis nula con un error de $5\,\%$. $\beta = \mathbb{P}(\bar{x} > 1580|\mu = \mu_1)$ donde $\mu_1 < 1600$.

(c) El valor encontrado en la muestra es de 1570 que es menor que 1580. Se concluye que el p-valor es menor que $5\,\%$. Podemos rechazar la hipótesis nula con un error menor que $5\,\%$.

Ejercicio 2.5

Las hipótesis son $H_o :\ \mu = 500$ contra $H_1 :\ \mu > 500$. Se supone la desviación estándar conocida e igual a 100 litros. El p-valor del test es:

$$\mathbb{P}(\bar{x}_{\geq}|\mu = 500) = \mathbb{P}(t_{59} \geq (525 - 500)/100/sqrt59) = \mathbb{P}(t_{59} \leq 1,92) \approx 0{,}027.$$

Se rechaza la hipótesis nula con un error de $2{,}7\,\%$ y se concluye que el nuevo proceso es eficaz.

Ejercicio 2.6

El intervalo de confianza es: $[0,30-1,96*\sqrt{0,3 \times 0,7/n}; 0,30+1,96*\sqrt{0,3 \times 0,7/n}] \Longrightarrow 1,96\sqrt{0,30 \times 0,70/n} = 0,031 \Longrightarrow n = (1,96\sqrt{0,30 \times 0,70/0,031})^2 \Longrightarrow n = 840$.

Para un nivel de confianza de $1\,\%$ y $n = 64$ tenemos el intervalo:

$$[0{,}35 - 2,326\sqrt{0,35 \times 0,65/64}; 0{,}35 + 2,326\sqrt{0,35 \times 0,65/64}] = [0,211; 0,489].$$

Obtenemos un intervalo muy ancho. El error es igual a $2,326\sqrt{0,35 \times 0,65/64} = 4,96\,\%$. Para disminuirlo tenemos que aumentar el tamaño de la muestra.

Ejercicio 2.7

Se debe aplicar un test de hipótesis: $H_o :\ \mu = 100$ contra $H_1 :\ \mu < 100$. Calculamos el p-valor del test:

$$\mathbb{P}(\bar{x} \leq 97|\mu = 100) = \mathbb{P}(t_{349} \leq (97 - 100)/78/\sqrt{349}) = \mathbb{P}(t_{349} \leq -0,718) \approx 0,236.$$

No se rechaza la hipótesis nula y se concluye que no hay evidencia para que el laboratorio tenga que revisar su proceso de producción.

Ejercicio 2.8

(a) $H_o :\ \mu = \mu_o$ contra $H_1 :: \mu \neq \mu_o$, donde $\mu_o = 1,45$ millones.

(b) $\mathcal{R} = \{\bar{x} \leq c_1\} \cup \{\bar{x} \geq c_2\}$ y p-valor$=2\mathbb{P}(\bar{x} \geq \bar{x}|\mu = 1,6\ millones)$.

(c) $c_1 = 1,35$ millones y $c_2 = 1,55$ millones. Se rechaza H_o con un error de $5\,\%$. El p-valor$=2 \times 0,0062 = 0,0124$. Se puede rechazar H_o con un error mayor o igual a $1{,}24\,\%$.

Ejercicio 2.9
Un test de hipótesis permite responder. Las hipótesis son: $H_o: \ \mu \leq 5$ contra $H_1: \mu > 5$. La región crítica: $\mathcal{R} = \{\bar{x} \geq 5,18\}$, por lo cual se rechaza H_o. El fabricante no cumple la normativa.
Ejercicio 2.10
$\bar{x}_1 = 3,077$; $s_1 = 0,750$; $n_1 = 12$; $\bar{x}_2 = 3,506$; $s_2 = 0,521$; $n_2 = 14$. El estadístico del test: $T = \frac{(\bar{x}_2 - \bar{x}_1)/\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}{\sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1+n_2-2}}} \sim t_{n_1+n_2-2}$, bajo H_o. En la muestra $T = -1,56$ y $\mathbb{P}(t_{24} \leq -1{,}71) = 5\,\%$. No se rechaza H_o. El p-valor vale 0,0659.
Ejercicio 2.11 Son 50 alumnos. El p-valor dela F es nulo. Se concluye que hay diferencia entre las facultades.

Tabla 4.13. ANOVA de las Facultades

Fuente de variabilidad	Suma de cuadrados	Grados de libertad	Cuadrados medios	F	p-valor
Facultad	4125,7	3	1375,2	24,83	0,000
Residuos	2547,9	46	55,39		
Total	6673,6	49			

Ejercicio 2.12
$V = 3,174^2 = 10,074$; $W = (20 \times 1,71^2 + 20 \times 1,94^2 + 30 \times 1,856^2 + 30 \times 1,394)/100 = 2,954$; $B = V_W = 7,12$; $F = \frac{B/3}{W/96} = 77,13$; p-valor=$\mathbb{P}(F_{3,96} \geq 77,13) = 0$. Hay efecto de la dosis sobre la producción.

Capítulo 3

Ejercicio 3.1

(a) Ver las tablas adjuntas completas.
(b) n=499+4+1=584;
(c) Los p-valores son las probabilidades $\mathbb{P}(|t_{499}| \geq t - \textit{student observada})$.
(d) Las PSU de Matemática y Ciencia y NEM son significativas. La PSU de lenguaje no es tan significativa (el p-valor es del orden de 8 %, que es mayor que 5 %). Sin embargo, el coeficiente de correlación múltiple no es muy alto, lo que da cierta reserva para hacer predicciones.

Variable	Estimación	Desviación estándar	t-student	p-valor
Constante	-53,98	12,2152	-4,419	0,000
Matemática	0,0312	0,0106	2,958	0,003
NEM	0,0245	0,0078	3,129	0,002
Lenguaje	0,0135	0,0077	1,744	0,082
Ciencia	0,753	0,0096	7,843	0,000

Fuente	Suma cuadrados	Grados libertad	Cuadrados medios	F	p-valor
Regresión	12540	4	3135	24,78	0,0000
Residuos	63131	499	126,51		
Total	75671	583			

Ejercicio 3.2

(a) Ver las tablas adjuntas completas.
(b) n=20. Los p-valores son las probabilidades $\mathbb{P}(|t_{499}| \geq t - student\ observada)$.
(c) La calidad del modelo es parecido al anterior. Quizás habría que eliminar la variable Temperatura también.

