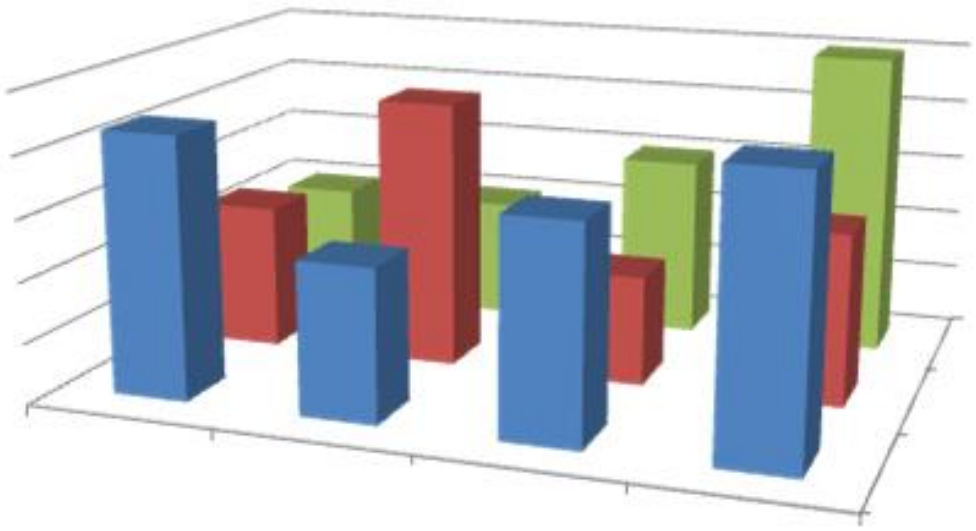


ESTADÍSTICA
EN CIENCIAS DE LA SALUD

Autores: Martínez-Beneito, M A; Botella-Rocamora, P; Alacreu-García, M.

Título: Estadística en Ciencias de la Salud.

Edita: Los autores

ISBN: 978-84-09-49335-7

Valencia, 2023

Estadística en Ciencias de la Salud

Martínez-Beneito, M. A[1]., Botella-Rocamora, P[2]., Alacreu-García, M[3].

[1] Departamento de Estadística e Investigación Operativa. Universidad de Valencia.
[2] Dirección General de Salud Pública. Conselleria de Sanitat Universal i Salut Pública. Generalitat Valenciana.
[3] Departamento de Matemáticas, Física y Ciencias Tecnológicas. Universidad CEU Cardenal Herrera.

Índice general

Capítulo 1

Estadística descriptiva

1.1. ¿Qué es la Bioestadística?

Concepto de Bioestadística

Se entiende como *bioestadística* la aplicación de técnicas estadísticas a las ciencias de la naturaleza, entre las que se encuentran todas las ciencias de la salud. Para que esta definición tenga sentido habremos de entender plenamente qué es la estadística. Podemos encontrar múltiples definiciones de estadística en la literatura, sin embargo encontramos particularmente adecuada para los objetivos que se van a acometer en este curso la siguiente:

La *Estadística* estudia los métodos y procedimientos para recoger, clasificar, resumir, hallar regularidades y analizar los datos, siempre y cuando la variabilidad e incertidumbre sean una causa intrínseca de los mismos; así como de realizar inferencias a partir de ellos, con la finalidad de ayudar a la toma de decisiones y en su caso formular predicciones.

¿Para qué sirve la estadística en el ámbito de las Ciencias de la Salud?

La estadística tiene una gran utilidad para vuestra formación en cualquier titulación en el ámbito de las Ciencias de la Salud. El transmitiros esa utilidad es el principal objetivo del presente texto, no obstante nos gustaría destacar los siguientes motivos por los que encontramos particularmente útil la estadística en vuestra formación.

- **La estadística os va a servir de ayuda para el resto de asignaturas de vuestra carrera.** En distintas ocasiones encontraréis que algunos conceptos de otras materias que debéis superar en vuestra formación están basados en conceptos estadísticos, por tanto la estadística os ayudará a completar y entender mejor ciertos aspectos de vuestra formación.

- **La estadística será una herramienta en el futuro ejercicio de vuestra profesión.** Al igual que en la vida real en el ejercicio de vuestra profesión encontraréis distintos hallazgos, procedimientos y conceptos basados en análisis estadísticos previos. La simple interpretación de un análisis de sangre requiere una madurez estadística suficiente para poder ser interpretado de forma adecuada.

- **La estadística os abre la puerta a la literatura científica.** Todas las disciplinas en las que se realizan estudios cuantitativos han de justificar sus hallazgos en términos estadísticos, es más, la validez de sus afirmaciones dependen de que la estadística lo juzgue como tal. Este es el motivo de que la literatura científica, y en concreto la relacionada con las ciencias de la salud, esté plagado de conceptos y términos estadísticos que intentaremos transmitiros a lo largo de este texto.

- **Supone una herramienta para el análisis de situaciones con componente aleatoria.** La estadística es la ciencia que trabaja y cuantifica la incertidumbre. En aquellas situaciones en las que el resultado de un procedimiento es incierto, lo que suele ser bastante más habitual de lo que podríamos pensar, la estadística se muestra como una herramienta imprescindible para tomar decisiones basadas en información objetiva y que ofrezcan garantías de ser adecuadas.

¿Son las Ciencias de la Salud ciencias exactas?

Tal y como podéis suponer las ciencias de la salud no son, para nada, ciencias exactas. Ni la respuesta a tratamientos idénticos por parte de distintos pacientes son siempre iguales, ni los tratamientos que se habrán de administrar a pacientes con la misma enfermedad han de ser necesariamente los mismos, o incluso existen pacientes que presentarán efectos secundarios a ciertos medicamentos y habrá otros que no. Es por ello que estas ciencias necesitan de la estadística como guía y sustento dada la aleatoriedad a la que están sujetos la mayoría de sus procesos. Además, la estadística supone una herramienta de incalculable valor para las ciencias de la salud a la hora de establecer protocolos para determinados procedimientos ya que es capaz de cuantificar la conveniencia de los resultados de distintas alternativas, por ejemplo de distintos tratamientos, y así poder tomar la mejor decisión de forma fundamentada.

1.2. Datos

Los *datos* son todas aquellas unidades de información relevantes a la hora de hacer un estudio estadístico, constituyen la materia prima de trabajo de la bioestadística. En concreto la labor de la bioestadística será transformar los datos que disponemos en información útil para el propósito de nuestra investigación. Los datos para un estudio estadístico vienen recogidos como variables sobre unidades experimentales. Las *unidades experimentales* (muestra) son todos aquellos individuos que albergan información sobre el objeto de interés de nuestro estudio y que por ello son incluidos en éste. En la definición anterior entendemos el concepto de individuo de forma bastante amplia, de forma que pueden ser individuos para un estudio bien personas, o grupos de personas como por ejemplo municipios, los trabajadores de cierta empresa..., o grupos que no estén formados necesariamente por personas un conjunto de muestras serológicas. A su vez las *variables* son todas aquellas características que resultan de interés de las unidades experimentales y que se incluyen en el estudio estadístico para su análisis. Los datos de un estudio estadístico proceden de la respuesta que tienen unas unidades experimentales sobre una característica de interés o variable.

Ejemplo 1.1.

Se desea realizar un estudio sobre hipertensión arterial en población anciana. Queremos estudiar este problema y qué características de los pacientes pueden tener relación o no con él. Identifica las unidades experimentales del estudio y las variables de interés.

Las unidades experimentales serían todos aquellos ancianos integrantes del estudio. Las variables de nuestro estudio serían: la presión arterial de los ancianos que es la variable de interés sobre la que queremos aprender y otras variables que desearíamos conocer si están relacionadas o no con la hipertensión como edad, sexo, consumo de calorías diarias,...

Persona	**Presión Art.**	**Edad**	**Sexo**	**Cons.calorías**	**...**
...	...	...	...	...	...
...	...	...	...	...	...
...	...	...	...	...	...

El primer paso en todo estudio será la planificación de cómo han de ser recogidos los datos de forma que las conclusiones que se puedan extraer de ellos sean "válidas" y por tanto conduzcan a conclusiones adaptadas a la realidad. En el capítulo 3 estudiaremos con más detalle como se ha de realizar este proceso y cuales son las precauciones que hemos de seguir para que las conclusiones que obtengamos de nuestros estudios sean correctas.

Cómo se organizan los datos: variables y unidades experimentales

Normalmente, en estadística aplicada a ciencias de la salud, los datos disponibles se refieren a personas, es decir, disponemos de variables medidas en personas, que serían en este caso las unidades experimentales. Por ejemplo, en el registro de admisión de un hospital se puede tomar para cada uno de los pacientes atendidos (los pacientes serán las unidades experimentales) datos sobre su edad, sexo, motivo de ingreso, municipio de residencia, ... Cada uno de estos datos de interés se recogerían como una variable distinta. La forma en la que se suelen almacenar los datos es mediante una tabla en la que la información de cada individuo se recoge en una fila, mientras que cada característica de cada unidad experimental se representa en una columna (es decir, las *variables* se presentan en columnas y las *unidades experimentales* en filas).

La información medida para cada unidad experimental, lo que llamamos variables, está sujeta a variabilidad y rodeada de incertidumbre. Si pensamos en el color de ojos de una persona, su altura, su tensión arterial,...varía de un individuo a otro. Por este hecho, nos solemos referir a las variables como *Variables Aleatorias*, ya que somos incapaces de predecir qué valores tomará cada individuo, al menos antes de realizar cualquier análisis estadístico.

Ejemplo 1.2.

Representa mediante una tabla cómo se almacenaría para su análisis estadístico la información del registro de admisión de un hospital.

La información tal y como se ha comentado vendrá organizada en una tabla en la que los individuos se corresponden con sus filas y las variables con las columnas, más o menos de la siguiente forma:

Identif.	**Edad**	**Sexo**	**Municipio**	**Motivo**	**...**
000001	29	H	Valencia	Migraña	...
000002	68	M	Sagunto	Desmayo	...
...	...	...	...	...	...

El primer paso para analizar estadísticamente unos datos es el resumen de los mismos y su representación gráfica. Mediante esta visualización nos podremos hacer una pequeña idea de cómo son los datos que manejamos y podremos detectar posibles errores que suelen pasar desapercibidos si prescindimos de esta representación. Esta parte preliminar del análisis se conoce como *estadística descriptiva* a diferencia de la *estadística inferencial* que es la encargada de estudiar determinadas características de interés de la población de estudio partir de los datos.

En este tema desarrollaremos principalmente la estadística descriptiva, es decir aprenderemos a resumir y representar gráficamente cada una de las variables que aparecen en una tabla de datos. La forma de resumir y representarlas depende del tipo de variable, así que en primer lugar estudiaremos los tipos de variables que existen, y a continuación estableceremos las herramientas adecuadas para el resumen de las mismas.

Tipos de variables aleatorias

Tal y como se ha comentado antes de llevar a cabo el análisis descriptivo de los datos se ha de tener claro de qué tipo es cada una de las variables de que disponemos. Así, podemos clasificar la variables según el siguiente criterio:

- *Variables cuantitativas*: Son aquellas que responden a la pregunta ¿cuánto?, y pueden ser expresadas numéricamente (es decir, siempre tomarán un va-lor numérico). A su vez se dividen en:
  - *Variables continuas*: Podrán tomar cualquier valor (entero o no) dentro de un rango determinado de valores.
  - *Variables discretas*: Sólo podrán tomar ciertos valores concretos (habitualmente números enteros).
- *Variables cualitativas o categóricas*: Responden a la pregunta ¿de qué tipo? Pueden tomar cualquier valor, numérico o de cualquier otro tipo. Cada uno de los posibles valores que puede tomar estos tipos de variables de dicen *Categorías*. Las variables cualitativas a su vez se dividen en:
  - *Variables ordinales*: Serán aquellas variables de tipo cualitativo en el que las posibles respuestas admiten una ordenación lógica.
  - *Variables nominales*: Serán aquellas variables de tipo cualitativo en el que las posibles respuestas NO admiten ningún tipo de ordenación lógica.

Ejemplo 1.3.

Clasifica las siguientes variables según los criterios anteriores:
Color de los ojos de una muestra de alumnos

Marrón	***Verde***	***Verde***	***Azul***	***Marrón***	***Azul***
Marrón	***Marrón***	***Azul***	***Marrón***	***Azul***	***Marrón***
Verde	***Marrón***	***Verde***	***Marrón***	***Azul***	***Verde***
Marrón	***Marrón***	***Azul***	***Marrón***	***Verde***	***Marrón***
Verde	***Verde***	***Marrón***	***Marrón***	***Marrón***	***Marrón***
Azul					

Estatura de alumnos de la muestra anterior

1.48	***1.56***	***1.56***	***1.59***	***1.60***	***1.61***	***1.63***	***1.64***	***1.64***
1.67	***1.68***	***1.68***	***1.68***	***1.68***	***1.69***	***1.70***	***1.71***	***1.72***
1.72	***1.75***	***1.76***	***1.76***	***1.77***	***1.77***	***1.79***	***1.81***	***1.81***
1.84	***1.84***	***1.88***	***1.94***					

En cuanto a la variable del color de ojos resulta obvio que no es de tipo cuantitativo, además no parece lógico establecer ningún tipo de orden en las categorías que componen las posibles respuestas de esta variable (a diferencia, por ejemplo de una variable cuyas posibles respuestas fueran: alto, medio y bajo), por tanto esta variable es de tipo *Categórica nominal*.
Respecto a la estatura de los estudiantes, tal y como se nos presenta es una variable de tipo cuantitativo. Además, como la estatura de cualquier persona puede tomar cualquier valor (independientemente de que pueda ser redondeado) esta variable es de tipo *Cuantitativa continua*.

Ejemplo 1.4.

Clasifica las siguientes variables según el criterio que acabamos de introducir: Gravedad de un infarto (leve, moderado, fuerte), Número de ataques de asma semanales, Sexo, Presión arterial, Estatura, Peso, Estado de dolor tras la toma de un fármaco (Peor, Igual, Mejor), Provincia, Edad, Número de preguntas acertadas en un test, Grupo sanguíneo.

La clasificación correcta de las variables sería la siguiente:

Cuantitativas	
Continuas	Discretas
Presión arterial Estatura Peso Edad	Número de ataques de asma semanales Número de preguntas acertadas en un test
Cualitativas	
Ordinales	Nominales
Gravedad de un infarto (leve, moderado, fuerte) Estado de dolor tras la toma de un fármaco (Peor, Igual, Mejor)	Sexo Provincia Grupo sanguíneo

1.3. Estadística descriptiva univariante: variables cualitativas

La estadística descriptiva resume un conjunto de datos proporcionando información mediante tablas, parámetros y/o gráficos. En cualquier análisis estadístico, la estadística descriptiva es la primera parte y más importante, pues permite conocer el comportamiento de las variables, consideradas una a una, o la posible relación existente entre ellas. En esta sección nos centraremos en el análisis univariante de las variables, es decir, el estudio individual de cada una de éstas. Tal y como se introdujo en la sección anterior el análisis descriptivo de las variables incluidas en el estudio dependerá del tipo de variables que que-ramos resumir, por tanto vamos a dividir los métodos descriptivos en función de este criterio.

1.3.1. Frecuencias absolutas y relativas

Podremos resumir individualmente variables de tipo cualitativo mediante las frecuencias absolutas y relativas de sus categorías.

- **Frecuencias absolutas**. Se definen las *frecuencias absolutas* (f_a) de una variable cualitativa como el número de ocasiones en las que se ha dado cada una de las categorías de la variable que queramos resumir.

- **Frecuencias relativas**. Por otro lado las *frecuencias relativas* (f_r) se definen como la proporción de veces que se ha dado cada uno de las categorías de la variable. Por tanto las frecuencias absolutas y relativas de una variable cumplen la siguiente relación:

$$f_r = \frac{f_a}{Número\ de\ unidades\ experimentales}$$

Ejemplo 1.5.

Calcula las frecuencias absolutas y relativas de la variable de color de ojos del ejemplo 1.3.

Las frecuencias absolutas no son más que el número de veces que se ha dado cada una de las posibles respuestas de la variable, es decir cada una de éstas tendrá asociada una frecuencia, mientras que para el cálculo de las frecuencias relativas habremos de dividir estos valores por el número total de unidades experimentales (31), por tanto tenemos:

Valor	f_a	f_r	%
Azul	7	0.226	22.6
Verde	8	0.258	25.8
Marrón	16	0.516	51.6

1.3.2. Representación gráfica

En cuanto a la representación gráfica de las variables cualitativas destacamos dos tipos de gráfico por ser los que se utilizan con mayor frecuencia.

- **Diagrama de sectores**. El primero de ellos, el *diagrama de sectores*, se utiliza para visualizar de forma sencilla las frecuencias relativas de las variables. En los gráficos de sectores se divide una figura, habitualmente de forma circular, de forma que el área correspondiente a

cada posible respuesta de la variable será proporcional a la frecuencia relativa de la variable. Esta representación se puede adornar de etiquetas en el interior o exterior del gráfico, además suele ser habitual incluir para cada categoría de la variable la frecuencia relativa (o si se desea absoluta de la variable). En cualquier caso todos estos adornos de la representación, así como otros detalles (color, forma del gráfico,...) son complementos que facilitan la visualización de los resultados y que los programas habituales de estadística suelen incorporar. La elección de como se ha de personalizar este tipo de gráficos es una decisión personal en función del detalle que se desea que incluya la representación final.

Ejemplo 1.6.

Representa mediante un diagrama de sectores la variable de color de ojos del ejemplo 1.3.

Para calcular los grados que les corresponde a cada porción del gráfico, hay que plantear reglas tres:

$100\,\%$ —— 360^{o}
$22.6\,\%$ —— Azul $\quad \rightarrow Azul = 81{,}36$

$100\,\%$ —— 360^{o}
$25.8\,\%$ —— Verde $\quad \rightarrow Verde = 92{,}88$

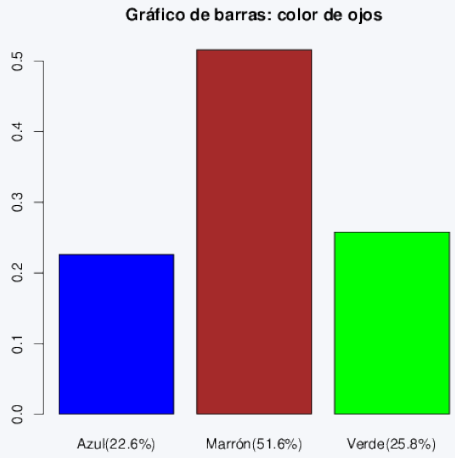
$100\,\%$ —— 360^{o}
$51.6\,\%$ —— Marrn $\quad \rightarrow Marrn = 185{,}76$

- **Gráfico de barras**. El segundo tipo de representaciones gráficas que vamos a contemplar son los *gráficos de barras*. En este tipo de gráfico se representa una barra vertical (u horizontal si

se desea) para cada una de las categorías de la variable de altura proporcional a su frecuencia, bien absoluta o relativa. Al igual que los diagramas de sectores los gráficos de barras se suelen personalizar al gusto del usuario de forma que su configuración resulte lo más ilustrativa posible. Los gráficos de barras suelen ser preferibles a los diagramas de sectores ya que según se ha podido comprobar el ojo humano esta particularmente entrenado para comparar longitudes y no para comparar áreas, sin embargo dada la popularidad de estos últimos en la literatura conviene conocer su interpretación y ser conscientes de su posible uso.

Ejemplo 1.7.

Representa mediante un gráfico de barras la variable de color de ojos del ejemplo 1.3.



1.4. Estadística descriptiva univariante: variables cuantitativas

Para resumir variables de tipo cuantitativo tenemos un abanico de herra-mientas bastante más amplio que para el caso cualitativo. Podemos tabular los datos en tablas de frecuencias, o bien calcular medidas de resumen específicas de este tipo de variables, que se pueden clasificar a grandes rasgos de la siguiente forma:

- *Medidas de centralización*: Resumen la localización alrededor de la cual se distribuyen los datos. Durante este curso introduciremos la media, moda y mediana.

- *Medidas de orden o posición (localización)*: Informan sobre distintas características de los datos a partir de la ordenación de los valores observados. Las medidas de orden que estudiaremos son los percentiles y cuartiles.

- *Medidas de dispersión*: Resumen la variabilidad que presentan los datos alrededor de alguno de los estadísticos de centralización. Estudiaremos como medidas de dispersión el rango, rango intercuartílico, varianza y desviación típica.

- *Medidas de forma*: Informan sobre el comportamiento de la distribución de los datos (simetría, uni/multimodalidad,...). Las comentaremos únicamente a nivel gráfico en este curso.

1.4.1. Tabulación de variables cuantitativas

Otra forma de resumir las variables cuantitativas que ayuda a comprender su comportamiento es su representación mediante una *tabla de frecuencias*. Para ello, a partir del rango de valores de la variable que queramos estudiar, se crea una división adecuada en intervalos más pequeños y se resume la cantidad de datos que se han observado en cada uno de esos intervalos. Sobre cuántos intervalos es necesario hacer para resumir un conjunto de datos, no hay una respuesta cerrada, se aconseja construir entre 5 y 15 intervalos aproximadamente, dependiendo de la cantidad de datos disponible. Para cada uno de esos intervalos, se calcula cuántos valores hay en cada uno de ellos (*frecuencias absolutas*) y qué proporción sobre el total de los datos implica esa cantidad de valores (*frecuencias relativas*). Esta representación mediante una tabla de frecuencias ayuda a visualizar el comportamiento de la variable (qué valores son más frecuentes y cuáles lo son menos) a lo largo de todo su rango de valores observados.

A los intervalos en los que se divide la variable se le llaman *clases* y el número de los mismos se denomina *número de clases*.

Ejemplo 1.8.

Resume mediante una tabla de frecuencias los valores de la variable de estaturas del Ejemplo 1.3

Clases	$\mathbf{f_a}$	$\mathbf{f_r}$	%
[1,4, 1,5)	1	0,032	3,2
[1,5, 1,6)	3	0,097	9,7
[1,6, 1,7)	11	0,355	35,5
[1,7, 1,8)	10	0,323	32,3
[1,8, 1,9)	5	0,161	16,1
[1,9, 2,0)	1	0,032	3,2

En el caso de variables cuantitativas discretas, como el número de valores que puede tomar la variable es limitado, se puede considerar cada valor como una clase (como si se tratara de una variable cualitativa).

Ejemplo 1.9.

Resumen mediante una tabla de frecuencias los valores de la variable cuantitativa discreta: Número de hijos para una muestra de 60 familias.

Número de hijos	f_a	f_r	%
0	4	0,067	6,7
1	12	0,200	20,0
2	28	0,467	46,7
3	10	0,167	16,7
4	5	0,083	8,3
5	1	0,017	1,7

1.4.2. Medidas de centralización

Las medidas de centralización nos informan sobre la localización alrededor de la que se encuentran los valores de la variable en estudio. Hay diferentes estadísticos que nos informan sobre este valor, entre los que destacamos:

- **Moda**. La *moda* de una variable será aquel valor que se repita un mayor número de veces. Cuando la variable que queramos estudiar apenas tome valores repetidos este estadístico será de poca utilidad (cuando la variable en estudio sea cuantitativa continua, se suele hablar del intervalo o rango que más valores contiene como la moda. Esta idea se estudiará en la tabulación de variables cuantitativas).

- **Media**. Supongamos que tenemos una variable cuantitativa a la que llamamos X y tenemos recogidos n valores de esta variable que denotamos con $x_1, x_2, ..., x_n$. La *media* de estos valores se representa $\overline{x}$ y se calcula mediante la fórmula:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + ... + x_n}{n}$$

- **Mediana**. La *mediana* es el valor que cumple que la mitad de los valores de la variable son inferiores a él y la otra mitad son superiores. Si el número de datos en la muestra es impar será el valor central de la muestra ordenada (muestra en la que las unidades experimentales aparecen ordenadas según el valor que toman). Si el número de datos es par la mediana se define como la media de los dos valores centrales de la muestra ordenada.

Observaciones:

- La media es muy sensible a la existencia de valores extremos de la variable (particularmente altos o bajos): ya que todas las observaciones intervienen en el cálculo de la media, la aparición de una observación extrema, hará que la media se desplace en esa dirección.

- Si consideramos una variable discreta, por ejemplo, el número de hijos en las familias de la ciudad de Valencia el valor de la media puede no pertenecer al conjunto de valores posibles de la variable; Por ejemplo $\overline{x} = 2{,}5$ hijos por familia.

Ejemplo 1.10.

Calcula la estatura media de los estudiantes del ejemplo 1.3.

En este caso la variable X en la que estamos interesados será la estatura de los estudiantes en nuestra muestra. En ese caso tenemos $\{x_1 = 1{,}48, x_2 = 1{,}56, ..., x_{31} = 1{,}94\}$ y n (el número de elementos en la muestra) será 31. Así, la media de la variable anterior vendrá dada por la siguiente expresión:

$$\bar{x} = \frac{x_1 + x_2 + ... + x_{31}}{31} = \frac{1{,}48 + 1{,}56 + ... + 1{,}94}{31} = 1{,}708$$

Por tanto la estatura media en nuestra muestra será 1.708 metros de altura.

1.4.3. Medidas de orden o posición (localización)

Estas medidas indican, como refleja su nombre, el orden o posición de una observación entre los valores de una variable cuantitativa. Para el cálculo de estas medidas debemos ordenar de forma ascendente los valores de la muestra, al resultado de dicha reubicación de los valores se le conoce como *muestra ordenada*.

- **Mínimo**: El *mínimo* es el valor menor.
- **Máximo**: El *máximo* es el valor mayor.
- **Percentil al p %**. El *percentil al p %* es el valor que cumple que el p % de las observaciones de la muestra son inferiores a él (y por tanto el resto son superiores a él). Para su cálculo deberíamos hallar la posición que ocupa dicho valor en la muestra ordenada. Dicha posición la podemos calcular mediante la siguiente expresión:

  $$Pos = (n + 1) \cdot \frac{p}{100}$$

  Si la posición Pos resulta un número entero, indica que el valor en esa posición de la muestra ordenada es el percentil buscado. Si, en cambio, la posición resulta un número con decimales, el percentil se calculará utilizando por un lado su parte entera ($[Pos]$), y por otra su parte decimal ($deci(Pos)$) mediante la fórmula:

  $$deci(Pos) \cdot X_{[Pos]+1} + (1 - deci(Pos)) \cdot X_{[Pos]}$$

  donde el término X_j en la expresión anterior se refiere al j-ésimo término de la muestra de valores ordenados. Es decir, combinaremos los valores de las observaciones $[Pos]$ y $[Pos] + 1$ de la muestra ordenada en función de si la parte decimal de Pos está más cercana a uno de estos dos valores que al otro.

  Los percentiles al 25 %, 50 % y 75 % reciben nombres concretos dada su importancia (y se denotan por $P_{25} = Q_1$, $P_{50} = Q_2$ y $P_{75} = Q_3$). A estos percentiles se les dice *cuartiles* (primer, segundo y tercer cuartil respectivamente) ya que dividen la muestra en cuatro partes de igual tamaño. Si pensamos con calma nos podremos dar cuenta de que ya hemos definido anteriormente al segundo cuartil ya que este estadístico no es más que el valor que es superior al 50 % de las observaciones de la variable y esa propiedad era la condición que había de cumplir necesariamente la media-na. Por tanto al hablar del percentil al 50 %, del segundo cuartil o de la mediana de una variable nos estamos refiriendo exactamente a la misma cantidad.

Ejemplo 1.11.

Calcula el primer cuartil, tercer cuartil, mediana, percentil al 12 % y percentil al 51 % de los datos de estaturas del Ejemplo 1.3

El primer cuartil no es más que el percentil al 25 % de las estaturas, por tanto para su cálculo habremos de obtener la posición número:

$$(31+1)\cdot\frac{25}{100}=8$$

de la muestra ordenada. Como tenemos la suerte de que la muestra que tenemos en dicho ejemplo ya se nos proporciona ordenada, resulta sencillo comprobar que este valor coincide con la octava posición de dicha muestra, por tanto $Q_1 = 1{,}64$ metros de estatura.
El tercer cuartil no es más que el percentil al 75 % de las estaturas, por tanto para su cálculo habremos de obtener la posición número:

$$(31+1)\cdot\frac{75}{100}=24$$

de la muestra ordenada. Resulta sencillo comprobar que este valor coincide con $Q_3 = 1{,}77$ metros de estatura.
Procediendo de la misma forma se puede calcular la mediana (Q_2) teniendo en cuenta que ésta no es más que el percentil al 50 %, en ese caso atendiendo a la observación número 16 de la muestra se comprueba que dicho valor vale 1.70 metros.
En cuanto al percentil al 12 % habremos de calcular la posición

$$(31+1)\cdot\frac{12}{100}=3{,}84$$

de la muestra ordenada. Es decir la posición que buscamos de la muestra ordenada estaría entre la tercera y la cuarta observación, en concreto deberíamos construir el percentil al 12 % tomando el 84 % de la cuarta observación (ya que 3.84 está más próximo de 4 que de 3) y un 16 % de la tercera observación. Es decir:

$$0{,}16\cdot X_3 + 0{,}84\cdot X_4 = 0{,}16\cdot 1{,}56 + 0{,}84\cdot 1{,}59 = 1{,}5852$$

Por tanto sólo el 12 % de los alumnos en nuestra muestra tienen una altura inferior a 1.5852 metros.
Por último, el percentil al 51 % se calculará de la siguiente forma:

$$(31+1)\cdot\frac{51}{100}=16{,}32$$

(participarán los datos situados en la muestra ordenada en la 16ª y la 17ª posición). El percentil queda:

$$0{,}68\cdot X_{16} + 0{,}32\cdot X_{17} = 0{,}68\cdot 1{,}70 + 0{,}32\cdot 1{,}71 = 1{,}7032$$

Por tanto el 51 % de los alumnos en nuestra muestra tienen una altura inferior a 1.7032 metros.

1.4.4. Medidas de dispersión

Los estadísticos de dispersión en general nos informan de la variabilidad de los datos, es decir si éstos son más dispersos o por el contrario se suelen agrupar de forma más o menos precisa en torno a cierto valor. Algunas medidas de dispersión importantes serían las siguientes:

- **Rango**. El *rango* es la diferencia entre el máximo y el mínimo valor de la variable.

$$Rango = M\acute{a}ximo - M\acute{i}nimo$$

- **Rango intercuartílico**. El *rango intercuartílico* se define como la dife-rencia entre el tercer y primer cuartil.

$$R.I.C. = Q_3 - Q_1 = P_{75} - P_{25}$$

  La principal ventaja que presenta el rango intercuartílico frente al rango es que este último se suele ver bastante afectado por la presencia de cualquier valor anómalo (anormalmente alto o bajo), mientras que el rango intercuartílico es bastante menos sensible a ese tipo de observaciones. Por tanto, en ocasiones suele ser preferible utilizar el rango intercuartílico en lugar del rango como medida de dispersión de los datos.

- **Desviación típica**. La *desviación típica* resume la distancia que suele darse entre cada observación y la media. En su cálculo, a diferencia del Rango y el Rango intercuartílico, en las que únicamente se incluyen dos observaciones (o bien el máximo y mínimo, o bien el primer y tercer cuartil), intervienen todos y cada uno de los valores. Se calcula mediante la siguiente expresión:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}} = \sqrt{\frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + ... + (x_n - \overline{x})^2}{n-1}}$$

  Suele ser habitual denotar a la desviación típica como s. Y su interpretación habitual es la distancia a la que soleremos encontrar las observaciones respecto de la media.

- **Varianza**. La *Varianza* es el cuadrado de la desviación típica. Se puede calcular mediante la fórmula:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1} = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + ... + (x_n - \overline{x})^2}{n-1}$$

  Suele ser habitual denotar a la varianza como s^2. Su interpretación no es tan clara como la de la desviación típica, simplemente debemos conocer que valores mayores de la varianza corresponderán a muestras que tienen mayor variabilidad. Aunque la interpretación de los valores de la varianza no es demasiado intuitiva conviene conocer su existencia ya que es un indicador bastante utilizado en la literatura.

- **Coeficiente de variación**. El *coeficiente de variación* es una medida de dispersión que viene definida por el cociente entre la desviación típica y la media, multiplicado por 100.

$$CV = \frac{s}{\overline{x}} \cdot 100$$

  La justificación de este indicador es que habitualmente las variables con valores más grandes (su media será mayor) son también las variables con mayor dispersión (su desviación típica será mayor). Al hacer el cociente de la desviación típica y la media estamos anulando dicho efecto y por tanto el coeficiente de variación nos permitirá la comparación de la variabilidad de variables medidas en escalas o unidades distintas.

Ejemplo 1.12.

Halla el rango y rango intercuartílico de los datos de estaturas del Ejemplo 1.3

El rango de los valores observados valdrá:

$$Rango = 1{,}94 - 1{,}48 = 0{,}46$$

mientras que el rango intercuartílico valdrá:

$$R.I.C. = 1{,}77 - 1{,}64 = 0{,}13$$

Ejemplo 1.13.

Halla la desviación típica y la varianza de la variable de esta-turas del Ejemplo 1.3

La media según vimos en el Ejemplo 1.10 vale 1.708 metros, por tanto la varianza será:

$$s^2 = \frac{(1{,}48 - 1{,}708)^2 + (1{,}56 - 1{,}708)^2 + ... + (1{,}94 - 1{,}708)^2}{(31-1)} =$$

$$= \frac{0{,}3122}{30} = 0{,}0104$$

Y la desviación típica:

$$s = \sqrt{s^2} = \sqrt{0{,}0104} = 0{,}1020$$

1.4.5. Valores atípicos (*outliers*)

Los *valores atípicos* en un conjunto de datos son aquellos que son mucho mayores o mucho menores que el resto de valores. Hay diferentes criterios para definir qué se entiende por *mucho mayor* o *mucho menor*, pero en este curso utilizaremos un criterio basado en los cuartiles. Consideraremos valores atípicos por exceso a aquellos que sean mayores al tercer cuartil (Q_3) más 1,5 veces el rango intercuartílico ($R.I.C.$) y valores atípicos por defecto a aquellos que sean menores al primer cuartil (Q_1) menos 1,5 veces el rango intercuartílico ($R.I.C.$).

Así, en general, podemos decir que son valores atípicos todos los que no se encuentren en el intervalo:

$$[Q_1 - 1{,}5 \cdot R.I.C., Q_3 + 1{,}5 \cdot R.I.C.]$$

Ejemplo 1.14.

Determina los valores atípicos, si los hay, del conjunto de datos de estaturas del Ejemplo 1.3

En el ejemplo de estaturas, hemos calculado previamente:

$$Q_1 = 1{,}64$$

$$Q_3 = 1{,}77$$

$$R.I.C. = 0{,}13$$

Por tanto, serían valores atípicos los que se encontraran fuera del intervalo:

$$(1{,}64 - 1{,}5 \cdot 0{,}13, 1{,}77 + 1{,}5 \cdot 0{,}13) = (1{,}445, 1{,}965)$$

En los datos de estaturas del Ejemplo 1.3 no hay ningún valor fuera de este intervalo, por tanto no hay ningún valor atípico.

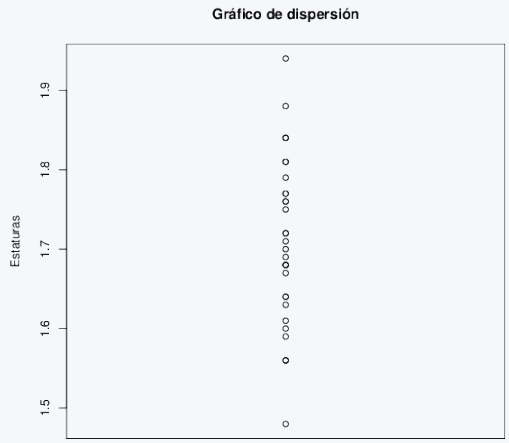
1.4.6. Representación gráfica

A continuación describimos las principales representaciones gráficas de datos cuantitativos. Estas representaciones nos ayudarán a visualizar los datos y de esta forma conocer sus principales características.

- **Gráfico de dispersión:** En relación a la representación gráfica de varia-bles cuantitativas, el primer factor que habremos de tener en cuenta a la hora de optar por una u otra representación será el nivel de detalle de los datos originales que queremos que refleje la representación. Así, en caso de querer que el gráfico conserve la mayor cantidad de información posible de la albergada originalmente en los datos, optaremos por un *gráfico de dispersión*, en el que se representarán todos los datos disponibles sobre una escala apropiada.

Ejemplo 1.15.

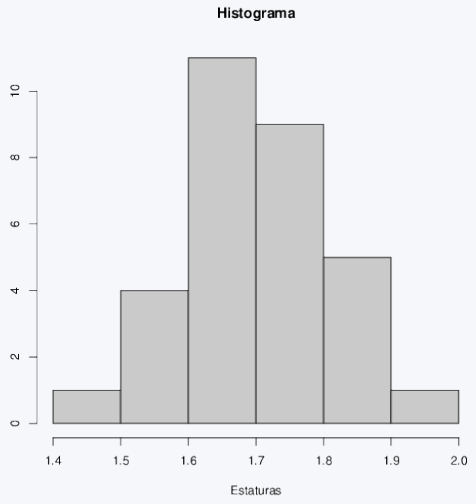
Representa mediante un gráfico de dispersión los datos de estaturas del Ejemplo 1.3



- **Histograma:** En caso de no ser necesario que aparezca el valor exacto de cada dato en la representación o cuando el número de observaciones sea demasiado grande, en cuyo caso la concentración de puntos en un gráfico de dispersión impediría observar con claridad las localizaciones que aglutinan una mayor cantidad de observaciones, puede ser más conveniente recurrir a un histograma para representar los datos. Para la elaboración de un *histograma* se ha de considerar una partición del rango de valores que ocupan los datos, de la misma forma que se ha descrito en la construcción de tablas de frecuencias de variables cuantitativas. Una vez resumida la variable en la tabla de frecuencias, en cada uno de los intervalos (clases) que la componen se representará una columna de altura proporcional a la frecuencia absoluta de ese intervalo (o de forma similar para la frecuencia relativa)

Ejemplo 1.16.

Representa mediante un histograma los datos de estaturas del Ejemplo 1.3



- **Diagrama de cajas:** Aún así nos puede ser suficiente con una representación todavía más esquemática de cómo se distribuyen los datos, en ese caso se puede optar por un *Diagrama de cajas*. En éste aparece en la parte central una *caja* cuyos extremos están delimitados por el primer y tercer cuartil, mientras que la mediana aparece como una línea que divide la caja anterior. A su vez los llamados *bigotes* de la caja, aparecen unidos por un segmento que cruza la caja anterior y que da una idea aproximada del rango de los datos. Hay diferentes criterios para representar los bigotes, pero el que estudiaremos en este curso será el que se detalla a continuación: el *bigote inferior* representará o bien 1,5 veces el $R.I.C.$ por debajo del primer cuartil o bien el valor mínimo si éste no es un valor atípico; y el *bigote superior* representará o bien 1,5 veces el $R.I.C.$ por encima del tercer cuartil o bien el valor máximo si éste no es un valor atípico. Si hay valores atípicos en el conjunto de datos, se representan mediante puntos aislados fuera del diagrama. Nuevamente, los detalles de cada uno de las representaciones anteriores (orientación horizontal/vertical de la re-presentación, colores,...) se dejan a la elección del usuario en función de las características de los datos y los requerimientos de la información que se quiera representar.

Ejemplo 1.17.

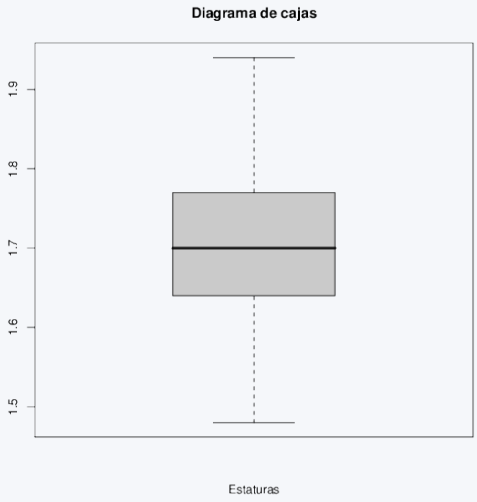
Representa mediante un diagrama de cajas los datos de estatu-ras del Ejemplo 1.3

$$\acute{M}inimo = 1{,}48$$
$$P_{25} = Q_1 = 1{,}64$$
$$Mediana = P_{50} = Q_2 = 1{,}70$$
$$P_{75} = Q_3 = 1{,}77$$
$$M\acute{a}ximo = 1{,}94$$
$$RIC = 1{,}77 - 1{,}64 = 0{,}13$$

Intervalo que determinará los valores atípicos:

$$(1{,}445, 1{,}965)$$

Como no hay valores atípicos los bigotes representarán el valor $\acute{M}inimo$ y el $M\acute{a}ximo$.



Ejemplo 1.18.

A continuación se relacionan las edades de una muestra de usuarios de un centro de rehabilitación fisioterapeútica: (51 63 61 44 63 57 53 63 44 59 51 56 58 59 71 25 28 82 85 72 58 72 58)***. Representa mediante un diagrama de cajas estos datos***

En primer lugar ordenaremos la muestra y calcularemos los estadísticos que necesitamos:

25 28 44 44 51 51 53 56 57 58 58 58 59 59 61 63 63 63 71 72 72 82 85

$$M\acute{i}nimo = 25$$

$$P_{25} = Q_1 = 51$$

$$Mediana = P_{50} = Q_2 = 58$$

$$P_{75} = Q_3 = 63$$

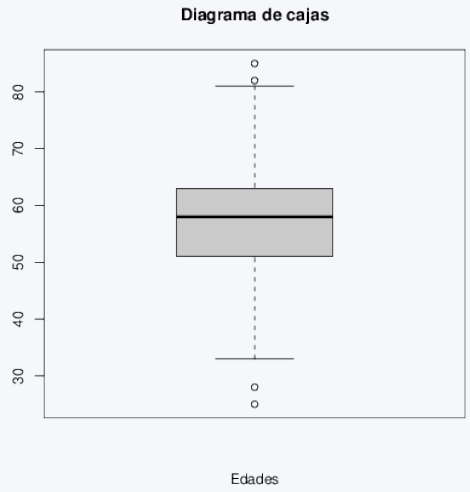
$$M\acute{a}ximo = 85$$

$$RIC = 63 - 51 = 12$$

Intervalo que determinará los valores atípicos:

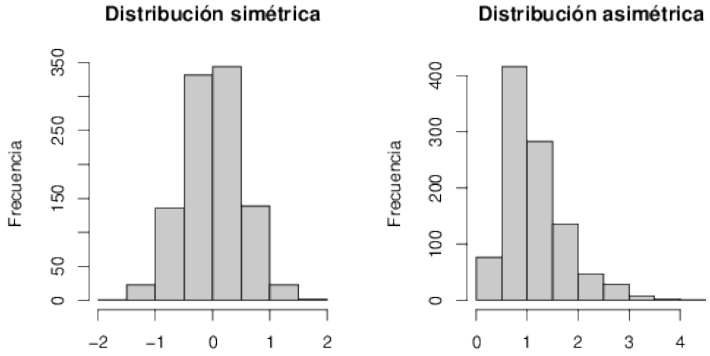
$$(51 - 1{,}5 \cdot 12, 63 + 1{,}5 \cdot 12) = (33, 81)$$

Hay dos valores en la muestra que son atípicos por defecto (el 25 y el 28) y otros dos valores que son atípicos por exceso (el 82 y el 85). Por tanto, los bigotes los representarán el valor 33 y el 81 y los valores atípicos aparecerán en la representación gráfica como puntos aislados.

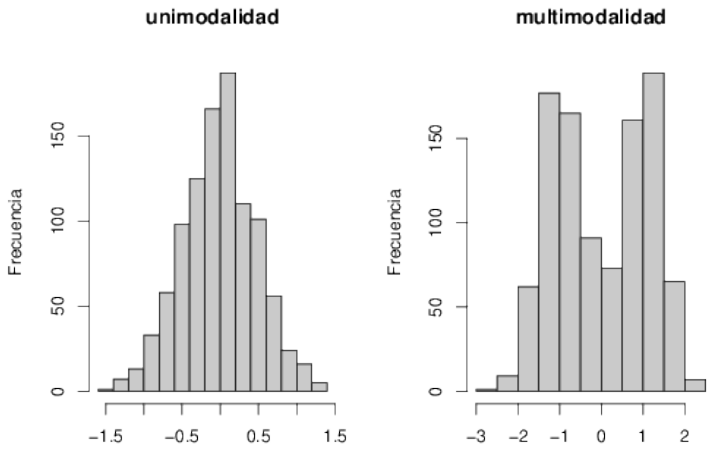


1.4.7. Medidas de forma (idea a nivel gráfico)

Las representaciones gráficas son extremadamente útiles ya que nos permiten apreciar información que no nos proporcionan los estadísticos de localización ni los de dispersión. Existen medidas cuyo valor numérico describe el tipo de comportamiento que a continuación vamos a comentar, pero en este curso nos centraremos únicamente en la idea que nos proporciona su representación gráfica. Así, por ejemplo, las representaciones gráficas nos permiten evaluar la *simetría* o asimetría de la distribución de los datos alrededor de su valor central.



Además, las representaciones gráficas también resultan útiles para evidenciar datos cuya distribución es *multimodal*, es decir que presentan más de una moda, entendiendo el concepto de moda de una forma amplia: aquella localización en torno a la cual tienden a agruparse los datos. Así, en la siguiente representación podemos apreciar un conjunto de datos unimodal junto a otra variable bimodal.



1.5. Estadística descriptiva bivariante

En esta sección abordaremos la representación gráfica de variables cuando estemos interesados en visualizar la relación entre dos de ellas en lugar de interesarnos la forma de cada una de ellas por separado. Nuevamente distinguiremos el tipo de representación que resultará más adecuada en función del tipo de variables que queramos visualizar

1.5.1. Dos variables cualitativas

La forma más adecuada de describir la relación entre dos variables categóricas es a partir de la construcción de una *tabla de contingencia*. Para ello se introduce en cada fila de la tabla las categorías de una de las variables y las categorías de la otra variable se asocian a cada una de las columnas de la tabla, en cada celda de la tabla aparecerá el número de observaciones correspondientes a la combinación oportuna de ambas variables. En cuanto a la representación gráfica de la relación entre dos variables categóricas se puede optar o bien por un gráfico de barras o bien un diagrama de sectores para cada una de las categorías de una de las variables.

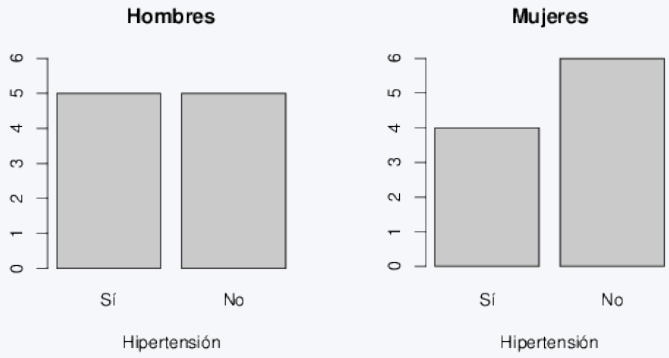
Ejemplo 1.19.

Resume en una tabla de contingencia y mediante una re-presentación gráfica la relación entre las variables Sexo y Hipertensión de un estudio en el que se han relacionado dichos factores (datos originales no mostrados)

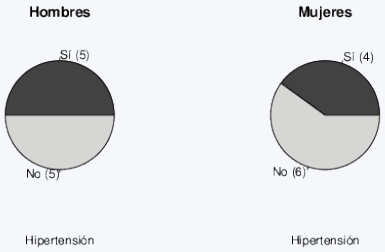
En la tabla de contingencia simplemente hemos contado cuantas veces se ha dado cada combinación de ambas variables.

Sexo \ Hipertensión	Sí	No	Total
Hombre	5	5	10
Mujer	4	6	10
Total	9	11	20

Respecto a la representación gráfica podemos optar o bien por gráficos de barras:



o bien por gráficos de sectores:



para cada una de las categorías de la variable sexo, aunque podríamos haber optado por realizar dichos gráficos para cada categoría de la varia-ble hipertensión. La elección de una representación u otra dependerá del objetivo concreto que se persiga o el matiz concreto de los datos que se quiera evidenciar.

1.5.2. Una variable cualitativa y otra cuantitativa

La descripción conjunta de una variable categórica y otra cuantitativa se reduce a la descripción de la variable cuantitativa, tal y como se ha descrito en la sección de análisis univariante, para cada una de las categorías de la variable cualitativa.

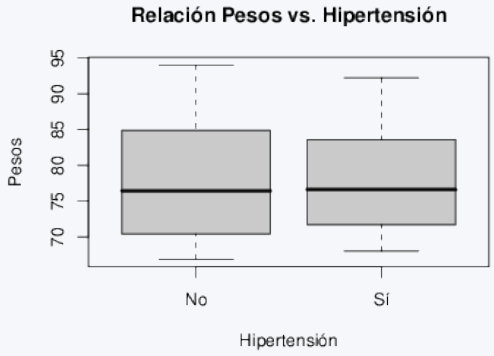
Ejemplo 1.20.

Resumen de la relación entre las variables Hipertensión y Peso del estudio anterior

A nivel numérico se han calculado las medianas de las observaciones de los pesos, tanto para el grupo de hipertensos por un lado, como para aquellos que no lo son por otro (de la misma forma se podría haber obtenido más estadísticos como cuartiles, medias, desviaciones típicas,...).

- Mediana grupo de hipertensos: 76,6 kg.
- Mediana grupo de no hipertensos: 76,4 kg

En cuanto a la representación gráfica se ha representado un diagrama de cajas para visualizar los valores que se han dado en ambos grupos:



Aunque la mediana del peso para ambos grupos es similar en la representación gráfica se aprecia que la variabilidad en el grupo de hipertensos es algo menor tal y como se aprecia en su menor rango y rango intercuartílico.

1.5.3. Dos variables cuantitativas

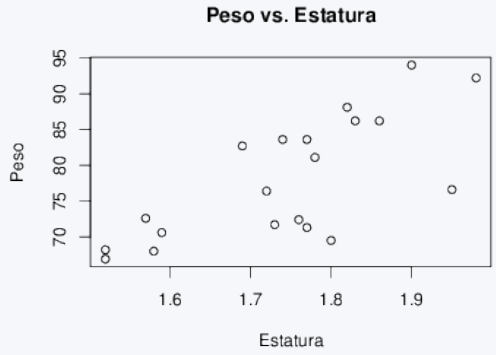
La descripción conjunta de dos variables cuantitativas se lleva a cabo a partir de la obtención del coeficiente de covarianza y del coeficiente de correlación de Pearson. En este curso profundizaremos en el cálculo y el uso del segundo de estos coeficientes en el tema de *Regresión lineal*.

Como representación gráfica utilizamos la *nube de puntos* (también llamada *Gráfico de dispersión bivariante*) que forman las dos variables cuantitativas representadas simultáneamente sobre un sistema de ejes cartesiano.

Ejemplo 1.21.

Representa gráficamente la relación entre las variables de estatura y peso del estudio anterior

La nube de puntos obtenida a partir de ambas variables tiene la siguiente forma:



Podemos apreciar que en general las personas de mayor estatura coinciden con aquellas de mayor peso y ambas variables siguen una relación que se asemeja a la de una recta, es decir una relación lineal.

1.6. Ejercicios Capítulo 1

Ejercicio 1.1.

Clasifica las siguientes variables según su tipo: cualitativas nominales, cualitativas ordinales, cuantitativas continuas o cuantitativas discretas.

- Talla de camiseta (S,M,L,XL,XXL)
- Número de calzado
- Temperatura corporal de un paciente
- Día de la semana
- Número de hijos
- Último libro leído
- Grado de aceptación de una decisión (de acuerdo, neutral, en desacuerdo)
- Marca de café preferida
- Línea del autobús que tomo más frecuentemente
- Número de asignaturas aprobadas el último curso.

Ejercicio 1.2.

En una farmacia se está recogiendo información sobre el grado de satisfacción de los clientes respecto a su servicio nocturno, concretamente se está preguntando cuál es la opinión de los clientes en cuanto la relación calidad-precio de este servicio nocturno. Las respuestas dadas por los clientes encuestados han sido codificadas según los códigos:

- 0: Muy desfavorable
- 1: Desfavorable
- 2: Favorable
- 3: Muy favorable

Se ha preguntado a un total de 50 clientes, y sus respuestas codificadas numéricamente han sido las siguientes:

0 1 3 0 1 1 2 3 0 0 3 3 3 2 1 2 0 3 0 2 1 0 0 2 3
2 2 2 1 1 2 2 0 3 0 2 2 0 3 3 0 3 0 1 2 2 2 0 2 1

1. Indica de qué tipo de variable se trata.
2. Resume los datos en la forma que consideres más adecuada.

Ejercicio 1.3.

En una encuesta a personas con hipertensión arterial, se les ha preguntado el número de veces que han recibido control de su presión arterial en los últimos 6 meses. Las respuestas se muestran a continuación:

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

1. Indica de qué tipo de variable se trata.
2. Resume los datos de esta variable en una tabla de frecuencias.

Ejercicio 1.4.

Un médico de cabecera en un área rural está interesado en conocer cuándo se producen un mayor número de demandas de asistencia a domicilio para reforzar el horario que más lo necesita. Para ello ha recogido datos sobre las últimas demandas que ha tenido y las ha catalogado como visitas de mañana, tarde, noche o festivo dependiendo de la hora y el día en el que se han producido. Los datos que ha obtenido son los siguientes:

Mañana	Mañana	Noche	Festivo	Noche	Tarde	Noche
Mañana	Mañana	Noche	Tarde	Festivo	Tarde	Mañana
Mañana	Mañana	Tarde	Mañana	Noche	Tarde	Tarde
Mañana	Tarde	Festivo	Mañana	Noche	Festivo	Mañana
Tarde	Festivo	Tarde	Noche			

- Identifica las unidades experimentales, la variable de estudio y el tipo de ésta.
- ¿Puedes calcular la mediana y rango de los datos?
- Calcula las frecuencias absolutas y relativas de cada tipo de visita.
- Realiza una gráfico de sectores y un gráfico de barras.

Ejercicio 1.5.

Se han tomado muestras a 40 niños de entre 1 y 5 años del nivel de cobre en orina, obteniéndose los siguientes valores:

0.10	0.30	0.34	0.36	0.42	0.42	0.45	0.48	0.50	0.52
0.55	0.58	0.62	0.63	0.64	0.65	0.65	0.66	0.69	0.70
0.72	0.73	0.74	0.74	0.75	0.76	0.77	0.78	0.81	0.83
0.85	0.86	0.88	0.90	0.94	0.98	1.04	1.12	1.16	1.24

- Identifica las unidades experimentales, la variable de estudio y el tipo de ésta.
- Calcula la mediana y rango de los datos.
- Calcula el primer y tercer cuartil, rango intercuartílico, percentil 10, percentil 95.
- Consideras alguno de los valores como atípico.
- Realiza un histograma y un diagrama de cajas.

Ejercicio 1.6.

Se dispone del peso (en gramos) de 16 niños de un mes de edad. Los datos se muestran a continuación:

4123	4336	4160	4165	4422	3853	3281	3990
4096	4166	3596	4127	4017	3769	4240	4194

1. Indica de qué tipo de variable se trata.
2. Calcula los siguientes estadísticos:

- Mínimo
- Máximo
- P_{10}
- $P_{25}(= Q_1)$
- $P_{50}(= Q_2)$
- $P_{75}(= Q_3)$
- P_{90}
- Media
- Mediana
- Moda
- Rango
- Rango IC
- Varianza (s^2)
- Desviación típica (s)
- Coeficiente de variación (CV)

Ejercicio 1.7.

En una farmacia se realiza seguimiento de la Hipertensión Arterial de algunos pacientes. Se dispone de 30 mediciones de la tensión arterial sistólica (TAS) realizadas en el día de hoy, las cuales se muestran a continuación:

173,03	165,54	141,59	158,66	158,81	156,49	150,29	154,53	162,50	158,49
151,11	166,13	147,47	152,83	166,99	135,62	138,77	168,11	162,04	176,77
159,97	152,99	161,92	167,70	143,35	154,06	160,82	180,08	172,93	158,72

1. Indica de qué tipo de variable se trata
2. Resume los datos de esta variable en una tabla de frecuencias
3. Calcula los siguientes estadísticos:

- Mínimo
- Máximo
- P_{10}
- $P_{25}(= Q_1)$
- $P_{50}(= Q_2)$
- $P_{75}(= Q_3)$
- P_{90}
- Media
- Mediana
- Moda
- Rango
- Rango IC
- Varianza (s^2)
- Desviación típica (s)
- Coeficiente de variación (CV)

4. Realiza un histograma y un diagrama de cajas.

Capítulo 2

Variables aleatorias y distribución Normal

2.1. Variable aleatoria y distribución

Una *variable*, como hemos estudiado en el tema anterior, es una característica que puede ser medida y que puede adoptar valores diferentes para cada uno de los elementos que constituyen la población de estudio. El atributo *aleatoria*, marca precisamente la presencia de incertidumbre en el valor de la variable de cada unidad experimental.

El comportamiento de los posibles valores varía de unas variables a otras (sea cuantitativa o cualitativa). Podemos encontrar variables en las que algunos valores aparecen con mayor frecuencia (o probabilidad) y otros lo hacen con una frecuencia menor. Otras variables, en cambio, presentan valores que se repiten aproximadamente con la misma frecuencia. Este hecho hace que el comportamiento que tienen los posibles valores de una variable aleatoria en relación a la frecuencia con la que los podemos encontrar en las unidades experimentales sea objeto de posible estudio.

Probabilidad de los valores de una variable aleatoria

Casi todos tenemos una idea intuitiva más o menos acertada del término probabilidad, frases como 'Este resultado es más probable que el otro' o 'La probabilidad de que se dé ese resultado es muy baja' no nos son para nada ajenas. Aun así dentro de esta asignatura y en cualquier discusión relacionada con la estadística es importante disponer de un concepto de probabilidad algo más riguroso y ese es el propósito que nos disponemos a cumplir a continuación.

Se define la *probabilidad* de cualquier valor x_0 de una variable aleatoria X como la frecuencia relativa que esperaríamos que tomara el valor x_0 en caso de disponer una muestra de la variable X de tamaño infinito. Obviamente cuanto mayor sea el número de observaciones que dispongamos las frecuencias relativas de los valores de cualquier variable reproducirán mejor la probabilidad de dichos valores. Como consecuencia de la definición anterior la probabilidad de cualquier valor tomará necesariamente un valor entre 0 y 1 (toda frecuncia relativa toma valores en dicho intervalo) y la suma de las probabilidades de todos los valores posibles de cualquier variable aleatoria simpre valdrá 1 (resulta sencillo demostrar que las frecuencias relativas de cualquier variable aleatoria también cumplen siempre esta propiedad).

