Estadística básica

Nelson

Resumen

Este es un resumen del libro *Estadística básica con R*, temario correspondiente a la asignatura Estadística Básica de primer curso del grado en Matemáticas de la UNED.

El resumen incluye todos los teoremas, definiciones, proposiciones, lemas y textos relevantes del libro respetando su nomenclatura y numeración para facilitar su uso en relación a la fuente original.

El resumen no incluye ejemplos (salvo textos relevantes), demostraciones ni ejercicios.

Puede contener erratas (consulte la fuente original).

Índice

1. Introducción al R **3**
- 1.1. Introducción . . . 3
- 1.2. El editor de objetos R . . . 3
- 1.3. Datos en R . . . 3
  - 1.3.1. Vectores . . . 3
  - 1.3.2. Factores . . . 4
  - 1.3.3. Matrices . . . 4
  - 1.3.4. Estructura de datos . . . 4
  - 1.3.5. Listas . . . 4
  - 1.3.6. Nombres a las filas y columnas de matrices y vectores . . . 4
- 1.4. Gráficos . . . 4
  - 1.4.1. Funciones gráficas de alto nivel . . . 4
  - 1.4.2. Funciones gráficas de bajo nivel . . . 4
- 1.5. Otras cuestiones . . . 5
- 1.6. Interfaz . . . 5
- 1.7. Modificar y crear funciones . . . 5
- 1.8. Librerías de R . . . 5

2. Estadística descriptiva **5**
- 2.1. Introducción a la estadística . . . 5
  - 2.1.1. Población e individuo . . . 5
  - 2.1.2. Muestras aleatorias . . . 5
  - 2.1.3. Variable aleatoria y modelo probabilístico . . . 6
  - 2.1.4. Diferentes estadísticas . . . 6
- 2.2. Conceptos fundamentales de la Estadística descriptiva . . . 6
- 2.3. Distribuciones unidimensionales de frecuencias . . . 7
  - 2.3.1. Representaciones gráficas de las distribuciones unidimensionales de frecuencias . . . 7
  - 2.3.2. Medidas de tendencia central de caracteres cuantitativos . . . 8
  - 2.3.3. Medidas de dispersión . . . 9
  - 2.3.4. Medidas de asimetría . . . 10
  - 2.3.5. Medidas de posición y dispersión con R . . . 10
- 2.4. Distribuciones bidimensionales de frecuencias 10
  - 2.4.1. Representaciones gráficas de las distribuciones bidimensionales de frecuencias . . . 11
  - 2.4.2. Ajuste por mínimos cuadrados . . . 11
  - 2.4.3. Precisión del ajuste por mínimos cuadrados . . . 12

3. Probabilidad **13**
- 3.1. Introducción . . . 13
- 3.2. Espacio muestral . . . 13
- 3.3. Conceptos de probabilidad . . . 13
- 3.4. Propiedades elementales de la probabilidad 14
- 3.5. Asignación de probabilidad en espacios muestrales discretos . . . 14
- 3.6. Modelo uniforme . . . 14
- 3.7. Probabilidad condicionada . . . 15
- 3.8. Independencia de sucesos . . . 15
- 3.9. Teorema de la probabilidad total . . . 15
- 3.10. Teorema de Bayes . . . 16

4. Modelos probabilísticos **16**
- 4.1. Introducción . . . 16
- 4.2. Distribución de probabilidad . . . 16
  - 4.2.1. Funciones básicas de R en probabilidades . . . 17
- 4.3. Variables aleatorias multivariantes . . . 17
- 4.4. Modelos unidimensionales discretos . . . 17
  - 4.4.1. Distribución binomial . . . 17
  - 4.4.2. Distribución de Poisson . . . 18
  - 4.4.3. Distribución geométrica . . . 18
  - 4.4.4. Distribución hipergeométrica . . . 18
  - 4.4.5. Distribución binomial negativa . . . 19
- 4.5. Modelos unidimensionales continuos . . . 19
  - 4.5.1. Distribución normal . . . 19
  - 4.5.2. Distribución uniforme . . . 19
  - 4.5.3. Distribución beta . . . 19
  - 4.5.4. Distribuciones Gamma y exponencial 20
  - 4.5.5. Distribución de Cauchy . . . 20
- 4.6. Modelos bidimensionales . . . 20
  - 4.6.1. Distribución normal bivariante . . . 20
- 4.7. Teorema central del límite . . . 20

5. Estimadores. Distribución en el muestreo **20**
- 5.1. Introducción . . . 20
- 5.2. Método de la máxima verosimilitud . . . 21
- 5.3. Distribuciones asociadas a poblaciones normales . . . 21
  - 5.3.1. Distribución χ^2 de Pearson . . . 21
  - 5.3.2. Distribución t de Student . . . 22
  - 5.3.3. Distribución F de Snedecor . . . 22
- 5.4. Estimación de la media de una población normal . . . 23
- 5.5. Estimación de la media de una población no necesariamente normal. Muestras grandes 23
- 5.6. Estimación de la varianza de una población normal . . . 24
- 5.7. Estimación del cociente de varianzas de dos poblaciones normales independientes . . . 24
- 5.8. Estimación de la diferencia de medias de dos poblaciones normales independientes . . 24
- 5.9. Estimación de la diferencia de medias de dos poblaciones independientes no necesariamente normales. Muestras grandes . . . 25
- 5.10. Datos apareados . . . 25
- 5.11. Tamaño muestral para una precisión dada . 26

6. Intervalos de confianza **26**
- 6.1. Introducción . . . 26
  - 6.1.1. Cálculo de intervalos de confianza con R . . . 26
- 6.2. Intervalo de confianza para la media de una población normal . . . 27
- 6.3. Intervalo de confianza para la media de una población no necesariamente normal. Muestras grandes . . . 27
- 6.4. Intervalo de confianza para la varianza de una población normal . . . 28
- 6.5. Intervalo de confianza para el cociente de varianzas de dos poblaciones normales independientes . . . 28

6.6. Intervalo de confianza para la diferencia de medias de dos poblaciones normales independientes 28
6.7. Intervalo de confianza para la diferencia de medias de dos poblaciones independientes no necesariamente normales. Muestas grandes 29
6.8. Intervalos de confianza para datos apareados 29

7. Contraste de hipótesis 29
7.1. Introducción y conceptos fundamentales . . 29
7.2. Contraste de hipótesis relativas a la media de una población normal 31
7.3. Contraste de hipótesis relativas a la media de una población no necesariamente normal. Muestras grandes 32
7.4. Contraste de hipótesis relativas a la varianza de una población normal 33
7.5. Contraste de hipótesis relativas a las varianzas de dos poblaciones normales independientes 34
7.6. Contraste de hipótesis relativas a la diferencia de medias de dos poblaciones normales independientes 34
7.7. Contraste de hipótesis relativas a la diferencia de medias de dos poblaciones independientes no necesariamente normales. Muestras grandes 35
7.8. Contrastes de hipótesis para datos apareados 36

8. Contrastes no paramétricos 37
8.1. Introducción 37
8.2. Pruebas χ^2 37
8.2.1. Pruebas χ^2 con R 37
8.2.2. Contraste de bondad del ajuste . . . 37
8.2.3. Contraste de homogeneidad de varias muestras 38
8.2.4. Contraste de independencia de caracteres 39
8.3. Test relativos a una muestra y datos apareados 39
8.3.1. El contraste de los signos 39
8.3.2. El contraste de los rangos signados de Wilcoxon 41
8.4. Tests relativos a dos muestras independientes 42
8.4.1. El contraste de Wilcoxon-Mann-Whitney 42
8.4.2. El contraste de la Mediana 43

9. Análisis de la Varianza 43
9.1. Introducción 43
9.2. Análisis de la varianza para un factor: Diseño Completamente Aleatorizado 43
9.3. Análisis de la varianza con R 45
9.4. Análisis de las condiciones 45
9.5. Comparaciones múltiples 45
9.6. Comparaciones múltiples con R 46

10. Regresión lineal y correlación 46
10.1. Introducción 46
10.2. Modelo de la regresión lineal simple 46
10.2.1. Interpretación de los coeficientes de regresión 46
10.3. Contraste de la Regresión Lineal Simple . . 46
10.3.1. Análisis de la variación explicada frente a la no explicada por la recta de regresión 46
10.3.2. Contraste de hipótesis para β_1 . . . 47
10.4. Regresión lineal con R 48
10.5. Correlación lineal 48
10.5.1. Estimación por punto de ρ 48
10.5.2. Contraste de hipótesis sobre ρ 48
10.6. Modelo de la regresión lineal múltiple . . . 48
10.6.1. Contraste de la regresión lineal múltiple 49

1. Introducción al R

1.1. Introducción

Como ocurre con todos los paquetes estadísticos, R también utiliza un lenguaje propio. Toda instrucción que pueda ser ejecutada desde la línea de comandos se denomina *expresión*. Las expresiones se ejecutan con `Enter`.

Éstas pueden tener una longitud de más de una línea. Cuando se presiona `Enter` después de una expresión sintácticamente incompleta, ésta no se ejecuta ni se producen mensajes de error; aparece en *promt* + al comienzo de una nueva línea de comandos invitándonos así a completar la expresión no concluida.

Los elementos básicos de R son los *objetos* y, por tanto, a ellos se referirán las expresiones de R. De hecho, los objetos son ficheros capaces de ser editados y, en su caso, ejecutados.

Un *objeto* es el resultado de ejecutar una expresión en la que aparece el operador `<-`. Dicho de otra manera una expresión que nos interese guardar puede ser *salvada* con el operador `<-`, también denominado *asignación*.

El nombre asignado a un objeto debe empezar por una letra y puede incluir cualquier combinación de letras mayúsculas o minúsculas, números y puntos.

Los dos tipos de objetos más utilizados son el *dato* y la *función*. Las funciones constan de un nombre seguido de dos paréntesis, `nombre();` entre los paréntesis se incluyen sus *argumentos*. Si sólo ejecutamos su nombre obtendremos su definición y si ejecuta `?nombre` obtendrá ayuda sobre su utilización.

Una de las funciones más sencillas y por medio de la cual *salimos* del programa es `q()`. Al ejecutarla, el ordenador nos preguntará si queremos conservar los cálculos que hayamos realizado en la sesión. Si respondemos *Sí*, al comenzar la sesión siguiente podremos volver a utilizar los resultados de la sesión recién finalizada.

También se puede usar la línea de comandos como una potente calculadora matemática como realizar las habituales operaciones matemáticas, obtener el valor de las funciones más conocidas como la potencial, la exponencial, la raíz cuadrada,... se pueden resolver sistemas de ecuaciones o hacer integración numérica y otras muchas aplicaciones matemáticas.

Dos funciones que queremos destacar son, la función `objects()`, mediante la cual podemos listar los objetos de R existentes.

Y la función `rm()`, utilizada para suprimir objetos; para ello debemos utilizar como argumentos suyos, los objetos a eliminar.

Para recuperar alguna instrucción ejecutada anteriormente, basta con pulsar la tecla ↑ tantas veces como sea necesario hasta que la instrucción aparezca. Luego deberemos ejecutarla pulsando `Enter`.

1.2. El editor de objetos R

Para crear o modificar objetos o funciones R, debemos utilizar las denominadas *funciones editoras* `edit()` y, preferiblemente `fix()`.

También pueden crearse funciones desde la línea de comandos. Si queremos crear una nueva función, deberemos asignarla con `nombre_nueva_función <- function(x) {expresión}`, colocando entre las llaves la definición de la función.

Cualquier asignación que hagamos eliminará asignaciones previas con el mismo nombre.

Un conjunto de instrucciones de R se denomina *script*, el cual puede ser guardado en el propio programa. Aquella parte que marquemos puede ser ejecutada con las teclas `Ctrl + R`.

1.3. Datos en R

En R se denomina *dato* al resultado de ejecutar una expresión R, es decir, un tipo de objeto.

Cada tipo de dato tiene asociados determinados *atributos*; el más importante es su *modo*. Consideramos cuatro clases de *modos*:

- *logical*(lógico): Modo binario en donde los valores posibles son `T` ó `F` (Verdadero o Falso).
- *numeric*(numérico): Modo en donde los valores posibles son números reales.
- *complex*(complejo): Modo en donde los valores posbiles son números complejos.
- *character*(carácter): Modo en donde los valores posibles son caracteres separados por comillas.

Consideramos cinco tipos de datos diferentes:

- *vector*(vector): Conjunto de elementos en un orden específico. Todos los elementos de un vector deben ser del mismo modo. Los más utilizados son los *vectores numéricos*, es decir, vectores cuyos elementos son números.
- *matrix*(matriz): Disposición bidimensional de elementos de un mismo modo.
- *factor*(factor): Vector cuyos elementos son valores procedentes de un número finito de *categorías*.
- *data frame*(estructura de datos): Disposición bidimensional de elementos cuyas columnas pueden estar formadas por elementos de distinto *modo*.
- *list*(lista): Expresión más general de dato, la cual puede contener colecciones arbitrarias de datos.

1.3.1. Vectores

Todos los elementos de un vector deben ser del mismo modo. El otro atributo considerado en un vector es su *longitud*.

Si queremos conocer el modo o la longitud de un vector se deben usar las funciones `mode()` y `length()`.

La forma más sencilla de crear un vector es utilizando la función `c()` y asignandola al nombre deseado, `nombre_vector <- c(elementos_vector)`.

Si los elementos de un vector son del modo *carácter*, debemos incluir dichos elementos entre comillas.

Los elementos de un vector pueden ser otros vectores.

Una situación habitual es que tengamos nuestros datos en un fichero *ascii*. En ese caso, con objeto de crear un vector debemos utilizar la función `scan()`, y asignarla al nombre que queramos dando como argumento la dirección del fichero, `nombre_vector<-scan(dirección_fichero)`.

1.3.2. Factores

El *factor* es un vector de datos no númericos formado por datos procedentes de categorías.

1.3.3. Matrices

Una matriz es una disposición bidimensional en donde todos los elementos deben ser del mismo *modo*.

Para crear una matriz utilizaremos la función `matrix()` con dos argumentos, la función `c()`, la cual tendrá a su vez como argumentos todos los datos a introducir, y el número de columnas que deberá tener la matriz. La matriz se construirá por columnas. `nombre_matriz <-matrix(c(elementos_matriz),ncol= numero_columnas)`.

Podemos utilizar, el lugar del argumento `ncol`, el argumento `nrow`, el cual asigna el número de filas que deberá tener la matriz. No obstante, ésta se seguirá formando por columnas, es decir, con los valores aportados por la función `c()` va completando columnas. Si quisiéramos que la completara por filas, utilizaríamos el agumento `byrow=T`.

SI tenemos dos o más vectores del mismo *modo* y además tienen la misma longitud, se pueden combinar para formar una matriz utilizando la función `cbind()`, que une los vectores por columnas. De forma análoga se podría utilizar la función `rbind` para que los uniera por filas.

Otro atributo importante de una matriz es su dimensión, que se puede averiguar mediante la función `dim()`.

Para poner nombre a las filas y columnas se utiliza, dentro de la función `matrix()`, el argumento `dimnames` el cual debe ser una lista de exactamente dos componentes, el primero de los cuales da los nombres de las filas de la matriz y el segundo la de los componentes. También es posible poner nombres a las filas y columnas de matrices ya creadas ejecutando la expresión `dimnames(nombre_matriz) <- list(c(nombres_filas), c(nombres_columnas))`.

Si queremos fomrar una matriz a partir de los datos de un fichero usamos `scan()` y le asignamos un nombre, `nombre_matriz <- scan(dirección_fichero), ncol= numero_columnas`.

1.3.4. Estructura de datos

Las estructuras de datos o *data frames* puede contener datos de varios *modos* en diferentes columnas.

Para crear *estructuras de datos* podemos utilizar dos funciones: Una, la función `data.frame()`, la cual una al igual que la función `matrix()`, objetos de varias clases, también por columnas.

Por otro lado, para leer datos procedentes de un fichero externo debemos utilizar la función `read.table()`, asignandole el nombre deseado. El argumento `header=T` es para indicar que queremos incorporar la primera línea de nombres de las variables. `nombre_estructura <-read.table("dirección_fichero", header=T)`.

1.3.5. Listas

La *lista* puede admitir datos de diferentes *modos* y de diferentes longitudes, e inclusive otras listas. La mayorá de funciones R que realizan un análisis esstadístico presentan sus resultados en una lista.

Para crear una lista se utiliza la función `list()` en donde cada uno de sus argumentos se convierte en un componente de la lista.

1.3.6. Nombres a las filas y columnas de matrices y vectores

Mediante la función `names()` podemos crear un *vector de nombres* de la misma longitud que el vector, `names(nombre_vector)<-c(nombres_elementos_vector)`.

1.4. Gráficos

La ventana de gráficos se abre de forma automática al ejecutar alguna función que los realice.

1.4.1. Funciones gráficas de alto nivel

Las funciones de gráficas de alto nivel producen un gráfico totalmente nuevo, incluidos los ejes y sus etiquetas, borrando previamente el gráfico que pudiera existir en la ventana de gráficos.

Las funciones gráficas de alto nivel se dividen en varios grupos dependiendo de lo que queramos representar. Si queremos representar funciones matemáticas, primero debemos crear un vector de valores correspondientes a las abscisas, es decir, el dominio de la función en los que va a ser evaluada ésta y , luego, realizar el gráfico deseado de pares de puntos utilizando la función `plot(vector_dominio, expresión_función, type="l")`, el argumento `type="l"` (en donde hemos utilizado la letra l y no el número 1) especifica que el gráfico de la función debe aparecer mediantre trazos sólidos.

Si queremos representar pares de datos `(x,y)` mediante un diagrama de dispresión, simplemente utilizaríamos la función `plot(x,y)`, donde `x` e `y` son vectores.

1.4.2. Funciones gráficas de bajo nivel

Se utilizan para modificar gráficos ya existentes y por lo tanto no suprimer el gráfico existente en la ventana de gráficos.

Las funciones `points()` y `lines()` son las funciones gráficas de bajo nivel corespondientes a `type="p"` y `type="l"` respectivamente.

La función `abline` permite añadir una línea recta a un gráfico ya existente, especificando los valoder de la ordenada en el origen y la pendiente.

1.5. Otras cuestiones

R distingue entre mayúsculas y minúsculas, por el contrario, uno o varios espacios son interpretados de la misma manera.

R no ejecuta lo que haya en una línea detrás del simbolo `#`, por lo que, en ocasiones, se incluyen comentarios después de una expresión mediante ese símbolo.

Un símbolo que aparece con frecuencia es el símbolo ~, que corresponde con el de código ASCII 126, con lo que se obtendrá manteniendo presionada la tecta `ALT` y, al mismo tiempo, tecleando en el teclado numérico el número 126.

1.6. Interfaz

Existen varios interfaces. Uno de ellos se denomina *DAS+R* y se puede obtener de http://www.statistik.tuwien.ac.at/StatDA/DASplusR

Una vez hayamos bajado la carpeta .zip, se instala ejecutando, si la tenemos por ejemplo en c: `install.packages(repos=NULL, ç://DASplusR.zip")`

Pero el interfas más habitualmente utilizado es *Rcmdr*, denominado *R-commander*, y que se instala en una sesión R (estando conectado a Internet) con la pestaña superior de la consola R, *Paquetes*.

Estando en una sesión de R, obtenemos el primer interfaz ejecutando `library(DASplusR)`. EL interfas *Rcmdr* se obtiene ejecutando `library(Rcmdr)`.

1.7. Modificar y crear funciones

Las funciones se pueden modificar a nuestra conveniencia con las *funciones editoras* `fix()` y `edit()` (la primera de ellas lo que hace es llamar a la segundo pero al salir la modifica, luego es más recomendable).

Al modificar una función previamente incluida en R, habitualmente no querremos prescindir de ella, por lo que definiremos primero una nueva y modificaremos la nueva.

Tenga cuidado con el editor. Si ha cometido algún error al programar, al cerrarlo se perderá todo lo que hubiera hecho, por lo que aconsejamos copiarlo antes de cerrarlo.

Recuerde que al salir de R, debe decir que Sí quiere conservar los cambios porque, en caso contrario, no le quedarán salvados.

1.8. Librerías de R

Muchas librerías ya estarán incorporadas a la versión de R que utilicemos por lo que, si queremos comprobar si tenemos en nuestro programa una determinada librería ejecutaremos `library(nombre_libreria)`.

Si el programa no dice nada es que la tenemos y la hemos *abierto* por lo que podemos utilizar las funciones que contiene.