Variable	Estimación	Desviación estándar	t-student	p-valor
Constante	43,9534	19,505	?	0,0386
Número empresas		0,0156	3,229	0,0052
Temperatura	-0,4485	0,3355	-1,337	0,2000
Población	-0,0240	?	-1,572	0,1339
R=0,905				

Fuente	Suma cuadrados	Grados libertad	Cuadrados medios	F	p-valor
Regresión	8.619,6	3	?	24,18	0,0000
Residuos	1.901,2	?	118,83		
Total	?	19			

Variable	Estimación	Desviación estándar	t-student	p-valor
Constante	47,373	20,8274	2,2745	0,0362
Número empresas	0,0266	0,0040	6,6509	0,000
Temperatura	-0,5826	0,3486	-1,6712	0,1130
R=0,89				

Fuente	Suma cuadrados	Grados libertad	Cuadrados medios	F	p-valor
Regresión	8.325,8	2	4.162,9	32,24	0,0000
Residuos	2.195,0	17	129,12		
Total	?	19			

Ejercicio 3.3

(a) Ver la tabla adjunta.
(b) Ver la tabla adjunta.
(c) n=55.
(d) $R = 887994/902773 = 0,984$.
(e) $F = 222000/29558 = 7,51$ con un p-valor nulo, donde $F \sin F_{4,50}$. El modelo es significativo.
(f) $\mathbb{P}(t_{50} \leq -2,01) = \mathbb{P}(t_{50} \geq 2,01) = 0,025 \Longrightarrow I = [0,734 - 2,01 \times 0,4741;\ 0,734 + 2,01 \times 0,4741] = [-0,215;\ 1,683]$. Confirmamos que la variable X_2 no es significativa en el modelo.
(g) El p-valor es menor que 5 %. Se rechaza la hipótesis nula.

Variable	Estimación	Desviación estándar	t-student	p-valor
Constante	23,45	14,90	1,57	0,122
X_1	0,9321	0,08602	10,84	0,000
X_2	0,734	0,4721	1,5554	0,126
X_3	-0,4982	0,1520	3,278	0,002
X_4	3,486	2,274	1,533	0,132

Fuente	Suma cuadrados	Grados libertad	Cuadrados medios	F	p-valor
Regresión	887994	4	222000	7,51	0,0000
Residuos	14779	50	29558		
Total	902773	54			

Ejercicio 3.4

(a) La correlación múltiple R es relativamente alta y el p-valor de la F de Fisher es casi nulo: $\mathbb{P}(F_{2,17} \geq 14,739) = 0,0002$. Se concluye que el modelo es globalmente significativo. Sin embargo, la variable z no es significativa en el modelo.
(b) $I_1 = [-7,436;\ -1,752]$ y $I_2 = [-0,125;\ 0,149]$.
(c) $Var(\hat{e}_i) = var(y) \times (1 - R^2) = 11,2$.
(d) Quizás se cumple, pero los residuos dependen de la variable y, lo que muestra que en el modelo faltan variables explicativas o que el modelo no es lineal.

Ejercicio 3.5

(a) R es alto y $F \sim F_{3,96}$ tiene un p-valor nulo. El modelo es globalmente significativo.
(b) No. La variable Trayecto no es significativa (p-valor=0,9372).
(c) $I = [0,0131;\ 0,0163]$.
(d) El segundo o tercero.

Ejercicio 3.6

(a) El coeficiente de correlación múltiple es la raíz del cociente de la varianza de los $\hat{y}_i$ y de la varianza de los y_i. Mide el grado de ajuste del modelo.

(b) $F = \frac{(n-p)\times R^2}{(p-1)\times(1-R^2)}$, donde n=20; p=4; $F = 1,90$. Los grados de libertad son 3 y 16. Interprétalo y concluye.

(c) Es razonable eliminar la variable "Ejercicios" del modelo, dado que el p-valor de su coeficiente es alto. De hecho, el coeficiente de correlación múltiple prácticamente no cambia.

(d) Se deberían examinar los gráficos con residuos, valores estimados del colesterol y el histograma de los residuos.

Ejercicio 3.7

(a) El t-Student de cada coeficiente es el cociente del coeficiente y de su desviación estándar. Tiene 91-5= 86 grados de libertad. Si el valor absoluto de la t es alto, la variables asociada es significativa en el modelo.

(b) La F tiene 5 y 86 grados de libertad. Mide el "error relativo" entre el modelo constante y el modelo propuesto.

(c) $I = [-0,028 - 1,96 \times 0,128;\ -0,028 + 1,96 \times 0,128] = [-0,279;\ 0,223]$. El intervalo esta centrado en un valor cercano al 0. Se concluye que la aceleración no es significativa.

Capítulo 4

Ejercicio 4.1

(a) Se puede aplicar un árbol de clasificación con la variable respuesta binaria "Sobrevive o no" después de ingresar en el hospital con un ataque al corazón; las variables de segmentación presión arterial, pulso y edad son numéricas y "si es el primer ataque o no" es binaria.

(b) Se puede aplicar un árbol de clasificación con la variable respuesta binaria "bancarrota o no" y las variables de segmentación son indicadores económicos de la empresa.

(c) Se puede aplicar un árbol de regresión con la variable respuesta "IMC". La variable de segmentación edad es numérica y el género es binaria.

(d) Se puede aplicar un árbol de regresión con la variable respuesta el "rendimiento del alumno" y variables de segmentación cualitativas.

(e) Se puede aplicar un árbol de clasificación con la variable respuesta binaria (fraude o no). Las variables de segmentación pueden ser de distintos tipos.

(f) Se puede aplicar un árbol de clasificación con la variable respuesta binaria "Fue a la competencia o no". Las variables de segmentación pueden ser de distintos tipos.

(g) Se puede aplicar un árbol de regresión con la variable respuesta "el grado de éxito del profesor". Las variables de segmentación pueden ser de distintos tipos.

La variable respuesta es el tabaquismo, que es una variable binaria. Las variables de segmentación son el tiempo dedicado al deporte, el género y la edad. Tenemos un árbol de clasificación. La variable respuesta es el rendimiento de los estudiantes, que es cuantitativa. Tenemos un árbol de regresión.

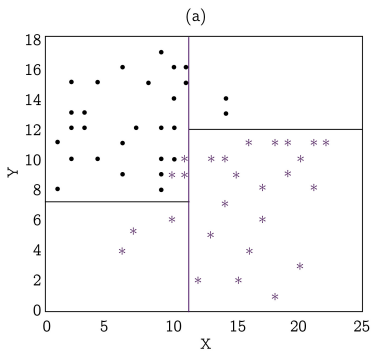
Ejercicio 4.2

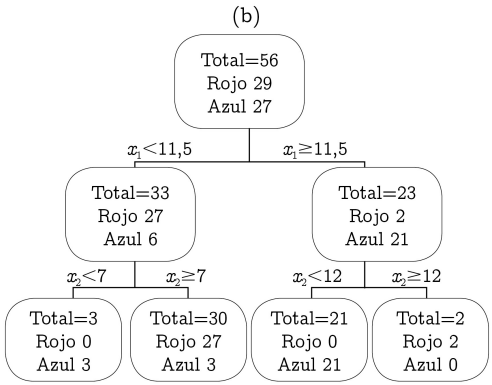
(a) La líneas posibles son dadas en la Figura 4.6(a), donde se obtienen tres errores de clasificaciones, o sea, 3/56=5,4 %.

(b) El árbol de clasificación asociado se encuentra en la Figura 4.6(b).