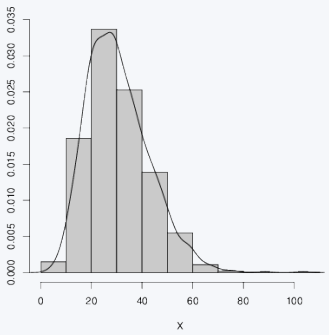
Cuando la variable que estudiemos sea cualitativa el concepto de probabilidad se define sin ningún tipo de problema ni ambigüedad. Las frecuencias de cualquier valor razonable siempre tomará valores superiores a 0. De todas formas para las variables cuantitativas continuas este concepto no está tan claro. Según hemos visto la suma de las probabilidades de todos los valores de la variable ha de sumar 1, y en este caso tenemos infinitos valores posibles. Si estos valores tienen probabilidad superior a 0 dicho criterio no se cumplirá (la suma de infinitos valores superiores a 0 es infinito). Por tanto cuando nos refiramos a cualquier variable cuantitativa continua sus valores tendrán probabilidad 0 (acaso, en una muestra de valores de la altura de los alumnos de esta clase ¿cuantos alumnos esperaríamos ver de altura 173.5093427594369.... centímetros?). Para este tipo de variables habremos de hablar de la probabilidad de un conjunto de valores y no de un valor único, por ejemplo la probabilidad de que la altura de cualquier alumno esté entre 173 y 176 cm. Dicha probabilidad no será nula y así el concepto de probabilidad para variables continuas recobra sentido de esta forma.

Distribución

Se define la *función de densidad de probabilidad de una variable aleatoria* como aquella función que para cualquier valor de la variable (ya sea un valor concreto o un intervalo por ejemplo) nos devuelve la probabilidad de dicho valor. En caso de que la variable que tengamos no sea cuantitativa continua la función de distribución valdrá en cualquier punto el valor de su probabilidad. En el caso de variables continuas la función de densidad será aquella función que para cualquier intervalo de valores $[a, b]$ el área que encierra la función entre a y b es exactamente la probabilidad de que la variable aleatoria tome un valor en dicho intervalo. La función de densidad de una variable nos informa completamente de que valores y con que frecuencia se distribuye la variable, así muchas veces nos referiremos a la *función de densidad de probabilidad de una variable aleatoria* como *distribución* de la variable aleatoria.

Ejemplo 2.1.

Idea de distribución de una variable



En este ejemplo se puede apreciar en el histograma que los valores más frecuentes de esta variable se encuentran entre 20 y 30 (se corresponden con la barra más alta del histograma), seguidos de los valores entre 30 y 40, y los menos frecuentes aquellos próximos a 0 y los mayores de 70. La linea que aparece en el histograma corresponde a la función de densidad, o distribución, de la variable aleatoria cuyos valores se han representado en el histograma. Cuantos más valores dispongamos de dicha variable más parecidos serán los valores del histograma y los de la distribución de la variable.
Podemos obtener, a partir de la curva que se ha representado superpuesta en la representación anterior y que se muestra a continuación, al menos las mismas ideas que podríamos extraer del histograma correspondiente (valores con mayor frecuencia, simetría,...). De hecho, dicho histograma no es más que una aproximación (basada en la información que nos proporcionan los datos) de la función de distribución que aparece en la figura anterior.

En el caso de las variables categóricas, para conocer su distribución es suficiente con conocer las probabilidades asociadas a cada una de las posibles categorías. En el caso de las variables cuantitativas, la distribución de la variable puede ser dada en forma de *curva* (o la función *matemática* que la genera).

Tal y como hemos comentado la función de distribución de una variable resume la probabilidad de que la variable tome cualquier valor (o rango de valores). Así, el area que encierra la función de distribución entre a y b coincide con la probabilidad (o frecuencia relativa que esperaríamos observar) de que la variable tome valores entre a y b. A partir de esta propiedad podemos establecer un resultado interesante que tendrá importantes repercusiones en el futuro según veremos: ¿Qué área encerrará la función de distribución entre el menor y el mayor valor posible (de menos infinito a infinito)? dicha área será equivalente a la probabilidad (o frecuencia relativa esperada) de valores de la variable que observaríamos entre menos infinito e infinito. Como todos los valores de la variable estarán incluidos en este rango dicha probabilidad o frecuencia valdrá 1 y en consecuncia el área que encierra cualquier función de distribución entre menos infinito e infinito será uno, sea cual sea la función de distribución que tengamos.

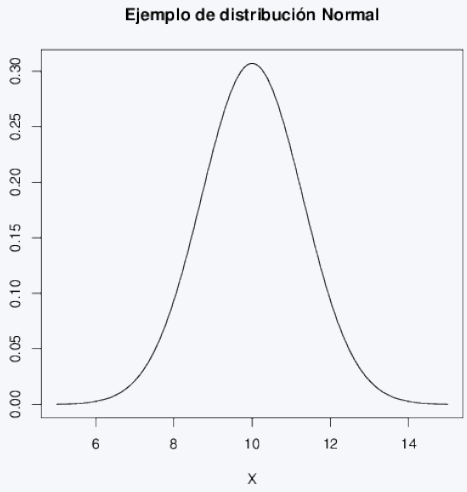
2.2. La distribución Normal

La *distribución Normal* es una familia de curvas (funciones de distribución) con las siguientes características:

- Simétricas
- Con forma de '*campana*' (no todas las curvas con esa forma siguen una distribución Normal como veremos en los próximos temas)
- Es la distribución que se presenta con mayor frecuencia en variables cuantitativas. Muchas características biológicas la siguen (*pesos de hombres y mujeres adultos, presión arterial en personas ancianas, errores aleatorios en muchos tipos de medidas de laboratorio,...*)

Ejemplo 2.2.

Ejemplos de distribución Normal

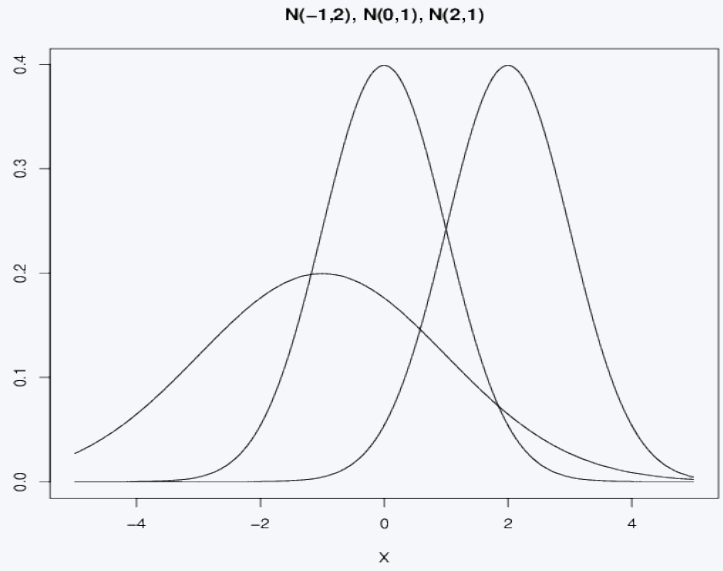


En este ejemplo se muestra la representación gráfica de una distribución Normal. Toda variable que siga esta distribución habrá de presentar un histograma similar a la curva que acabamos de presentar. Más parecido cuantos más valores dispongamos de la variable.

Una distribución Normal queda definida por dos parámetros: su *media* representada por la letra griega μ y su *desviación típica* que se suele representar por la letra griega σ. Una distribución Normal con media μ y desviación típica σ se denota mediante la expresión: $N(\mu, \sigma)$. Estos dos parámetros definen la forma concreta de la distribución. En concreto la media (μ) define la localización del centro de la campana (si ésta está más desplazada hacia la izquierda o la derecha), mientras que la desviación típica (σ) define la forma de la misma. Así una distribución Normal con mayor desviación típica que otra (según vimos en el tema anterior mayor desviación típica correspondía a datos más dispersos) tendrá una forma más "ensanchada" y "aplastada" que una distribución Normal con menor desviación típica, más "estrecha" y "apuntada". A continuación se representan algunas distribuciones Normales con distintos parámetros.

Ejemplo 2.3.

Ejemplos de distribuciones Normales con distintos parámetros



En este ejemplo se representan tres distribuciones Normales con diferentes parámetros, $X_1 \sim N(-1, 2), X_2 \sim N(0, 1), X_3 \sim N(2, 1)$. Se puede observar que las que tienen igual valor en su desviación típica pero distinta media tienen exactamente la misma forma (igual amplitud, altura,...) aunque centradas en diferentes valores (cada una en su media), mientras que cuando varía la desviación típica varía también la forma de la curva.

Dados los valores de una variable aleatoria en un conjunto de unidades experimentales, existen *pruebas de normalidad* que determinan si ciertos datos siguen o no este tipo de distribución. Informalmente, diremos que una variable sigue una distribución Normal si la mayoría de valores de esta variable se concentran alrededor de un valor tomando, con la misma probabilidad, valores mayores y menores a éste y de forma menos frecuente cuanto más nos alejamos de los valores centrales (los más frecuentes).

Propiedades de la distribución Normal

1. Queda definida por dos parámetros que son la media μ y la desviación típica σ.
2. Para expresar que los valores de una variable cuantitativa X sigue una distribución Normal con media μ y desviación típica σ diremos:
$$X \sim N(\mu, \sigma)$$
3. Los valores de una variable con esta distribución tiene su máxima frecuencia alrededor de μ (la curva de la distribución Normal alcanza su máximo en μ), y en este valor coinciden el valor de su media, su mediana y su moda.

4. Esta distribución es completamente simétrica respecto al eje vertical que corte al eje de abcisas (eje x) en el valor μ.

5. En el intervalo $[\mu - \sigma,\ \mu + \sigma]$ se encuentran, aproximadamente, el 68.26 % de los valores más frecuentes de la distribución y en el intervalo $[\mu - 2\sigma,\ \mu + 2\sigma]$ el 95.44 % de los valores más frecuentes de la distribución.

6. Es asintótica respecto al eje de abcisas, es decir, nunca llega a cruzar este eje y a medida que los valores del eje x se acercan hacia $-\infty$ o ∞ la distribución Normal se aproxima más y más a él. Ésto en términos prácticos significa que ningún valor (entendemos por valor en este caso un intervalo de valores al ser la variable continua) tiene probabilidad exactamente igual a 0 para esta distribución, aunque en términos prácticos a partir de ciertos valores la frecuencia con la que aparecerán estos valores será *casi* nula (por ejemplo 0,00000000000000000000000000001).

7. A la distribución $N(0,1)$ se le conoce como *distribución Normal Estándar* (o *Tipificada*). Existen unas tablas que permiten el cálculo de forma sencilla de los valores de las probabilidades correspondientes a la distribución $N(0,1)$. Si la variable en estudio no sigue una distribución $N(0,1)$, sino por ejemplo cualquier variable que se distribuya: $N(\mu,\sigma)$, veremos en la siguiente sección como pasar calcular sus probabilidades a partir de la distribución tipificada.

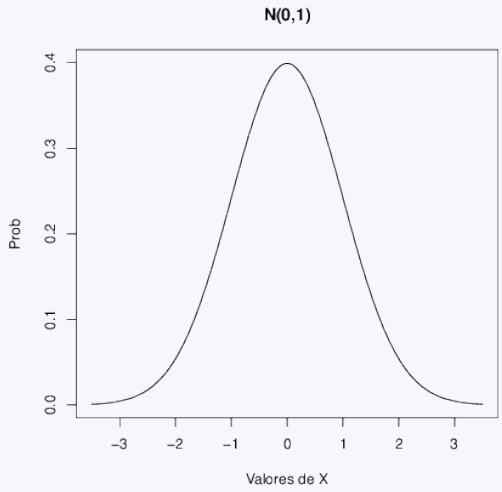
2.2.1. Distribución Normal Estándar ($N(0,1)$)

En esta sección intentaremos dar respuesta a cualquier pregunta relacionada con las probabilidades de cualquier variable que siga una distribución $N(0,1)$.

Ejemplo 2.4.

Ejemplo de distribución Normal(0,1)

Supongamos que tenemos una variable Z que sigue una distribución $N(0,1)$ ($Z \sim N(0,1)$). La representación gráfica de la distribución de frecuencias sería:



¿Qué valores toma una variable con distribución $N(0,1)$?

Como se puede apreciar en el ejemplo 2.4, los valores que tiene una variable que sigue esta distribución se situa alrededor de 0, y por tanto son tanto positivos como negativos. Los valores más frecuentes se sitúan cerca de 0, entre -1 y 1 (donde la distribución y por tanto la frecuencia relativa es más alta), y a medida que nos alejamos del 0, bien hacia valores positivos altos o bien hacia negativos bajos la probabilidad (o frecuencia) de los valores disminuye.

En el siguiente ejemplo se muestran 40 valores de una variable que sigue esta distribución:

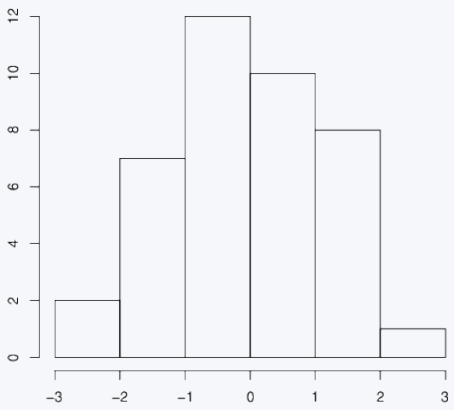
Ejemplo 2.5.

Ejemplo de valores de una distribución Normal(0,1)

Los valores que se indican a continuación han sido obtenidos de una variable que sigue una distribución $N(0,1)$:

0.73	-0.91	-0.11	-1.05	-0.05	-0.28	-2.43	0.69	-0.30	0.23
1.89	-1.83	0.68	-0.23	-1.72	-0.02	-2.15	0.33	1.49	-1.34
1.55	1.51	1.02	-1.84	-1.04	-0.14	1.25	0.44	1.45	0.46
0.61	1.38	-0.18	-0.16	-0.58	-1.43	0.37	2.61	-0.90	0.95

A continuación resumiremos estos datos mediante una representación gráfica en forma de histograma:



Podemos observar que el histograma tiene aproximadamente la forma de la distribución $N(0,1)$ (forma de campana, centrado en 0, aproximadamente simétrico, con valores entre -3 y 3 y con la frecuencia que disminuye a medida que nos alejamos de los valores centrales). Cuantos más valores tengamos en nuestra muestra más se asemejará el histograma a la distribución $N(0,1)$.

A continuación presentamos la table de probabilidades de la distribución $N(0,1)$ a partir de la cual podremos calcular la probabilidad de que una variable con esta distribución tome cualquier colección de valores (acotados en un intervalo, superiores a cierto valor,...)

Tabla de probabilidades de la distribución N(0,1)

$[P(Z < z)]$

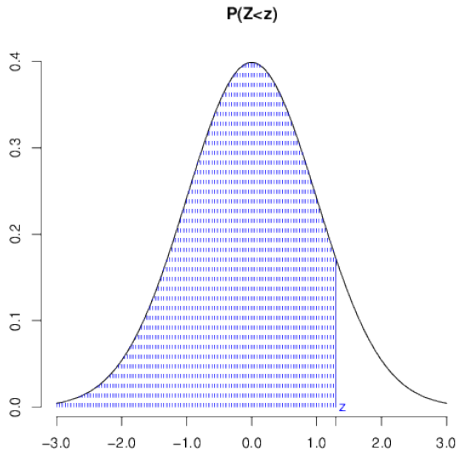
	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5039	0.5079	0.5119	0.5159	0.5199	0.5239	0.5279	0.5318	0.5358
0.1	0.5398	0.5437	0.5477	0.5517	0.5556	0.5596	0.5635	0.5674	0.5714	0.5753
0.2	0.5792	0.5831	0.5870	0.5909	0.5948	0.5987	0.6025	0.6064	0.6102	0.6140
0.3	0.6179	0.6217	0.6255	0.6293	0.6330	0.6368	0.6405	0.6443	0.6480	0.6517
0.4	0.6554	0.6590	0.6627	0.6664	0.6700	0.6736	0.6772	0.6808	0.6843	0.6879
0.5	0.6914	0.6949	0.6984	0.7019	0.7054	0.7088	0.7122	0.7156	0.7190	0.7224
0.6	0.7257	0.7290	0.7323	0.7356	0.7389	0.7421	0.7453	0.7485	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7703	0.7733	0.7763	0.7793	0.7823	0.7852
0.8	0.7881	0.7910	0.7938	0.7967	0.7995	0.8023	0.8051	0.8078	0.8105	0.8132
0.9	0.8159	0.8185	0.8212	0.8238	0.8263	0.8289	0.8314	0.8339	0.8364	0.8389
1.0	0.8413	0.8437	0.8461	0.8484	0.8508	0.8531	0.8554	0.8576	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8707	0.8728	0.8749	0.8769	0.8790	0.8810	0.8829
1.2	0.8849	0.8868	0.8887	0.8906	0.8925	0.8943	0.8961	0.8979	0.8997	0.9014
1.3	0.9032	0.9049	0.9065	0.9082	0.9098	0.9114	0.9130	0.9146	0.9162	0.9177
1.4	0.9192	0.9207	0.9221	0.9236	0.9250	0.9264	0.9278	0.9292	0.9305	0.9318
1.5	0.9331	0.9344	0.9357	0.9369	0.9382	0.9394	0.9406	0.9417	0.9429	0.9440
1.6	0.9452	0.9463	0.9473	0.9484	0.9494	0.9505	0.9515	0.9525	0.9535	0.9544
1.7	0.9554	0.9563	0.9572	0.9581	0.9590	0.9599	0.9607	0.9616	0.9624	0.9632
1.8	0.9640	0.9648	0.9656	0.9663	0.9671	0.9678	0.9685	0.9692	0.9699	0.9706
1.9	0.9712	0.9719	0.9725	0.9731	0.9738	0.9744	0.9750	0.9755	0.9761	0.9767
2.0	0.9772	0.9777	0.9783	0.9788	0.9793	0.9798	0.9803	0.9807	0.9812	0.9816
2.1	0.9821	0.9825	0.9829	0.9834	0.9838	0.9842	0.9846	0.9849	0.9853	0.9857
2.2	0.9860	0.9864	0.9867	0.9871	0.9874	0.9877	0.9880	0.9883	0.9886	0.9889
2.3	0.9892	0.9895	0.9898	0.9900	0.9903	0.9906	0.9908	0.9911	0.9913	0.9915
2.4	0.9918	0.9920	0.9922	0.9924	0.9926	0.9928	0.9930	0.9932	0.9934	0.9936
2.5	0.9937	0.9939	0.9941	0.9942	0.9944	0.9946	0.9947	0.9949	0.9950	0.9952
2.6	0.9953	0.9954	0.9956	0.9957	0.9958	0.9959	0.9960	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9971	0.9972	0.9973
2.8	0.9974	0.9975	0.9975	0.9976	0.9977	0.9978	0.9978	0.9979	0.9980	0.9980
2.9	0.9981	0.9981	0.9982	0.9983	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986
3.0	0.9986	0.9986	0.9987	0.9987	0.9988	0.9988	0.9988	0.9989	0.9989	0.9989

Uso de la tabla de la distribución $N(0,1)$

En adelante, Z representará una variable que sigue una distribución $N(0,1)$, es decir:

$$Z \sim N(0,1)$$

Estamos interesados en conocer cómo es el comportamiento de esta distribución $N(0,1)$, y para ello vamos a querer responder a preguntas del tipo: *¿Qué probabilidad tiene la variable Z de obtener un valor inferior a cierto valor z?, ¿y superior?, ¿y entre un valor z_1 y un valor z_2?* La tabla anterior responde directamente a la pregunta *¿Qué probabilidad tiene la variable Z de tomar un valor inferior a la cantidad z?* para cualquier z positivo. Esta pregunta la podríamos escribir matemáticamente mediante la expresión: $P(Z < z)$. Gráficamente, lo que queremos calcular viene representado por la siguiente figura:



Estaríamos interesados en conocer el área que encierra la función de distribución a la izquierda de la abcisa z. Para calcular esta probabilidad, recurrimos a la tabla y buscamos en la primera columna la unidad y el primer decimal de z, mientras que en la primera fila buscamos el segundo dígito decimal de z. Donde se cruzan la fila y la columna que forman conjuntamente el valor z se encuentra la probabilidad de que un valor cualquiera de la variable $N(0,1)$ sea inferior al valor z en cuestión.

Ejemplo 2.6.

¿Cuál es la probabilidad de que si elegimos un valor al azar de la variable Z sea inferior a 1,45***?***

Esta pregunta la podemos expresar en lenguaje *matemático* de forma muy reducida de la siguiente forma:

$$P(Z < 1{,}45)$$

Probabilidad de que un valor de la variable Z (que sigue una distribución $N(0,1)$) sea inferior a 1,45. Esta información (la probabilidad acumulada a la izquierda de cierto valor) es la que proporciona directamente la tabla. Como 1,45 es un número positivo tenemos suerte ya que es uno de los valores que podemos buscar directamente en la tabla. Obtendremos la respuesta buscando en la primera columna el valor 1,4 (que se corresponde con el dígito principal y el primer decimal), y en la primera fila el valor 0,05 (que se corresponde con el segundo decimal). Donde se cruzan estos dos valores en el interior de la tabla obtenemos la cantidad 0,9264, y este valor es precisamente la probabilidad que nos pedían. Así, podemos responder:

$$P(Z < 1{,}45) = 0{,}9264$$

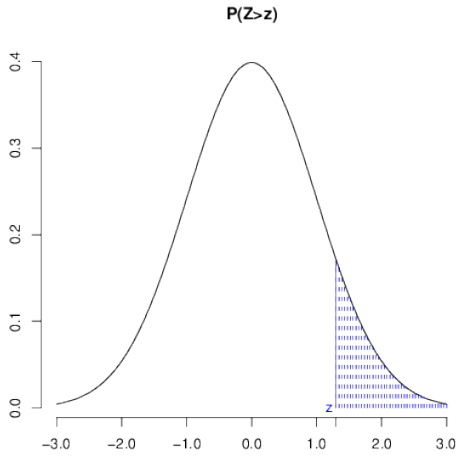
Podemos interpretar este resultado como *La probabilidad de que un valor de la variable Z sea inferior a* 1,45 *es* 0,9264, y también como que: *el* 92,64 % *de los valores de la variable Z son menores que el valor* 1,45. Además, podemos interpretar también que 1,45 es el percentil 92,64 de la distribución $N(0,1)$, pues el 92,64 % de los valores de la variable son inferiores a él.

Si miramos con detenimiento la tabla de la distribución Normal podremos observar que conforme aumenta el valor de z (el percentil que buscamos en la tabla) aumenta el valor de la probabilidad a la izquierda de dicho valor ($P(Z < z)$). Esto resulta lógico si pensamos que, en la figura anterior, cuanto mas desplazamos z hacia la derecha mayor será el área que aparece resaltada a su izquierda. Por otro lado, los valores de las probabilidades en la tabla de la distribución Normal Estandar parecen estabilizarse conforme aumentamos el valor de z. El valor en el que se estabiliza dicha probabilidad es 1. Este hecho también parece lógico ya que si tomamos un valor de z muy alto tendremos practicamente asegurado el que todo valor de la variable Normal Estandar será inferior a él. Por tanto si tuviéramos una muestra de valores de esta variable, la frecuencia relativa de veces en el que $Z < z$ será muy próxima a 1 y en consecuencia también lo será su probabilidad.

En la tabla de la distribución Normal Estándar únicamente aparecen las probabilidades asociadas a los valores de la variable que son inferiores a un número positivo cualquiera (*percentil*). Si estamos interesados en la probabilidad contraria, es decir, en *la probabilidad que tiene la variable Z de obtener un valor superior al percentil z* (que escribiríamos como $P(Z > z)$), simplemente tenemos que restar a 1 la probabilidad de que sea inferior. Es decir,

$$P(Z > z) = 1 - P(Z < z)$$

Gráficamente, esta idea viene representada en la figura:



En esta figura se aprecia que el área asociada a que $Z > z$ corresponde a la región complementaria a la utilizada cuando queríamos evaluar la probabilidad de $Z < z$. Como la extensión de ambas regiones ha de sumar 1 (tal y como vimos al final de la sección 2.1) tendremos:

$$P(Z > z) + P(Z < z) = 1$$

entonces:

$$P(Z > z) = 1 - P(Z < z)$$

tal y como acabábamos de señalar.

Ejemplo 2.7.

¿Cuál es la probabilidad de que si elegimos un valor al azar de la variable Z ***sea superior a*** $1{,}45$***?***

Esta pregunta la traducimos en la expresión:

$$P(Z > 1{,}45)$$

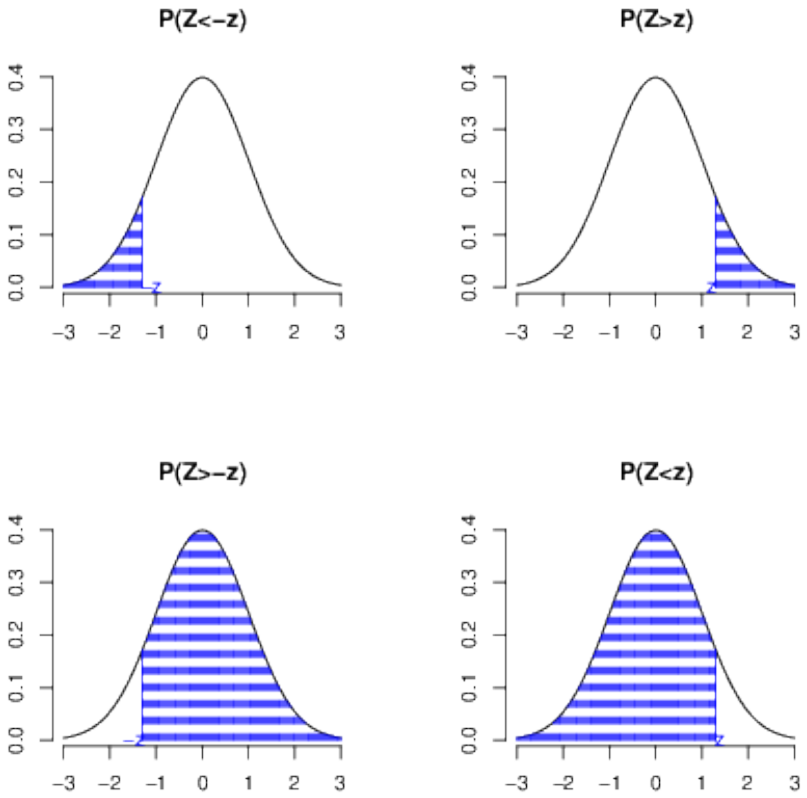
Puesto que sabemos por el ejemplo anterior que $P(Z < 1{,}45) = 0{,}9264$, tendremos:

$$P(Z > 1{,}45) = 1 - P(Z < 1{,}45) = 1 - 0{,}9264 = 0{,}0736$$

Por otro lado, tal y como hemos comentado en la tabla de la distribución Normal Estandar únicamente aparecen valores positivos de los percentiles, mientras que, como hemos comentado con anterioridad, dicha distribución tiene tanto valores positivos como negativos. El hecho de que en la tabla se muestren únicamente valores positivos está basado en la *simetría* de esta distribución respecto al 0. Conociendo el comportamiento de la distribución en la parte positiva podemos deducir cuál será el comportamiento en la parte negativa. Si z es un número positivo cualquiera (y si $-z$ representa su homólogo negativo), se cumple que:

$$P(Z < -z) = P(Z > z) \quad y \quad P(Z > -z) = P(Z < z)$$

La representación gráfica que a continuación mostramos ayuda a clarificar estas ideas. Tanto en la primeta como en la segunda fila de la figura siguiente vemos que el área señalada en azul para las figuras de la izquierda coincide con el área señalada en las figuras de la derecha, lo que demuestra las igualdades anteriores.



La aplicación de las igualdades que acabamos de señalar nos permitirá calcular las probabilidades a la izquierda o a la derecha de cualquier número negativo en una distribución Normal Estandar.

Ejemplo 2.8.

Considerando que la variable Z sigue una distribución $N(0,1)$, calcula las probabilidades: $P(Z < -2{,}35)$ ***y*** $P(Z > -2{,}56)$

- Primera probabilidad:

$$P(Z < -2{,}35) = P(Z > 2{,}35) = 1 - P(Z < 2{,}35) = 1 - 0{,}9906 = 0{,}0094$$

- Segunda probabilidad:

$$P(Z > -2{,}56) = P(Z < 2{,}56) = 0{,}9947$$

En las dos expresiones anteriores las dos últimas igualdades han sido determinadas únicamente mirando la tabla de la distribución Normal Estandar.

En el caso de probabilidades compuestas, es decir, si queremos calcular por ejemplo la probabilidad de que un valor de la variable Z esté entre dos valores z_1 y z_2 (que puede ser expresado como $P(z_1 < Z < z_2)$), tendríamos varias formas de resolverlo, pero una sencilla podría ser:

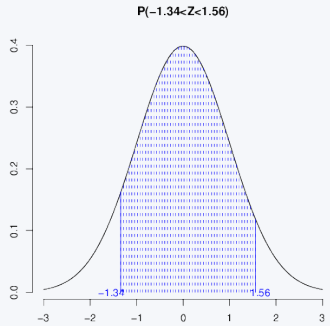
$$P(z_1 < Z < z_2) = P(Z < z_2) - P(Z < z_1)$$

En el ejemplo que se muestra a continuación, se muestra el cálculo de una probabilidad de este tipo con el apoyo de representaciones gráficas.

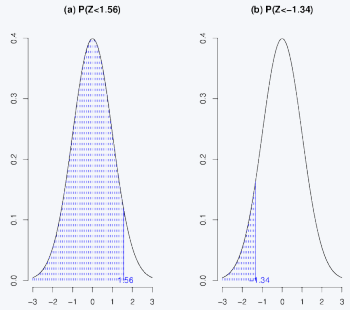
Ejemplo 2.9.

Si la variable Z ***sigue una distribución*** $N(0,1)$***, calcula la probabilidad:*** $P(-1{,}34 < Z < 1{,}56)$

Para este tipo de probabilidades entre dos valores, nos apoyaremos de representaciones gráficas. La probabilidad que buscamos sería la que se representa gráficamente en la siguiente figura:



Como el área anterior no puede deducirse directamente mirando en la tabla habremos de elaborar un poco más nuestros cálculos para poder obtenerla. Podemos calcular el área seleccionada de varias formas. Por ejemplo, podríamos calcular el área que se muestra en la figura **(a)** ($P(Z < 1{,}56)$) y a continuación restarle el área que se muestra en la figura **(b)** ($P(Z < -1{,}34)$).



Así, podríamos calcular:

$$P(-1{,}34 < Z < 1{,}56) = P(Z < 1{,}56) - P(Z < -1{,}34) =$$

$$= 0{,}9406 - (1 - 0{,}9098) = 0{,}8504$$

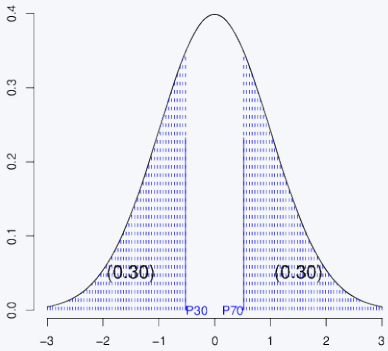
Para finalizar, en ocasiones estaremos interesados en realizar el proceso inverso al que hemos realizado hasta ahora, es decir, en lugar de hallar la probabilidad a la izquierda de un percentil estaremos interesados en hallar el valor del percentil al que le corresponde cierta probabilidad. En ese caso buscaremos la probabilidad deseada en el interior de la tabla y comprobaremos a qué valor de percentil corresponde (según la fila y columna en la que esté situada la probabilidad de interés).

La forma más sencilla de entender este mecanismo es mediante un par de ejemplos que mostramos a continuación.

Ejemplo 2.10.

La variable Z ***sigue una distribución*** $N(0,1)$***. Calcula el percentil*** 30 ***y el percentil*** 70 ***para los valores de esta variable***

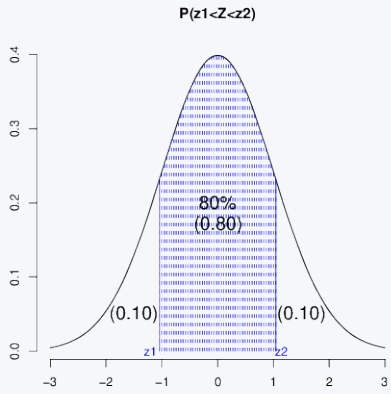
El percentil 70 es el valor que cumple que el $70\,\%$ de los valores de la variable son inferiores a él. Este percentil será un número positivo, ya que la distribución $N(0,1)$ es simétrica respecto de 0 (su media) y por tanto 0 es su mediana (que es el percentil al $50\,\%$). Los percentiles superiores a 50 serán números positivos y los inferiores números negativos. Para buscar el percentil 70 simplemente tenemos que buscar en el interior de la tabla la probabilidad 0,70 (o el valor más cercano en su defecto) y comprobar a qué valor corresponde (comprobando en qué fila y columna se halla ese valor). Así, podemos comprobar que $P_{70} = 0{,}52$. El percentil 30 por simetría se corresponderá con el percentil 70, por tanto: $P_{30} = -0{,}52$



Ejemplo 2.11.

Supongamos de nuevo que la variable Z sigue una distribución $N(0,1)$. Calcula un intervalo centrado (en 0) que contenga el $80\,\%$ de los valores de la variable

Estamos interesados en buscar los valores z_1 y z_2 que cumplan que $P(z_1 < Z < z_2) = 0{,}80$. Gráficamente:



A partir del gráfico podemos deducir algunas ideas:

- Si el intervalo está *centrado* en 0, el valor z_1 será el homólogo a z_2 pero en negativo.
- Si queremos que el intervalo contenga el $80\,\%$ de las observaciones (probabilidad 0,80), fuera del intervalo quedarán el $20\,\%$ restantes, repartido en los dos extremos ($10\,\%$ de valores por encima del valor z_2 y $10\,\%$ de valores por debajo del valor z_1).
- El valor que puede aparecer en la tabla $N(0,1)$ es el valor z_2, puesto que z_1 es negativo, y z_2 es el valor que cumple que el $90\,\%$ de las observaciones son inferiores a él (o lo que es lo mismo, el percentil 90).

Buscando el percentil 90 en la tabla, es decir, buscando en el interior de la tabla el valor de la probabilidad 0,90 (o en su defecto el valor más cercano) observamos que le corresponde al valor 1,28, y por tanto $z_2 = 1{,}28$. En consecuencia $z_1 = -1{,}28$ y el intervalo que nos piden es $(-1{,}28, 1{,}28)$

2.2.2. Aritmética de variables normales

En la sección anterior hemos aprendido a responder cualquier pregunta (en términos de probabilidad) de las variables que siguen una distribución $N(0,1)$. Comentamos al principio de este tema, que multitud de variables biológicas siguen distribuciones normales, pero es evidente, que los parámetros de estas distribuciones normales no son necesariamente $\mu = 0$ y $\sigma = 1$, pensar por ejemplo en las alturas de los alumnos de la clase. En esta sección aprenderemos a responder para cualquier variable que siga una distribución Normal el mismo tipo de preguntas que nos planteábamos para variables normales estándar.

Cualquier variable X que siga una distribución $N(\mu, \sigma)$, se puede transformar fácilmente en una $N(0,1)$, simplemente restando a todos sus valores la media (μ) y dividiendo por su desviación

típica (σ). Los nuevos valores que se obtienen de esta transformación de la variable X siguen una distribución Normal estándar ($N(0,1)$). Es decir,

$$X \sim N(\mu,\sigma) \Rightarrow Z = \frac{X-\mu}{\sigma} \sim N(0,1)$$

A esta proceso de conversión de cualquier variable Normal en una Normal Estandar se le llama *tipificación*.

El proceso de tipificación que acabamos de ver se puede revertir, es decir, si a los valores de una variable Z que sigue una distribución $N(0,1)$, los multiplicamos por cualquier valor σ y a continuación les sumamos otro valor μ, los valores que se obtienen de esta transformación de la variable Z siguen una distribución $N(\mu,\sigma)$. Es decir,

$$Z \sim N(0,1) \Rightarrow X = \sigma \cdot Z + \mu \sim N(\mu,\sigma)$$

Ejemplo 2.12.

La longitud del fémur de cualquier feto a las 25 ***semanas de gestación sigue una distribución Normal con media*** $44mm$ ***y desviación típica*** $2mm$***. Si tomamos una embarazada al azar con*** 25 ***semanas de gestación ¿qué probabilidad tenemos de que el fémur de su feto mida más de*** $46mm$***? ¿y de que mida entre*** $47mm$ ***y*** $49mm$***?.***

Llamaremos X a la variable *longitud del fémur de un feto a las* 25 *semanas de gestación*. Como $X \sim N(44,2)$, tendremos que la variable $Z = \frac{X-44}{2}$seguirá una distribución Normal Estandar.
Calculamos la probabilidad de que al elegir un feto de 25 semanas de gestación al azar, su fémur mida más de 46:

$$P(X > 46) = P(\frac{X-44}{2} > \frac{46-44}{2}) = P(Z > 1{,}0) = 1 - 0{,}8413 = 0{,}1587$$

La respuesta es 0,1587 (o lo que es lo mismo, el 15,87 % de los fetos de 25 semanas de gestación tienen una longitud de fémur superior a $46mm$).
Respecto a la probabilidad de que el fémur mida entre $47mm$ y $49mm$, procedemos de la misma forma:

$$P(47 < X < 49) = P(\frac{47-44}{2} < \frac{X-44}{2} < \frac{49-44}{2}) =$$

$$= P(1{,}5 < Z < 2{,}5) = P(Z < 2{,}5) - P(Z < 1{,}5) = 0{,}9937 - 0{,}9331 = 0{,}0606$$

Por tanto la probabilidad de que cualquier feto tenga un fémur entre 47 y 49 milímetros será 0,0606 (un 6,06 %).

Ejemplo 2.13.

Siguiendo con la variable longitud de fémur de un feto a las 25 ***semanas de gestación del ejemplo anterior. Calcula también un intervalo (centrado en la media) que contenga el*** $80\,\%$ ***de los valores de longitud de fémur.***

Para calcular el intervalo, buscamos x_1 y x_2 que cumplan:

$$P(x_1 < X < x_2) = 0{,}80$$

Si operamos con esta expresión:

$$0{,}80 = P(x_1 < X < x_2) = P(\frac{x_1 - 44}{2} < \frac{X - 44}{2} < \frac{x_2 - 44}{2}) =$$

$$= P(z_1 < Z < z_2)$$

donde Z sigue una distribución $N(0,1)$, $z_1 = \frac{x_1 - 44}{2}$ y $z_2 = \frac{x_2 - 44}{2}$, y (z_1, z_2) es un intervalo que contiene el $80\,\%$ de los valores de la distribución $N(0,1)$. Este intervalo (z_1, z_2), lo podemos conocer, pues se trata de un intervalo de la $N(0,1)$, y concretamente este ha sido calculado en el ejemplo 2.11: es el intervalo $(-1{,}28, 1{,}28)$. Así:

$$z_1 = -1{,}28 \Rightarrow \frac{x_1 - 44}{2} = -1{,}28 \Rightarrow x_1 = (-1{,}28 \cdot 2) + 44 = 41{,}44$$

$$z_2 = 1{,}28 \Rightarrow \frac{x_2 - 44}{2} = 1{,}28 \Rightarrow x_2 = (1{,}28 \cdot 2) + 44 = 46{,}56$$

Por tanto la respuesta es $(41{,}44, 46{,}56)$, entre estos dos valores podemos encontrar el $80\,\%$ de las longitudes de fémur de fetos de 25 semanas de gestación.

Como conclusión general podemos establecer que la respuesta a cualquier pregunta sobre las probabilidades de una variable con distribución Normal no estandar pasará por la tipificación de la variable. Tras dicha tipificación podremos recurrir a la tabla de la distribución Normal Estandar, la principal herramienta de que disponemos para conocer el valor de la probabilidad que estamos buscando.

2.3. Ejercicios Capítulo 2

Ejercicio 2.1.

Consideramos que la variable Z sigue una distribución $N(0,1)$. Calcula las siguientes probabilidades:

1. $P(Z < 1{,}56)$
2. $P(Z < 2{,}78)$
3. $P(Z > 3{,}00)$
4. $P(Z > 1{,}01)$
5. $P(Z < -1{,}5)$
6. $P(Z > -2{,}61)$
7. $P(Z < -0{,}32)$
8. $P(Z > -1{,}63)$
9. $P(0{,}83 < Z < 1{,}64)$
10. $P(-1{,}25 < Z < 2{,}37)$
11. $P(-2{,}36 < Z < -1{,}33)$
12. Valor z_1 tal que $P(Z < z_1) = 0{,}648$
13. Valor z_1 tal que $P(Z < z_1) = 0{,}468$
14. Valor z_1 tal que $P(Z > z_1) = 0{,}9978$

Ejercicio 2.2.

Si disponemos de una variable $Z \sim N(0,1)$. Calcula:

1. Intervalo centrado en 0 que contenga entre sus valores una probabilidad de 0,90
2. Intervalo centrado en 0 que contenga entre sus valores una probabilidad de 0,95
3. Intervalo centrado en 0 que contenga entre sus valores una probabilidad de 0,99
4. Intervalo de la $N(0,1)$ que contiene el 95 % de los valores mayores, es decir, que deja fuera el 5 % de los valores menores.
5. Intervalo de la $N(0,1)$ que contiene el 95 % de los valores menores, es decir, que deja fuera el 5 % de los valores mayores.

Ejercicio 2.3.

De nuevo para la variable $Z \sim N(0,1)$, calcula los siguientes percentiles:

1. Percentil 50.

2. Percentil 75.
3. Percentil 90.
4. Percentil 10.
5. Percentil 25.

Ejercicio 2.4.

Se sabe que el peso de los niños de 1 año de edad sigue (aproximadamente) una distribución $N(7, 2)$ (en kg).

1. Calcula un intervalo centrado que cubra el peso del 95 % de los niños de 1 año.
2. Si acude a la clínica de un pediatra una madre con un niño de 1 año que pesa 10,5 kg. ¿En qué percentil se encuentra el niño en cuanto a peso?, es decir, ¿qué porcentaje de niños de esa edad pesa menos que él?
3. Calcula un intervalo centrado que contenga el 99 % de los valores del peso de niños de un año.
4. Calcula un intervalo que contenga el 90 % de los pesos más altos (dejando fuera el 10 % de los pesos más bajos)
5. ¿Qué porcentaje de niños de 1 año pesa menos de 3,5 kg?
6. ¿Qué porcentaje de niños de 1 año pesa más de 4,5 kg?
7. ¿Qué porcentaje de niños de 1 año pesa entre 6 y 8 kg? ¿y entre 8 y 9?, ¿y entre 4 y 5?

Ejercicio 2.5.

Se sabe que la estatura de los alumnos matriculados en primero de la universidad CEU-Cardenal Herrera (en centímetros) tiene una distribución $N(175, 8)$:

1. Calcula un intervalo para la estatura de los alumnos centrado en la media y que incluya al 95 % de éstos.
2. Calcula un intervalo para la estatura de los alumnos que contenga el 95 % de los alumnos de menor estatura.
3. Calcula un intervalo para la estatura de los alumnos que contenga el 95 % de los alumnos de mayor estatura.
4. Que porcentaje de alumnos mide más de 190cm.
5. Que porcentaje de alumnos mide menos de 182cm.
6. Que porcentaje de alumnos mide entre 170 y 185cm.

Ejercicio 2.6.

El diámetro máximo de los hematíes de una persona con malaria por Plamodium vivax presenta las siguientes características: Si la célula esta infectada dicha variable se distribuye de forma Normal con media 7,6 micras y desviación típica 0,81 micras, y si la célula no está infectada dicha variable se distribuye de forma Normal con media 9,6 micras y desviación típica 1,0 micras. Calcula:

1. Proporción de células no infectadas con un diámetro máximo mayor que 9,4 micras.
2. Proporción de células no infectadas con un diámetro máximo inferior a 7 micras.
3. Proporción de células infectadas con un diámetro máximo inferior a 9,4 micras.
4. Da un intervalo centrado que contenga el 95 % de las células infectadas y repite el proceso para las células no infectadas.

Ejercicio 2.7.

La longitud de un feto, en la semana 20 de gestación, sigue una distribución normal con media μ= 22.5 cm y desviación típica, σ= 2.85 cm.
a) Calcula la probabilidad de que, monitorizado un feto al azar, tenga una longitud inferior a 15 cm.
b) Calcula un intervalo centrado que contenga el 80 % de las longitudes de fetos con 20 semanas de gestación.
c) Calcula el percentil 25 e interpreta el resultado en el contexto del ejercicio.

Capítulo 3

Introducción a la Inferencia estadística

3.1. Población y muestra

Llamamos *población estadística*, *universo* o *colectivo* al conjunto de referencia del que extraemos las observaciones, es decir, el conjunto de todas las posibles unidades experimentales. Por más que nos refiramos muchas veces a este concepto como población, este conjunto no tiene que ser necesariamente un grupo de personas o animales (pensemos en las variables *Cantidad de plomo en el agua de las ciudades de una comunidad*, *Disposición de TAC en los hospitales españoles*, *Número de errores en las historias clínicas de un hospital*).

Llamamos *muestra* a un subconjunto de elementos de la población que habitualmente utilizaremos para realizar un estudio estadístico. Se suelen tomar muestras cuando es difícil, imposible o costosa la observación de todos los elementos de la población estadística, es decir, su uso se debe a que frecuentemente la población es demasiado extensa para trabajar con ella. El número de elementos que componen la muestra es a lo que llamamos *tamaño muestral* y se suele representar por la letra minúscula n.

Nuestro propósito será llegar a conocer ciertas características de la **población** a partir de la **muestra** que dispongamos. A este proceso le llamamos ***inferencia***

Ejemplo 3.1.

Estudio de enfermos renales

Si quisiéramos conocer las características de los enfermos renales en cuanto a *calidad de vida, tipo de tratamiento, edad de aparición de la enfermedad, sexo, variables que influyen en el éxito de un trasplante,...*, difícilmente podríamos acceder a todos y cada uno de los enfermos renales que existen (sería la *población* en estudio), pero posiblemente podríamos conseguir a través de algunos hospitales o centros de hemodiálisis los datos de una cantidad determinada de este tipo de enfermos (por ejemplo $n = 200$ enfermos). Nuestro objetivo no sería conocer las características de esos 200 enfermos en concreto, pero utilizaríamos el conocimiento sobre estos 200 enfermos para *obtener conclusiones* sobre todos los enfermos renales (nuestra *población* a estudio). Este proceso es lo que se conoce como *inferencia estadística*.

3.2. Muestreo y muestra aleatoria

El *muestreo* estudia la relación entre una población y las posibles muestras tomadas de ella. Podemos decir que el muestreo es el procedimiento de selección de una porción de la población para hacer inferencia sobre alguna de sus características. Para que a partir de una muestra, estudiemos las características de la población, es necesario que la muestra sea *representativa* de la misma, es decir, que mantenga aproximadamente y en la medida de lo posible las mismas características de interés que la población de estudio. Por ello es necesario cuidar la selección de la muestra, ya que no nos sirve cualquier forma de seleccionarla. El muestreo es una de las partes del análisis estadístico en el que habremos de ser particularmente cuidadosos.

Existen multitud de mecanismos para seleccionar una muestra que sea representativa de la población, y éstos dependen principalmente de los recursos disponibles y de la naturaleza de los elementos que componen la población. Hay dos preguntas *fundamentales* en la selección de una muestra:

- *¿Cuántos elementos debe tener la muestra?, es decir, ¿Cuál debe ser el tamaño de la misma?*
- *¿De qué forma seleccionamos esos elementos?*

A la primera pregunta la mejor respuesta siempre es: *cuantos más mejor*. Normalmente son los recursos disponibles para llevar a cabo el estudio o la población accesible la que limita este tamaño. Si queremos estudiar una población y lo vamos a hacer a partir de una muestra, es evidente que a mayor tamaño de la muestra más nos aproximamos a la población y por tanto podremos formular conclusiones más precisas acerca de la misma.

Respecto a la segunda pregunta, hay multitud de formas distintas de seleccionar la muestra, pero aquí comentaremos a grandes rasgos algunos tipos particulares de *muestreo*.

1. El primer tipo de muestreo que comentaremos es el que se conoce como *Muestreo aleatorio*. Este muestreo consiste en seleccionar los elementos que componen la muestra totalmente *al azar*. Este método supone que cualquier elemento de la población puede ser incluido en la muestra y que todos tienen exactamente la misma probabilidad de serlo. Se puede realizar o bien ayudándonos de una *tabla de números aleatorios* o bien mediante un *generador de números aleatorios (ordenador)*. En cualquier caso, sería necesario enumerar a todos los elementos de la población, y en algunos casos, la población ni siquiera es numerable (por ejemplo, en un estudio medioambiental, la selección de peces en un río). Por este motivo, en multitud de ocasiones este muestreo es *adaptado* para obtener un método que, en la medida de lo posible, se acerque a él (la selección de elementos en la muestra sea lo más aleatoria posible).

Ejemplo 3.2.

Estudio de la presión arterial en personas mayores de 65 ***años.***

Si quisiéramos estudiar la presión arterial media de las personas mayores de 65 años y queremos extraer una muestra de tamaño $n = 100$ (porque únicamente disponemos recursos económicos, materiales, personales,... para estudiar a este número de personas) mediante un muestreo aleatorio simplemente tendríamos que buscar un *censo* de todas estas personas y seleccionar a 100 de todas ellas totalmente al azar. (*Ésto es la teoría, ahora habría que buscar* ese censo *y tener en cuenta si querrían participar o no, pero aquí estamos estudiando la teoría, la práctica debería aproximarse, en la medida de lo posible, a esta teoría*)

2. El *Muestreo estratificado* se utiliza fundamentalmente cuando existe una variable categórica cuya influencia es determinante en los resultados del estudio o puede confundir los mismos (esta variable se llama *factor confusor*). La población es dividida en sub-poblaciones definidas por la categoría de la variable confusora y dentro de cada sub-población se toma una muestra aleatoria. El tamaño de cada una de las sub-muestras vendrá dado por el tamaño de cada sub-población en relación con el tamaño de la población total.

Ejemplo 3.3.

Continúa ejemplo **Presión Arterial** ***en mayores de 65 años.***

Siguiendo con este ejemplo, supongamos que es conocido que la hipertensión es más frecuente en hombres que en mujeres. Supongamos también que la población de estudio está compuesta por: 55 % mujeres-45 % de hombres. Si por *azar* en nuestra muestra de $n = 100$ personas obtuviéramos 50 hombres y 50 mujeres, los hombres estarían sobre-representados en nuestra muestra (puesto que en la población las proporciones de hombres-mujeres son $45\,\% - 55\,\%$) y como además, los hombres suelen tener con más frecuencia hipertensión, el nivel medio de la presión arterial que obtendríamos de nuestra muestra podría ser superior al nivel medio de la población (que es el valor al que nos gustaría acercarnos).
Como el factor confusor (*sexo*) tiene cierta influencia en la variable de interés (*presión arterial*), si quisiéramos controlar el posible efecto confusor de la misma podríamos realizar un *muestreo estratificado*. Este muestreo consistiría en: 1.- Partir la población de personas mayores de 65 años en dos sub-poblaciones: *hombres* y *mujeres*; 2.- Como la población total está compuesta por un 45 % de hombres y un 55 % de mujeres, de la sub-población de hombres extraeríamos un 45 % de los elementos de la muestra, y de la sub-población de mujeres extraeríamos el 55 % restante; 3.- Así, la muestra final estaría formada por 45 hombres seleccionados al azar de entre todos los hombres de la población y 55 mujeres seleccionadas al azar de entre todas las mujeres de la población.

3. Una tercera solución si el factor de confusión es numérico u ordinal será el *Muestreo sistemático*. En éste se ordena la muestra según los valores del factor confusor, y selecciona todos los individuos separados cierto número de posiciones entre sí (dentro de la muestra ordenada), tomando el primer elemento de forma aleatoria entre los primeros. De esta forma aseguramos que los valores que observaremos de la variable a estudiar corresponderán a todo el rango de valores del efecto confusor.

Ejemplo 3.4.

Continúa ejemplo* Presión Arterial *en mayores de 65 años.

Retomando de nuevo el ejemplo anterior, supongamos que es conocido que la hipertensión es más frecuente a medida que aumenta la *edad* de las personas. En este caso podría ser una variable confusora la variable *Edad*. Si por *azar* en nuestra muestra de $n = 100$ personas seleccionáramos más, o menos, personas mayores de los que hay proporcionalmente en la población, podríamos obtener una presión arterial media a partir de nuestra muestra que podría ser superior, o inferior respectivamente, al nivel medio de la población (que es a lo que nos gustaría acercarnos). Así, como la variable *edad* tiene cierta influencia en la variable de interés (*presión arterial*), si quisiéramos controlar su efecto confusor podríamos realizar un *muestreo sistemático* que consistiría en: 1.- Ordenar la población por la variable confusora, es decir, del de menor edad al de mayor edad; 2.- Si por ejemplo la población total está formada por 1000 personas y nosotros queremos seleccionar a 100, tendríamos que tomar una persona de cada 10; 3.- De entre las 10 primeras personas seleccionamos una al azar, y a partir de esa persona seleccionamos una cada 10. Así finalmente, la muestra estaría compuesta por 100 personas de todas las edades en la misma proporción aproximada que en la población.

3.3. Estadísticos, estimadores y parámetros

Un *estadístico* es una medida usada para describir alguna característica de una muestra (media, mediana, desviación típica,...) y un *parámetro* es una medida usada para describir las mismas características pero de la población (media, mediana, desviación típica,...). Cuando el *estadístico* se calcula en una muestra con idea de hacer inferencia sobre la misma característica en la población, se le llama *estimador*. La inferencia estadística pretende aproximarse a los parámetros de la población a partir de los estimadores de la muestra. Para distinguir los estimadores (valores muestrales) de los parámetros (valores poblacionales) los representaremos a partir de ahora con diferentes símbolos:

Característica	**Muestra (Estadístico)**	**Población (Parámetro)**
Variable Cuantitativa		
Media	$\overline{x}$	μ
Desviación típica	s	σ
Varianza	s^2	σ^2
Variable Cualitativa		
Porcentaje	$\widehat{P}$	P

3.4. Consistencia, insesgadez y precisión

Los estadísticos muestrales nos proporcionan información sobre los parámetros poblacionales correspondientes si la muestra se ha recogido correctamente. Hay dos características de los estadísticos que los hacen especialmente deseables:

- Diremos que un estadístico es un estimador *consistente* de un parámetro poblacional si al aumentar el tamaño de la muestra la diferencia entre el estadístico y el parámetro tiende a desaparecer.

- Diremos que un estadístico es un estimador *insesgado* de un parámetro poblacional si su valor esperado es igual a ese parámetro ($\mathrm{E}(\widehat{\theta})=\theta$). Es decir, a veces $\widehat{\theta}$ sobreestima el parámetro y otras veces lo subestima, pero del concepto de esperanza se deduce que si se repite muchas veces el método del muestreo, entonces, en promedio, el resultado es igual al parámetro poblacional.

- Diremos que un estimador es *preciso* si al calcular el estadístico para distintas muestras los valores de éste son parecidos.

Interesará que los estimadores que tomemos de los parámetros sean consistentes, insesgados y lo más precisos que se pueda.

Ejemplo 3.5.

Estimador consistente y preciso

Si tenemos una variable X que sigue una distribución $N(\mu, \sigma)$:
Se puede demostrar matemáticamente que $\overline{x}$ es un estimador consistente de μ. Significa que si tomamos muestras más y más grandes, las medias muestrales se aproximarán más y más a la media de la población (μ).
$\overline{x}$ es un estimador de μ más preciso que $X_{[1]}$ (primer valor de la muestra ordenada=mínimo). Si tomamos muestras de un tamaño considerable, la variabilidad que puede haber entre los mínimos de esas muestras, siempre será mayor que la que obtendremos de las diferentes medias muestrales.

3.5. Variación entre muestras

Si tomamos varias muestras aleatorias de cierta población, cualquier estimador tomará valores distintos para cada una de ellas. A esta variación en las estimaciones, efecto del azar sobre la elección de la muestra, se le llama *variación muestral*. La variación muestral dependerá de la variabilidad de la variable que tengamos y también del tamaño de la muestra.

Ejemplo 3.6.

Ejemplos de variación entre muestras

- Si tomamos distintas muestras de la temperatura corporal en población sana tendremos una variación muestral bastante baja (la variabilidad de esta variable es baja entre diferentes personas).

  Si tomamos la tensión arterial en la población española obtendremos una variación muestral bastante más elevada. (Variabilidad de esta variable alta).

- Si tomamos muestras de tamaño 10 y calculamos medias muestrales, por ejemplo, se parecerán entre ellas menos, que si las muestras que tomamos son de tamaño 1000. Es decir la variación muestral será en general más baja cuanto más grande sea la muestra utilizada para calcular nuestro estimador.

3.6. Distribución de estadísticos en el muestreo

3.6.1. Error estándar de la media muestral

El *Teorema Central del Límite* nos asegura que si nuestra muestra es *razonablemente grande* la distribución de la **media muestral** de cualquier variable sigue una distribución *Normal* y que además, la desviación típica de esta media tiene como expresión:

$$\frac{\sigma}{\sqrt{n}}$$

donde σ es la desviación típica de la variable original y n es el tamaño de la muestra. A la expresión anterior se le llama *error estándar* de la media.

Se entiende que el error estándar sería la desviación típica resultante de la obtención de las medias de distintas muestras aleatorias de la población. El error estándar será el efecto de la variabilidad muestral sobre el valor que obtenemos de la media en cada muestra, es decir la desviación típica de la media se conoce como error estándar.

Supongamos que tenemos una variable cuantitativa cualquiera X, cuya media en la población es μ y cuya desviación típica (también en la población) es σ. Si se toman varias muestras de tamaño *suficientemente grande* y llamamos $\overline{X}$ a la variable que guarda las medias muestrales para cada una de las muestras, por el Teorema Central del Límite tenemos asegurado:

$$\overline{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$

Ejemplo 3.7.

Comportamiento de las medias muestrales (de tamaño 50***) de una variable con media*** 10 ***y desviación típica*** 1,5***.***

Supongamos que tenemos una variable que en la población tiene media $\mu = 10$ y desviación típica $\sigma = 1{,}5$. Si el comportamiento de esta variable fuera aproximadamente *Normal*, la mayoría de valores de esta variable estarían alrededor del valor 10 más/menos dos desviaciones típicas por arriba y por abajo de este valor (es decir, entre $10 - 3 = 7$ y $10 + 3 = 13$ estarían la mayor parte de los valores de la variable)
¿Cómo se comportarían las medias muestrales si extrajéramos varias muestras de tamaño 50*?*
Pues según el *Teorema Central del Límite*, las medias muestrales seguirán una distribución *Normal* con media $\mu = 10$ y desviación típica $\frac{\sigma}{\sqrt{n}} = \frac{1{,}5}{\sqrt{50}} = \frac{1{,}5}{7{,}0711} = 0{,}2121$
Por tanto, las medias muestrales estarían alrededor del valor 10, pero con más/menos dos desviaciones típicas por arriba y por abajo (es decir, entre 9,5758 y 10,4242 estarían la mayor parte de las medias de las muestras). Así observamos que en general las medias muestrales son más precisas que las variables de las que provienen y serán más precisas cuantos más valores tengamos en nuestra muestra.

3.6.2. Error estándar de un porcentaje

En el caso de que la variable de interés sea una variable nominal no tiene sentido que nos planteemos el error estándar de su media (de hecho la media de una variable nominal no tiene tampoco sentido) sino el de su porcentaje de individuos en cada uno de sus valores. En este caso si P es el porcentaje de respuestas en ese valor su error estándar será:

$$\sqrt{\frac{P \cdot (100 - P)}{n}}$$

En la expresión anterior se ha supuesto que la variable P está expresada en tantos por 100, si estuviera expresada en tantos por uno (es decir P es un valor entre 0 y 1) únicamente habríamos de cambiar en ella el valor 100 por 1 y la expresión seguiría siendo válida.