En algunas ocasiones tendremos claro qué librería queremos incorporar a R pero en otras ocasiones no. En estas últimas, una buena manera de acutar es buscar en Google el nombre (en inglés) del método que queramos utilizar precedido de una R entre corchetes.

En la dirección http://lib.stat.cmu.edu/R/CRAN/web/packages/ tenemos la relación de librerías .ºficiales". *Pinchando* en su nombre tendremos, entre otras cosas, un fichero en pdf que nos da indicaciones de lo que hace.

Pero lo más interesante es que podemos incorporarla fácilmente a nuestro R. Para ello, con R abierto y conectados a Internet, desplegamos la pestaña superior `Paquetes` y elegimos la opción `Seleccionar espejo CRAN`; aquí elegimos, preferiblemente, algún lugar cercano a donde tengamos instalado el ordenador. Después, dentro de la misma pestaña `Paquetes`, elegimos la opción `Instalar paquete(s)` y allí seleccionamos el paquete que estamos buscando.

Bastará hacer esto una sola vez. Luego ya estará instalado en R como cualquier otro paquete y para abrirlo sólo tendremos que ejecutar la función `library()`.

2. Estadística descriptiva

2.1. Introducción a la estadística

2.1.1. Población e individuo

Los fenómenos aleatorios se presentan en un mundo real formado por *individuos*, en los que se observa el fenómeno aleatorio en estudio. El conjunto de todos los individuos recibe el nombre de *población*.

Los términos *población* e *individuo*, no deben ser entendidos necesariamente en uns entido de población humana y persona humana, sino, respectivamente, como colectivo del que queremos sacar conclusiones y como elemento o unidad que compone la población.

Una cuestión muy importante es la de determinar con precisión lo que constituye la población ya que de ella se elegirán unos cuantos individuos con objeto de obtener conclusiones acerca de toda la población.

La definición de lo que constituye la población depende del expreimentador y de la naturaleza del problema que se investiga. No obstante, una vez definida, de ella se tomarán las observaciones y se deberán sacar las conclusiones. Al conjunto de individuos que elegimos de la población lo denominaremos *muestra*.

Es muy importante fijar la población con toda precisión, ya que solamente la obtención de una muestra representativa de la población permitirá obtener conclusiones fiables sobre ella.

Habitualmente la muestra representativa se obtendrá por un procedimiento aleatorio, lo cual permitirá medir y controlar los posibles errores en términos de probabilidades, pero insistimos que lo importante es obtener una muestra representativa de la población sea o no un procedimiento aleatorio.

2.1.2. Muestras aleatorias

Trabajar con muestras aleatorias es una cuestión de suma importancia a la hora de obtener buenas concluisones, ya que la muestra será la *materia prima* a utilizar en la

elaboración de inferencias, y solamente de buena mateira prima se pueden obtener buenos productos.

Al elegir una muestra aleatoria debemos asegurarnos de que el método que elegimos para selecionar no está sesgado por las caracteristicas del propio individuo.

Este laborioso proceso se simplifica notablemente con la utilización de programas que generan *números aleatorios* aunque en estos siempre está presente la arbitreriedad del inicio o *semilla* de la elección.

De todas formas, en los trabajos de campo, la selección aleatoria es más complicada, por lo que deberemos admitir que un muestreo aleatorio es un ideal que el investigador debe esforzarse en conseguir y que probablemente nunca llegará a alcanzar completamente. La propia inferencia estadística proporciona técnicas que permiten chequear si la muestra obtenida puede considerarse aleatoria o no.

2.1.3. Variable aleatoria y modelo probabilístico

Habitualmente la situación que se presenta es la de una característica o valor poblacional objeto de investigación, al que denominaremos parámetro poblacional o simplemente *parámetro*, estando éste asociado a una variable de estudio.

Desde un punto de vista técnico, esta *variable* en estudio que deberemos identificar en el expreimento que estemos realizando, se corresponde con lo que matemáticamente se deomina *variable aleatoria* X y que, como aquí, de forma habitual denominaremos simplemente variable.

Con objeto de hacer *inferencias* sobre el parámetro en estudio, es decir, o bien poder llegar a dar un valor como estimación suya (*estimación por punto*), o bien dar un intervalo numérico en el que versímilmente se encuentre (*estimación por intervalos de confianza*), o bien poder decidir si puede considerarse razonable un valor u otro para dicho parámetro (*constante de hipótesis*), el investigador selecciona al azar de la población unos cuantos individuos, digamos n, los cuales, como dijimos antes, constituyen la *muestra*, siendo n el *tamaña muestral*, en los que se observará la variable en estudio.

Se obtendrán así n realizaciones de la variable aleatoria en estudio X, que representaremos por $(X_1, ..., X_n)$, entendiendoes cada $X_i, i = 1, ..., n$ como el valor que toma la variable en estudio en el individuo seleccionado al azar en el i-ésimo lugar.

Aquí sólo consideraremos la situación en la que cada individuo es seleccionado de forma independiente e idéntica a como lo son los demás. Matemáticamente esto significa que las n variables aleatorias son lo que se dice *independientes* e *idénticamente distribuidas*.

La realización de las observaciones en los individuos de la muestra dará origen a los *datos*.

Los valores posibles de cada variable aleatoria junto con las probabilidades con los que los toma, se denomina distribución o ley de probabilidad de la variable aleatoria en estudio, o más brevemente *modelo probabilístico*.

2.1.4. Diferentes estadísticas

El propósito de la *Inferencia estadística* es el de obtener concluisones de la población en estudio en base a la muestra obtenida de ella, mientras que el objetivo de la *Estadística descriptiva* es el de, dados los datos, ordenarlos, simplificarlos, resumirlos, clasificarlos, etc. determinando de esta manera un conjunto de valores que, además de proporcionar una rápida impresión de sus principales características, permitan hacer comparaciones con otros conjuntos de datos.

Existe aún una tercera posibilidad: utilizar información *a priori* sobre el parámetro a la hora de hacer nuestras inferencias. Esta situación, denominada *Inferencia Bayesiana*, no será tratada aquí.

2.2. Conceptos fundamentales de la Estadística descriptiva

Caracteres

Cada uno de los individuos de la población en estudio posee uno o varios *caracteres*. La observación de uno o más de esos caracteres en los individuos de la muestra es lo que dará origen a los datos.

Los caracteres pueden ser de dos clases: *cuantitativos*, cuando son tales que su observación en un individuo determinado proporciona un valor numérico como medida asociada, o *cualitativos*, cuando su observación en los individuos no suministra un número, sino la pertenecencia a una clase determinada.

Modalidades de los caracteres

A las posibilidades, tipos o clases que pueden presenta los caracteres las denominaremos *modalidades*.

Las modalidades de un carácter deben ser a la vez incompatibles y exhaustivas. Es decir, las diversas modalidades de un carácter deben cubrir todas las posibilidades que éste puede presentar y, además, deben ser disjuntas (un individuo no puede presentar más de una de ellas y debe presentar alguna de ellas).

Así, al estudiar algún carácter, el investigador deberá considerar todas las posibles modalidades del carácter, con objeto de poder clasificar a todos los individuos que observe.

La matriz de datos

Habitualmente, la información primaria sobre los individuos, es decir, la forma más elemental en la que se expresan los datos es la de una matriz, en la que aparecen en la primera columna los individuos identificados de alguna manera y en las siguientes columnas las observaciones de los diferentes caracteres en estudio para cada uno de los individuos. Dicha matriz recibe el nombre de *matriz de datos*.

Clases de datos

Es habitual denominar a los caracteres *variables estadísticas* o simplemente *variables*, calificándolas de cualitativas o cuantitativas según sea el correspondiente carácter, y hablar de los *valores de la variable* al referirnos a sus modalidades aunque solamente tendremos verdaderos valores numéricos cuando analicemos variables cuantitativas.

En ocasiones, con objeto de facilitar la toma de los datos, el investigador los agrupa en intervalos. Observemos, no obstante, que siempre se producirá una pérdida de información al agrupar los datos en intervalos.

Consideremos, por tanto, tres tipos posibles de datos:

1. Datos correspondientes a un carácter cualitativo.

2. Datos sin agrupar correspondientes a un carácter cuantitativo.

3. Datos agrupados en intervalos correspondientes a un carácter cuantitativo.

Agrupamiento en intervalos

Si los intervalos o *clases*, como a veces se denominan, son:

$$[c_0 - c_1), [c_1 - c_2), ..., [c_{j-1} - c_j), ..., [c_{k-1}, c_k)$$

llamaremos *extremos* de la clase j-ésima a c_{j-1} y a c_j; *amplitud* del intervalo a la diferencia de sus extremos, hablando de intervalos de amplitud constante o variable según tengan o no todos la misma amplitud y, por último, llamaremos *centro* o *marca de clase* correspondiente al intervalo j-ésimo al punto medio del intervalo; es decir, a $c'_j = (c_j + c_{j-1})/2$.

A lo largo del texto consideraremos que el dato c_j pertenece al intervalo $j + 1, j = 1, ..., k - 1$, siendo el c_k del k-ésimo.

Podemos considerar como regla general la de construir, siempre que sea posible, intervalos de amplitud constante, sugiriendo sobre el número k de intervalos a considerar el propuesto por Sturges (1926).

$$k = 1 + 3,322 \log_{10} n$$

siendo n el número total de datos.

Una vez determinado el número k de intervalos a considerar, y si es posible tomarlos de igual amplitud, ésta será

$$c = \frac{x_{(n)} - x_{(1)}}{k}$$

en donde $x_{(n)}$ es el dato mayor y $x_{(1)}$ el menor.

2.3. Distribuciones unidimensionales de frecuencias

En este apartado consideraremos que tenemos datos correspondientes a un solo carácter, el cual llamaremos variable estadística y representaremos por X.

Llamaremos *frecuencia total* al número de datos n. Llamaremos *frecuencia absoluta* n_i de la modalidad M_i (valor x_i o intervalo I_i) de la variable X al número de datos que presentan modalidad M_i (valor x_i o valor intervalo I_i). Si existen k modalidades posibles, se verificará

$$\sum_{i=1}^{k} n_i = n_1 + n_2 + ... + n_k = n.$$

Llamaremos *frecuencia relativa* f_i de la modalidad M_i (valor x_i o intervalo I_i) de la variable X al cociente $f_i = n_i/n$, verificándose,

$$\sum_{i=1}^{k} f_i = f_1 + f_2 + ... + f_k = 1.$$

Llamaremos *frecuencia absoluta acumulada* N_i hasta la modalidad M_i (valor x_i o intervalo I_i) a la suma

$$N_i = n_1 + ... + n_i = \sum_{j=1}^{i} n_j.$$

Claramente es $N_k = \sum_{j=1}^{k} n_j = n$.

Llamaremos *frecuencia relativa acumulada* F_i hasta la modalidad M_i (valor x_i o intervalo I_i) al cociente $F_i = N_i/n$, o lo que es lo mismo, a

$$F_i = f_1 + ... + f_i = \sum_{j=1}^{i} f_i$$

siendo $F_k = \sum_{j=1}^{k} f_j = 1$.

Distribuciones unidimensionales de frecuencias

La tabla formada por las distintas modalidades (valores o intervalos) del carácter X y por las frecuencias absolutas (relativas, absolutas acumuladas o relativas acumuladas) recibe el nomrbe de *distribución de frecuencias absolutas* (*relativas, absolutas acumuladas* o *relativas acumuladas* respectivamente).

2.3.1. Representaciones gráficas de las distribuciones unidimensionales de frecuencias

La representación gráfica de una distribución de frecuencias depende del tipo de datos que la constituya.

Datos correspondientes a un carácter cualitativo

La representación gráfica de este tipo de datos está basada en la proporcionalidad de las áreas a las frecuencias aboslutas o relativas. Veremos dos tipos de representaciones:

Diagrama de sectores:
Esta representación consiste en dividir un círculo en tantos sectores circulares como modalidades presente el carácter cualitativo, asignando un ángulo central a cada sector circular proporcional a la frecuencia absoluta n_i, consiguiendo de esta manera un sector con área proporcional también a n_i.

Ésta representación se obtendrá en R, primero introduciendo los datos en un vector y luego ejecutando la función `pie()`. Si queremos que denomine de una manera concreta a los sectores, debemos crear primero un vector de nombres. También podemos crear un vector de colores y ponerle título al gráfico con el argumento `main`. Obtendriamos

el gráfico deseado al ejecutar

```
x2<-c(datos1, dato2, ...)
n2<-c(nombre1, nombre2, ...)
c2<-c(color1, color2, ...)
pie(x2,labels=n2,col=c2,main="Título")
```

Apuntamos el hecho de que el argumento `col` y el argumento `main` lo son de todas las funciones gráficas de R.

Diagramas de rectángulos:
Esta representación gráfica consiste en construir tantos rectángulos como modalidades presenete el carácter cualitativo en estudio, todos ellos con base de igual amplitud. La altura se toma igual a la frecuencia absoluta o relativa, consiguiendo de esta manera rectángulos con áreas proporcionales a las frecuencias que se quieren representar.

Para obtener dicha representación en R se ejecuta la función `barplot()`, en donde la única variación con respecto a la función `pie()`, es que `labels` no es un argumento de la función sino que el argumento correspondiente para añadir nombres a las clases es `names`.

```
barplot(x2,names=n2,col=c2,main="Título")
```

Datos correspondientes a un carácter cuantitativo agrupado en intervalos

La representación habitual es el *Histograma* en donde sobre cada intervalo se levanta un rectángulo con un área igual a la frecuencia, absoluta o relativa según la distribución que estemos considerando, por lo que hay que tener en cuenta si los intervalos tiene igual o distinta amplitud. La representación gráfica se consigue con la función `hist()` aunque veremos que esta función está pensada para datos sin agrupar. El *Polígono de frecuencias acumuladas* podría realizarse de forma análoga a como se obtendrá la función de distribución empírica en el siguiente apartado.

Para expresar esto en R, primero introducimos en un vector las marcas de clase y en otro las frecuencias absolutas. Con la función `rep()`, replicamos las marcas de clase, tantas veces como sea la frecuencia absoluta del intervalo del que es marca de clase obteniendo así los datos a representar. Finalmente, indicamos cuáles queremos que sean los puntos de corte de los intervalos en la representación gráfica con un vector y los colores con otro vector.

De esta manera, el área del histograma no sumará 1 como habitualmente deseamos. Para conseguir esto, debemos utilizar el argumento `prob=T`. De esta forma quedaría

```
m1<-c(marca1, marca2,...)
n1<-c(frec1,frec2,...)
datos<-rep(m1,n1)
d1<-c(punto1, punto2,...)
c1<-c(color1,color2,...)
hist(datos,breaks=d1,col=c1,prob=T,main="Título")
```

Datos correspondientes a un carácter cuantitativo sin agrupar en intervalos

Las representaciones gráficas habituales serán, si son pocos los valores distintos de la variable, el *Diagrama de barras*, con la misma filosofía del diagrama de rectángulo antes estudiado y, si hay muchos valores distintos, el *Histograma*, o su versión modificada, el *Diagrama de hojas y ramas (steam and leaf plot)*. En el caso de frecuencias acumuladas la representación gráfia será el *Diagrama de frecuencias acumuladas*, denominado *Función de distribución empírica* si las frecuencias acumuladas a representar son relativas.

2.3.2. Medidas de tendencia central de caracteres cuantitativos

Estas medidas reciben el nombre de *promedios, medidas de posición* o *medidas de tendencia central* que, aunque algunas de ellas puedan aplicarse a caracteres cualitativos, habitualmente lo son sobre caracteres cuantitativos.

Medida aritmética

Llamando $x_1, ..., x_k$ a los datos distintos de un carácter cuantitativo en estudio, o las marcas de clase de los intervalos en los que se han agrupado dichos datos, y $n_1, ..., n_k$ a las correspondientes frecuencias absolutas de dichos valores o marcas de clase, llamaremos *media artimética* de la distribución de frecuencias al valor

$$a = \frac{\sum_{i=1}^{k} x_i n_i}{n}$$

en donde n es la frecuencia total.

Mediana

La *mediana* es otra medida de posición, la cual se define como aquel valor de la variable tal que, supuestos ordenados los valores de ésta en orden creciente, la mitad son menores o iguales y la otra mitad mayores o iguales.

Datos sin agrupar:
Si $N_{j-1} < \frac{n}{2} < N_j$ entonces la mediana es

$$M_e = c_j$$

Si la situación es $N_{j-1} = \frac{n}{2} < N_j$ entonces la mediana es

$$M_e = \frac{c_{j-1} + c_j}{2}$$

Datos agrupados:
Cuando existe una frecuencia absoluta acumulada N_j tal que $n/2 = N_j$, la mediana es

$$M_e = c_j$$

Si la situación es $N_{j-1} < \frac{n}{2} < N_j$, entonces la mediana está en el intervalo $[c_{j-1}, c_j)$, tomándose en ese caso, por razonamientos de proporcionalidad, como mediana el valor

$$M_e = c_{j-1} + \frac{\frac{n}{2} - N_{j-1}}{n_j} \cdot a_j$$

siendo a_j la amplitud del intervalo $[c_{j-1}, c_j)$.

Moda

La *moda* se define como aquel valor de la variable al que corresponde máxima frecuencia (absoluta o relativa). Para calcularla, también será necesario distinguir si los datos están o no agrupados.

Datos sin agrupar:
Para datos sin agrupar, la determinación del valor o valores modales es muy sencilla. Basta observar a que valor le corresponde una mayor n_i. Ése será la moda.

Datos agrupados:
Si los datos se presentan agrupado en intervalos es necesario, a su vez, distinguir si éstos tienen o no igual amplitud.

Si tienen amplitud constante c, una vez identificado el intervalo modal $[c_{j-1}, c_j)$, es decir el intervalo al que corresponde mayor frecuencia absoluta $n_j = \text{máx}\{n_1, ..., n_k\}$, la moda se define, también por razones geométricas, como

$$M_d = c_{j-1} + \frac{n_{j+1}}{n_{j-1} + n_{j+1}} \cdot c$$

Si los intervalos tuvieran distinta amplitud a_j, primero debemos *estandarizar* las frecuencias absolutas n_j, determinando los cocientes $l_j = \frac{n_j}{a_j}$ para $j = 1, ..., k$ y luego aplicar la regla definida poara el caso de intervalos de amplitud constante a los l_j. Es decir, primero calcular el $l_j = \text{máx}\{l_1, ..., l_k\}$ para determinar el intervalo modal $[c_{j-1}, c_j)$ y luego aplicar la fórmula

$$M_d = c_{j-1} + \frac{l_{j+1}}{l_{j-1} + l_{j+1}} \cdot a_j$$

siendo a_j la amplitud del intervalo modal $[c_{j-1}, c_j)$.

Cuantiles

El *cuantil* $p_{r/k}, r = 1, 2, ..., k-1$ se define como aquel valor de la variable que divide la distribución de frecuencias, previamente ordenada de forma creciente, en dos partes, estando el $(100 \cdot r/k)\,\%$ de ésta formado por valores menores que $p_{r/k}$.

Si $k = 4$ los (tres) cuantiles reciben el nombre de *cuartiles*. Si $k = 10$ los (nueve) cuantiles reciben el nombre de *deciles*. Por último, si $k = 100$ los (noventa y nueve) cuantiles reciben el nombre de *centiles*.

Obsérvese que siempre que r y k mantengan la misma proporción (r/k) obtendremos el mismo valor. En este sentido, la mediana M_e es el segundo cuartil, o el quinto decil, etc.

Para el cálculo de los cuantiles de nuevo hay que considerar si los datos vienen o no agrupados en intervalos.

Datos sin agrupar:
Si los datos vienen sin agrupar y es $N_{j-1} < \frac{r}{k} \cdot n < N_j$ el r-ésimo cuantil de orden k será

$$p_{r/k} = c_j$$

valor al que corresponde la frecuencia absoluta acumulada N_j.

Si la situación fuera de la forma $N_{j-1} = \frac{r}{k} \cdot n < N_j$ tomaríamos, en esta situación indeterminada,

$$p_{r/k} = \frac{c_{j-1} + c_j}{2}$$

Datos agrupados:
Si los datos se presentan agrupados y, para algún j, fuera $\frac{r}{k} \cdot n = N_j$ el r-ésimo cuantil de orden k sería

$$p_{r/k} = c_j$$

Por último, si fuera $N_{j-1} < \frac{r}{k} \cdot n < N_j$ el intervalo a considerar sería el $[c_{j-1}, c_j)$, al que corresponde frecuencia absoluta n_j y absoluta acumulada N_j, siendo entonces el cuantil el dado por la expresión

$$p_{r/k} = c_{j-1} + \frac{\frac{r}{k} \cdot n - N_{j-1}}{n_j} \cdot a_j$$

para $r = 1, ..., k-1$ en donde a_j es la amplitud del intervalo $[c_{j-1}, c_j)$.