(c) El coeficiente de Gini G_2 para el nivel 2 (de bajo de la raíz): $G_2 = \frac{33}{56}\gamma_1 + \frac{23}{56}\gamma_2$ donde $\gamma_1 = \frac{27}{33}\frac{6}{33} = 0,1488$ y $\gamma_2 = \frac{2}{23}\frac{21}{23} = 0,0794$. Luego $G_2 = 0,1203$.
Para el nivel 3, a la izquierda, $G_{3(i)} = \frac{3}{33}\gamma_3 + \frac{30}{33}\gamma_4$, donde $\gamma_3 = 0$ y $\gamma_4 = \frac{27}{30}\frac{3}{30} = 0,09$. Luego $G_{3(i)} = 0,0818$. A la derecha, $G_{3(d)} = \frac{21}{23}\gamma_5 + \frac{2}{23}\gamma_6 = 0$, donde $\gamma_5 = 0$ y $\gamma_6 = 0$. La derecha tiene nodos puros.

(d) La nueva observación $X_1 = 8$ y $X_2 = 14$ se clasifica "roja" con un 10 % de error.





Ejercicio 4.3

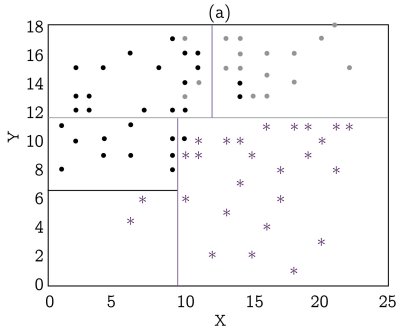
(a) La líneas posibles son dadas en la Figura 4.6(a), donde se obtienen cinco errores de clasificaciones, o sea 5/73=6,8 %.

(b) El árbol de clasificación asociado se encuentra en la Figura 4.6(b).

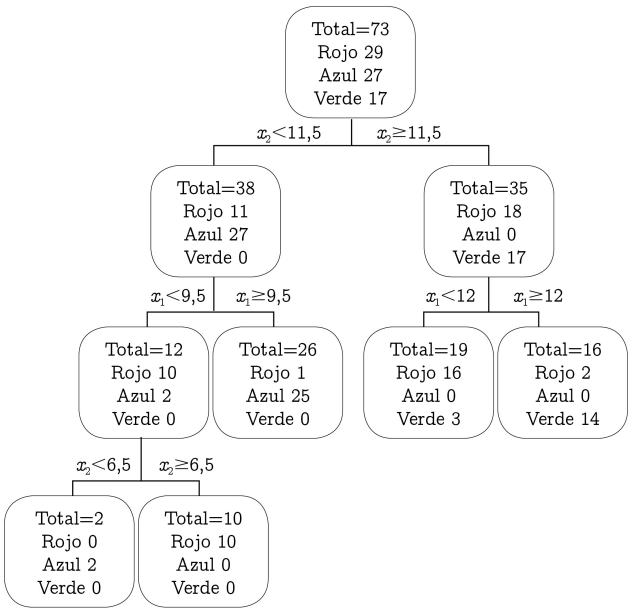
(c) La nueva observación $X_1 = 16$ y $X_2 = 16$ se clasifica "verde" con un 12,5 % de error.

(d) La nueva observación $X_1 = 5$ y $X_2 = 5$ se clasifica "Azul" con un 0 % de error.

(e) La nueva observación $X_1 = 5$ y $X_2 = 5$ se clasifica "Rojo" con un 2/12=16,7 % de error.



(b)



Ejercicio 4.4

(a) El coeficiente η es el cociente de la varianza intergrupo con la varianza total. Varía entre 0 y 1. Cuando esta cercano a 1, los grupos son diferentes; y cuando está cercano a 0, no hay mucha diferencia entre los grupos en relación con la varianza al interior de los grupos.

(b) La dependencia que segmenta PP de Mu+PS. Tiene la varianza intergrupo (o η) más alta.

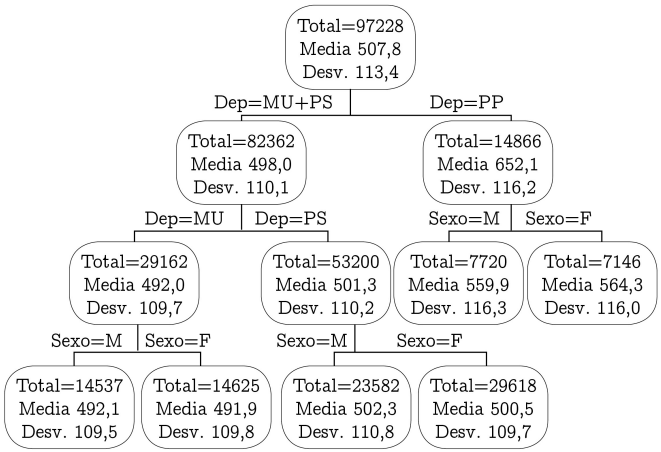
(c) El árbol corresponde a (a). Las medias y desviaciones estándares se obtienen de la Tabla. Por ejemplo. en el nodo Dep=MU+PS, la media es igual a $\frac{53200\times501,3+29162\times492,0}{53200+29162} = 498,0$ y la desviación estándar es igual a $\sqrt{\frac{53200\times110,2^2+29162\times109,7^2}{53200+29162}} = 110,1$.

(d) Hay que podar los nodos con p-valores mayores a 5 %. Se obtiene el árbol (b).

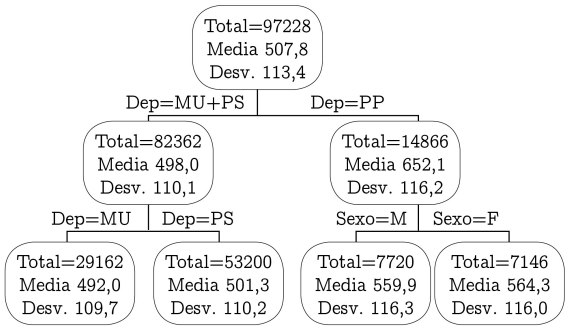
(e) 559,9 con un intervalo de confianza: $[559,9 - 1,96 \times \frac{116,3}{\sqrt{7720}};\ 559,9 + 1,96 \times \frac{116,3}{\sqrt{7720}}] = [557,3;\ 562,5]$.

(f) 492,0 con un intervalo de confianza: $[492 - 1,96 \times \frac{109,7}{\sqrt{29162}};\ 492 + 1,96 \times \frac{109,7}{\sqrt{29162}}] = [490,7;\ 493,3]$.