Supongamos que tenemos una variable categórica y que nos interesa estimar el porcentaje de una de sus categorías en la población, al que llamamos P. Si tomamos varias muestras de tamaño *suficientemente grande* (n) y en cada una de esas muestras obtenemos una estimación del porcentaje de interés, si llamamos $\widehat{P}$ a la variable que guarda los porcentajes de esas muestras, se cumple que esta variable aleatoria sigue la siguiente distribución:

$$\widehat{P} \sim N(P, \sqrt{\frac{P \cdot (100 - P)}{n}})$$

3.6.3. Utilidad del Teorema Central del Límite

Ejemplo 3.8.

Se supone que el peso de los niños de un año de edad siguen una distribución normal de media $\mu = 10$ ***Kg y desviación típica*** σ***=2 Kg. Se extrae una muestra de 25 niños cuyo peso medio ha resultado ser*** $\overline{x}$***=12.5 Kg. A la vista del resultado, ¿parece cierto el supuesto de que el peso medio poblacional de los niños de un año de edad esté entorno a los 10 Kg?***

Solución:
Si la muestra de niños es representativa de la realidad (cosa que supuestamente es así), el peso medio muestral debería estar "cerca" del peso medio poblacional del que procede (10 kg). Ya que, por el Teorema Central del Límite, sabemos que $\overline{X} \sim N(10,\ 2/\sqrt{(25)})$ Es decir, la probabilidad de encontrar muestras de 25 niños con pesos medios muestrales superiores al observado (12.5) debería ser común ($\approx< 0.5$).
Hacemos el cálculo:
P($\overline{X}$ >12.5)=P(Z> $\frac{(12,5-10}{\frac{2}{\sqrt{25}}}$)=P(Z>6.25)≈1-1=0.
Es decir, 12.5 Kg es un peso medio extremadamente extraño si procede de la población $N(10,\ 2/\sqrt{(25)})$. Por lo tanto, podemos afirmar que el peso medio real de los niños de un año de edad es significativamente mayor que 10 Kg.

3.7. Ejercicios Capítulo 3

Ejercicio 3.1.

Se supone que la longitud de un feto, en la semana 20 de gestación, sigue una distribución normal con media $\mu = 23.5$ cm y desviación típica $\sigma = 2.85$ cm. Los resultados de las ecografías, en la semana 20 de gestación, de 9 mujeres dan las siguientes longitudes de feto:

21.9, 24.7, 15.0, 21.7, 25.9, 22.6, 23.5, 17.8, 22.1

Calcula $\mathrm{P}(\overline{X} < \overline{x})$ y razona si la muestra puede provenir de la población de longitudes descrita en el ejercicio.

Ejercicio 3.2.

Cierta empresa afirma que las baterías de las bombas de insulina que fabrica para suministrar a los hospitales, siguen una distribución normal con una duración media de 1.200 horas y una desviación típica de 400 horas. Supón que el hospital le compra a la empresa nueve bombas de insulina y que su duración media ha sido de 1050 horas. Calcula $\mathrm{P}(\overline{X} < 1050)$ ¿Qué conclusión deduces del resultado?.

Ejercicio 3.3.

Los creadores de un nuevo molino de viento afirman que puede generar una media de 800 kilovatios diarios de energía. Se supone que la generación diaria de energía sigue una distribución normal que tiene una desviación típica $\sigma = 120$ kilovatios. Se toma una muestra aleatoria de 100 días y se obtiene una media muestral de 768 kilovatios. a) Calcula la probabilidad de que la media muestral sea inferior a la observada. b) A la vista del resultado, que puedes comentar sobre la eficiencia anunciada por los creadores del molino.

Ejercicio 3.4.

En la memoria anual de cierta compañía de seguros, se estimó que el 15 % de sus pacientes afiliados necesitaron realizarse alguna prueba diagnóstica durante el año pasado. En el primer mes del año siguiente se considera una muestra de 100 pacientes de los cuales 16 necesitaron realizarse alguna prueba diagnóstica. Calcula $\mathrm{P}(\widehat{p} > 16\,\%)$. ¿Existen indicios para pensar que el próximo año aumentará significativamente el porcentaje de pacientes que se realizarán pruebas diagnósticas?

Capítulo 4

Intervalos de confianza

4.1. Intervalo de confianza

El proceso de inferencia es aquel mediante el cual se pretende estimar el valor de un parámetro a partir del valor de un estadístico. Esta estimación puede ser puntual o bien por intervalo. La mejor *estimación puntual* de un parámetro es simplemente el valor del estadístico correspondiente, pero es poco informativa porque la probabilidad de no dar con el valor correcto es muy elevada, es por eso que se acostumbra a dar una estimación por intervalo, en el que se espera encontrar el valor del parámetro con una elevada probabilidad. Esta estimación recibe el nombre de estimación mediante *intervalos de confianza*.

Ejemplo 4.1.

Algunos parámetros y sus estimadores puntuales

A continuación detallamos algunos parámetros y sus respectivos estimadores puntuales:

- μ representa la media poblacional de una variable cuantitativa y su estimador puntual es la media muestral $\overline{X}$
- σ representa la desviación típica poblacional de una variable cuantitativa y su estimador puntual es la desviación típica muestral S (de la misma forma, el estimador de la varianza poblacional σ^2 es la varianza muestral S^2)
- P representa el porcentaje de valores de una categoría de interés en una variable categórica y su estimador puntual es el porcentaje de esta característica en la muestra $\widehat{P}$

La estimación por *intervalos de confianza* consiste en determinar un posible rango de valores o intervalo (a, b), en el que, con una determinada probabilidad, sus límites contendrán el valor del parámetro poblacional que andamos buscando. Para cada muestra obtendremos un intervalo distinto que, para el X % de ellas, contendrá el verdadero valor del parámetro. A este intervalo se le denomina *intervalo de confianza*.

En este capítulo estudiaremos la estimación por intervalos de confianza para una *proporción* o *porcentaje* (P) en el caso de disponer de una variable categórica y la *media* (μ) cuando dispongamos de una variable cuantitativa.

Evidentemente esta técnica no tiene porqué dar siempre un resultado correcto, tal y como hemos comentado para algunas muestras el intervalo correspondiente contendrá el verdadero valor del parámetro y para otras no. A la probabilidad de que hayamos acertado al decir que el intervalo contiene al parámetro se la denomina *nivel de confianza* (o simplemente *confianza*). También se denomina nivel de significación a la probabilidad de errar en esta afirmación, es decir la significación (probabilidad de errar con nuestro

intervalo) será igual a 1-(nivel de confianza), ya que el nivel de confianza corresponde a la probabilidad de que el intervalo contenga el valor verdadero del parámetro.

Según se introdujo en el tema anterior la variabilidad muestral hace que al obtener varias muestras de la población y calcular los estadísticos sobre éstas (como media, desviación típica, varianza,...) obtengamos valores distintos para cada muestra, por tanto podemos hablar de la distribución de estos estadísticos en un conjunto de muestras, de la misma forma que hablamos de la distribución de cualquier otra variable aleatoria. El conocer las distribuciones de los estimadores anteriores nos permitirá asociar a cada muestra un intervalo de confianza para el parámetro poblacional correspondiente.

Concretamente, el objetivo de este curso es trabajar con la estimación de la media poblacional de una variable cuantitativa (μ) y la de el porcentaje de una característica de interés en la población a partir de una variable categórica (P). En general siempre querremos estimar cantidades poblacionales, por ejemplo μ, P, y no sus equivalentes muestrales $\overline{x}, \hat{P}$ ya que de estos últimos conoceremos sus valores exactos y en consecuencia no necesitan ser estimados (se conocen sin ningún tipo de ambigüedad). El reto que nos proponemos es, a partir de los valores muestrales, conocer tanto como sea posible los valores poblacionales. Para ello, utilizaremos las distribuciones de los correspondientes estimadores:

- **Intervalo de confianza para un porcentaje poblacional** P: utilizaremos la distribución en el muestreo del estadístico $\widehat{P}$

$$\widehat{P} \sim N(P, \sqrt{\frac{P \cdot (100 - P)}{n}}) \Rightarrow \frac{\widehat{P} - P}{\sqrt{\frac{P \cdot (100-P)}{n}}} \sim N(0, 1)$$

- **Intervalo de confianza para una media poblacional** μ: utilizaremos la distribución en el muestreo del estadístico $\overline{x}$

  1. Si la desviación típica poblacional σ es conocida podemos utilizar la expresión introducida en el tema 3

$$\overline{x} \sim N(\mu, \frac{\sigma}{\sqrt{n}}) \Rightarrow \frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$$

  2. Si la desviación típica poblacional σ es desconocida (que es lo habitual), y por tanto a lo sumo conoceremos S que es un estimador de σ. En ese caso, debemos introducir una nueva distribución llamada *Distribución t de Student*, pues la distribución del estadístico $\overline{x}$ cuando usamos la desviación típica muestral S es:

$$\frac{\overline{x} - \mu}{\frac{S}{\sqrt{n}}} \sim t_{n-1}$$

     donde t_{n-1} representa la distribución *t de Student* con $n-1$ grados de libertad. Esta distribución se estudiará en el siguiente punto de este tema.

4.2. Distribución t-Student

Cuando nos disponemos a hacer inferencia sobre la media poblacional (μ) a partir de la media muestral ($\overline{x}$), resulta lógico utilizar el Teorema Central del Límite, es decir, que

$$\frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$$

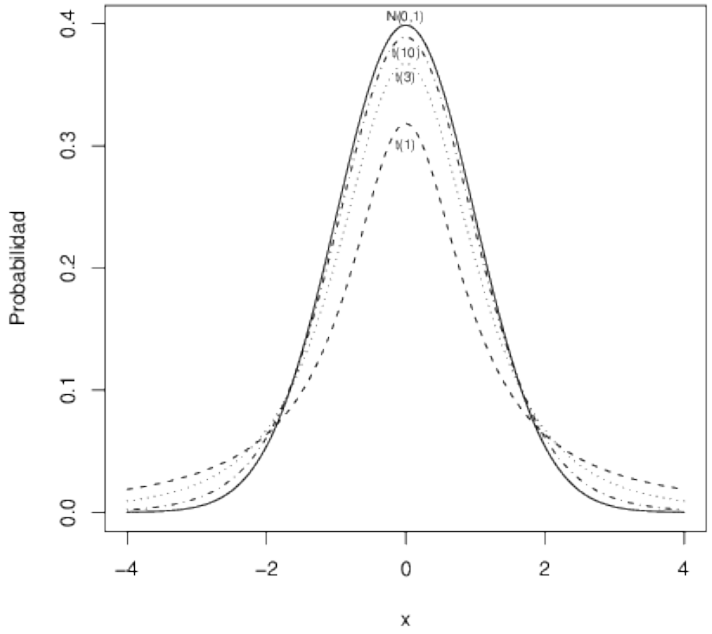
En esta expresión σ representa la desviación típica poblacional, de la que habitualmente no tendremos información sobre ella, es decir será un valor desconocido. Si tenemos una muestra de tamaño *suficientemente grande*, podemos estimar el valor de la desviación típica poblacional σ, a partir de la desviación típica muestral S con una precisión *aceptable*. Por tanto la expresión anterior seguirá siendo válida. Pero si la muestra que tenemos, no es *suficientemente grande*, la estimación que tendremos de σ a partir de S

no será lo suficientemente precisa, y por tanto la expresión anterior no será válida. En consecuencia, si σ no es conocida y el tamaño muestral que disponemos no es *suficientemente grande*, la expresión $\frac{\overline{x}-\mu}{\frac{S}{\sqrt{n}}}$, que es la que realmente usaremos para calcular el intervalo de confianza que pretendemos obtener, no seguirá una distribución $N(0,1)$ sino otra distribución similar (pero diferente), una distribución t *de Student*.

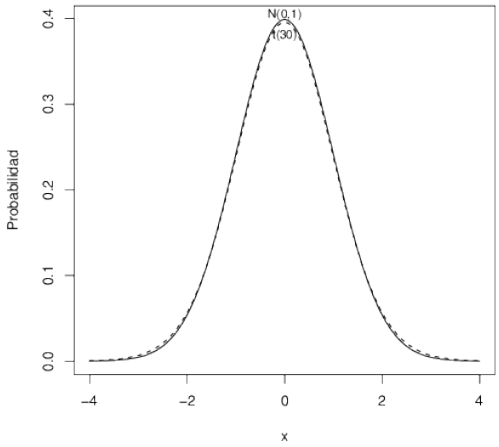
La distribución t *de Student* es una distribución con las siguientes características:

- Forma de campana.
- La máxima probabilidad se concentra alrededor del valor 0 (que es su media, moda y mediana) y disminuye a medida que nos alejamos de este valor central.
- Su forma se define por un parámetro g llamado *grados de libertad*, y que modula la mayor o menor variabilidad de los valores de esta distribución.

Como consecuencia de las características anteriores resulta que la distribución t tiene una forma muy similar a la distribución Normal Estándar, pero en función de los grados de libertad cambia su forma. A continuación se muestra la representación de varias distribuciones t con diferentes grados de libertad junto a una distribución Normal Estándar.



Podemos observar, que a medida que aumentan los grados de libertad, la distribución t se va aproximando a la distribución Normal estándar. En la siguiente figura representamos conjuntamente una distribución t *de Student* con 30 *grados de libertad* y la distribución Normal estándar. Tal y como se puede apreciar ambas distribuciones son prácticamente indistinguibles a nivel gráfico. Esto justifica que para tamaños de muestra superiores a 30 valores el intervalo de confianza que obtendríamos la distribución t serían prácticamente iguales que si empleamos una distribución Normal Estandar. Por tanto, cuando en el párrafo anterior decíamos que podríamos emplear la distribución Normal cuando el tamaño de la muestra que dispongamos sea *suficientemente grande*, en función de la siguiente figura podemos considerar que tamaños muestrales superiores a 30 unidades son suficientemente grandes, mientras que para tamaños muestrales menores sería más prudente utilizar la distribución t en lugar de la Normal.



A continuación reproducimos una tabla de la distribución t de la misma forma que hicimos para la distribución Normal Estándar. Cada fila de esta tabla se refiere a un número de grados de libertad diferente, que aparecen en la primera columna. A su vez cada una de las columnas de la tabla corresponde a un valor concreto de probabilidad. Para cada combinación de fila y columna la tabla reproduce aquel valor que para los grados de libertad correspondientes deja a su izquierda la probabilidad determinada por la columna a la que pertenece.

Tabla de probabilidades de la distribución *t de Student*
$[P(t < T)]$

$g \diagdown T$	0.650	0.700	0.750	0.800	0.850	0.900	0.950	0.9750	0.990	0.995
1	0.509	0.726	1.000	1.376	1.962	3.077	6.313	12.706	31.820	63.656
2	0.444	0.617	0.816	1.060	1.386	1.885	2.919	4.302	6.964	9.924
3	0.424	0.584	0.764	0.978	1.249	1.637	2.353	3.182	4.540	5.840
4	0.414	0.568	0.740	0.940	1.189	1.533	2.131	2.776	3.746	4.604
5	0.408	0.559	0.726	0.919	1.155	1.475	2.015	2.570	3.364	4.032
6	0.404	0.553	0.717	0.905	1.134	1.439	1.943	2.446	3.142	3.707
7	0.401	0.549	0.711	0.896	1.119	1.414	1.894	2.364	2.997	3.499
8	0.399	0.545	0.706	0.888	1.108	1.396	1.859	2.306	2.896	3.355
9	0.397	0.543	0.702	0.883	1.099	1.383	1.833	2.262	2.821	3.249
10	0.396	0.541	0.699	0.879	1.093	1.372	1.812	2.228	2.763	3.169
11	0.395	0.539	0.697	0.875	1.087	1.363	1.795	2.200	2.718	3.105
12	0.394	0.538	0.695	0.872	1.083	1.356	1.782	2.178	2.680	3.054
13	0.393	0.537	0.693	0.870	1.079	1.350	1.770	2.160	2.650	3.012
14	0.393	0.536	0.692	0.868	1.076	1.345	1.761	2.144	2.624	2.976
15	0.392	0.535	0.691	0.866	1.073	1.340	1.753	2.131	2.602	2.946
16	0.392	0.535	0.690	0.864	1.071	1.336	1.745	2.119	2.583	2.920
17	0.391	0.534	0.689	0.863	1.069	1.333	1.739	2.109	2.566	2.898
18	0.391	0.533	0.688	0.862	1.067	1.330	1.734	2.100	2.552	2.878
19	0.391	0.533	0.687	0.860	1.065	1.327	1.729	2.093	2.539	2.860
20	0.390	0.532	0.686	0.859	1.064	1.325	1.724	2.085	2.527	2.845
21	0.390	0.532	0.686	0.859	1.062	1.323	1.720	2.079	2.517	2.831
22	0.390	0.532	0.685	0.858	1.061	1.321	1.717	2.073	2.508	2.818
23	0.390	0.531	0.685	0.857	1.060	1.319	1.713	2.068	2.499	2.807
24	0.389	0.531	0.684	0.856	1.059	1.317	1.710	2.063	2.492	2.796
25	0.389	0.531	0.684	0.856	1.058	1.316	1.708	2.059	2.485	2.787
26	0.389	0.530	0.684	0.855	1.057	1.314	1.705	2.055	2.478	2.778
27	0.389	0.530	0.683	0.855	1.056	1.313	1.703	2.051	2.472	2.770
28	0.389	0.530	0.683	0.854	1.055	1.312	1.701	2.048	2.467	2.763
29	0.389	0.530	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756
30	0.389	0.530	0.682	0.853	1.054	1.310	1.697	2.042	2.457	2.749
40	0.388	0.528	0.680	0.850	1.050	1.303	1.683	2.021	2.423	2.704
60	0.387	0.527	0.678	0.847	1.045	1.295	1.670	2.000	2.390	2.660
120	0.386	0.525	0.676	0.844	1.040	1.288	1.657	1.979	2.357	2.617
∞	0.385	0.524	0.674	0.841	1.036	1.281	1.644	1.959	2.326	2.575

4.3. Intervalo de confianza para una media

En este caso estaremos interesados en encontrar un procedimiento para calcular el intervalo para el parámetro μ que, en caso de disponer de varias muestras de la población, contendría el verdadero valor del parámetro cierto porcentaje de veces (confianza del intervalo). La confianza del intervalo se suele representar como del $(1-\alpha)\,\%$, es decir, en adelante asumiremos que α representa la proporción de muestras para las que el intervalo que calculemos no contendrá el verdadero valor del parámetro.

4.3.1. Intervalo de confianza para una media: desviación típica poblacional conocida

Tal y como hemos señalado anteriormente, en caso conocer la desviación típica de la población tenemos garantizadas las siguientes relaciones:

$$\overline{x} \sim N(\mu, \frac{\sigma}{\sqrt{n}}) \Rightarrow \frac{\overline{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

Según vimos en el tema 2, para cualquier variable que siga una distribución $N(0,1)$, si queremos hallar un intervalo (centrado en 0) que contenga el $100\cdot(1-\alpha)\,\%$ de los valores de la variable podemos delimitarlo mediante los valores el $Z_{1-\frac{\alpha}{2}}$ y $Z_{\frac{\alpha}{2}}(=-Z_{1-\frac{\alpha}{2}})$, que nos proporciona la tabla de la distribución Normal, donde $Z_{1-\frac{\alpha}{2}}$ es el valor de la distribución $N(0,1)$ que cumple que el $100\cdot(1-\frac{\alpha}{2})\,\%$ de los valores de esta distribución son inferiores a él. En consecuencia tenemos que el intervalo:

$$-Z_{1-\frac{\alpha}{2}} \leq \frac{\overline{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq Z_{1-\frac{\alpha}{2}}$$

contiene el $100\cdot(1-\alpha)\,\%$ de los valores que podría tomar $\frac{\overline{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$. Aplicando la aritmética de variables normales que también introducimos en el Tema 2, tenemos:

$$-Z_{1-\frac{\alpha}{2}}\cdot\frac{\sigma}{\sqrt{n}} \leq \overline{x}-\mu \leq Z_{1-\frac{\alpha}{2}}\cdot\frac{\sigma}{\sqrt{n}}$$

y

$$-Z_{1-\frac{\alpha}{2}}\cdot\frac{\sigma}{\sqrt{n}}-\overline{x} \leq -\mu \leq Z_{1-\frac{\alpha}{2}}\cdot\frac{\sigma}{\sqrt{n}}-\overline{x}$$

Finalmente,

$$\overline{x}-Z_{1-\frac{\alpha}{2}}\cdot\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{x}+Z_{1-\frac{\alpha}{2}}\cdot\frac{\sigma}{\sqrt{n}}$$

Por tanto, el intervalo en el que μ estará contenido con un $100\cdot(1-\alpha)\,\%$ de confianza es:

$$[\overline{x}-Z_{1-\frac{\alpha}{2}}\cdot\frac{\sigma}{\sqrt{n}}\ ,\ \overline{x}+Z_{1-\frac{\alpha}{2}}\cdot\frac{\sigma}{\sqrt{n}}]$$

De esta forma hemos obtenido la expresión del intervalo de confianza al $100\cdot(1-\alpha)\,\%$ para la media poblacional μ.

Ejemplo 4.2.

En un estudio se pretende estimar la edad media a la que se diagnostica la* Diabetes Mellitus *en la Comunitat Valenciana. Para ello se dispone de una muestra de* 21 *pacientes a los que se les ha preguntado la edad de diagnóstico de la enfermedad. A partir de estos* 21 *pacientes se ha obtenido una edad media de* 48,78 *años. Si es conocido, a raíz de otros estudios, que la desviación típica de esta variable (Edad de diagnóstico de la enfermedad***) es* 16,32*, calcula un intervalo de confianza al* 95 % *para la edad media de diagnóstico de esta enfermedad en la región de estudio.***

Datos para realizar la estimación: $n = 21$, $\overline{x} = 48{,}78$ y $\sigma = 16{,}32$
Como queremos obtener un intervalo con un 95 % de confianza, tenemos $1 - \alpha = 0{,}95$, y por tanto $\alpha = 0{,}05$ y $\frac{\alpha}{2} = 0{,}025$, así, $(1 - \frac{\alpha}{2}) = 0{,}975$. Por tanto debemos buscar el valor de la Normal estándar que cumple que el 97,5 % de los valores son inferiores a él. Este valor de la $N(0, 1)$ es $Z_{1-\frac{\alpha}{2}} = 1{,}96$, y por tanto, el intervalo es:

$$\left[48{,}78 - 1{,}96 \cdot \frac{16{,}32}{\sqrt{21}} \ , \ 48{,}78 + 1{,}96 \cdot \frac{16{,}32}{\sqrt{21}}\right] =$$

$$[41{,}79 \ , \ 55{,}76]$$

Con un 95 % de confianza, la edad media a la que se diagnostica la *Diabetes Mellitus* en la Comunitat Valenciana será un valor contenido en el intervalo [41,79 , 55,76].

4.3.2. Intervalo de confianza para una media: desviación típica poblacional desconocida

Tal y como hemos señalado anteriormente, en caso de no conocer la desviación típica de la población no podremos conocer el valor de la siguiente expresión:

$$\frac{\overline{x} - \mu}{\sigma/\sqrt{n}}$$

sino que habremos de aproximarla por:

$$\frac{\overline{x} - \mu}{S/\sqrt{n}}$$

que según hemos visto sigue una distribución t con $n-1$ grados de libertad. Conociendo dicha distribución podremos proceder de manera análoga a como lo hemos hecho cuando conocíamos el valor de la desviación típica poblacional. Es decir, si $t_{(n-1,1-\frac{\alpha}{2})}$ es el valor de la distribución *t de Student* con $n-1$ *grados de libertad* que cumple que el $(1 - \frac{\alpha}{2}) \cdot 100\,\%$ de los valores de esta distribución son inferiores a él, entonces tenemos que con una confianza del $(1 - \alpha) \cdot 100\,\%$ se darán las siguientes desigualdades:

$$-t_{(n-1,1-\frac{\alpha}{2})} \leq \frac{\overline{x} - \mu}{\frac{S}{\sqrt{n}}} \leq t_{(n-1,1-\frac{\alpha}{2})}$$

Por tanto, aplicando la aritmética de variables vista en el Tema 2:

$$-t_{(n-1,1-\frac{\alpha}{2})} \cdot \frac{S}{\sqrt{n}} \leq \overline{x} - \mu \leq t_{(n-1,1-\frac{\alpha}{2})} \cdot \frac{S}{\sqrt{n}}$$

y

$$t_{(n-1,1-\frac{\alpha}{2})} \cdot \frac{S}{\sqrt{n}} - \overline{x} \leq -\mu \leq t_{(n-1,1-\frac{\alpha}{2})} \cdot \frac{S}{\sqrt{n}} - \overline{x}$$

Finalmente,

$$\overline{x} - t_{(n-1,1-\frac{\alpha}{2})} \cdot \frac{S}{\sqrt{n}} \leq \mu \leq \overline{x} + t_{(n-1,1-\frac{\alpha}{2})} \cdot \frac{S}{\sqrt{n}}$$

Por tanto, el intervalo en el que μ estará contenido con un $100 \cdot (1-\alpha)\,\%$ de confianza es:

$$\left[\overline{x} - t_{(n-1,1-\frac{\alpha}{2})} \cdot \frac{S}{\sqrt{n}} \;\;,\;\; \overline{x} + t_{(n-1,1-\frac{\alpha}{2})} \cdot \frac{S}{\sqrt{n}}\right]$$

Ejemplo 4.3.

En un estudio se pretende estimar la edad media a la que se diagnostica la* Diabetes Mellitus *en la Comunitat Valenciana. Para ello se dispone de una muestra de* 21 *pacientes a los que se les ha preguntado la edad de diagnóstico de la enfermedad. A partir de estos* 21 *pacientes se ha obtenido una edad media de* 48,78 *años y una desviación típica de* 16,32*. Calcula un intervalo de confianza al* 95 % *para la edad media de diagnóstico de esta enfermedad en la región de estudio.

Tenemos como datos para realizar la estimación: $n = 21$, $\overline{x} = 48{,}78$ y $S = 16{,}32$.
Como queremos obtener un intervalo con un 95 % de confianza, tenemos $1-\alpha = 0{,}95$, y por tanto $\alpha = 0{,}05$ y $\frac{\alpha}{2} = 0{,}025$, así, $(1-\frac{\alpha}{2}) = 0{,}975$ y debemos buscar el valor de la distribución *t de Student* con $n-1 = 20$ *grados de libertad* que cumple que el 97,5 % de los valores son inferiores a él. Este valor de la *t de Student* es $t_{(20,0{,}975)} = 2{,}085$, y por tanto, el intervalo que queríamos calcular tomará la siguiente expresión:

$$\left[48{,}78 - 2{,}085 \cdot \frac{16{,}32}{\sqrt{21}} \;\;,\;\; 48{,}78 + 2{,}085 \cdot \frac{16{,}32}{\sqrt{21}}\right] =$$
$$[41{,}35 \;\;,\;\; 56{,}20]$$

Con un 95 % de confianza, la edad media a la que se diagnostica la *Diabetes Mellitus* en la Comunitat Valenciana estará contenida en el intervalo [41,35 , 56,20], es decir, entre 41 y 56 años aproximadamente.

Por tanto a modo de resumen, el cálculo de un intervalo de confianza para una media poblacional se calcula como se indica a continuación:

Parámetro a estimar	Estimación puntual	Desviación típica poblacional (σ) conocida	Intrvalo de confianza al $100 \cdot (1-\alpha)\,\%$ para μ
μ	$\overline{x}$	Sí	$\left[\overline{x} - Z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \;\;,\;\; \overline{x} + Z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right]$
μ	$\overline{x}$	No	$\left[\overline{x} - t_{(n-1,1-\frac{\alpha}{2})} \cdot \frac{S}{\sqrt{n}} \;\;,\;\; \overline{x} + t_{(n-1,1-\frac{\alpha}{2})} \cdot \frac{S}{\sqrt{n}}\right]$

4.4. Intervalo de confianza para un porcentaje

En el caso de disponer de una variable cualitativa, en la que su media no tiene demasiado sentido, suele ser habitual plantearse el cálculo del intervalo de confianza para el porcentaje de individuos en cada una de sus categorías. En esta sección nos ocuparemos del caso en que dispongamos de una variable binaria y queramos hacer inferencia sobre el porcentaje de dicha característica en la población (P), a partir del porcentaje de esa misma característica en nuestra muestra ($\widehat{P}$).

Calculemos el intervalo para P con nivel de confianza $100 \cdot (1-\alpha)\,\%$. La distribución del estadístico $\widehat{P}$ es:

$$\widehat{P} \sim N(P, \sqrt{\frac{P \cdot (100-P)}{n}})$$

por tanto si tipificamos dicha variable resulta:

$$\frac{\widehat{P} - P}{\sqrt{\frac{P \cdot (100-P)}{n}}} \sim N(0,1)$$

si n es razonablemente grande, de la misma forma que sucedía para la media de poblaciones normales, podremos aproximar la desviación típica de la expresión anterior tomando $P = \widehat{P}$ en ella. En ese caso tenemos:

$$\frac{\widehat{P} - P}{\sqrt{\frac{\widehat{P}\cdot(100-\widehat{P})}{n}}} \sim N(0,1)$$

así, procediendo de forma análoga al caso de la media de una población normal tenemos que, con una confianza del $100 \cdot (1-\alpha)\,\%$, el estadístico anterior se hallará en el siguiente intervalo:

$$-Z_{1-\frac{\alpha}{2}} \leq \frac{\widehat{P} - P}{\sqrt{\frac{\widehat{P}\cdot(100-\widehat{P})}{n}}} \leq Z_{1-\frac{\alpha}{2}}$$

por tanto, aplicando la aritmética de variables tenemos:

$$-Z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\widehat{P}\cdot(100-\widehat{P})}{n}} \leq \widehat{P} - P \leq Z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\widehat{P}\cdot(100-\widehat{P})}{n}}$$

y

$$-Z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\widehat{P}\cdot(100-\widehat{P})}{n}} - \widehat{P} \leq -P \leq Z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\widehat{P}\cdot(100-\widehat{P})}{n}} - \widehat{P}$$

Finalmente,

$$\widehat{P} - Z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\widehat{P}\cdot(100-\widehat{P})}{n}} \leq P \leq \widehat{P} + Z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\widehat{P}\cdot(100-\widehat{P})}{n}}$$

Por tanto, el intervalo en el que P estará contenido con un $(1-\alpha)\times 100\,\%$ de confianza será:

$$\left[\widehat{P} - Z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\widehat{P}\cdot(100-\widehat{P})}{n}} \;\;,\;\; \widehat{P} + Z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\widehat{P}\cdot(100-\widehat{P})}{n}}\right]$$

Ejemplo 4.4.

Un estudio pretende estimar el porcentaje de hipertensos que hay entre las personas mayores de 65 años en la Comunidad Valenciana. Además de una estimación puntual de este porcentaje, interesa obtener un intervalo de confianza al $95\,\%$ ***para este parámetro de la población (*** P***). Para llevar a cabo este estudio, han sido seleccionadas*** 350 ***personas mayores de 65 años en toda la Comunidad, resultando tras realizar las pruebas correspondientes que 167 padecen de hipertensión.***

P=% de hipertensos entre las personas mayores de 65 años en la Comunidad Valenciana. $n = 350$
Estimador puntual: $\widehat{P} = \frac{167}{350} \cdot 100\,\% = 47{,}71\,\%$
Como queremos obtener un intervalo con un $95\,\%$ de confianza, entonces $1-\alpha = 0{,}95$ y por tanto $\alpha = 0{,}05$. Así, $(1-\frac{\alpha}{2}) = 0{,}975$ y en consecuencia debemos buscar el valor de la Normal estándar que cumple que el $97{,}5\,\%$ de los valores son inferiores a él. Este valor de la $N(0,1)$ es $Z_{1-\frac{\alpha}{2}} = 1{,}96$, y por tanto, el intervalo que buscamos es:

$$\left[47{,}71 - 1{,}96 \cdot \sqrt{\frac{47{,}71\cdot(100-47{,}71)}{350}} \;\;,\;\; 47{,}71 + 1{,}96 \cdot \sqrt{\frac{47{,}71\cdot(100-47{,}71)}{350}}\right] =$$

$$[42{,}48 \;\;,\;\; 52{,}94]$$

Así, con un $95\,\%$ de confianza, el porcentaje de hipertensos entre las personas mayores de 65 años en la Comunidad Valenciana estará contenido en el intervalo $[42{,}48\;,\;52{,}94]$, es decir, aproximadamente entre el $42{,}5\,\%$ y $53\,\%$ de la población.

4.5. Cálculo del tamaño muestral para obtener un error de estimación prefijado

En ocasiones, antes de comentar un estudio, nos planteamos cuál es el tamaño que debe tener la muestra que vamos a seleccionar. La respuesta estadística siempre es *"lo más grande posible"*. Sin embargo, cuando tenemos un objetivo concreto, como cometer un error no mayor de un umbral determinado, es posible calcular el tamaño muestra necesario para cumplir ese requisito con un nivel de confianza $(1-\alpha)$ determinado.

4.5.1. Tamaño muestral necesario para la estimación de una media poblacional con un error determinado.

En esta sección nos plantearemos el cálculo del tamaño muestral necesario para estimar una media poblacional (μ) con un error máximo e. Para ello tendremos que fijar previamente el nivel de confianza con el que queremos trabajar $(1-\alpha)\times 100\,\%$ y conocer (o tener una estimación aproximada a partir de estudios previos o una pre-muestra) de la desviación típica poblacional σ.

Sabemos, que la fórmula para hallar el intervalo de confianza para una media poblacional con desviación típica poblacional conocida y con una confianza del $(1-\alpha)\times 100\,\%$ es:

$$\left[\overline{x} - Z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \ , \ \overline{x} + Z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right]$$

Considerando que el error de la estimación (e) es la amplitud del intervalo, queremos:

$$e \leq 2 \cdot Z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

Así, despejando de esta expresión, podemos obtener:

$$n \geq \left(\frac{2 \cdot Z_{1-\frac{\alpha}{2}} \cdot \sigma}{e}\right)^2 = \frac{4 \cdot (Z_{1-\frac{\alpha}{2}})^2 \sigma^2}{e^2}$$

Ejemplo 4.5.

Supongamos que queremos estimar el nivel medio de hemoglobina para los pacientes oncológicos sometidos a tratamiento de Quimioterapia. Supongamos también que queremos obtener esta estimación con un error máximo de $0{,}2$ ***unidades y que queremos trabajar con una confianza del 95 %. Como no disponemos, como es habitual, del valor de la desviación típica de esta variable en la población, hemos tomado una pre-muestra de esta población y hemos obtenido una desviación típica de esta pre-muestra de*** $0{,}6$***. Calcula el tamaño muestral necesario para llevar a cabo la estimación bajo las exigencias anteriores.***

Partimos de los siguientes datos:

$$e = 0{,}2; \qquad 1-\alpha = 0{,}95 \Rightarrow 1 - \frac{\alpha}{2} = 0{,}975 \Rightarrow Z_{1-\frac{\alpha}{2}} = 1{,}96; \qquad \sigma \approx 0{,}6$$

Aplicando la fórmula anterior obtenemos:

$$n \geq \frac{4 \cdot (1{,}96)^2 (0{,}6)^2}{(0{,}2)^2} = 138{,}3 \approx 139 \ \ \textit{pacientes}$$

4.5.2. Tamaño muestral necesario para la estimación de un porcentaje poblacional con un error determinado.

En esta sección nos plantearemos, de forma similar a la sección anterior, el cálculo del tamaño muestral necesario para estimar un porcentaje poblacional (P) con un error máximo e. Para ello, también debemos fijar previamente el nivel de confianza con el que queremos trabajar $(1-\alpha)\times 100\,\%$ y tener una estimación aproximada, a partir de estudios previos o una pre-muestra, de la magnitud del porcentaje que queremos estimar (si estará alrededor del $10\,\%$, $35\,\%$, $50\,\%$....). Si no tenemos esta información nos pondremos en el peor de los casos, es decir, en el que tiene una estimación con mayor variabilidad, que coincide con $P \approx 50\,\%$.

Sabemos, que la fórmula para hallar el intervalo de confianza para una media poblacional con desviación típica poblacional conocida y con una confianza del $(1-\alpha)\times 100\,\%$ es:

$$\left[\widehat{P} - Z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\widehat{P}\cdot(100-\widehat{P})}{n}} \;\;,\;\; \widehat{P} + Z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\widehat{P}\cdot(100-\widehat{P})}{n}}\right]$$

donde en la expresión $\sqrt{\frac{\widehat{P}\cdot(100-\widehat{P})}{n}}$ hemos aproximado por el valor de $\widehat{P}$ el verdadero porcentaje poblacional P.

Considerando que el error de la estimación (e) es, como en el caso anterior, la amplitud del intervalo, queremos:

$$e \leq 2 \cdot Z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{P\cdot(100-P)}{n}}$$

Así, despejando de esta expresión, podemos obtener:

$$n \geq \frac{4\cdot(Z_{1-\frac{\alpha}{2}})^2 \cdot P\cdot(100-P)}{e^2}$$

Ejemplo 4.6.

Supongamos que queremos estimar el porcentaje de niños menores de 15 años que tienen alguna caries en sus dientes definitivos en la Comunidad Valenciana. Supongamos también que queremos obtener esta estimación con un error máximo de un $6\,\%$ ***y que queremos trabajar con una confianza del 95 %. Si dispusiéramos de una estimación previa de este porcentaje (o bien por estudios previos en otras comunidades, o bien obtenido a partir de una pre-muestra lo utilizaríamos como valor de*** P***. Supongamos que en esta ocasión no es así, no tenemos ninguna idea previa sobre el valor que toma este porcentaje, así que nos pondremos en el peor de los casos y supondremos que nuestro porcentaje está alrededor del*** $50\,\%$***. Calcula el tamaño muestral necesario para llevar a cabo la estimación bajo las exigencias anteriores.***

Partimos de los siguientes datos:

$$e = 6\,\%; \qquad 1-\alpha = 0{,}95 \Rightarrow 1-\frac{\alpha}{2} = 0{,}975 \Rightarrow Z_{1-\frac{\alpha}{2}} = 1{,}96; \qquad P \approx 50\,\%$$

Aplicando la fórmula anterior obtenemos:

$$n \geq \frac{4\cdot(1{,}96)^2\cdot 50\cdot(100-50)}{6^2} = 1067{,}1 \approx 1067 \;\; ninos$$

Ejemplo 4.7.

Consideremos, con el mismo escenario que en ejemplo anterior, que sí tenemos una estimación previa, por un estudio realizado en la Comunidad de Galicia, del porcentaje aproximado de niños en esas edades con alguna caries en sus dientes definitivos. En esa región se ha obtenido un porcentaje del $10\,\%$

En este caso, aunque no conozcamos ese porcentaje en nuestra región de estudio, cabría pensar que no distaría muchísimo del que se ha obtenido en Galicia y, aprovechando esta estimación, podríamos considerar que el porcentaje que queremos estimar estará alrededor del valor $P \approx 10\,\%$. El tamaño muestral necesario cambiaría sustancialmente, tal y como se indica a continuación:

$$n \geq \frac{4 \cdot (1{,}96)^2 \cdot 10 \cdot (100-10)}{6^2} = 384{,}2 \approx 384 \ \ ninos$$

4.6. Ejercicios Capítulo 4

Para todos los problemas que se proponen a continuación reflexiona sobre cuál es en cada uno de ellos:

- Población en estudio y muestra
- Variable en estudio y tipo de la misma (cuantitativa o cualitativa)
- Parámetro de interés para el que se calcula el intervalo de confianza
- Interpretación del intervalo de confianza obtenido.

Ejercicio 4.1.

Se llevó a cabo un experimento diseñado para estimar el número medio de latidos por minuto del corazón en los niños de 5 años de edad. A partir de una muestra aleatoria de 49 niños de 5 años se encontró que el número medio de latidos por minuto era de 90 pulsaciones. Si podemos considerar que la varianza de esta variable en la población en estudio es $\sigma^2 = 100$, calcula un intervalo de confianza al 95 % para el número medio de latidos por minuto de esta población e interpreta el resultado.

Ejercicio 4.2.

Un estudio se planteó como objetivo comparar la capacidad física de niños de 7 y 9 años. Con este fin se diseñó una prueba que ponía a prueba la capacidad física de los niños y se fijaron los criterios que determinaban si cada niño había superado la prueba física o no. Se sometió a dicha prueba a una muestra de 21 niños de 7 años, de los cuales 6 consiguieron superarla. Por otro lado, aplicando la prueba a una muestra aleatoria de 16 niños de 9 años se obtuvo que 12 de ellos también consiguieron pasarla. Calcula un intervalo de confianza al 95 % para el porcentaje de niños de 7 años que pueden pasar la prueba en la población y otro intervalo con la misma confianza para los niños de 9 años. Interprétalos y explica qué conclusiones podrías obtener si comparas ambos intervalos.¿Crees que existen diferencias significativas entre ambos grupos?

Ejercicio 4.3.

Se obtuvieron, a partir de una muestra de 15 hombres adultos físicamente activos, los siguientes valores de la *presión parcial de oxígeno en la sangre (PaQ2)* en reposo:

$$75,\ 80,\ 84,\ 74,\ 84,\ 78,\ 89,\ 72,\ 83,\ 76,\ 75,\ 87,\ 78,\ 79,\ 88$$

A partir de estos datos, calcula el intervalo de confianza al 95 por ciento para el nivel medio de esta variable de la población e interpreta su resultado.

Ejercicio 4.4.

El coordinador de un centro de salud estaba interesado en estimar el tiempo medio que los pacientes pasan en la sala de espera entre sus registros en admisión y su atención por un miembro del equipo médico. Para ello, seleccionó aleatoriamente una muestra de 100 pacientes, a partir de los cuales obtuvo un tiempo medio de permanencia en la sala de espera de 23 minutos y una desviación típica de 10 minutos. Calcula un intervalo de confianza al 90 % para el tiempo medio de espera en el centro de salud e interpreta el resultado.

Ejercicio 4.5.

En un estudio sobre la duración de la hospitalización realizado por varios hospitales en cooperación, se extrajo al azar una muestra aleatoria de 64 pacientes con úlcera péptica de una lista de todos los pacientes con esta enfermedad admitidos alguna vez en los hospitales. A partir de esta muestra se determinó, para cada uno, la duración de su hospitalización. Se encontró que la duración media de la hospitalización de esta muestra fue de 8,25 días. Si se sabe por otros estudios previos que la desviación típica del tiempo medio de hospitalización para esta población es de $\sigma = 3$ días, halla el intervalo de confianza al 90 % para la duración de la hospitalización media poblacional e interpreta su resultado.

Ejercicio 4.6.

Un proyecto de investigación se plantea llevar a cabo un estudio actualizado de caracterización de los niños de diez años, en el que una de las variables de interés es su peso. A partir de una muestra de 25 niños de diez años de edad se obtuvo un peso medio y una desviación típica de 36,5 y 5 kg respectivamente. Con estos datos, halla un intervalo de confianza al 90 % para el peso medio de niños de diez años de la población e interpreta su resultado.

Ejercicio 4.7.

Repite ahora el problema anterior, pero considerando que la desviación típica que conocemos es la de la población $\sigma = 5$. Obtén el intervalo de confianza al 90 % para el peso medio de niños de diez años de la población y compara el resultado con el obtenido en el ejercicio anterior.

Ejercicio 4.8.

Los siguientes valores son las concentraciones de bilirrubina en suero de una muestra de 10 pacientes admitidos a un hospital para el tratamiento de la hepatitis:

20,5, 14,8, 21,3, 12,7, 15,2, 26,6, 23,4, 22,9, 15,7, 19,2

Con estos valores construye un intervalo de confianza al 95 % para la concentración media de bilirrubina en suero de este tipo de pacientes e interpreta el resultado.

Ejercicio 4.9.

Un encargado del archivo de expedientes médicos de un hospital se planteó llevar a cabo un estudio sobre la calidad de la información de los expedientes de los pacientes del hospital. Para ello extrajo al azar una muestra de 100 expedientes de pacientes y encontró que en el 8 por ciento de ellos la carátula tenía, al menos, un detalle de información que contradecía el resto de la información que aparecía en el expediente. Construye un intervalos de confianza al 99 % para el porcentaje de los expedientes que contienen dichas discrepancias.

Ejercicio 4.10.

En un equipo de rehabilitación estaban diseñando una actividad destinada a pacientes con una determinada incapacidad física. Antes de ponerla a prueba necesitaban una estimación del tiempo medio que este tipo de pacientes requeriría para realizar la actividad. Con este objetivo se expuso la actividad en prueba a 9 pacientes que padecían la incapacidad física en estudio y se les pidió que la llevaran a cabo como parte de un experimento. El tiempo promedio requerido por estos pacientes para realizar la tarea fue de 7 minutos con una desviación típica de 2 minutos. Construye un intervalo de confianza 95 % para el tiempo medio requerido para que este tipo de pacientes efectúe la tarea e interpreta el resultado.

Ejercicio 4.11.

Una muestra de 100 hombres adultos aparentemente normales, de 25 años de edad, mostró una presión sistólica sanguínea media de 125 unidades. Si se sabe que la desviación típica de esta medida en la población es de 15 unidades, calcula el intervalo de confianza al 99 % para la media de esta variable en la población e interpreta el resultado.

Ejercicio 4.12.

Una compañía de seguros se planteó realizar un estudio de mercado en una determinada comunidad. Tenía interés en estimar el porcentaje de familias en las que al menos uno de los miembros de la misma tenía contratado alguna forma de seguro relacionado con la salud. Para ello, recogió información de una muestra aleatoria de 150 familias en la comunidad en estudio, la cuál reveló que en el 87 por ciento de los casos por lo menos uno de los miembros de la familia tenía contratado alguno de estos seguros. Con esta información, construye un intervalo de confianza al 90 % para el porcentaje de interés e interpreta su resultado.

Ejercicio 4.13.

Un grupo de investigación se planteó el estudio del porcentaje de personas con asma que tiene reacciones positivas de la piel al polvo de su casa. Para ello tomó una muestra de 140 pacientes asmáticos, en los que se obtuvo que el 35 por ciento tuvo estas reacciones positivas de la piel al polvo de casa. Con estos resultados construye el intervalo de confianza al 95 % para el porcentaje real de asmáticos que pueden tener estas reacciones alérgicas positivas e interprétalo.

Ejercicio 4.14.

Un centro de investigación está diseñando un proyecto mediante el cuál quieren estimar el tamaño medio de los cálculos biliares (piedras en la vesícula) de los pacientes que requieren una Colecistectomía (eliminación de la vesícula biliar), ya que están diseñando una nueva técnica para llevar a cabo la intervención. El tamaño medio de estos cálculos biliares se quiere estimar con un error inferior o igual a 4 mm y con una confianza del 99 %. Conocen, por otros estudios, que la desviación típica de esta variable se puede aproximar con el valor de 0.85 cm. Necesitan que calcules cuántos cálculos biliares deben analizar (tamaño de la muestra) para llevar a cabo su objetivo de estimación según sus condiciones de error.

Ejercicio 4.15.

Es conocido por un estudio reciente que la prevalencia de pacientes con Diabetes Tipo 2 que sufren como complicación una Nefropatía está alrededor del 13 %. Un grupo de investigación está interesado en llevar a cabo una estimación de esta misma prevalencia, pero con pacientes que tienen Diabetes Tipo 1. Quieren estimar esta prevalencia de Nefropatía en pacientes con Diabetes Tipo 1 con un error máximo del 3.5 % y con una confianza del 90 %. Calcula el tamaño muestral necesario para llevar a cabo este objetivo en los dos siguientes supuestos:

1. Considerando como valor aproximado de prevalencia que se quiere estimar la estimación de la misma en los pacientes con Diabetes Tipo 2.
2. Considerando que no se dispone de información previa para los valores esperados para esta prevalencia.

Ejercicios recopilatorios

Ejercicio 4.16.

Se dispone de una muestra de 14 niños de 5 años a los que se les ha medido la longitud de la tibia obteniendo los siguientes valores (en cm):

21.7 28.2 26.8 26.5 30.5 28.4 25.9 28.8 28.5 30.9 30.8 26.7 30.6 27.9

a) Calcula la media, la desviación típica, la mediana, el percentil 25 y el percentil 75 de estos datos.
b) Calcula un intervalo de confianza al 99 % para la media de la variable en estudio de la población e interpreta su resultado.
c) Tomando estos datos como estudio previo (o pre-muestra), calcula el tamaño muestral necesario para obtener una estimación de la longitud media de la tibia en niños de 5 años con un error inferior o igual a 1.0 cm (y con una confianza del 99 %).

Ejercicio 4.17.

Estudios epidemiológicos revelan que en Italia, alrededor del 10 % de los mayores de 65 años tienen diabetes. Para averiguar si en España la prevalencia de la enfermedad es significativamente diferente a la de Italia, se extrae una muestra de 500 mayores de 65 años y se determina que 38 de ellos tienen diabetes.
a) Indica cuál es la población de estudio, cuál es la variable de interés y clasifica de qué tipo de variable se trata.
b) Calcula un intervalo de confianza al 85 % para el porcentaje de diabéticos, mayores de 65 años, en España y determina, razonadamente, si la prevalencia de la diabetes de los dos paises es diferente.
c) Olvida todo lo anterior e imagina ahora que el porcentaje de diabéticos españoles, mayores de 65 años, sigue una distribución normal con media poblacional $\mu = 8$ y desviación típica poblacional $\sigma = 1{,}5$. Calcula la probabilidad de que en una provincia española el porcentaje de diabéticos varíe entre 5.3 y 9.7.

Ejercicio 4.18.

Los historiales de una clínica de adelgazamiento revelan que el último grupo de 9 pacientes adelgazaron (en kg) los siguientes valores:

9, 11, 11, 12, 10, 10, 7, 8 , 15

a) Calcula el percentil 33 e interpreta su resultado.
b) Averigua un intervalo de confianza al 80 % e interpreta su resultado en el contexto del ejercicio.

Ejercicio 4.19.

Entre los historiales de una clínica fisioterapéutica pertenecientes a los pacientes que han comenzado un tratamiento de filtraciones de rodilla en el primer trimestre se han seleccionado 12 historiales al azar y se ha extraído de ellos la edad de los pacientes obteniendo los siguientes valores:

10, 11, 11, 13, 14, 14, 15, 15, 16, 19, 21, 29

a) Calcula el percentil 40 e interpreta su resultado.
b) Realiza los cálculos necesarios para justificar si existen valores atípicos en la muestra dada.
c) Calcula un intervalo de confianza al 80 % e interpreta su resultado en el contexto del ejercicio.

Ejercicio 4.20.

Un estudio pretende estimar el porcentaje de personas ancianas con anemia en la Comunidad Valenciana. Para ello ha seleccionado una muestra de 584 ancianos de esta región, de los que en 139 se ha detectado algún tipo de anemia.
a) Calcula un intervalo de confianza al 98 % para el porcentaje de ancianos con anemia en la Comunidad Valenciana e interpreta su resultado.
b) Considera los datos del enunciado como datos de una pre-muestra y calcula el tamaño muestral necesario para estimar el porcentaje de interés con un error inferior o igual al 5 % y con la misma confianza del 98 %.

Ejercicio 4.21.

Se desea estimar el porcentaje de personas diabéticas que padecen hipertensión en España. Con este objetivo se ha tomado una muestra de 345 personas diabéticas de las que 87 padecían hipertensión.
a) Indica cuál es la población en estudio, cuál es la variable en estudio y el tipo de la misma.
b) Calcula un intervalo de confianza al 96 % para el porcentaje de interés en la población e interprétalo.
c) Estudios realizados recientemente publicados han estimado que el porcentaje de diabéticos en EEUU que padecen hipertensión es del 22 %. A la vista del intervalo calculado, ¿podemos concluir que en España ese porcentaje es significativamente diferente?

Ejercicio 4.22.

Un proyecto de investigación pretende, entre sus objetivos, poder estimar el nivel medio de hematocrito en hombres nadadores profesionales. Con este fin ha recogido una muestra de 11 sujetos de esta población obteniendo para ellos los siguientes valores de hematocrito:

46.5, 48.9, 43.6, 48.8, 49.5, 42.8, 45.9, 47.2, 46.9, 44.4, 47.7

a) Indica cuál es la población en estudio, cuál es la variable en estudio y el tipo de la misma.
b) Calcula un intervalo de confianza al 98 % para la media de la variable en la población e interpreta el resultado.
c) Los investigadores que han realizado un estudio previo afirman que el valor medio de hematocrito en este tipo de profesionales es superior a 44 ¿el intervalo que has obtenido en el apartado anterior confirma este resultado o no? Razona tu respuesta.
d) Calcula el percentil 77 de los datos de la muestra e interprétalo.

Capítulo 5

Introducción a los contrastes de hipótesis

Junto con los intervalos de confianza los *contrastes (o tests) de hipótesis* son la herramienta más importante de la inferencia estadística, es decir, una de las técnicas más importantes para extraer información a partir de los datos. Según hemos visto en el capítulo anterior los intervalos de confianza nos permiten dar estimaciones de cualquier parámetro estadístico incorporando la incertidumbre que todavía tenemos sobre dicho parámetro y que los datos que disponemos no son capaces de precisar en mayor medida. Por el contrario los contrastes de hipótesis son capaces de responder a preguntas concretas que nos podemos formular sobre los parámetros poblacionales de interés, por ejemplo: ¿La cantidad media diaria de sal ingerida por hipertensos es mayor que la que ingieren las personas con presión arterial normal?, ¿La temperatura media corporal de los pacientes que han sufrido cierta infección bacteriana es superior a los 36.7 grados centigrados?, ¿La proporción de personas diabéticas con problemas de vista es superior a la de la población general?. Resulta evidente que un mecanismo capaz de dar respuesta a cuestiones como las anteriores sería una herramienta muy valiosa, en consecuencia los contrastes o tests de hipótesis son una de las utilidades más valoradas y extendidas en la realización de estudios estadísticos.

Ejemplo 5.1.

Un estudio pretende estudiar la edad media a la que se diagnostica la Endometriosis en mujeres de un área de salud concreta. Para ello se dispone de una muestra aleatoria de 16 mujeres diagnosticadas de esta enfermedad en ese distrito, cuyas edades de diagnóstico se muestran a continuación:

22	*24*	*21*	*20*	*26*	*28*	*22*	*21*	*18*	*13*	*23*	*27*	*29*	*16*	*31*	*19*

¿Que podríamos decir a partir de dicha muestra?

Como queremos conocer la edad media de diagnóstico de esta enfermedad calculamos este valor para nuestra muestra, así como su desviación típica:

$$\bar{X} = 22{,}5; \quad S = 4{,}87; \quad S_{\bar{X}} = \frac{S}{\sqrt{n}} = 1{,}22$$

Podemos también calcular un intervalo de confianza para la edad media poblacional utilizando la distribución t con 15 grados de libertad. De dicha forma obtendríamos un intervalo de confianza al 95 % para la edad media de diagnóstico en el área de salud en cuestión (μ):

$$[19{,}91; 25{,}09]$$

Por tanto, con una confianza del 95 % el intervalo anterior contendrá a μ. Pero a parte de esta información también podemos estar interesados en preguntas del tipo:

- ¿La edad media de diagnóstico en ese distrito es superior a 24 años? (por ejemplo porque consideren que alrededor de 24 años de edad es un momento importante a partir del cuál muchas mujeres se plantean su maternidad).
- ¿Podemos afirmar que la edad media de diagnóstico es distinta de 25 años? ¿y de 18? (por ejemplo para poder compararse con la edad media de otras zonas en las que es conocido que diagnostican esta enfermedad de media a los 25 o 18 años)

Los contrastes de hipótesis darán respuesta a este y otros muchos tipos de preguntas.

5.1. Elementos fundamentales en contrastes de hipótesis

Las hipótesis

En cualquier contraste de hipótesis tendremos 2 alternativas complementarias en las que se especificarán distintos valores de un parámetro poblacional y a la vista de los datos habremos de optar por una de ellas. Por ejemplo, si deseamos conocer si el valor de un parámetro μ puede ser igual a 24 o por el contrario es inadmisible a la vista de los datos que disponemos, nuestras hipótesis serán:

$$\mu = 24 \;\; y \;\; \mu \neq 24$$

Estas 2 hipótesis que hemos señalado no jugarán el mismo papel dentro de cualquier contraste de hipótesis por tanto cada una de ellas recibirá un nombre específico:

- *Hipótesis nula*, a la que habitualmente nos referimos como H_0.
- *Hipótesis alternativa*, a la que habitualmente nos referimos como H_A o H_1.

A la hipótesis nula siempre se le concederá el beneficio de la duda e intentaremos encontrar en nuestra muestra evidencias en contra de ella. Así, al terminar el contraste habremos de optar por aceptar H_0 (si no tenemos evidencia suficiente en su contra) o rechazarla (si los datos hacen que la descartemos).

Se podría hacer un símil entre el papel de la hipótesis nula en un contraste de hipótesis y el acusado de un juicio: ambos tienen presunción de inocencia y si los datos no aportan evidencias suficientes en contra de su veracidad nos veremos obligados a aceptarlos. En consecuencia, si en un contraste de hipótesis rechazamos la hipótesis nula será porque disponemos de evidencias suficientes en su contra, es decir estamos razonablemente seguros de que dicha hipótesis es falsa. Por el contrario si aceptamos H_0 será porque no hemos encontrado evidencias suficientes en su contra, pero esto no implica que estemos más o menos seguros de que realmente dicha hipótesis sea cierta, podría darse el caso de que H_0 fuera falsa pero que los datos no aportan evidencia suficiente como para que lleguemos a dicha conclusión. En los juicios también pasa algo parecido con los acusados, si alguien resulta absuelto en un juicio no será porque hemos determinado su inocencia sino porque no hemos encontrado pruebas suficientes que lo inculpen.

En el siguiente cuadro se resumen las conclusiones a que conduce cada posible resultado de un contraste de hipótesis:

Resultado del contraste	**Conclusión**
Rechazar H_0	Podemos descartar H_0
No rechazar H_0	Aceptamos la posibilidad de que H_0 sea cierta aunque también lo podría ser H_1

Tal y como se puede apreciar en el cuadro anterior el rechazar H_0 conduce a conclusiones mucho más valiosas que el aceptarlo. Cuando aceptamos H_0 seguimos sin saber cuál de las dos opciones, la hipótesis nula o la alternativa, admitimos como cierta; por el contrario, cuando rechazamos H_0 estamos admitiendo implícitamente como cierta H_1, de esta forma nos decantamos por una de las dos hipótesis. Por este motivo suele ser bastante más valorado un resultado en el que se rechaza la hipótesis nula que aquel en el que se acepta. Es decir, el objetivo habitual que se perseguirá a la hora de hacer cualquier contraste de hipótesis será el intentar descartar la hipótesis nula que nos planteemos.

Ejemplo 5.2.