Si el intervalo a considerar fuera el primero $[c_0, c_1)$, se tomaría en la expresión anterior $N_{j-1} = 0$.

2.3.3. Medidas de dispersión

Las *medidas de dispersión* tienen como propósito estudiar lo concentrada que está la distribución en torno a algún promedio.

Recorrido

Si $x_{\text{máx}}$ (también representado por $x_{(n)}$) es el dato mayor, o la última marca de clase si es que los datos vienen agrupados en intervalos, y $x_{\text{mín}}$ (ó $x_{(1)}$) el dato menor, o primera marca de clase, llamaremos *Recorrido* a

$$R = x_{\text{máx}} - x_{\text{mín}}$$

La principal ventaja del recorrido es la de proporcionar una medida de la dispersión de los datos fácil y rápida de calcular. A veces se utiliza también el *Recorrido intercuartílico*, definido como la diferencia entre el tercer y el primer cuartil.

Varianza

Denotando de nuevo por $x_1, ..., x_k$ los datos o las marcas de clase, llamaremos *Varianza* a

$$s^2 = \frac{1}{n} \sum_{i=1}^{k} (x_i - a)^2 n_i = \frac{1}{n} \sum_{i=1}^{k} x_i^2 n_i - a^2$$

siendo a la media aritmética de la distribución.

Al valor

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{k} (x_i - a)^2 n_i = \frac{n \cdot s^2}{n-1}$$

se le denomina *cuasivarianza*.

Desviación típica

La varianz tiene un problema, y es que está expresada en unidades al cuadrado. Esto puede producir una falsa imagen de la dispersión de la distribución. En su lugar suele utilizarse su raíz cuadrada, denominada *Desviación típica.*

Coeficiente de variación de Pearson

La desviación típica sirve para medir de forma eficaz la dispersión de un conjunto de datos entorno a su media. Desgraciadamente, esta medida puede resultar engañosa cuando tratamos de comparar la dispersión de dos conjuntos de datos. Podemos observar una misma desviación típica en dos grupos diferentes y podría parecernos que los datos tienen la misma dispersión sin ser necesariamente así. El *Coeficiente de variación de Pearson* elimina esa posible confusión al ser una medida de la variación de datos pero en relación con su media (supuestamente mayor que cero). Se define como

$$V_p = \frac{s}{a} \cdot 100$$

siendo s y a respectivamente la desviación típica y la media aritmética de la distribución en estudio y donde el factor 100 tiene como único objetivo el evitar operar con valores decimales.

De la definición de V_p se deduce fácilmente que aquella distribución a la que corresponda mayor coeficiente tendrá mayor dispersión.

2.3.4. Medidas de asimetría

Diremos que una distribución es *simétrica* cuando su mediana, su moda y su media aritmética coincidan.

Diremos que una distribución es *asimétrica a la derecha* si las frecuencias (absolutas o relativas) descienden más lentamente por la derecha que por la izquierda.

Si las frecuencias descienden más lentamente por la izquierda que por la derecha diremos que la distribución es *asimétrica a la izquierda.*

Coeficiente de asimetría de Pearson

El *coeficiente de asimetría de Pearson* se define como

$$A_p = \frac{a - M_d}{s}$$

siendo cero cuando la distribución es simétrica, positivo cuando existe asimetría a la derecha y negativo cuando existe asimetría a la izquierda.

De la definición se observa que este coeficiente sólo se podrá utilizar cuando la distribución sea unimodal.

Coeficiente de asimetría de Fisher

El *coeficiente de asimetría de Fisher* se define como

$$A_f = \frac{\sum_{i=1}^{k}(x_i - a)^3 \cdot n_i}{n \cdot S^3}$$

siendo x_i los valores de la variable o las marcas de clase y $S = \sqrt{S^2}$, llamada a veces cuasideviación típica.

La interpretación del coeficiente de Fisher es semejante a la del coeficiente de Pearson: se dice que la distribución es simétrica cuando vale cero, siendo el coeficiente positivo o negativo cuando exista asimetría a la derecha o izquierda respectivamente.

2.3.5. Medidas de posición y dispersión con R

Las medidas de posición y dispersión son la *Media*, obtenida con la función `mean()`; la *Mediana*, cuyo valor obtenemos con `median()`; la *Cuasivarianza* (no la varianza) para la que debemos ejecutar la función `var()`; su raíz cuadrada, la *Cuasidesviación típica*, obtenida con `sd()`, y los cuantiles, que se consiguen con `cuantile()`.

Un buen resumen de muchas de las medidas de posición se obtiene de una vez con la función `summary()`.

Un gráfico con el que podemos visualizar la dispersión y simetría de los datos es el *Diagrama de cajas* ejecutado por la función de R `boxplot()`. Consiste en representar una caja en donde el lado inferior sea el primer cuartil, el superior el tercer cuartil apareciendo dividida la caja por la mediana de los datos. Se añaden dos segmentos a la caja así formada para unirla al máximo y mínimo valor. Aquellos datos inferiores al primer cuartil menos 1,5 veces el recorrido intercuartílico, o superiores al tercer cuartil más 1,5 veces el recorrido intercuartílico se consideran anómalos y se representan por pequeños círculos fuera del diagrama de cajas.

2.4. Distribuciones bidimensionales de frecuencias

En esta sección estudiaremos la situación en la que los datos son observaciones de dos caracteres efectuadas en los individuos de una determinada población. Ambos caracteres pueden ser cuantitativos, cualitativos o uno de cada.

En estas situaciones los datos se recogen en lo que de forma genérica se denomina una *Tabla de doble entrada* o *Tabla de contingencia*, cuya expresión general es la siguiente:

Carácter$B$ / Carácter$A$	B_1	...	B_j	...	B_k	
A_1	n_{11}	...	n_{1j}	...	n_{1k}	$n_{1.}$
...	...	...	...	...	...	...
A_i	n_{i1}	...	n_{ij}	...	n_{ik}	$n_{i.}$
...	...	...	...	...	...	...
A_l	n_{l1}	...	n_{lj}	...	n_{lk}	$n_{l.}$
	$n_{,1}$	...	$n_{.j}$	...	$n_{.k}$	n

en donde n_{ij}, denominada *frecuencia absoluta* del par (A_i, B_j), respresenta el número de individuos, de entre los n, que poseen a la vez la modalidad A_i del carácter A y la modalidad B_j del carácter B.

Ésta es la forma habitual en la que se presentan los datos bidimensionales, aunque si dividimos las n_{ij} por n obtendremos una distribución bidimensional de frecuencias

relativas f_{ij}, en donde sería

$$\sum_{i=1}^{l}\sum_{j=1}^{k} f_{ij} = 1.$$

Distribuciones marginales

Las tablas formadas con la primera y última columnas y con la primera y última filas de la tabla anterior

Carácter$A$	frec. absol.	Carácter$B$	frec. absol.
A_1	$n_{1.}$	B_1	$n_{,1}$
...	...	...	...
A_i	$n_{i.}$	B_j	$n_{.j}$
...	...	...	...
A_l	$n_{l.}$	B_k	$n_{.k}$
	n		n

se denominan *distribuciones marginales* (absolutas), respectivamente, de los caracteres A y B.

Cunado existan más de dos variables en consideración también será posible agrupar un número m de ellas formando distribuciones marginales m-dimensionales. También podrían hacerse grupos de variables para conseguir distribuciones de frecuencias condicionadas *multivariantes*.

Distribuciones condicionadas

En general, la distribución (de frecuencias absolutas) condicionada del carácter A por la modalidad B_j del carácter B será,

A/B_j	
A_1	n_{1j}
...	...
A_i	n_{ij}
...	...
A_l	n_{lj}
	$n_{.j}$

y la del carácter B por la modalidad A_i del carácter A será,

B/A_i	
B_1	n_{i1}
...	...
B_j	n_{ij}
...	...
B_k	n_{ik}
	$n_{i.}$

existiendo $k+l$ distribuciones condicionadas, si es que los caracteres B y A presentan, respectivamente, k y l modalidades cada uno.

2.4.1. Representaciones gráficas de las distribuciones bidimensionales de frecuencias

Las distribuciones marginales y condicionadas son distribuciones de frecuencias unidimensionales, y por tanto, su representación gráfica se ajustará a los desarrollado en la sección 2.3.1.

Por otro lado, de las distribuciones bidimensionales sólo consideraremos representaciones gráficas en el caso de que ambos caracteres sean cuantitativos.

Datos agrupados en intervalos correspondientes a un carácter cuantitativo

La representación habitual de este tipo de datos es un *Histograma tridimensional*, el cual se construye utilizando los mismos criterios que el histograma visto anteriormente.

Se utiliza un sistema de ejes coordenados en tres dimensiones, en donde los dos primeros ejes se reservan para las dos variables, representándose en altura la frecuencia, absoluta o relativa según la distribución que estemos representando.

Esto en el caso de que los intervalos de ambas variables tengan igual amplitud; si no, las alturas de los paralelepípedos deberán ser tales que su volumen resultante sea igual a la frecuencia.

Datos sin agrupar correspondientes a un carácter cuantitativo

La representación gráfica, denominada *Diagrama de barras tridimensional*, se hace utilizando también un sistema de ejes coordenados en tres dimensiones, levantando en cada par de valores (x_i, y_i) de la variable bidimensional (X, Y), una barra de altura igual a su frecuencia (absoluta o relativa).

No obstante, si no existen pares de valores repetidos, suele utilizarse el denominado diagrama de dispersión o *nube de puntos*, el cual consiste en representar en un sistema de ejes coordenados de dos dimensiones tantos puntos como datos, asignando a cada dato (x_i, y_i) el punto de coordenadas (x_i, y_i).

En R, se obtiene el diagrama de dispersión usando la función `plot()`, con los argumentos `main="Título"` para el título, `xlim=c(inf,sup)` y `ylim=c(inf,sup)` para limitar el recorrido del gráfico, `pch=çarácter"` para cambiar los puntos por un carácter, `pch=número` para cambiar el símbolo de los puntos (hay del 0 al 18), `xlab="nombre"` y `ylab="nombre"` para poner nombre a los ejes y por último `axes=F` para no poner el marco al gráfico.

2.4.2. Ajuste por mínimos cuadrados

Una relación entre datos no es *funcional* en el sentido de que no se puede determinar una fórmula exacta que nos dé la relación de un dato en función del otro aunque exista unos determinados valores entre los que razonablemente debería estar un dato en función del otro.

Aunque no exista tal ecuación si fueramos capaces de determinar una recta

$$y_{t_i} = \beta_0 + \beta_1 x_i$$

próxima a la nube de puntos, los valores y_i de un individuo debería estar altededor del valor que nos dé la recta y_{t_i} para x_i.

Éste es el objetivo de ésta sección: determinar la ecuación de una recta $y_{t_i} = \beta_0 + \beta_1 x_i$, lo más próxima posible a una nube de puntos $(x_1, y_1), ..., (x_n, y_n)$ en el sentido de *mínimos cuadrados*, es decir, determinar los valores de β_0 y β_1 que hagan mínima la suma de los cuadrados de las

desviaciones e_i entre los valores observados y_i y los teóricos dados por la recta y_{t_i}

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - y_{t_i})^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

Matemáticamente este problema se resuelve considerando la euación anterior como una función de β_0, derivando respecto a β_0 e igualando a cero dicha ecuación. A continuación, se considera la ecuación de la suma de los cuadrados como una función de β_1, se deriva respecto a β_1 y se iguala a cero esta ecuación. Se obtiene así un sistema de ecuaciones con dos incógnitas de donde despejamos y obtenemos los valores para β_1 y β_0,

$$\beta_1 = \frac{n\sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

y

$$\beta_0 = \frac{\sum_{i=1}^{n} y_i - \beta_1 \sum_{i=1}^{n} x_i}{n}$$

En R, obtenemos dicha recta mediante la función `lm(y x)` y usando `abline()` añadimos la recta de regresión a la nube de puntos, a esta última función se le pueden añadir losa argumentos `col=número` para cambiar el color de la recta. También podemos usar `abline(número1, número2, lty=número, col=número)` para añadir una recta donde `número1` es la ordenada en origen, `número2` es la pendiente y los argumentos `lty` y `col` son, respectivamente, el grosor y el color de la linea. Finalmente, con `legen(número1,número2, c("nombre"))` añadimos un rótulo, "nombre", en las coordenadas (número1, número2).

2.4.3. Precisión del ajuste por mínimos cuadrados

Diferentes nubes de puntos pueden parecer menos concentradas alrededor de su recta de ajsute que otras, la causa de esa falta de concentración puede ser que am bas variables no estén relacionadas linealmente.

Es probable que para cierto tipo de datos se ajuste *mejor* una función de tipo *exponencial* de la forma

$$y_{t_i} = a \cdot b^{x_i}$$

con $b < 1$. Es decir, se ajustase mejor a los datos

$$\{(x_i, \log y_i) : i = 1, ..., n\}$$

una recta de la forma

$$\log y_{t_i} = A + Bx_i$$

con pendiente $B = \log b$ negativa (y ordenada en el origen $A = \log a$).Valores que se obtendrán por las mismas expresiones que antes,

$$B = \log b = \frac{n\sum_{i=1}^{n}(x_i \log y_i) - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} \log y_i)}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

y

$$A = \log a = \frac{\sum_{i=1}^{n} \log y_i - B\sum_{i=1}^{n} x_i}{n}$$

obteniéndose trivialmente los valores de a y b a partir de los de A y B por las expresiones

$$a = \exp\{A\} \quad \text{y} \quad b = \exp\{B\}$$

Es decir, que para algunos datos no siempre será una recta la mejor función a determionar. En ocasiones será una *función exponencial*

$$y_{t_i} = a \cdot x_i^b$$

la adecuada, lo que llevará a ajustar una recta

$$\log x_{t_i} = \log a + b \log x_i$$

a los datos $\{(\log x_i, \log y_i) : i = 1, ..., n\}$.

Otras veces será necesario utilizar una *parábola*, o en general un *polinomio de grado n*, para conseguir un buen ajuste.

Necesitamos un valor que nos dé una medida de lo próxima que está la función que hemos ajustado a la nube de puntos de los datos; es decir, una medida de la *bondad del ajuste*.

Como el criterio que hemos utlizado para ajustar una función a la nube de puntos ha sido el de mínimos cuadrados, es decir, el de elegir como valores para los parámetros que definen la función f aquellos que minimicen la suma de cuadrados de las desviaciones

$$\sum_{i=1}^{n} (y_i - y_{t_i})^2$$

parece razonable que una vez determinados dichos parámetros, calculemos cuánto vale dicha suma de cuadrados para cada una de las funciones determinadas, eligiendo aquella para la que se obtenga un menor valor.

Este valor recibe el nombre de *Varianza residual*

$$V_r = \frac{1}{n}\sum_{i=1}^{n} (y_i - y_{t_i})^2.$$

La función óptima sería, en principio, aquella para la que su varianza residual fuera cero; es decir, aquella que pasara por todos los puntos y_i.

No obstante, si esto se consigue utilizzando una función muy complicada el ajuste se considera inadecuado porque es preferible poder explicar el fenómeno en estudio con funciones lo más simples posibles.

Es conveniente usar otro valor que permita decidir si el ajuste es o no adecuado en sí mismo.

Surge así el concepto de *Coeficiente de determinación* definido como

$$R^2 = 1 - \frac{V_r}{s_y^2}$$

siendo V_r la varianza residual y $s_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - a_y)^2$ la varianza (marginal) de las y_i.

Este coeficiente está comprendido entre 0 y 1, hablándose de un buen ajuste cuando R^2 esté cerca de 1, y de un mal ajuste cuando sea cercano a 0.

Por último, en el caso de que se ajuste una recta, tenemos el *Coeficiente de correlación* de Pearson, definido como

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{\sqrt{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2}}$$

para el caso en que los n pares de datos vengan aislados, y que en el caso que éstos aparezcan en forma de distribución bidimensional de frecuencias, la fórmula anterior resulta ser igual a

$$r = \frac{\sum_{i=1}^{l}\sum_{j=1}^{k}(x_i - a_x)(y_j - a_y)n_{ij}/n}{\sqrt{\frac{1}{n}\sum_{i=1}^{l}(x_i - a_x)^2 n_{i.}}\sqrt{\frac{1}{n}\sum_{j=1}^{k}(y_j - a_y)^2 n_{.j}}}$$

en donde a_x y a_y son las medias marginales.

Este coeficiente toma valores entre -1 y 1. Estos dos valores extremos indicarían una relación funcional entre los valores de X e Y y un valor igual a 0 indicaría que, mediante la recta de mínimos cuadrados no vamos a poder *explicar* adecuadamente a la variable Y en función de X.

Si es $r > 0$, a medida que aumentemos los valores de las X aumentarán los de las Y, hablándose de *correlación positiva*, utilizando la expresión *correlación negativa* para cuando es $r < 0$.

Si se ha realizado el ajuste de una recta

$$y_{t_i} = \beta_0 + \beta_1 x_i$$

el coeficiente de determinación es igual al cuadrado del coeficiente de correlación, el cual se podrá calcular en este caso por la expresión

$$R^2 = (r)^2 = \frac{\beta_1^2\left(\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2/n\right)}{\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2/n}$$

El coeficiente de correlación se obtiene en R ejecutando la función `cor(x,y)`, donde `x` e `y` son los vectores con los datos.

3. Probabilidad

3.1. Introducción

La probabilidad que habitualmente manejaremos en estadística vendrá ligada a un modelo probabilístico el cual suponemos rige nuestro fenómeno aleatorio en estudio.

Vamos a denominar *Nivel I* al nivel donde están los elementos básicos: el conjunto de individuos analizados, que denominaremos *Espacio muestral*, Ω, y la *Probabilidad* P con la que estos van a ser seleccionados para dar lugar a la muestra. P va a estar definida no sólo sobre Ω sino sobre el conjunto de todos los subconjuntos posibles de Ω, este conjunto sobre el que va estar definida P, lo denominaremos *Espacio de sucesos* $\mathcal{A}$.

Tenemos, por lo tanto, los tres elementos básicos del cálculo de probabilidades: $(\Omega, \mathcal{A}, P)$, que reciben el nombre de *Espacio probabilístico*.

3.2. Espacio muestral

El conjunto de todos los resultados posibles diferentes de un determinado experimento aleatorio se denomina *Espacio muestral* asociado a dicho experimento y se suele representar por Ω. A los elementos de Ω se les denomina *sucesos elementales*. El espacio muestras asociado a la selección aleatoria de individuos de una población determinada, será el conjunto de personas de esa población y, los sucesos elementales, dichos individuos.

Si $\mathcal{A}$ el conjunto de las partes de Ω, es decir, el conjunto de todos los subconjuntos de Ω (también denominado σ-álgebra), cualquier elemento de $\mathcal{A}$, contendrá una cierta incertidumbre, por lo que trataremos de asignarle un número entre 0 y 1 como medida de su incertidumbre. En cálculo de probabilidades dichos conjuntos reciben el nombre de *sucesos*, siendo la medida de la incertidumbre su probabilidad.

Por tanto, asociado a todo experimento aleatorio existen tres conjuntos: El espacio muestral Ω, la clase de sucesos, es decir, el conjunto de los elementos con incertidumbre asociados a nuestro experimento aleatorio $\mathcal{A}$, y la función $P : \mathcal{A} \mapsto [0, 1]$, la cual asignará a cada suceso (elemento de $\mathcal{A}$) un número entre 0 y 1 como medida de su incertidumbre.

La elección del espacio muestral asociado a un experimento aleatorio no tiene por qué ser única, sino que dependerá de qué sucesos elementales queramos considerar como distintos y del problema de la asignación de la probabilidad sobre esos sucesos elementales.

Cuando Ω es finito o numerable, la clase $\mathcal{A}$ es cerrada para las operaciones entre çonjuntos"(entre sucesos) como unión, intersección, complementario, etc. En otras ocasiones en las que Ω sea un conjunto continuo, deberá ser $\mathcal{A}$ un conjunto estrictamente más pequeño que el conjunto de las partes de Ω.