(a) PSUM



(b) PSUM


Total=97228
Media 507,8
Desv. 113,4
Dep=MU+PS
Dep=PP
Total=82362
Media 498,0
Desv. 110,1
Total=14866
Media 652,1
Desv. 116,2
Dep=MU
Dep=PS
Sexo=M
Sexo=F
Total=29162
Media 492,0
Desv. 109,7
Total=53200
Media 501,3
Desv. 110,2
Total=7720
Media 559,9
Desv. 116,3
Total=7146
Media 564,3
Desv. 116,0

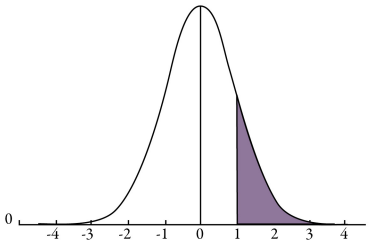
Anexo 2: Tablas Estadísticas

La tabla de distribución Normal $\mathcal{N}(0,1)$ no es fácil de usar. Tomando en cuenta que la distribución $\mathcal{N}(0,1)$ es simétrica con respecto de 0, la tabla presenta solamente las probabilidades de los valores positivos. Por ejemplo, para $\mathbb{P}(Z \geq 1,03)$, se busca la parte entera y la decimal de 1,03 que son, respectivamente, 1,0 en la primera columna y se busca la centesimal que es 0,03 en la primera fila. El valor de la probabilidad se encuentra, entonces, al cruce de la fila en el valor 1,0 y de la columna en el valor 0,03. La probabilidad es 0,1515.

Si se quiere $\mathbb{P}(Z \leq -1)$, basta tomar $1 - \mathbb{P}(Z \geq 1) = 0,159$.

Si $X \sim \mathcal{N}(2,3)$ y se quiere $\mathbb{P}(X \leq 2,5)$ se normaliza X: $Z = \frac{X-2}{3}$ y se calcula $\mathbb{P}(Z \leq \frac{2,5-2}{3}) = \mathbb{P}(Z \leq 0,17) = 1 - \mathbb{P}(Z \geq 0,167)$. En la tabla $\mathbb{P}(Z \geq 0,17) = 0,4325$. Se deduce $\mathbb{P}(X \leq 2,5) = 0,5675$.

DISTRIBUCIÓN NORMAL



Ejemplo:

X~N(1,2)

Z=(X-1)/2~N(0,1)

P(Z>1)=0,1587

Área gris de bajo de la curva

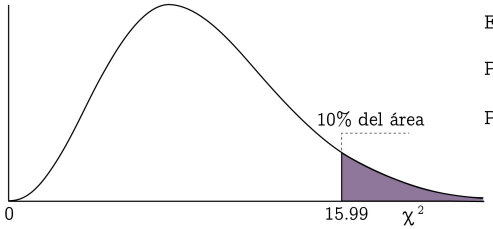
P(X>3)=0,1587

Decimal	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,4960	0,4920	0,4880	0,4840	0,4801	0,4761	0,4721	0,4681	0,4641
0,1	0,4602	0,4562	0,4522	0,4483	0,4443	0,4404	0,4364	0,4325	0,4286	0,4247
0,2	0,4207	0,4168	0,4129	0,4090	0,4052	0,4013	0,3974	0,3936	0,3897	0,3859
0,3	0,3821	0,3783	0,3745	0,3707	0,3669	0,3632	0,3594	0,3557	0,3520	0,3483
0,4	0,3446	0,3409	0,3372	0,3336	0,3300	0,3264	0,3228	0,3192	0,3156	0,3121
0,5	0,3085	0,3050	0,3015	0,2981	0,2946	0,2912	0,2877	0,2843	0,2810	0,2776
0,6	0,2743	0,2709	0,2676	0,2643	0,2611	0,2578	0,2546	0,2514	0,2483	0,2451
0,7	0,2420	0,2389	0,2358	0,2327	0,2296	0,2266	0,2236	0,2206	0,2177	0,2148
0,8	0,2119	0,2090	0,2061	0,2033	0,2005	0,1977	0,1949	0,1922	0,1894	0,1867
0,9	0,1841	0,1814	0,1788	0,1762	0,1736	0,1711	0,1685	0,1660	0,1635	0,1611
1,0	0,1587	0,1562	0,1539	0,1515	0,1492	0,1469	0,1446	0,1423	0,1401	0,1379
1,1	0,1357	0,1335	0,1314	0,1292	0,1271	0,1251	0,1230	0,1210	0,1190	0,1170
1,2	0,1151	0,1131	0,1112	0,1093	0,1075	0,1056	0,1038	0,1020	0,1003	0,0985
1,3	0,0968	0,0951	0,0934	0,0918	0,0901	0,0885	0,0869	0,0853	0,0838	0,0823
1,4	0,0808	0,0793	0,0778	0,0764	0,0749	0,0735	0,0721	0,0708	0,0694	0,0681
1,5	0,0668	0,0655	0,0643	0,0630	0,0618	0,0606	0,0594	0,0582	0,0571	0,0559
1,6	0,0548	0,0537	0,0526	0,0516	0,0505	0,0495	0,0485	0,0475	0,0465	0,0455
1,7	0,0446	0,0436	0,0427	0,0418	0,0409	0,0401	0,0392	0,0384	0,0375	0,0367
1,8	0,0359	0,0351	0,0344	0,0336	0,0329	0,0322	0,0314	0,0307	0,0301	0,0294
1,9	0,0287	0,0281	0,0274	0,0268	0,0262	0,0256	0,0250	0,0244	0,0239	0,0233
2,0	0,0228	0,0222	0,0217	0,0212	0,0207	0,0202	0,0197	0,0192	0,0188	0,0183
2,1	0,0179	0,0174	0,0170	0,0166	0,0162	0,0158	0,0154	0,0150	0,0146	0,0143
2,2	0,0139	0,0136	0,0132	0,0129	0,0125	0,0122	0,0119	0,0116	0,0113	0,0110
2,3	0,0107	0,0104	0,0102	0,0099	0,0096	0,0094	0,0091	0,0089	0,0087	0,0084
2,4	0,0082	0,0080	0,0078	0,0075	0,0073	0,0071	0,0069	0,0068	0,0066	0,0064
2,5	0,0062	0,0060	0,0059	0,0057	0,0055	0,0054	0,0052	0,0051	0,0049	0,0048
2,6	0,0047	0,0045	0,0044	0,0043	0,0041	0,0040	0,0039	0,0038	0,0037	0,0036
2,7	0,0035	0,0034	0,0033	0,0032	0,0031	0,0030	0,0029	0,0028	0,0027	0,0026
2,8	0,0026	0,0025	0,0024	0,0023	0,0023	0,0022	0,0021	0,0021	0,0020	0,0019
2,9	0,0019	0,0018	0,0018	0,0017	0,0016	0,0016	0,0015	0,0015	0,0014	0,0014
3,0	0,0013	0,0013	0,0013	0,0012	0,0012	0,0011	0,0011	0,0011	0,0010	0,0010