En el ejemplo anterior podríamos plantearnos el contraste:

$$H_0 : \mu = 24$$

$$H_1 : \mu \neq 24$$

¿Hasta qué punto los datos de la muestra invalidan la hipótesis nula? ¿Los datos de que disponemos nos conducen a rechazar H_0***?.***

Las dos preguntas anteriores se responderían mediante el contraste de hipótesis correspondiente a las dos hipótesis anteriores. En principio aceptaremos H_0 (le concedemos el beneficio de la duda) y habremos de valorar si los datos nos proporcionan suficientes evidencias en contra de la hipótesis nula. El intervalo de confianza para μ que hemos calculado anteriormente ([19.91; 20.09]) parece apuntar que el valor 24 en principio podría ser un valor admisible para μ pero desearíamos obtener un procedimiento que cuantificará la fiabilidad con la que puedo aceptar dicho valor, o no. El proceso de cálculo de contrastes de hipótesis que estamos introduciendo nos permitirirá establecer dicha fiabilidad, y en función de ella rechazaremos, o no, la hipótesis nula.

La unilateralidad o bilateralidad del contraste

Tal y como hemos podido comprobar hasta ahora todas las hipótesis que hemos formulado han sido expresadas 'matemáticamente' como relaciones de igualdad o desigualdad entre un parámetro y un valor concreto. Como norma general, y por razones que justificaremos con mayor detalle en la próxima sección, la hipótesis nula se corresponderá siempre con una igualdad. Sin embargo, la hipótesis alternativa no ha de responder siempre a una relación de desigualdad completa ($\neq$) sino que puede responder simplemente a una desigualdad parcial ($<$ o $>$). El utilizar una u otra desigualdad dependerá del problema en particular,

en concreto de aquello que queramos demostrar. Aquellos contrastes en los que la hipótesis alternativa se defina mediante el signo $\neq$ se llaman *Contrastes bilaterales*, ya que nos valen ambos sentidos de la desigualdad (tanto si el primer término es mayor que el segundo, o menor). Por el contrario aquellos contrastes en los que la hipótesis nula sea de la forma $<$ o $>$ se conocen como *Contrastes unilaterales*.

Ejemplo 5.3.

Plantea la hipótesis nula y alternativa para los siguientes contrastes de hipótesis:
· En un estudio se desea demostrar que el hecho de ser diabético altera también la presión arterial media de estos pacientes.
· En un estudio se desea demostrar que un nuevo farmaco antipirético es realmente efectivo, es decir, realmente baja la temperatura media de enfermos que presentan fiebre

En el primer caso queremos comparar dos valores, el valor medio de presión arterial en diabéticos μ_d frente a dicho valor en la población no afectado por esta enfermedad μ_n. En concreto deseamos conocer si ambos valores coinciden o no. Estas dos alternativas definen las dos hipótesis de nuestro contraste. Tal y como hemos mencionado anteriormente la hipótesis nula se corresponde con la igualdad, en ese caso tenemos:

$$H_0 : \mu_d = \mu_n$$

La hipótesis alternativa vendrá determinada por aquello que estamos interesados en demostrar, en este caso que las dos cantidades anteriores son distintas. Así:

$$H_1 : \mu_d \neq \mu_n$$

En el segundo contraste que se plantea, nuevamente se desean comparar dos cantidades, la temperatura media corporal antes de consumir el fármaco μ_a frente a la misma temperatura media algún tiempo tras de su consumo μ_t. Nuevamente la hipótesis nula viene dada por la igualdad de ambas cantidades (aquello que desearíamos descartar), entonces:

$$H_0 : \mu_a = \mu_t$$

En esta ocasión como queremos demostrar que la temperatura media trás la ingesta del fármaco disminuye. No querremos demostrar que existe una desigualdad en cualquiera de las dos sentidos posibles (menor o mayor), sino que querremos demostrar:

$$H_1 : \mu_a > \mu_t$$

Es decir esta será nuestra hipótesis alternativa.

La significatividad

Según hemos comentado previamente el objetivo fundamental de los contrastes de hipótesis será cuantificar la fiabilidad con la que podemos aceptar la hipótesis nula. Dicha fiabilidad, según veremos en la próxima sección, se mide como la probabilidad que tendríamos de equivocarnos en nuestra decisión si rechazáramos la hipótesis nula. Obviamente cuando dicha probabilidad sea 'alta' no rechazaremos H_0 ya que tendríamos un gran riesgo de equivocarnos. Por el contrario si la probabibilidad de errar en caso de rechazar H_0 fuera muy 'baja' podríamos rechazarla sin temor. Esta es la idea fundamental de los contrastes de hipótesis.

En cualquier caso para llevar a cabo el procedimiento anterior hemos de determinar cual será el umbral para la probabilidad por debajo del cual consideraremos que el riesgo de equivocarnos es 'bajo' o no. Dicho valor se conoce como la *significatividad* del contraste y habitualmente se denota como α. La interpretación de este parámetro sería: Máxima probabilidad de equivocarnos que estamos dispuestos a asumir en caso de que rechacemos la hipótesis nula.

En la práctica totalidad de estudios estadísticos el valor que se suele elegir para α es 0.05, aunque

también suelen tomarse $\alpha = 0{,}01$ o $\alpha = 0{,}10$ dependiendo de si queremos asumir menos o más riesgo de equivocarnos, respectivamente, en caso de rechazar la hipótesis nula. La utilización de estos valores se ha definido por consenso de la comunidad ciéntifica y resulta muy inusual la utilización de otros valores de significatividad distintos a los anteriores y su utilización requiere la existencia de alguna razón de peso que habría de ser debidamente justificada.

5.2. Mecánica de los contrastes de hipótesis

Una vez hemos descrito los elementos fundamentales de los contrastes de hipótesis estamos en condiciones de describir la mecánica habitual para llevar a cabo este proceso. Dividimos este proceso en las siguientes fases:

1. Búsqueda de pivote.
2. Cálculo del pivote y su probabilidad.
3. Delimitación de la región de rechazo.
4. Aceptación/rechazo de la hipótesis nula.

A continuación describimos con más detalle cada una de estas fases.

Búsqueda de pivote

Llamaremos pivote a un estadístico, función de los datos que dispongamos, que tenga una distribución conocida cuando asumamos como cierta la hipótesis nula.

Ejemplo 5.4.

Vamos a continuar con el ejemplo que hemos venido planteando sobre la edad de diagnóstico de la Endometriosis en un área de salud. Supongamos que consideramos que esta edad sigue una distribución Normal con una media μ (desconocida), pero que conocemos que su desviación típica es $\sigma = 5$. Vamos a hallar un pivote apropiado para el contraste de hipótesis que se plantea como hipótesis alternativa que la edad media de diagnóstico es diferente de 24.

Partimos de que las edades de diagnóstico siguen una distribución Normal con media desconocida (μ) y desviación típica conocida ($\sigma = 5$):

$$X_1, ..., X_{16} \sim N(\mu, 5)$$

Hemos de calcular una transformación de los datos anteriores de forma que conozcamos su distribución resultante bajo la hipótesis nula ($H_0 : \mu = 24$). Según vimos la distribución de la media de un conjunto de valores tiene como distribución:

$$\bar{X} \sim N(\mu, \frac{5}{\sqrt{16}})$$

, que bajo la hipótesis nula queda completamente determinada como:

$$\bar{X} \sim N(24, \frac{5}{\sqrt{16}})$$

Por tanto la media será un pivote apropiado para llevar a cabo el contraste de hipótesis que nos estamos planteando.

Cálculo del valor del pivote y su probabilidad

Una vez hemos determinado qué función de los datos puede ser válida como pivote estaremos en disposición de calcular el valor concreto de nuestro pivote y localizar dicho valor dentro de la distribución

de probabilidad que seguiría bajo la hipótesis nula. La idea subyacente que desarrollaremos en los siguientes pasos del contraste de hipótesis es que si el pivote cae en una región anómala, es decir de baja probabilidad, de la distribución anterior (en la que está implícita la hipótesis nula) será síntoma de que la hipótesis nula no es demasiado compatible con los datos que hemos observado. De esta forma nos veremos abocados a rechazar la hipótesis nula.

Ejemplo 5.5.

Siguiendo con el ejemplo anterior calcula el valor del pivote y represéntalo en relación a su función de distribución

La media del conjunto de valores de este ejemplo según vimos vale 22.5. Respecto a la ubicación de este valor en relación a su distribución resultará más conveniente su representación respecto a la distribución tipificada ya que de esta forma podremos evaluar la probabilidad que tendríamos de haber obtenido valores mayores y/o menores bajo la hipótesis nula. Para tipificar el valor anterior habremos de restarle su valor esperado bajo la hipótesis nula (24) y dividirlo por su desviación típica. De esta forma le correspondería el siguiente valor de una normal tipificada:

$$\frac{22{,}5-24}{5/\sqrt{16}}=\frac{-1{,}5}{5/4}=-1{,}2$$

Si recurrimos a la tabla de la distribución normal tipificada podremos determinar que el valor anterior deja a su izquierda una probabilidad de 0.1151. Es decir, asumiendo la hipótesis nula el valor que hemos observado de la media sería un valor relativamente bajo ya que sólo un 11.51 % de los valores de la normal tipificada son inferiores a éste.

Delimitación de la región de rechazo

Una vez disponemos de la distribución correspondiente al pivote bajo la hipótesis nula podremos delimitar aquellos valores de esta distribución que nos parecen más anómalos. En caso de que el pivote sea uno de estos valores deberíamos rechazar la hipótesis nula ya que los datos (pivote) no parecen ser demasiado compatibles con dicha hipótesis.

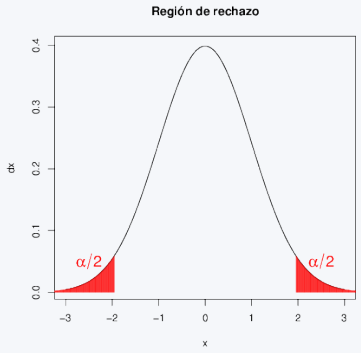
La región de rechazo dependerá de los siguientes factores concretos que hemos introducido en la sección anterior: la unilateralidad o bilateralidad del contraste y la significatividad. En concreto, la unilateralidad/bilateralidad del test nos dirá si debemos coger una o las dos colas de la distribución (respectivamente) como región de rechazo de la hipótesis nula. Es decir, si el contraste que manejamos es bilateral, en cuyo caso la hipótesis alternativa será una desigualdad completa ($H_1 : \mu \neq \mu_0$), cualquier valor del pivote (en nuestro ejemplo $\bar{X}$) que se situe muy alejado del valor que estamos contrastando μ_0 (en nuestro caso 24) apuntará a que los datos y la hipótesis nula no son compatibles y, por tanto, nos obligará a rechazar dicha hipótesis. En ese caso la región de rechazo estará formada por todos aquellos valores muy superiores a 24 o aquellos muy inferiores. En cualquier caso la región de rechazo constará de dos trozos o colas de la distribución. Por el contrario, si la hipótesis alternativa que manejamos es unilateral, por ejemplo del tipo $H_1 : \mu < \mu_0$, no todos los valores de nuestra media muestral $\bar{X}$ apuntarán hacia la hipótesis alternativa, sino aquellos que aporten evidencias de que la media de los datos es inferior a μ_0 y, en ese caso, la región de rechazo de la hipótesis nula estará formada sólo por una de las colas de la distribución, la correspondiente a los valores más bajos.

Por otro lado, la significatividad es el otro factor que va a determinar la extensión de la región de rechazo. Según hemos comentado previamente la significatividad se corresponde con el riesgo de equivocarnos que estamos dispuestos a asumir en caso de rechazar la hipótesis nula. En ese caso a valores de la significatividad más bajos seremos más restrictivos para rechazar la hipótesis nula, o de forma equivalente, habremos de definir regiones de rechazo más pequeñas.

Ejemplo 5.6.

Continuando con el ejemplo anterior, vamos a encontrar la región de rechazo en caso de que deseamos hacer el contraste con una significatividad $\alpha = 0{,}05$

Como resulta más fácil trabajar con la distribución normal tipificada y disponemos del valor correspondiente de nuestro pivote en la distribución tipificada (-1.2) vamos a delimitar la región de rechazo en la distribución tipificada. Como el test planteado es bilateral ($H_1 : \mu \neq 24$) habremos de delimitar dos regiones de rechazo, una para los valores del pivote muy superiores a 24, equivalentemente para los valores de la distribución tipificada muy superiores a 0. La otra región de rechazo se corresponderá con aquellos valores del pivote muy inferiores a 24, equivalentemente aquellos valores de la distribución tipificada mucho menores que 0. La región de rechazo habrá de abarcar el 5 % de valores más extremos, y por tanto anómalos, de esta distribución. De esta forma admitiremos como región de rechazo aquella situada por debajo del percentil 2.5 ($\alpha/2$) y por encima del percentil 97.5 ($\alpha/2$).



De esta forma hemos delimitado una región que abarca el 5 % de los valores más discordantes con la hipótesis nula. Si los datos que disponemos conducen a un pivote que cayera en dicha región deberíamos rechazarlo ya que la probabilidad de que dicho pivote se situara en dicha región únicamente por azar es sólo del 5 %, que es el riesgo de equivocarnos que estamos dispuestos a asumir. Es decir, si la hipótesis nula fuera cierta, únicamente un 5 % de los valores del pivote estarían situados en ese 5 % de valores que vamos a descartar, y ese es el error que asumimos al rechazar la hipótesis nula cuando obtenemos un pivote en esa región (la consideramos una probabilidad despreciable ligada, por tanto, a un error asumible).

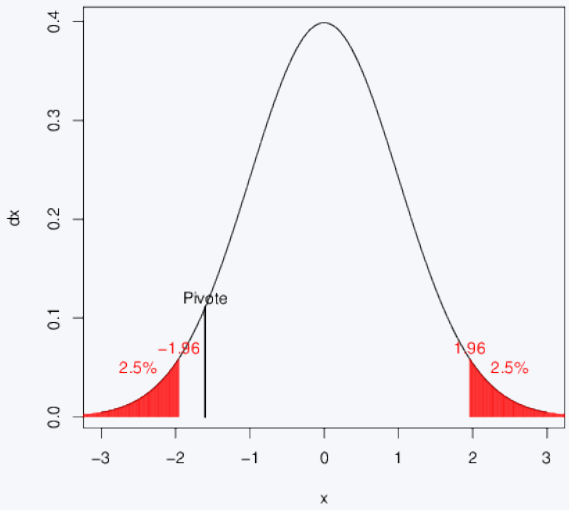
Aceptación/rechazo de la hipótesis nula

Una vez hemos calculado el valor de nuestro pivote, habitualmente sobre una distribución tipificada, y la región de rechazo correspondiente a nuestro contraste estaremos en condiciones de concluir el contraste de hipótesis. Así, si el pivote recae dentro de la región de rechazo concluiremos el contraste descartando la hipótesis nula y admitiendo por tanto la hipótesis alternativa como verdadera. Por el contrario si el pivote no cae dentro de la región de rechazo no dispondremos de evidencias suficientes como para descartar la hipótesis nula y concluiremos que dicha hipótesis puede ser cierta, aunque también podría serlo la hipótesis alternativa. En este último caso una forma apropiada de expresar nuestra conclusión final sería: 'Los datos no aportan evidencia suficiente como para descartar la hipótesis nula, por lo que aceptamos que pueda ser cierta'.

Ejemplo 5.7.

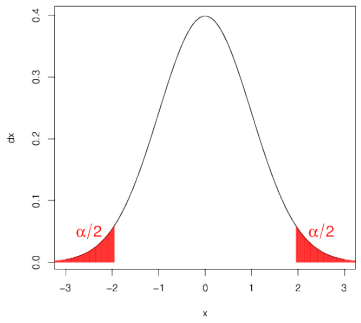
A partir del valor del pivote y la región de rechazo calculada en los ejemplos anteriores vamos a concluir el contraste de hipótesis correspondiente.

Tal y como se aprecia en la siguiente figura, el pivote recae fuera de la región de rechazo de la hipótesis nula. Por tanto, a la vista de los datos que disponemos no tenemos evidencia suficiente como para descartar la hipótesis nula.

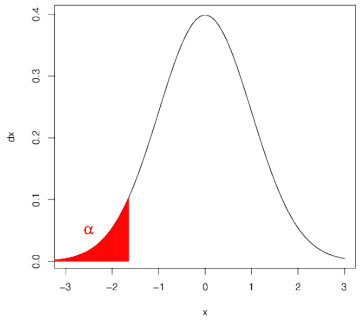


A continuación se resume el cálculo de la región de rechazo según el nivel de significatividad (α) y el carácter unilateral/bilateral del contraste:

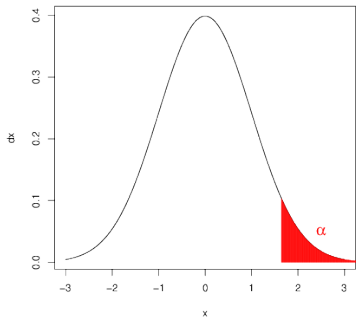
- Contraste bilateral ($H_1 : \mu \neq \mu_0$)



- Contraste unilateral ($H_1 : \mu < \mu_0$)



- Contraste unilateral ($H_1 : \mu > \mu_0$)



Notar que en el procedimiento que hemos descrito no resulta indispensable el cálculo de la probabilidad asociada al pivote (tal y como hemos hemos hecho en el segundo paso de los cuatro que hemos descrito), ya que podíamos haber situado simplemente el pivote (valor -1.2 en el ejemplo) dentro de la distribución tipificada sin necesidad de conocer su probabilidad asociada (en este caso 0.1151). Sin embargo esta probabilidad nos va a proporcionar un estadístico de gran importancia en los contrastes de hipótesis, el P-valor.

5.3. Resolución de contrastes mediante el cálculo del P-valor

En todo contraste de hipótesis aceptaremos o rechazaremos al hipótesis nula dependiendo del valor que hayamos establecido de significatividad (α). En concreto, si la significatividad es más alta admitimos mayor riesgo de equivocarnos cuando rechacemos la hipótesis nula y en consecuencia rechazaremos dicha hipótesis con mayor facilidad. El *P-valor* de un contraste de hipótesis se define como la probabilidad de error que tendríamos que estar dispuestos a asumir en caso de rechazar la hipótesis nula con los datos de que disponemos. La importancia del P-valor viene dada porque nos proporciona un resultado mucho más informativo que el que nos proporciona el propio resultado del contraste, ya que este último termina diciendo únicamente si aceptamos o no la hipótesis nula, ya sea con una gran holgura, o sin ella. Sin embargo el P-valor cuantifica el riesgo a equivocarnos que tendríamos que asumir si quisiéramos rechazar H_0 con nuestros datos. Por tanto, se suele interpretar el P-valor como una medida de la evidencia que aportan los datos a favor (o en contra) de la hipótesis nula. En concreto, aquellos valores bajos del P-valor se corresponden con datos que no apoyan la hipótesis nula, ya que la probabilidad de equivocarnos que tendríamos que asumir para rechazarla sería baja.

El P-valor supone, además, una herramienta alternativa para la resolución de contrastes de hipótesis. Así, supongamos pues que disponemos del valor del P-valor p de cierto contraste y supongamos que dicho valor es inferior a la significatividad del contraste, es decir $p < \alpha$. En ese caso la probabilidad de equivocarnos que tendríamos que asumir para rechazar la hipótesis nula (el P-valor) es menor que la probabilidad de equivocarnos que estamos dispuestos a asumir (la significatividad), por tanto podremos rechazar la hipótesis nula. Por el contrario, si el P-valor es mayor que la significatividad, la probabilidad de equivocarnos que tendríamos que asumir para rechazar la hipótesis nula (P-valor) sería superior a la que estamos dispuestos a asumir (α), por lo que no podríamos rechazar dicha hipótesis.

$$P - valor \leq \alpha \Rightarrow \ Rechazamos\ H_0$$

$$P - valor > \alpha \Rightarrow \ (No\ \ podemos\ \ rechazar\ H_0\ \ )Aceptamos\ H_0$$

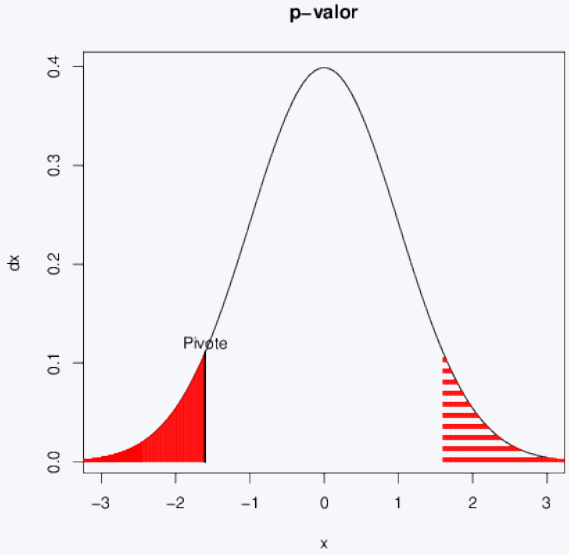
En consecuencia, la comparación del P-valor con la significatividad nos proporciona un criterio alternativo para la resolución de contrastes de hipótesis.

Ahora sólo nos queda ilustrar cómo se calcula el P-valor de un contraste de hipótesis, veámoslo con un ejemplo.

Ejemplo 5.8.

Vamos a calcular el P-valor del contraste propuesto en el ejemplo 5.2

En los ejemplos anteriores vimos que el pivote de dicho contraste valía -1.2 y que dicho valor dejaba a su izquierda una probabilidad de 0.1151. Además, vimos que la región de rechazo era bilateral, es decir, se compone de los valores más altos y más bajos de la distribución.
Según hemos definido el P-valor habremos de calcular el menor valor de la significatividad de forma que el pivote caiga en la región de rechazo. O, lo que es lo mismo, *hasta dónde tendría que llegar la región de rechazo para que el pivote estuviera dentro de ella.* Conforme aumentemos el valor de la significatividad crecerá la extensión de la región de rechazo hacia el centro de la distribución. Por tanto, habremos de calcular el valor de la significatividad correspondiente a la región de rechazo delimitida por el valor del pivote, es decir la región de rechazo formada por todos aquellos valores inferiores a -1.2 y aquellos valores superiores a 1.2 (dado el caráter bilateral del contraste).



Como la probabilidad de aquellos valores inferiores a -1.2 era 0.1151 y, por simetría, la probabilidad de aquellos valores superiores a 1.2 también valdrá 0.1151, la probabilidad de ambas regiones conjuntamente ascenderá a $2 \cdot 0{,}1151 = 0{,}2302$. Por tanto, ese valor corresponderá al P-valor del contraste que nos hemos planteado. Notar que como el P-valor (0.2302) es mayor que la significatividad, no deberíamos rechazar la hipótesis nula. Por tanto el resultado que habríamos obtenido por el método del P-valor coincide, obviamente, con el que habríamos obtenido con el primero de los métodos de resolución de contrastes que hemos expuesto.

Vamos a ilustrar con mayor detalle la información que nos proporciona el P-valor en la resolución de contrastes de hipótesis. Supongamos que efectuamos 2 contrastes de hipótesis con una significatividad de 0.05. En el primero de ellos obtenemos un P-valor de 0.053, mientras que en el segundo el P-valor resulta 0.53. En ambos casos el contraste concluiría con la aceptación de la hipótesis nula, puesto que en los dos casos el P-valor obtenido es mayor que el nivel de significatividad 0.05 que hemos definido. Sin embargo, en el primero de los contrastes el P-valor está muy cerca de la significatividad y, por tanto, podremos darnos cuenta de que estamos muy cerca de haber podido rechazar H_0. Por el contrario, en el segundo caso si quisiéramos rechazar la hipótesis nula, tendríamos que haber asumido un error de 0,53 (nos equivocaríamos en más de la mitad de las veces que lo hiciéramos con un valor así), por lo que en este caso no rechazaremos

H_0 bajo ningún concepto.

Como resumen de esta sección damos las pautas para hallar el P-valor:

1. Calculamos el valor del *Pivote* y buscamos en la distribución correspondiente con qué percentil se corresponde.

2. A partir de este valor podemos calcular la probabilidad de obtener un valor superior (si es positivo) o inferior (si es negativo) al pivote.

   - Si el contraste es unilateral, esta probabilidad que hemos obtenido es el P-valor.

   - Si el contraste es bilateral, multiplicaremos esta probabilidad por 2 y este nuevo valor será el P-valor

3. Comparamos el P-valor (p) con el valor del nivel de significatividad del contraste (α):

   - Si $p \leq \alpha$, rechazamos H_0

   - Si $p > \alpha$, no rechazamos H_0

5.4. Contrastes para una media

Uno de los ejemplos más sencillos y, a su vez habituales, de contrastes de hipótesis es el ejercicio de comparación del valor de una media con un valor concreto. Este es el objetivo que nos planteamos en esta sección.

Ejemplo 5.9.

Es conocido por diversos estudios que la población general de mujeres españolas tienen un peso medio aproximado de 57 kg. Nuestro objetivo es estudiar el peso medio de las mujeres que siguen dieta vegetariana. Hemos realizado un estudio en el que se ha recogido una muestra de 20 mujeres que siguen dicha dieta. A partir de la muestra recogida se ha obtenido un peso medio de 54.54 kg y una desviación típica de 5 kg. ¿Podemos concluir a partir de los datos que las mujeres vegetarianas tienen un peso medio significativamente inferior a la población general? Plantea el problema anterior como un contraste de hipótesis.

En el problema que se nos plantea, deseamos conocer cierta característica de μ, el valor esperado poblacional de las mujeres que siguen dieta vegetariana. En concreto, deseamos comparar este parámetro con el valor de referencia 57, que es el peso medio esperado de las mujeres que no siguen la dieta. En el contraste de hipótesis que se nos plantea la hipótesis nula, como siempre, viene dada por el signo de igualdad, es decir:

$$H_0 : \mu = 57$$

mientras que, al contrario, la hipótesis alternativa (que es aquello que querríamos demostrar), tal y como ha sido planteada la pregunta del enunciado correspondería a:

$$H_1 : \mu < 57$$

Ya que queremos demostrar que las mujeres que siguen la dieta vegetariana pesan *menos* que las mujeres que no siguen esta dieta. Visto de otra forma, en principio consideramos la hipótesis nula como válida (admitimos que las mujeres vegetarianas pesan igual que las que no siguen esta dieta) y a la vista de los datos querremos conocer si dicha hipótesis es admisible o no. Dado que la hipótesis alternativa sólo se compone de aquellos valores menores que el valor de referencia nos encontramos ante un contraste de hipótesis unilateral.
Una vez hemos planteado el contraste de hipótesis, hemos de determinar con qué herramientas contamos para poder dar respuesta a la pregunta que nos planteamos, los datos. Disponemos como datos de nuestro problema: El número de mujeres que integran nuestra muestra, $n = 20$, el peso medio de las mujeres en nuestra muestra, $\bar{X} = 54{,}54$ y la desviación típica de los pesos de estas mujeres, $s = 5$. En principio, el peso medio de las mujeres de nuestra muestra (54.54) parece apuntar a que las mujeres que siguen dieta vegetariana podrían pesar menos que las mujeres de la población general (57), pero ¿la diferencia de peso que hemos observado entre las mujeres de nuestra muestra y la media de la población general, es realmente concluyente, o puede haberse dado simplemente por azar?. Esta pregunta se respondería mediante el contraste de hipótesis que nos hemos planteado.

Ejemplo 5.10.

Vamos a resolver el contraste anterior mediante el método de las regiones de Aceptación/Rechazo o método de la región de Rechazo, para el nivel de significatividad α=0.05

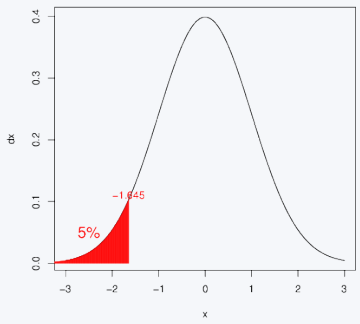
Lo primero que hemos de hacer para resolver el contraste es determinar un pivote válido. Como nos planteamos una cuestión sobre el peso medio poblacional de las mujeres vegetarianas, un buen candidato como pivote podría ser la media muestral de este grupo de mujeres.
Bajo la hipótesis nula sabemos que $\mu = 57$. Como no conocemos la desviación típica poblacional σ, en este caso tenemos:

$$\frac{\bar{X} - 57}{5/\sqrt{20}} \sim T_{19}$$

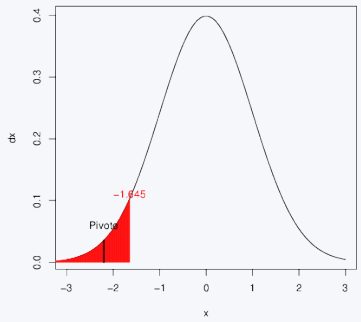
En nuestros datos la media muestral vale 54,54 kilos, por tanto, el valor del pivote tipificado en nuestro problema será:

$$\frac{\bar{X} - 57}{1{,}12} = \frac{54{,}54 - 57}{1{,}12} = \frac{-2{,}46}{1{,}12} = -2{,}2 \sim T_{19}$$

Una vez hemos determinado el valor del pivote, hemos de delimitar la región de rechazo de nuestro problema y, a continuación, comprobar si el pivote cae dentro o fuera de esta región. El valor de la significatividad (0,05) nos informa sobre qué dimensión debe tener la región de rechazo. Además, como la hipótesis alternativa es ($H_1 : \mu < 57$), sabemos que contempla sólo los valores más pequeños que 57 kilos (estos se corresponden en la distribución tipificada con los valores negativos). Nuestra región de rechazo habrá de tener, por tanto, la siguiente forma:



Como el percentil al 5 % de la T_{19} vale: $P_5 = -P_{0,95} = -1{,}729$ tenemos que la región de rechazo estará formada por todos aquellos valores inferiores a $-1{,}729$. Es decir, valores del pivote tipificado inferiores a $-1{,}729$ sólo tienen una probabilidad del 5 % de producirse por azar si la hipótesis nula fuera cierta. Por tanto, asumiendo ese 5 % de error, consideraremos que todos estos valores se deben a que la población que hemos observado tiene realmente un peso inferior al de la población general.



Como el valor del pivote $-2{,}2$ está incluido en la región de rechazo, puesto que $(-2{,}2 < -1{,}729)$, podemos rechazar la hipótesis nula $H_0 : \mu = 57$ y admitimos la alternativa como hipótesis valida $(H_1 : \mu < 57)$. Así, hemos demostrado que las mujeres vegetarianas pesan de media menos, de forma significativa, que las mujeres de la población general. Al concluir este resultado tenemos una probabilidad de habernos equivocado del 5 % (la significatividad), que es el riesgo que hemos asumido en nuestro contraste. Si quisiéramos estar más seguros de nuestra afirmación deberíamos asumir un valor de la significatividad más bajo, por ejemplo del 1 %.

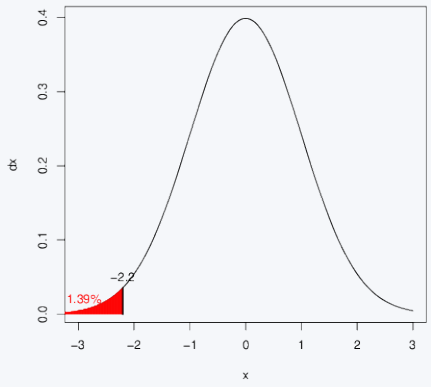
Ejemplo 5.11.

Vamos a resolver nuevamente el contraste planteado en el ejemplo anterior, pero ahora mediante el método del P-valor

Según hemos determinado en el ejemplo anterior el pivote de dicho contraste valía -2.2, y la región de rechazo estaba formada por la cola izquierda de la distribución T_{19}. El límite superior de dicha región venía dado por la significatividad del contraste, dependiendo de ésta situaremos el límite de la región de rechazo o más a la izquierda o más a la derecha.
Para hallar el P-valor hemos de hacer coincidir el límite de la región de rechazo con el valor del pívote y determinar cual es el área determinada por dicha región. Así, querríamos determinar cual es el área que acumula la distribución T_{19} por debajo del valor -2.2. Es decir, queremos hallar:

$$P(T_{19} < -2{,}2)$$

Para ello hemos de valernos de la tabla de la distribución T_{19}. Como en dicha tabla sólo aparecen números positivos buscaremos al área a la izquierda de 2.2, dicha área vale 0.9825, aproximadamente. Por tanto, el área a la derecha de 2.2 valdrá $1 - 0{,}9825 = 0{,}0175$, aproximadamente. Por simetría de la distribución T_{19} podremos comprobar que el área a la izquierda de -2.2 es exactamente la misma que el área a la derecha de 2.2. Por tanto el P-valor que buscábamos (área por debajo de -2.2) vale 0.0175.



Como el P-valor (probabilidad que tendríamos que asumir de equivocarnos en nuestra decisión si rechazáramos la hipótesis nula) es inferior a la significatividad que habíamos establecido, 0.05 (riesgo de equivocarnos que estamos dispuestos a asumir en caso de rechazar la hipótesis nula) podremos rechazar la hipótesis nula. Obviamente, el resultado que hemos obtenido por ambos métodos ha sido el mismo, rechazar la hipótesis nula, puesto que son equivalentes ambas formas de resolución. Además, mediante el P-valor sabemos que el riesgo de equivocarnos en nuestra decisión es de 0.0175, mientras que con el método de la región de rechazo sólo sabíamos que dicho riesgo era inferior al 5 %.

En los contrastes de hipótesis para una media tenemos la misma casuística que en el caso de los intervalos de confianza estudiados anteriormente.

- Si conocemos la desviación típica de la población σ, o bien el tamaño muestral es suficiente para poder estimar esta desviación típica con la desviación típica muestral S (por ejemplo, $n > 30$), utilizaremos como pivote y distribución:

$$\frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

- Si por el contrario, la desviación típica de la población σ es desconocida, y el tamaño de la muestra no permite aproximar de forma razonable esta desviación típica con la desviación típica muestral (por ejemplo, $n < 30$), utilizaremos como pivote y distribución:

$$\frac{\overline{x} - \mu}{\frac{S}{\sqrt{n}}} \sim t_{n-1}$$

5.5. Contrastes para un porcentaje

Un segundo ejemplo de contraste de hipótesis muy habitual es el contraste sobre un porcentaje (P). En este caso los datos disponibles son el porcentaje muestral ($\widehat{P}$) y el tamaño de la muestra (n). El pivote que se utiliza en estos casos es:

$$\frac{\widehat{P} - P}{\sqrt{\frac{P \cdot (100 - P)}{n}}}$$

A continuación vamos a ver un ejemplo de este tipo de contrastes.

Ejemplo 5.12.

Es conocido por diversos estudios que el 15 % de los hombres europeos mayores de 50 años padece de hipertensión. Nuestro objetivo es estudiar el porcentaje de hombres mayores de 50 años españoles que padecen hipertensión. Hemos realizado un estudio en el que se ha recogido una muestra de 300 hombres españoles mayores de 50 años. A partir de la muestra recogida se ha obtenido que 48 de ellos padecían hipertensión. ¿Podemos concluir a partir de los datos de que disponemos, que el porcentaje de hipertensos entre los hombres españoles mayores de 50 años es significativamente diferente al de Europa?. Plantea el contraste de hipótesis necesario para resolver la cuestión anterior.

Queremos conocer cierta característica de P, el porcentaje de hombres españoles, mayores de 50 años, con hipertensión. En concreto queremos saber saber si tenemos evidencias de si dicho valor es necesariamente distinto del 15 %, o por el contrario no tenemos evidencias suficientes como para hacer dicha afirmación. Así, la hipótesis nula de nuestro contraste vendrá dada por la igualdad, es decir:

$$H_0 : P = 15\,\%$$

Por el contrario, como estamos interesados en demostrar que dicho parámetro es distinto de 15, la hipótesis alternativa valdrá:

$$H_1 : P \neq 15\,\%$$

De esta forma, estamos admitiendo que en principio los hombres españoles deberían tener el mismo porcentaje de hipertensión que el resto de europeos, y a la vista de los datos deduciremos si podemos seguir manteniendo este afirmación, o no.
Como datos para resolver este contraste tenemos: el porcentaje de personas hipertensas observadas en nuestra muestra, $\hat{P} = 100 \cdot 48/300 = 16\,\%$, y el tamaño muestral de nuestra muestra, 300 personas.

Ejemplo 5.13.

Vamos a resolver el contraste anterior mediante el método de la región de rechazo, para el nivel de significatividad $\alpha = 0{,}05$

Comenzaremos buscando un pivote apropiado para el contraste. Como queremos determinar alguna característica del porcentaje de hipertensos en la población, seguramente el porcentaje de hipertensos en la muestra nos podrá ser de utilidad. La distribución de dicho estadístico es:

$$\hat{P} \sim N\left(P, \sqrt{\frac{P(100-P)}{n}}\right)$$

Bajo la hipótesis nula tenemos $P = 15$, entonces admitiendo dicha hipótesis la distribución de $\hat{P}$ resulta:

$$\hat{P} \sim N\left(15, \sqrt{\frac{15(100-15)}{300}}\right) \rightarrow \frac{\hat{P}-15}{\sqrt{\frac{15(100-15)}{300}}} \sim N(0,1)$$

Como $\hat{P} = 16$ el pivote toma al valor:

$$\frac{\hat{P}-15}{\sqrt{\frac{15(100-15)}{300}}} = \frac{16-15}{\sqrt{\frac{15*85}{300}}} = \sqrt{\frac{300}{15*85}} = 0{,}485$$

Como el test que nos planteamos es bilateral hemos de delimitar dos regiones, una delimitida por el percentil 97.5 en adelante y la otra por los valores menores que el percentil 2.5. Es decir, la región de rechazo estará formada por los valores del pivote superiores a 1.96 y los valores inferiores a -1.96. Como el pivote (0.485) cae fuera de la región de rechazo no podemos rechazar la hipótesis nula. En ese caso concluimos que no tenemos evidencias suficientes como para asegurar que el porcentaje de hipertensos en España sea distinto al resto de la Unión Europea.

Ejemplo 5.14.

Vamos a resolver, ahora, el contraste anterior mediante el método del P-valor, para el nivel de significatividad $\alpha = 0{,}05$

En el contraste anterior hemos determinado que el valor del pivote es 0.485. Para calcular el P-valor hemos de hacer coincidir el límite de la región de rechazo con el pivote y calcular la probabilidad asociada a dicha región de rechazo. Si hacemos coincidir el límite de la región de rechazo con el pivote tendríamos que la región de rechazo estará formada por dos regiones, todos aquellos valores superiores a 0.485 y por simetría (ya que el contraste es bilateral) todos aquellos valores inferiores a -0.485. Para hallar la probabilidad de que un valor sea superior a 0.485 podemos ir a la tabla de la distribución normal ya que en ella aparece la probabilidad de que un valor de esta distribución sea inferior a 0.485 ((0,6843 + 0,6879)/2 = 0.686). Simplemente haciendo 1-0.686=0.314 tendremos la probabilidad que buscamos. Como la probabilidad de la cola izquierda de la región de rechazo por simetría tendrá la misma probabilidad que la derecha tenemos que la probabilidad cubierta por la región de rechazo, y en consecuencia el P-valor, valdrá: 2*0.314=0.628.
Como el P-valor es mayor que la significatividad no podemos rechazar la hipótesis nula. Además, en caso de rechazarla la probabilidad que tendríamos de equivocarnos es de 0.628. Por tanto, el P-valor no sólo nos asegura que no podemos rechazar la hipótesis nula, sino que si lo hicíeramos tendríamos una gran probabilidad de equivocarnos. Por tanto, nuevamente el P-valor nos proporciona cierta información que el contraste mediante la región de rechazo no nos proporcionaba.

5.6. Errores de tipo I y tipo II

Ante un contraste de hipótesis se pueden dar todas las combinaciones que discribimos a continuación.

	No rechazamos H_0	**Rechazamos** H_0
H_0 **Cierta**	Acierto	Error de tipo I
H_0 **Falsa**	Error de tipo II	Acierto

Podemos acertar en nuestra decisión de acertar o rechazar la hipótesis nula, o por el contrario podemos equivocarnos en nuestra decisión. En ningún caso sabremos si hemos acertado o no en nuestra decisión, aunque sí podremos conocer la probabilidad que tenemos de equivocarnos en nuestra decisión. En concreto, sabremos la probabilidad que tendríamos de equivocarnos en caso de que rechazáramos la hipótesis nula, esto es lo que en su momento definimos como significatividad. Por tanto la probabilidad en que incurrimos en lo que en la tabla anterior hemos definido como error de tipo I es lo que conocemos como significatividad. En un contraste de hipótesis no se le da la misma importancia al error de tipo I que al de tipo II, de forma análoga a los juicios de la vida real. En dichos juicios consideramos asumible que una persona que ha cometido un delito no resulte condenada (si no tenemos pruebas suficientes que lo incriminen ...). Sin embargo, lo que sí consideramos inaceptable es que una persona inocente pueda ser condenada. En los contrastes de hipótesis pasa algo parecido, queremos controlar a toda costa el error de tipo I (no queremos rechazar en general la hipótesis nula si ésta es cierta). De hecho, la significatividad nos garantiza que la probabilidad de dicho error va a ser siempre baja. Sin embargo, al error de tipo II en general, tal y como hemos visto a lo largo del tema se le presta bastante menos atención. Consideramos que en general vamos a aceptar la hipótesis nula aun siendo falsa si no tenemos datos suficientes como para descartarla. Por tanto consideramos que el error de tipo II es cuestión exclusiva de los datos y que para disminuir dicho error la única posibilidad con la que contamos es aumentar el número de datos de que disponemos.

Se suele denominar potencia de un contraste a $\beta = 1 - P(error\ de\ tipo\ II)$ es decir, un contraste será más potente cuanto menor sea su probabilidad de error de tipo II, o dicho de otro modo cuanta más sensibilidad tenga para detectar el que la hipótesis nula sea falsa cuando realmente lo sea. Entre dos contrastes diferentes (por ejemplo basado en pivotes distintos) de una misma hipótesis siempre preferiremos aquel de mayor potencia.

5.7. Ejercicios Capítulo 5

Para todos los problemas que se proponen a continuación reflexiona sobre cuál es en cada uno de ellos:

- Población en estudio
- Variable en estudio y tipo de la misma (cuantitativa o cualitativa)
- Parámetro de interés para el que se plantea el contraste de hipótesis
- Interpretación de los resultados obtenidos en el contexto del ejercicio.

Ejercicio 5.1.

Un grupo de investigación tiene interés en estimar la edad media a la que aparecen determinados trastornos relacionados con la Diabetes Tipo II. Para ello ha seleccionado las historias clínicas de algunos de estos pacientes diagnosticados con este problema y ha obtenido sus edades de diagnóstico.

58 62 64 67 69 70 72 73 73 75 80

Plantea el contraste de hipótesis adecuado para contrastar si la edad media de diagnóstico es significativamente diferente de 65 años, con una significatividad de $\alpha = 0{,}05$. Resuelve el contraste por el método de la región de rechazo y explica con claridad la conclusión del mismo.

Calcula el p-valor del contraste. ¿Llegarías a la misma conclusión que en el apartado anterior?

Comprueba que calculando el intervalo de confianza al 95 % para la media poblacional correspondiente, se obtiene la misma conclusión que has obtenido con el contraste de hipótesis.

Ejercicio 5.2.

Se tiene interés en estimar el porcentaje de personas con alguna discapacidad física en España (sabemos que en el resto de Europa es alrededor de un 3 %). Con este objetivo se ha tomado una muestra de 125 personas españolas aleatoriamente y se ha obtenido que en ella hay 5 personas con alguna tipo de discapacidad física.

¿Puede concluirse a partir de estos datos que el porcentaje de personas con alguna discapacidad física en España es diferente al del resto de Europa (3 %)?

Para responder a esta pregunta plantea el contraste de hipótesis correspondiente tomando como nivel de significatividad $\alpha = 0{,}05$. Resuelve el contraste, tanto con el método de la región de rechazo como mediante el cálculo del p-valor y explica tus conclusiones.

Comprueba que calculando el intervalo de confianza al 95 % para el porcentaje poblacional correspondiente, se obtiene la misma conclusión que has obtenido mediante el contraste de hipótesis.

Ejercicio 5.3.

Un equipo de cardiólogos tiene interés en estudiar la presión arterial en personas con diagnóstico de Alzheimer que toman un fármaco en fase de pruebas. Estos enfermos suelen tener una presión arterial media de 160 en condiciones normales, es decir, sin el uso del nuevo fármaco en prueba. Con el objetivo de valorar si el nuevo fármaco consigue disminuir la presión arterial de estos enfermos se toma la presión arterial de 15 personas con esta enfermedad que toman el nuevo fármaco y se obtiene en ellas una presión arterial media de 148 y una desviación típica de 26.

¿Puede concluirse a partir de los datos que en enfermos con este síndrome que toman el nuevo fármaco tienen una presión arterial media menor que 160?

a) Para responder a esta pregunta plantea el contraste de hipótesis correspondiente tomando como nivel de significatividad $\alpha = 0{,}05$. Resuelve el contraste según la región de rechazo y aceptación y calcula también el p-valor del contraste. Comprueba que llegas a la misma conclusión con las dos metodologías.

b) Ahora, repite el ejercicio considerando que es conocido que la desviación típica de la presión arterial de los enfermos de Alzheimer, en general, es 26 en condiciones normales. Reflexiona sobre los cambios que este dato produce en la resolución del ejercicio.

Ejercicio 5.4.

Se llevó a cabo un estudio sobre nutrición en un país en desarrollo. Se tomó una muestra aleatoria de 500 adultos de este país y se obtuvo un consumo medio de calorías de 1985 con una desviación típica de 210.

¿Puede concluirse a partir de estos datos que el consumo medio de calorías de la población adulta de este país es menor que 2000?

Para responder a esta pregunta plantea el contraste de hipótesis correspondiente tomando como nivel de significatividad $\alpha = 0{,}01$. Resuelve el contraste según la región de rechazo y aceptación y calcula también el p-valor del contraste. Comprueba que llegas a la misma conclusión con las dos metodologías.

Ejercicio 5.5.

Se puso en marcha en un barrio interior de una ciudad un programa de salud con el objetivo de estudiar la prevalencia de diferentes enfermedades de interés en la población. A partir de una muestra de 1500 residentes de ese barrio se obtuvo que 125 de ellos obtuvieron resultados positivos en cuanto a la anemia de células falciformes

¿Proporcionan estos datos evidencia suficiente que indique que el porcentaje (prevalencia) de individuos con dicha enfermedad en la población es mayor del 6 %?

Para responder a esta pregunta plantea el contraste de hipótesis correspondiente tomando como nivel de significatividad $\alpha = 0{,}05$. Resuelve el contraste según el método de la región de rechazo y calcula también el p-valor del contraste. Comprueba que llegas a la misma conclusión con las dos metodologías.

Ejercicio 5.6.

Se está realizando dentro de un programa de control de calidad en ciudades, el control del nivel de cloro en el agua de una determinada población. Se sabe que el nivel ideal es 325 unid. En este programa se revisan 150 grifos públicos y se midió el cloro en cada uno de ellos, obteniendo un nivel medio en la muestra de 332 y una desviación típica de 52.

¿Puede considerarse a partir de los datos que la media del nivel de cloro es distinta de 325 unidades?

Para responder a esta pregunta plantea el contraste de hipótesis correspondiente tomando como nivel de significatividad $\alpha = 0{,}05$. Resuelve el contraste según el método de las regiones de rechazo/aceptación y calcula también el p-valor del contraste. Comprueba que llegas a la misma conclusión con las dos metodologías.

Comprueba, además, que calculando el intervalo de confianza al 95 % para la media poblacional correspondiente se obtiene la misma conclusión que has obtenido con el contraste de hipótesis.

Ejercicio 5.7.

En una determinada comunidad autónoma, el porcentaje de personas en lista de espera era de un 8 %. Tras aplicar una nueva política en la gestión de las listas de espera, la consejería de Sanidad tiene interés en comprobar si dicha política había tenido algún efecto. Con tal fin ha tomado una muestra de 250 personas de la comunidad, de los que 10 han resultado estar en lista de espera en la Sanidad Pública por algún motivo.

¿Puede considerar la consejería de sanidad de la comunidad en cuestión que el porcentaje de personas en lista de espera actualmente es menor del 8 %?

Para responder a esta pregunta plantea el contraste de hipótesis correspondiente tomando como nivel de significatividad $\alpha = 0{,}05$. Resuelve el contraste según la región de rechazo y aceptación y calcula también el p-valor del contraste. Comprueba que llegas a la misma conclusión con las dos metodologías.

Ejercicios recopilatorios

Ejercicio 5.8.

En un proyecto que pretende estudiar a los enfermos de Parkinson en estadío 2, se tiene interés en estimar la longitud media de paso de estos enfermos tras la aplicación de tratamiento fisioterapeútico. Para valorar la eficacia del tratamiento se ha recogido una muestra de 13 de estos enfermos a los que se les ha estimado, tras la aplicación de tratamiento fisioterapeútico, la longitud de paso obteniendo los siguientes resultados (en cm):

41.9 55.2 61.8 47.9 49.5 52.4 54.7 38.8 47.5 50.9 50.8 61.7 55.6

a)¿Pueden concluir estos investigadores que la longitud de paso media en personas con Parkinson (en estadío 2) que reciben tratamiento fisioterapeútico es significativamente mayor de 45.9 cm? (lo comparamos con este valor porque la longitud media de paso en enfermos que no reciben tratamiento es de 45.9 cm.) Plantea y resuelve el contraste de hipótesis adecuado para responder a esta pregunta y explica tus conclusiones. Indica la fórmula que utilizas y la distribución. Calcula el p-valor del contraste. Utiliza un nivel de significatividad $\alpha = 0{,}05$.
b)Calcula la mediana y el rango intercuartílico de los datos recogidos.

Ejercicio 5.9.

Se está valorando la regeneración de cartílago en rodilla que consigue un nuevo tratamiento aplicado sobre enfermos con osteoartritis. El estudio ha mostrado los siguientes valores sobre la regeneración de cartílago (en cm^2) para 10 pacientes:

1.23, 1.53, 0.98, 0.56, 1.35, 1.45, 1.11, 1.01, 1.66, 0.78

a) Indica cuál es la población en estudio, cuál es la variable en estudio y el tipo de la misma.
b) El tratamiento estándar utilizado los últimos años conseguía una regeneración media de cartílago en este tipo de enfermos de 1 cm. ¿Pueden concluir estos investigadores que la regeneración media con el nuevo tratamiento es significativamente superior a la obtenida por el tratamiento estándar (1 cm^2)? Plantea y resuelve el contraste de hipótesis adecuado mediante la técnica del p-valor y explica tus conclusiones. Indica la fórmula que utilizas y la distribución. Utiliza un nivel de significatividad α=0.05.
c) Calcula el percentil 65 de los datos de la muestra e interprétalo.

Ejercicio 5.10.

Estudios de los últimos años han reflejado que en EEUU el 16 % de los niños padecen obesidad. Expertos españoles en la materia piensan que en España el porcentaje es superior. Para contrastar esta hipótesis han planteado un estudio en España, seleccionando 725 niños de los que 138 han sido considerados obesos.
a) Indica cuál es la población en estudio, cuál es la variable en estudio y el tipo de la misma.
b) ¿Pueden concluir estos investigadores que el porcentaje de niños obesos en España es significativamente superior al 16 %? Plantea y resuelve el contraste de hipótesis adecuado mediante la técnica de las regiones de aceptación-rechazo y explica tus conclusiones. Indica la fórmula que utilizas y la distribución. Utiliza un nivel de significatividad α=0.01.

Ejercicio 5.11.

Es sabido que un gran porcentaje de hemipléjicos padecen dolor de hombro durante los 12 meses siguientes al ictus. Se desea investigar si una novedosa terapia rehabilitadora, basada en la fisioterapia, reduce significativamente el tiempo medio de dolor de hombro. Para llevar a cabo el estudio se aplica la terapia a 8 hemipléjicos, con dolor de hombro, durante el tiempo necesario hasta corregir el problema de dolor. A continuación aparece el tiempo, en meses, que han recibido la terapia:

5.4, 7.3, 14.5, 8.1, 10, 11.7, 9.2, 7.4

a) Determina cuál es la población de estudio, la muestra, la variable de interés, el tipo de la misma y el parámetro de interés. b) Plantea el contraste necesario para averiguar si la terapia es efectiva y resuélvelo según la metodología de las regiones de rechazo y aceptación. (α=0.05). Explica las conclusiones que se deducen en el contexto del ejercicio.

Capítulo 6

Comparación de dos grupos

El objetivo de este capítulo es conocer las técnicas adecuadas para la comparación de los parámetros de dos poblaciones. Concretamente abordaremos el estudio de la comparación de dos varianzas poblacionales, dos medias y dos proporciones.

6.1. Comparación de dos proporciones

La situación en estudio estará compuesta por dos poblaciones independientes, *Población 1* y *Población 2*. El objetivo es comparar los porcentajes de cierta respuesta de una variable cualitativa de interés en cada una de ellas, a los que llamaremos P_1 y P_2. Para llevar a cabo esta comparación dispondremos de dos muestras, una de cada una de las poblaciones en estudio, con tamaños que denotaremos n_1 y n_2. En cada una de esas muestras podremos obtener el porcentaje de interés: $\widehat{P}_1$ y $\widehat{P}_2$.

Para llevar a cabo la comparación, podemos tanto plantear un contraste de hipótesis (unilateral o bilateral dependiendo de la situación):

$$H_0 : P_1 = P_2$$

$$H_1 : P_1 \neq P_2 \ \ (H_1 : P_1 < P_2 \ \ ; \ \ H_1 : P_1 > P_2)$$

o bien calcular un intervalo de confianza para la diferencia de ambos porcentajes, es decir, para $P_1 - P_2$.

En ambos casos, necesitamos una distribución en el muestreo de los estadísticos involucrados en este problema. Una aproximación normal ampliamente utilizada es la que se plantea a continuación:

$$\widehat{P}_1 - \widehat{P}_2 \sim N(P_1 - P_2, \sqrt{\frac{P_1 \cdot (100 - P_1)}{n_1} + \frac{P_2 \cdot (100 - P_2)}{n_2}})$$

A partir de esta distribución, podemos resolver tanto los contrastes de hipótesis como calcular los intervalos de confianza planteados. Para ello, será útil tipificar esta expresión para poder trabajar con la distribución $N(0, 1)$:

$$\frac{(\widehat{P}_1 - \widehat{P}_2) - (P_1 - P_2)}{\sqrt{\frac{P_1 \cdot (100 - P_1)}{n_1} + \frac{P_2 \cdot (100 - P_2)}{n_2}}} \sim N(0, 1) \tag{6.1}$$

Para usar la expresión anterior, debemos aproximar la desviación típica poblacional (denominador) por la desviación típica muestral. Aunque hay diferentes criterios para realizar esta aproximación, que pueden incluso ser diferentes en contrastes de hipótesis e intervalos de confianza, clásicamente se utiliza la siguiente:

$$\frac{(\widehat{P}_1 - \widehat{P}_2) - (P_1 - P_2)}{\sqrt{\frac{\widehat{P}_1 \cdot (100 - \widehat{P}_1)}{n_1} + \frac{\widehat{P}_2 \cdot (100 - \widehat{P}_2)}{n_2}}} \sim N(0, 1) \tag{6.2}$$

Esta expresión será utilizada como *Pivote* en los contrastes de hipótesis y a partir de ella se puede deducir la siguiente fórmula para los intervalos de confianza para la diferencia de porcentajes de dos poblaciones:

$$\left[\ (\widehat{P}_1 - \widehat{P}_2) \pm Z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\widehat{P}_1 \cdot (100 - \widehat{P}_1)}{n_1} + \frac{\widehat{P}_2 \cdot (100 - \widehat{P}_2)}{n_2}} \ \right] \tag{6.3}$$

Ejemplo 6.1.

Se ha planificado un ensayo clínico de un nuevo producto farmacéutico contra la hipertensión. Se ha probado el fármaco tradicional sobre 64 personas de las cuales 12 han presentado efectos secundarios, mientras que el nuevo fármaco ha sido probado sobre 51 personas y 5 de ellas han presentado efectos secundarios. Realiza el contraste de hipótesis adecuado para contestar si el porcentaje de personas que tendrán efectos secundarios con el fármaco nuevo es significativamente diferente al porcentaje de personas que lo tendrán con el fármaco tradicional. Utiliza como nivel de significatividad $\alpha = 0{,}05$

El contraste que debemos plantear es el siguiente.

$$H_0 : P_1 = P_2$$

$$H_1 : P_1 \neq P_2$$

Los datos que podemos obtener del enunciado son los siguientes:

$$\widehat{P}_1 = \frac{12}{64} \cdot 100 = 18{,}75\,\% \quad n_1 = 64$$

$$\widehat{P}_2 = \frac{5}{51} \cdot 100 = 9{,}80\,\% \quad n_2 = 51$$

El pivote que utilizaremos para resolver este contraste es el aproximado dado por la expresión (6.2). En este pivote sustituiremos por un lado los estadísticos obtenidos de los datos. Y por otro el valor de $P_1 - P_2$ lo sustituiremos por 0, ya que según la hipótesis nula estamos asumiendo cierto que $P_1 = P_2$. Así, el pivote en nuestro caso quedará:

$$Pivote = \frac{(18{,}75 - 9{,}80) - (0)}{\sqrt{\frac{18{,}75 \cdot (100 - 18{,}75)}{64} + \frac{9{,}80 \cdot (100 - 9{,}80)}{51}}} = 1{,}40$$

A partir de este valor podemos resolver el contraste de cualquiera de las dos formas que explicamos en el capítulo anterior, o bien a través de las regiones de rechazo, o bien a través del P-valor.
En este caso, si recurrimos al P-valor, podemos comprobar que la probabilidad de encontrar un valor superior a 1,40 en la distribución $N(0, 1)$ es $1 - 0{,}9192 = 0{,}0808$. Como el contraste es bilateral:

$$P - valor = 0{,}0808 \cdot 2 = 0{,}1616$$

Como $P - valor > \alpha$, no podemos rechazar H_0, y por tanto no podemos concluir que el nuevo fármaco provoque a un porcentaje de personas efectos secundarios significativamente diferente a lo que lo hace el fármaco tradicional.

Ejemplo 6.2.

Con los datos del ejemplo anterior se pide: calcular un intervalo de confianza al 95 % para la diferencia de porcentajes de personas que tendrán efectos secundarios entre el fármaco tradicional y el nuevo

Los datos que podemos obtener del enunciado son los siguientes:

$$\widehat{P}_1 = \frac{12}{64} \cdot 100 = 18{,}75\,\% \quad n_1 = 64$$

$$\widehat{P}_2 = \frac{5}{51} \cdot 100 = 9{,}80\,\% \quad n_2 = 51$$

y aplicando la fórmula para los intervalos de confianza dada por la expresión (6.3) hallamos el intervalo:

$$\left((18{,}75 - 9{,}80) \pm 1{,}96 \cdot \sqrt{\frac{18{,}75 \cdot (100 - 18{,}75)}{64} + \frac{9{,}80 \cdot (100 - 9{,}80)}{51}} \right) =$$

$$= (-3{,}62; 21{,}52)$$

$$P_1 - P_2 \in (-3{,}62\,\%; 21{,}52\,\%)$$

Con un 95 % de confianza la diferencia de porcentajes de personas que tienen efectos secundarios entre el fármaco tradicional y fármaco nuevo estará contenida en este intervalo. Como el intervalo está formado por un extremo de signo negativo y otro de signo positivo, no tenemos evidencias suficientes para afirmar que existen diferencias significativas entre los dos fármacos en cuanto al porcentaje de personas que presentan efectos secundarios.

6.2. Comparación de dos varianzas

Supongamos que queremos comparar las varianzas poblacionales (σ_1^2 y σ_2^2) de dos poblaciones **normales**. Para ello dispondremos de dos muestras, una de cada población, y por tanto tendremos conocimiento de los dos tamaños de ambas muestras (n_1 y n_2) y de las varianzas muestrales obtenidas (S_1^2 y S_2^2).