En todo caso podemos pensar en $\mathcal{A}$ como en el conjunto que contiene todos los elementos de interés, es decir, todos los sucesos a los que les corresponda una probabilidad.

Apuntemos algunas peculiaridades del cálculo de probabilidades respectoa la teoría de conjuntos. El conjunto vacío $\emptyset$ recibe el nombre de *suceso imposible*, definido como aquel subconjunto de Ω que no contiene ningún suceso elemental y que corresponde a la idea de aquel suceso que no puede ocurrir. De forma análoga, el espacio total Ω recibe el nombre de *suceso seguro* al recoger dicha denominación la idea que representa. LLamaremos *sucesos incompatibles* a aquellos cuya intersección sea el suceso imposible. Por último, digamos que la inclusión de sucesos, $A \subset B$, se interpreta aquí como que siempre que se cumpla el suces A se cumple el suceso B.

3.3. Conceptos de probabilidad

Concepto frecuentista

Es un hecho, empíricamente comprobado, que la frecuencia relativa de un suceso tiende a estabilizarse cuando la frecuencia total aumenta.

Surge así el *concepto frecuentista* de la probabilidad de un suceso como un número ideal al que converge su frecuencia relativa cuando la frecuencia total tiende a infinito.

El problema radica en que, al no poder repetir la experiencia infinitas veces, la probabilidad de un suceso ha de ser aproximada por su frecuencia relativa para un n suficientemente grande, y ¿cuán grande es un n grande?¿qué hacer con aquellas experiencias que sólo se pueden repetir una vez?

Concepto clásico

Está basado en el concepto de resultados igualmente verosímiles y motivado por el denominado *Principio de la razón insuficiente*, el cual postula que si no existe un fundamento para preferir una, entre varias posibilidades, todas deber ser consideradas equiprobables.

Laplace recogió esta idea y formuló la regla clásica del cociente entre casos favorables y casos posibles, supuestos éstos igualmente verosímiles.

El problema aquí surge porque en definitiva *igualmente verosímil* es lo mismo que *igualmente probable*, es decir, se justifica la premisa con el resultado.

Concepto subjetivo

Se basa en la idea de que la probabilidad que una persona dé a un suceso debe depender de su juicio y experiencia personal, pudiendo dar dos personas distintas probabilidades diferentes a un mismo suceso.

El principal problema a que da lugar esta definición es que dos personas diferentes pueden dar probabilidades diferentes a un mismo suceso.

Definición formal de probabilidad

Llamaremos *Probabilidad* a una aplicación

$$P : \mathcal{A} \mapsto [0, 1]$$

tal que

- *Axioma 1:* Para todo suceso A de $\mathcal{A}$ sea $0 \leqslant P(A)$.
- *Axioma 2:* Sea $P(\Omega) = 1$.
- *Axioma 3:* Para toda colección de sucesos incompatibles, $\{A_i\}$ con $A_i \cap A_j = \emptyset, i \neq j$ debe ser

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

Obsérvese que esta definición no dice cómo asignar las probabilidades ni siquiera a los sucesos elementales. Sólo dice que cualquier asignación que hagamos debe verificar estos tres axiomas para que pueda llamarse Probabilidad.

3.4. Propiedades elementales de la probabilidad

Toda probabilidad cumple una seria de propiedades, las cuales se obtienen como consecuencia de los axiomas que debe cumplir. Veamos las más importantes:

a) $P(\emptyset) = 0$.

b) Se cumple la *aditividad finita* para sucesos incompatibles. Es decir,

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

si $A_i \cap A_j = \emptyset, i \neq j$.

c) La probabilidad del complementario de un suceso A es

$$P(A*) = 1 - P(A).$$

d) Si dos sucesos son tales que $A \subset B$, entonces es $P(A) \leqslant P(B)$.

e) Si dos sucesos no son incompatibles, la probabilidad de su unión debe calcularse por la siguiente regla:

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

3.5. Asignación de probabilidad en espacios muestrales discretos

Por las propiedades anteriores es suficiente conocer la probabilidad de los sucesos elementales ya que, entonces, se podrá determinar la de cualquier otro suceso.

Es decir, el problema radica en *asignar* una probabilidad a los sucesos elementales: asignar un número entre 0 y 1 a cada uno de los sucesos elementales, de tal forma que su suma sea 1.

En principio, cualquier asignación que cumpla los tres axiomas mencionados en la definición de probabilidad es válida. No obstante, el propósito del cálculo de probabilidades, como soporte de la estadística, es el de construir un esquema matemático que refleje de la forma más exacta posible el fenómeno aleatorio real que estemos estudiando, por lo que la asignación de probabilidades que hagamos debe ser lo más ajustada posible a la realidad que estamos observando.

En otras ocasiones, la observación del mismo fenómeno en otra población semejante a la que estamos estudiando, o inclusive en la objeto de estudio en un tiempo anterior, permitirá obtener una distribución de frecuencias a partir de la cual asingar una probabilidad.

A veces es precisamente la asiganción de la probabilidad la que determina el espacio muestral en el sentido en que si consideramos un espacio muestral en donde todos los sucesos se puedan considerar equiprobables tendremos una sitación mucho más sencilla que si consideramos otro espacio muestral donde los sucesos dejen de ser equiprobables.

3.6. Modelo uniforme

Dentro de las posibles asignaciones de probabilidad el *Modelo uniforme* destaca por ser de las más utilizadas y por obtenerse de ella interesantes propiedades. En esta sección estudiaremos un caso particular el cual corresponde con una situación en la que los sucesos elementales dels espacio muetral (que suponemos finito) puedan ser considerados como equiprobables.

En todos estos casos de modelos uniformes, en especial aquellos que el espacio muestral es finito, $\Omega = \{\omega_1, \omega_2, ..., \omega_n\}$, al ser los sucesos elementales incompatibles y equiprobables, será

$$1 = P(\Omega) = P(\omega_1) + ... + P(\omega_n) = n \cdot P(\omega_i)$$

con lo que $P(\omega_i) = 1/n, \forall i = 1, ..., n$. Por tanto, si un suceso A es unión de k sucesos elementales, será

$$P(A) = \frac{k}{n} = \frac{\text{casos favorables a } A}{\text{casos posibles}}$$

Si de un grupo de N elementos tomamos n, y nos importa el orden de los n elementos seleccionados, tendremos *variaciones* y si no nos importa el orden, tendremos *combinaciones*. Además, si admitimos la posibilidad de que entre estos n pueda haber elementos repetidos, hablaremos, respectivamente, de *variaciones* y de *combinaciones con repetición*.

Por último, si solamente queremos contar el número posible de reordenaciones de un conjunto de elementos, hablaremos de *permutaciones con* o *sin repetición*.

Las fórmulas son:

Variaciones de N elementos tomados de n en n

$$V_{N,n} = N \cdot (N-1) \cdot ... \cdot (N-n+1) = \frac{N!}{(N-n)!}$$

Variaciones con repetición de N elementos tomados de n en n

$$RV_{N,n} = N^n$$

Combinaciones de N elementos tomados de n en n

$$C_{N,n} = \binom{N}{n} = \frac{N!}{n!(N-n)!}$$

Combinaciones con repetición de N elementos tomados de n en n

$$RC_{N,n} = \binom{N+n-1}{n} = \binom{N+n-1}{N-1} = \frac{(N+n-1)!}{n!(N-1)!}$$

Permutaciones de N elementos

$$P_N = N! = N \cdot (N-1) \cdot ... \cdot 2 \cdot 1$$

Permutaciones con repetición de N elementos, uno de los cuales se repite n_1 veces, otro n_2 veces, ..., otro n_r veces

$$RP_N^{n_1,...,n_r} = \frac{N!}{n_1! \cdot n_2! \cdot ... \cdot n_r!}$$

3.7. Probabilidad condicionada

Mediante un espacio probabilístico damos una información matemática a un fenómeno aleatorio que estamso observando. Parece por lo tanto razonable que si observamos algo que aporta información a nuestro fenómeno aleatorio, esto deba alterar el espacio probabilístico de partida.

Es decir, en el nuevo espacio probabilístico deberá hablarse de probabilidad *condicionada* por el suceso A, de forma que se recojan hechos tan evidentes como que ahora la probabilidad (condicionada) de obtener un suceso, el que sea, se habrá visto afectada por el hecho de haberse observado ya un suceso A.

Definición

Dado un espacio probabilístico $(\Omega, \mathcal{A}, P)$ y un suceso $B \in \mathcal{A}$ tal que $P(B) > 0$, llamaremos *probabilidad condicionada* del suceso A por el suceso B a

$$P(A/B) = \frac{P(A \cap B)}{P(B)}.$$

A partir de esta definición podemos deducir que

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

y como los sucesos A y B pueden intercambiarse en la expresión anterior, será (lógicamente si es $P(A) > 0$),

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

por lo que tenemos una expresión más para calcular la probabilidad condicionada

$$P(A/B) = \frac{P(B/A) \cdot P(A)}{P(B)}.$$

3.8. Independencia de sucesos

Existen situaciones en las que la información suministrada por la ocurrencia de un suceso B no altera para nada el cálculo de la probabilidad de otro suceso A. Son aquellas en las que el suceso A es *independiente* de B. Es decir, cuando

$$P(A/B) = P(A).$$

Como entonces, por la útlima expresión de la probabilidad condicionada, es

$$P(B/A) = \frac{P(A) \cdot P(B)}{P(A)} = P(B)$$

y, por tanto, se podría decir que también B lo es de A, hablaremos de *sucesos independientes* cuando esta situación ocurra.

Definición

Dos sucesos A y B de un mismo espacio probabilístico $(\Omega, \mathcal{A}, P)$ se dicen independientes cuando

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

3.9. Teorema de la probabilidad total

Teorema 3.1

Sea un espacio probabilístico $(\Omega, \mathcal{A}, P)$ y $\{A_n\} \subset \mathcal{A}$ una partición de sucesos de Ω. Es decir,

$$\bigcup_n A_n = \Omega \quad \text{y} \quad A_i \cap A_j = \emptyset \quad \forall i \neq j.$$

Entonces, para todo suceso $B \in \mathcal{A}$ es

$$P(B) = \sum_n P(B/A_n) \cdot P(A_n).$$

Resultado que se puede parafrasear diciendo que la probabilidad de un suceso que se puede dar de varias maneras es igual a la suma de los productos de las probabilidades de éste en cada una de las maneras, $P(B/A_n)$, por las probabilidades de que se den estas maneras, $P(A_n)$.

3.10. Teorema de Bayes

Teorema 3.2

Sea un espacio probabilístico $(\Omega, \mathcal{A}, P), \{A_n\} \subset \mathcal{A}$ una partición de sucesos de Ω y $B \in \mathcal{A}$ un suceso con probabilidad positiva. Entonces, para todo suceso A_i es

$$P(A_i/B) = \frac{P(A_i) \cdot P(B/A_i)}{\sum_n P(A_n) \cdot P(B/A_n)}.$$

Si las cosas que pueden ocurrir las tenemos clasificadas en los sucesos A_i de los cuales conocemos sus probabilidades $P(A_i)$, denominadas *a priori*, y se observa un suceso B, la fórmula de Bayes nos da las probabilidades *a posteriori* de los sucesos A_i, ajustadas o modificadas por B.

4. Modelos probabilísticos

4.1. Introducción

La situación que el investigador tiene planteada habitualmente es la de analizar una determinada variable X en los individuos de una población. Ésta puede ser *unidimensional* o *multidimensional*.

Más en concreto, el investigador tendrá el propósito de estudiar alguna característica relacionada con dicha variable, para ello usará técnicas inferenciales cuyos resultados dependerán de la *distribución de probabilidad* o *modelo probabilístico* supuesto como ley que rige el fenómeno aleatorio en estudio.

4.2. Distribución de probabilidad

La selección aleatoria de los individuos de una población puede formalizarse mediante un espacio probabilístico $(\Omega, \mathcal{A}, P)$ en el que el espacio muestral esté contenido por los individuos de la población y tal que sobre el conjunto $\mathcal{A}$ de los sucesos esté definida una probabilidad P, de forma que todos los sucesos elementales sean equiprobables: Modelo Uniforme.

Habitualmente, estaremos interesados no en el espacio probabilístico, sino en una transformación suya, tal que no sólo nos dé los valores de la característica en estudio para los individuos de la población

$$X : \Omega \to \mathbb{R}$$

sino que conserve la probabilidad P, aglutinando la nueva P_X las probabilidades de los sucesos elementales ω_i a los que corresponda el mismo valor mediante X,

$$P_X(A) = P\{\omega \in \Omega : X(\omega) \in A\} = P\{X^{-1}(A)\}$$

La función X recibe el nombre de *variable aleatoria* y P_X el de su *distribución de probabilidad*.

Cuando se consideran a la vez varia variables aleatorias, $X_1, ..., X_p$, de forma que en los individuos de la población se observan varios caracteres, queda constituido lo que se denomina una *variable aleatoria multidimensional*, o *vector aleatorio* $X = (X_1, ..., X_p)$.

Asociada a toda variable aleatoria existe una función $F(x)$, denominada *función de distribución* de X, la cual va midiendo la probabilidad *acumulada* por X hasta el punto x. Es decir

$$F(x) = P\{\omega \in \Omega : X(\omega) \leqslant x\}.$$

Esta función tiene la propiedad de caracterizar la distribución de probabilidad de X, P_X.

Si una variable aleatoria toma valores aislados se denomina *discreta*. Si por el contrario puede tomar cualquier valor de un intervalo, la variable aleatoria recibe el nombre de *continua*. Estos calificativos se aplican también a su distribución, hablando de *distribuciones discretas* o *continuas*.

A partir de las propiedades de las probabilidades se puede deducir que las funciones de distribución son

1. No crecientes.
2. Continuas por la derecha.
3. $\lim_{x\to-\infty} F(x) = 0$ y $\lim_{x\to\infty} F(x) = 1$.

Las variables aleatorias discretas X, las cuales tienen una función de distribución en escalera, tienen asociada una función, denominada *función de masa*, $p_X(x)$, la cual da la probabilidad de los valores de dicha variable aleatoria; es decir,

$$p_X(x) = P_X(\{x\}) = P\{\omega : X(\omega) = x\}.$$

Además,

$$p_X(x) = F(x) - F(-x)$$

en donde $F(-x)$ es el límite por la izquierda de F en x. Se ve, por tanto, que la función de masa recoge el valor del salto de la función de distribución, e inversamente,

$$F(X) = \sum_{y\leqslant x} p_X(y).$$

De manera análoga, las variables aleatorias continuas X tienen asociada una función, denominada *función de densidad*, $f_X(x)$, la cual indica la *velocidad* a la que crece su función de distribución, siendo

$$f_X(x) = \frac{d}{dx}F(x)$$

e inversamente,

$$F(x) = \int_{-\infty}^{x} f_X(y)\, dy.$$

Características de una distribución de probabilidad

Dada una variable aleatoria discreta X, con función de masa p_X, llamaremos *media* o *esperanza* de X a la suma de los valores que toma por las probabilidades con que los toma

$$\mu_X = E[X] = \sum_x x p_X(x)$$

y *varianza* de X a

$$\sigma_X^2 = V(X) = \sum_x (x - \mu_X)^2 p_X(x).$$

Dada una variable aleatoria continua X, con función de densidad f_X, llamaremos *media* o *esperanza* de X a la integral

$$\mu_X = E[X] = \int_{-\infty}^{\infty} x f_X(x)\, dx$$

y *varianza* de X a

$$\sigma_X^2 = V(X) = \int_{-\infty}^{\infty} (x-\mu_X)^2 f_X(x)\, dx.$$

En ambos casos, llamaremos *desviación típica* de X a la raíz cuadrada de la varianza:

$$\sigma_X = D(X) = \sqrt{\sigma_X^2}.$$

4.2.1. Funciones básicas de R en probabilidades

- p*distribu*`(x,par)`, con la que calculamos el valor de la función de distribución del modelo *distribu* en el punto `x`. Es decir, $F(x)$, siendo F la función de distribución de *distribu*.
- d*distribu*`(x,par)` con la que calculamos el valor de la función de masa o densidad de la distribución *distribu* en el punto `x`.
- q*distribu*`(p,par)` con la que podemos calcular el `p`-cuantil de la distribución *distribu*. Es decir, $F^{-1}(\mathtt{p})$, siendo F la función de distribución de *distribu*.
- r*distribu*`(n,par)` mediante la que podemos conseguir `n` valores obtenidos al azar según el modelo *distribu*.

El segundo argumento utilizado en las cuatro funciones anteriores, `par`, quiere indicar que es ahí en donde deberemos incluir el parámetro o parámetros de la distribución considerada.

En lugar de *distribu*, en las cuatro funciones de R antes mencionadas podemos utiilzar los modelos probabilísticos que estudiaremos en las secciones siguientes.

4.3. Variables aleatorias multivariantes

Una *variable aleatoria multivariante* $(X_1, ..., X_p)$ no es más que un vector de variables aleatorias unidimensionales, pudiendo generalizarse los conceptos vistos hasta ahora.

Centrándonos en el caso de una variable aleatoria bidimensional, (X, Y), podemos idealizar distribuciones de frecuencias relativas mediante una *función de masa bidimensional*

$$p_{XY}(x,y) = P\{X = x, Y = y\}$$

y mediante una *función de densidad bidimensional* $f_{XY}(x,y)$, tendremos caracterizada la distribución de probabilidad de una variable aleatoria bidimensional (X, Y) discreta o continua, para la cual también tendrán sentido las características poblacionales tales como las medias marginales (por ejemplo la de X)

$$\mu_X = \begin{cases} \sum\limits_x x p_X(x) = \sum\limits_x x \sum\limits_y p_{XY}(x,y) & \textit{Caso discreto} \\ \int_{-\infty}^{\infty} x f_X\, dx = \int_{-\infty}^{\infty} x \int_{-\infty}^{\infty} f_{XY}(x,y)\, dy & \textit{Caso continuo} \end{cases}$$

la *covariana poblacional*

$$\mu_{11} = \begin{cases} \sum\limits_x \sum\limits_y (x-\mu_X)(y-\mu_Y) p_{XY}(x,y) & \textit{Caso discreto} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x-\mu_X)(y-\mu_Y) f_{XY}(x,y)\, dx\, dy & \textit{Caso continuo} \end{cases}$$

y el *coficiente de correlación poblacional*, ρ, definido por

$$\rho = \frac{\mu_{11}}{\sigma_X \sigma_Y}$$

en donde σ_X y σ_Y son las desviaciones típicas marginales.

Independencia de variables aleatorias

Diremos que las variables aleatorias discretas $\{X_1, ..., X_n\}$ son *independientes*, si y sólo si la función de masa conjunto es el producto de las funciones de masa marginales,

$$p_{(X_1,...,X_n)}(x_1, ..., x_n) = \prod_{i=1}^{n} p_{X_i}(x_i).$$

Análogamente, diremos que las variables aleatorias continuas $\{X_1, ..., X_n\}$ son *independientes*, si y sólo si la función de densidad conjunta es el producto de las marginales,

$$f_{(X_1,...,X_n)}(x_1, ..., x_n) = \prod_{i=1}^{n} f_{X_i}(x_i).$$

4.4. Modelos unidimensionales discretos

4.4.1. Distribución binomial

Esta distribución modeliza el *número de exitos* en experimentos denominados, de forma genérica, *pruebas de Bernoulli*. Estas pruebas consisten en la realización de ensayos repetidos e independientes, existiendo en cada ensayo solamente dos resultados posibles (denominados de forma genérica *éxito* y *fracaso*) y manteniéndose constante la probabilidad éxito a lo largo de los ensayos.

Si el número de pruebas de Bernoulli que se realizan es n y la probabilidad de *éxito* en cada una de ellas es p, la variable de interés X es el *número de éxitos en las n pruebas*, siendo la función de masa de esta distribución,

$$p_X(x) = P\{X = x\} = \binom{n}{x} p^x (1-p)^{n-x}, \quad i = 0, 1, ..., n$$

en donde deben ser $1 \leqslant n$ y $0 < p < 1$. En este caso diremos que X sigue una distribución (o tiene un modelo) binomial de parámetros n y p y lo representaremos por $X \rightsquigarrow B(n,p)$.

Su media y su varianza son, respectivamente,

$$E[X] = np \quad \text{y} \quad V(X) = np(1-p).$$

El caso particular de que sólo se considere una prueba de Bernoulli, es decir, que sea $X \rightsquigarrow B(1,p)$ recibe el nombre de *Distribución de Bernoulli*.