DISTRIBUCIÓN BINOMIAL

						p					
n	*k*	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
1	0	0.9500	0.9000	0.8500	0.8000	0.7500	0.7000	0.6500	0.6000	0.5500	0.5000
1	1	0.0500	0.1000	0.1500	0.2000	0.2500	0.3000	0.3500	0.4000	0.4500	0.5000
2	0	0.9025	0.8100	0.7225	0.6400	0.5625	0.4900	0.4225	0.3600	0.3025	0.2500
2	1	0.0950	0.1800	0.2550	0.3200	0.3750	0.4200	0.4550	0.4800	0.4950	0.5000
2	2	0.0025	0.0100	0.0225	0.0400	0.0625	0.0900	0.1225	0.1600	0.2025	0.2500
3	0	0.8574	0.7290	0.6141	0.5120	0.4219	0.3430	0.2746	0.2160	0.1664	0.1250
3	1	0.1354	0.2430	0.3251	0.3840	0.4219	0.4410	0.4436	0.4320	0.4084	0.3750
3	2	0.0071	0.0270	0.0574	0.0960	0.1406	0.1890	0.2389	0.2880	0.3341	0.3750
3	3	0.0001	0.0010	0.0034	0.0080	0.0156	0.0270	0.0429	0.0640	0.0911	0.1250
4	0	0.8145	0.6561	0.5220	0.4096	0.3164	0.2401	0.1785	0.1296	0.0915	0.0625
4	1	0.1715	0.2916	0.3685	0.4096	0.4219	0.4116	0.3845	0.3456	0.2995	0.2500
4	2	0.0135	0.0486	0.0975	0.1536	0.2109	0.2646	0.3105	0.3456	0.3675	0.3750
4	3	0.0005	0.0036	0.0115	0.0256	0.0469	0.0756	0.1115	0.1536	0.2005	0.2500
4	4	0.0000	0.0001	0.0005	0.0016	0.0039	0.0081	0.0150	0.0256	0.0410	0.0625
5	0	0.7738	0.5905	0.4437	0.3277	0.2373	0.1681	0.1160	0.0778	0.0503	0.0313
5	1	0.2036	0.3281	0.3915	0.4096	0.3955	0.3602	0.3124	0.2592	0.2059	0.1563
5	2	0.0214	0.0729	0.1382	0.2048	0.2637	0.3087	0.3364	0.3456	0.3369	0.3125
5	3	0.0011	0.0081	0.0244	0.0512	0.0879	0.1323	0.1811	0.2304	0.2757	0.3125
5	4	0.0000	0.0005	0.0022	0.0064	0.0146	0.0284	0.0488	0.0768	0.1128	0.1563
5	5	0.0000	0.0000	0.0001	0.0003	0.0010	0.0024	0.0053	0.0102	0.0185	0.0313
6	0	0.7351	0.5314	0.3771	0.2621	0.1780	0.1176	0.0754	0.0467	0.0277	0.0156
6	1	0.2321	0.3543	0.3993	0.3932	0.3560	0.3025	0.2437	0.1866	0.1359	0.0938
6	2	0.0305	0.0984	0.1762	0.2458	0.2966	0.3241	0.3280	0.3110	0.2780	0.2344
6	3	0.0021	0.0146	0.0415	0.0819	0.1318	0.1852	0.2355	0.2765	0.3032	0.3125
6	4	0.0001	0.0012	0.0055	0.0154	0.0330	0.0595	0.0951	0.1382	0.1861	0.2344
6	5	0.0000	0.0001	0.0004	0.0015	0.0044	0.0102	0.0205	0.0369	0.0609	0.0938
6	6	0.0000	0.0000	0.0000	0.0001	0.0002	0.0007	0.0018	0.0041	0.0083	0.0156
7	0	0.6983	0.4783	0.3206	0.2097	0.1335	0.0824	0.0490	0.0280	0.0152	0.0078
7	1	0.2573	0.3720	0.3960	0.3670	0.3115	0.2471	0.1848	0.1306	0.0872	0.0547
7	2	0.0406	0.1240	0.2097	0.2753	0.3115	0.3177	0.2985	0.2613	0.2140	0.1641
7	3	0.0036	0.0230	0.0617	0.1147	0.1730	0.2269	0.2679	0.2903	0.2918	0.2734
7	4	0.0002	0.0026	0.0109	0.0287	0.0577	0.0972	0.1442	0.1935	0.2388	0.2734
7	5	0.0000	0.0002	0.0012	0.0043	0.0115	0.0250	0.0466	0.0774	0.1172	0.1641
7	6	0.0000	0.0000	0.0001	0.0004	0.0013	0.0036	0.0084	0.0172	0.0320	0.0547
7	7	0.0000	0.0000	0.0000	0.0000	0.0001	0.0002	0.0006	0.0016	0.0037	0.0078
8	0	0.6634	0.4305	0.2725	0.1678	0.1001	0.0576	0.0319	0.0168	0.0084	0.0039
8	1	0.2793	0.3826	0.3847	0.3355	0.2670	0.1977	0.1373	0.0896	0.0548	0.0313
8	2	0.0515	0.1488	0.2376	0.2936	0.3115	0.2965	0.2587	0.2090	0.1569	0.1094
8	3	0.0054	0.0331	0.0839	0.1468	0.2076	0.2541	0.2786	0.2787	0.2568	0.2188
8	4	0.0004	0.0046	0.0185	0.0459	0.0865	0.1361	0.1875	0.2322	0.2627	0.2734
8	5	0.0000	0.0004	0.0026	0.0092	0.0231	0.0467	0.0808	0.1239	0.1719	0.2188
8	6	0.0000	0.0000	0.0002	0.0011	0.0038	0.0100	0.0217	0.0413	0.0703	0.1094
8	7	0.0000	0.0000	0.0000	0.0001	0.0004	0.0012	0.0033	0.0079	0.0164	0.0313
8	8	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0002	0.0007	0.0017	0.0039

DISTRIBUCIÓN χ^2



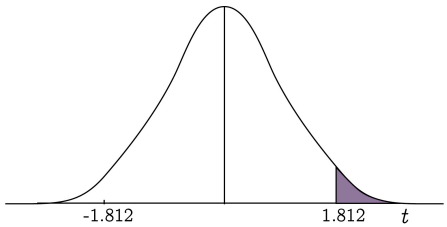
Ejemplo:

Para $\varnothing = 10$ grados de libertad:

$P[\chi^2 > 15.99] = 0.10$

π / ϕ	0.995	0.99	0.975	0.95	0.9	0.75	0.5	0.25	0.1	0.05	0.025	0.01	0.005	π / ϕ
1	3.93E-05	1.57E-04	9.82E-04	3.93E-03	1.58E-02	0.102	0.455	1.323	2.71	3.84	5.02	6.63	7.88	1
2	1.00E-02	2.01E-02	5.06E-02	0.103	0.211	0.575	1.386	2.77	4.61	5.99	7.38	9.21	10.60	2
3	7.17E-02	0.115	0.216	0.352	0.584	1.213	2.37	4.11	6.25	7.81	9.35	11.34	12.84	3
4	0.207	0.297	0.484	0.711	1.064	1.923	3.36	5.39	7.78	9.49	11.14	13.28	14.86	4
5	0.412	0.554	0.831	1.145	1.610	2.67	4.35	6.63	9.24	11.07	12.83	15.09	16.75	5
6	0.676	0.872	1.237	1.635	2.20	3.45	5.35	7.84	10.64	12.59	14.45	16.81	18.55	6
7	0.989	1.239	1.690	2.17	2.83	4.25	6.35	9.04	12.02	14.07	16.01	18.48	20.3	7
8	1.344	1.647	2.18	2.73	3.49	5.07	7.34	10.22	13.36	15.51	17.53	20.1	22.0	8
9	1.735	2.09	2.70	3.33	4.17	5.90	8.34	11.39	14.68	16.92	19.02	21.7	23.6	9
10	2.16	2.56	3.25	3.94	4.87	6.74	9.34	12.55	15.99	18.31	20.5	23.2	25.2	10
11	2.60	3.05	3.82	4.57	5.58	7.58	10.34	13.70	17.28	19.68	21.9	24.7	26.8	11
12	3.07	3.57	4.40	5.23	6.30	8.44	11.34	14.85	18.55	21.0	23.3	26.2	28.3	12
13	3.57	4.11	5.01	5.89	7.04	9.30	12.34	15.98	19.81	22.4	24.7	27.7	29.8	13
14	4.07	4.66	5.63	6.57	7.79	10.17	13.34	17.12	21.1	23.7	26.1	29.1	31.3	14
15	4.60	5.23	6.26	7.26	8.55	11.04	14.34	18.25	22.3	25.0	27.5	30.6	32.8	15
16	5.14	5.81	6.91	7.96	9.31	11.91	15.34	19.37	23.5	26.3	28.8	32.0	34.3	16
17	5.70	6.41	7.56	8.67	10.09	12.79	16.34	20.5	24.8	27.6	30.2	33.4	35.7	17
18	6.26	7.01	8.23	9.39	10.86	13.68	17.34	21.6	26.0	28.9	31.5	34.8	37.2	18
19	6.84	7.63	8.91	10.12	11.65	14.56	18.34	22.7	27.2	30.1	32.9	36.2	38.6	19
20	7.43	8.26	9.59	10.85	12.44	15.45	19.34	23.8	28.4	31.4	34.2	37.6	40.0	20
21	8.03	8.90	10.28	11.59	13.24	16.34	20.3	24.9	29.6	32.7	35.5	38.9	41.4	21
22	8.64	9.54	10.98	12.34	14.04	17.24	21.3	26.0	30.8	33.9	36.8	40.3	42.8	22
23	9.26	10.20	11.69	13.09	14.85	18.14	22.3	27.1	32.0	35.2	38.1	41.6	44.2	23
24	9.89	10.86	12.40	13.85	15.66	19.04	23.3	28.2	33.2	36.4	39.4	43.0	45.6	24
25	10.52	11.52	13.12	14.61	16.47	19.94	24.3	29.3	34.4	37.7	40.6	44.3	46.9	25
26	11.16	12.20	13.84	15.38	17.29	20.8	25.3	30.4	35.6	38.9	41.9	45.6	48.3	26
27	11.81	12.88	14.57	16.15	18.11	21.7	26.3	31.5	36.7	40.1	43.2	47.0	49.6	27
28	12.46	13.56	15.31	16.93	18.94	22.7	27.3	32.6	37.9	41.3	44.5	48.3	51.0	28
29	13.12	14.26	16.05	17.71	19.77	23.6	28.3	33.7	39.1	42.6	45.7	49.6	52.3	29
30	13.79	14.95	16.79	18.49	20.6	24.5	29.3	34.8	40.3	43.8	47.0	50.9	53.7	30
40	20.7	22.2	24.4	26.5	29.1	33.7	39.3	45.6	51.8	55.8	59.3	63.7	66.8	40
50	28.0	29.7	32.4	34.8	37.7	42.9	49.3	56.3	63.2	67.5	71.4	76.2	79.5	50
60	35.5	37.5	40.5	43.2	46.5	52.3	59.3	67.0	74.4	79.1	83.3	88.4	92.0	60
70	43.3	45.4	48.8	51.7	55.3	61.7	69.3	77.6	85.5	90.5	95.0	100.4	104.2	70
80	51.2	53.5	57.2	60.4	64.3	71.1	79.3	88.1	96.6	101.9	106.6	112.3	116.3	80
90	59.2	61.8	65.6	69.1	73.3	80.6	89.3	98.6	107.6	113.1	118.1	124.1	128.3	90
100	67.3	70.1	74.2	77.9	82.4	90.1	99.3	109.1	118.5	124.3	129.6	135.8	140.2	100
Z_α	-2.58	-2.33	-1.96	-1.64	-1.28	-0.674	0.000	0.674	1.282	1.645	1.96	2.33	2.58	Z_α

DISTRIBUCIÓN DE STUDENT



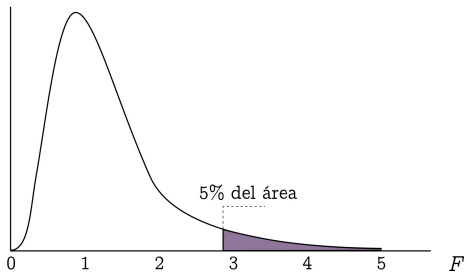
Ejemplo:

Para $\varnothing = 10$ grados de libertad:

$P[\ t > 1.812] = 0.05$
$P[\ t < -1.812] = 0.05$

α / r	0,25	0,2	0,15	0,1	0,05	0,025	0,01	0,005	0,0005
1	1,000	1,376	1,963	3,078	6,314	12,706	31,821	63,656	636,578
2	0,816	1,061	1,386	1,886	2,920	4,303	6,965	9,925	31,600
3	0,765	0,978	1,250	1,638	2,353	3,182	4,541	5,841	12,924
4	0,741	0,941	1,190	1,533	2,132	2,776	3,747	4,604	8,610
5	0,727	0,920	1,156	1,476	2,015	2,571	3,365	4,032	6,869
6	0,718	0,906	1,134	1,440	1,943	2,447	3,143	3,707	5,959
7	0,711	0,896	1,119	1,415	1,895	2,365	2,998	3,499	5,408
8	0,706	0,889	1,108	1,397	1,860	2,306	2,896	3,355	5,041
9	0,703	0,883	1,100	1,383	1,833	2,262	2,821	3,250	4,781
10	0,700	0,879	1,093	1,372	1,812	2,228	2,764	3,169	4,587
11	0,697	0,876	1,088	1,363	1,796	2,201	2,718	3,106	4,437
12	0,695	0,873	1,083	1,356	1,782	2,179	2,681	3,055	4,318
13	0,694	0,870	1,079	1,350	1,771	2,160	2,650	3,012	4,221
14	0,692	0,868	1,076	1,345	1,761	2,145	2,624	2,977	4,140
15	0,691	0,866	1,074	1,341	1,753	2,131	2,602	2,947	4,073
16	0,690	0,865	1,071	1,337	1,746	2,120	2,583	2,921	4,015
17	0,689	0,863	1,069	1,333	1,740	2,110	2,567	2,898	3,965
18	0,688	0,862	1,067	1,330	1,734	2,101	2,552	2,878	3,922
19	0,688	0,861	1,066	1,328	1,729	2,093	2,539	2,861	3,883
20	0,687	0,860	1,064	1,325	1,725	2,086	2,528	2,845	3,850
21	0,686	0,859	1,063	1,323	1,721	2,080	2,518	2,831	3,819
22	0,686	0,858	1,061	1,321	1,717	2,074	2,508	2,819	3,792
23	0,685	0,858	1,060	1,319	1,714	2,069	2,500	2,807	3,768
24	0,685	0,857	1,059	1,318	1,711	2,064	2,492	2,797	3,745
25	0,684	0,856	1,058	1,316	1,708	2,060	2,485	2,787	3,725
26	0,684	0,856	1,058	1,315	1,706	2,056	2,479	2,779	3,707
27	0,684	0,855	1,057	1,314	1,703	2,052	2,473	2,771	3,689
28	0,683	0,855	1,056	1,313	1,701	2,048	2,467	2,763	3,674
29	0,683	0,854	1,055	1,311	1,699	2,045	2,462	2,756	3,660
30	0,683	0,854	1,055	1,310	1,697	2,042	2,457	2,750	3,646
40	0,681	0,851	1,050	1,303	1,684	2,021	2,423	2,704	3,551
60	0,679	0,848	1,045	1,296	1,671	2,000	2,390	2,660	3,460
120	0,677	0,845	1,041	1,289	1,658	1,980	2,358	2,617	3,373
∞	0,674	0,842	1,036	1,282	1,645	1,960	2,326	2,576	3,290