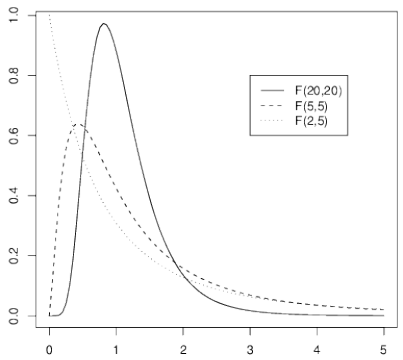
El contraste de hipótesis que nos plantearemos será el siguiente:

$$H_0 : \sigma_1^2 = \sigma_2^2$$

$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

6.2.1. Distribución F de Snedecor

Para resolver este contraste de hipótesis, necesitamos incorporar una nueva distribución, la distribución *F de Snedecor* o *F de Fisher-Snedecor*. Esta distribución, a diferencia de las distribuciones *Normal* o *t-Student* solo está definida para valores positivos y no tiene forma de *campana simétrica*. La distribución F está regida por dos parámetros, m y n, llamados *grados de libertad*, y se suele representar como $F_{(m,n)}$. A continuación se muestran algunas representaciones gráficas de distribuciones F con diferentes grados de libertad.



Para trabajar con esta distribución, necesitaremos tablas numéricas de ayuda que contengan los percentiles de la misma para algunos grados de libertad. Estas tablas se muestran en el Anexo de *Tablas Estadísticas*. Concretamente utilizaremos las tablas de los percentiles: 0,900, 0,950, 0,975, 0,990, 0,995. Para hallar los percentiles opuestos a estos (0,100, 0,050, 0,025, 0,010, 0,005)) no hay más que tener en cuenta que, para cualquier probabilidad γ:

$$F_{(m,n),\gamma} = \frac{1}{F_{(n,m),1-\gamma}}$$

Por tanto, para calcular $F_{(m,n),\frac{\alpha}{2}}$ podemos usar la siguiente expresión

$$F_{(m,n),\frac{\alpha}{2}} = \frac{1}{F_{(n,m),1-\frac{\alpha}{2}}}$$

.

6.2.2. Resolución del contraste de hipótesis

Una vez introducida la distribución F, para resolver el contraste de hipótesis planteado necesitaremos la distribución en el muestreo de un estadístico que sea conocida bajo la hipótesis nula. Concretamente utilizaremos el estadístico **F** que definimos a continuación:

$$\mathbf{F} = \frac{\frac{S_1^2}{\sigma_1^2}}{\frac{S_2^2}{\sigma_2^2}} = \frac{S_1^2 \cdot \sigma_2^2}{S_2^2 \cdot \sigma_1^2} = \left(\frac{S_1^2}{S_2^2}\right) \cdot \left(\frac{\sigma_2^2}{\sigma_1^2}\right) \sim F_{(n_1-1,n_2-1)}$$

Este estadístico, **F**, es el que utilizaremos como *Pivote*, y bajo la hipótesis nula ($H_0 : \sigma_1^2 = \sigma_2^2$), la expresión $\left(\frac{\sigma_2^2}{\sigma_1^2}\right)$ tomará el valor 1. Así, la expresión **F** bajo la hipótesis nula se calculará simplemente como $\mathbf{F} = \left(\frac{S_1^2}{S_2^2}\right)$ y su distribución conocida será una $F_{(n_1-1,n_2-1)}$

Notar que para la aplicación de estas técnicas las poblaciones de origen deben ser asumidas como *Normales*. Existen otros test para contrastar la igualdad de varianzas que no exigen la condición de normalidad (por ejemplo el *test de Levene*). Estos tests no serán vistos en la parte teórica de esta asignatura, aunque alguno de ellos se verá en la parte práctica con software informático.

Ejemplo 6.3.

Se quiere valorar el tiempo (en minutos) que tardan en realizar una determinada tarea pacientes operados por dos técnicas quirúrgicas diferentes. Se supone que esta variable sigue una distribución* Normal *en ambas poblaciones (*Intervenidos por técnica 1 *e* Intervenidos por técnica 2*). Se han tomado datos de* 21 *pacientes intervenidos por la* técnica 1 *y de* 16 *pacientes intervenidos por la* técnica 2 *obteniendo unas varianzas en ambas muestras de* 50 *y* 24 *respectivamente. Contrasta si la varianza de esta variable en ambas poblaciones es significativamente diferente (considera* $\alpha = 0{,}05$*).

Con los datos disponibles: $n_1 = 21 \quad S_1^2 = 50$ y $n_2 = 16 \quad S_2^2 = 24$ planteamos el contraste de hipótesis:

$$H_0 : \sigma_1^2 = \sigma_2^2$$
$$H_1 : \sigma_1^2 \neq \sigma_2^2$$

El pivote, bajo la hipótesis nula resulta:

$$F = \left(\frac{S_1^2}{S_2^2}\right) = \frac{50}{24} = 2{,}08$$

y debe seguir una distribución $F_{(20,15)}$
A continuación calculamos las regiones de rechazo: $F_{(20,15)0{,}975} = 2{,}7559$ y $F_{(20,15)0{,}025} = \frac{1}{F_{(15,20)0{,}975}} = \frac{1}{2{,}5731} = 0{,}3886$
Dado que el pivote (2,08) pertenece a la región de aceptación, se puede asumir que no hay diferencias significativas entre las varianzas de la variable en ambas poblaciones. Es decir, entre las varianzas de los tiempos en realizar la actividad de los intervenidos con las técnicas quirúrgicas 1 y 2.

6.3. Comparación de dos medias

Existen diferentes técnicas estadísticas que permiten comparar los valores de dos variables cuantitativas. Concretamente, este problema se plantea como una comparación de los valores medios de ambas variables y suele resolverse, en caso de que las condiciones lo permitan, mediante la utilización de diferentes versiones del test t de Student. Estas técnicas requieren, entre alguna otra condición, comportamiento *Normal* de las variables que van a ser comparadas. El estudio de esta condición mediante alguna técnica estadística se explicará en el siguiente capítulo, aunque a priori puede valorarse mediante el conocimiento de los valores de las variables a comparar. Si por ejemplo, estamos analizando el **número de crías** de cerdas de dos razas en una granja y los valores habituales son 1,3,5,6,...díficilmente el comportamiento de esta variable va a ser aproximadamente *Normal*. Sin embargo, si estamos midiendo el **peso de las crías al nacer** de las dos razas, es muy probable que esta variable sí tenga un comportamiento similar a la Normalidad.

Para aquéllas variables cuyo comportamiento no sea *Normal* se suelen aplicar técnicas estadísticas llamadas *Técnicas no paramétricas*, en contraposición a las sí vamos a estudiar a continuación, que se enmarcan dentro de las llamadas *Técnicas paramétricas*. Más concretamente, en el problema de comparación de dos variables cuantitativas que no tienen comportamiento Normal es habitual utilizar el test de Wilcoxon. En el caso que nos ocupa, es decir, en el caso de que las variables que van a ser compardas sí tienen un comportamiento aproximadamente Normal, explicaremos a continuación diferentes versiones de la prueba paramétrica **test t de Student**.

Para llevar a cabo la comparación de los valores de dos variables cuantitativas que tienen un comportamiento Normal, es habitual realizar la comparación de sus valores medios, es decir, de sus medias poblacionales.

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2 \;\; (H_1 : \mu_1 < \mu_2 \;\; ; \;\; H_1 : \mu_1 > \mu_2)$$

Esta comparación se realiza de dos formas diferentes según las muestras disponibles sean *Muestras independientes* o *Muestras dependientes o relacionadas*. Dos muestras están relacionadas si existe entre los elementos de ambas muestras una relación que pudiera establecer dependencia entre los valores obtenidos entre las dos variables. Como ejemplo podemos citar aquéllos experimentos que realizan una medición a los individuos y, tras una intervención o simplemente tras un periodo de tiempo, vuelven a realizar otra medición. Los individuos son los mismos para las mediciones realizadas antes y después, por lo que son *dependientes* unas de las otras en cada individuo. Este tipo de muestras (relacionadas), requieren un tratamiento estadístico especial, a diferencia del resto de casos en los que las muestras a comparar son totalmente independientes.

6.3.1. Muestras independientes. Varianzas poblacionales conocidas e iguales

La igualdad de varianzas es una **premisa** para valorar la igualdad de medias. Asumimos que σ_1^2 y σ_2^2 son parámetros desconocidos pero iguales, es decir $\sigma_1^2 = \sigma_2^2 = \sigma^2$, donde σ^2 representa la varianza común de ambas poblaciones. Para asumir que las varianzas poblacionales son iguales, o bien tenemos información previa que nos permita asumir como cierta esta hipótesis, o bien podríamos realizar un contraste de igualdad de varianzas y en el caso de no rechazar en este contraste la igualdad de las mismas podríamos asumir que esta hipótesis es cierta (o no se desvía mucho de serlo).

La distribución en el muestreo de partida es la siguiente:

$$\overline{x}_1 - \overline{x}_2 \sim N(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}})$$

de la cuál podemos deducir:

$$\frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1) \tag{6.4}$$

Si la varianza común $\sigma^2 = \sigma_1^2 = \sigma_2^2$ es conocida, esta distribución en el muestreo es la que podemos utilizar directamente en los contrastes de hipótesis sobre la diferencia de medias y a partir de ella podemos extraer la siguiente fórmula para el cálculo de intervalos de confianza (a un nivel $(1-\alpha)\cdot 100\,\%$ de confianza) para la diferencia de medias $(\mu_1 - \mu_2)$:

$$\left[\; (\overline{x}_1 - \overline{x}_2) \pm Z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \; \right] \tag{6.5}$$

6.3.2. Muestras independientes. Varianzas poblacionales desconocidas pero pudiéndose asumir iguales

En cualquier caso, es habitual no conocer las varianzas poblacionales y los datos disponibles en esta situación son los basados en las dos muestras, es decir, solemos conocer n_1, $\overline{x}_1$, S_1^2, n_2, $\overline{x}_2$, S_2^2

Calcularemos un estimador S^2 para la varianza común σ^2 a partir de las dos varianzas muestrales S_1^2 y S_2^2 de la siguiente forma:

$$S^2 = \frac{(n_1 - 1)\cdot S_1^2 + (n_2 - 1)\cdot S_2^2}{(n_1 - 1) + (n_2 - 1)} = \frac{(n_1 - 1)\cdot S_1^2 + (n_2 - 1)\cdot S_2^2}{(n_1 + n_2 - 2)} \tag{6.6}$$

Y utilizaremos la siguiente distribución en el muestreo:

$$\frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S^2}{n_1} + \frac{S^2}{n_2}}} = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{S \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2} \tag{6.7}$$

A partir de esta expresión, que es la que se puede utilizar como pivote en los contrastes de hipótesis correspondientes también podemos extraer la correspondiente fórmula para el cálculo de los intervalos de confianza (a un nivel $(1-\alpha)\cdot 100\,\%$ de confianza) para la diferencia de medias $(\mu_1 - \mu_2)$:

$$\left[\; (\overline{x}_1 - \overline{x}_2) \pm t_{\left((n_1+n_2-2)\;(1-\frac{\alpha}{2})\right)} \cdot S \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \; \right] \tag{6.8}$$

Ejemplo 6.4.

Se está realizando un estudio que pretende comparar el nivel medio de calcio en plasma sanguíneo en hombres y mujeres. Así, de los 18 casos que disponemos 10 de ellos son hombres y 8 de ellos mujeres, obteniendo que el nivel medio para los hombres es 3.6 mmol/l con una desviación típica de los datos de 0.9 mmol/l mientras que para las mujeres el nivel medio es 2.9 con una desviación típica en los datos de 1.2 mmol/l. ¿Es significativa la diferencia obtenida en el nivel medio de calcio entre hombres y mujeres (α=0.05)?

Los datos disponibles consisten en:
Hombres: $n_h = 10$, $\overline{x}_h = 3{,}6$, $S_h = 0{,}9$
Mujeres: $n_m = 8$, $\overline{x}_m = 2{,}9$, $S_m = 1{,}2$

En primer lugar comprobamos si podemos asumir como cierta la hipótesis de que las varianzas de esta variable en ambas poblaciones son iguales, es decir, que $\sigma_h^2 = \sigma_m^2$. Para ello planteamos el contraste:

$$H_0 : \sigma_h^2 = \sigma_m^2$$
$$H_1 : \sigma_h^2 \neq \sigma_m^2$$

El pivote, bajo la hipótesis nula resulta:

$$pivote = \left(\frac{S_1^2}{S_2^2}\right) = \frac{(0{,}9)^2}{(1{,}2)^2} = 0{,}5625 \sim F_{n_1-1,n_2-1} = F_{9,7}$$

.
En esta distribución, las regiones de rechazo vienen determinadas por:
los valores que quedan a la derecha de $F_{(9,7)0,975} = 4{,}8232$ y los valores que quedan a la izquierda de $F_{(9,7)0,025} = \frac{1}{F_{(7,9)0,975}} = \frac{1}{4{,}1970} = 0{,}2383$
Dado que el pivote (0,5625) no está en la región de rechazo, no podemos concluir que haya diferencias significativas entre las varianzas de la variable en ambas poblaciones, es decir, entre las varianzas del nivel de calcio en plasma sanguíneo entre hombres y mujeres y por tanto, no existe diferencias significativas entre las varianzas y se pueden comparar las medias en ambos grupos poblacionales (que es el objetivo del problema planteado).
A continuación, puesto que podemos asumir que las varianzas de la variable en ambas poblaciones son iguales, calculamos una estimación de la varianza común mediante la expresión (6.6):

$$S = \sqrt{\frac{(10-1)\cdot(0{,}9)^2 + (8-1)\cdot(1{,}2)^2}{(10+8-2)}} = 1{,}04$$

μ_h y μ_m representarán el nivel medio de calcio en plasma sanguíneo en hombres y mujeres, respectivamente. Para realizar la comparación planteamos el contraste que corresponde:

$$H_0 : \ \mu_h = \mu_m$$

$$H_1 : \ \mu_h \neq \mu_m$$

Si la hipótesis nula es cierta, y utilizando la expresión (6.7):

$$\frac{(3{,}6-2{,}9)-(0)}{1{,}04\cdot\sqrt{\frac{1}{10}+\frac{1}{8}}} = 1{,}42 \ \ \sim t_{16}$$

La región de rechazo de este contraste bilateral está compuesta por aquellos valores menores a $-2{,}119$ y mayores a 2,119. Puesto que el valor de nuestro pivote es 1.42, no podemos rechazar la hipótesis nula. Por tanto, podemos concluir que el nivel medio del calcio en plasma sanguíneo no es significativamente diferente en hombres y mujeres.

Ejemplo 6.5.

Si en la misma situación que en el ejemplo anterior, decidimos plantearlo desde el punto de vista de los intervalos de confianza, calcularíamos un intervalo de confianza, por ejemplo al 95 % para la diferencia de niveles medios de calcio en plasma sanguíneo entre hombres y mujeres

Con los datos del ejemplo anterior:
Hombres: $n_h = 10$, $\overline{x}_h = 3{,}6$, $S_h = 0{,}9$
Mujeres: $n_m = 8$, $\overline{x}_m = 2{,}9$, $S_m = 1{,}2$
Con el cálculo realizado $S = 1{,}04$, y con la fórmula dada por la expresión (6.8), obtendríamos el intervalo de confianza al 95 %:

$$\mu_h - \mu_m \in \left[\ (3{,}6 - 2{,}9) \pm 2{,}119 \cdot 1{,}04 \cdot \sqrt{\frac{1}{10} + \frac{1}{8}} \ \right] = [\ -0{,}3453\ ,\ 1{,}7453\ ]$$

$$\mu_h - \mu_m = 0$$

$$\mu_h = \mu_m$$

Con un 95 % de confianza, la diferencia de medias estará contenida en este intervalo. Como en este intervalo un extremo es negativo y el otro positivo no podemos concluir que haya diferencias significativas entre los niveles medios de calcio en plasma sanguíneo de hombres y mujeres.

6.3.3. Muestras independientes. Varianzas poblacionales desconocidas y no pudiéndose asumir iguales

Cuando las varianzas de ambas poblaciones son desconocidas, pero además, no pueden ser asumidas iguales se pueder utilizar una distribución t de Student con grados de libertad aproximados y dependientes de las varianzas muestrales. Concretamente, podemos utilizar la siguiente distribución en el muestreo:

$$\frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim t_{gl} \tag{6.9}$$

donde $gl = \dfrac{\left(\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}\right)^2}{\dfrac{\left(\frac{S_1^2}{n1}\right)^2}{n_1 - 1} + \dfrac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1}}$

Así, si tras realizar la comprobación de la hipótesis de homogeneidad de varianzas resultara que no es posible asumirla dado que en el contraste de hipótesis de varianzas se ha rechazado la hipótesis nula (de igualdad de varianzas), el pivote que debeeríamos utilizar para la comparación de medias es el que se muestra en la ecuación (6.9).

De la misma forma, en esta situación la expresión del intervalo de confianza para la diferencia de medias $(\mu_1 - \mu_2)$ viene dado por la fórmula:

$$\left[\ (\overline{x}_1 - \overline{x}_2) \pm t_{((gl)\ (1-\frac{\alpha}{2}))} \cdot \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \ \right] \tag{6.10}$$

donde gl, que de nuevo representa los grados de libertad aproximados de la distribución t de Student se obtienen a partir de la expresión $gl = \dfrac{\left(\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}\right)^2}{\dfrac{\left(\frac{S_1^2}{n1}\right)^2}{n_1 - 1} + \dfrac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1}}$

Ejemplo 6.6.

Se está realizando un estudio que pretende comparar el nivel medio de calcio en plasma sanguíneo en hombres y mujeres. Así, de los 18 casos que disponemos 10 de ellos son hombres y 8 de ellos mujeres, obteniendo que el nivel medio para los hombres es 3.6 mmol/l con una desviación típica de los datos de 0.9 mmol/l mientras que para las mujeres el nivel medio es 2.9 con una desviación típica en los datos de 2.1 mmol/l. ¿Es significativa la diferencia obtenida en el nivel medio de calcio entre hombres y mujeres (α=0.05)?

Los datos disponibles consisten en:
Hombres: $n_h = 10$, $\overline{x}_h = 3{,}6$, $S_h = 0{,}9$
Mujeres: $n_m = 8$, $\overline{x}_m = 2{,}9$, $S_m = 2{,}1$

En primer lugar comprobamos si podemos asumir como cierta la hipótesis de que las varianzas de esta variable en ambas poblaciones son iguales, es decir, que $\sigma_h^2 = \sigma_m^2$. Para ello planteamos el contraste:

$$H_0 : \sigma_h^2 = \sigma_m^2$$
$$H_1 : \sigma_h^2 \neq \sigma_m^2$$

El pivote, bajo la hipótesis nula resulta:

$$pivote = \left(\frac{S_1^2}{S_2^2}\right) = \frac{(0{,}9)^2}{(2{,}1)^2} = 0{,}1837 \sim F_{n_1-1,n_2-1} = F_{9,7}$$

.
En esta distribución, las regiones de rechazo vienen determinadas por:
los valores que quedan a la derecha de $F_{(9,7)0,975} = 4{,}8232$ y los valores que quedan a la izquierda de $F_{(9,7)0,025} = \frac{1}{F_{(7,9)0,975}} = \frac{1}{4{,}1970} = 0{,}2383$
Dado que el pivote (0,1837) no está en la región de aceptación, no podemos asumir la igualdad de las varianzas de la variable en ambas poblaciones, es decir, entre las varianzas del nivel de calcio en plasma sanguíneo entre hombres y mujeres existen diferencias significativas.

A continuación, para realizar la comparación de medias, planteamos el contraste que corresponde:

$$H_0 : \ \mu_h = \mu_m$$

$$H_1 : \ \mu_h \neq \mu_m$$

Puesto que no podemos asumir que las varianzas de la variable en ambas poblaciones sean iguales, se calcula el pivote mediante la expresión (6.9):

$$\frac{(3{,}6 - 2{,}9) - (0)}{\sqrt{\frac{0{,}9^2}{10} + \frac{2{,}1^2}{8}}} = 0{,}88 \quad \sim t_{gl}$$

donde

donde $gl = \dfrac{\left(\dfrac{0{,}9^2}{10} + \dfrac{2{,}1^2}{8}\right)^2}{\dfrac{\left(\frac{0{,}9^2}{10}\right)^2}{9} + \dfrac{\left(\frac{2{,}1^2}{8}\right)^2}{7}} = 9{,}07 \approx 9$

La región de rechazo de este contraste bilateral está compuesta por aquellos valores menores a $-2{,}262$ y mayores a 2,262. Puesto que el valor de nuestro pivote es 0.88, no podemos rechazar la hipótesis nula. Por tanto, podemos concluir que el nivel medio del calcio en plasma sanguíneo no es significativamente diferente en hombres y mujeres.

6.3.4. Muestras dependientes o pareadas

Diremos que 2 muestras son pareadas **si existe alguna relación entre los elementos de ambas muestras** que pudiera establecer dependencia entre los valores obtenidos de la variable de estudio. Por ejemplo, si queremos evaluar los efectos de una dieta sobre el peso corporal en cierta población tomaremos el peso a un conjunto de individuos antes de someterlos a dieta. Tras el periodo de dieta pesamos nuevamente a los integrantes del estudio obteniendo así una segunda medición del peso en cada individuo. Así obtenemos 2 muestras de pesos de la población, pero estas 2 muestras tienen una peculiaridad y es que los individuos que las componen están relacionados, es más son los mismos individuos. En este caso diremos que las muestras están pareadas. Para este tipo de problemas en lugar de plantearnos un contraste habitual sobre la igualdad de medias como el que acabamos en apartados anteriores, restaríamos las 2 mediciones efectuadas a cada persona (o cada par de mediciones relacionadas), de esta forma obtendremos una única muestra de diferencias y contrastaremos si la media de estas diferencias es distinta de 0 o no. Así conseguimos que las observaciones de la variable sean independientes entre sí, reduciendo así cualquier efecto que pudiera tener esta dependencia sobre los resultados del estudio.

Las técnicas a utilizar, por tanto, son las vistas en el tema 5 si se quiere plantear mediante un contraste de hipótesis y las del tema 4 si se quisiera plantear desde la perspectiva de los intervalos de confianza.

Ejemplo 6.7.

Se planifica un ensayo clínico para valorar la eficacia de un nuevo tratamiento antihipertensivo. Este tratamiento se sospecha que podría tener unos efectos secundarios considerables, por ello hemos conseguido únicamente 14 pacientes dispuestos a integrar el estudio. Hemos tomado la presión arterial de los pacientes antes y después de someterse al tratamiento obteniendo los siguientes valores:

	Antes Trat.	***Después Trat.***	***Diferencia***
Paciente 1	***188***	***176***	***12***
Paciente 2	***210***	***208***	***2***
Paciente 3	***202***	***193***	***9***
Paciente 4	***188***	***185***	***3***
Paciente 5	***176***	***177***	***-1***
Paciente 6	***171***	***174***	***-3***
Paciente 7	***186***	***176***	***10***
Paciente 8	***192***	***182***	***10***
Paciente 9	***200***	***196***	***4***
Paciente 10	***176***	***157***	***19***
Paciente 11	***197***	***191***	***6***
Paciente 12	***185***	***183***	***2***
Paciente 13	***194***	***189***	***5***
Paciente 14	***207***	***191***	***16***
Media			$\overline{x}_d = 6{,}71$
Desv.Típica			$S_d = 6{,}29$

Contrasta la hipótesis que el tratamiento realmente ha producido una disminución significativa (α=0.05) de la presión arterial.

Para contrastar la efectividad del tratamiento trabajaremos con la variable de diferencias *Antes - Después*. Si el tratamiento no ha producido ningún efecto la media de esta variable en la población (la llamaremos μ_d) debería ser 0, mientras que si el tratamiento ha tenido el efecto esperado la media de esta variable en la población debería ser mayor que 0 (indicando que la presión arterial media antes del tratamiento es mayor que la de después). Así, plantearemos el contraste:

$$H_0 : \mu_d = 0$$

$$H_1 : \mu_d > 0$$

Utilizando $\frac{\overline{x}_d - \mu_d}{\frac{S_d}{\sqrt{n}}} = \frac{6{,}71-0}{\frac{6{,}29}{\sqrt{14}}} = 3{,}99 \sim t_{n-1} = t_{13}$. Como $\alpha = 0{,}05$ y el contraste es unilateral, comprobamos que el percentil 95 % para la distribución t_{13} es aproximadamente 1,77 y por tanto nuestro pivote toma un valor de la región de rechazo para este contraste. Así, rechazamos la hipótesis nula y por tanto encontramos evidencias suficientes para concluir que el tratamiento produce una disminución significativa de la presión arterial media.

6.4. Ejercicios Capítulo 6

Diferencia de Porcentajes

Ejercicio 6.1.

Un experimento se plantea el estudio de la efectividad de una nueva vacuna frente al SIDA. El experimento se encuentra en la primera fase, en la que se está valorando dicha efectividad en monos. La vacuna se administró a 60 monos, de los que, tras estar en contacto con el virus VIH, se comprobó que 4 resultaron infectados. Por otro lado, se trabajó con un grupo control de 50 monos que no recibieron la vacuna y que también estuvieron en contacto con el virus VIH, de los que los 15 resultaron infectados. Las diferencias como puedes apreciar son notables, pero ¿podemos concluir que la vacuna es efectiva? Es decir, ¿el porcentaje de infección en monos vacunados es significativamente inferior que en los no vacunados?

Platea el contraste de hipótesis adecuado para responder a esta pregunta y resuélvelo tanto por el método de las regiones de aceptación/rechazo como por el método del p-valor utilizando como nivel de significatividad α=0.01.

Ejercicio 6.2.

En un estudio sobre niños de un año de edad se seleccionaron niños de los dos grupos étnicos predominantes que constituían la clientela de un determinado departamento de salud con el objetivo de comparar la prevalencia de un tipo de anemia en ambos grupos. En el *grupo étnico 1* se seleccionaron 450 niños, de los cuales 105 presentaron indicios de anemia, mientras que en el *grupo étnico 2* se seleccionaron 375 niños de los cuales 120 presentaron rasgos de anemia. ¿Proporcionan estos datos evidencia suficiente que indique que existe una diferencia en las dos poblaciones con respecto al porcentaje de anémicos en las mismas? Plantea y resuelve el contraste de hipótesis adecuado para responder a esta pregunta utilizando un nivel de significatividad $\alpha = 0{,}05$.

Ejercicio 6.3.

Se está probando la eficacia de dos tipos de ejercicio para mejorar los síntomas de la artritis reumatoide. El primer tratamiento (T1) ha sido probado en 150 pacientes con esta enfermedad obteniendo que 87 de ellos mejoran tras un mes de práctica. El segundo tratamiento en prueba (al que llamaremos T2) ha sido probado en 170 pacientes de los que 90 han mejorado tras un mes de práctica. Calcula un intervalo de confianza al 99 % para la diferencia de porcentajes de mejoría de ambos tratamientos e interpreta los resultados. A la vista del resultado, ¿crees que el porcentaje de personas que mejoran con el tratamiento T1 es significativamente superior al del T2?. Razona la respuesta.

Ejercicio 6.4.

Con los datos del ejercicio anterior plantea el contraste de hipótesis correspondiente para averiguar si el porcentaje de personas que mejoran con el tratamiento T1 es significativamente superior al porcentaje de personas que mejoran con el tratamiento T2 (considera *alpha*=0.01)

Ejercicio 6.5.

Un organismo sanitario trata de valorar la calidad de servicio de dos hospitales públicos que dependen del mismo (a los que llamaremos *Hospital A* y *Hospital B*). Para ello seleccionó, al azar, una muestra de 150 personas de entre todos los pacientes hospitalizados en el *Hospital A* durante dos últimos años, de los que 129 valoraron el servicio como *Muy favorable* (calificación máxima). De una muestra de 160 pacientes seleccionados de forma similar del *Hospital B*, 144 de ellos calificaron el servicio recibido como *Muy favorable*. Calcula el intervalo de confianza al 99 % para la diferencia de porcentajes de máxima satisfacción entre los usuarios de ambos hospitales . A la vista del resultado, ¿piensas que existen diferencias significativas entre ambos porcentajes?. Platea el contraste de hipótesis adecuado y resuélvelo tanto por el método de las regiones de aceptación/rechazo como por el método del p-valor utilizando como nivel de significatividad α=0.01.

Ejercicio 6.6.

En una encuesta conducida por un grupo de salud bucodental, se les pidió a 500 adultos que dieran la razón de su última visita al dentista. De los 220 que tenían una educación inferior a la secundaria, 44 señalaron que lo habían hecho por razones preventivas. De los restantes 280, los cuales tenían educación secundaria o un nivel superior, 150 señalaron que lo habían hecho por la misma razón. Platea el contraste de hipótesis adecuado y resuélvelo tanto por el método de las regiones de aceptación/rechazo como por el método del p-valor utilizando como nivel de significatividad α=0.1. Construye un intervalo de confianza al 90 % por ciento para la diferencia entre los porcentajes de personas que acuden al dentista por razonas preventivas de las dos poblaciones en estudio (personas con estudios inferiores y iguales o superiores a educación secundaria). Interpreta el significado del intervalo. A la vista del resultado, ¿crees que existen diferencias significativas entre los porcentajes de pacientes que acuden al dentista por razones preventivas en las dos poblaciones? Razona la respuesta.

Ejercicio 6.7.

En una muestra de 1350 personas que residen en un barrio periférico de una gran ciudad se ha realizado un estudio para conocer la prevalencia de cierta alergia. De las pruebas realizadas, 95 proporcionaron resultados positivos. Al mismo tiempo, se tomó una muestra de 2010 personas para el resto de la ciudad en la que se observaron 113 casos. ¿Proporcionan estos resultados evidencia suficiente (α=0.05) que indique que el porcentaje de individuos con dicha enfermedad en dicho barrio es diferente a dicho porcentaje en el resto de la ciudad? Platea el contraste de hipótesis adecuado y resuélvelo tanto por el método de las regiones de aceptación/rechazo como por el método del p-valor utilizando como nivel de significatividad α=0.05.

Ejercicio 6.8.

Se va a realizar un estudio sobre enfermedad cardiovascular (relacionada con contaminación atmosférica) en distintas zonas (Norte y Sur) de una gran comunidad autónoma. El norte está caracterizado por una gran cantidad de industria y por tanto tiene más contaminación, mientras que el sur por el contrario no tiene tanta industria y su contaminación es menor. Se toma una muestra de 1350 personas que residen en la zona norte, de las que 95 resultó tener alguna enfermedad cardiovascular. Al mismo tiempo, se tomó una muestra de 2010 personas de la zona sur, en la que se observaron 113 casos. ¿Proporcionan estos resultados evidencia suficiente (con α =0.05) que indique que el porcentaje de individuos con alguna de estas enfermedades en la zona norte es mayor a dicho porcentaje en la zona sur?

Diferencia de Varianzas

Ejercicio 6.9.

Las mediciones de cierto hueso del cuerpo humano de una muestra de hombres y mujeres adultos dieron los resultados que se detallan a continuación: de una muestra de 11 hombres estudiados se obtuvo una media de 13.21 cm y una desviación típica de 1.05 cm; de una muestra de 9 mujeres se obtuvo una media de 11.00 cm y una desviación típica de 1.01 cm. ¿Podemos concluir que la varianza de la longitud de este hueso es significativamente diferente en hombre y mujeres? (Supondremos que estas variables tienen un comportamiento *normal*) Plantea el contraste correspondiente (con $\alpha = 0{,}01$) y explica tus conclusiones.

Ejercicio 6.10.

Veinte pacientes que sufren de epilepsia se dividieron al azar en dos grupos iguales con el fin de estudiar posibles diferencias en cuanto al número de convulsiones entre dos tratamientos diferentes. El grupo A recibió un tratamiento que incluía dosis diarias de vitamina D. El grupo B recibió el mismo tratamiento con la excepción de que a este grupo se le dio un placebo en lugar de la vitamina D. Las medias del número de convulsiones observadas durante el periodo de tratamiento en los dos grupos fueron 15 (grupo A) y 24 (grupo B) y las desviaciones típicas 3 y 3.5, respectivamente. ¿Proporcionan estos datos evidencia suficiente que indique que la varianza del número de convulsiones es diferente entre los que toman o no Vitamina D? (Supondremos que estas variables tienen un comportamiento *normal*).Plantea el contraste correspondiente (con $\alpha = 0{,}05$) y explica tus conclusiones.

Ejercicio 6.11.

A dos grupos de niños se les hicieron pruebas de agudeza visual. El grupo 1 estaba formado por 11 niños que recibieron cuidados de la salud por parte de médicos privados. La calificación media para este grupo fue de 26 con una desviación estándar de 5. El segundo grupo, que incluía 16 niños que recibieron cuidados de la salud por parte del departamento de salud pública, tuvo una calificación promedio de 21 con un desviación estándar de 6. ¿Podemos concluir que la varianza de ambos grupos es significativamente diferente? (Supondremos que estas variables tienen un comportamiento *normal*).Plantea el contraste correspondiente (con $\alpha = 0{,}05$) y explica tus conclusiones.

Diferencia de Medias

Ejercicio 6.12.

Las mediciones de cierto hueso del cuerpo humano de una muestra de hombres y mujeres adultos dieron los resultados que se detallan a continuación: de una muestra de 11 hombres estudiados se obtuvo una media de 13.21 cm y una desviación típica de 1.05 cm; de una muestra de 9 mujeres se obtuvo una media de 11.00 cm y una desviación típica de 1.01 cm. Realiza la prueba adecuada, bajo un nivel de significación de α=0.01, para valorar si la longitud media del hueso en hombres es significativamente mayor que en mujeres. Razona tu respuesta.

Ejercicio 6.13.

Veinte pacientes que sufren de epilepsia se dividieron al azar en dos grupos iguales con el fin de estudiar posibles diferencias en cuanto al número de convulsiones entre dos tratamientos diferentes. El grupo A recibió un tratamiento que incluía dosis diarias de vitamina D. El grupo B recibió el mismo tratamiento con la excepción de que a este grupo se le dio un placebo en lugar de la vitamina D. Las medias del número de convulsiones observadas durante el periodo de tratamiento en los dos grupos fueron 15 (grupo A) y 24 (grupo B) y sus desviaciones típicas 3 (grupo A) y 3.5 (grupo B). ¿Proporcionan estos datos evidencia suficiente que indique que la vitamina D es efectiva para disminuir el número medio de convulsiones? (α=0.05) Razona la respuesta.

Repite el ejercicio suponiendo que se tiene la siguiente información poblacional sobre las desviaciones típicas: $\sigma_A = \sigma_B = 3$. Reflexiona sobre los cambios que esta información supone sobre la resolución del ejercicio.

Ejercicio 6.14.

Queremos contrastar el efecto de una nueva dieta que prometen revolucionaria, y para ello sometemos a esta dieta a 12 personas durante 3 días obteniendo los siguientes resultados sobre el peso antes y después de esta dieta:

Persona	Peso Antes	Peso Después
1	86.2	84.9
2	53.6	53.7
3	69.9	68.8
4	71.4	71.0
5	51.8	52.5
6	95.4	93.8
7	84.0	83.2
8	60.2	57.8
9	92.6	91.1
10	50.2	48.9
11	49.4	49.5
12	90.0	90.4

¿Reduce esta dieta el peso medio de forma significativa? Plantea el contraste adecuado utilizando $\alpha = 0{,}05$ y razona la respuesta. Calcula el p-valor del contraste.

Ejercicio 6.15.

Un proyecto trataba de valorar los resultados de las pruebas de agudeza visual, según el organismo que las practicara. Para ello repartió aleatoriamente los niños disponibles a dos grupos. El grupo 1 realizó las pruebas de agudeza visual en un centro de salud ocular privado, mientras que el grupo 2 realizó dichas pruebas en el departamento de salud pública. El grupo 1 estaba formado por 11 niños que obtuvieron una calificación media de 26 con una desviación típica de 5. El segundo grupo incluía 14 niños y obtuvo una calificación promedio de 21 con un desviación típica de 6. Plantea y resuelve el contraste de hipótesis adecuado para evaluar si el sistema privado valora significativamente diferente la agudeza visual de los niños respecto al sistema público. ($\alpha = 0{,}1$). Razona la respuesta.

Ejercicio 6.16.

Con los datos del ejercicio anterior, vamos a suponer ahora que el número de datos por grupo y los valores medios muestrales obtenidos en cada grupo son los mismos, pero que realmente conocemos la desviación típica poblacional de la variable *agudeza visual* en la población y que toma el valor de 5.5. Halla, en este nuevo escenario, el intervalo de confianza al 90 % por ciento para la diferencia entre las medias poblacionales.

Ejercicio 6.17.

La piel de los cadáveres puede utilizarse para proporcionar injertos temporales de piel en personas con quemaduras graves. La mejoría que experimentan los pacientes con este tipo de injertos está en relación directa con el tiempo de supervivencia del injerto, que finalmente será rechazado por el sistema inmunológico del paciente. Un equipo médico investiga la eficacia de tales injertos con respecto al sistema antígeno HL-A. A cada paciente se le practican dos injertos, uno con alta HL-A compatibilidad y otro con baja compatibilidad. El tiempo de supervivencia, en días, de los injertos se muestra en la tabla adjunta.

Persona	Compatibilidad Alta	Compatibilidad Baja
1	37	29
2	19	13
3	57	15
4	93	26
5	16	11
6	23	18
7	20	26
8	63	43
9	29	18
10	60	42
11	18	19

Plantea el contraste de hipótesis adecuado para estudiar si los injertos con alta compatibilidad dan mejores resultados que los de baja compatibilidad $\alpha = 0{,}05$ y explica tus conclusiones.

Ejercicio 6.18.

Se estudió la eficacia de un medicamento analgésico en 30 mujeres que sufrían calambres tras el parto. Se eligieron aleatoriamente 15 de estas mujeres y se les administró el medicamento y a las 15 restantes se les administró un placebo (sustancia inerte). Las cápsulas conteniendo el medicamento o el placebo se administraron antes de desayunar y al mediodía. La mejoría experimentada se midió en una escala entre 0 (ninguna mejoría en absoluto) y 56 (mejoría completa durante 8 horas). En las mujeres tratadas (Medicamento), se obtuvo una mejoría media de 31.96 puntos y con una desviación típica de 12.05, mientras que en las mujeres no tratadas (Placebo) se obtuvo una mejoría media de 25.32 puntos y una desviación típica de 13.78. ¿Proporcionan estos datos evidencias suficientes para concluir que el tratamiento es efectivo? Plantea el contraste que corresponda utilizando un nivel de significatividad $\alpha = 0{,}05$ y razona la respuesta.

Ejercicio 6.19.

Un proceso habitual en las industrias conserveras consiste en tratar las verduras con agua hirviendo antes de enlatarlas. El problema radica en la gran pérdida de vitaminas que sufren las verduras así tratadas. Se cree que un método consistente en un lavado previo de las verduras con vapor de agua puede evitar la pérdida de vitaminas. Para comparar ambos métodos, se analizaron 10 grupos de judías provenientes de granjas diferentes. La mitad de las judías de un grupo se trataron con agua hirviendo y la otra mitad con vapor de agua. Se midió el contenido vitamínico de cada mitad después del lavado, obteniéndose los resultados siguientes:

Grupo	Vapor	Agua
1	35	33
2	48	40
3	65	55
4	33	41
5	61	62
6	54	54
7	49	40
8	37	35
9	58	59
10	65	56

Plantea el contraste de hipótesis adecuado para estudiar si el método de lavado con vapor de agua es mejor que el de agua hirviendo utilizando $\alpha = 0{,}05$, calcula el p-valor del contraste y explica tus conclusiones.

Ejercicio 6.20.

Veinticuatro animales de laboratorio con deficiencia de vitamina D se dividieron en dos grupos iguales. El grupo 1 recibió un tratamiento consistente en una dieta que proporcionaba la vitamina D. El segundo grupo no fue tratado. Al término del periodo experimental, se hicieron las determinaciones del nivel de vitamina D, obteniéndose los siguientes resultados:

$$Grupo\ \ tratado: \quad \overline{x}_1 = 11{,}1\ mg/100ml \quad S_1 = 1{,}5$$

$$Grupo\ \ no\ \ tratado: \quad \overline{x}_2 = 7{,}8\ mg/100ml \quad S_1 = 2{,}0$$

Calcula un intervalo de confianza al 99 % para la diferencia de los niveles medios de vitamina D entre los animales tratados y no tratados e interpreta el resultado del mismo.

Ejercicio 6.21.

Un grupo de investigadores del cáncer de mama reunió los siguientes datos en cuanto al tamaño de dos tipos de tumores diferentes (A y B):

Tipo de tumor	Tamaño muestral	Media muestral	Desv.Típica muestral
A	21	3.85 cm	1.95 cm
B	16	2.80 cm	1.70 cm

1. Calcula un intervalo de confianza al 95 % para la diferencia de los tamaños medios poblacionales de ambos tipos de tumores e interpreta los resultados.
2. Plantea el contraste de hipótesis adecuado para determinar si hay diferencias significativas entre los tamaños medios de ambos tumores utilizando $\alpha = 0{,}05$.
3. Comprueba que de ambas formas obtienes las mismas conclusiones.

Ejercicio 6.22.

Un epidemiólogo desea comparar dos vacunas para la rabia. Las personas que previamente habían recibido dichas vacunas se dividieron en dos grupos. El grupo 1 recibió una dosis de refuerzo de la vacuna del tipo 1 y el grupo 2 recibió una dosis de refuerzo de la vacuna del tipo 2. Las respuestas de los anticuerpos se registraron dos semanas después. Las medias, desviaciones típicas y tamaño de las muestras para los dos grupos fueron los siguientes:

Grupo	Tamaño muestral	Media muestral	Desv.Típica muestral
1	10	4.5	1.3
2	9	2.5	1.1

¿Indican estos datos que existe diferencia en la efectividad de las dos vacunas utilizadas para dosis de refuerzo? (α=0.05). Calcula p-valor del contraste.

Ejercicios recopilatorios

Ejercicio 6.23.

En un nuevo Departamento de Salud están realizando un estudio para decidir a cuál de dos laboratorios encargan las vacunas anti-gripales para la siguiente campaña gripal. Para tomar la decisión han conseguido datos de la campaña gripal anterior de la aplicación de las vacunas de los dos laboratorios. Según los datos disponibles, de 154 personas que fueron vacunadas con la vacuna del primer laboratorio (Lab1) 12 padecieron finalmente la gripe, mientras que de 169 personas que fueron vacunadas con la vacuna del segundo laboratorio (Lab2) 18 padecieron la gripe. Calcula un intervalo de confianza al 98 % para la diferencia de porcentajes de afectados por gripe entre una y otra vacuna. Interpreta los resultados y explica con claridad tus conclusiones.

Ejercicio 6.24.

En un Departamento de Salud están realizando un estudio para decidir cuál de dos marcas de parches transdérmicos de morfina recomiendan a sus pacientes con enfermedades crónicas que los necesitan. Por un lado está la marca A, que han usado durante los últimos años y ha funcionado muy bien y por otro, la marca B, que acaba de salir al mercado con un precio algo más competitivo y cuyos fabricantes afirman que mejora el tiempo medio que el paciente pasa sin dolor respecto de la marca convencional (A). Para tomar una decisión, los investigadores responsables de la decisión han tomado 23 pacientes con enfermedades crónicas que requieren estos parches y han probado la marca A sobre 11 de ellos y la marca B sobre 10. A partir de estos pacientes han obtenido que el tiempo medio que los pacientes que han usado el parche marca A han permanecido sin dolor es de 54.5 horas con una desviación típica de 7.5 horas, mientras que para los que han usado el parche marca B se ha obtenido un tiempo medio sin dolor de 58.9 horas con una desviación típica de 6.9 horas.

Realiza las pruebas previas necesarias y el contraste de hipótesis adecuado para comprobar si los nuevos parches (marca B) proporcionan a los pacientes un tiempo medio sin dolor significativamente mayor que los parches convencionales (de marca A). Utiliza α= 0.05. Explica las conclusiones que de este análisis se derivan.

Ejercicio 6.25.

Se desea investigar si el porcentaje de diabéticos con altos niveles de triglicéridos es significativamente superior al porcentaje de no diabéticos que tienen altos niveles de triglicéridos. Los datos muestrales, de que se disponen para hacer el estudio, indican que de 180 diabéticos, 103 tenía altos niveles de triglicéridos, mientras que de 190 individuos no diabéticos 80 tenían altos niveles de triglicéridos. Plantea el contraste necesario para resolver la investigación e interpreta el resultado, en el contexto del ejercicio, en función del p-valor. (α= 0.1).

Ejercicio 6.26.

Estudios epidemiológicos han señalado que el consumo moderado de bebidas alcohólicas fermentadas tiene un efecto protector sobre la aparición y desarrollo de enfermedades cardiovasculares. Por ello, se desea investigar si el consumo moderado de cerveza, de forma habitual, aumenta significativamente la concentración sérica media de HDL (colesterol bueno). El estudio debes llevarlo a cabo según los datos siguientes:

Individuos que NO consumen ningún tipo de alcohol: n_1= 13, $\overline{x}_1$= 35.7 mg/dl, s_1= 2.3 mg/dl.
Individuos que consumen cerveza moderadamente: n_2= 10, $\overline{x}_2$= 42.5 mg/dl, s_2= 1.2 mg/dl.

Realiza las pruebas previas necesarias y el contraste de hipótesis adecuado para comprobar si son ciertas las sospechas de la investigación. Utiliza α= 0.05. Explica las conclusiones que de este análisis se derivan.

Capítulo 7

Análisis de la varianza

7.1. Introducción al análisis de la varianza (ANOVA)

En el capítulo 6 estudiamos la comparación de las medias de dos poblaciones. En este capítulo introducimos el análisis de la varianza, cuyo objetivo es comparar dos o más medias simultáneamente. Si por ejemplo queremos valorar el nivel medio de mercurio en sangre de los habitantes de tres ciudades diferentes (C1, C2 y C3) querríamos comparar 3 medias simultáneamente: μ_1, μ_2 y μ_3. Detrás de la comparación de estas tres medias podemos ver el estudio de la relación de dos variables, una cuantitativa (nivel de mercurio en sangre) y otra categórica (ciudad). A la variable categórica (ciudad) se le suele llamar *factor* y en este caso se trata de un factor con tres categorías (C1, C2 y C3). Para llevar a cabo esta comparación plantearíamos un *modelo de análisis de la varianza de una vía* (de un *factor*). Si tuviéramos más de un factor para estudiar, por ejemplo queremos comparar diferentes ciudades y diferenciar entre hombres y mujeres, tendríamos dos factores en estudio: ciudad (C1, C2 y C3) y sexo (H y M), y el modelo sería de Análisis de la varianza con más de un factor. En esta asignatura vamos a introducir únicamente la comparación cuando tenemos un factor.

Siguiendo con el ejemplo de comparar el nivel medio de mercurio en sangre de los habitantes de 3 ciudades, podríamos pensar en comparar los valores medios de los habitantes de estas ciudades dos a dos, es decir, podríamos comparar el de C1 con C2, C1 con C3 y C2 con C3. El problema es que si cada una de esas comparaciones la realizamos con un nivel de significatividad $\alpha = 0{,}05$, la comparación global no tendría este nivel de significatividad, sino uno mayor (en este caso alrededor de 0,14). Recordamos que α representa la probabilidad de rechazar la hipótesis nula siendo cierta, y en general solemos tomar 0,05 como esta probabilidad de equivocarnos si rechazamos H_0. En el caso de realizar las tres comparaciones que hemos comentado, estamos diciendo que la probabilidad de rechazar al menos uno de ellos siendo ser cierta H_0 aumentaría hasta un 0,14, por tanto, la probabilidad de equivocarnos globalmente si encontramos diferencias en alguno de ellos sería de un 14 %.

Para solventar este problema, algunos autores proponen algunas correcciones sobre cada una de las comparaciones. Por ejemplo, *Bonferroni* propone realizar cada una de las comparaciones tomando un nivel de α menor de forma que el nivel global con las tres comparaciones sea 0,05. Así, por ejemplo, si realizamos cada una de las tres comparaciones que hemos comentado utilizando un nivel de significatividad de $\alpha = 0{,}0167$, el nivel de significatividad global sería aproximadamente 0,05. El problema que tiene esta aproximación, es que para rechazar la hipótesis nula en cada una de las comparaciones anteriores tendríamos que observar diferencias muy muy grandes, puesto que el p-valor obtenido tendría que ser inferior a 0,0167, y por tanto, esta técnica es muy conservadora si lo que queremos es saber si para la variable que estamos midiendo (*nivel de mercurio en sangre*) existen diferencias significativas entre diferentes grupos (*ciudades*).

El análisis de la varianza, realiza las comparaciones simultáneamente y con el α global que deseemos, y es capaz de determinar si en general existen diferencias significativas entre los grupos que estamos comparando o no las hay.

7.2. Contraste de hipótesis

Supongamos que tenemos una variable cuantitativa Y cuyo valor medio queremos comparar en diferentes grupos (definidos por las k categorías de una variable categórica llamada *factor*). El contraste de hipótesis que nos planteamos es:

$$H_0 : \mu_1 = \mu_2 = ... = \mu_k$$

$$H_1 : \mu_i \neq \mu_j \text{ para algún } i, j$$

Si el p-valor obtenido de este contraste es menor que α, se rechazaría la hipótesis nula y se concluiría que *al menos dos de las medias difieren entre sí*. Posteriormente, si resulta de interés, habría que valorar cuál o cuáles son las medias entre las que hay diferencias.

7.2.1. Datos

Los datos disponibles habitualmente en este tipo de problemas son una muestra para cada uno de los k grupos:

1: $Y_{11}, Y_{12}, ..., Y_{1n_1}$

2: $Y_{21}, Y_{22}, ..., Y_{2n_2}$

...

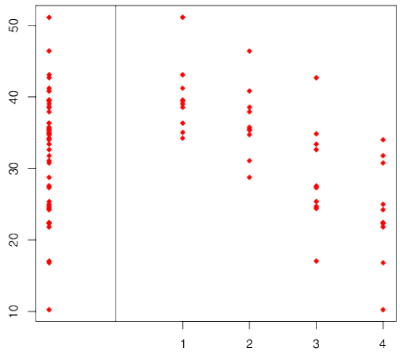
k: $Y_{k1}, Y_{k2}, ..., Y_{kn_k}$

Si el diseño de partida es *equilibrado* los tamaños de las muestras de cada grupo serán iguales, es decir $n_1 = n_2 = ... = n_k$, pero no necesariamente tendrá que serlo.

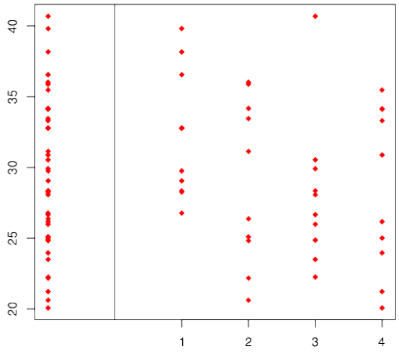
7.2.2. Idea intuitiva del funcionamiento del contraste

Aunque en esta asignatura no pretendemos calcular el estadístico de este contraste manualmente, sí queremos dar algunas ideas intuitivas sobre el funcionamiento del mismo y la justificación de por qué a la técnica que utilizamos para comparar *medias* de diferentes grupos se le llama *Análisis de la varianza*.

A continuación se muestra una figura en la que se pueden observar una muestra de tamaño 8 de cada una de las siguientes distribuciones: $N(45, 5)$, $N(40, 5)$, $N(30, 5)$,$N(20, 5)$.



Y en la siguiente figura podemos visualizar una muestra, también de tamaño 8, de cada una de cuatro distribuciones iguales: $N(35, 5)$ (o lo que es lo mismo, cuatro muestras de la misma población con media 35 y desviación típica 5).



En ambas figuras podemos observar en la parte izquierda (separada por una línea) las 32 observaciones juntas, y en la parte derecha separadas para cada uno de los grupos. En la figura superior, en la que las medias difieren de un grupo a otro, podemos observar que la varianza que tenemos de los datos dentro de cada grupo es menor que la varianza que tenemos en el conjunto de datos total (izda.). Mientras que en el caso del gráfico inferior, en el que las medias de todos los grupos coinciden, no hay mucha diferencia entre la variabilidad que tenemos dentro de cada grupo con la variabilidad que tenemos en el conjunto de datos total. Esta idea refleja, que comparando varianzas (dentro de cada grupo con la total) podemos detectar que el comportamiento es diferente cuando las medias coinciden y cuando no. En esta idea, entre otras (algunas por supuesto más teóricas), se basa la técnica ANOVA, que mediante la comparación de la variabilidad total, la que hay dentro de cada grupo y la que hay entre grupos permite obtener conclusiones sobre el comportamiento de las medias (es decir, sobre si existen diferencias significativas entre las medias o no).

7.2.3. Resolución del contraste de hipótesis

Supongamos que tenemos datos de un total de n individuos, repartidos en k grupos diferentes definidos por las k categorías de una variable categórica (es decir, $n = n_1 + n_2 + ... + n_k$ individuos) La resolución de este contraste de hipótesis está basada en la comparación de dos varianzas (llamadas *varianza entre grupos* y *varianza intra-grupos*), que, bajo la hipótesis nula de igualdad de medias deben ser iguales. Estas varianzas se comparan mediante su cociente en un estadístico al que llamamos F. Si la hipótesis nula es cierta se observarán valores del estadístico cercanos a 1, mientras que si la hipótesis nula no es cierta, se observarán valores grandes del estadístico F. Por tanto, estamos ante un contraste de hipótesis unilateral. Para resolver este contraste se utiliza el estadístico F, que bajo la hipótesis nula sigue una distribución F *de Snedecor* con $k-1$ y $n-k$ grados de libertad. A partir del valor calculado del estadístico F, y bien mediante la región de rechazo o bien mediante el p-valor a partir de la distribución $F_{(k-1,n-k)}$ tendríamos como solución el rechazar o no poder rechazar la hipótesis nula. Un resultado en el que rechazaríamos la hipótesis nula indicaría que tenemos evidencias suficientes para concluir que existen diferencias significativas entre las medias, es decir, al menos existiría un par de medias que diferirían entre sí. Mientras que un resultado en el que no pudiéramos rechazar la hipótesis nula indicaría que no tenemos evidencias para concluir que existan diferencias significativas entre la medias.

7.3. Hipótesis necesarias para la aplicación del ANOVA

7.3.1. Muestreo aleatorio

Todos los individuos que componen las observaciones de cada uno de los grupos deben haber sido elegidos de la población y asignados aleatoriamente a cada uno de ellos.

7.3.2. Normalidad

Los valores de la variable se distribuyen *normalmente* (o *siguen una distribución Normal*) en cada uno de los grupos definidos por el factor, es decir, para cada grupo i, $Y_{ij} \sim N(\mu_i, \sigma^2)$. La violación de este supuesto no afecta mucho a las conclusiones del análisis de la varianza si el tamaño de las muestras de cada grupo es relativamente grande (por ejemplo más de 30 datos por grupo). Si el número de datos por grupo no es excesivamente grande es recomendable realizar un contraste de hipótesis como el siguiente:

$$H_0 : \text{Los valores del grupo k siguen una distribución Normal}$$

$$H_1 : \text{Los valores del grupo k no siguen una distribución Normal}$$

y en caso de no rechazar la hipótesis nula podríamos decir que *no es un disparate* asumir que el comportamiento de los datos es *normal*. Existen diferentes pruebas para contrastar la normalidad, pero una muy extendida en uso es la llamada ***Prueba de Shapiro-Wilks***. Si el p-valor obtenido en esta prueba es superior al nivel de significatividad (normalmente $\alpha = 0{,}05$) no podemos rechazar la hipótesis de normalidad y por tanto podemos asumirla como cierta.

7.3.3. Homocedasticidad

La *Homocedasticidad*, o lo que es lo mismo, la *Homogeneidad de varianzas*, asume que las varianzas de todos los grupos a comparar son homogéneas (es decir, que no se detectan diferencias significativas entre las varianzas de los grupos a comparar). La violación de esta hipótesis impide asumir como correctos los resultados que de este análisis se deriven, y por tanto impiden la utilización de ANOVA. En los casos en los que no se pueda asumir esta premisa como cierta se recomienda utilizar otros métodos de comparación de medias, como los llamados *no paramétricos*, entre los que se encuentra la prueba de *Kruskall-Wallis* y que no forma parte del temario de esta asignatura. Para poder asumir como cierta esta hipótesis, es recomendable realizar el siguiente contraste de hipótesis:

$$H_0 : \sigma_1^2 = \sigma_2^2 = ... = \sigma_k^2$$

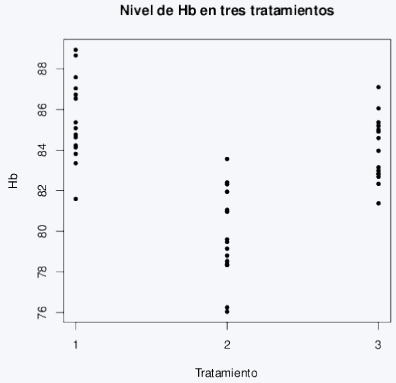
$$H_1 : \sigma_i^2 \neq \sigma_j^2 \text{ para algún } i, j$$

Existen diferentes pruebas para llevar a cabo este contraste, pero las más extendidas en uso son el ***Test de Barlett*** y el ***Test de Levene***. Si al realizar alguna de estas dos pruebas obtenemos un p-valor superior al nivel de significatividad (normalmente $\alpha = 0{,}05$) no podremos rechazar la hipótesis de homogeneidad de varianzas y por tanto podemos asumirla como cierta.

Ejemplo 7.1.

Supongamos que estamos interesados en comprobar si existen diferencias significativas en el nivel medio de hemoglobina (Hb) en tres tratamientos diferentes para personas con cierto tipo de anemia diagnosticada. Con el fin de realizar la comparación correspondiente se toman 45 pacientes con este tipo de anemia diagnosticada y se reparten al azar entre los tres tratamientos (15 en cada grupo). A continuación mostramos el análisis correspondiente para llevar a cabo esta comparación.