En ADD, *Tabla 1*, vienen recogidos los valores de la función de masa para algunos valores del parámetro p, no obstante, el cálculo de probabilidades asociadas a distribuciones binomiales se hace hoy en dia con R usando el comando `binom()`. Así,

- `pbinom(x,n,p)`, valor de la función de distribución en `x` de la binomial `(n,p)`.
- `dbinom(x,n,p)`, valor de la función de masa en `x` de la binomial `(n,p)`.
- `qbinom(q,n,p)`, cuantil de orden `q` de la binomial `(n,p)` (hasta él la probabilidad acumulada es `q`).
- `rbinom(m,n,p)`, muestra aleatoria de tamaño `m` de la binomial `(n,p)`.

4.4.2. Distribución de Poisson

Se utiliza, por lo general, para modelizar el número de veces que ocurren sucesos raros. La distribución de masa de la variable *número de éxitos* de esta distribución es

$$p_X(x) = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x = 0, 1, 2, ...$$

siendo $\lambda > 0$ el único parámetro del que depende esta distribución. Si una variable aleatoria X tiene como modelo probabilístico una *Distribución de Poisson* de parámetro λ, lo expresaremos poniendo $X \rightsquigarrow \mathcal{P}(\lambda)$.

De hecho, este parámetro es su media y su varianza, $E[X] = \lambda = V(X)$, coincidencia nada habitual en los modelos probabilísticos y que puede servir como indicación de que una variable puede ser modelizada por esta distribución.

Existen tablas de la distribución de Poisson que nos dan su función de masa para distintos valores del parámetro. En ADD, *Tabla 2*, aparecen los más utilizados.

Aproximación de la distribución binomial por la de Poisson

La función de masa de la distribución binomial $B(n, p)$ converge, cuando n crece, a la función de masa de una distribución de Poisson (λ) con $\lambda = np$.

Por tanto, cuando queramos calcular probabilidades binomiales con n grande, podremos utilizar las tablas de la distribución de Poisson.

Esta aproximación es buena cuando p es muy pequeño con respecto a np y a su vez np es también muy pequeño con respecto a n. Como indicación de estas cantidades, se toma como buena aproximación cuando al menos es $np < 5$ y $p < 0,1$.

Para ejecutar con R aplicaciones de esta distribución, el comando que debemos utilizar es `pois()`. Así,

- `ppois(x,a)`, valor de la función de distribución en `x` de la Poisson(a).
- `dpois(x,a)`, valor de la función de masa en `x` de la Poisson(a).
- `qpois(p,a)`, cuantil de orden `p` de la Poisson(a).
- `rpois(n,a)`, muestra aleatoria de tamaño `n` de la Poisson(a).

4.4.3. Distribución geométrica

La *Distribución geométrica* de parámetro p es un modelo que se asocia también con pruebas de Bernoulli, es decir, del tipo éxito/fracaso con probabilidad de éxito p aunque modelizando ahora la variable *número de fallos antes del primer éxito*. Su función de masa es

$$p_X(x) = (1-p)^x p, \quad x = 0, 1, 2, ...$$

en donde debe ser $0 < p \leqslant 1$. Se expressa con $X \rightsquigarrow Geom(p)$.

Su media y su varianza son respectivamente

$$E[X] = \frac{1-p}{p} \quad \text{y} \quad V(X) = \frac{1-p}{p^2}.$$

El comando a utilizar en R es `geom()`. Así,

- `pgeom(x,p)`, función de distribución en `x` de la geométrica(p).
- `dgeom(x,p)`, función de masa en `x` de la geométrica(p).
- `qgeom(q,p)`, cuantil de orden `q` de la geométrica(p).
- `rgeom(n,p)`, muestra aleatoria de tamaño `n` de la geométrica(p).

4.4.4. Distribución hipergeométrica

Este modelo se utiliza para situaciones que se adaptan al siguiente esquema: Se supone una caja con N piezas de las cuales D son *defectuosas* y $N - D$ *no defectuosas*. Se extraen sin reemplazamiento n piezas (o las n de una vez) de la caja y estamos interesados en modelizar el *número de defectuosas extraídas en las n seleccionadas*.

El cálculo de probabilidades que nos da el valor de esta probabilidad que será la función de masa de este modelo,

$$p_X(x) = \frac{\binom{D}{x}\binom{N-D}{n-x}}{\binom{N}{n}}, \quad \text{máx}\{0, n-N+D\} \leqslant x \leqslant \text{mín}\{n, D\}.$$

El que una variable X siga esta distribución lo expresaremos como $X \rightsquigarrow Hiper(D, N, n)$.

Su media y su varianza son respectivamente

$$E[X] = \frac{Dn}{N} \quad \text{y} \quad V(X) = \frac{D(N-D)n(N-n)}{N^2(N-1)}$$

El comando a utilizar en R es `hyper()`. Así,

- `phyper(x,D,N-D,n)`, función de distribución en `x` de ña hipergeométrica(D,N,n).
- `dhyper(x,D,N-D,n)`, función de masa en `x` de la hipergeométrica(D,N,n).
- `qhyper(p,D,N-D,n)`, cuantil de orden `p` de la hipergeométrica(D,N,n).
- `rhyper(n,D,N-D,n)`, muestra aleatoria de tamaño `n` de la hipergeométrica(D,N,n).

4.4.5. Distribución binomial negativa

Esta es una generalización de la distribución geométrica antes estudiada. Con una *distribución binomial negativa* modelizamos de nuevo un experimento de Bernoulli, del tipo éxito/fracaso con probabilidad de éxito p, pero analizando ahora la variable X =*número de fallos antes del éxito n-ésimo*. La función de masa de este modelo, expresado de la forma $X \rightsquigarrow BN(n,p)$, será

$$p_X(x) = \binom{n+x-1}{n-1}(1-p)^x p^n, \quad x = 0,1,2,...$$

en donde debe ser $0 < n$ y $0 < p \leqslant 1$.

El comando a utilizar en R es `nbinom()`. Así,

- `pnbinom(x,n,p)`, función de distribución en `x` de la binomial negativa(n,p).
- `dnbinom(x,n,p)`, función de masa en `x` de la binomial negativa(n,p).
- `qnbinom(q,n,p)`, cuantil de orden `q` de la binomial negativa(n,p).
- `rnbinom(n,n,p)`, muestra aleatoria de tamaño `n` de la binomial negativa(n,p).

4.5. Modelos unidimensionales continuos

4.5.1. Distribución normal

La *Distribución normal* se define como aquella distribución cuya función de densidad es

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}, \quad -\infty < x < \infty$$

donde debere ser $0 < \sigma$.

La distribución normal depende de dos parámetros, μ y σ, y se expresa de la forma $X \rightsquigarrow N(\mu,\sigma)$.

Se puede demostrar que si $X \rightsquigarrow N(\mu,\sigma)$ entonces $E[X] = \mu$ y $V(X) = \sigma^2$.

El caso de la distribución normal de media $\mu = 0$ y desviación típica $\sigma = 1$, es de singular importancia, de ahí que, en ocasiones, a la $N(0,1)$ se la denomine *normal estándar*.

La relación existente entre una variable $X \rightsquigarrow N(\mu,\sigma)$ y una $Z \rightsquigarrow N(0,1)$ es muy sencilla:

$$Z = \frac{X-\mu}{\sigma} \quad \text{o bien} \quad X = \mu + \sigma Z$$

El paso de una $N(\mu,\sigma)$ a una $N(0,1)$ se denomina *tipificación*, y es muy importante en estadística ya que la función distribución de una normal no admite una expresión explícita, sino que es necesario acudir a unas tablas calculadas a tal efecto.

No obstante, por el proceso de tipificación, no será preciso dar tablas de todas las normales, sino solamente de la $N(0,1)$. Éste es el contenido de la *Tabla 3* de ADD.

Si es $X \rightsquigarrow N(\mu,\sigma)$, el coficiente de asimetría es $E[(X-\mu)^3]/\sigma^3 = 0$ y el de apuntamiento o curtosis es $E[(X-\mu)^4]/\sigma^4 = 3$, valores que, calculados en una muestra, permiten un rápido análisis de si los datos se distribuyen como una normal.

El comando a utilizar en R es `norm()`. Así,

- `pnorm(x,a,b)`, función distribución en `x` de la normal(a,b).
- `dnorm(x,a,b)`, función de densidad en `x` de la nromal(a,b).
- `qnorm(p,a,b)`, cuantil de orden `p` de la normal(a,b).
- `rnorm(n,a,b)`, muestra aleatoria de tamañano `n` de la norma(a,b).

Si no indicamos nada, R toma por defecto los valores $a = 0, b = 1$.

4.5.2. Distribución uniforme

La *distribución uniforme*, de parámetros (a,b), es un modelo que asigna, de forma continua, igual probabilidad a todas las partes del intervalo (a,b) en el que está definida. Su función de densidad es

$$f_X(x) = \frac{1}{b-a}, \quad a \leqslant x \leqslant b$$

y lo representaremos por $X \rightsquigarrow Unif(a,b)$.

Su media y su varianza son, respectivamente,

$$E[X] = \frac{a+b}{2} \quad \text{y} \quad V(X) = \frac{(b-a)^2}{12}$$

El comando a utilizar en R es `unif()`. Así,

- `punif(x,a,b)`, función de distribución en `x` de la uniforme(a,b).
- `dunif(x,a,b)`, función de densidad en `x` de la uniforma(a,b).
- `qunif(p,a,b)`, cuantil de orden `p` de la uniforma(a,b).
- `runif(n,a,b)`, muestra aleatoria de tamaño `n` de la uniforme(a,b).

Si no indicamos nada, R toma por defecto los valores $a = 0, b = 1$.

4.5.3. Distribución beta

La *distribución beta* de parámetros (a,b) tiene por función de densidad,

$$f_X(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}, \quad 0 < x < 1$$

en donde Γ es la función $\Gamma(p) = \int_0^\infty t^{p-1}e^{-t}\,dt$, definida para $p > 0$.

Deben ser $a > 0$ y $b > 0$ y su media y su varianza son, respectivamente,

$$E[X] = \frac{a}{a+b} \quad \text{y} \quad V(X) = \frac{ab}{(a+b)^2(a+b+1)}$$

El comando a utilizar en R es `beta()`. Así,

- `pbeta(x,a,b)`, función de distribución en `x` de la beta(a,b).
- `dbeta(x,a,b)`, función densidad en `x` de la beta(a,b).
- `qbeta(p,a,b)`, cuantil de orden `p` de la beta(a,b).
- `rbeta(n,a,b)`, muestra aleatoria de tamaño `n` de la beta(a,b).

4.5.4. Distribuciones Gamma y exponencial

La *distribución Gamma* de parámetros (a, b) tiene por función de densidad,

$$f_X = \frac{1}{b^a \Gamma(a)} x^{a-1} e^{-x/b}, \quad x > 0$$

en donde deben ser $a > 0$ y $b > 0$. Su media y su varianza son, respectivamente,

$$E[X] = ab \quad \text{y} \quad V(X) = ab^2.$$

El caso particular de $a = 1$ se denomina *distribución exponencial*, tanto esta distribución como su generalización, la distribución gamma, son muy habituales en la modelización de tiempos que transcurren hasta que un determinado suceso acontece.

El comando a utilizar en R es `gamma()`. Así,

- `pgamma(x,a,b)`, función de distribución en `x` de la gamma(a,b).
- `dgamma(x,a,b)`, función de densidad en `x` de la gamma(a,b).
- `qgamma(p,a,b)`, cuantil de orden `p` de la gamma(a,b).
- `rgamma(n,a,b)`, muestra aleatoria de tamaño `n` de la gamma(a,b).

Si no especificamos el valor del parámetro, R toma $b = 1$.

4.5.5. Distribución de Cauchy

La *distribución de Cauchy* de parámetros (a, b) tienepor función de densidad,

$$f_X(x) = \frac{b}{\pi} \frac{1}{b^2 + (x-a)^2}, \quad -\infty < x < \infty$$

en donde debe ser $b > 0$, mientras que a puede ser cualquier número real.

El comando a utilizar en R es `cauchy()`. Así,

- `pcauchy(x,a,b)`, función de distribución en `x` de la Cauchy(a,b).
- `dcauchy(x,a,b)`, función de densidad en `x` de la Cauchy(a,b).
- `qcauchy(p,a,b)`, cuantil de orden `p` de la Cauchy(a,b).
- `rcauchy(n,a,b)`, muestra aleatoria de tamaño `n` de la Cauchy(a,b).

Si no le damos valores a los parámetros, R toma por defecto, $a = 0$ y $b = 1$.

4.6. Modelos bidimensionales

4.6.1. Distribución normal bivariante

Una variable aleatoria bidimensional (X, Y) se dice que sigue una distribución normal bivariante de medias (μ_1, μ_2) y de varianzas-covarianzas $(\sigma_1^2, \sigma_2^2, \mu_{11})$, si su función de densidad es

$$f(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \cdot$$

$$\cdot \exp\left\{ -\frac{1}{2(1-\rho^2)} \left[\frac{(x-\mu_1)^2}{\sigma_1^2} - \frac{2\rho(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2} \right] \right\}$$

si $-\infty < x < \infty, -\infty < y < \infty$.

En este caso, las medias y varianzas marginales son $E[X] = \mu_1, V(X) = \sigma_1^2, E[Y] = \mu_2, V(Y) = \sigma_2^2$ y la covarianza $\mu_{11} = \rho\sigma_1\sigma_2$.

4.7. Teorema central del límite

Si $X_1, X_2, ...$ es una sucesión de variables aleatorias independientes, idénticamente distribuidas y con varianza común σ^2 finita, entonces la variable aleatoria

$$\frac{X_1 + X_2 + ... + X_n - n\mu}{\sqrt{n}\sigma}$$

en donde μ es la media común, tiene como distribución asintótica una $N(0, 1)$.

Es decir, si tenemos un gran número de observaciones independientes $X_1, X_2, ...$, sa cual sea su distribución común (mientras tenga varianza finita), para n suficientemente grande podemos aproximar la distribución de la variable aleatoria anterior por una $N(0, 1)$.

Obsérvese que dividiendo por n en la expresión anterior, podemos expresar el resultado diciendo que para n grande es

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

siendo $\overline{x} = (X_1 + ... + X_n)/n$ la media aritmética de las observaciones (que más adelante denominaremos *media muestral*). Es decir, que la distribución de $\overline{x}$ es, aproximadamente,

$$\overline{x} \approx N\left(\mu, \frac{\sigma}{\sqrt{n}}\right).$$

Una aplicación directa de dicho teorema es la aproximación de una distribución binomial $X \rightsquigarrow N(n, p)$ por una normal,

$$X \approx N\left(np, \sqrt{np(1-p)}\right).$$

5. Estimadores. Distribución en el muestreo

5.1. Introducción

El proceso de selección de una muestra aleatoria simple conlleva el que las X_i sean variables aleatorias independientes e idénticamente distribuidas con distribución común la de X.

Así pues, formalmente una muestra aleatoria simple de una variable aleatoria X es una variable aleatoria n-dimensional $(X_1, ..., X_n)$ cuyas variables aleatorias unidimensionales que la componen son independientes y con la misma distribución.

Por tanto, si X es continua con función de densidad f, la función de densidad conjunta de $(X_1, ..., X_n)$ será

$$f(x_1, ..., x_n) = \prod_{i=1}^{n} f(x_i) = f(x_1) \cdot ... \cdot f(x_n)$$

y si X es discreta con función de masa p, la función de masa conjunta será

$$p(x_1, ..., x_n) = \prod_{i=1}^{n} p(x_i).$$

Esta distribución de $(X_1, ..., X_n)$ se denomina *distribución muestral* y dado que la situación habitual que se plantea en inferencia es la de ser la distribución de X no totalmente conocida, sino dependiente de algún parámetro θ desconocido, la distribución muestral también dependerá de θ, haciéndose referencia explícita de éste en su expresión, f_θ o p_θ.

Otra cuestión es la estimación. Estaremos interesados bien en asignar un valor numérico al parámetro θ (*estimación por punto*), o bien inferir un conjunto de valores plausibles para θ (*estimación por intervalos de confianza* y *contraste de hipótesis*). En este proceso será imprescindible contar más que con la muestra, con una función suya $T(X_1, ..., X_n)$ denominada *estimador* o *estadístico*.

Así, si θ es la media de la población, parece razonable utilizar la *media muestral*

$$T(X_1, ..., X_n) = \overline{x} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

en su estimación, entendida ésta como función (media aritmética) de los valores que observemos.

Así, la *varianza muestral*

$$s^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{x})^2$$

es un buen estimador de la varianza poblacional $\sigma^2 = V(X)$.

El *coeficiente de correlación muestral*

$$r = \frac{n \sum_{i=1}^{n} X_i Y_i - (\sum_{i=1}^{n} X_i)(\sum_{i=1}^{n} Y_i)}{\sqrt{n \sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i^2)} \sqrt{n \sum_{i=1}^{n} Y_i^2 - (\sum_{i=1}^{n} Y_i^2)}}$$

lo es del coeficiente de correlación poblacional ρ, etc.

Al ser $T(X_1, ..., X_n)$ una variable aleatoria, tendrá una distribución de probabilidad que denominaremos *distribución en el muestreo* de T.

Así, un estimador cuya media sea el parámetro a estimar: $E[T] = \theta$ (denominado *centrado* o *insesgado*) es deseable, puesto que esta propiedad nos expresa una *cancelación* entre los valores mayores y menores que toma, respecto al parámetro.

Un estimador con poca varianza nos indicará una mayor probabilidad de obtención de valores *cercanos* al parámetro.

5.2. Método de la máxima verosimilitud

La idea del *Método de la máxima verosimilitud* consiste en dar como estimación del parámetro aquel valor (de entre los posibles) que haga máxima la probabilidad del suceso observado, es decir, de la muestra obtenida. Es decir, aquel que maximice la función de masa o densidad de la muestra observada, $p_\theta(x_1, ..., x_n)$ ó $f_\theta(x_1, ..., x_n)$.

Pero al decir *de la muestra observada* estamos diciendo que, en esa función los valores $x_1, ..., x_n$ están fijos y lo que en realidad hacemos variar es θ con objeto de maximizar la función.

Para resaltas este hecho, a la función de probabilidad de la muestra la representaremos por $L(\theta)$, y la denominaremos *función de verosimilitud* de la muestra,

$$L(\theta) = \begin{cases} p_\theta(x_1, ..., x_n) & \text{si es discreta la variable} \\ f_\theta(x_1, ..., x_n) & \text{si es continua la variable.} \end{cases}$$

El método de la máxima verosimilitud propone como estimador de θ aquel $\hat{\theta}$ que maximice la función de verosimilitud,

$$L(\hat{\theta}) = \max_{\theta} L(\theta).$$

Como el máximo de una función y el de su logaritmo se alcanzan en el mismo punto, habitualmente determinaremos el $\hat{\theta}$ tal que

$$\log L(\hat{\theta}) = \max_{\theta} \log L(\theta).$$

El cálculo de este máximo se determina de forma habitual en la que se determinan los máximos de una función.

5.3. Distribuciones asociadas a poblaciones normales

5.3.1. Distribución χ^2 de Pearson

Sean $X_1, ..., X_n$, n variables aleatorias independientes, cada una de las cuales sigue una distribución $N(0, 1)$.

LLamaremos *distribución χ^2_n de Pearson* a la distribución de la variable aleatoria suma de los cuadrados de las n variables $N(0, 1)$

$$Y = X_1^2 + X_2^2 + ... + X_n^2.$$

El subíndice n de la χ^2_n corresponde al número de variables aleatorias independientes cuya suma forma la χ^2_n y se denomina *grados de libertad* de la variable. Tiene que ser, por lo tanto, un número entero positivo.

La χ^2_n es una distribución continua cuya función de densidad es

$$f(y) = \frac{y^{n/2-1} e^{-y/2}}{2^{n/2} \Gamma(n/2)}, \quad y > 0$$

siendo su media $E[Y] = n$ y su varianza $V(Y) = 2n$.

La función de distribución o, más en concreto, uno menos ella, $P\{\chi^2_n > p\}$, es decir, la probabilidad cola, es la función más usada en inferencia. Con el objeto de obtener estas probabilidades cola, en la *Tabla 4* de ADD, aparecen algunos de sus valores; o mejor dicho, las abscisas a las que corresponden probabilidades cola p, expresadas por líneas según los grados de libertad n.