DISTRIBUCIÓN DE FISHER



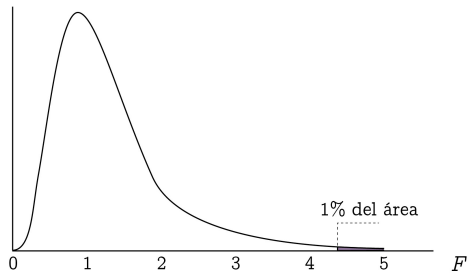
Ejemplo:

Para nl = 9,n2 = 12 grados de libertad:

P[$F > 2.80$] = 0.05

Grados de libertad del denominador	Grados de libertad del numerador										
	1	2	3	4	5	6	7	8	9	10	12
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	243.91
2	18.513	19.000	19.164	19.247	19.296	19.330	19.353	19.371	19.385	19.396	19.413
3	10.128	9.5521	9.2766	9.1172	9.0135	8.9406	8.8867	8.8452	8.8123	8.7855	8.7446
4	7.7086	6.9443	6.5914	6.3882	6.2561	6.1631	6.0942	6.0410	5.9988	5.9644	5.9117
5	6.6079	5.7861	5.4095	5.1922	5.0503	4.9503	4.8759	4.8183	4.7725	4.7351	4.6777
6	5.9874	5.1433	4.7571	4.5337	4.3874	4.2839	4.2067	4.1468	4.0990	4.0600	3.9999
7	5.5914	4.7374	4.3468	4.1203	3.9715	3.8660	3.7870	3.7257	3.6767	3.6365	3.5747
8	5.3177	4.4590	4.0662	3.8379	3.6875	3.5806	3.5005	3.4381	3.3881	3.3472	3.2839
9	5.1174	4.2565	3.8625	3.6331	3.4817	3.3738	3.2927	3.2296	3.1789	3.1373	3.0729
10	4.9646	4.1028	3.7083	3.4780	3.3258	3.2172	3.1355	3.0717	3.0204	2.9782	2.9130
11	4.8443	3.9823	3.5874	3.3567	3.2039	3.0946	3.0123	2.9480	2.8962	2.8536	2.7876
12	4.7472	3.8853	3.4903	3.2592	3.1059	2.9961	2.9134	2.8486	2.7964	2.7534	2.6866
13	4.6672	3.8056	3.4105	3.1791	3.0254	2.9153	2.8321	2.7669	2.7144	2.6710	2.6037
14	4.6001	3.7389	3.3439	3.1122	2.9582	2.8477	2.7642	2.6987	2.6458	2.6022	2.5342
15	4.5431	3.6823	3.2874	3.0556	2.9013	2.7905	2.7066	2.6408	2.5876	2.5437	2.4753
16	4.4940	3.6337	3.2389	3.0069	2.8524	2.7413	2.6572	2.5911	2.5377	2.4935	2.4247
17	4.4513	3.5915	3.1968	2.9647	2.8100	2.6987	2.6143	2.5480	2.4943	2.4499	2.3807
18	4.4139	3.5546	3.1599	2.9277	2.7729	2.6613	2.5767	2.5102	2.4563	2.4117	2.3421
19	4.3807	3.5219	3.1274	2.8951	2.7401	2.6283	2.5435	2.4768	2.4227	2.3779	2.3080
20	4.3512	3.4928	3.0984	2.8661	2.7109	2.5990	2.5140	2.4471	2.3928	2.3479	2.2776
21	4.3248	3.4668	3.0725	2.8401	2.6848	2.5727	2.4876	2.4205	2.3660	2.3210	2.2504
22	4.3009	3.4434	3.0491	2.8167	2.6613	2.5491	2.4638	2.3965	2.3419	2.2967	2.2258
23	4.2793	3.4221	3.0280	2.7955	2.6400	2.5277	2.4422	2.3748	2.3201	2.2747	2.2036
24	4.2597	3.4028	3.0088	2.7763	2.6207	2.5082	2.4226	2.3551	2.3002	2.2547	2.1834
25	4.2417	3.3852	2.9912	2.7587	2.6030	2.4904	2.4047	2.3371	2.2821	2.2365	2.1649
26	4.2252	3.3690	2.9752	2.7426	2.5868	2.4741	2.3883	2.3205	2.2655	2.2197	2.1479
27	4.2100	3.3541	2.9604	2.7278	2.5719	2.4591	2.3732	2.3053	2.2501	2.2043	2.1323
28	4.1960	3.3404	2.9467	2.7141	2.5581	2.4453	2.3593	2.2913	2.2360	2.1900	2.1179
29	4.1830	3.3277	2.9340	2.7014	2.5454	2.4324	2.3463	2.2783	2.2229	2.1768	2.1045
30	4.1709	3.3158	2.9223	2.6896	2.5336	2.4205	2.3343	2.2662	2.2107	2.1646	2.0921
40	4.0847	3.2317	2.8387	2.6060	2.4495	2.3359	2.2490	2.1802	2.1240	2.0772	2.0035
50	4.0343	3.1826	2.7900	2.5572	2.4004	2.2864	2.1992	2.1299	2.0734	2.0261	1.9515
60	4.0012	3.1504	2.7581	2.5252	2.3683	2.2541	2.1665	2.0970	2.0401	1.9926	1.9174
120	3.9201	3.0718	2.6802	2.4472	2.2899	2.1750	2.0868	2.0164	1.9588	1.9105	1.8337
∞	3.8841	3.0369	2.6456	2.4127	2.2551	2.1400	2.0514	1.9807	1.9226	1.8739	1.7964

DISTRIBUCIÓN DE FISHER



Ejemplo:

Para nl = 9,n2 = 12 grados de libertad:

P[$F > 4.39$] = 0.01

Grados de libertad del denominador	Grados de libertad del numerador										
	1	2	3	4	5	6	7	8	9	10	12
1	4052.2	4999.5	5403.4	5624.6	5763.6	5859.0	5928.4	5981.1	6022.5	6055.8	6106.3
2	98.503	99.000	99.166	99.249	99.299	99.333	99.356	99.374	99.388	99.399	99.416
3	34.116	30.817	29.457	28.710	28.237	27.911	27.672	27.489	27.345	27.229	27.052
4	21.198	18.000	16.694	15.977	15.522	15.207	14.976	14.799	14.659	14.546	14.374
5	16.258	13.274	12.060	11.392	10.967	10.672	10.456	10.289	10.158	10.051	9.8883
6	13.745	10.925	9.7795	9.1483	8.7459	8.4661	8.2600	8.1017	7.9761	7.8741	7.7183
7	12.246	9.5466	8.4513	7.8466	7.4604	7.1914	6.9928	6.8400	6.7188	6.6201	6.4691
8	11.259	8.6491	7.5910	7.0061	6.6318	6.3707	6.1776	6.0289	5.9106	5.8143	5.6667
9	10.561	8.0215	6.9919	6.4221	6.0569	5.8018	5.6129	5.4671	5.3511	5.2565	5.1114
10	10.044	7.5594	6.5523	5.9943	5.6363	5.3858	5.2001	5.0567	4.9424	4.8491	4.7059
11	9.6460	7.2057	6.2167	5.6683	5.3160	5.0692	4.8861	4.7445	4.6315	4.5393	4.3974
12	9.3302	6.9266	5.9525	5.4120	5.0643	4.8206	4.6395	4.4994	4.3875	4.2961	4.1553
13	9.0738	6.7010	5.7394	5.2053	4.8616	4.6204	4.4410	4.3021	4.1911	4.1003	3.9603
14	8.8616	6.5149	5.5639	5.0354	4.6950	4.4558	4.2779	4.1399	4.0297	3.9394	3.8001
15	8.6831	6.3589	5.4170	4.8932	4.5556	4.3183	4.1415	4.0045	3.8948	3.8049	3.6662
16	8.5310	6.2262	5.2922	4.7726	4.4374	4.2016	4.0259	3.8896	3.7804	3.6909	3.5527
17	8.3997	6.1121	5.1850	4.6690	4.3359	4.1015	3.9267	3.7910	3.6822	3.5931	3.4552
18	8.2854	6.0129	5.0919	4.5790	4.2479	4.0146	3.8406	3.7054	3.5971	3.5082	3.3706
19	8.1849	5.9259	5.0103	4.5003	4.1708	3.9386	3.7653	3.6305	3.5225	3.4338	3.2965
20	8.0960	5.8489	4.9382	4.4307	4.1027	3.8714	3.6987	3.5644	3.4567	3.3682	3.2311
21	8.0166	5.7804	4.8740	4.3688	4.0421	3.8117	3.6396	3.5056	3.3981	3.3098	3.1730
22	7.9454	5.7190	4.8166	4.3134	3.9880	3.7583	3.5867	3.4530	3.3458	3.2576	3.1209
23	7.8811	5.6637	4.7649	4.2636	3.9392	3.7102	3.5390	3.4057	3.2986	3.2106	3.0740
24	7.8229	5.6136	4.7181	4.2184	3.8951	3.6667	3.4959	3.3629	3.2560	3.1681	3.0316
25	7.7698	5.5680	4.6755	4.1774	3.8550	3.6272	3.4568	3.3239	3.2172	3.1294	2.9931
26	7.7213	5.5263	4.6366	4.1400	3.8183	3.5911	3.4210	3.2884	3.1818	3.0941	2.9578
27	7.6767	5.4881	4.6009	4.1056	3.7848	3.5580	3.3882	3.2558	3.1494	3.0618	2.9256
28	7.6356	5.4529	4.5681	4.0740	3.7539	3.5276	3.3581	3.2259	3.1195	3.0320	2.8959
29	7.5977	5.4204	4.5378	4.0449	3.7254	3.4995	3.3303	3.1982	3.0920	3.0045	2.8685
30	7.5625	5.3903	4.5097	4.0179	3.6990	3.4735	3.3045	3.1726	3.0665	2.9791	2.8431
40	7.3141	5.1785	4.3126	3.8283	3.5138	3.2910	3.1238	2.9930	2.8876	2.8005	2.6648
50	7.1706	5.0566	4.1993	3.7195	3.4077	3.1864	3.0202	2.8900	2.7850	2.6981	2.5625
60	7.0771	4.9774	4.1259	3.6490	3.3389	3.1187	2.9530	2.8233	2.7185	2.6318	2.4961
120	6.8509	4.7865	3.9491	3.4795	3.1735	2.9559	2.7918	2.6629	2.5586	2.4721	2.3363
∞	6.7515	4.7029	3.8719	3.4055	3.1014	2.8848	2.7214	2.5929	2.4888	2.4023	2.2664

Bibliografía

[1] Aliaga M., Gunderson B., *Interactive Statistics*, Prentice Hall, 2002.
[2] Batanero C., Godino J., *Análisis de datos y su didáctica*, Universidad de Granada, 2001
[3] Batanero C., *Didáctica de la Estadística*, Universidad de Granada, 2001.
[4] Brook R. et al., *The Fascination of Statistics*, Marcel Dekker, 1986.
[5] Cuesta M., Herrero F., *http://www.psico.uniovi.es/Dpto_Psicologia/metodos/tutor.1/indice.html*, Departamento de Psicología, Universidad de Oviedo.
[6] Gil O., *Excursiones por el Álgebra Lineal*, J.C. Sáez Editor, Santiago, 2011.
[7] Lacourly N., *Introducción a la Estadística*, J.C. Sáez Editor, Santiago, 2011.
[8] Lladser M., *Variables Aleatorias y Simulación Estocástica*, Editorial JC. Sáez, Santiago, 2011.
[9] Moore D., McCabe G., *Introduction to the Practice of Statistics*, (3rd Ed) W H Freeman & Co, 1998.
[10] Naiman A., R. Rosenberg R. & Zirkel G., *Understanding Statistics*, Mc Graw-Hill, 1996.
[11] Newman J., *The World of Mathematics*, Simon & Schuster, New York, 1956.
[12] Osses A., *Análisis numérico*, J.C. Sáez Editor, Santiago, 2011.
[13] Pearson K., *On Lines and Planes of Closest Fit to Systems of Points in Space*, Philosophical Magazine 2 (6): 559 - 572, 1901. *http://stat.smmu.edu.cn/history/pearson1901.pdf*.
[14] Romagnoli P., *Probabilidades Doctas con discos y bolitas*, J,C. Sáez Editor, Santiago, 2011.
[15] Ycart B., Curso por Internet, *http://ljk.imag.fr/membres/Bernard.Ycart/emel/index.html*.
[16] Yule G. U. *An Introduction to the theory of statistics*, London, C. Griffin, 1922.

Índice de figuras

Índice de nombres propios

Índice de palabras