En primer lugar mostramos los datos obtenidos del nivel de Hb. en cada uno de los tres grupos:



Como primera apreciación tras la representación gráfica, sí que parece que el tratamiento 2 obtenga unos niveles de Hb inferiores, de media, a los de los otros dos tratamientos. El análisis adecuado para comprobar si existen diferencias significativas entre los niveles medios de Hb. en los tres tratamientos es un ANOVA,

$$H_0 : \mu_{T1} = \mu_{T2} = \mu_{T3}$$

$$H_1 : \mu_i \neq \mu_j \text{ para algún } i, j$$

pero en primer lugar debemos comprobar que se cumplen las hipótesis de aplicabilidad de esta técnica:

- Muestreo aleatorio: Esta hipótesis se cumple, pues los 45 pacientes disponibles son asignados aleatoriamente a cada uno de los tres tratamientos.
- Normalidad: Para comprobar esta hipótesis puedo realizar, por ejemplo, un test de Shapiro-Wilks en la que comprobaría si es asumible esta hipótesis, o por lo contrario, los datos la violan de forma clara. Con este fin, para cada uno de los tres grupos plantearía el siguiente contraste de hipótesis:

H_0 : Los datos que provienen del tratamiento T1 siguen una distribución Normal

H_1 : Los datos que provienen del tratamiento T1 no siguen una distribución Normal

H_0 : Los datos que provienen del tratamiento T2 siguen una distribución Normal

H_1 : Los datos que provienen del tratamiento T2 no siguen una distribución Normal

H_0 : Los datos que provienen del tratamiento T3 siguen una distribución Normal

H_1 : Los datos que provienen del tratamiento T3 no siguen una distribución Normal

Supongamos conocidos los p-valores obtenidos para este contraste en cada uno de los grupos fueran:

1. Tratamiento 1: Test de Normalidad Shapiro-Wilk (p-valor=0.7645)
2. Tratamiento 2: Test de Normalidad Shapiro-Wilk (p-valor=0.5438)
3. Tratamiento 3: Test de Normalidad Shapiro-Wilk (p-valor=0.8341)

Como en todos los casos, el p-valor para este contraste es mayor que 0.05, en ninguno de los casos puede ser rechazada la hipótesis de normalidad, y por tanto, podemos asumir que se cumplen (al menos estamos seguros que no se desvían demasiado de cumplirse)

- Homocedasticidad u Homogeneidad de varianzas: En este caso, planteamos el contraste de hipótesis:

$$H_0 : \sigma^2_{T1} = \sigma^2_{T2} = \sigma^2_{T3}$$
$$H_1 : \sigma^2_i \neq \sigma^2_j \text{ para algún } i, j$$

  Supongamos conocido el Test de Levene y el resultado obtenido es: Test de igualdad de varianzas de Levene (p-valor=0.2859). Puesto que el p-valor es mayor que 0.05, no podemos rechazar la hipótesis nula y por tanto no podemos rechazar que la igualdad de varianzas se cumpla. Así, podemos asumir esta hipótesis como cierta (como en el caso anterior, al menos estamos seguros de que en caso de no ser cierta no se desvía mucho de la misma).

Una vez comprobado que las hipótesis necesarias para poder aplicar el ANOVA se cumplen, ya estamos en disposición de aplicar esta técnica y obtener conclusiones sobre la igualdad o no de los niveles medios de la variable cuantitativa sobre los distintos grupos.
Planteamos el contraste ANOVA:

$$H_0 : \mu_{T1} = \mu_{T2} = \mu_{T3} (\text{ Nivel medio de Hb igual en los tres trat.})$$

$$H_1 : \mu_i \neq \mu_j \text{ para algún } i, j \text{ (Nivel medio de Hb no igual en al menos 2 trat.)}$$

A partir de los datos obtenemos el estadístico F que tiene un valor pivote=11.229. Bajo la hipótesis nula el estadístico F sigue una distribución F con $3 - 1 = 2$ y $45 - 3 = 42$ grados de libertad, es decir una $F_{(2,42)}$. La región de rechazo (para $\alpha = 0{,}05$) sería aproximadamente $(3{,}23, +\infty)$, y por tanto con el valor de nuestro pivote rechazaríamos la hipótesis nula.
Conclusión: Tenemos evidencias para concluir que existen diferencias significativas entre los niveles medios de Hb que proporcionan al menos dos de los tratamientos.

7.4. Comparaciones múltiples

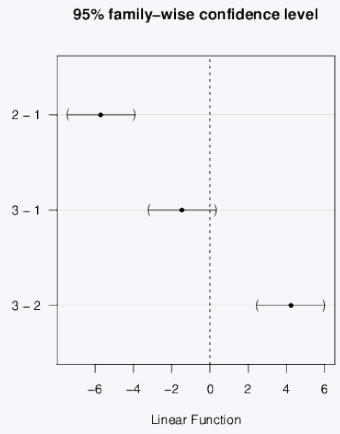
Una vez realizado el análisis de la varianza, si se detectan diferencias estadísticamente significativas entre las medias de los grupos comparados, en ocasiones tiene interés el determinar entre qué pares de medias existen esas diferencias y para ello se utilizan las llamadas pruebas de *comparaciones múltiples*. Existen diferentes pruebas para llevar a cabo esta comparación de todos los pares de medias (dos a dos), pero en todos los casos se tiene en cuenta que se van a realizar multitud de comparaciones y se consideran cada comparación de forma adecuada para que el nivel de significatividad global sea el deseado (por ejemplo $\alpha = 0{,}05$). Las pruebas más utilizadas son **Tukey**, y en menor medida **Sheffé** y **Dunnett**, y en todas ellas tenemos finalmente para cada comparación, o bien un p-valor o bien un intervalo de confianza para la diferencia de las medias de cada comparación.

Ejemplo 7.2.

Siguiendo con el ejemplo anterior, ahora queremos comprobar dónde se encuentran las diferencias detectadas por la técnica ANOVA. Sabemos que los tres tratamientos no obtienen el mismo nivel medio de Hb pero....¿entre qué tratamientos se encuentra las diferencias? Con ayuda de R-Commander obtenemos la comparación de todas las medias dos a dos bajo el criterio de Tukey. Los resultados se muestran a continuación:

Comparación	Diferencia media estimada	Intervalo 95 % Lím.Inf.	Intervalo 95 % Lím.Sup.	p-valor
T2-T1	-5.7075	-7.4544	-3.9606	$< 0{,}001$ ***
T3-T1	-1.4692	-3.2161	0.2777	0.114
T3-T2	4.2383	2.4914	5.9853	$< 0{,}001$ ***

Podemos observar que existen diferencias significativas entre los tratamientos 2 y 1, y también entre los tratamientos 3 y 2, mientras que no existen diferencias significativas entre los tratamientos 3 y 1. Este resultado lo podemos apreciar o bien a través de los intervalos de confianza para la diferencia de niveles medio de cada par de grupos, o bien a través del p-valor. Por un lado, en los resultados mostrados para la comparación de los tratamientos 2 y 1 y de los tratamientos 2 y 3 el intervalo estimado no contiene al 0, es decir, tiene o bien los dos extremos positivos, o bien los dos negativos y además el p-valor es inferior al nivel de significatividad (0.05). Por otro lado, en los resultados mostrados para la comparación de los tratamientos 1 y 3 observamos un intervalo para la diferencia de medias que contiene al 0 y un p-valor superior al nivel de significatividad (0.05).
Así, tendríamos dos grupos de tratamientos homogéneos entre sí, el formado por los tratamientos 1 y 3 que tienen niveles medios que difieren significativamente y el formado por el tratamiento 2, que tiene un nivel medio diferente a los que forman el otro grupo.
Estos resultados, también pueden ser observados a nivel gráfico en la siguiente representación (también proporcionada por el R-Commander):



7.5. Ejercicios Capítulo 7

Ejercicio 7.1.

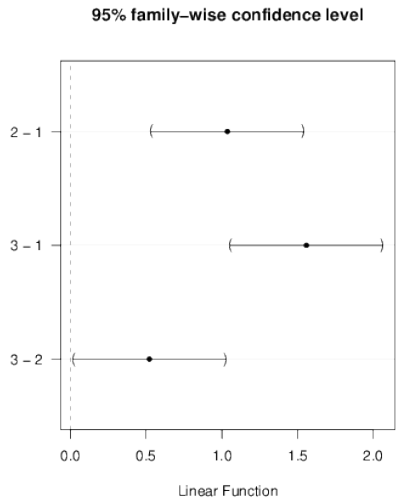
Un epidemiólogo desea comparar tres variantes de una vacuna para la meningitis. Se seleccionaron 75 personas que posteriormente fueron repartidas al azar en los tres grupos que posteriormente recibieron cada una de las variantes. El objetivo es investigar si las variantes de la vacuna, proporcionan diferentes números medios de anticuerpos. Las respuestas de los anticuerpos se registraron dos semanas después para cada persona. A continuación se muestran algunos resultados obtenidos junto con algunas preguntas que debes resolver ($\alpha = 0{,}05$):

- Se dispone de los siguientes resultados:
  - Test de Levene (p-valor=0.3758)
  - Variante 1: Test de Shapiro-Wilk (p-valor=0.4567)
  - Variante 2: Test de Shapiro-Wilk (p-valor=0.4538)
  - Variante 3: Test de Shapiro-Wilk (p-valor=0.0834)

  Plantea el contraste correspondiente a cada p-valor y explica las conclusiones que se deriva de cada uno de ellos. ¿Se cumplen todas las hipótesis de aplicabilidad del ANOVA? Justifica tu respuesta
- Tras aplicar el ANOVA los resultados obtenidos son: *pivote* = 35,3.
  - ¿En qué distribución nos fijaremos para calcular la región de rechazo?
  - ¿Cuál es el contraste de hipótesis asociado a este pivote y cuál es la conclusión del mismo?
  - A partir del ANOVA podemos saber qué tratamientos difieren entre sí y cuáles no?
- Tras el ANOVA se han obtenido los siguientes resultados y la siguiente representación gráfica. Interprétalos:

  Contrastes de Tukey

Comparación	Diferencia media estimada	Intervalo 95 % Lím.Inf.	Intervalo 95 % Lím.Sup.
2-1	1.03	0.53	1.54
3-1	1.56	1.05	2.06
3-2	0.42	0.02	1.02



Ejercicio 7.2.

Se desea estudiar si hay diferencias significativas en el peso medio de los niños de 8 años que realizan la comida principal (mediodía) en diferentes situaciones. Concretamente se han considerado tres grupos de niños: niños que comen con los padres(P), niños que comen con familiares (abuelos, etc,...)(F) y niños que comen en comedor escolar (C). Para ello se dispone de una muestra aleatoria de 100 niños de 8 años (35 del primer grupo, 32 del segundo y 33 del tercero), para los que se ha registrado el peso y el ámbito en el que realiza la comida a mediodía.

1. Sobre estos datos se ha realizado un primer análisis estadístico con un nivel de significación de $\alpha = 0{,}05$ y se han obtenido los siguientes resultados:
   - Test de Levene (p-valor=0.09578)
   - Grupo P: Test de Shapiro-Wilk (p-valor=0.0672)
   - Grupo F: Test de Shapiro-Wilk (p-valor=0.5581)
   - Grupo C: Test de Shapiro-Wilk (p-valor=0.0848)

   Explica para qué nos sirven estos resultados y plantea el contraste correspondiente a cada uno de los p-valores anteriores. Explica las conclusiones que se derivan de cada uno de ellos. ¿Qué conclusión global podemos obtener de estos resultados? Justifica tu respuesta.
2. Tras realizar la prueba estadística ANOVA sobre estos datos se ha obtenido un valor del estadístico *pivote* = 0,515, pero se desconoce el valor del p-valor. A partir de este valor debes plantear y resolver el contraste asociado a la técnica ANOVA y explicar las conclusiones que se derivarán de ella (Ayuda: puedes resolverlo mediante al cálculo de las regiones de aceptación-rechazo)

Ejercicio 7.3.

Un psicólogo clínico desea evaluar la eficacia de un fármaco para reducir la ansiedad. Para ello, selecciona al azar 15 pacientes de su consulta que sufren este problema y forma aleatoriamente tres grupos del mismo tamaño. A cada grupo le administra aleatoriamente una dosis de fármaco (10 mg, 20 mg y 30 mg). Al cabo de un tiempo les mide su nivel de ansiedad. Tras la experiencia, el psicólogo realiza un análisis estadístico bajo un nivel de significación de α= 0.05 y obtiene los siguientes resultados:

Test de Levene (p-valor= 0.3966)
Dosis 10mg : Test Shapiro - Wilk (p-valor= 0.3254)
Dosis 20mg : Test Shapiro - Wilk (p-valor= 0.1185)
Dosis 30mg : Test Shapiro - Wilk (p-valor= 0.3254)

a) Plantea el contraste correspondiente a cada uno de los p-valores anteriores y explica las conclusiones que se derivan de cada uno de ellos. ¿Se cumplen todas las hipótesis de aplicabilidad del ANOVA? Justifica tu respuesta.
b) Tras resolver la prueba ANOVA se obtiene que pivote = 67.5. Plantea el contraste de hipótesis al que contesta este pivote y explica la conclusión del mismo.
c) Los contrastes de Tukey dan lugar a la siguiente tabla. Interpreta los resultados.

Comparación	Diferencia media estimada	Intervalo 95 % Lím.Inf.	Intervalo 95 % Lím.Sup.
$\mu_{20} - \mu_{10}$	-3	-4.38	-1.62
$\mu_{30} - \mu_{10}$	-6	-7.38	-4.62
$\mu_{30} - \mu_{20}$	-3	-4.38	-1.62

Ejercicio 7.4.

Se desea valorar si existen diferencias significativas en el tiempo medio de recuperación de una intervención quirúrgica, para la extirpación de un tumor en la vejiga, según tres técnicas quirúrgicas: A (Laparoscopia), B (Cirugía abierta clásica), C (Cirugía abierta innovadora). Para llevar a cabo el

estudio se tomó una muestra de 58 pacientes con este tipo de tumor y se les aplicó, al azar, una de las tres técnicas. 25 pacientes fueron intervenidos por laparoscopia, 13 pacientes fueron intervenidos por cirugía abierta clásica y 20 por innovadora.$\alpha = 0{,}05$ Según los resultados que aparecen a continuación,

Test de Levene (p-valor= 0.6082)
Técnica A: Test Shapiro - Wilk (p-valor= 0.3444)
Técnica B: Test Shapiro - Wilk (p-valor= 0.5688)
Técnica C: Test Shapiro - Wilk (p-valor= 0.3060)
ANOVA: pivote=80.13

Comparación	Diferencia media estimada	Intervalo 95 % Lím.Inf.	Intervalo 95 % Lím.Sup.
B-A	8.43	6.06	10.80
C-A	10.28	8.20	12.36
C-B	1.85	-0.61	4.32

a) ¿Se cumplen los criterios de aplicabilidad del ANOVA? Justifica tu respuesta, escribiendo los contrastes necesarios para la discusión.

b) Escribe el contraste al que contesta la prueba ANOVA y razona a qué conclusión conduce su resultado.

c) A partir de la prueba ANOVA, ¿podemos saber qué técnicas quirúrgicas difieren entre si? Justifica la respuesta.

d) Comenta las conclusiones que se deducen de las comparaciones de Tukey.

Ejercicio 7.5.

Una empresa de gestión hospitalaria desea investigar si existen diferencias significativas en el tiempo medio de hospitalización (expresado en días), tras una intervención quirúrgica de las mismas características, en 3 hospitales de la ciudad (A, B y C). Tras el análisis estadístico adecuado, R-Commader otorgó los siguientes resultados.$\alpha = 0{,}05$

```
> AnovaModel.3 <- aov(TiempoHosp ~ Hospital, data=Datos)
> summary(AnovaModel.3)
             Df Sum Sq Mean Sq F value  Pr(>F)
Hospital      2  210.6  105.31   6.577   ***
Residuals    30  480.4   16.01
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figura 7.3: .

```
> leveneTest(Datos$TiempoHosp, Datos$Hospital, center=median)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  2  0.0128 0.4531
      30
```

Figura 7.1: .

```
> shapiro.test(Datos$TiempoHosp[Datos$Hospital=="A"])

        Shapiro-Wilk normality test

data:  Datos$TiempoHosp[Datos$Hospital == "A"]
W = 0.9066, p-value = 0.2224

> shapiro.test(Datos$TiempoHosp[Datos$Hospital=="B"])

        Shapiro-Wilk normality test

data:  Datos$TiempoHosp[Datos$Hospital == "B"]
W = 0.9634, p-value = 0.8131

> shapiro.test(Datos$TiempoHosp[Datos$Hospital=="C"])

        Shapiro-Wilk normality test

data:  Datos$TiempoHosp[Datos$Hospital == "C"]
W = 0.9312, p-value = 0.4227
```

Figura 7.2: .

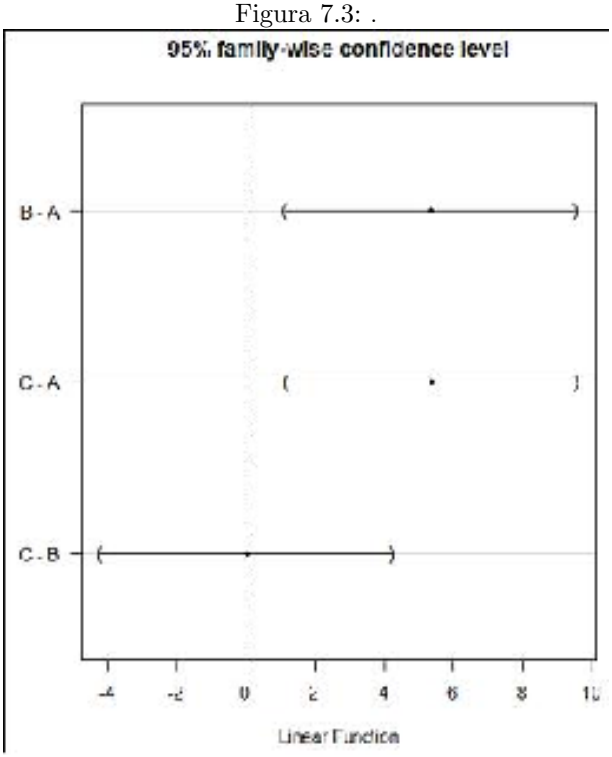


Figura 7.4: .

a) Plantea el contraste correspondiente a cada p-valor y explica las conclusiones que se deriva de cada uno de ellos. ¿Se cumplen todas las hipótesis de aplicabilidad del ANOVA? Justifica tu respuesta

b) Tras aplicar el ANOVA:

- ¿Cuál es el contraste de hipótesis asociado a esta prueba?
- ¿Con qué distribución se ha calculado el pivote?
- A partir del ANOVA, ¿podemos saber qué hospitales difieren entre sí y cuáles no?.
- ¿Cuál es la conclusión que podemos obtener?

c) Interpreta el gráfico.

Ejercicio 7.6.

Estudios epidemiológicos han señalado que el consumo moderado de bebidas alcohólicas fermentadas tiene un efecto protector sobre la aparición y desarrollo de enfermedades cardiovasculares. Por ello, se desea investigar si el consumo moderado habitual de cerveza o vino produce diferencias significativamente en la concentración sérica media de HDL (colesterol bueno) respecto a las personas que sólo consumen agua. Para llevar a cabo el estudio se han seleccionado aleatoriamente 43 pacientes. $\alpha = 0{,}01$

```
> leveneTest(MisDatos$HDL, MisDatos$Bebida, center=median)
Levene's Test for Homogeneity of Variance (center = median)
      Df F value  Pr(>F)
group  2  4.2525 0.02116 *
      40
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figura 7.5: .

```
> shapiro.test(MisDatos$HDL[MisDatos$Bebida=="cerveza"])
        Shapiro-Wilk normality test

data:  MisDatos$HDL[MisDatos$Bebida == "cerveza"]
W = 0.9076, p-value = 0.1118

> shapiro.test(MisDatos$HDL[MisDatos$Bebida=="vino"])
        Shapiro-Wilk normality test

data:  MisDatos$HDL[MisDatos$Bebida == "vino"]
W = 0.9783, p-value = 0.8977

> shapiro.test(MisDatos$HDL[MisDatos$Bebida=="agua"])
        Shapiro-Wilk normality test

data:  MisDatos$HDL[MisDatos$Bebida == "agua"]
W = 0.9716, p-value = 0.5598
```

Figura 7.6: .

```
> AncvaHdl.1 <- aov(HDL ~ Bebida, data=Datos)
> summary(AncvaHdl.1)
             Df Sum Sq Mean Sq F value Pr(>F)
Bebida        2 1881.0   917.1   149.8 <2e-16
Residuals    40  245.7     6.1
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figura 7.7: .