En dicha tabla sólo aparecen valores hasta $n = 30$ grados de libertad. La razón es que para mayor número de grados de libertad, esta distribución se aproxima por una normal. En concreto, se verifica que

$$\sqrt{2\chi_n^2} \approx N(\sqrt{2n-1}, 1).$$

Si queremos utilizar R en el cálculo de probabilidades relacionadas con esta distribución, el comando a utilizar es `chisq()`. Así,

- `pchisq(x,n)`, función de distribución en `x` de la χ^2 con `n` grados.
- `dchisq(x,n)`, función de densidad en `x` de la χ^2 con `n` grados.
- `qchisq(p,n)`, cuantil de orden `p` de la χ^2 con `n` grados.
- `rchisq(m,n)`, muestra aleatoria de tamaño `m` de la χ^2 con `n` grados.

5.3.2. Distribución t de Student

En poblaciones normales $N(\mu, \sigma)$, la distribución en el muestreo de la media muestral $\overline{x}$ es también normal, aunque dependiente de σ, siendo este parámetro habitualmente desconocido.

Este hecho hace inviable el cálculo de probabilidades relacionadas con dicha media muestral, a menos que las muestras sean lo suficientemente grandes como para poder aplicar el teorema central del límite.

Si $X, X_1, ..., X_n$ son $n+1$ variables aleatorias independientes e idénticamente distribuidas $N(0, \sigma)$, llamaremos *distribución t de Student* a la distribución de variable aleatoria

$$T = \frac{X}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} X_i^2}}.$$

El número de variables aleatorias independientes del denominador recibe el nombre de *grados de libertad* de la t de Student, por lo que deberá ser un número entero positivo. A esta distribución se la suele representar por t_n.

La distribución t_n es de tipo continuo siendo su función de densidad,

$$f(y) = \frac{1}{\sqrt{n\pi}} \frac{\Gamma((n+1)/2)}{\Gamma(n/2)} \left(1 + \frac{y^2}{n}\right)^{-(n+1)/2}, \ -\infty < y < \infty$$

la cual tiene como un aspecto semejante a una $N(0, 1)$. De hecho, para muestras grandes converge a una $N(0, 1)$.

Se puede demostrar que si $Y \rightsquigarrow t_n$, entonces es $E[Y] = 0$ y $V(Y) = n/(n-2)$.

El cálculo de probabilidades relacionadas con esta distribución está tabulado en la *Tabla 5* de ADD.

Si queremos calcular probabilidades con R, el comando a utilizar es `t()`. Así,

- `pt(x,n)`, función de distribución en `x` de la t-Student con `n` grados.
- `dt(x,n)`, función de densidad en `x` de la t-Student con `n` grados.
- `qt(p,n)`, cuantil de orden `p` de la t-Student con `n` grados.
- `rt(m,n)`, muestra aleatoria de tamaño `m` de la t-Student con `n` grados.

5.3.3. Distribución F de Snedecor

Sean $X_1, ..., X_{n_1}, Y_1, ..., Y_{n_2}, n_1 + n_2$ variables aleatorias independientes e idénticamente distribuidas $N(0, \sigma)$. Llamaremos *distribución F de Snedecor* a la distribución de la variable aleatoria

$$F = \frac{\frac{1}{n_1}\sum_{i=1}^{n_1} X_i^2}{\frac{1}{n_2}\sum_{i=1}^{n_2} Y_i^2}.$$

El número de sumandos del numerados y denominador recibe el nombre de *grados de libertad* de la F de Snedecor, por lo que se suele representar esta distribución como F_{n_1,n_2}. Lógicamente, estos números deben ser enteros positivos.

Esta distribución también es de tipo continuo cuya función de densidad, de aspecto semejante a la χ^2, es

$$f(y) = \frac{\Gamma(n_1/2 + n_2/2)}{\Gamma(n_1/2)\Gamma(n_2/2)} \frac{n_1^{n_1/2} n_2^{n_2/2} y^{n_1/2-1}}{(n_1 y + n_2)^{(n_1+n_2)/2}}, \quad y > 0$$

apareciendo las probabilidades cola de esta distribución en ADD *Tabla 6*.

Un hecho importante en relación con la búsqueda de abscisas de esta distribución, es que, como por definición es

$$F_{n,m} = \frac{1}{F_{m,n}}$$

si representamos por $F_{m,n;p}$ el valor de la abscisa de la función de densidad de una F de Snedecor con (m, n) grados de libertad que deja a la derecha una área de probabilidad p,

$$P\{F_{m,n} > F_{m,n;p}\} = p$$

entonces es

$$F_{n,m;1-p} = \frac{1}{F_{m,n;p}}$$

Si queremos calcular con R probabilidades relacionadas con esta distribución, el comando a utilizar es `f()`. Así,

- `pf(x,n1,n2)`, función de distribución en `x` de la $F(n_1, n_2)$.
- `df(x,n1,n2)`, función de densidad en `x` de la $F(n_1, n_2)$.
- `qf(p,n1,n2)`, cuantil de orden `p` de la $F(n_1, n_2)$.
- `rf(n,n1,n2)`, muestra aleatoria de tamaño `n` de la $F(n_1, n_2)$.

5.4. Estimación de la media de una población normal

En esta sección estudiaremos cúal debe ser el estimador a utilizar para estimar la media μ, cuando para la variable en estudio X se supone como modelo una $N(\mu, \sigma)$, así como su distribución en el muestreo.

En ADD aparecen resumidos los resultados que se irán obteniendo para cada situación, para su rápida consulta.

Teorema de Fisher

Sea $X_1, ..., X_n$ una muestra aleatoria simple de una población $N(\mu, \sigma)$. Entonces, si $\overline{x}$ y S^2 son, respectivamente, la media y cuasivarianza muestrales se tiene que

a) $\overline{x} \rightsquigarrow N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$.

b) $\frac{(n-1)S^2}{\sigma^2} \rightsquigarrow \chi^2_{n-1}$.

c) $\overline{x}$ y $\frac{(n-1)S^2}{\sigma^2}$ son independientes.

A partir de este teorema obtenemos los siguientes resultados.

σ conocida

Cuando la varianza poblacional es conocida, es razonable utilizar la media muestra $\overline{x}$ para estimar μ. Su distribución en el muestreo es una normal de media μ y desviación típica $\sigma/\sqrt{n}$. Es decir,

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

σ desconocida

Como una t de Student es el cociente entre una $N(0,1)$ y la raíz cuadrada de una χ^2 dividida por sus grados de libertad, si es que ambas distribuciones son independientes, del teorema de Fisher obtenemos que

$$\frac{\frac{\overline{x}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S^2/\sigma^2}{n-1}}} \rightsquigarrow t_{n-1}$$

de donde simplificando se obtiene que

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

Obsérvese que, como para muestras grandes (digamos n>30) la distribución t de Student se aproxima por una $N(0,1)$, las probabilidades del cociente anterior se buscarán en las tablas de la normal en esos casos.

5.5. Estimación de la media de una población no necesariamente normal. Muestras grandes

Si no se conoce el modelo y las muestras son suficientemente grandes se deben utilizar los resultados del siguiente apartado. Si no se conoce el modelo y las muestras no se pueden considerar grandes, o éstas son de datos *cualitativos* los métodos a utilizar son *no paramétricos*, los cuales se estudiarán en el capítulo 8.

Otra cosa es que la población no sea normal pero sea conocida. En estos casos habrá que utilizar métodos específicos de la población en cuestión, determinando primero el estimador adecuado al parámetro en estudio y luego calculando su distribución en el muestreo.

Vemos en esta sección dos situaciones en las que para muestras grandes se obtienen distribuciones límite normales.

Población no necesariamente normal

Si no conocemos o no queremos suponer un modelo determinado para la variable en estudio, siempre que ésta tenga varianza finita σ^2, podemos utilizar el teorema central del límite obteniendo para muestras suficientemente grandes, digamso $n > 30$, que

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

Esto claro está, si σ **es conocido**, ya que si no lo es, tampoco nos servirá este resultado.

Si el tamaño muestral es algo mayor, digamos $n > 100$, podemos sustituir la varianza por un estimador suyo, obteniendo en el caso de que σ **sea desconocido** que

$$\frac{\overline{x} - \mu}{S/\sqrt{n}} \approx N(0,1)$$

Población binomial

Si estamos interesados en estimar una proporción poblacional p es razonable establecer un modelo binomial para la variable dicotómica en estudio con p con probabilidad de éxito $X \rightsquigarrow B(1,p)$.

En este caso, las observaciones $X_1, ..., X_n$ serán unos o ceros según presenten o no los n individuos de la muestra la característica en estudio.

Utilizando el método de la máxima verosimilitud se puede deducir que el estimador de p es la *proporción muestral*

$$\hat{p} = \frac{\sum_{i=1}^n X_i}{n}$$

siendo su distribución en el muestreo tal que $n\hat{p} \rightsquigarrow B(n,p)$.

Por las buenas propiedades asintóticas que tienen los estimadores de máxima verosimilitud, se puede demostrar que si las muestras son suficientemente grandes, digamos $n > 100$, es

$$\frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \approx N(0,1)$$

Población Poisson

Si se admite un modelo $X \rightsquigarrow \mathcal{P}(\lambda)$, el estimador de máxima verosimilitud para λ basado en una muestra aleatoria simple de tamaño n de X, era la media muestral $\overline{x}$, el cual tiene una distribución en el muestreo tal que $n\overline{x} \rightsquigarrow \mathcal{P}(n\lambda)$.

No obstante, para muestras grandes, digamso $n > 100$, es posible utilizar su distribución asintótica, que es

$$\frac{\overline{x} - \lambda}{\sqrt{\overline{x}/n}} \approx N(0,1)$$

5.6. Estimación de la varianza de una población normal

μ desconocida

Si la media μ es desconocida, el teorema de Fisher ya nos indicaba que el estimador de la varianza en este supuesto debía ser la cuasivarianza muestral S^2 ya que, entre otras razones, al ser

$$\frac{(n-1)S^2}{\sigma^2} \rightsquigarrow \chi^2_{n-1}$$

será

$$E\left[\frac{(n-1)S^2}{\sigma^2}\right] = n-1$$

es decir, $E[S^2] = \sigma^2$, lo cual supone que S^2 posee una propiedad deseable en los estimadores.

Por tanto, si μ es desconocida el estimador a utilizar para estimar la varianza σ^2 es S^2 con distribución en el muestreo

$$\frac{(n-1)S^2}{\sigma^2} \rightsquigarrow \chi^2_{n-1}$$

μ conocida

Si μ es conocida, el caso anterior nos sugiere que el estimador a considerar sea similar al allí considerado pero utilizando, en lugar de la media muestral, la media poblacional μ ya que es conocida. Es decir, parece razonable utilizar el estimador

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2.$$

Además, como las X_i siguen distribución $N(\mu, \sigma)$, seguirá $(X_i - \mu)/\sigma$ una distribución $N(0,1)$, con lo que será

$$\frac{\sum_{i=1}^{n}(X_i - \mu)}{\sigma^2} \rightsquigarrow \chi^2_n$$

por definición de la distribución χ^2_n.

5.7. Estimación del cociente de varianzas de dos poblaciones normales independientes

Supondremos que $X_1, ..., X_{n_1}$ una muestra aleatoria sumple, de tamaño n_1, de una $N(\mu_1, \sigma_1)$ y que $Y_1, ..., Y_{n_2}$ una muestra aleatoria simple, de tamaño n_2, de una $N(\mu_2, \sigma_2)$, siendo ambas independientes y con medias muestrales, respectivamente, $\overline{x}_1$ y $\overline{x}_2$.

Como la cuasivarianza muestral es un buen estimador de la varianza poblacional, parece razonable estimar σ_1^2/σ_2^2 mediante el cociente S_1^2/S_2^2.

No obstante, con objeto de hacer inferencias sobre el cociente de las varianzas poblacionales, es necesario conocer la distribución en el muestreo del estimador utilizado en dichas inferencias.

Aplicando el teorema de Fisher a cada una de las dos poblaciones, sabemos que es

$$\frac{\sum_{i=1}^{n_1}(X_i - \overline{x}_1)^2}{\sigma_1^2} = \frac{(n_1-1)S_1^2}{\sigma_1^2} \rightsquigarrow \chi^2_{n_1-1}$$

y

$$\frac{\sum_{j=1}^{n_2}(Y_j - \overline{x}_2)^2}{\sigma_2^2} = \frac{(n_2-1)S_2^2}{\sigma_2^2} \rightsquigarrow \chi^2_{n_2-1}$$

siendo además ambas χ^2 independientes.

Por tanto, será

$$\frac{\frac{1}{n_1-1}\frac{\sum_{i=1}^{n_1}(X_i - \overline{x}_1)^2}{\sigma_1^2}}{\frac{1}{n_2-1}\frac{\sum_{j=1}^{n_2}(Y_j - \overline{x}_2)^2}{\sigma_2^2}} = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \rightsquigarrow F_{n_1-1,n_2-1}$$

al ser el cociente de dos χ^2 independientes divididas por sus grados de libertad, una F de Snedecor.

Por tanto, si las medias poblacionales μ_1 y μ_2 **son desconocidas** será

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \rightsquigarrow F_{n_1-1,n_2-1}.$$

Si fueran μ_1 y μ_2 **conocidas**, como una χ^2 es la suma de cuadrados de normales $N(0,1)$ independientes y

$$\sum_{i=1}^{n_1}\left(\frac{X_i - \mu_1}{\sigma_1}\right)^2 \rightsquigarrow \chi^2_{n_1} \quad \text{y} \quad \sum_{j=1}^{n_2}\left(\frac{Y_j - \mu_2}{\sigma_2}\right)^2 \rightsquigarrow \chi^2_{n_2}$$

siendo además ambas distribuciones independientes por proceder las observaciones de poblaciones independientes, tendremos, por los mismo razonamientos anteriores, que

$$\frac{\frac{1}{n_1}\frac{\sum_{i=1}^{n_1}(X_i - \mu_1)^2}{\sigma_1^2}}{\frac{1}{n_2}\frac{\sum_{j=1}^{n_2}(Y_j - \mu_2)^2}{\sigma_2^2}} \rightsquigarrow F_{n_1,n_2}$$

Aunque los papeles de la primera y segunda poblaciones son intercambiables, suele tomarse como población 1, la de mayor cuasivarianza muestral debido a que se obtienen mejores inferencias con $S_1^2 > S_2^2$, por la simetría de la distribución F de Snedecor.

5.8. Estimación de la diferencia de medias de dos poblaciones normales independientes

La situación que nos ocupa en esta sección, es la de dos muestras independientes, la muestra $X_1, ..., X_{n_1}$ de

una $N(\mu_1, \sigma_1)$ y la muestra $Y_1, ..., Y_{n_2}$ procedente de una $N(\mu_2, \sigma_2)$.

Aplicando el teorema de Fisher a cada una de las muestras se obtiene que es

$$\overline{x}_1 \rightsquigarrow N(\mu_1, \sigma_1/\sqrt{n_1}) \quad \text{y} \quad \overline{x}_2 \rightsquigarrow N(\mu_2, \sigma_2/\sqrt{n_2})$$

siendo ambas normales independientes, por proceder de poblaciones independientes. Por tanto, será

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

por lo que, si σ_1 y σ_2 **son conocidas**, será

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

Si σ_1 y σ_2 **son desconocidas y las muestras son pequeñas**, digamos $n_1 + n_2 \leqslant 30$, habrá que recurrir a una t de Student.

No obsante,en el caso que nos ocupa habrá que diferenciar dos situaciones, admitiendo, la mayoriá de veces, una de las dos por medio de un contraste de hipótesis.

σ_1 y σ_2 se suponen iguales.

Aplicando el teorema de Fisher a cada una de las poblaciones independientes, tenemos,

Población 1	Población 2
$\frac{(n_1-1)S_1^2}{\sigma_1^2} \rightsquigarrow \chi^2_{n_1-1}$	$\frac{(n_2-1)S_2^2}{\sigma_2^2} \rightsquigarrow \chi^2_{n_2-1}$
$\overline{x}_1 \rightsquigarrow N\left(\mu_1, \frac{\sigma_1}{\sqrt{n_1}}\right)$	$\overline{x}_2 \rightsquigarrow N\left(\mu_2, \frac{\sigma_2}{\sqrt{n_2}}\right)$
independientes	independientes

de donde se deduce que debe ser

$$\begin{cases} \frac{(n_1-1)S_1^2}{\sigma_1^2} + \frac{(n_2-1)S_2^2}{\sigma_2^2} \rightsquigarrow \chi_{n_1+n_2-2} \\ \\ \frac{\overline{x}_1 - \overline{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \\ \\ \text{independientes} \end{cases}$$

Si al construir la t de Student, como cociente entre la $N(0,1)$ y la χ^2 independientes, utilizamos la suposición de ser $\sigma_1 = \sigma_2$, quedará

$$\frac{\overline{x}_1 - \overline{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \rightsquigarrow t_{n_1+n_2-2}$$

σ_1 y σ_2 no se suponen iguales.

En este caso, al construir la t de Student, no se puede llevar a cabo la simplificación antes realizada. De hecho, esta situación no está resuelta completamente, existiendo varias aproximaciones a los grados de libertad f de la distribución t_f del cociente

$$\frac{\overline{x}_1 - \overline{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \approx t_f$$

Aquí consideraremos la *aproximación de Welch* que define f como el entero más próximo a

$$\frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1+1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2+1}} - 2$$

5.9. Estimación de la diferencia de medias de dos poblaciones independientes no necesariamente normales. Muestras grandes

Si las muestras son lo suficientemente grandes como para no poder aplicar el teorema central del límite, digamos $n_1 + n_2 > 30$ en el primer caso que sigue y $n_1 + n_2 > 100$ en el segundo y tercero, además de ser en los tres casos n_1 aproximadamente igual a n_2, tendremos que,

Si σ_1 y σ_2 son conocidas

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

Si σ_1 y σ_2 son desconocidas

$$\frac{\overline{x}_1 - \overline{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \approx N(0,1)$$

Poblaciones binomiales

Si además de ser las muestras grandes, las poblaciones independientes son binomiales: $X_1 \rightsquigarrow B(1, p_1)$ y $X_2 \rightsquigarrow B(1, p_2)$, será

$$\hat{p}_1 - \hat{p}_2 \approx N\left(p_1 - p_2, \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\right)$$

5.10. Datos apareados

En esta sección se estudia el caso en que tengamos pares de datos dependientes $(X_1, Y_1), ..., (X_n, Y_n)$.

En este esquema no podemos suponer ambas poblaciones X e Y como independientes, sino formando una población bidimensional.

La manera de abordar este tipo de datos consiste en definir la variable diferencia $D = X - Y$, la cual ya es unidimensional, y aplicar a los n datos, supuestamente obtenidos de ella, $X_1 - Y_1, ..., X_n - Y_n$, los resultados de las secciones 5.4, 5.5 y 5.6.

5.11. Tamaño muestral para una precisión dada

En estadística es frecuente que se quieran determinar resultados con una determinada precisión, medida ésta en términos de probabilidad.

En general, la situación que se tendrá dependerá del problema que se trate, por lo que aquí no obtendremos expresiones del tamaño muestral en las diversas situaciones posibles. Del enunciado del problema es de donde se obtendrá, para cada caso, una ecuación en n y, utilizando la distribución en le muestreo estadístico que aparezca en la ecuación, se despejará la incognita n.

6. Intervalos de confianza

6.1. Introducción

Los estimadores de la sección anterior estimaban parámetros poblacionales *puntualmente*, no obstante, rara vez coincidirá esta *estimación puntual* con el desconocido valor del parámetro, sin duda, es mucho más interesante concluir la inferencia con un intervalo de posibles valores del parámetro, al que denominaremos *intervalo de confianza*, de manera que, antes de tomar la muestra, el desconocido valor del parámetro se encuentre en dicho intervalo con una probabilidad todo lo alta que deseemos.

Con objeto de aumentar la precisión de la inferencia, serán deseables intervalos de confianza lo más cortos posible.

No obstante, la longitud del intervalo de confianza dependerá de lo alta que queramos sea la probabilidad con la que dicho intervalo, cuyos extremos son aleatorios, cubra a θ. La longitud del intervalo depende de la probabilidad $1-\alpha$ elegida en su construcción, a la que denominaremos *coeficiente de confianza*, y del tamaño muestral (a mayor tamaño muestral n, menor será la longitud del intervalo).

Definición

Supongams que X es la variable aleatoria en estudio, cuya distribución depende de un parámetro desconocido θ, y $X_1, ..., X_n$ una muestra aleatoria simple de dicha variable.

Si $T_1(X_1, ..., X_n)$ y $T_2(X_1, ..., X_n)$ son dos estadísticos tales que

$$P\{T_1(X_1, ..., X_n) \leqslant \theta \leqslant T_2(X_1, ..., X_n)\} = 1 - \alpha$$

el intervalo

$$[T_1(x_1, ..., x_n), T_2(x_1, ..., x_n)]$$

recibe el nombre de *intervalo de confianza* para θ de *coeficiente de confianza* $1-\alpha$.

Obsérvese que tiene sentido hablar de que, antes de tomar la muestra, el intervalo aleatorio

$$[T_1(X_1, ..., X_n), T_2(X_1, ..., X_n)]$$

cubra al verdadero y desconocido valor del parámetro θ con probabilidad $1-\alpha$ pero, una vez elegida una muestra particular $x_1, ..., x_n$, el intervalo no aleatorio

$$[T_1(x_1, ..., x_n), T_2(x_1, ..., x_n)]$$

cubrirá o no a θ, pero ya no tiene sentido hablar de la probabilidad con que lo cubre.

Obsérvese también que el intervalo de confianza es un subconjunto de los posibles valores del parámetro precisamente por ser no aleatorio.

Así mismo mencionemos que cualquier par de estimadores T_1 y T_2 que cumplan la condición impuesta en la definición anterior darán lugar a un intervalo de confianza.

Un resumen de cada uno de estos intervalos aparece en ADD facilitando una rápida consulta.

Respecto a la notación que utilizaremos, tanto en los intervalos de confianza como en el resto del resumen, digamos que denotaremos por $z_p, t_{n;p}, \chi^2_{n;p}$ y $F_{n_1,n_2;p}$, respectivamente, el valor de la abscisa de una distribución $N(0,1)$, t_n de Student, χ^2_n de Pearson y F_{n_1,n_2} de Snedecor, que deja a su derecha, bajo la correspondiente función de densidad, un área de probabilidad p.

6.1.1. Cálculo de intervalos de confianza con R

El intervalo de confianza de un parámetro se corresponde con la región de aceptación de un test bilateral. Por esta razón se utiliza la misma función en R para obtener intervalos de confianza y test de hipótesis sobre un parámetro.

En concreto, la función de R que nos va a proporcionar los itnervalos, es la función `t.test()`. Con ella vamos a poder determinar los intervalos de confianza para la media, para datos apareados y para la diferencia de medias, pero no para aquellos casos en los que la varianza, varianzas o medias poblacionales sean conocidas sino para cuando haya que estimarlas a partir de datos. También queremos advertir que, para poder aplicar esta función, es necesario conocer los datos individualmente ya que no podremos utilizarla cuando sólo conozcamos los valores de las medias o cuasivarianzas muestrales y no los datos de donde éstas proceden.

La función a utilizar en el caso de intervalos de confianza es

```
t.test(x,y=NULL,paired=FLASE,var.equal=FALSE,
conf.level=0.95)
```

Entrando a describir cad uno de sus argumentos, en primer lugar diremos que los valores de aparecen después del simbolo `=` son los que toma la función por defecto y que, por lo tanto, no será necesario especificar si son los valores que deseamos ejecutar. En `x` incorporamos los datos de la muestra, si se trata de inferencias para una sola muestra; si se trata de datos apareados o de dos muestras independientes, introduciremos los datos de la segunda muestra en el argumento `y`.

Si especificamos `paired=F`, estamos en una situación de datos no apareados. Un caso de datos apareados debe especificarse con `paired=T`.

El argumento `var.equal` nos permite indicar qué tipo de situación tenemos en el caso de comparación de dos poblaciones independientes. Si es `var.equal=T` tendremos una situación en la que las varianzas de ambas poblaciones se suponen iguales, y el intervalo será el habitual basado en una t de Student. Si especificamos `var.equal=F` las

varianzas de ambas poblaciones no se suponen iguales y, en este caso, estamos requiriendo un intervalo basado en una t de Student pero en donde los grados de libertad se determinan por la aproximación de Welch.

El último argumento permite especificar el coeficiente de confianza.

El intervalo de confianza para el cociente de varianzas poblacionales se obtiene con la función

```
var.test(x,y,conf.level=0.95)
```

en donde incorporamos los datos en los argumentos `x` e `y`. De nuevo aquí necesitaremos conocer los datos concretos y no admite esta función la situacion de ser las medias poblacionales conocidas.

Por último, en la obtención de intervalos de dos poblaciones binomiales, debemos utilizar la función de R `prop.test()`,

```
prop.test(x,n,conf.level=0.95, correct=TRUE)
```

Los argumentos de esta función son: `x`, vector de éxitos. `n`, vector de número de pruebas realizadas. El último argumento `correct` es utilizado para indicar al ordenador que utilice una corrección de Yates, no considerada en este texto, por lo que deberemos indicar `correct=F` si queremos replicar los resultados obtenidos sin la ayuda de R.

6.2. Intervalo de confianza para la media de una población normal

Tanto en esta sección como en las siguientes, determinaremos intervalos de confianza de *colas iguales*. Es decir, aquellos tales que, si el coeficiente de confianza es $1-\alpha$, dejan a cada uno de los extremos la mitad de la probabilidad, $\alpha/2$.

En esta sección suponemos que los n datos procedes de una población $N(\mu,\sigma)$, y lo que pretendemos determinar es el itnervalo de confianza para la media μ.

Tanto si la varianza poblacional σ^2 es conocida como si no lo es, el estimador natural de μ es la media muestral $\overline{x}$, por lo que determinar un intervalo de confianza para μ significa buscar un número c tal que

$$P\{|\overline{x}-\mu|<c\}=1-\alpha.$$

σ conocida

La distribución en el muestreo de $\overline{x}$ es, en este caso,

$$\overline{x}\rightsquigarrow N\left(\mu,\frac{\sigma}{\sqrt{n}}\right)$$

con lo que, tipificando será

$$P\left\{|Z|<\frac{c}{\sigma/\sqrt{n}}\right\}=1-\alpha$$

es decir,

$$c=\frac{z_{\sigma/2}\cdot\sigma}{\sqrt{n}}.$$

El intervalo buscado será, por tanto,

$$\left[\overline{x}-z_{\sigma/2}\frac{\sigma}{\sqrt{n}},\overline{x}+z_{\sigma/2}\frac{\sigma}{\sqrt{n}}\right].$$

σ desconocida

En este caso la media muestral tiene por distribución

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

con lo que

$$P\{|\overline{x}-\mu|<c\}=1-\alpha$$

será equivalente a

$$P\left\{|t_{n-1}|<\frac{c}{S/\sqrt{n}}\right\}=1-\alpha$$

de donde se obtiene

$$c=\frac{t_{n-1;\sigma/2}\cdot S}{\sqrt{n}}.$$

Así pues, el intervalo de confianza para la media resulta

$$\left[\overline{x}-t_{n-1;\sigma/2}\frac{S}{\sqrt{n}},\overline{x}+t_{n-1;\sigma/2}\frac{S}{\sqrt{n}}\right]$$

en donde S^2 es la cuasivarianza muestral.

6.3. Intervalo de confianza para la media de una población no necesariamente normal. Muestras grandes

Población no necesariamente normal

Si no suponemos modelo alguno para la variable aleatoria en estudio, excepto que tenga varianza σ^2 finita y que la muestra de tamaño n sea suficientemente grande, tenemos dos situaciones posibles dependiendo del conocimiento o no de la varianza poblacional.

Si σ **es conocida** el intervalo de confianza para μ de coeficiente de confianza $1-\alpha$ será

$$I=\left[\overline{x}-z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\overline{x}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

y si σ **es desconocida**

$$I=\left[\overline{x}-z_{\alpha/2}\frac{S}{\sqrt{n}},\overline{x}+z_{\alpha/2}\frac{S}{\sqrt{n}}\right]$$

siendo S la cuasideviación típica muestral.

Población binomial

Si suponemos que $X\rightsquigarrow B(1,p)$ y que la muestra es suficientemente grande, el intervalo de confianza para p de coeficiente de confianza $1-\alpha$ es

$$I=\left[\hat{p}-z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\hat{p}+z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$$

en donde $\hat{p}$ es la proporción muestral.

Población Poisson

Suponiendo que $X\rightsquigarrow\mathcal{P}(\delta)$ y que la muestra es suficientemente grande, el intervalo de confianza para δ de coeficiente de confianza $1-\alpha$ es

$$I=\left[\overline{x}-z_{\alpha/2}\sqrt{\overline{x}/n},\overline{x}+z_{\alpha/2}\sqrt{\overline{x}/n}\right].$$

6.4. Intervalo de confianza para la varianza de una población normal

Dada una muestra aleatoria simple $X_1, ..., X_n$ de una población $N(\mu, \sigma)$, vamos a determinar el intervalo de confianza para σ^2, distinguiendo dos casos según sea desconocida o no la media de la población μ.

μ desconocida

Como es

$$\frac{(n-1)S^2}{\sigma^2} \rightsquigarrow \chi^2_{n-1}$$

podemos encontrar en las tablas de la χ^2 dos abscisas tales que

$$P\left\{\chi^2_{n-1;1-\sigma/2} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{n-1;\sigma/2}\right\} = 1-\alpha$$

de donde, despejando, se obtiene

$$P\left\{\frac{(n-1)S^2}{\chi_{n-1;\sigma/2}} < \sigma^2 < \frac{(n-1)S^2}{\chi_{n-1;1-\alpha/2}}\right\} = 1-\alpha$$

es decir, el intervalo de confianza buscado será

$$I = \left[\frac{(n-1)S^2}{\chi_{n-1;\alpha/2}}, \frac{(n-1)S^2}{\chi^2_{n-1;1-\alpha/2}}\right]$$

con S^2 la cuasivarianza muestral.

μ conocida

En este caso, el intervalo de confianza será

$$I = \left[\frac{\sum_{i=1}^n (X_i-\mu)^2}{\chi_{n;\alpha/2}}, \frac{\sum_{i=1}^n (X_i-\mu)^2}{\chi^2_{n;1-\alpha/2}}\right].$$

6.5. Intervalo de confianza para el cociente de varianzas de dos poblaciones normales independientes

Supondremos que $X_1, ..., X_{n_1}$ e $Y_1, ..., Y_{n_2}$ son dos muestras de tamaños n_1 y n_2 extraídas respectivamente de dos poblaciones independientes $N(\mu_1, \sigma_1)$ y $N(\mu_2, \sigma_2)$.

μ_1 y μ_2 conocidas

En este caso, el intervalo de colas iguales es

$$I = \left[\frac{n_2 \sum_{i=1}^{n_1}(X_i-\mu_1)^2 / \sum_{j=1}^{n_2}(Y_j-\mu_2)^2}{n_1 \cdot F_{n_1,n_2;\alpha/2}}, \frac{n_2 \sum_{i=1}^{n_1}(X_i-\mu_1)^2 / \sum_{j=1}^{n_2}(Y_j-\mu_2)^2}{n_1 \cdot F_{n_1,n_2;1-\alpha/2}}\right]$$

μ_1 y μ_2 desconocidas

Si las medias poblacionales son desconocidas y las muestras proporcionan cuasivarianzas muestrales S_1^2 y S_2^2 respectivamente, el intervalo de confianza que se obtiene es

$$I = \left[\frac{S_1^2/S_2^2}{F_{n_1-1,n_2-1;\alpha/2}}, \frac{S_1^2/S_2^2}{F_{n_1-1,n_2-1;1-\alpha/2}}\right].$$

6.6. Intervalo de confianza para la diferencia de medias de dos poblaciones normales independientes

Suponemos que $X_1, ..., X_{n_1}$ e $Y_1, ..., Y_{n_2}$ son dos muestras de tamaños n_1 y n_2 respectivamente, extraídas de dos poblaciones normales independientes $N(\mu_1, \sigma_1)$ y $N(\mu_2, \sigma_2)$.

σ_1 y σ_2 conocidas

En este caso sabemos que es

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

de donde el intervalo de confianza buscado será

$$I = \left[\overline{x}_1 - \overline{x}_2 - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, \overline{x}_1 - \overline{x}_2 + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right]$$

σ_1 y σ_2 desconocidas. Muestras pequeñas

En esta situación habrá que distinguir según sean

(a) $\sigma_1 = \sigma_2$ En cuyo caso obtendremos como intervalo de confianza

$$I = \left[\overline{x}_1 - \overline{x}_2 \mp t_{n_1+n_2-2;\alpha/2}\sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right]$$

(b) $\sigma_1 \neq \sigma_2$ En este caso, la aproximación de Welch proporciona como intervalo de confianza

$$I = \left[\overline{x}_1 - \overline{x}_2 - t_{f;\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}, \overline{x}_1 - \overline{x}_2 + t_{f;\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}\right]$$

en donde S_1^2 y S_2^2 son las cuasivarianzas muestrales y f el entero más próximo a

$$\frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1+1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2+1}} - 2$$

6.7. Intervalo de confianza para la diferencia de medias de dos poblaciones independientes no necesariamente normales. Muestas grandes

Si ahora $X_1, ..., X_{n_1}$ e $Y_1, ..., Y_{n_2}$ son dos muestras de tamaños n_1 y n_2 suficientemente grandes, extraídas de dos poblaciones independientes de medias μ_1 y μ_2 respectivamente, de las que sólo suponemos que tienen varianzas σ_1^2 y σ_2^2 finitas, tendremos que

Si σ_1 y σ_2 son conocidas

El intervalo de confianza para $\mu_1 - \mu_2$ con un coeficiente de confianza $1 - \alpha$ es

$$I = \left[\overline{x}_1 - \overline{x}_2 - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, \overline{x}_1 - \overline{x}_2 + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right]$$

Si σ_1 y σ_2 son desconocidas

El intervalo de confianza se obtendrá sustituyendo las desconocidas varianzas por las cuasivarianzas muestrales, S_1^2 y S_2^2, obteniendose

$$I = \left[\overline{x}_1 - \overline{x}_2 - z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}, \overline{x}_1 - \overline{x}_2 + z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \right]$$

Poblaciones binomiales

Por último, si admitimos, además de que los tamaños muestrales n_1 y n_2 son grandes, el que las poblaciones independientes son binomiales $X_1 \rightsquigarrow B(1, p_1)$ y $X_2 \rightsquigarrow B(1, p_2)$, la distribución aproximada en el muestreo de la diferencia de proporciones muestrales es

$$\hat{p}_1 - \hat{p}_2 \approx \left(p_1 - p_2, \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \right)$$

conduce al siguiente intervalo de confianza para la diferencia de proporciones poblacionales $p_1 - p_2$

$$I = \left[\hat{p}_1 - \hat{p}_2 \mp z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \right].$$

6.8. Intervalos de confianza para datos apareados

Si la muestra que tenemos es de datos emparejados $(X_1, Y_1), ..., (X_n, Y_n)$, en el sentido de proceder de una población bidimensional, la forma de actuar consiste en definir la variable unidimensional diferencia $D = X_i - Y_i$ y aplicar a sus parámetros los intervalos de confianza antes determinados.

7. Contraste de hipótesis

7.1. Introducción y conceptos fundamentales

El primer punto a considerar en un contraste de hipótesis es precisamente el establecer las *hipótesis* que se quieren contrastar, es decir, comparar.

Una de las hipótesis, generalmente la que corresponde a la situación estándar, recibe el nombre de *hipótesis nula* H_0, mientras que la otra recibe el nombre de *hipótesis alternativa* H_1, siendo el *contraste de hipótesis* el proceso de decisión basado en técnicas estadísticas mediante el cual decidimos (inferimos) cuál de las hipótesis creemos correcta, aceptándola y rechazando en consecuencia la otra.

Si X representa la variable en observación, el contraste de hipótesis concluirá formulando una regla de actuación, denominada también contraste de hipótesis o por no ser excesivamente redundantes, *test de hipótesis*, la cual estará basada en una muestra de X de tamaño n, $X_1, ..., X_n$, o más en concreto en una función suya denominada *estadístico del contraste* $T(X_1, ..., X_n)$, y que habitualmente será una función del estimador natural asociado al parámetro del que se quiere contrastar las hipótesis.

En la realización de un contraste de hipótesis suuele ser habitual suponer un modelo probabilístico para la variable X, en cuyo caso hablaremos de *contrastes paramétricos*, en contraposición con los denominados *contraste no paramétricos*, en los que sólo serán necesarias suposiciones generales sobre el modelo probabilístico, tales como la simetría o continuidad de éste.

En todo caso, será imprescindible determinar la distribución en el muestreo estadístico T del test, pudiendo formularse de la siguiente forma: si fuera cierta la hipótesis nula H_0, la muestra, o mejor T, debería de comportarse de una determinada manera (tener una determinada distribución de probabilidad). Si extraída una muestra al azar, acontece un suceso para T que tenía poca probabilidad de ocurrir si fuera cierta H_0, puede haber ocurrido una de las dos cosas siguientes: o bien hemos tenido tan mala suerte de haber elegido una muestra *muy rara* o, lo más probable, que la hipótesis nula era falsa. La filosofía del contraste de hipótesis consiste en admitir la segunda posibilidad, rechazando en ese caso H_0, aunque acotando la probabilidad de la primer posibilidad.

Errores de tipo I y de tipo II

Debemos considerar los dos errores posibles que podemos cometer al realizar un contraste de hipótesis, los cuales son el de rechazar la hipótesis nula H_0 cuando es cierto, denominado *error de tipo I*, o el de aceptar H_0 cuando es falsa, denominado *error de tipo II*.

La estadística matemática ha deducido tests de hipótesis, es decir reglas de actuación, siguiendo el criterio de fijar una cota superior para la probabilidad de error tipo I, denominada *nivel de significación*, que maximizan $1 - P\{\text{error de tipo II}\}$, expresión ésta última denominada *potencia del contraste*.

Región crítica y región de aceptación

Los tests de hipótesis, expresados siempre en función de un estadístico T adecuado al problema en cuestión, son de la forma

$$\begin{cases} \text{Aceptar } H_0 & \text{si } T \in C* \\ \text{Rechazar } H_0 & \text{si } T \in C \end{cases}$$

en donde C y $C*$ son dos conjuntos disjuntos en los que se ha dividido el conjunto de valores posibles de T. C recibe el nombre de *región crítica* del test, y se corresponde con el conjunto de valores de T en donde se rechaza la hipótesis nula H_0.

El conjunto complementario, $C*$, se denomina *región de aceptación* y se corresponde con el conjunto de valores del estadístico para los cuales se acepta H_0.

Por complementar la terminología propia de los contrastes de hipótesis, diremos que un test es *bilateral* cuando C esté formada por dos intervalos disjuntos y *unilateral* cuando la región crítica sea un intervalo.

Por último, se dice que una hipótesis, nula o alternativa, es *simple* cuando esté formada por un solo valor de parámetro. Si está formada por más de uno, se denomina *compuesta*.

Es decir, si $H_0 : \mu = \mu_0$ fuera cierta, cabría esperar que $\overline{x}$ tomara un valor cercano a μ_0; en concreto del intervalo $[\mu_0 - c, \mu_o + c]$, con gran probabilidad, $1 - \alpha$, dependiendo el valor de c d esta probabilidad.

Si observada una muestra concreta, $\overline{x}$ no cae en el intervalo anterior, rechazaremos H_0, siendo, en consecuencia el mencionado intervalo, la región de aceptación del test.

Determinemos el valor de la constante c: si queremos que la probabilidad de cometer un error del tipo I, es decir, el nivel de significación sea α, deberá ser

$$P\{\overline{x} \in C\} = P\{|\overline{x} - \mu_0| > c\} = \alpha$$

es decir,

$$P\{|\overline{x} - \mu_0 < c|\} = 1 - \alpha$$

cuando H_0 es cierta, es decir cuando $\mu = \mu_0$.

La deducción exacta de cada contraste óptimo depende de la situación concreta que se tenga: hipótesis de normalidad, muestras grandes, etc., ya que cada una de estas situaciones implica una distribución en el muestreo del estadístico a considerar.

Relación entre intervalos de confianza y tests de hipótesis

Si admitimos un modelo poblacional normal, es decir que $X \rightsquigarrow N(\mu, \sigma)$, aceptamos $H_0 : \mu = \mu_0$ cuando

$$\frac{|\overline{x} - \mu_0|}{S/\sqrt{n}} \leqslant t_{n-1;\alpha/2}$$

o bien, haciendo operaciones, cuando

$$\mu_0 \in \left[\overline{x} - t_{n-1;\alpha/2}\frac{S}{\sqrt{n}}, \overline{x} + t_{n-1;\alpha/2}\frac{S}{\sqrt{n}}\right]$$

es decir, cuando la hipótesis nula pertenece al intervalo de confianza correspondiente.