```
Multiple Comparisons of Means: Tukey Contrasts
Linear Hypotheses:
                        Estimate lwr      upr
cerveza - agua == 0 14.5583  12.2219 16.8947
vino - agua == 0    14.5685  12.2648 16.8723
vino - cerveza == 0  0.0102  -2.1579  2.1783
```

Figura 7.8: .

a) Plantea el contraste correspondiente a cada p-valor y explica las conclusiones que se deriva de cada uno de ellos. ¿Se cumplen todas las hipótesis de aplicabilidad del ANOVA? Justifica tu respuesta

b) Tras aplicar el ANOVA:

- ¿Cuál es el contraste de hipótesis asociado a esta prueba?
- ¿Con qué distribución se ha calculado el pivote?
- A partir del ANOVA, ¿podemos saber qué bebidas difieren entre sí y cuáles no?.
- ¿Cuál es la conclusión que podemos obtener?

c) Interpreta el resultado de las comparaciones múltiples.

Ejercicio 7.7.

Un proyecto de investigación pretende comparar la resistencia media de tres tipos de vendas: las vendas tipo I, las de tipo II y las de tipo III. Con este fin, se dispuso en un laboratorio un experimento que consistía en tirar de un trozo de venda desde ambos lados y medir la fuerza horizontal necesaria para romperla. Se tomaron 60 piezas de venda, de las cuales 20 fueron de tipo I, otras 20 de tipo II y otras 20 de tipo III. El encargado del análisis estadístico les proporcionó, a modo de

resumen, la información que a continuación se detalla, pero no llegó a explicarles el significado de estos resultados, y para eso te necesitan a ti. Utiliza y ordena los resultados proporcionados para explicar detalladamente el procedimiento seguido para realizar el análisis. En todos los casos indica cuál es el contraste de hipótesis asociado, para qué sirve y cuáles son sus conclusiones. Si falta algún cálculo para poder dar una conclusión, realízalo y concluye. Explica con claridad tus conclusiones. Utiliza = 0.01.

Comparación	Diferencia media estimada	Intervalo 99 % Lím.Inf.	Intervalo 99 % Lím.Sup.
II-I	0.35	-0.08	0.92
III-I	1.50	1.17	1.97
III-II	0.90	0.37	1.27

Test de Levene (p-valor= 0.1876)
ANOVA: pivote=44.5
Vendas tipo I: Test Shapiro - Wilk (p-valor= 0.085)
Vendas tipo II: Test Shapiro - Wilk (p-valor= 0.024)
Vendas tipo III: Test Shapiro - Wilk (p-valor= 0.342)

Ejercicio 7.8.

Se desea investigar la eficacia de tres tratamientos diferentes (A, B y C) sobre la rinitis alérgica. Para ello, se seleccionan 43 pacientes que padecen la enfermedad y se les asignan aleatoriamente uno de los tres tratamiento. Tras un mes de medicación se recoge el tiempo que transcurre hasta la siguiente crisis alérgica (tiempo de eficacia). A continuación se analizan los datos estadísticamente mediante la técnica ANOVA ($\alpha = 0{,}01$) y se obtienen los siguientes resultados:

Resultado 1: Test de Leven (p-valor= 0.9123).

Resultado 2:

Test Shapiro Wilk para el tratamiento A (p-valor= 0.1184)

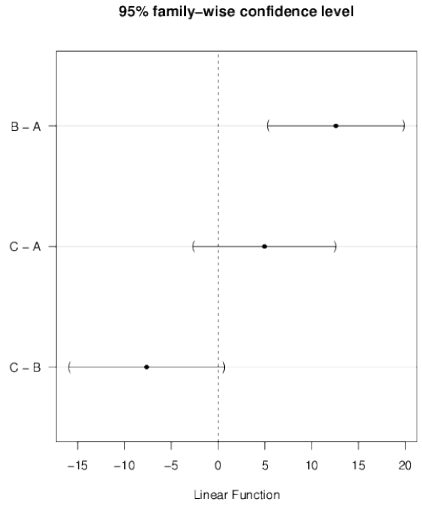
Test Shapiro Wilk para el tratamiento B (p-valor= 0.5452)

Test Shapiro Wilk para el tratamiento C (p-valor= 0.4468)

Resultado 3:

pivote ANOVA= 9.001

Resultado 4:



A partir de estos resultados, contesta las siguientes preguntas:

a) Plantea el contraste correspondiente al p-valor del Resultado 1 y explica las conclusiones que se derivan de él.

b) Explica la conclusión que se deriva de los p-valores del Resultado 2:

c) ¿Se cumplen las hipótesis de aplicabilidad del ANOVA? Justifica tu respuesta.

d) Plantea el contraste correspondiente al pivote del Resultado 3. Determina qué distribución sigue este pivote. Calcula la región de aceptación y rechazo del contraste y explica la conclusión en el contexto del ejercicio.

e) A partir del ANOVA, ¿podemos saber qué tratamientos difieren entre sí y cuáles no?

f) A partir del Resultado 4, determina la relación que hay entre μ_A, μ_B y μ_C ajustando el símbolo adecuado entre las medias ($<$, $>$ o $=$) y explica la conclusión correspondiente en el contexto del ejercicio.

Capítulo 8

Test Chi-cuadrado

En capítulos anteriores estudiamos el test t de comparación de dos medias en muestras independientes y extendimos esta técnica a la comparación de dos o más medias mediante la técnica ANOVA. En este tipo de problemas estudiamos la relación entre una variable categórica (que define dos o más grupos) y una variable cuantitativa, cuyo valor medio queremos comparar en los diferentes grupos definidos por la variable categórica.

También estudiamos la comparación de dos porcentajes. Está técnica nos permite comparar los porcentajes de una categoría concreta de una variable categórica en los diferentes grupos definidos por otra variable categórica. En este caso, estamos estudiando la relación de dos variables categóricas, una con dos categorías y la otra con una categoría de interés. En este tema, extenderemos esta comparación de dos porcentajes a la comparación simultánea de varios porcentajes, definida por el cruce de dos variables categóricas con cualquier número de categórías cada una de ellas.

Por tanto, nos planteamos la **relación o influencia entre dos variables cualitativas o categóricas** (cada una con dos o más categorías). El test χ^2 nos proporciona una prueba para valorar si existe relación (influencia) entre ellas. Diremos que existe relación entre 2 variables categóricas o que dichas variables son dependientes si las proporciones de respuesta de cada categoría que se dan en una de las variables dependen de la categoría de la otra variable.

Ejemplo 8.1.

Relación de la exposición al tabaco con la presencia de migrañas en personas menores de 25 años

Un estudio se está planteando la posible relación entre la exposición al tabaco y la aparición de migrañas en personas menores de 25 años. Con este fin, se han medido dos variables sobre una muestra de jóvenes menores de 25 años: su exposición al tabaco (medida como *Fumador*, *Fumador pasivo* y *No Fumador*) y si padece habitualmente migrañas (medida como *No* y *Sí*).
En este caso, si por ejemplo la proporción de jóvenes con migraña fuera distinta dependiendo de si se tratara de jóvenes fumadores, fumadores pasivos o no fumadores, es decir, si la propoción de jóvenes con migraña dependiera de la categoría de exposición al tabaco que estuviéramos valorando, hablaríamos de una posible relación entre ambas variables.
Sin embargo, si, independientemente de que nos centráramos en jóvenes fumadores, fumadores pasivos o no fumadores la proporción de jóvenes con migraña fuera la misma (o muy similar) diríamos que no parece existir relación entre ambas variables.

8.1. Tabla de contingencia: distribuciones marginales y conjunta

Hasta ahora hemos resumido las variables categóricas mediante la proporción de veces que se ha dado cada una de sus posibles respuestas (frecuencias relativas) independientemente de los valores que toman otras variables. A esta distribución de la respuesta, que ignora el valor de otras variables, le llamamos **distribución marginal** de la variable.

Para valorar si dos variables son dependientes o independientes, habremos de atender a su *tabla de contingencia*. En dicha tabla cada fila y cada columna representan las categorías de cada una de las dos variables que estamos resumiendo, y en cada casilla de la tabla de contingencia disponemos del número de veces que hemos observado la correspondiente combinación de ambas variables en nuestra muestra. Es decir, en la tabla de contingencia se muestran las frecuencias absolutas (o relativas) de todas las combinaciones de las categorías de ambas variables dos a dos. A los valores de las casillas de la tabla de contingencia, los cuales resumen el comportamiento relativo de las dos variables conjuntamente le llamamos **distribución conjunta** de las variables.

Ejemplo 8.2.

Supongamos que el estudio sobre la relación de migrañas y nivel de exposición al tabaco hemos recogido información sobre 300 jóvenes menores de 25 años. A continuación mostramos cómo quedaría la tabla de contingencia

	Fumadores	Fumadores Pasivos	No Fumadores	*Total*	
No migrañas	39	43	188	***270***	***(90 %)***
Sí migrañas	11	7	12	***30***	***(10 %)***
Total	***50***	***50***	***200***	***300***	
	(16.67 %)	***(16.67 %)***	***(66.66 %)***		

Podemos observar en el interior de la tabla la distribución conjunta de ambas variables. Observamos, en frecuencias absolutas, el comportamiento conjunto de ambas variables.
En los totales, tanto por fila como por columna, se muestran en letra negrita las distribuciones marginales (como frecuencias absolutas y porcentajes).

A la vista de una tabla de contingencia podemos analizar la posible relación de dependencia o, por el contrario, la independencia entre las dos variables cualitativas en estudio.

Si las variables fueran independientes, las distribuciones marginales que observamos en la tabla se deben reproducir de forma aproximada también en cada fila y/o en cada columna. Si por ejemplo consideramos las distribución marginal de la variable que defina las filas de la tabla de contingencia, las proporciones que definen esta distribución marginal debemos encontrarla en cada una de las columnas de la tabla. De la misma forma, las proporciones de la distribución marginal de la variable que defina las columnas de la tabla de contingencia esperamos encontrarlas en cada una de las filas.

Ejemplo 8.3.

Análisis de las distribuciones marginales y conjunta de la tabla del ejemplo anterior.

En la tabla de contingencia mostrada en el ejemplo anterior, observamos que la distribución marginal de la variable Migrañas (sí/no) queda definida con un 10 % de jóvenes que padecen migraña, frente a un 90 % de jóvenes que no las padecen. Si las dos variables fueran independientes, esperaríamos que esta misma proporción se repitiera (aproximadamente) en cada una de las columnas de la tabla, es decir, en cada uno de los grupos que define la variable exposición al tabaco. Así, esperaríamos que en los jóvenes fumadores el 10 % padeciera migrañas y el 90 % no, en los jóvenes fumadores pasivos los mismos porcentajes, y exactamente los mismos para los jóvenes no fumadores (10 % padeciendo migrañas frente a un 90 % que no las padezcan). Si ésto fuera así, diríamos que la proporción de jóvenes con migraña es la misma en cualquier grupo de exposición al tabaco, por lo que las variables son independendientes y no está relacionado el hecho de tener mayor o menor exposición al tabaco con la aparición de migrañas. En la medida en la que estos porcentajes se desvíen de este comportamiento, aumentarán los indicios de que existe una posible relación de dependencia entre las variables.

Si queremos realizar este análisis reflexivo desde las perspectiva de la distribución marginal de la variable Exposición al tabaco, observamos que en total el 16.67 % de los jóvenes son fumadores, el 16.67 % fumadores pasivos y el 66.66 % restante son no fumadores. Esta misma distribución de porcentajes esperaríamos observar tanto en los jóvenes que padecen migrañas como en los que no las padecen. En la medida en la que los porcentajes de una fila y la otra se distancien crecerán nuestras evidencias en contra de la independencia de estas variables.

Vamos a observar exactamente cómo se comportan los porcentajes por columna, por ejemplo, y compararlos con el comportamiento total (distribución marginal). Como hemos comentado, observamos que en total el 10 % de los jóvenes padecen migrañas frente al 90 % que nos las padecen. Si exploramos estos porcentajes para cada uno de los grupos que define la variable Exposición al tabaco obtenemos:

	Fumadores	Fumadores Pasivos	No Fumadores	*Total*
No migrañas	78 %	86 %	94 %	(***90 %***)
Sí migrañas	22 %	14 %	6 %	(***10 %***)

Esperábamos el 90 % y 10 % en cada columna, y observamos que en este caso ésto no es así. En los fumadores se observa mayor porcentaje de jóvenes con migrañas (22 %), en fumadores pasivos algo menos (14 %), pero todavía por encima del 10 % esperado. Por último, en no fumadores se observa un porcentaje de jóvenes con migrañas mucho menor que en los otros dos grupos (y también menor que el 10 % esperable si las variables fueran independientes). En este caso, observamos que este comportamiento no es acorde con la *independencia* de las dos variables, ya que la presencia de migrañas depende del grupo de exposición al tabaco que se considere. Por tanto, estos datos apuntan a una posible relación entre ambas variables (las variables son dependientes)

8.2. Valores Observados y Valores Esperados

Los valores que se encuentran en cada casilla de la tabla de contingencia se llaman *Valores Observados*.

El análisis detallado de una tabla de contingencia puede apuntar a una posible relación de dependencia entre las variables o bien hacia la independencia de las mismas. Como hemos comentado, si el comportamiento de las proporciones en cada fila, o en cada columna, es igual (o similar) al que muestra la distribución marginal de la variable en cuestión, estará apuntando a la independencia de las variables. Por tanto, sabemos qué comportamiento cabe esperar si las variables fueran independientes: en cada casilla de la tabla esperamos un valor que se corresponda con el porcentaje que le otorga la distribución marginal.

Supongamos que disponemos de dos variables categóricas y la tabla de contingencia para ambas variables tiene la forma:

		Variable 2				*Total*
		Cat.1	Cat.2	...	Cat.n_c	*Casos*
Variable 1	Cat.1	O_{11}	O_{12}	...	O_{1m}	TF_1
	Cat.2	O_{21}	O_{22}	...	O_{2m}	TF_2
	...	...	...	...	...	...
	Cat.n_f	O_{n_f1}	O_{n_f2}	...	$O_{n_fn_c}$	TF_{n_f}
Total	Casos	TC_1	TC_2	...	TC_{n_c}	**N**

Notar que $\mathbf{N} = \sum_{i=1}^{n_f} TF_i = \sum_{j=1}^{n_c} TC_j$.

Con esta notación, el valor esperado para la categoría i de la primera variable y la categoría j de la segunda puede ser calculado como:

$$E_{ij} = \frac{TF_i \cdot TC_j}{N}$$

Ejemplo 8.4.

Valores Esperados del ejemplo anterior.

Si enfocamos este estudio según la distribución marginal de la variable migrañas (no/sí), como hemos analizado en el ejemplo anterior esperaríamos un 10 % de jóvenes con migrañas frente a un 90 % de jóvenes que no las padecieran, independientemente del grupo de exposición al tabaco que analizarámos. Por tanto, los 300 jóvenes que componen nuestra muestra, de los que 50 eran fumadores, 50 fumadores pasivos y 200 no fumadores, esperaríamos que se distribuyeran en la tabla de contingencia de la siguiente forma:

	Fumadores	Fumadores Pasivos	No Fumadores	*Total*	
No migrañas	45	45	180	***270***	(***90 %***)
Sí migrañas	5	5	20	***30***	(***10 %***)
Total	***50*** ***(16.67 %)***	***50*** ***(16.67 %)***	***200*** ***(66.66 %)***	***300***	

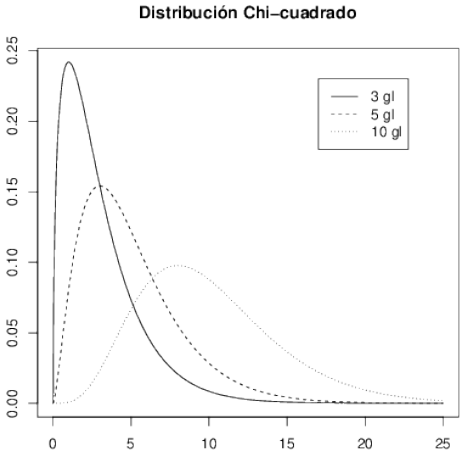
De esta forma se obtendrían las mismas distribuciones marginales (para las dos variables) y se obtendrían los mismos porcentajes por filas y por columnas.

A esta tabla se le llama la tabla de *Valores Esperados*, atendiendo que es el comportamiento que esperaríamos si las variables fueran independientes (es decir, bajo la hipótesis de independencia).

En la medida en que los Valores Observados se asemejen de los Valores Esperados se estaría apuntando a una independencia de las variables. Y por el contrario, en la medida en que los Valores Observados se alejen de los Valores Esperados se estaría apuntando a una dependencia de las variables. Pero, ¿cómo podemos valorar ante cualquier ejemplo real hasta qué punto los valores Observados y Esperados son razonablemente parecidos o por el contrario difieren los suficiente como para afirmar que hay relación de dependencia entre las variables? Necesitamos una herramienta estadística que nos permita valorar estas diferencias y nos ayude a transformarlas en un valor asociado a una determinada probabilidad.

8.3. Distribución Chi-cuadrado

La distribución χ^2 es una distribución asimétrica y con una única cola ya que únicamente toma valores superiores a 0. Esta distribución puede tomar las siguientes formas:



En el gráfico anterior observamos varias distribuciones, y es que la distribución χ^2 al igual que la distribución t tiene como parámetro los *grados de libertad*. Así, en la representación anterior observamos esta distribución con 3, 5 y 10 grados de libertad respectivamente. Observamos que cuanto mayor es el número de grados de libertad la distribución 2 admite valores mayores, es decir una variable χ^2 con un número de grados de libertad bajo tomará valores bajos mientras que una variable con un número alto de grados de libertad en su distribución tomará valores más altos con mayor probabilidad.

8.4. Test de independencia de dos variables categóricas χ^2

El test χ^2 se plantea el siguiente contraste de hipótesis a partir de dos variables categóricas:

H_0 : *Las variables son independientes* (No existe relación entre ellas)
H_1 : *Las variables no son independientes* (Existe relación entre ellas)

Todo contraste de hipótesis lleva asociado, para su resolución, un pivote, cuya distribución es conocida bajo la hipótesis nula. En este caso necesitamos definir la siguiente notación para definir el pivote:

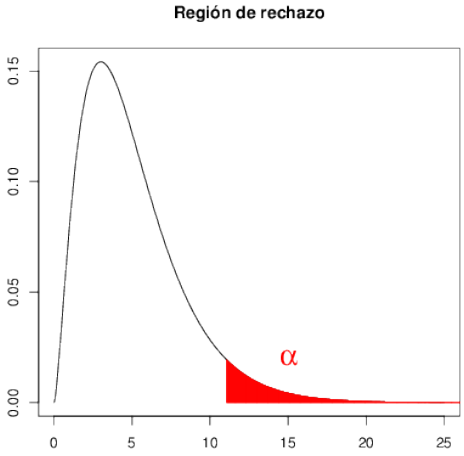
- n_f denotará el número de filas de la tabla de contingencia (el número de categorías de la variable que se sitúa por filas)
- n_c denotará el número de columnas de la tabla de contingencia (el número de categorías de la variable que se sitúa por columnas)
- i denotará cada una de las filas de la tabla de contingencia
- j denotará cada una de las columnas de la tabla de contingencia
- O_{ij} denotará el Valor Observado en la casilla (i, j) de la tabla de contingencia, es decir, en la casilla que se corresponde con la fila i y la columna j.
- E_{ij} denotará el Valor Esperado en la casilla (i, j) de la tabla de contingencia bajo la hipótesis nula, es decir, el valor que esperaríamos en esa casilla si las variables fueran independientes.

Con esta notación, el pivote se construye de la siguiente forma:

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad \sim \quad \chi^2_g$$

donde g representa los grados de libertad de la distribución χ^2 que viene dado por $g = (n_f - 1) \cdot (n_c - 1)$.

Cuando las variables son independientes, los valores observados en los datos se asemejan a los valores esperados (que son los que esperaríamos si las variables fueran independientes) y, en ese caso, el pivote toma valores pequeños. Cuando las variables no son independientes y por tanto existe entre ellas algún tipo de relación, los valores observados en los datos se alejan de los valores esperados y, en ese caso, el pivote toma valores más grandes. Por tanto, valores bajos del pivote apoyan la hipótesis de independencia (H_0) y valores altos de este pivote apoyan la hipótesis de dependencia (H_1). Por este motivo, se trata de un contraste de hipótesis unilateral, en el que la región de rechazo está formado por el $\alpha \times 100\,\%$ de valores mayores en la distribución χ^2 que corresponda.



Ejemplo 8.5.

Continuando con el ejemplo que hemos empleado durante todo el capítulo, vamos a contrastar si existe o no una relación significativa entre las variables presencia de migrañas y nivel de exposición al tabaco en jóvenes menores de 25 años. Utilizaremos como nivel de significatividad $\alpha = 0{,}05$

Tenemos, como datos, nuestra tabla de *Valores Observados*:

	Fumadores	Fumadores Pasivos	No Fumadores	*Total*
No migrañas	39	43	188	***270***
Sí migrañas	11	7	12	***30***
Total	***50***	***50***	***200***	***300***

A partir de ella calculamos la tabla de *Valores Esperados*:

	Fumadores	Fumadores Pasivos	No Fumadores	*Total*
No migrañas	45	45	180	***270***
Sí migrañas	5	5	20	***30***
Total	***50***	***50***	***200***	***300***

Ayuda: Para simplificar el cálculo de esta tabla puede usarse como regla para calcular el valor de la casilla (i, j) el cálculo del total de la fila i multiplicado por el total de la columna j y dividido por el total de individuos en la tabla. Por ejemplo, la primera casilla tiene un valor de 45 que puede verse como $\frac{50 \cdot 270}{300}$.

Planteamos el contraste de hipótesis:

H_0 : *Las variables son independientes* (No existe relación entre ellas)
H_1 : *Las variables no son independientes* (Existe relación entre ellas)

Calculamos el pivote:

$$\chi^2 = \frac{(39-45)^2}{45} + \frac{(43-45)^2}{45} + \frac{(188-180)^2}{180} + \frac{(11-5)^2}{5} + \frac{(7-5)^2}{5} + \frac{(12-20)^2}{20} = 12{,}44$$

$$\chi^2 = 12{,}44 \sim \chi^2_2$$

donde los grados de libertad se han calculado teniendo en cuenta que la tabla tiene 2 filas y 3 columnas: $g = (2-1) \cdot (3-1) = 2$

- Región de Rechazo: El percentil 0.95 de la distribución Chi-cuadrado con 2 grados de libertad es 5,99, por lo que la región de rechazo es $(5{,}99, +\infty)$. Dado que el pivote se encuentra en la región de rechazo rechazamos la hipótesis nula.
- P-valor: Calculamos la probabilidad (aproximada) de obtener en la distribución χ^2_2 un valor superior al pivote:
$$p - valor = P(\chi^2_2 > 12{,}44) \approx (1 - 0{,}995) = 0{,}005.$$
Puesto que el p-valor es menor que el nivel de significatividad (α), rechazamos la hipótesis nula.

Por tanto, podemos concluir que existe, tal y como intuíamos al analizar los valores de la tabla de contigencia, una relación significativa entre las variables migraña y nivel de exposición al tabaco. La mayor o menor presencia de migrañas en los jóvenes menores de 25 años depende de su grado de exposición al tabaco.

8.5. Ejercicios Capítulo 8

Ejercicio 8.1.

En una empresa que utilizaba para la fabricación de pinturas cierto producto químico se detectó que algunos empleados comenzaron a tener ciertos problemas de salud relacionados con alteraciones respiratorias. Se estaba contemplando la posibilidad de que el producto químico pudiera tener algo que ver con los problemas respiratorios. Para valorar esta hipótesis se seleccionó al azar a 500 empleados de la empresa, los cuales fueron clasificados en base a su nivel de exposición al producto y si tenían o no los síntomas de tales alteraciones respiratorias. Los resultados se presentan en la siguiente tabla:

	Contacto directo	Contacto limitado	No contacto	*Total*
Sí alteraciones resp.	185	33	17	***235***
No alteraciones resp.	120	73	72	***265***
Total	***305***	***106***	***89***	***500***

1. Explica cuál es tu impresión sobre la hipótesis de trabajo únicamente analizando la tabla de datos.
2. ¿Tenemos evidencias que indiquen, a nivel de significación 0.05, la existencia de relación entre el nivel de exposición y la presencia de síntomas de alteraciones respiratorias entre los empleados? Plantea y resuelve el contraste de hipótesis adecuado tanto por el método de las regiones de aceptación/rechazo, como por el método del p-valor. Explica las conclusiones obtenidas.

Ejercicio 8.2.

Un estudio realizado por logopedas tenía como objetivo valorar la relación del grupo socioeconómico de las familias de los niños y la presencia o ausencia de cierto defecto en la pronunciación. Para valorar esta relación seleccionó aleatoriamente a 500 niños de escuela primaria, los cuales fueron clasificados con el grupo socioeconómico de sus familias (como *Alto, Medio-Alto, Medio-Bajo,Bajo* y la presencia o ausencia del defecto en la pronunciación. Los resultados fueron los siguientes:

	Alto	Medio-Alto	Medio-Bajo	Bajo	*Total*
Defecto pronun. presente	8	24	32	27	***91***
Defecto pronun. ausente	42	121	138	108	***409***
Total	***50***	***145***	***170***	***135***	***500***

1. Explica cuál es tu impresión sobre la hipótesis de trabajo únicamente analizando la tabla de datos.
2. ¿Son compatibles estos datos con la hipótesis de que el defecto en la pronunciación no está relacionado con el estado socioeconómico ($\alpha = 0{,}05$)?. Plantea y resuelve el contraste de hipótesis adecuado para responder a esta pregunta y calcula el p-valor del mismo.

Ejercicio 8.3.

Se está llevando a cabo un estudio para comparar dos fármacos distintos, a los que llamaremos F1 y F2, donde ambos son tratamientos para dolor agudo provocado por migraña. A cada paciente se le clasifica, tras una hora de la aplicación del tratamiento, como: *Elimina dolor*, *Reduce intensidad*, y *No nota nada*. Se administra el tratamiento F1 a 32 pacientes y el F2 a 28 pacientes. De los 12 casos en los que se *Elimina dolor* 7 corresponden al fármaco F1 mientras el resto corresponden al fármaco F2; de los 30 casos en los que *Reduce intensidad*, 17 corresponden al fármaco F1 y el resto al fármaco F2; y de los 18 casos en los que *No nota nada*, 8 corresponden al fármaco F1 y el resto al fármaco F2. ¿Son igualmente efectivos ambos fármacos para el tratamiento de las migrañas? Plantea y resuelve el contraste de hipótesis adecuado para responder a esta pregunta y explica tus conclusiones. Utiliza como nivel de significatividad $\alpha = 0{,}05$.

Ejercicio 8.4.

El hábito de fumar es fuertemente nocivo para la salud, ya que no sólo daña el corazón, los pulmones y cada una de las células de cuerpo, sino que puede destruir tus músculos progresivamente, y que sus efectos en el organismo, culminan deteriorando la masa muscular. Ante la sospecha de que el hábito de fumar pueda influir en la masa muscular, se tomaron dos muestras, una de fumadores y otra de no fumadores, y se midió la síntesis de proteínas musculares. Para cada uno de ellos se midió el nivel de proteína (NP) total en sangre en g/dL en relación con los percentiles de la población (menor que el percentil 10, entre el 10 y el 90 y mayor que el percentil 90). El resultado se expresa en la tabla siguiente:

	$NP < P_{10}$	$P_{10} < NP < P_{90}$	$NP > P_{90}$
Persona fumadora	117	529	19
Persona no fumadora	124	1147	117

¿Existen evidencias significativas a favor de la sospecha a la vista de los resultados de la muestra? Plantea y resuelve el contraste de hipótesis adecuado (considerando $\alpha = 0{,}01$). Calcula el p-valor del contraste y explica tus conclusiones.

Ejercicio 8.5.

Un estudio odontológico pretendía evaluar la relación existente entre el motivo de la primera visita al dentista en niños de entre 3 y 6 años y su comportamiento en cuanto al *miedo* que presentaban los niños atendidos. Para ello, se tomaron niños con edades comprendidas entre los 3 y los 6 años que acudían por primera vez al dentista. Para cada uno de ellos se evaluó el motivo de su primera visita como *Prevención*, *Accidente* o *Curación*. El grado de *Miedo* para cada uno de ellos se clasificó como *Alto*, *Medio* y *Bajo*. Los datos obtenidos se muestran a continuación:

	N.Miedo Alto	N.Miedo Medio	N.Miedo Bajo	Total
Prevención	2	3	45	50
Accidente	3	7	5	15
Curación	8	12	40	60
Totales	13	22	90	125

Plantea y resuelve el contraste de hipótesis adecuado para valorar si existe o no una relación significativa entre el Motivo de la primera visita y el Grado de Miedo entre niños de 3 y 6 años. Calcula el p-valor del contraste y explica con claridad tus conclusiones. Utiliza $\alpha = 0{,}10$.

Ejercicios recopilatorios

Ejercicio 8.6.

Se desea estudiar si la infección por rotavirus tiene una relación significativa con la edad de los niños. Para ello, se ha considerado una muestra aleatoria de niños entre 1 y 5 años, para los que se han medido las dos variables de interés: el grupo de edad al que pertenece cada niño y si ha tenido o no infección por rotavirus. Los datos obtenidos se resumen en la siguiente tabla:

	Anticuerpos contra rotavirus	
Grupo de Edad	Sí	No
Menores de 1 año	30	120
Entre 1 y 5 años	40	110
Mayores de 5 años	30	70

a) Únicamente observando la tabla comenta razonadamente si crees que existe relación significativa entre estas dos variables o no.

b) Plantea y resuelve el contraste de hipótesis adecuado para comprobar si existe una relación significativa entre las dos variables involucradas en este problema e interpreta con claridad las conclusiones que se obtienen de este análisis. Utiliza como nivel de significatividad $\alpha = 0{,}10$.

Ejercicio 8.7.

Se desea estudiar si el sellado de fosas y fisuras dentales tiene relación significativa con la reducción de las caries dentales en niños de entre 1 y 5 años. Para ello, se ha considerado una muestra aleatoria de esta población y se han recogido los siguientes datos:

	Sellado de las fosas y fisuras	
Caries dentales	Sí	No
0	120	90
Entre 1 y 2 caries	50	60
2 o más	50	65

a) Plantea y resuelve el contraste de hipótesis adecuado para comprobar si existe una relación significativa entre las dos variables involucradas, mediante la metodología de las regiones de aceptación y rechazo. Utiliza como nivel de significatividad a=0.01

b) Resuelve el contraste del apartado anterior mediante el p-valor.

c) Interpreta los resultados en el contexto del ejercicio.

Ejercicio 8.8.

Se lleva a cabo una investigación odontológica para valorar si el Grado de dolor (Suave , Intenso y Muy Intenso) que sufren los pacientes tras una intervención quirúrgica, de las mismas características, tiene asociación con el Tipo de cirugía (Tradicional ó Innovadora) que se ha practicado al paciente. Durante el estudio los pacientes fueron clasificados del siguiente modo:

	Tipo de cirugía	
Grado de dolor	Tradicional	Innovadora
Suave	15	13
Intenso	11	6
Muy Intenso	20	3

a) Plantea y resuelve el contraste de hipótesis adecuado para comprobar si existe una relación significativa entre las dos variables involucradas, mediante la metodología de las regiones de aceptación y rechazo. Utiliza como nivel de significatividad a=0.05

b) Resuelve el contraste del apartado anterior mediante el p-valor.

c) Interpreta los resultados en el contexto del ejercicio.

Capítulo 9

Regresión lineal simple

En este tema abordamos el estudio de la relación entre **dos variables cuantitativas**. El estudio de esta relación nos puede indicar si existe dependencia o no entre ambas variables y, en el caso de que exista, de qué tipo es.

Ejemplo 9.1.

Supongamos que queremos explorar la relación entre la Edad (en meses) y la Talla (en cm) de niños con edades comprendidas entre los 3 y los 9 meses, a partir de una muestra de 14 niños.

Los datos diponibles para ambas variables se muestran a continuación:

Niño	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Edad (meses)	3	6	5	5	3	4	9	8	9	7	6	5	8	6
Talla (cm)	55	68	64	66	62	65	74	75	73	69	73	68	73	71

Respecto a estas dos variables (cuantitativas) nos podemos plantear cuestiones como las siguientes:

- *¿Cuál es la relación de la talla con la edad?.*
- *¿Podemos predecir cuál será la talla que tendrá un niño que, por ejemplo, tiene 6 meses de edad? Y, en caso de poder hacerlo, ¿sería bueno ese pronóstico?.*

Si se piensa que una variable puede depender (estar relacionada) con la otra se puede cuantificar esta relación. A este proceso se le denomina *Regresión*. Los objetivos de la *Regresión*, a grandes rasgos, son:

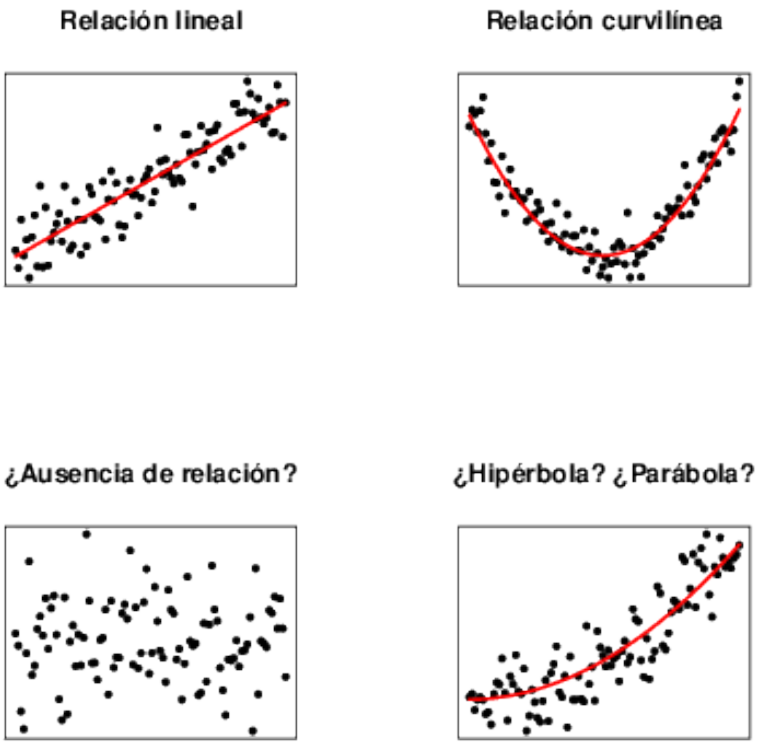
- Estudiar si dos variables aleatorias están relacionadas.
- Estudiar el tipo de relación, si existe, que las une.
- Predecir los valores de una de ellas a través de los valores de la otra.

La relación que pretendemos explorar entre las dos variables cuantitativas no se corresponde con una *relación determinística*, es decir, una relación perfecta exacta, como la que puede existir entre el espacio recorrido (E) por un móvil con velocidad cte (V) en un tiempo (T), que puede ser calculado como $E = V \cdot T$. La relación que abordamos explorar es una *relación aleatoria*, en la que sabemos que dado un valor de una de las variables, los valores que puede tomar la otra no están determinados con exactitud, pero sí podemos conocer de forma aproximada su distribución de probabilidad.

Para introducir una *idea intuitiva de regresión*, supongamos que disponemos de n parejas de valores de las dos variables cuantitativas en estudio, a las que denotaremos por X e Y. Los datos disponibles son los pares (x_i, y_i) para $i = 1, ..., n$. Cada par de valores define un punto que puede ser representado en el plano cartesiano, dando lugar todos ellos en conjunto a lo que conocemos como una *nube de puntos*. Si a dicha nube de puntos se le puede ajustar alguna curva se dice, entonces, que se puede llevar a cabo una

*regresión.*A tal curva se le llama *línea de regresión.* A la variable ubicada en el eje horizontal (abcisas) se la llama *variable independiente* y a la ubicada en el eje vertical (ordenadas) *variable dependiente.*

Las nubes de puntos originadas a partir de dos variables cuantitativas pueden tener infinidad de formas. Algunas de estas formas se muestran a continaución:



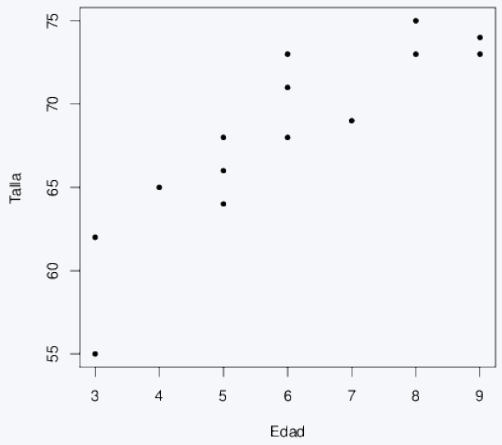
Dependiendo de la forma de la nube de puntos intentaremos ajustar una línea de regresión diferente. Cuando la forma de la nube de puntos se asemeja a una recta, decimos que una de las variables está relacionada *linealmente* con la otra variable, y denominamos al proceso de estudiar dicha relación *Regresión Lineal.*

A partir del estudio de la relación lineal entre dos variables podemos determinar la fuerza de la asociación (lineal) entre las dos variables. Además, si existe una relación lineal entre ellas, es posible predecir los valores de una de ellas (*Variable Dependiente*) en función de los valores de la otra (*Variable Independiente*).

Ejemplo 9.2.

Continuando con el ejemplo anterior vamos a obtener la representación gráfica de la nube de puntos que definen los datos recogidos por ambas variables.

A partir de los datos de la muestra anterior se obtiene el siguiente diagrama de dispersión (o nube de puntos).



Como podemos apreciar se trata de una nube de puntos a la que se le podría ajustar una recta, por tanto, tiene sentido apliar una *regresión lineal* para cuantificar la relación entre ambas variables.

Si existe relación entre dos variables X e Y, no necesariamente tiene que ser por que una sea causa de la otra. En general, se puede dar cualquiera de las situaciones siguientes:

- Que una de ellas sea realmente causa de la otra.
- Que ambas variables se influyan mutuamente.
- Que ambas variables dependan de una causa común (una tercera variable que no se esté considerando).

9.1. Coeficiente de correlación lineal

El coeficiente de correlación lineal mide la fuerza de asociación lineal con que dos variables aleatorias están ligadas (linealmente). Esta fuerza es medida por el coeficiente de correlación lineal poblacional (ρ). El coeficiente de correlación lineal poblacional es adimensional (no depende de las unidades de medida) y puede tomar valores en el intervalo $[-1, 1]$. Para interpretar el coeficiente de correlación lineal debemos interpretar por separado su *magnitud* y su *signo*:

- Su **signo** indica el *sentido de la asociación*:
  - $\rho > 0 \Rightarrow$ Asociación positiva. Al aumentar los valores de una de las variables aumentan los valores de la otra.
  - $\rho < 0 \Rightarrow$ Asociación negativa. Al aumentar los valores de una de las variables disminuyen los valores de la otra.
- Su **magnitud** indica la *fuerza de la asociación*:

- ρ cercano a 0 $\Rightarrow$ Independencia lineal o falta de asociación lineal.
- ρ cercano a 1 o -1 $\Rightarrow$ Fuerte asociación lineal.

Como todos los parámetros poblacionales, no es posible conocer el valor del coeficiente de correlación lineal poblacional existente entre dos variables cuantitativas, su valor se estima mediante el *coeficiente de correlación muestral* (r) obtenido a partir de los datos de la muestra (x_i, y_i) para las $i = 1, ..., n$ observaciones. El coeficiente de *correlación lineal muestral* se interpreta, a partir de su signo y su magnitud, exactamente igual que el coeficiente de correlación lineal poblacional, y s calcula a partir de la fórmula:

$$r = \frac{(xy)}{\sqrt{(xx)(yy)}}$$

donde (xx), (yy) y (xy) son las llamadas *Sumas de Cuadrados* que se obtienen de la siguiente forma:

$$(xx) = \sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}; \qquad (yy) = \sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n}$$

$$(xy) = \sum_{i=1}^{n} x_i \cdot y_i - \frac{(\sum_{i=1}^{n} x_i) \cdot (\sum_{i=1}^{n} y_i)}{n}$$

Ejemplo 9.3.

Continuando con el ejemplo de la Edad y la Talla de los niños de 3 a 9 meses mostramos el cálculo y la interpretación del coeficiente de correlación lineal muestral.

Si consideramos como variable X la variable Edad, y como variable Y la variable Talla, a partir de los datos de este ejemplo podemos obtener:

$$(xx) = 52 \qquad (yy) = 402{,}86 \qquad (xy) = 127$$

$$r = \frac{127}{\sqrt{(52) \cdot (402{,}86)}} = 0{,}88$$

Entre las variables Edad y Talla el coefciente de correlación lineal muestral obtenido toma el valor 0.88. Puesto que se trata de un valor positivo indica que la asociación lineal entre ambas variables es positiva, es decir, a medida que aumentan los valores de la Edad también aumentan los valores de la Talla (y viceversa). Su magnitud (0.88, cercana al valor 1), indica que se trata de una asociación lineal fuerte.

Obtenido un coeficiente de correlación muestral, ¿hasta qué punto podemos afirmar que existe una relación lineal significativa entre ambas variables? Existirá una relación lineal significativa siempre que el coeficiente de correlación poblacional (ρ) sea significativamente diferente de 0. Con este fin vamos a definir el llamado Test de Independencia lineal para el coeficiente de correlación que contrasta precisamente si el coeficiente de correlación lineal es significativamente diferente de 0.

9.1.1. Test de independencia lineal para el coeficiente de correlación lineal (ρ)

Este test se plantea el siguiente contraste de hipótesis:

$$H_0 : \rho = 0$$
$$H_1 : \rho \neq 0$$

La resolución de este contraste se lleva a cabo mediante la defnición del pivote correspondiente que, bajo la hipótesis nula sigue una distribución t de Student con $n - 2$ grados de libertad. La expresión de este pivote se muestra a continuación:

$$Pivote = \sqrt{\frac{(n-2)r^2}{1-r^2}} \sim t_{n-2}$$

Ejemplo 9.4.

A partir de los datos del ejemplo con el que estamos ilustrando este tema nos preguntamos si existe una relación lineal significativa entre la Edad y la Talla de los niños de 3 a 9 meses de edad (consideraremos, como es habitual, un nivel de significatividad de $\alpha = 0{,}05$***.***

En primer lugar nos planteamos el contraste adecuado:

$$H_0 : \rho = 0$$
$$H_1 : \rho \neq 0$$

Conocemos que, para este problema, el coeficiente de correlación lineal muestral toma el valor de $r = 0{,}88$ y que el valor de n en este caso es 14, ya que disponemos de 14 parejas de valores (o de datos de 14 individuos). Pasamos a calcular el pivote asociado a este contraste:

$$Pivote = \sqrt{\frac{(14-2)(0{,}88)^2}{1-(0{,}88)^2}} = 6{,}41$$

Bajo la hipótesis nula, el pivote sigue una distribución t_{12}. El p-valor asociado a este pivote, en esta distribución y en un contraste bilateral es $< 0{,}001$, por lo que se rechaza la hipótesis nula (independencia lineal entre las variables) y podemos concluir que existe una relación lineal significativa entre la Edad y la Talla de los niños entre 3 y 9 meses.

9.1.2. Coeficiente de determinación

El coeficiente de determinación poblacional es un indicador de la bondad del ajuste que proporciona la recta de regresión. Este coeficiente se calcula como ρ^2 y representa la proporción de la variabilidad total de Y que es capaz de explicar la variable X. Este coeficiente toma valores dentro del intervalo $[0, 1]$. En cuanto a su interpretación nos podemos apoyar en las siguientes indicaciones:

- Si ρ^2 es cercano a 0, la variabilidad observada en Y no se explica por su relación con X.
- Si ρ^2 es cercano a 1, la variabilidad observada en Y se debe, en gran parte, a la variación de X.

La estimación del coeficiente de determinación poblacional ρ^2 es su correspondiente pero a nivel muestral r^2. Habitualemente, para expresar la información que proporciona se suele pasar a porcentaje (simplemente multiplicándolo por 100).

Ejemplo 9.5.

Coeficiente de determinación en el ejemplo anterior

A partir de los datos del ejemplo de la Edad y la Talla hemos obtenido un coeficiente de correlación muestral de $r = 0{,}88$. Por tanto, el coeficiente de determinación muestral será $r^2 = (0{,}88)^2 = 0{,}7744$. Un $77.44\,\%$ de la variabilidad observada en la Talla queda explicada por su relación con la Edad (y cómo varía esta última variable).

9.2. El modelo de regresión lineal

Cuando el coeficiente de correlación lineal muestral indica que existe una relación lineal entre las variables en estudio tiene sentido ajustar un modelo de *regresión lineal* entre las dos variables. Ajustar un

modelo de regresión lineal entre las variables es equivalente a ajustar la recta que mejor ajusta la nube de puntos que definen las dos variables conjuntamente.

En este momento es necesario fijar cuál será la variable dependiente y cuál será la variable independiente. La variable dependiente es la variable que nos gustaría predecir conociendo los valores de la otra variable, de la variable independiente. En determinados ejemplos, es indiferente fijar una u otra como variable dependiente. Sin embargo, en otros es importante reflexionar sobre qué variable debe tomar cada posición. Si por ejemplo, una de las variables es difícil de medir, mientras que la otra resulta mucho más sencilla de obtener, lo lógico es fijar como variable dependiente la de difícil medición y como variable independiente la más sencilla. De esta forma, si la relación lineal lo permite, podremos estimar (con un determinado error) qué valor tomará la variable difícil simplemente conociendo el valor que toma en la variable que se mide de forma más sencilla. También ayuda el pensar qué variable pensamos que depende de la otra. Si simplemente la relación es mutua no importa el papel asignado a cada una de ellas, pero si resulta más lógico pensar que los valores de una de ellas dependen de los valores de la otra, y no al revés, esa debe ser la variable dependiente.

A la variable independiente la denotaremos con la letra X y a la variable dependiente la denotaremos con la letra Y. Gráficamente, la variable X se suele representar en el *eje de abscisas* (también conocido como *eje x*) y la variable Y en el *eje de ordenadas* (también conocido como *eje y*).

Puesto que tratamos de ajustar una recta a la nube de puntos que definen las dos variables, debemos recordar que cualquier recta tiene como expresión:

$$Y = A + B \cdot X$$

donde A representa la ordenada en el origen y B representa la pendiente de la recta. La ordenada en el origen indica el valor que toma la Y cuando la X toma el valor 0. La pendiente representa el valor que se incrementa la variable Y por cada valor que aumenta la variable X.

El ajuste de una recta sin más, no representaría la relación aleatoria que queremos estimar (se asemeja más a la expresión de una relación determinista). Sabemos que las variables X e Y estarán relacionadas pero, sin embargo, sabemos que el concocer el valor de una de ellas (de X por ejmplo) no nos conducirá de forma determinista al valor que tomará la otra (la Y). Por este motivo, entendemos que la relación lineal (en la población) que unirá a las variables X e Y s puede expresar a partir del siguiente modelo:

$$Y = A + B \cdot X + \varepsilon$$

donde ε es el responsable de la varianza. Conocido el valor de X esperamos que la variable Y tome un valor aproximado de $A + B \cdot X$ con un determinado error que viene dado por ε. Concretamente se asume que $\varepsilon \sim N(0, \sigma^2)$. Al parámetro B, pendiente de la recta de regresión lineal, se le conoce también como *coeficiente de regresión*.

Los parámetros A y B representan los coeficientes de la recta de regresión que mejor se ajusta a la nube de puntos que definen las variables X e Y en la población (y σ^2 la varianza asociada a esta relación). Como siempre, estos valores son parámetros de la población difíciles o imposibles de determinar, por lo que, como es habitual, obtenemos una estimación de los mismos a partir de la muestra.

Calcularemos la recta de regresión que mejor se ajusta a los datos proporcionados por nuestra muestra (mediante una técnica llamada *Mínimos cuadrados*) y la pendiente de esta recta (b) y su ordenada en el origen (a) serán estimadores puntuales de estos mismos parámetros en la población. Los estimadores a y b se calcularán a partir de las fórmulas que indicamos a continuación:

$$b = \frac{(xy)}{(xx)}; \qquad a = \overline{y} - b \cdot \overline{x}$$

donde (xy), y (xx) son las sumas de cuadrados definidas en la sección anterior, y $\overline{x}$ y $\overline{y}$ son las medias muestrales de los valores de X e Y en la muestra, respectivamente. La varianza poblacional de la recta, σ^2, será estimada mediante el estimador s^2 que se calcula mediante la fórmula:

$$s^2 = \frac{(yy) - \frac{(xy)^2}{(xx)}}{n-2}$$

La varianza s^2, y por tanto la desviación típica s, representan una estimación del error de predicción con la recta estimada.

A continuación mostramos, a modo de resumen, los parámetros poblacionales asociados a la recta de regresión lineal y su correspondencia con sus estimadores muestrales:

	Parámetro poblacional	Estimador puntual (muestra)
Ordenada en el origen	A	a
Pendiente	B	b
Recta de regresión	$Y = A + B \cdot X$	$\widehat{Y} = a + b \cdot X$
Varianza del modelo	σ^2	s^2
Desviación típica del modelo	σ	s

Ejemplo 9.6.

Continuamos con el ejemplo de la Edad y la Talla de los niños de 3 a 9 meses. Vamos a calcular a continuación la estimación de la recta de regresión lineal que nos permitiría estimar la Talla de los niños en función de su Edad.

En este caso, tal y como está planteado el ejemplo, vamos a considerar que la variable dependiente es la Talla (pues es la que queremos predecir) y la variable independiente es la Edad. Así, conociendo el valor de la edad de un niño, el modelo de regresión nos permitirá estimar su talla esperada (con un determinado error). Partiendo de los siguientes cálculos: $(xy) = 127$, $(xx) = 52$, $\overline{x} = 6$ y $\overline{y} = 68{,}29$, podemos obtener los coeficientes de la recta de regresión:

$$b = \frac{127}{52} = 2{,}44 \qquad a = 68{,}29 - (2{,}44)(6) = 53{,}65$$

$$s^2 = \frac{(402{,}86) - \frac{(127)^2}{(52)}}{12} = 7{,}72 \quad \Rightarrow \quad s = \sqrt{s^2} = \sqrt{7{,}72} = 2{,}778$$

La ecuación de la recta ajustada es:

$$\widehat{Talla} = a + b \cdot Edad = 53{,}65 + 2{,}44 \cdot Edad$$

Y el error que esperamos cometer, de media, cuando realicemos una predicción de la talla de un niño a partir de su edad utilizando esta recta es de 2.778 centímetros.

Mediante la recta de regresión estimada podemos realizar predicciones de qué valores tomará la variable Y (en promedio) para un individuo que tome una valor x_0 para la variable X. En ese caso nuestra predicción para el valor de Y de ese individuo será:

$$\widehat{Y} = a + b \cdot x_0$$

Estas predicciones deben ser utilizadas únicamente para individuos que tomen valores de la variable X en el rango en el que se han tomado los datos.

Ejemplo 9.7.

Estimación de la Talla media espera de un niño de 5.5 meses y de otro niño de 15 meses.

La primera estimación podemos hacerla a partir de la recta de regresión:

$$\widehat{Y} = 53{,}65 + 2{,}44 \cdot 5{,}5 = 67{,}07$$

Para un niño de 5.5 meses esperaríamos una Talla media de 67.07 cm. El error que esperaríamos cometer con esta predicción es de 2.778 cm.
Sin embargo, para la predicción de la Talla de un niño de 15 meses no deberíamos utilizar la misma recta, pues el rango de edades que hemos contemplado en nuestro estudio y del que tenemos datos no abarca los 15 meses, por lo que en ese caso no sabemos qué comportamiento va a tener la relación entre ambas variables. No debemos utilizar la recta de regresión estimada para realizar esta estimación.

9.2.1. Test de independencia lineal para el coeficiente de regresión (B)

De la misma forma que podemos contrastar si la relación lineal entre las dos variables es significativa, podemos realizar un contraste equivalente mediante la pendiente del modelo. Si la pendiente de la recta de regresión poblacional fuera 0, indicaría ausencia de relación lineal entre las variables (el incremento de la variable independiente no tendría efecto sobre la variable dependiente). Por este motivo, es posible plantear un test de independencia lineal equivalente al planteado mediante el coeficiente de correlación lineal poblacional, pero en este caso a partir de la pendiente poblacional:

Este test se planteará el siguiente contraste de hipótesis:

$$H_0 : B = 0$$
$$H_1 : B \neq 0$$

La resolución de este contraste se lleva a cabo mediante la defínción del pivote correspondiente que, bajo la hipótesis nula sigue una distribución t de Student con $n-2$ grados de libertad. La expresión de este pivote se muestra a continuación:

$$Pivote = \frac{b - B}{\sqrt{\frac{s^2}{(xx)}}} \sim t_{n-2}$$

Ejemplo 9.8.

A partir de los datos del ejemplo con el que estamos ilustrando este tema nos preguntamos si existe una relación lineal significativa entre la Edad y la Talla de los niños de 3 a 9 meses de edad. En este caso, lo realizaremos mediante el test de independencia mediante la pendiente del modelo lineal (consideraremos, como es habitual, un nivel de significatividad de $\alpha = 0{,}05$***.)***

En primer lugar nos planteamos el contraste adecuado:

$$H_0 : B = 0$$
$$H_1 : B \neq 0$$

Conocemos que, para este problema, el coeficiente de correlación lineal muestral toma el valor de $r = 0{,}88$ y que el valor de n en este caso es 14, ya que disponemos de 14 parejas de valores (o de datos de 14 individuos). Pasamos a calcular el pivote asociado a este contraste:

$$Pivote = \frac{2{,}44 - 0}{\sqrt{\frac{7{,}72}{52}}} = 6{,}33$$

Bajo la hipótesis nula, el pivote sigue una distribución t_{12}. El p-valor asociado a este pivote, en esta distribución y en un contraste bilateral es $< 0{,}001$, por lo que se rechaza la hipótesis nula (independencia lineal entre las variables) y podemos concluir que existe una relación lineal significativa entre la Edad y la Talla de los niños entre 3 y 9 meses.

Además de resolver el test de hipótesis de independencia lineal con la pendiente de la recta de regresión lineal, sería equivalente calcular el intervalo de confianza, con la confianza deseada, y comprobar si el valor 0 está contenido en este intervalo. La expresión que nos permite calcular el intervalo de confianza al $(1 - \alpha) \times 100\,\%$ es la siguiente:

$$(b - t_{(1-\frac{\alpha}{2},n-2)}\sqrt{\frac{s^2}{(xx)}} \quad , \quad b + t_{(1-\frac{\alpha}{2},n-2)}\sqrt{\frac{s^2}{(xx)}})$$

Ejemplo 9.9.

Calculamos ahora el intervalo de confianza al 95 % para el coeficiente de regresión que relaciona la Talla con la Edad de los niños.

$$(2{,}44 - 2{,}179 \cdot \sqrt{\frac{7{,}72}{52}} \ , \ 2{,}44 - 2{,}179 \cdot \sqrt{\frac{7{,}72}{52}}) = (1{,}60 \ , \ 3{,}28)$$

Con un 95 % de confianza, la pendiente de la recta de regresión que explica la Talla de los niños en función de su Edad está contenida en este intervalo. Puesto que el 0 no está contenido en este intervalo, existe una relación de dependencia lineal significativa entre la Talla y la Edad de los niños entre 3 y 9 meses.

9.3. Ejercicios Capítulo 9

Ejercicio 9.1.

Reflexiona, en cada uno de los casos que se exponen a continuación, sobre cuál de las dos variables de los siguientes pares, fijarías como variable dependiente y cuál como independiente. Además, expresa también si consideras que la relación que se obtendría es positiva o negativa:

1. potencia de un coche y precio.
2. tensión arterial y consumo de sal.
3. consumo de tabaco y duración de vida.

Ejercicio 9.2.

Se desea estudiar si la altura de los hijos se puede explicar linealmente en función de la altura de sus padres.

Padres	1.70	1.77	1.68	1.75	1.80	1.75	1.69	1.72	1.71	1.73
Hijos	1.74	1.78	1.72	1.77	1.78	1.77	1.71	1.76	1.73	1.74

Ayuda: (xx)=0.0128, (yy)=0.0058, (xy)=0.0078. **Utiliza como nivel de significativad** $\alpha = 0{,}05$.

1. Determina cuál es la variable independiente y cuál es la variable dependiente
2. Calcula el coeficiente de correlación muestral. Determina, en función del test de independencia sobre el coeficiente de correlación lineal, si la asociación entre las dos variables es significativa. Indica el sentido de asociación.
3. Estima la recta de regresión que mejor ajusta a los datos e interpreta sus coeficientes en el contexto del ejercicio.
4. Estima el error de predicción que se comete cuando se utiliza la recta de regresión del apartado anterior.
5. Plantea y resuelve el contraste de independencia sobre el coeficiente de regresión. Explica si las conclusiones que obtienes son equivalentes a las obtenidas con el test de independencia del coeficiente de correlación lineal.
6. Estima el coeficiente de determinación e interpreta su resultado en el contexto del ejercicio.

Ejercicio 9.3.

Un estudiante que busca piso ha tomado los siguientes datos de los precios de alquiler semanal y de la superficie de los pisos en metros cuadrados.

Superficie	60	60	80	90	100	110	120
Precio	70	85	80	90	85	110	115

Ayuda: (xx)= 3285.714, (yy)= 1571.429, (xy)= 1957.143. **Utiliza como nivel de significativad** $\alpha = 0{,}05$.

1. Determina cuál es la variable independiente y cuál es la variable dependiente
2. Calcula el coeficiente de correlación muestral. Determina, en función del test de independencia sobre el coeficiente de correlación lineal, si la asociación entre las dos variables es significativa. Indica el sentido de asociación.
3. Estima la recta de regresión que mejor ajusta a los datos e interpreta sus coeficientes en el contexto del ejercicio.

4. Estima el error de predicción que se comete cuando se utiliza la recta de regresión del apartado anterior.
5. Plantea y resuelve el contraste de independencia sobre el coeficiente de regresión. Explica si las conclusiones que obtienes son equivalentes a las obtenidas con el test de independencia del coeficiente de correlación lineal.
6. Estima el coeficiente de determinación e interpreta su resultado en el contexto del ejercicio.

Ejercicio 9.4.

Una encuesta de salarios entre graduados proporciona los datos siguientes:

Edad (años)	28	28	32	35	38	44	49	52	58	62	66	70
Salario (miles de euros)	2.2	2.2	3.8	4.2	4.2	5.3	7.3	6.4	6.7	5.3	6.0	5.1

Ayuda: (xx)= 2445.667, (yy)= 29.58917, (xy)= 196.1833. **Utiliza como nivel de significativad** $\alpha = 0{,}01$.

1. Determina cuál es la variable independiente y cuál es la variable dependiente
2. Calcula el coeficiente de correlación muestral. Determina, en función del test de independencia sobre el coeficiente de correlación lineal, si la asociación entre las dos variables es significativa. Indica el sentido de asociación.
3. Estima la recta de regresión que mejor ajusta a los datos e interpreta sus coeficientes en el contexto del ejercicio.
4. Estima el error de predicción que se comete cuando se utiliza la recta de regresión del apartado anterior.
5. Plantea y resuelve el contraste de independencia sobre el coeficiente de regresión. Explica si las conclusiones que obtienes son equivalentes a las obtenidas con el test de independencia del coeficiente de correlación lineal.
6. Estima el coeficiente de determinación e interpreta su resultado en el contexto del ejercicio.

Ejercicio 9.5.

Se desea estudiar si el nivel en sangre de estradiol tiene relación lineal con la edad de las mujeres, con el objetivo de predecir y modificar su nivel farmacológicamente en edades que lo necesiten. Para ello, se considera una muestra de 10 mujeres de las que se ha tomado su edad (en años) y su nivel de estradiol (en pg/ml):

Edad	14.3	21.2	25.7	35.2	38.2	41.8	47.2	51.3	54.5	62.7
Estradiol	193.7	195.2	185.3	152.7	120.7	88.3	75.2	47.5	25.1	24.2

Ayuda: (xx)= 2146.769, (yy)= 42024.19, (xy)= -9222.199. **Utiliza como nivel de significativad** $\alpha = 0{,}05$.

1. Determina cuál es la variable independiente y cuál es la variable dependiente
2. Calcula el coeficiente de correlación muestral. Determina, en función del test de independencia sobre el coeficiente de correlación lineal, si la asociación entre las dos variables es significativa. Indica el sentido de asociación.
3. Estima la recta de regresión que mejor ajusta a los datos e interpreta sus coeficientes en el contexto del ejercicio.
4. Estima el error de predicción que se comete cuando se utiliza la recta de regresión del apartado anterior.
5. Plantea y resuelve el contraste de independencia sobre el coeficiente de regresión. Explica si las conclusiones que obtienes son equivalentes a las obtenidas con el test de independencia del coeficiente de correlación lineal.

6. Estima el coeficiente de determinación e interpreta su resultado en el contexto del ejercicio.

Ejercicio 9.6.

La siguiente tabla contiene la tensión arterial sistólica (TAS), medida en mm de Hg y la hemoglobina glicosilada (HBA_1), expresada en %, de una muestra de 5 pacientes.

TAS	150	120	145	140	130
HBA_1	8.2	6.1	7.2	7	6

AYUDA: (xx)= 3.24, (yy)= 580, (xy)= 39.5. **Utiliza como nivel de significativad** $\alpha = 0{,}05$.

1. Determina cuál es la variable independiente y cuál es la variable dependiente
2. Calcula el coeficiente de correlación muestral. Determina, en función del test de independencia sobre el coeficiente de correlación lineal, si la asociación entre las dos variables es significativa. Indica el sentido de asociación.
3. Estima la recta de regresión que mejor ajusta a los datos e interpreta sus coeficientes en el contexto del ejercicio.
4. Estima el error de predicción que se comete cuando se utiliza la recta de regresión del apartado anterior.
5. Plantea y resuelve el contraste de independencia sobre el coeficiente de regresión. Explica si las conclusiones que obtienes son equivalentes a las obtenidas con el test de independencia del coeficiente de correlación lineal.
6. Estima el coeficiente de determinación e interpreta su resultado en el contexto del ejercicio.

Ejercicio 9.7.

Se desea investigar si el peso de las personas tiene influencia lineal sobre el colesterol LDL.

LDL (mg/dl)	131	143	178	189	121	99
Peso (Kg)	67	71.24	89.56	92.50	81.70	65.80

AYUDA: (xx)= 673.10, (yy)= 5903.5, (xy)= 1659.1. **Utiliza como nivel de significativad** $\alpha = 0{,}1$.

1. Determina cuál es la variable independiente y cuál es la variable dependiente
2. Calcula el coeficiente de correlación muestral. Determina, en función del test de independencia sobre el coeficiente de correlación lineal, si la asociación entre las dos variables es significativa. Indica el sentido de asociación.
3. Estima la recta de regresión que mejor ajusta a los datos e interpreta sus coeficientes en el contexto del ejercicio.
4. Estima el error de predicción que se comete cuando se utiliza la recta de regresión del apartado anterior.
5. Plantea y resuelve el contraste de independencia sobre el coeficiente de regresión. Explica si las conclusiones que obtienes son equivalentes a las obtenidas con el test de independencia del coeficiente de correlación lineal.
6. Estima el coeficiente de determinación e interpreta su resultado en el contexto del ejercicio.

Ejercicio 9.8.

Se desea estudiar si un nuevo fármaco es eficaz en el control de la glucemia para diabéticos. Para ello, se dispone de la dosis de tratamiento recibida por cada paciente (en ml) y el nivel medio de glucemia tras la ingesta del tratamiento. Los datos se muestran en la siguiente tabla:

Dosis (ml)	65	52	80	64	54	50
Glucemia	125	110	150	132	115	102

AYUDA: (xx)= 636.83, (yy)= 1485.33, (xy)= 951.33. **Utiliza como nivel de significativad** $\alpha = 0{,}05$.

1. Determina cuál es la población de estudio.
2. Determina cuál es la variable independiente y cuál es la variable dependiente
3. Calcula el coeficiente de correlación muestral. Determina, en función del test de independencia sobre el coeficiente de correlación lineal, si la asociación entre las dos variables es significativa. Indica el sentido de asociación.
4. Estima la recta de regresión que mejor ajusta a los datos e interpreta sus coeficientes en el contexto del ejercicio.
5. Estima el error de predicción que se comete cuando se utiliza la recta de regresión del apartado anterior.
6. Plantea y resuelve el contraste de independencia sobre el coeficiente de regresión. Explica si las conclusiones que obtienes son equivalentes a las obtenidas con el test de independencia del coeficiente de correlación lineal.
7. Estima el coeficiente de determinación e interpreta su resultado en el contexto del ejercicio.

Ejercicio 9.9.

La dexametasona es un corticoide que se utiliza en el tratamiento del asma, sin embargo su utilización aumenta la glucemia. Se desea investigar si la glucemia se puede explicar linealmente en función de la dosis de dexametasona (ml/día) para pacientes con este tipo de enfermedad. Para ello, se han obtenido los siguientes datos de 5 pacientes:

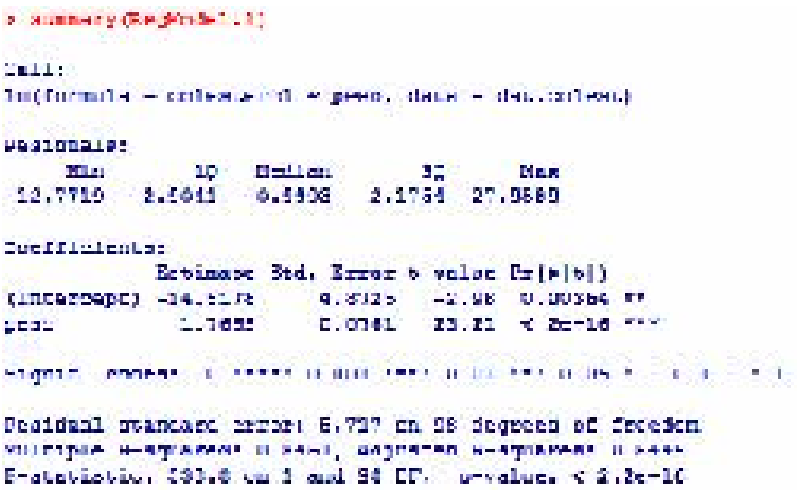
dexametasona (ml/día)	1	4	3	5	10
glucemia (mg/dl)	132	152	141	153	173

AYUDA: (xx)= 45.2, (yy)= 946.8, (xy)= 203.4. **Utiliza como nivel de significativad** $\alpha = 0{,}01$.

1. Determina cuál es la población de estudio.
2. Determina cuál es la variable independiente y cuál es la variable dependiente
3. Calcula el coeficiente de correlación muestral. Determina, en función del test de independencia sobre el coeficiente de correlación lineal, si la asociación entre las dos variables es significativa. Indica el sentido de asociación.
4. Estima la recta de regresión que mejor ajusta a los datos e interpreta sus coeficientes en el contexto del ejercicio.
5. Estima el error de predicción que se comete cuando se utiliza la recta de regresión del apartado anterior.
6. Plantea y resuelve el contraste de independencia sobre el coeficiente de regresión. Explica si las conclusiones que obtienes son equivalentes a las obtenidas con el test de independencia del coeficiente de correlación lineal.
7. Estima el coeficiente de determinación e interpreta su resultado en el contexto del ejercicio.

Ejercicio 9.10.

Se desea investigar si el colesterol, medido en mg/dl, depende linealmente del peso de las personas, medido en Kg. Para ello, se lleva a cabo un estudio sobre una muestra de 100 pacientes. Los datos analizados, con el programa R-Commander, han proporcionado la siguiente salida:



1. Determina cuál es la población de estudio.
2. Determina cuál es la variable independiente y cuál es la variable dependiente
3. Indica cuál es el coeficiente de determinación e interpreta su resultado en el contexto del ejercicio.
4. Calcula el coeficiente de correlación muestral. Determina, en función del test de independencia sobre el coeficiente de correlación lineal, si la asociación entre las dos variables es significativa. Indica el sentido de asociación.
5. Escribe la recta de regresión que mejor ajusta a los datos e interpreta sus coeficientes en el contexto del ejercicio.
6. Indica el error de predicción e interpreta su resultado.
7. Indica qué distribución se ha utilizado para calcular el p-valor que aparece en la línea del peso. Escribe el contraste de hipótesis al que da respuesta. Explica si las conclusiones que obtienes son equivalentes a las obtenidas con el test de independencia del coeficiente de correlación lineal.

Apéndice A

Anexo I: Soluciones Numéricas Ejercicios

A.1. Soluciones numéricas Ejercicios Capítulo 1

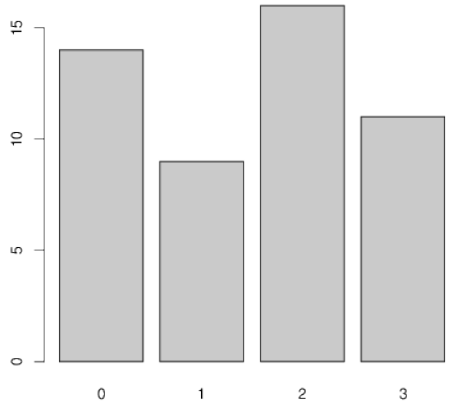
Ejercicio 1.1

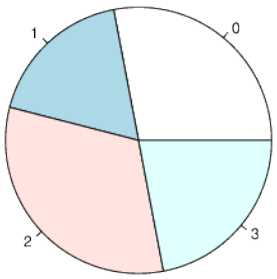
- Talla de la camiseta (S, M, L, XL, XXL): *Cualitativa ordinal*
- Número de calzado: *Cuantitativa discreta (¿cualitativa ordinal?)*.
- Temperatura corporal de un paciente: *Cuantitaiva continua*.
- Día de la semana: *Cualitativa ordinal*.
- Número de hijos: *Cuantitativa discreta*.
- Último libro leido: *Cualitativa nominal*.
- Grado de aceptación de una decisión (de acuerdo, neutral, en desacuerdo): *Cualitativa ordinal*.
- Marca de café preferida: *Cualitativa nominal*.
- Línea de autobús que toma más frecuentemente: *Cualitativa nominal*.
- Número de asignaturas aprobadas el último curso: *Cuantitativa discreta*.

Ejercicio 1.2

1. Se trata de una variable *cualitativa ordinal*.
2. Tabla de frecuencias:
'

Respuesta	f_a	f_r
0 (Muy desfavorable)	14	0.28
1 (Desfavorable)	9	0.18
2 (Favorable)	16	0.32
3 (Muy favorable)	11	0.22





Ejercicio 1.3

1. Se trata de una variable *cuantitativa discreta*.

2. Tabla de frecuencias:

Respuesta	f_a	f_r	%
0	7	0.175	17.5 %
1	4	0.100	10.0 %
2	6	0.150	15.0 %
3	7	0.175	17.5 %
4	4	0.100	10.0 %
5	4	0.100	10.0 %
6	8	0.200	20.0 %

Ejercicio 1.4

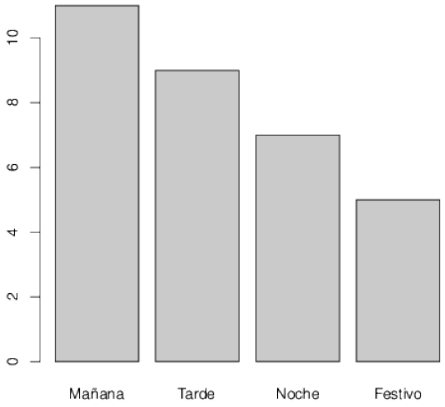
- *Unidades experimentales*: Cada una de las peticiones de asistencia a domicilio posibles; *Variable de estudio:* Horario de las asistencias a domicilio; *Tipo de la variable de estudio:* Cualitativa - Nominal.

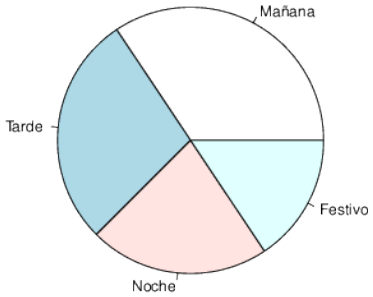
- Mediana y rango de los datos: No se pueden calcular estos estadísticos en variables cualitativas.

- Tabla de frecuencias:

Respuesta	f_a	f_r
Mañana	11	0.34
Tarde	9	0.28
Noche	7	0.22
Festivo	5	0.16

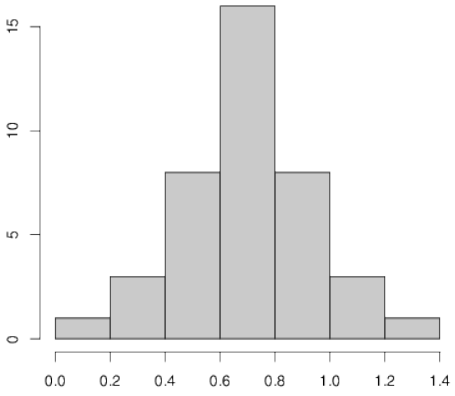
- Representaciones gráficas:

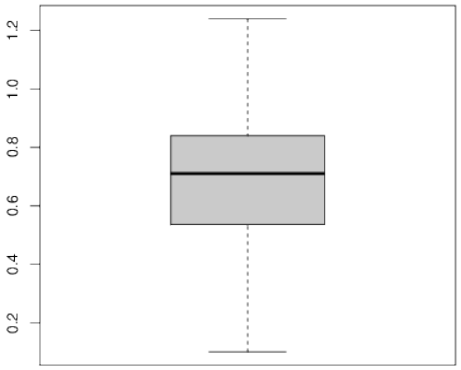




Ejercicio 1.5

- *Unidades experimentales:* Niños entre 1 y 5 años; *Variable de estudio:* Nivel de cobre en orina; *Tipo de la variable de estudio:* Cuantitativa - Continua.
- *Mediana*: 0.71 ; *Rango de datos:* 1.14
- *Primer cuartil:* 0.53; *Tercer cuartil:* 0.84; *Rango intercuartílico:* 0.31; *Percentil 10:* 0.37; *Percentil 95:* 1.16;
- *Intervalo que determina si existen o no valores atï¿½picos:* (0.065,1.305).No existen valores atï¿½picos en la muestra.
- Representaciones gráficas





Ejercicio 1.6

1. Variable cuantitativa continua.
2.
   - Mínimo:3281

- Máximo:4422
- P10: 3501.5
- P25 (Q1): 3887.25
- P50 (Q2)(Mediana): 4125
- P75 (Q3): 4187
- P90: 4361.8.0
- Media ($\bar{x}$): 4033
- Moda (aproximada por intervalo): 4000-4200
- Rango: 1141
- Rango IC: 299.75
- Varianza (S^2): 82981.2
- Desviación típica (S): 288.0646
- Coeficiente de variación (CV): 7.1419 %

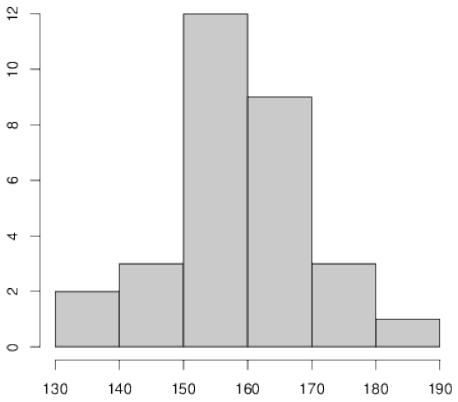
Ejercicio 1.7

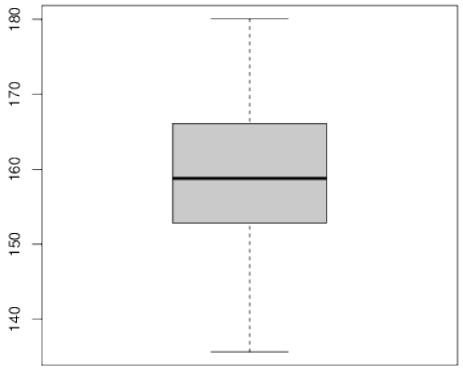
1. Variable cuantitativa continua
2. Tabla de frecuencias:
   Datos ordenados:
   135.62 138.77 141.59 143.35 147.47 150.29 151.11 152.83 152.99 154.06 154.53 156.49 158.49 158.66 158.72 158.81 159.97 160.82 161.92 162.04 162.5 165.54 166.13 166.99 167.70 168.11 172.93 173.03 176.77 180.08

Intervalo	f_a	f_r	%
$[130, 140)$	2	0.067	6.7 %
$[140, 150)$	3	0.100	10.0 %
$[150, 160)$	12	0.400	40.0 %
$[160, 170)$	9	0.300	30.0 %
$[170, 180)$	3	0.100	10.0 %
$[180, 190)$	1	0.033	3.3 %

3. - Mínimo: 135.6
   - Máximo: 180.1
   - P10: 141.766
   - P25 (Q1): 152.400
   - P50 (Q2)(Mediana): 158.765
   - P75 (Q3): 166.345
   - P90: 173.020
   - Media: 158.6
   - Moda (aproximada con intervalo): 150-160
   - Rango: 44.5
   - RIC: 13.945
   - Varianza (S^2): 117.145
   - Desviación típica (S) 10.823
   - Coeficiente de variación (CV): 6.82 %

4. Representaciones gráficas:
   Intervalo para comprobar si hay valores atípicos: (131.48, 187.26) No hay valores atípicos.





A.2. Soluciones numéricas Ejercicios Capítulo 2

Ejercicio 2.1

$Z \sim N(0,1)$

1. $P(Z < 1{,}56) = 0{,}9406$
2. $P(Z < 2{,}78) = 0{,}9972$
3. $P(Z > 3{,}00) = 0{,}0014$
4. $P(Z > 1{,}01) = 0{,}1563$
5. $P(Z < -1{,}5) = 0{,}0669$
6. $P(Z > -2{,}61) = 0{,}9954$
7. $P(Z < -0{,}32) == 0{,}3745$
8. $P(Z > -1{,}63) = 0{,}9484$
9. $P(0{,}83 < Z < 1{,}64) = 0{,}1527$
10. $P(-1{,}25 < Z < 2{,}37) =) = 0{,}8854$
11. $P(-2{,}36 < Z < -1{,}33)) = 0{,}0826$
12. Valor z_1 tal que $P(Z < z_1) = 0{,}648$: $z_1 = 0{,}38$
13. Valor z_1 tal que $P(Z < z_1) = 0{,}468$: $z_1 = -0{,}08$
14. Valor z_1 tal que $P(Z > z_1) = 0{,}9978$: $z_1 = -2{,}85$

Ejercicio 2.2

$Z \sim N(0,1)$

1. I(90 %): $(-1{,}645, 1{,}645)$
2. I(95 %): $(-1{,}96, 1{,}96)$
3. I(99 %): $(-2{,}58, 2{,}58)$
4. $(-1{,}645, +\infty)$
5. $(-\infty, 1{,}645)$

Ejercicio 2.3

$Z \sim N(0,1)$

1. $P_{5O} = 0{,}00$
2. $P_{75} = 0{,}67$
3. $P_{90} = 1{,}28$
4. $P_{10} = -1{,}28$
5. $P_{25} = -0{,}67$

Ejercicio 2.4

$X \sim N(7,2)$

1. $(3{,}08, 10{,}92)$
2. 0,9599 (95,99 %)
3. $(1{,}84, 12{,}16)$
4. $(4{,}42, +\infty)$

5. $1 - 0{,}9599 = 0{,}0401$ (4,01 %)
6. 0,8944(89,44 %)
7.
   - Entre 6 y 8: (38,3 %)
   - Entre 8 y 9: (14,98 %)
   - Entre 4 y 5: (9,19 %)

Ejercicio 2.5

$X \sim N(175, 8)$

1. (159 , 191)
2. $(-\infty$, 188)
3. (162 , $+\infty)$
4. 3 %
5. 81 %
6. 63 %

Ejercicio 2.6

$X_i \sim N(7{,}6, 0{,}81)$; $X_{ni} \sim N(9{,}6, 1{,}0)$

1. 0,5793
2. 0,0047
3. 0,9867
4.
   - Células infectadas:(6,01 , 9,19)
   - Células no infectadas: (7,64 , 11,56)

Ejercicio 2.7

$X \sim N(22{,}5; 2{,}85)$

a) 0,0043
b) (18,852 , 26,148)
c) 20,59

A.3. Soluciones numéricas Ejercicios Capítulo 3

Ejercicio 3.1

$X \sim N(23{,}5; 2{,}85)$; $\overline{X} \sim N(23{,}5; \frac{2{,}85}{\sqrt{9}})$; $\overline{x} = 21{,}69$
$P(\overline{X} < \overline{x}) = 0{,}0281$

Ejercicio 3.2

$X \sim N(1200; 400)$; $\overline{X} \sim N(1200; \frac{400}{\sqrt{9}})$
$P(\overline{X} < 1050) = 0{,}1293$

Ejercicio 3.3

$X \sim N(800, 120)$; $\overline{X} \sim N(800, \frac{120}{\sqrt{100}})$
a)$P(\overline{X} < 768) = 0{,}0038$

Ejercicio 3.4

$\widehat{P} \sim N(15, \sqrt{\frac{15*85}{100}})$
a)$P(\widehat{p} > 16) = 0{,}3898$

A.4. Soluciones numéricas Ejercicios Capítulo 4

Ejercicio 4.1

IC(95 %)=[87.2; 92.8]

Ejercicio 4.2

$IC(95\,\%)_{7a\tilde{n}os} = [9{,}25\,\%; 47{,}89\,\%]$
$IC(95\,\%)_{9a\tilde{n}os} = [53{,}78\,\%; 96{,}22\,\%]$

Ejercicio 4.3

IC(95 %)=[77.14; 83.13]

Ejercicio 4.4

$t_{99;0,95} \approx 1{,}66$
IC(90 %)=[21.34; 24.66]

Ejercicio 4.5

$z_{0,95} = 1{,}64$
IC(90 %)=[7.64; 8.87]

Ejercicio 4.6

$t_{24;0,95} = 1{,}71$
IC(90 %)=[34.79; 38.21]

Ejercicio 4.7

IC(90 %)=[34.86; 38.14]

Ejercicio 4.8

$t_{9;0,975} = 2{,}262$
IC(95 %)=[16.01; 22.45]

Ejercicio 4.9

IC(99 %)=[1 %; 15 %]

Ejercicio 4.10

$t_{8;0,975} = 2{,}306$
IC(95 %)=[5.46; 8.54]

Ejercicio 4.11

$z_{0,995} = 2{,}58$
IC(99 %)=[121.13; 128.87]

Ejercicio 4.12

IC(90 %)=[82.50 %; 91.50 %]

Ejercicio 4.13

IC(95 %)=[27.1 %; 42.9 %]

Ejercicio 4.14

$n \approx 121$

Ejercicio 4.15

1.$n \approx 994$
2.$n \approx 2196$

Ejercicio 4.16

a) $\overline{x} = 28{,}01; s = 2{,}48; Me = 28{,}3; P_{25} = 26{,}65; P_{75} = 30{,}53$
b) IC(99 %)=[26.02, 30.01]
c) n $\approx$ 164

Ejercicio 4.17

b)$z_{0,925} = 1{,}44$
IC(85 %)=[5.89 %; 9.31 %]
c) p$\sim$N(8,1.5)
P(5.3$\leq$p$\leq$9.7)=0.8347

Ejercicio 4.18

a)$P_{33} = 9{,}3$
b)$t_{8;0,9} = 1{,}396$
IC(80 %)=[9.24, 11.42]

Ejercicio 4.19

a)$P_{40} = 14$
b)29 es atípico porque $\notin$ [1.38, 28.38]
c)$t_{11;0,9} = 1{,}363$
IC(80 %)=[13.59, 17.74]

Ejercicio 4.20

a)$z_{0,99} = 2{,}33$
IC(98 %)=[19.69 %; 27.91 %]
b) n $\approx$ 1576

Ejercicio 4.21

b)$z_{0,98} = 2{,}05$
IC(96 %)=[20.42 %; 30.01 %]

Ejercicio 4.22

b)$t_{10;0,99} = 2{,}763$
IC(98 %)=[44.72, 48.41]
d)$P_{77} = 48{,}82$

A.5. Soluciones numéricas Ejercicios Capítulo 5

Ejercicio 5.1

a) pivote= 2.3 $\notin$ [-2.228, 2.228] $\Rightarrow$ Se rechaza H_0
b) p-valor$\approx$ 0.04 $<$ 0.05 $\Rightarrow$ Se rechaza H_0
c) 65 $\notin$ [65.13, 73.59] $\Rightarrow$ Se rechaza H_0

Ejercicio 5.2

a) pivote= 0.66 $\in$ [-1.96, 1.96] $\Rightarrow$ No se rechaza H_0
b) p-valor= 0.5094 $>$ 0.05 $\Rightarrow$ No se rechaza H_0
c) 3 $\in$ [0.56, 7.44] $\Rightarrow$ No se rechaza H_0

Ejercicio 5.3

a) pivote= -1.79 $\notin$ [-1.761, $+\infty$) $\Rightarrow$ Se rechaza H_0
p-valor$\approx$ 0.04 $<$ 0.05 $\Rightarrow$ Se rechaza H_0
b) pivote= -1.79 $\notin$ [-1.65, $+\infty$) $\Rightarrow$ Se rechaza H_0
p-valor= 0.0368 $<$ 0.05 $\Rightarrow$ Se rechaza H_0

Ejercicio 5.4

a) pivote= -1.6 $\in$ [-2.326, $+\infty$) $\Rightarrow$ No se rechaza H_0
b) p-valor$\approx$ 0.075 $>$ 0.01 $\Rightarrow$ No se rechaza H_0

Ejercicio 5.5

a) pivote= 3.8 $\notin$ ($-\infty$, 1.64] $\Rightarrow$ Se rechaza H_0
p-valor$\approx$ 0.0001 $<$ 0.05 $\Rightarrow$ Se rechaza H_0

Ejercicio 5.6

a) pivote= 1.65 $\in$ [-1.96, 1.96] $\Rightarrow$ No se rechaza H_0
b) p-valor= 0.1 $>$ 0.05 $\Rightarrow$ No se rechaza H_0

Ejercicio 5.7

a) pivote= -2.33 $\notin$ [-1.64, $+\infty$) $\Rightarrow$ Se rechaza H_0
b) p-valor= 0.01 $<$ 0.05 $\Rightarrow$ Se rechaza H_0

Ejercicio 5.8

a) pivote= 2.98 $\notin$ ($-\infty$, 1.782] $\Rightarrow$ Se rechaza H_0
p-valor$\approx$ 0.0075$<$0.05 $\Rightarrow$ Se rechaza H_0
b) Me= 50.9; RIC=7.7

Ejercicio 5.9

b) pivote= 1.50; p-valor$\approx$ 0.075 $>$ 0.05 $\Rightarrow$ No se rechaza H_0
$P_{65} = 1{,}37$

Ejercicio 5.10

pivote= 2.23 $\in$ ($-\infty$, 2.33] $\Rightarrow$ No se rechaza H_0

Ejercicio 5.11

a) pivote= -2.76 $\notin$ [-1.894, $+\infty$) $\Rightarrow$ Se rechaza H_0

A.6. Soluciones numéricas Ejercicios Capítulo 6

Ejercicio 6.1

pivote= -3.22 $\notin$ [-2.33,∞) $\Rightarrow$ Se rechaza H_0
p-valor$\approx$ 0.00001<0.01 $\Rightarrow$ Se rechaza H_0

Ejercicio 6.2

pivote= -2.77 $\notin$ [-1.96,1.96] $\Rightarrow$ Se rechaza H_0
p-valor= 0.0058<0.05 $\Rightarrow$ Se rechaza H_0

Ejercicio 6.3

0 $\in$ IC(99 %)=[-9.28,19.40]$\Rightarrow$ No se rechaza H_0

Ejercicio 6.4

pivote= 0.91 $\in$ (-∞,2.33] $\Rightarrow$ No se rechaza H_0
p-valor= 0.1815>0.01 $\Rightarrow$ No se rechaza H_0

Ejercicio 6.5

0 $\in$ IC(99 %)=[-13.53,5.53]$\Rightarrow$ No se rechaza H_0

Ejercicio 6.6

0 $\notin$ IC(90 %)=[-40.16, -26.98]$\Rightarrow$ Se rechaza H_0

Ejercicio 6.7

pivote= 1.64 $\in$ [-1.96,1.96] $\Rightarrow$ No se rechaza H_0
p-valor= 0.1>0.005 $\Rightarrow$ No se rechaza H_0

Ejercicio 6.8

pivote= 1.64 $\in$ (-∞,1.65] $\Rightarrow$ No se rechaza H_0
p-valor= 0.05$\geq$0.05 $\Rightarrow$ No se rechaza H_0

Ejercicio 6.9

pivote= 1.08 $\in$ [0.17, 7.21] $\Rightarrow$ No se rechaza H_0

Ejercicio 6.10

pivote=0.735 $\in$ [0.2484, 4.0260] $\Rightarrow$ No se rechaza H_0

Ejercicio 6.11

pivote=0.69 $\in$ [0.28,3.0602] $\Rightarrow$ No se rechaza H_0

Ejercicio 6.12

Requisito sobre la igualdad de varianzas comprobado en el ejercicio 6.9
pivote=4.85 $\notin$ (-∞,2.552] $\Rightarrow$ Se rechaza H_0

Ejercicio 6.13

a)Requisito sobre la igualdad de varianzas poblacionales comprobado en el ejercicio 6.10
pivote=-6.17 $\notin$ [-1.734,$+\infty$) $\Rightarrow$ Se rechaza H_0
b) No necesita comprobar el requisito sobre la igualdad de varianzas poblacionales.
pivote=-6.71 $\notin$ [-1.65,$+\infty$) $\Rightarrow$ Se rechaza H_0

Ejercicio 6.14

No necesita comprobar el requisito sobre la igualdad de varianzas poblacionales.
pivote=2.8 $\notin$ (-∞,1.795] $\Rightarrow$ Se rechaza H_0
p-valor$\approx$ 0.008 $<$ 0.05 $\Rightarrow$ Se rechaza H_0

Ejercicio 6.15

Requisito sobre la igualdad de varianzas poblacionales: pivote= 0.69 $\in$ [0.347, 2.6710] $\Rightarrow$ No se rechaza H_0.
pivote=2.22 $\notin$ [-1.713,1.713] $\Rightarrow$ Se rechaza H_0
p-valor$\approx$ 0.04 $<$ 0.1 $\Rightarrow$ Se rechaza H_0

Ejercicio 6.16

IC(90 %)=[1.34,8.66]$\Rightarrow$ $\mu_1 - \mu_2 > 0 \Rightarrow \mu_1 > \mu_2$

Ejercicio 6.17

No necesita comprobar el requisito sobre la igualdad de varianzas poblacionales.
pivote=2.49 $\notin$ (-∞,1.812] $\Rightarrow$ Se rechaza H_0
p-valor$\approx$ 0.02 $<$ 0.05 $\Rightarrow$ Se rechaza H_0

Ejercicio 6.18

Requisito sobre la igualdad de varianzas poblacionales: pivote= 0.76 $\in$ [0.339, 2.9493] $\Rightarrow$ No se rechaza H_0.
pivote=1.40 $\in$ (-∞,1.701] $\Rightarrow$ No se rechaza H_0

Ejercicio 6.19

No necesita comprobar el requisito sobre la igualdad de varianzas poblacionales.
pivote=1.62 $\in$ (-∞,1.833] $\Rightarrow$ No se rechaza H_0
p-valor$\approx$ 0.08 $>$ 0.05 $\Rightarrow$ No se rechaza H_0

Ejercicio 6.20

Requisito sobre la igualdad de varianzas poblacionales: pivote= 0.56 $\in$ [0.18, 5.33] $\Rightarrow$ No se rechaza H_0.
IC(99 %)=[1.27,5.33]$\Rightarrow$ $\mu_1 - \mu_2 > 0 \Rightarrow \mu_1 > \mu_2$

Ejercicio 6.21

Requisito sobre la igualdad de varianzas poblacionales: pivote= 1.32 $\in$ [0.39, 2.7559] $\Rightarrow$ No se rechaza H_0.
1) IC(95 %)=[-0.1513, 2.2513]$\Rightarrow$ $\mu_1 - \mu_2 \approx 0 \Rightarrow \mu_1 \approx \mu_2$
2) pivote=1.71 $\in$ [-1.96,1.96] $\Rightarrow$ No se rechaza H_0
p-valor$\approx$ 0.0874 $>$ 0.05 $\Rightarrow$ No se rechaza H_0

Ejercicio 6.22

Requisito sobre la igualdad de varianzas poblacionales: pivote= 1.40 $\in$ [0.2438, 4.3572] $\Rightarrow$ No se rechaza H_0.

pivote=3.6

p-valor$\approx$ 0.00002 $<$ 0.05 $\Rightarrow$ Se rechaza H_0

Ejercicio 6.23

IC(98 %)=[-10.336, 4.62]$\Rightarrow$ $p_1 - p_2 \approx 0 \Rightarrow p1 \approx p_2$

Ejercicio 6.24

Requisito sobre la igualdad de varianzas poblacionales: pivote= 1.18 $\in$ [0.2646, 3.9639] $\Rightarrow$ No se rechaza H_0.

pivote=-1.46 $\in$ [-1.72, ∞) $\Rightarrow$ No se rechaza H_0

Ejercicio 6.25

pivote=2.94 $\notin$ (-∞, 1.28] $\Rightarrow$ Se rechaza H_0

p-valor= 0.0017$<$0.1 $\Rightarrow$ Se rechaza H_0

Ejercicio 6.26

Requisito sobre la igualdad de varianzas poblacionales: pivote= 3.67 $\in$ [0.2911, 3.8682] $\Rightarrow$ No se rechaza H_0.

pivote=-8.47 $\notin$ [-1.72, ∞) $\Rightarrow$ Se rechaza H_0

A.7. Soluciones numéricas Ejercicios Capítulo 7

Ejercicio 7.1

Pivote ANOVA: $(F) = 35{,}3\ F_{(2,72)}$
Región de rechazo $\approx (3{,}0873, +\infty)$

Ejercicio 7.2

Pivote ANOVA: $(F) = 0{,}515\ F_{(2,97)}$
Región de rechazo $\approx (3{,}0873, +\infty)$

Ejercicio 7.3

Pivote ANOVA: $(F) = 67{,}5\ F_{(2,12)}$
Región de rechazo $= (3{,}8853, +\infty)$

Ejercicio 7.4

Pivote ANOVA: $(F) = 80{,}13\ F_{(2,55)}$
Si considero $\alpha = 0{,}05$:
Región de rechazo $\approx (3{,}19, +\infty)$

Ejercicio 7.5

Pivote ANOVA: $(F) = 6{,}577\ F_{(2,30)}$
Si considero $\alpha = 0{,}05$:
Región de rechazo $= (3{,}3158, +\infty)$
$p - valor < \alpha$

Ejercicio 7.6

Pivote ANOVA: $(F) = 149{,}3\ F_{(2,40)}$
Si considero $\alpha = 0{,}01$:
Región de rechazo $= (5{,}178, +\infty)$
$p - valor < \alpha$

Ejercicio 7.7

Pivote ANOVA: $(F) = 44{,}5\ F_{(2,57)}$
Región de rechazo $\approx (4{,}977, +\infty)$

Ejercicio 7.8

a) H_0: $\sigma_A^2 = \sigma_B^2 = \sigma_C^2$
H_1: $\sigma_i^2 \neq \sigma_j^2$ para algún i, j.
Conclusión: como el p-valor $> \alpha$, no se rechaza H_0 y se cumple la hipótesis de homocedasticidad.

b) Conclusión: como todos los p-valores son $> \alpha$, no se rechaza H_0 y se cumple la hipótesis de normalidad.

c) Sí, porque la hipótesis de aleatoriedad se cumple por la asignación que se especifica en el enunciado y además se cumplen las hipótesis de homocedasticidad y normalidad por los dos apartados anteriores.

d) H_0: $\mu_A = \mu_B = \mu_C$
H_1: $\mu_i \neq \mu_j$ para algún i, j.
Distribución: $F_{2,40}$. Región de aceptación: [0, 5.178]. Región de rechazo: (5.178, ∞) Conclusión: como 9.001

$\in (5.178, \infty)$, se rechaza H_0, por lo que al menos dos de los tratamientos tienen tiempos medios de eficacia significativamente diferentes.

e) No, sólo sabemos que al menos dos de los tratamientos tienen tiempos medios de eficacia significativamente diferentes pero no cuáles. Esta aclaración la facilitan las comparaciones múltiples de Tukey.

f) $\mu_A < \mu_B$, $\mu_A = \mu_C$, $\mu_B = \mu_C$. Conclusión: El tiempo medio de eficacia del tratamiento B es significativamente superior al tiempo medio de eficacia del tratamiento A. El tiempo medio de eficacia del tratamiento C no tiene diferencias significativas con el tratamiento A y tampoco con el B.

A.8. Soluciones numéricas Ejercicios Capítulo 8

Ejercicio 8.1

Pivote:61.36 (p-valor aprox.:0.000)
Región de rechazo:$(5{,}99, +\infty)$

Ejercicio 8.2

Pivote:0.765 (p-valor aprox.:0.86)
Región de rechazo:$(7{,}81, +\infty)$

Ejercicio 8.3

Pivote:0.826 (p-valor aprox.:0.67)
Región de rechazo:$(5{,}99, +\infty)$

Ejercicio 8.4

Pivote:50.32 (p-valor aprox.:0.000)
Región de rechazo:$(9{,}21, +\infty)$

Ejercicio 8.5

Pivote:20.61 (p-valor aprox.:0.000)
Región de rechazo:$(7{,}78, +\infty)$

Ejercicio 8.6

b) Pivote= $3.55 \sim \chi^2_2 \in [0, 4{,}60517] \Rightarrow$ No se rechaza H_0
p-valor $\approx 0{,}15 > 0{,}1 \Rightarrow$ No se rechaza H_0

Ejercicio 8.7

a) Pivote= $7.08 \sim \chi^2_2 \in [9{,}21, \infty) \Rightarrow$ No se rechaza H_0
b) p-valor $\approx 0{,}03 > 0{,}01 \Rightarrow$ No se rechaza H_0

Ejercicio 8.8

a) Pivote= $6.52 \sim \chi^2_2 \in [5{,}99146, \infty) \Rightarrow$ Se rechaza H_0
b) p-valor $\approx 0{,}0375 < 0{,}05 \Rightarrow$ Se rechaza H_0

A.9. Soluciones numéricas Ejercicios Capítulo 9

Ejercicio 9.1

(No es necesario realizar ningún cálculo)

Ejercicio 9.2

- $(xx) = 0{,}0128$
- $(yy) = 0{,}0058$
- $(xy) = 0{,}0078$
- $n = 10$
- $\bar{x} = 1{,}73$
- $\bar{y}1{,}75$
- $r = 0{,}91$
- $r^2\,\% = 82{,}82\,\%$
- Recta de regresión estimada: $Y = a + b \cdot X$
  - $a = 0{,}70$
  - $b = 0{,}61$
- $s^2 = 0{,}0001$; $s = 0{,}01$
- Test de independencia con el coeficiente de correlación lineal:
  - $Pivote \approx 6{,}90$
  - Región de aceptación:$(-2{,}306, 2{,}306)$
  - p-valor=0,000
- Test de independencia con el coeficiente de regresión:
  - $Pivote = 6{,}03$
  - Región de aceptación:$(-2{,}306, 2{,}306)$
  - p-valor=0,000
- Intervalo de confianza para B: $(0{,}38, 0{,}84)$

Ejercicio 9.3

- $(xx) = 3285{,}714$
- $(yy) = 1571{,}429$
- $(xy) = 1957{,}143$
- $n = 7$
- $\bar{x} = 88{,}57$
- $\bar{y} = 90{,}71$
- $r = 0{,}86$
- $r^2\,\% = 74{,}74\,\%$
- Recta de regresión estimada: $Y = a + b \cdot X$
  - $a = 37{,}95$
  - $b = 0{,}60$

- $s^2 = 81{,}13$; $s = 9{,}01$
- Test de independencia con el coeficiente de correlación lineal:
  - $Pivote = 3{,}79$
  - Región de aceptación:$(-2{,}57, 2{,}57)$
  - p-valor$< \alpha$
- Test de independencia con el coeficiente de regresión:
  - $Pivote = 3{,}82$
  - Región de aceptación:$(-2{,}57, 2{,}57)$
  - p-valor$< \alpha$
- Intervalo de confianza para B: $(0{,}19, 1{,}00)$

Ejercicio 9.4

- $(xx) = 2445{,}667$
- $(yy) = 29{,}58917$
- $(xy) = 196{,}1833$
- $n = 12$
- $\bar{x} = 46{,}83$
- $\bar{y} = 4{,}89$
- $r = 0{,}73$
- $r^2 = 0{,}53$ 53 %
- Recta de regresión estimada: $Y = a + b \cdot X$
  - $a = 1{,}13$
  - $b = 0{,}08$
- $s^2 = 1{,}3852$; $s = 1{,}18$
- Test de independencia con el coeficiente de correlación lineal:
  - $Pivote = 3{,}37$
  - Región de aceptación:$(-3{,}169, 3{,}169)$
  - p-valor$< \alpha$
- Test de independencia con el coeficiente de regresión:
  - $Pivote = 3{,}36$
  - Región de aceptación:$(-3{,}169, 3{,}169)$
  - p-valor$< \alpha$
- Intervalo de confianza para B: $(0{,}0048, 0{,}1556)$

Ejercicio 9.5

- $(xx) = 2146,769$
- $(yy) = 42024,19$
- $(xy) = -9222,199$
- $n = 10$
- $\bar{x} = 39,21$
- $\bar{y} = 110,79$
- $r = -0{,}97$
- $r^2 = 0{,}94$ 94 %
- Recta de regresión estimada: $Y = a + b \cdot X$
  - $a = 279{,}23$
  - $b = -4{,}30$
- $s^2 = 300{,}88$; $s = 17{,}35$
- Test de independencia con el coeficiente de correlación lineal:
  - $Pivote = 11{,}47$
  - Región de aceptación:$(-2{,}31, 2{,}31)$
  - p-valor$< \alpha$
- Test de independencia con el coeficiente de regresión:
  - $Pivote = 11{,}49$
  - Región de aceptación:$(-2{,}31, 2{,}31)$
  - p-valor$< \alpha$
- Intervalo de confianza para B: $(-5{,}16, -3{,}43)$

Ejercicio 9.6

- $(xx) = 3{,}24$
- $(yy) = 580$
- $(xy) = 39{,}5$
- $n = 5$
- $\bar{x} = 6{,}9$
- $\bar{y} = 137$
- $r = 0{,}91$
- $r^2 = 0{,}8303$ 83,03 %
- Recta de regresión estimada: $Y = a + b \cdot X$
  - $a = 52{,}88$
  - $b = 12{,}19$
- $s^2 = 32{,}8138$; $s = 5{,}73$
- Test de independencia con el coeficiente de correlación lineal:
  - $Pivote = 3{,}83$
  - Región de aceptación:$(-3{,}182, 3{,}182)$

    - p-valor$< \alpha$
- Test de independencia con el coeficiente de regresión:
    - $Pivote = 3{,}83$
    - Región de aceptación:$(-3{,}182, 3{,}182)$
    - p-valor$< \alpha$
- Intervalo de confianza para B: $(2{,}06, 22{,}32)$

Ejercicio 9.7

- $(xx) = 673{,}1$
- $(yy) = 5903{,}5$
- $(xy) = 1659{,}1$
- $n = 6$
- $\bar{x} = 77{,}97$
- $\bar{y} = 143{,}5$
- $r = 0{,}83$
- $r^2 = 0{,}69$ 69 %
- Recta de regresión estimada: $Y = a + b \cdot X$
    - $a = -48{,}69$
    - $b = 2{,}46$
- $s^2 = 453{,}51$; $s = 21{,}30$
- Test de independencia con el coeficiente de correlación lineal:
    - $Pivote = 3{,}00$
    - Región de aceptación:$(-2{,}131, 2{,}131)$
    - p-valor$< \alpha$
- Test de independencia con el coeficiente de regresión:
    - $Pivote = 3{,}00$
    - Región de aceptación:$(-2{,}131, 2{,}131)$
    - p-valor$< \alpha$
- Intervalo de confianza para B: $(0{,}72, 4{,}21)$

Ejercicio 9.8

- $(xx) = 636{,}83$
- $(yy) = 1485{,}33$
- $(xy) = 951{,}33$
- $n = 6$
- $\bar{x} = 60{,}83$
- $\bar{y} = 122{,}33$
- $r = 0{,}98$
- $r^2 = 0{,}96$ 96 %

- Recta de regresión estimada: $Y = a + b \cdot X$
  - $a = 31{,}46$
  - $b = 1{,}49$
- $s^2 = 16{,}0459$; $s = 4{,}01$
- Test de independencia con el coeficiente de correlación lineal:
  - $Pivote = 9{,}41$
  - Región de aceptación:$(-2{,}776, 2{,}776)$
  - p-valor$< \alpha$
- Test de independencia con el coeficiente de regresión:
  - $Pivote = 9{,}39$
  - Región de aceptación:$(-2{,}776, 2{,}776)$
  - p-valor$< \alpha$
- Intervalo de confianza para B: $(1{,}05, 1{,}93)$

Ejercicio 9.9

- $(xx) = 45{,}2$
- $(yy) = 946{,}8$
- $(xy) = 203{,}4$
- $n = 5$
- $\bar{x} = 4{,}6$
- $\bar{y} = 150{,}2$
- $r = 0{,}98$
- $r^2 = 0{,}97$ 97 %
- Recta de regresión estimada: $Y = a + b \cdot X$
  - $a = 129{,}5$
  - $b = 4{,}50$
- $s^2 = 10{,}5$; $s = 3{,}24$
- Test de independencia con el coeficiente de correlación lineal:
  - $Pivote = 9{,}34$
  - Región de aceptación:$(-5{,}84, 5{,}84)$
  - p-valor$< \alpha$
- Test de independencia con el coeficiente de regresión:
  - $Pivote = 9{,}34$
  - Región de aceptación:$(-5{,}84, 5{,}84)$
  - p-valor$< \alpha$
- Intervalo de confianza para B: $(1{,}69, 7{,}31)$

Apéndice B

Anexo II: FORMULARIO

Estadística descriptiva

Media $\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{x_1 + x_2 + ... + x_n}{n}$	**Desviación típica** $s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}$
Coeficiente de variación $CV = \frac{s}{\overline{x}} \cdot 100$	**Percentil al p %** $Pos = (n+1) \cdot \frac{p}{100}$ $deci(Pos) \cdot X_{[Pos]+1} + (1 - deci(Pos)) \cdot X_{[Pos]}$
Rango $Rango = Máximo - Mínimo$	**Rango intercuartílico** $R.I.C. = Q_3 - Q_1 = P_{75} - P_{25}$
Valores atípicos $[Q_1 - 1{,}5 \cdot R.I.C., Q_3 + 1{,}5 \cdot R.I.C.]$	**Aritmética de variables normales** $X \sim N(\mu, \sigma) \Rightarrow Z = \frac{X-\mu}{\sigma} \sim N(0,1)$ $Z \sim N(0,1) \Rightarrow X = \sigma \cdot Z + \mu \sim N(\mu, \sigma)$

Intervalos de confianza

Una media

a) $\left[\overline{x} - Z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \ , \ \overline{x} + Z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right]$

b) $\left[\overline{x} - t_{(n-1,1-\frac{\alpha}{2})} \cdot \frac{S}{\sqrt{n}} \ , \ \overline{x} + t_{(n-1,1-\frac{\alpha}{2})} \cdot \frac{S}{\sqrt{n}}\right]$

Un porcentaje

$$\left[\widehat{P} - Z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\widehat{P} \cdot (100 - \widehat{P})}{n}} \ , \ \widehat{P} + Z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\widehat{P} \cdot (100 - \widehat{P})}{n}}\right]$$

Dos medias

a)Varianzas poblacionales conocidas e iguales:

$$\left[\ (\overline{x}_1 - \overline{x}_2) \pm Z_{1-\frac{\alpha}{2}} \cdot \sigma \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \ \right]$$

b) Varianzas poblacionales desconocidas.

b.1) Pueden ser asumidas iguales:

$$\left[\ (\overline{x}_1 - \overline{x}_2) \pm t_{\left((n_1+n_2-2) \ (1-\frac{\alpha}{2})\right)} \cdot S \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \ \right] \text{ donde } S = \sqrt{\frac{(n_1 - 1) \cdot S_1^2 + (n_2 - 1) \cdot S_2^2}{(n_1 + n_2 - 2)}}$$

b.2) No pueden ser asumidas iguales:

$$\left[\ (\overline{x}_1 - \overline{x}_2) \pm t_{\left((gl) \ (1-\frac{\alpha}{2})\right)} \cdot \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \ \right] \text{ donde } gl = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1}}$$

Dos porcentajes

$$\left[\ (\widehat{P}_1 - \widehat{P}_2) \pm Z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\widehat{P}_1 \cdot (100 - \widehat{P}_1)}{n_1} + \frac{\widehat{P}_2 \cdot (100 - \widehat{P}_2)}{n_2}} \ \right]$$

Tamaño muestral

Una media	**Un porcentaje**
$n \geq \frac{4 \cdot (Z_{1-\frac{\alpha}{2}})^2 \sigma^2}{e^2}$	$n \geq \frac{4 \cdot (Z_{1-\frac{\alpha}{2}})^2 \cdot P \cdot (100 - P)}{e^2}$

Contrastes de hipótesis

Una media

a) $\dfrac{\overline{x}-\mu}{\dfrac{\sigma}{\sqrt{n}}} \sim N(0,1)$; b) $\dfrac{\overline{x}-\mu}{\dfrac{S}{\sqrt{n}}} \sim t_{n-1}$

Un porcentaje

$$\frac{\widehat{P}-P}{\sqrt{\dfrac{P\cdot(100-P)}{n}}} \sim N(0,1)$$

Dos porcentajes

$$\frac{(\widehat{P}_1-\widehat{P}_2)-(P_1-P_2)}{\sqrt{\dfrac{\widehat{P}_1\cdot(100-\widehat{P}_1)}{n_1}+\dfrac{\widehat{P}_2\cdot(100-\widehat{P}_2)}{n_2}}} \sim N(0,1)$$

Dos varianzas

$\left(\dfrac{S_1^2}{S_2^2}\right)\cdot\left(\dfrac{\sigma_2^2}{\sigma_1^2}\right) \sim F_{(n_1-1,n_2-1)}$ Recordatorio: $F_{(m,n),\gamma}=\dfrac{1}{F_{(n,m),1-\gamma}}$

Dos medias

a) Varianzas poblacionales conocidas e iguales.

$$\frac{(\overline{x}_1-\overline{x}_2)-(\mu_1-\mu_2)}{\sigma\cdot\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}} \sim N(0,1)$$

b) Varianzas poblacionales desconocidas.

b.1) Pueden ser asumidas iguales:

$$\frac{(\overline{x}_1-\overline{x}_2)-(\mu_1-\mu_2)}{S\cdot\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}} \sim t_{n_1+n_2-2} \text{ donde } S=\sqrt{\frac{(n_1-1)\cdot S_1^2+(n_2-1)\cdot S_2^2}{(n_1+n_2-2)}}$$

b.2) No pueden ser asumidas iguales:

$$\frac{(\overline{x}_1-\overline{x}_2)-(\mu_1-\mu_2)}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}} \sim t_{gl} \text{ donde } gl=\frac{\left(\dfrac{S_1^2}{n_1}+\dfrac{S_2^2}{n_2}\right)^2}{\dfrac{\left(\frac{S_1^2}{n1}\right)^2}{n_1-1}+\dfrac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2-1}}$$

ANOVA

$pivote \sim F_{(k-1,n-k)}$

Chi-cuadrado

$$\sum_{i,j}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} \sim \chi_g^2$$

donde $g=(n_f-1)\cdot(n_c-1)$ Recordatorio: $E_{ij}=\dfrac{TF_i\cdot TC_j}{N}$

Regresión lineal

Sumas de cuadrados

$$(xx) = \sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}; \quad (yy) = \sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n}; \quad (xy) = \sum_{i=1}^{n} x_i \cdot y_i - \frac{(\sum_{i=1}^{n} x_i) \cdot (\sum_{i=1}^{n} y_i)}{n}$$

Coeficiente de correlación muestral

$$r = \frac{(xy)}{\sqrt{(xx)(yy)}}$$

Contraste de independencia para el coeficiente de correlación lineal

$$Pivote = \sqrt{\frac{(n-2)r^2}{1-r^2}} \sim t_{n-2}$$

Ecuación de la recta de regresión estimada

$$\widehat{Y} = a + b \cdot X; \quad b = \frac{(xy)}{(xx)}; \quad a = \overline{y} - b \cdot \overline{x};$$

Varianza y desviación típica residual

$$s^2 = \frac{(yy) - \frac{(xy)^2}{(xx)}}{n-2}; \quad s = \sqrt{\frac{(yy) - \frac{(xy)^2}{(xx)}}{n-2}}$$

Contraste de independencia para el coeficiente de regresión (pendiente)

$$Pivote = \frac{b - B}{\sqrt{\frac{s^2}{(xx)}}} \sim t_{n-2}$$

Intervalo de confianza para el coeficiente de regresión (pendiente)

$$\left(b - t_{(1-\frac{\alpha}{2}, n-2)} \sqrt{\frac{s^2}{(xx)}} \quad , \quad b + t_{(1-\frac{\alpha}{2}, n-2)} \sqrt{\frac{s^2}{(xx)}} \right)$$

Apéndice C

Anexo III: Tablas estadísticas

Tabla de probabilidades de la distribución N(0,1)
$[P(Z < z)]$

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5039	0.5079	0.5119	0.5159	0.5199	0.5239	0.5279	0.5318	0.5358
0.1	0.5398	0.5437	0.5477	0.5517	0.5556	0.5596	0.5635	0.5674	0.5714	0.5753
0.2	0.5792	0.5831	0.5870	0.5909	0.5948	0.5987	0.6025	0.6064	0.6102	0.6140
0.3	0.6179	0.6217	0.6255	0.6293	0.6330	0.6368	0.6405	0.6443	0.6480	0.6517
0.4	0.6554	0.6590	0.6627	0.6664	0.6700	0.6736	0.6772	0.6808	0.6843	0.6879
0.5	0.6914	0.6949	0.6984	0.7019	0.7054	0.7088	0.7122	0.7156	0.7190	0.7224
0.6	0.7257	0.7290	0.7323	0.7356	0.7389	0.7421	0.7453	0.7485	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7703	0.7733	0.7763	0.7793	0.7823	0.7852
0.8	0.7881	0.7910	0.7938	0.7967	0.7995	0.8023	0.8051	0.8078	0.8105	0.8132
0.9	0.8159	0.8185	0.8212	0.8238	0.8263	0.8289	0.8314	0.8339	0.8364	0.8389
1.0	0.8413	0.8437	0.8461	0.8484	0.8508	0.8531	0.8554	0.8576	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8707	0.8728	0.8749	0.8769	0.8790	0.8810	0.8829
1.2	0.8849	0.8868	0.8887	0.8906	0.8925	0.8943	0.8961	0.8979	0.8997	0.9014
1.3	0.9032	0.9049	0.9065	0.9082	0.9098	0.9114	0.9130	0.9146	0.9162	0.9177
1.4	0.9192	0.9207	0.9221	0.9236	0.9250	0.9264	0.9278	0.9292	0.9305	0.9318
1.5	0.9331	0.9344	0.9357	0.9369	0.9382	0.9394	0.9406	0.9417	0.9429	0.9440
1.6	0.9452	0.9463	0.9473	0.9484	0.9494	0.9505	0.9515	0.9525	0.9535	0.9544
1.7	0.9554	0.9563	0.9572	0.9581	0.9590	0.9599	0.9607	0.9616	0.9624	0.9632
1.8	0.9640	0.9648	0.9656	0.9663	0.9671	0.9678	0.9685	0.9692	0.9699	0.9706
1.9	0.9712	0.9719	0.9725	0.9731	0.9738	0.9744	0.9750	0.9755	0.9761	0.9767
2.0	0.9772	0.9777	0.9783	0.9788	0.9793	0.9798	0.9803	0.9807	0.9812	0.9816
2.1	0.9821	0.9825	0.9829	0.9834	0.9838	0.9842	0.9846	0.9849	0.9853	0.9857
2.2	0.9860	0.9864	0.9867	0.9871	0.9874	0.9877	0.9880	0.9883	0.9886	0.9889
2.3	0.9892	0.9895	0.9898	0.9900	0.9903	0.9906	0.9908	0.9911	0.9913	0.9915
2.4	0.9918	0.9920	0.9922	0.9924	0.9926	0.9928	0.9930	0.9932	0.9934	0.9936
2.5	0.9937	0.9939	0.9941	0.9942	0.9944	0.9946	0.9947	0.9949	0.9950	0.9952
2.6	0.9953	0.9954	0.9956	0.9957	0.9958	0.9959	0.9960	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9971	0.9972	0.9973
2.8	0.9974	0.9975	0.9975	0.9976	0.9977	0.9978	0.9978	0.9979	0.9980	0.9980
2.9	0.9981	0.9981	0.9982	0.9983	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986
3.0	0.9986	0.9986	0.9987	0.9987	0.9988	0.9988	0.9988	0.9989	0.9989	0.9989

Tabla de probabilidades de la distribución *t de Student*
$[P(t < T)]$

$g \diagdown T$	0.650	0.700	0.750	0.800	0.850	0.900	0.950	0.9750	0.990	0.995
1	0.509	0.726	1.000	1.376	1.962	3.077	6.313	12.706	31.820	63.656
2	0.444	0.617	0.816	1.060	1.386	1.885	2.919	4.302	6.964	9.924
3	0.424	0.584	0.764	0.978	1.249	1.637	2.353	3.182	4.540	5.840
4	0.414	0.568	0.740	0.940	1.189	1.533	2.131	2.776	3.746	4.604
5	0.408	0.559	0.726	0.919	1.155	1.475	2.015	2.570	3.364	4.032
6	0.404	0.553	0.717	0.905	1.134	1.439	1.943	2.446	3.142	3.707
7	0.401	0.549	0.711	0.896	1.119	1.414	1.894	2.364	2.997	3.499
8	0.399	0.545	0.706	0.888	1.108	1.396	1.859	2.306	2.896	3.355
9	0.397	0.543	0.702	0.883	1.099	1.383	1.833	2.262	2.821	3.249
10	0.396	0.541	0.699	0.879	1.093	1.372	1.812	2.228	2.763	3.169
11	0.395	0.539	0.697	0.875	1.087	1.363	1.795	2.200	2.718	3.105
12	0.394	0.538	0.695	0.872	1.083	1.356	1.782	2.178	2.680	3.054
13	0.393	0.537	0.693	0.870	1.079	1.350	1.770	2.160	2.650	3.012
14	0.393	0.536	0.692	0.868	1.076	1.345	1.761	2.144	2.624	2.976
15	0.392	0.535	0.691	0.866	1.073	1.340	1.753	2.131	2.602	2.946
16	0.392	0.535	0.690	0.864	1.071	1.336	1.745	2.119	2.583	2.920
17	0.391	0.534	0.689	0.863	1.069	1.333	1.739	2.109	2.566	2.898
18	0.391	0.533	0.688	0.862	1.067	1.330	1.734	2.100	2.552	2.878
19	0.391	0.533	0.687	0.860	1.065	1.327	1.729	2.093	2.539	2.860
20	0.390	0.532	0.686	0.859	1.064	1.325	1.724	2.085	2.527	2.845
21	0.390	0.532	0.686	0.859	1.062	1.323	1.720	2.079	2.517	2.831
22	0.390	0.532	0.685	0.858	1.061	1.321	1.717	2.073	2.508	2.818
23	0.390	0.531	0.685	0.857	1.060	1.319	1.713	2.068	2.499	2.807
24	0.389	0.531	0.684	0.856	1.059	1.317	1.710	2.063	2.492	2.796
25	0.389	0.531	0.684	0.856	1.058	1.316	1.708	2.059	2.485	2.787
26	0.389	0.530	0.684	0.855	1.057	1.314	1.705	2.055	2.478	2.778
27	0.389	0.530	0.683	0.855	1.056	1.313	1.703	2.051	2.472	2.770
28	0.389	0.530	0.683	0.854	1.055	1.312	1.701	2.048	2.467	2.763
29	0.389	0.530	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756
30	0.389	0.530	0.682	0.853	1.054	1.310	1.697	2.042	2.457	2.749
40	0.388	0.528	0.680	0.850	1.050	1.303	1.683	2.021	2.423	2.704
60	0.387	0.527	0.678	0.847	1.045	1.295	1.670	2.000	2.390	2.660
120	0.386	0.525	0.676	0.844	1.040	1.288	1.657	1.979	2.357	2.617
∞	0.385	0.524	0.674	0.841	1.036	1.281	1.644	1.959	2.326	2.575

Inversa de la función de distribución F de Snedecor [0.900]:

$$\mathcal{F}_{m,n}(x) = P[F_{m,n} \leq x], \quad \mathcal{F}_{m,n}(F_{m,n,\alpha}) = \alpha, \quad F_{m,n,\alpha} = \mathcal{F}^{-1}_{m,n}(\alpha), \quad \alpha = 0{,}900$$

$n \diagdown m$	1	2	3	4	5	6	7	8	9	10	12	15	20	50	100	∞
1	39.8635	49.5000	53.5932	55.8330	57.2401	58.2044	58.9060	59.4390	59.8576	60.1950	60.7052	61.2203	61.7403	62.6881	63.0073	63.3281
2	8.5263	9.0000	9.1618	9.2434	9.2926	9.3255	9.3491	9.3668	9.3805	9.3916	9.4081	9.4247	9.4413	9.4712	9.4812	9.4912
3	5.5383	5.4624	5.3908	5.3426	5.3092	5.2847	5.2662	5.2517	5.2400	5.2304	5.2156	5.2003	5.1845	5.1546	5.1443	5.1337
4	4.5448	4.3246	4.1909	4.1072	4.0506	4.0097	3.9790	3.9549	3.9357	3.9199	3.8955	3.8704	3.8443	3.7952	3.7782	3.7607
5	4.0604	3.7797	3.6195	3.5202	3.4530	3.4045	3.3679	3.3393	3.3163	3.2974	3.2682	3.2380	3.2066	3.1471	3.1263	3.1050
6	3.7759	3.4633	3.2888	3.1808	3.1075	3.0546	3.0145	2.9830	2.9577	2.9369	2.9047	2.8712	2.8363	2.7697	2.7463	2.7222
7	3.5894	3.2574	3.0741	2.9605	2.8833	2.8274	2.7849	2.7516	2.7247	2.7025	2.6681	2.6322	2.5947	2.5226	2.4971	2.4708
8	3.4579	3.1131	2.9238	2.8064	2.7264	2.6683	2.6241	2.5893	2.5612	2.5380	2.5020	2.4642	2.4246	2.3481	2.3208	2.2926
9	3.3603	3.0065	2.8129	2.6927	2.6106	2.5509	2.5053	2.4694	2.4403	2.4163	2.3789	2.3396	2.2983	2.2180	2.1892	2.1592
10	3.2850	2.9245	2.7277	2.6053	2.5216	2.4606	2.4140	2.3772	2.3473	2.3226	2.2841	2.2435	2.2007	2.1171	2.0869	2.0554
11	3.2252	2.8595	2.6602	2.5362	2.4512	2.3891	2.3416	2.3040	2.2735	2.2482	2.2087	2.1671	2.1230	2.0364	2.0050	1.9721
12	3.1765	2.8068	2.6055	2.4801	2.3940	2.3310	2.2828	2.2446	2.2135	2.1878	2.1474	2.1048	2.0597	1.9704	1.9379	1.9036
13	3.1362	2.7632	2.5603	2.4337	2.3467	2.2830	2.2341	2.1953	2.1638	2.1376	2.0966	2.0532	2.0070	1.9153	1.8817	1.8462
14	3.1022	2.7265	2.5222	2.3947	2.3069	2.2426	2.1931	2.1539	2.1220	2.0954	2.0537	2.0095	1.9625	1.8686	1.8340	1.7973
15	3.0732	2.6952	2.4898	2.3614	2.2730	2.2081	2.1582	2.1185	2.0862	2.0593	2.0171	1.9722	1.9243	1.8284	1.7929	1.7551
16	3.0481	2.6682	2.4618	2.3327	2.2438	2.1783	2.1280	2.0880	2.0553	2.0281	1.9854	1.9399	1.8913	1.7934	1.7570	1.7182
17	3.0262	2.6446	2.4374	2.3077	2.2183	2.1524	2.1017	2.0613	2.0284	2.0009	1.9577	1.9117	1.8624	1.7628	1.7255	1.6856
18	3.0070	2.6239	2.4160	2.2858	2.1958	2.1296	2.0785	2.0379	2.0047	1.9770	1.9333	1.8868	1.8368	1.7356	1.6976	1.6567
19	2.9899	2.6056	2.3970	2.2663	2.1760	2.1094	2.0580	2.0171	1.9836	1.9557	1.9117	1.8647	1.8142	1.7114	1.6726	1.6308
20	2.9747	2.5893	2.3801	2.2489	2.1582	2.0913	2.0397	1.9985	1.9649	1.9367	1.8924	1.8449	1.7938	1.6896	1.6501	1.6074
21	2.9610	2.5746	2.3649	2.2333	2.1423	2.0751	2.0233	1.9819	1.9480	1.9197	1.8750	1.8271	1.7756	1.6700	1.6298	1.5862
22	2.9486	2.5613	2.3512	2.2193	2.1279	2.0605	2.0084	1.9668	1.9327	1.9043	1.8593	1.8111	1.7590	1.6521	1.6113	1.5668
23	2.9374	2.5493	2.3387	2.2065	2.1149	2.0472	1.9949	1.9531	1.9189	1.8903	1.8450	1.7964	1.7439	1.6358	1.5944	1.5490
24	2.9271	2.5383	2.3274	2.1949	2.1030	2.0351	1.9826	1.9407	1.9063	1.8775	1.8319	1.7831	1.7302	1.6209	1.5788	1.5327
25	2.9177	2.5283	2.3170	2.1842	2.0922	2.0241	1.9714	1.9292	1.8947	1.8658	1.8200	1.7708	1.7175	1.6072	1.5645	1.5176
26	2.9091	2.5191	2.3075	2.1745	2.0822	2.0139	1.9610	1.9188	1.8841	1.8550	1.8090	1.7596	1.7059	1.5945	1.5513	1.5036
27	2.9012	2.5106	2.2987	2.1655	2.0730	2.0045	1.9515	1.9091	1.8743	1.8451	1.7989	1.7492	1.6951	1.5827	1.5390	1.4906
28	2.8938	2.5028	2.2906	2.1571	2.0645	1.9958	1.9427	1.9001	1.8652	1.8359	1.7895	1.7395	1.6852	1.5718	1.5276	1.4784
29	2.8870	2.4955	2.2831	2.1494	2.0566	1.9878	1.9345	1.8918	1.8568	1.8274	1.7808	1.7306	1.6759	1.5616	1.5169	1.4670
30	2.8807	2.4887	2.2761	2.1422	2.0492	1.9803	1.9269	1.8841	1.8490	1.8195	1.7727	1.7223	1.6673	1.5522	1.5069	1.4564
40	2.8354	2.4404	2.2261	2.0909	1.9968	1.9269	1.8725	1.8289	1.7929	1.7627	1.7146	1.6624	1.6052	1.4830	1.4336	1.3769
60	2.7911	2.3933	2.1774	2.0410	1.9457	1.8747	1.8194	1.7748	1.7380	1.7070	1.6574	1.6034	1.5435	1.4126	1.3576	1.2915
100	2.7564	2.3564	2.1394	2.0019	1.9057	1.8339	1.7778	1.7324	1.6949	1.6632	1.6124	1.5566	1.4943	1.3548	1.2934	1.2142
200	2.7308	2.3293	2.1114	1.9732	1.8763	1.8038	1.7470	1.7011	1.6630	1.6308	1.5789	1.5218	1.4575	1.3100	1.2418	1.1439
∞	2.7055	2.3026	2.0838	1.9449	1.8473	1.7741	1.7167	1.6702	1.6315	1.5987	1.5458	1.4871	1.4206	1.2633	1.1850	1.0000

Inversa de la función de distribución F de Snedecor [0.950]:

$$\mathcal{F}_{m,n}(x) = P[F_{m,n} \leq x], \quad \mathcal{F}_{m,n}(F_{m,n,\alpha}) = \alpha, \quad F_{m,n,\alpha} = \mathcal{F}^{-1}_{m,n}(\alpha), \quad \alpha = 0{,}950$$

$n \diagdown m$	1	2	3	4	5	6	7	8	9	10	15	20	50	100	∞
1	161.4476	199.5000	215.7073	224.5832	230.1619	233.9860	236.7684	238.8827	240.5433	241.8817	245.9499	248.0131	251.7742	253.0411	254.3144
2	18.5128	19.0000	19.1643	19.2468	19.2964	19.3295	19.3532	19.3710	19.3848	19.3959	19.4291	19.4458	19.4757	19.4857	19.4957
3	10.1280	9.5521	9.2766	9.1172	9.0135	8.9406	8.8867	8.8452	8.8123	8.7855	8.7029	8.6602	8.5810	8.5539	8.5264
4	7.7086	6.9443	6.5914	6.3882	6.2561	6.1631	6.0942	6.0410	5.9988	5.9644	5.8578	5.8025	5.6995	5.6641	5.6281
5	6.6079	5.7861	5.4095	5.1922	5.0503	4.9503	4.8759	4.8183	4.7725	4.7351	4.6188	4.5581	4.4444	4.4051	4.3650
6	5.9874	5.1433	4.7571	4.5337	4.3874	4.2839	4.2067	4.1468	4.0990	4.0600	3.9381	3.8742	3.7537	3.7117	3.6689
7	5.5914	4.7374	4.3468	4.1203	3.9715	3.8660	3.7870	3.7257	3.6767	3.6365	3.5107	3.4445	3.3189	3.2749	3.2298
8	5.3177	4.4590	4.0662	3.8379	3.6875	3.5806	3.5005	3.4381	3.3881	3.3472	3.2184	3.1503	3.0204	2.9747	2.9276
9	5.1174	4.2565	3.8625	3.6331	3.4817	3.3738	3.2927	3.2296	3.1789	3.1373	3.0061	2.9365	2.8028	2.7556	2.7067
10	4.9646	4.1028	3.7083	3.4780	3.3258	3.2172	3.1355	3.0717	3.0204	2.9782	2.8450	2.7740	2.6371	2.5884	2.5379
11	4.8443	3.9823	3.5874	3.3567	3.2039	3.0946	3.0123	2.9480	2.8962	2.8536	2.7186	2.6464	2.5066	2.4566	2.4045
12	4.7472	3.8853	3.4903	3.2592	3.1059	2.9961	2.9134	2.8486	2.7964	2.7534	2.6169	2.5436	2.4010	2.3498	2.2962
13	4.6672	3.8056	3.4105	3.1791	3.0254	2.9153	2.8321	2.7669	2.7144	2.6710	2.5331	2.4589	2.3138	2.2614	2.2064
14	4.6001	3.7389	3.3439	3.1122	2.9582	2.8477	2.7642	2.6987	2.6458	2.6022	2.4630	2.3879	2.2405	2.1870	2.1307
15	4.5431	3.6823	3.2874	3.0556	2.9013	2.7905	2.7066	2.6408	2.5876	2.5437	2.4034	2.3275	2.1780	2.1234	2.0658
16	4.4940	3.6337	3.2389	3.0069	2.8524	2.7413	2.6572	2.5911	2.5377	2.4935	2.3522	2.2756	2.1240	2.0685	2.0096
17	4.4513	3.5915	3.1968	2.9647	2.8100	2.6987	2.6143	2.5480	2.4943	2.4499	2.3077	2.2304	2.0769	2.0204	1.9604
18	4.4139	3.5546	3.1599	2.9277	2.7729	2.6613	2.5767	2.5102	2.4563	2.4117	2.2686	2.1906	2.0354	1.9780	1.9168
19	4.3807	3.5219	3.1274	2.8951	2.7401	2.6283	2.5435	2.4768	2.4227	2.3779	2.2341	2.1555	1.9986	1.9403	1.8780
20	4.3512	3.4928	3.0984	2.8661	2.7109	2.5990	2.5140	2.4471	2.3928	2.3479	2.2033	2.1242	1.9656	1.9066	1.8432
21	4.3248	3.4668	3.0725	2.8401	2.6848	2.5727	2.4876	2.4205	2.3660	2.3210	2.1757	2.0960	1.9360	1.8761	1.8117
22	4.3009	3.4434	3.0491	2.8167	2.6613	2.5491	2.4638	2.3965	2.3419	2.2967	2.1508	2.0707	1.9092	1.8486	1.7831
23	4.2793	3.4221	3.0280	2.7955	2.6400	2.5277	2.4422	2.3748	2.3201	2.2747	2.1282	2.0476	1.8848	1.8234	1.7570
24	4.2597	3.4028	3.0088	2.7763	2.6207	2.5082	2.4226	2.3551	2.3002	2.2547	2.1077	2.0267	1.8625	1.8005	1.7330
25	4.2417	3.3852	2.9912	2.7587	2.6030	2.4904	2.4047	2.3371	2.2821	2.2365	2.0889	2.0075	1.8421	1.7794	1.7110
26	4.2252	3.3690	2.9752	2.7426	2.5868	2.4741	2.3883	2.3205	2.2655	2.2197	2.0716	1.9898	1.8233	1.7599	1.6906
27	4.2100	3.3541	2.9604	2.7278	2.5719	2.4591	2.3732	2.3053	2.2501	2.2043	2.0558	1.9736	1.8059	1.7419	1.6717
28	4.1960	3.3404	2.9467	2.7141	2.5581	2.4453	2.3593	2.2913	2.2360	2.1900	2.0411	1.9586	1.7898	1.7251	1.6541
29	4.1830	3.3277	2.9340	2.7014	2.5454	2.4324	2.3463	2.2782	2.2229	2.1768	2.1045	2.0275	1.7748	1.7096	1.6376
30	4.1709	3.3158	2.9223	2.6896	2.5336	2.4205	2.3343	2.2662	2.2107	2.1646	2.0148	1.9317	1.7609	1.6950	1.6223
40	4.0847	3.2317	2.8387	2.6060	2.4495	2.3359	2.2490	2.1802	2.1240	2.0772	1.9245	1.8389	1.6600	1.5892	1.5089
60	4.0012	3.1504	2.7581	2.5252	2.3683	2.2541	2.1665	2.0970	2.0401	1.9926	1.8364	1.7480	1.5590	1.4814	1.3893
100	3.9361	3.0873	2.6955	2.4626	2.3053	2.1906	2.1025	2.0323	1.9748	1.9267	1.7675	1.6764	1.4772	1.3917	1.2832
200	3.8884	3.0411	2.6498	2.4168	2.2592	2.1441	2.0556	1.9849	1.9269	1.8783	1.7166	1.6233	1.4146	1.3206	1.1885
∞	3.8415	2.9957	2.6049	2.3719	2.2141	2.0986	2.0096	1.9384	1.8799	1.8307	1.6664	1.5705	1.3501	1.2434	1.0000

Inversa de la función de distribución F de Snedecor [0.975]:

$$\mathcal{F}_{m,n}(x) = P[F_{m,n} \leq x], \quad \mathcal{F}_{m,n}(F_{m,n,\alpha}) = \alpha, \quad F_{m,n,\alpha} = \mathcal{F}^{-1}_{m,n}(\alpha), \quad \alpha = 0{,}975$$

$n \diagdown m$	1	2	3	4	5	6	7	8	9	10	15	20	50	100	∞
1	647.7890	799.5000	864.1630	899.5833	921.8479	937.1111	948.2169	956.6562	963.2846	968.6274	984.8668	993.1028	1008.1171	1013.1748	1018.2583
2	38.5063	39.0000	39.1655	39.2484	39.2982	39.3315	39.3552	39.3730	39.3869	39.3980	39.4313	39.4479	39.4779	39.4879	39.4979
3	17.4434	16.0441	15.4392	15.1010	14.8848	14.7347	14.6244	14.5399	14.4731	14.4189	14.2527	14.1674	14.0099	13.9563	13.9021
4	12.2179	10.6491	9.9792	9.6045	9.3645	9.1973	9.0741	8.9796	8.9047	8.8439	8.6565	8.5599	8.3808	8.3195	8.2573
5	10.0070	8.4336	7.7636	7.3879	7.1464	6.9777	6.8531	6.7572	6.6811	6.6192	6.4277	6.3286	6.1436	6.0800	6.0153
6	8.8131	7.2599	6.5988	6.2272	5.9876	5.8198	5.6955	5.5996	5.5234	5.4613	5.2687	5.1684	4.9804	4.9154	4.8491
7	8.0727	6.5415	5.8898	5.5226	5.2852	5.1186	4.9949	4.8993	4.8232	4.7611	4.5678	4.4667	4.2763	4.2101	4.1423
8	7.5709	6.0595	5.4160	5.0526	4.8173	4.6517	4.5286	4.4333	4.3572	4.2951	4.1012	3.9995	3.8067	3.7393	3.6702
9	7.2093	5.7147	5.0781	4.7181	4.4844	4.3197	4.1970	4.1020	4.0260	3.9639	3.7694	3.6669	3.4719	3.4034	3.3329
10	6.9367	5.4564	4.8256	4.4683	4.2361	4.0721	3.9498	3.8549	3.7790	3.7168	3.5217	3.4185	3.2214	3.1517	3.0798
11	6.7241	5.2559	4.6300	4.2751	4.0440	3.8807	3.7586	3.6638	3.5879	3.5257	3.3299	3.2261	3.0268	2.9561	2.8828
12	6.5538	5.0959	4.4742	4.1212	3.8911	3.7283	3.6065	3.5118	3.4358	3.3736	3.1772	3.0728	2.8714	2.7996	2.7249
13	6.4143	4.9653	4.3472	3.9959	3.7667	3.6043	3.4827	3.3880	3.3120	3.2497	3.0527	2.9477	2.7443	2.6715	2.5955
14	6.2979	4.8567	4.2417	3.8919	3.6634	3.5014	3.3799	3.2853	3.2093	3.1469	2.9493	2.8437	2.6384	2.5646	2.4872
15	6.1995	4.7650	4.1528	3.8043	3.5764	3.4147	3.2934	3.1987	3.1227	3.0602	2.8621	2.7559	2.5488	2.4739	2.3953
16	6.1151	4.6867	4.0768	3.7294	3.5021	3.3406	3.2194	3.1248	3.0488	2.9862	2.7875	2.6808	2.4719	2.3961	2.3163
17	6.0420	4.6189	4.0112	3.6648	3.4379	3.2767	3.1556	3.0610	2.9849	2.9222	2.7230	2.6158	2.4053	2.3285	2.2474
18	5.9781	4.5597	3.9539	3.6083	3.3820	3.2209	3.0999	3.0053	2.9291	2.8664	2.6667	2.5590	2.3468	2.2692	2.1869
19	5.9216	4.5075	3.9034	3.5587	3.3327	3.1718	3.0509	2.9563	2.8801	2.8172	2.6171	2.5089	2.2952	2.2167	2.1333
20	5.8715	4.4613	3.8587	3.5147	3.2891	3.1283	3.0074	2.9128	2.8365	2.7737	2.5731	2.4645	2.2493	2.1699	2.0853
21	5.8266	4.4199	3.8188	3.4754	3.2501	3.0895	2.9686	2.8740	2.7977	2.7348	2.5338	2.4247	2.2081	2.1280	2.0422
22	5.7863	4.3828	3.7829	3.4401	3.2151	3.0546	2.9338	2.8392	2.7628	2.6998	2.4984	2.3890	2.1710	2.0901	2.0032
23	5.7498	4.3492	3.7505	3.4083	3.1835	3.0232	2.9023	2.8077	2.7313	2.6682	2.4664	2.3567	2.1374	2.0557	1.9677
24	5.7166	4.3187	3.7211	3.3794	3.1548	2.9946	2.8738	2.7791	2.7027	2.6396	2.4374	2.3273	2.1067	2.0243	1.9353
25	5.6864	4.2909	3.6943	3.3530	3.1287	2.9685	2.8478	2.7531	2.6766	2.6135	2.4110	2.3005	2.0787	1.9955	1.9055
26	5.6586	4.2655	3.6697	3.3289	3.1048	2.9447	2.8240	2.7293	2.6528	2.5896	2.3867	2.2759	2.0530	1.9691	1.8781
27	5.6331	4.2421	3.6472	3.3067	3.0828	2.9228	2.8021	2.7074	2.6309	2.5676	2.3644	2.2533	2.0293	1.9447	1.8527
28	5.6096	4.2205	3.6264	3.2863	3.0626	2.9027	2.7820	2.6872	2.6106	2.5473	2.3438	2.2324	2.0073	1.9221	1.8291
29	5.5878	4.2006	3.6072	3.2674	3.0438	2.8840	2.7633	2.6686	2.5919	2.5286	2.3248	2.2131	1.9870	1.9011	1.8072
30	5.5675	4.1821	3.5894	3.2499	3.0265	2.8667	2.7460	2.6513	2.5746	2.5112	2.3072	2.1952	1.9681	1.8816	1.7867
40	5.4239	4.0510	3.4633	3.1261	2.9037	2.7444	2.6238	2.5289	2.4519	2.3882	2.1819	2.0677	1.8324	1.7405	1.6371
60	5.2856	3.9253	3.3425	3.0077	2.7863	2.6274	2.5068	2.4117	2.3344	2.2702	2.0613	1.9445	1.6985	1.5990	1.4821
100	5.1786	3.8284	3.2496	2.9166	2.6961	2.5374	2.4168	2.3215	2.2439	2.1793	1.9679	1.8486	1.5917	1.4832	1.3473
200	5.1004	3.7578	3.1820	2.8503	2.6304	2.4720	2.3513	2.2558	2.1780	2.1130	1.8996	1.7780	1.5108	1.3927	1.2290
∞	5.0239	3.6889	3.1161	2.7858	2.5665	2.4082	2.2875	2.1918	2.1136	2.0483	1.8326	1.7085	1.4284	1.2956	1.0000

Inversa de la función de distribución F de Snedecor [0.990]:

$$\mathcal{F}_{m,n}(x) = P[F_{m,n} \leq x], \quad \mathcal{F}_{m,n}(F_{m,n,\alpha}) = \alpha, \quad F_{m,n,\alpha} = \mathcal{F}^{-1}_{m,n}(\alpha), \quad \alpha = 0{,}990$$

$n \diagdown m$	1	2	3	4	5	6	7	8	9	10	15	20	50	100	∞
1	4052.181	4999.500	5403.352	5624.583	5763.650	5858.986	5928.356	5981.070	6022.473	6055.847	6157.285	6208.730	6302.517	6334.110	6365.864
2	98.503	99.000	99.166	99.249	99.299	99.333	99.356	99.374	99.388	99.399	99.433	99.449	99.479	99.489	99.499
3	34.116	30.817	29.457	28.710	28.237	27.911	27.672	27.489	27.345	27.229	26.872	26.690	26.354	26.240	26.125
4	21.198	18.000	16.694	15.977	15.522	15.207	14.976	14.799	14.659	14.546	14.198	14.020	13.690	13.577	13.463
5	16.258	13.274	12.060	11.392	10.967	10.672	10.456	10.289	10.158	10.051	9.722	9.553	9.238	9.130	9.020
6	13.745	10.925	9.780	9.148	8.746	8.466	8.260	8.102	7.976	7.874	7.559	7.396	7.091	6.987	6.880
7	12.246	9.547	8.451	7.847	7.460	7.191	6.993	6.840	6.719	6.620	6.314	6.155	5.858	5.755	5.650
8	11.259	8.649	7.591	7.006	6.632	6.371	6.178	6.029	5.911	5.814	5.515	5.359	5.065	4.963	4.859
9	10.561	8.022	6.992	6.422	6.057	5.802	5.613	5.467	5.351	5.257	4.962	4.808	4.517	4.415	4.311
10	10.044	7.559	6.552	5.994	5.636	5.386	5.200	5.057	4.942	4.849	4.558	4.405	4.115	4.014	3.909
11	9.646	7.206	6.217	5.668	5.316	5.069	4.886	4.744	4.632	4.539	4.251	4.099	3.810	3.708	3.602
12	9.330	6.927	5.953	5.412	5.064	4.821	4.640	4.499	4.388	4.296	4.010	3.858	3.569	3.467	3.361
13	9.074	6.701	5.739	5.205	4.862	4.620	4.441	4.302	4.191	4.100	3.815	3.665	3.375	3.272	3.165
14	8.862	6.515	5.564	5.035	4.695	4.456	4.278	4.140	4.030	3.939	3.656	3.505	3.215	3.112	3.004
15	8.683	6.359	5.417	4.893	4.556	4.318	4.142	4.004	3.895	3.805	3.522	3.372	3.081	2.977	2.868
16	8.531	6.226	5.292	4.773	4.437	4.202	4.026	3.890	3.780	3.691	3.409	3.259	2.967	2.863	2.753
17	8.400	6.112	5.185	4.669	4.336	4.102	3.927	3.791	3.682	3.593	3.312	3.162	2.869	2.764	2.653
18	8.285	6.013	5.092	4.579	4.248	4.015	3.841	3.705	3.597	3.508	3.227	3.077	2.784	2.678	2.566
19	8.185	5.926	5.010	4.500	4.171	3.939	3.765	3.631	3.522	3.434	3.153	3.003	2.709	2.602	2.489
20	8.096	5.849	4.938	4.431	4.103	3.871	3.699	3.564	3.457	3.368	3.088	2.938	2.643	2.535	2.421
21	8.017	5.780	4.874	4.369	4.042	3.812	3.640	3.506	3.398	3.310	3.030	2.880	2.584	2.475	2.360
22	7.945	5.719	4.817	4.313	3.988	3.758	3.587	3.453	3.346	3.258	2.978	2.827	2.531	2.422	2.305
23	7.881	5.664	4.765	4.264	3.939	3.710	3.539	3.406	3.299	3.211	2.931	2.780	2.483	2.373	2.256
24	7.823	5.614	4.718	4.218	3.895	3.667	3.496	3.363	3.256	3.168	2.889	2.738	2.440	2.329	2.211
25	7.770	5.568	4.675	4.177	3.855	3.627	3.457	3.324	3.217	3.129	2.850	2.699	2.400	2.289	2.169
26	7.721	5.526	4.637	4.140	3.818	3.591	3.421	3.288	3.182	3.094	2.815	2.664	2.364	2.252	2.131
27	7.677	5.488	4.601	4.106	3.785	3.558	3.388	3.256	3.149	3.062	2.783	2.632	2.330	2.218	2.097
28	7.636	5.453	4.568	4.074	3.754	3.528	3.358	3.226	3.120	3.032	2.753	2.602	2.300	2.187	2.064
29	7.598	5.420	4.538	4.045	3.725	3.499	3.330	3.198	3.092	3.005	2.726	2.574	2.271	2.158	2.034
30	7.562	5.390	4.510	4.018	3.699	3.473	3.304	3.173	3.067	2.979	2.700	2.549	2.245	2.131	2.006
40	7.314	5.178	4.313	3.828	3.514	3.291	3.124	2.993	2.888	2.801	2.522	2.369	2.058	1.938	1.805
60	7.077	4.977	4.126	3.649	3.339	3.119	2.953	2.823	2.718	2.632	2.352	2.198	1.877	1.749	1.601
100	6.895	4.824	3.984	3.513	3.206	2.988	2.823	2.694	2.590	2.503	2.223	2.067	1.735	1.598	1.427
200	6.763	4.713	3.881	3.414	3.110	2.893	2.730	2.601	2.497	2.411	2.129	1.971	1.629	1.481	1.279
∞	6.635	4.605	3.782	3.319	3.017	2.802	2.639	2.511	2.407	2.321	2.039	1.878	1.523	1.358	1.000

Inversa de la función de distribución F de Snedecor [0.995]:

$$\mathcal{F}_{m,n}(x) = P[F_{m,n} \leq x], \quad \mathcal{F}_{m,n}(F_{m,n,\alpha}) = \alpha, \quad F_{m,n,\alpha} = \mathcal{F}^{-1}_{m,n}(\alpha), \quad \alpha = 0{,}995$$

$n \diagdown m$	1	2	3	4	5	6	7	8	9	10	15	20	50	100	∞
1	16210.72	19999.50	21614.74	22499.58	23055.80	23437.11	23714.57	23925.41	24091.00	24224.49	24630.21	24835.97	25211.09	25337.45	25464.46
2	198.50	199.00	199.17	199.25	199.30	199.33	199.36	199.37	199.39	199.40	199.43	199.45	199.48	199.49	199.50
3	55.55	49.80	47.47	46.19	45.39	44.84	44.43	44.13	43.88	43.69	43.08	42.78	42.21	42.02	41.83
4	31.33	26.28	24.26	23.15	22.46	21.97	21.62	21.35	21.14	20.97	20.44	20.17	19.67	19.50	19.32
5	22.78	18.31	16.53	15.56	14.94	14.51	14.20	13.96	13.77	13.62	13.15	12.90	12.45	12.30	12.14
6	18.63	14.54	12.92	12.03	11.46	11.07	10.79	10.57	10.39	10.25	9.81	9.59	9.17	9.03	8.88
7	16.24	12.40	10.88	10.05	9.52	9.16	8.89	8.68	8.51	8.38	7.97	7.75	7.35	7.22	7.08
8	14.69	11.04	9.60	8.81	8.30	7.95	7.69	7.50	7.34	7.21	6.81	6.61	6.22	6.09	5.95
9	13.61	10.11	8.72	7.96	7.47	7.13	6.88	6.69	6.54	6.42	6.03	5.83	5.45	5.32	5.19
10	12.83	9.43	8.08	7.34	6.87	6.54	6.30	6.12	5.97	5.85	5.47	5.27	4.90	4.77	4.64
11	12.23	8.91	7.60	6.88	6.42	6.10	5.86	5.68	5.54	5.42	5.05	4.86	4.49	4.36	4.23
12	11.75	8.51	7.23	6.52	6.07	5.76	5.52	5.34	5.20	5.09	4.72	4.53	4.17	4.04	3.90
13	11.37	8.19	6.93	6.23	5.79	5.48	5.25	5.08	4.94	4.82	4.46	4.27	3.91	3.78	3.65
14	11.06	7.92	6.68	6.00	5.56	5.26	5.03	4.86	4.72	4.60	4.25	4.06	3.70	3.57	3.44
15	10.80	7.70	6.48	5.80	5.37	5.07	4.85	4.67	4.54	4.42	4.07	3.88	3.52	3.39	3.26
16	10.58	7.51	6.30	5.64	5.21	4.91	4.69	4.52	4.38	4.27	3.92	3.73	3.37	3.25	3.11
17	10.38	7.35	6.16	5.50	5.07	4.78	4.56	4.39	4.25	4.14	3.79	3.61	3.25	3.12	2.98
18	10.22	7.21	6.03	5.37	4.96	4.66	4.44	4.28	4.14	4.03	3.68	3.50	3.14	3.01	2.87
19	10.07	7.09	5.92	5.27	4.85	4.56	4.34	4.18	4.04	3.93	3.59	3.40	3.04	2.91	2.78
20	9.94	6.99	5.82	5.17	4.76	4.47	4.26	4.09	3.96	3.85	3.50	3.32	2.96	2.83	2.69
21	9.83	6.89	5.73	5.09	4.68	4.39	4.18	4.01	3.88	3.77	3.43	3.24	2.88	2.75	2.61
22	9.73	6.81	5.65	5.02	4.61	4.32	4.11	3.94	3.81	3.70	3.36	3.18	2.82	2.69	2.55
23	9.63	6.73	5.58	4.95	4.54	4.26	4.05	3.88	3.75	3.64	3.30	3.12	2.76	2.62	2.48
24	9.55	6.66	5.52	4.89	4.49	4.20	3.99	3.83	3.69	3.59	3.25	3.06	2.70	2.57	2.43
25	9.48	6.60	5.46	4.84	4.43	4.15	3.94	3.78	3.64	3.54	3.20	3.01	2.65	2.52	2.38
26	9.41	6.54	5.41	4.79	4.38	4.10	3.89	3.73	3.60	3.49	3.15	2.97	2.61	2.47	2.33
27	9.34	6.49	5.36	4.74	4.34	4.06	3.85	3.69	3.56	3.45	3.11	2.93	2.57	2.43	2.29
28	9.28	6.44	5.32	4.70	4.30	4.02	3.81	3.65	3.52	3.41	3.07	2.89	2.53	2.39	2.25
29	9.23	6.40	5.28	4.66	4.26	3.98	3.77	3.61	3.48	3.38	3.04	2.86	2.49	2.36	2.21
30	9.18	6.35	5.24	4.62	4.23	3.95	3.74	3.58	3.45	3.34	3.01	2.82	2.46	2.32	2.18
40	8.83	6.07	4.98	4.37	3.99	3.71	3.51	3.35	3.22	3.12	2.78	2.60	2.23	2.09	1.93
60	8.49	5.79	4.73	4.14	3.76	3.49	3.29	3.13	3.01	2.90	2.57	2.39	2.01	1.86	1.69
100	8.24	5.59	4.54	3.96	3.59	3.33	3.13	2.97	2.85	2.74	2.41	2.23	1.84	1.68	1.49
200	8.06	5.44	4.41	3.84	3.47	3.21	3.01	2.86	2.73	2.63	2.30	2.11	1.71	1.54	1.31
∞	7.88	5.30	4.28	3.72	3.35	3.09	2.90	2.74	2.62	2.52	2.19	2.00	1.59	1.40	1.00

Inversa de la función de distribución χ^2 de Pearson:

$$\mathcal{X}_n(x) = P[\chi^2_n \leq x], \quad \mathcal{X}_n(\chi^2_{n,\alpha}) = \alpha, \quad \mathcal{X}_n^{-1}(\alpha) = \chi^2_{n,\alpha}$$

$n \diagdown \alpha$	0.05	0.1	0.2	0.3	0.5	0.7	0.8	0.9	0.95	0.975	0.99	0.995
1	0.00393	0.01579	0.06418	0.14847	0.45494	1.07419	1.64237	2.70554	3.84146	5.02389	6.63490	7.87944
2	0.10259	0.21072	0.44629	0.71335	1.38629	2.40795	3.21888	4.60517	5.99146	7.37776	9.21034	10.59663
3	0.35185	0.58437	1.00517	1.42365	2.36597	3.66487	4.64163	6.25139	7.81473	9.34840	11.34487	12.83816
4	0.71072	1.06362	1.64878	2.19470	3.35669	4.87843	5.98862	7.77944	9.48773	11.14329	13.27670	14.86026
5	1.14548	1.61031	2.34253	2.99991	4.35146	6.06443	7.28928	9.23636	11.07050	12.83250	15.08627	16.74960
6	1.63538	2.20413	3.07009	3.82755	5.34812	7.23114	8.55806	10.64464	12.59159	14.44938	16.81189	18.54758
7	2.16735	2.83311	3.82232	4.67133	6.34581	8.38343	9.80325	12.01704	14.06714	16.01276	18.47531	20.27774
8	2.73264	3.48954	4.59357	5.52742	7.34412	9.52446	11.03009	13.36157	15.50731	17.53455	20.09024	21.95495
9	3.32511	4.16816	5.38005	6.39331	8.34283	10.65637	12.24215	14.68366	16.91898	19.02277	21.66599	23.58935
10	3.94030	4.86518	6.17908	7.26722	9.34182	11.78072	13.44196	15.98718	18.30704	20.48318	23.20925	25.18818
11	4.57481	5.57778	6.98867	8.14787	10.34100	12.89867	14.63142	17.27501	19.67514	21.92005	24.72497	26.75685
12	5.22603	6.30380	7.80733	9.03428	11.34032	14.01110	15.81199	18.54935	21.02607	23.33666	26.21697	28.29952
13	5.89186	7.04150	8.63386	9.92568	12.33976	15.11872	16.98480	19.81193	22.36203	24.73560	27.68825	29.81947
14	6.57063	7.78953	9.46733	10.82148	13.33927	16.22210	18.15077	21.06414	23.68479	26.11895	29.14124	31.31935
15	7.26094	8.54676	10.30696	11.72117	14.33886	17.32169	19.31066	22.30713	24.99579	27.48839	30.57791	32.80132
16	7.96165	9.31224	11.15212	12.62435	15.33850	18.41789	20.46508	23.54183	26.29623	28.84535	31.99993	34.26719
17	8.67176	10.08519	12.00227	13.53068	16.33818	19.51102	21.61456	24.76904	27.58711	30.19101	33.40866	35.71847
18	9.39046	10.86494	12.85695	14.43986	17.33790	20.60135	22.75955	25.98942	28.86930	31.52638	34.80531	37.15645
19	10.11701	11.65091	13.71579	15.35166	18.33765	21.68913	23.90042	27.20357	30.14353	32.85233	36.19087	38.58226
20	10.85081	12.44261	14.57844	16.26586	19.33743	22.77454	25.03751	28.41198	31.41043	34.16961	37.56623	39.99685
21	11.59131	13.23960	15.44461	17.18227	20.33723	23.85779	26.17110	29.61509	32.67057	35.47888	38.93217	41.40106
22	12.33801	14.04149	16.31404	18.10072	21.33704	24.93902	27.30145	30.81328	33.92444	36.78071	40.28936	42.79566
23	13.09051	14.84796	17.18651	19.02109	22.33688	26.01837	28.42879	32.00690	35.17246	38.07563	41.63840	44.18128
24	13.84842	15.65868	18.06180	19.94323	23.33673	27.09596	29.55332	33.19624	36.41503	39.36408	42.97982	45.55851
25	14.61141	16.47341	18.93975	20.86703	24.33659	28.17192	30.67520	34.38159	37.65248	40.64647	44.31410	46.92789
26	15.37916	17.29188	19.82019	21.79240	25.33646	29.24633	31.79461	35.56317	38.88514	41.92317	45.64168	48.28988
27	16.15140	18.11390	20.70298	22.71924	26.33634	30.31929	32.91169	36.74122	40.11327	43.19451	46.96294	49.64492
28	16.92788	18.93924	21.58797	23.64746	27.33623	31.39088	34.02657	37.91592	41.33714	44.46079	48.27824	50.99338
29	17.70837	19.76774	22.47505	24.57699	28.33613	32.46117	35.13936	39.08747	42.55697	45.72229	49.58788	52.33562
30	18.49266	20.59923	23.36411	25.50776	29.33603	33.53023	36.25019	40.25602	43.77297	46.97924	50.89218	53.67196
40	26.50930	29.05052	32.34495	34.87194	39.33534	44.16487	47.26854	51.80506	55.75848	59.34171	63.69074	66.76596