Éste es un hecho bastante frecuente, aunque no una propiedad general, de los contrastes de tipo $H_0 : \theta = \theta_0$ frente a $H_0 : \theta \neq \theta_0$. El intervalo de confianza, de coeficiente de confianza uno menos el nivel de significación, constituye la región de aceptación del test.

Tests de hipótesis unilaterales

Supongamos que queremos contrastar las hipótesis $H_0 : \mu \leqslant \mu_0$ frente a $H_1 : \mu > \mu_0$. Ahora parece claro que la región crítica sea unilateral del tipo $\mu_0 + c$.

Si la probabilidad de error tipo I es de nuevo α, deberá ser

$$P_{\mu=\mu_0}\{\overline{x} > \mu_0 + c\} = \alpha.$$

Si admitimos la misma situación poblacional anterior, será la distribución de $\overline{x}$ de nuevo

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

con lo que en la expresión anterior, c deberá ser tal que

$$P\left\{t_{n-1} > \frac{c\sqrt{n}}{S}\right\} = \alpha$$

es decir,

$$c = t_{n-1;\alpha}\frac{S}{\sqrt{n}}$$

con lo que se llegaría en definitiva, a considerar como test de nivel α para contrastar $H_0 : \mu \leqslant \mu_0$ frente a $H_1 : \mu > \mu_0$ el siguiente,

$$\begin{cases} \text{Se acepta } H_0 & \text{si } \frac{\overline{x}-\mu}{S/\sqrt{n}} \leqslant t_{n-1;\alpha} \\ \text{Se rechaza } H_0 & \text{si } \frac{\overline{x}-\mu}{S/\sqrt{n}} > t_{n-1;\alpha} \end{cases}$$

P-valor

Una crítica respecto a la técnica de los tests de hipótesis es la dependencia de nuestros resultados en el nivel de significación α elegido antes de efectuar el contraste.

Parece razonable, que independientemente del nivel de significación que hubiéramos elegido, debamos aceptar H_0, si el nivel de significación más pequeño que hubiéramos tenido que elegir para rechazar H_0 es demasiado grande como para admitir tal probabilidad de error tipo I.

Este *nivel de significación observado* recibe el nombre de *p-valor* y se define con más precisión como el mínimo nivel de significación necesario para rechazar H_0.

Obsérvese que al realizar un contraste de hipótesis debemos fijar un nivel de significación antes de tomar la muestra y para ese nivel de significación elegido, aceptar o rechazar H_0. Es decir, siempre se llega a una conclusión.

El cálculo del p-valor permite valorar la decisiń ya tomada de rechazar o aceptar H_0, de forma que un p-valor grande (digamos 0,2 o más) confirma una decisión de aceptación de H_0. Tanto más nos lo confirma cuanto mayor sea el p-valor.

Por contra, un p-valor pequeño (digamos 0,01 o menos) confirma una decisión de rechazo de H_0. Tanto más se nos confirmará esta decisión de rechazo cuanto menor sea el p-valor.

En situaciones intermedias, el p-valor no nos indica nada en contreto salvo que quizás sería recomendable elegir otra muestra y volver a realizar el contraste.

Si una persona ha tomado una decisión que el p-valor contradice el individuo lógicamente cambiará su decisión. Por esta razón, muchas técnicas estadísticas aplicadas no fijan el nivel de significación, simplemente hacen aparecer al final de sus el p-valor (*tail probability*), sacandose conclusiones si éste se lo permite o simplemente indicándolo de forma que el lector las saque.

Contrastes óptimos

Ante una situación concreta que se nos plantee, la determinación del contraste óptimo dependerá fundamentalmente de las suposiciones que se hagan en el modelo y de las hipótesis que se desee contrastar.

Contraste de hipótesis con R

El intervalo de confianza de un parámetro se corresponde con la región de aceptación de un test de hipótesis bilateral. Por esta razón se utiliza una misma función de R para obtener intervalos de confianza y test de hipótesis sobre un parámetro. En concreto, la función de R que nos va a proporcionar los tests es la función `t.test()`,

```
t.test(x,y=NULL,alternative="two.sided",mu=0,
paired=FALSE,var.equal=FALSe,conf.level=0.95)
```

Los argumentos `x` e `y` se utilizan para indicar el o los vectores de datos a utilizar en el contraste. El tercer argumento `alternative` presenta tres opciones: `two.sided`, que es la que se utiliza por defecto y que corresponde al caso de contrastes bilaterales; `greater`, correspondiente al caso de hipótesis nula *menor o igual* frende a hipótesis alternativa *mayor*, y `less` para el caso de hipótesis nula de *mayor o igual* frente a alternativa *menor*. Con el argumento `mu` indicamos el valor de la hipótesis nula.

De nuevo *paired* sirve para indicar una situación de datos apareados y `var.equal` si las varianzas poblacionales pueden considerarse o no iguales. El útlimo argumento permite especificar el nivel de significación del test tomándose por defecto el valor 0,05.

7.2. Contraste de hipótesis relativas a la media de una población normal

Supongamos que tenemos una muestra aleatoria simple $X_1, ..., X_n$ procedente de una población $N(\mu, \sigma)$ y que queremos contrastar hipótesis relativas a la media de la población, μ.

En primer lugar consideraremos el caso de *igual* frente a *distina*, es decir, el caso en que queremos contrastar si puede admitirse para la media poblacional un determinado valor μ_0 o no.

$$\boxed{\begin{array}{l} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{array}}$$

Así, si suponemos σ **conocida**, fijado un nivel de significación α, aceptaremos $H_0 : \mu = \mu_0$ cuando y sólo cuando

$$\mu_0 \in \left[\overline{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \overline{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right]$$

con lo que podemos concluir diciendo que el test óptimo en esta situación es

$$\begin{cases} \text{Se acepta } H_0 & \text{si } \frac{|\overline{x}-\mu_0|}{\sigma/\sqrt{n}} \leqslant z_{\alpha/2} \\ \text{Se rechaza } H_0 & \text{si } \frac{|\overline{x}-\mu_0|}{\sigma/\sqrt{n}} > z_{\alpha/2} \end{cases}$$

Si se supone a σ **desconocida**, el test óptimo en este caso es

$$\begin{cases} \text{Se acepta } H_0 & \text{si } \frac{|\overline{x}-\mu_0|}{S/\sqrt{n}} \leqslant t_{n-1;\alpha/2} \\ \text{Se rechaza } H_0 & \text{si } \frac{|\overline{x}-\mu_0|}{S/\sqrt{n}} > t_{n-1;\alpha/2} \end{cases}$$

a nivel de significación α.

$$\boxed{\begin{array}{l} H_0 : \mu \leqslant \mu_0 \\ H_1 : \mu > \mu_0 \end{array}}$$

En estos casos, el objetivo es rechazar H_0 con un p-valor pequeño, lo que conduce a quedarnos con la hipótesis de interés H_1, con un error pequeño en la inferencia, el error de rechazar H_0 siendo cierta, error suministrado por el p-valor.

La distribución en el muestreo de $\overline{x}$ en los supuestos que se establecen, así como las consideraciones hechas al hablar de las hipótesis unilateraless, llevan a la estadística matemática a proponer como test óptimo para contrastar $H_0 : \mu \leqslant \mu_0$ frente a $H_1 : \mu > \mu_0$,

Si σ es conocida

El test óptimo indica que

$$\begin{cases} \text{Se acepta } H_0 & \text{si } \frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}} \leqslant z_{\alpha} \\ \text{Se rechaza } H_0 & \text{si } \frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}} > z_{\alpha} \end{cases}$$

Si σ es desconocida

En este caso, el test óptimo indica que

$$\begin{cases} \text{Se acepta } H_0 & \text{si } \frac{\overline{x}-\mu_0}{S/\sqrt{n}} \leqslant t_{n-1;\alpha} \\ \text{Se rechaza } H_0 & \text{si } \frac{\overline{x}-\mu_0}{S/\sqrt{n}} > t_{n-1;\alpha} \end{cases}$$

$$\boxed{\begin{array}{l} H_0 : \mu \geqslant \mu_0 \\ H_1 : \mu < \mu_0 \end{array}}$$

Los mismos razonamientos anteriores llevan a proponer los siguientes tests para las hipótesis simétricas aquí consideradas.

Si σ es conocida

$$\begin{cases} \text{Se acepta } H_0 & \text{si } \frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}} \geqslant z_{1-\alpha} \\ \text{Se rechaza } H_0 & \text{si } \frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}} < z_{1-\alpha} \end{cases}$$

Si σ es desconocida

$$\begin{cases} \text{Se acepta } H_0 & \text{si } \frac{\overline{x}-\mu_0}{S/\sqrt{n}} \geqslant t_{n-1;1-\alpha} \\ \text{Se rechaza } H_0 & \text{si } \frac{\overline{x}-\mu_0}{S/\sqrt{n}} < t_{n-1;1-\alpha} \end{cases}$$

7.3. Contraste de hipótesis relativas a la media de una población no necesariamente normal. Muestras grandes

La obtención de muestras suficientemente grandes, digamos mayores de 30, evita la obligación de suponer normalidad en la distribución, alcanzándose, no obstante, resultados análogos a cuando se verifica tal suposición.

La normalidad en la distribución asintótica de $\overline{x}$, añade la peculiaridad de hacer que los puntos críticos sean ahora abscisas de normales estándar, tanto si la varianza poblacional es conocida como si no lo es.

Población no necesariamente normal

Supongamos que $X_1, ..., X_n$ es una muestra aleatoria simple de tamaño suficientemente grane como para poder admitir como distribución asintótica de $\overline{x}$ la siguiente,

$$\overline{x} \approx N\left(\mu, \frac{\sigma}{\sqrt{n}}\right).$$

En este caso, consideramos los tres tipos de tests y distinguiendo, de nuevo, la situación en la que la varianza es conocida y la situación en la que es desconocida, tenemos los siguientes contrastes,

$$\boxed{\begin{array}{l} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{array}}$$

σ conocida

El test óptimo que se propone es la siguiente regla de actuación

$$\begin{cases} \text{Se acepta } H_0 & \text{si } \frac{|\overline{x}-\mu_0|}{\sigma/\sqrt{n}} \leqslant z_{\alpha/2} \\ \text{Se rechaza } H_0 & \text{si } \frac{|\overline{x}-\mu_0|}{\sigma/\sqrt{n}} > z_{\alpha/2} \end{cases}$$

σ desconocida

Si σ es desconocida, entonces el test óptimo es

$$\begin{cases} \text{Se acepta } H_0 & \text{si } \frac{|\overline{x}-\mu_0|}{S/\sqrt{n}} \leqslant z_{\alpha/2} \\ \text{Se rechaza } H_0 & \text{si } \frac{|\overline{x}-\mu_0|}{S/\sqrt{n}} > z_{\alpha/2} \end{cases}$$

$$\boxed{\begin{array}{l} H_0 : \mu \leqslant \mu_0 \\ H_1 : \mu > \mu_0 \end{array}}$$

Si σ es conocida

$$\begin{cases} \text{Se acepta } H_0 & \text{si } \frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}} \leqslant z_{\alpha} \\ \text{Se rechaza } H_0 & \text{si } \frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}} > z_{\alpha} \end{cases}$$

Si σ es desconocida

$$\begin{cases} \text{Se acepta } H_0 & \text{si } \frac{\overline{x}-\mu_0}{S/\sqrt{n}} \leqslant z_{\alpha} \\ \text{Se rechaza } H_0 & \text{si } \frac{\overline{x}-\mu_0}{S/\sqrt{n}} > z_{\alpha} \end{cases}$$

$$\boxed{\begin{array}{l} H_0 : \mu \geqslant \mu_0 \\ H_1 : \mu < \mu_0 \end{array}}$$

Si σ es conocida

$$\begin{cases} \text{Se acepta } H_0 & \text{si } \frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}} \geqslant z_{1-\alpha} \\ \text{Se rechaza } H_0 & \text{si } \frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}} < z_{1-\alpha} \end{cases}$$

Si σ es desconocida

$$\begin{cases} \text{Se acepta } H_0 & \text{si } \frac{\overline{x}-\mu_0}{S/\sqrt{n}} \geqslant z_{1-\alpha} \\ \text{Se rechaza } H_0 & \text{si } \frac{\overline{x}-\mu_0}{S/\sqrt{n}} < z_{1-\alpha} \end{cases}$$

Población binomial

La situación que se analiza en este apartado es la de una muestra aleatoria simple $X_1, ..., X_n$ de variables $B(1, p)$, es decir, variables que toman sólo los valores 1 *(éxito)*, 0 *(fracaso)*, siendo la probabilidad de éxito el parámetro p.

$$\boxed{\begin{array}{l} H_0 : p = p_0 \\ H_1 : p \neq p_0 \end{array}}$$

Como el resto de los contraste de *igual* frente a *distinta*, la región de aceptación del test óptimo se corresponde con el intervalo de confianza, aceptándose $H_0 : \mu = \mu_0$ cuando μ_0 pertenezca a dicho intervalo. Hacemos la observación de que allí, al ser la varianza de la propoción muestral, $p(1-p)/n$, desconocida por depender del parámetro p, la estimamos con la proporción muestral mediante $\hat{p}(1-\hat{p})/n$. Aquí sin embargo, dado que los test se realizan bajo H_0, es decir, suponiendo que es cierta la hipótesis nula, ésta implica suponer como varianza de $\hat{p}$ el valor $p_0(1-p_0)/n$ (si es $p_0 \neq 0$).

En el caso que aquí nos ocupa,

$$\begin{cases} \text{Se acepta } H_0 & \text{si } \frac{|\hat{p}-p_0|}{\sqrt{\frac{p_0(1-p_0)}{n}}} \leqslant z_{\alpha/2} \\ \text{Se rechaza } H_0 & \text{si } \frac{|\hat{p}-p_0|}{\sqrt{\frac{p_0(1-p_0)}{n}}} > z_{\alpha/2} \end{cases}$$

$$\boxed{\begin{array}{l} H_0 : p \leqslant p_0 \\ H_1 : p > p_0 \end{array}}$$

En esta situación el test óptimo es

$$\begin{cases} \text{Se acepta } H_0 & \text{si } \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \leqslant z_{\alpha} \\ \text{Se rechaza } H_0 & \text{si } \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} > z_{\alpha} \end{cases}$$

$$\boxed{\begin{array}{l} H_0 : p \geqslant p_0 \\ H_1 : p < p_0 \end{array}}$$

$$\begin{cases} \text{Se acepta } H_0 & \text{si } \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \geqslant z_{1-\alpha} \\ \text{Se rechaza } H_0 & \text{si } \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} < z_{1-\alpha} \end{cases}$$

Población Poisson

La situación que tenemos en este apartado es la de una muestra aleatoria simple $X_1, ..., X_n$, también de tamaño suficientemente grande, n al menos 30, extrída de una población de Poisson de parámetro λ, $\mathcal{P}(\lambda)$, siendo los contrastes a considerar relativos a dicho parámetro.

Obsérvese que estos contrastes aparecen en la sección relativa a los contrastes de la media; de hecho λ es la media de la distribución.

Recuérdese, sin embargo, que λ también es su varianza, por lo que si queremos hacer contrastes sobre la varianza de una población y, después de analizada ésta, se considera en ella un modelo de Poisson, los contrastes erlativos a su varianza son los que aparecen a continuación.

$$\boxed{\begin{array}{l} H_0 : \lambda = \lambda_0 \\ H_1 : \lambda \neq \lambda_0 \end{array}}$$

En este caso, el test óptimo sugiere que

$$\begin{cases} \text{Se acepta } H_0 & \text{si } \frac{|\overline{x}-\lambda_0|}{\sqrt{\lambda_0/n}} \leqslant z_{\alpha/2} \\ \text{Se rechaza } H_0 & \text{si } \frac{|\overline{x}-\lambda_0|}{\sqrt{\lambda_0/n}} > z_{\alpha/2} \end{cases}$$

$$\boxed{\begin{array}{l} H_0 : \lambda \leqslant \lambda_0 \\ H_1 : \lambda > \lambda_0 \end{array}}$$

En esta situación, el contraste óptimo indica seguir la siguiente regla de actuación:

$$\begin{cases} \text{Se acepta } H_0 & \text{si } \frac{\overline{x}-\lambda_0}{\sqrt{\lambda_0/n}} \leqslant z_{\alpha} \\ \text{Se rechaza } H_0 & \text{si } \frac{\overline{x}-\lambda_0}{\sqrt{\lambda_0/n}} > z_{\alpha} \end{cases}$$

$$\boxed{\begin{array}{l} H_0 : \lambda \geqslant \lambda_0 \\ H_1 : \lambda < \lambda_0 \end{array}}$$

$$\begin{cases} \text{Se acepta } H_0 & \text{si } \frac{\overline{x}-\lambda_0}{\sqrt{\lambda_0/n}} \geqslant z_{1-\alpha} \\ \text{Se rechaza } H_0 & \text{si } \frac{\overline{x}-\lambda_0}{\sqrt{\lambda_0/n}} < z_{1-\alpha} \end{cases}$$

7.4. Contraste de hipótesis relativas a la varianza de una población normal

En toda la sección supondremos que tenemos una muestra $X_1, ..., X_n$ de una población normal $N(\mu, \sigma)$ y que estamos interesados en realizar contrastes sobre la varianza de dicha distribución.

Apuntemos, además, que las hipótesis referentes a la desviación típica se contrastarían utilizando las raíces cuadradas de los tests que aparecen a continuación.

$$\boxed{\begin{array}{l} H_0 : \sigma^2 = \sigma_0^2 \\ H_1 : \sigma^2 \neq \sigma_0^2 \end{array}}$$

μ conocida

Si la media es conocida, el test óptimo a utilizar de nivel de significación α, es

$$\begin{cases} \text{Se acepta } H_0 & \text{si } \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma_0^2} \in \left[\chi^2_{n;1-\frac{\alpha}{2}}, \chi^2_{n;\frac{\alpha}{2}}\right] \\ \text{Se rechaza } H_0 & \text{si } \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma_0^2} \notin \left[\chi^2_{n;1-\frac{\alpha}{2}}, \chi^2_{n;\frac{\alpha}{2}}\right] \end{cases}$$

μ desconocida

En este caso la regla a utilizar será

$$\begin{cases} \text{Se acepta } H_0 & \text{si } \frac{(n-1)S^2}{\sigma_0^2} \in \left[\chi^2_{n-1;1-\frac{\alpha}{2}}, \chi^2_{n-1;\frac{\alpha}{2}}\right] \\ \text{Se rechaza } H_0 & \text{si } \frac{(n-1)S^2}{\sigma_0^2} \notin \left[\chi^2_{n-1;1-\frac{\alpha}{2}}, \chi^2_{n-1;\frac{\alpha}{2}}\right] \end{cases}$$

$$\boxed{\begin{array}{l} H_0 : \sigma^2 \leqslant \sigma_0^2 \\ H_1 : \sigma^2 > \sigma_0^2 \end{array}}$$

μ conocida

En este caso el test óptimo es

$$\begin{cases} \text{Se acepta } H_0 & \text{si } \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma_0^2} \leqslant \chi^2_{n;\alpha} \\ \text{Se rechaza } H_0 & \text{si } \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma_0^2} > \chi^2_{n;\alpha} \end{cases}$$

μ desconocida

$$\begin{cases} \text{Se acepta } H_0 & \text{si } \frac{(n-1)S^2}{\sigma_0^2} \leqslant \chi^2_{n-1;\alpha} \\ \text{Se rechaza } H_0 & \text{si } \frac{(n-1)S^2}{\sigma_0^2} > \chi^2_{n-1;\alpha} \end{cases}$$

$$\boxed{\begin{array}{l} H_0 : \sigma^2 \geqslant \sigma_0^2 \\ H_1 : \sigma^2 < \sigma_0^2 \end{array}}$$

μ conocida

En esta situación, el test óptimo indica que

$$\begin{cases} \text{Se acepta } H_0 & \text{si } \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma_0^2} \geqslant \chi^2_{n;1-\alpha} \\ \text{Se rechaza } H_0 & \text{si } \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma_0^2} < \chi^2_{n;1-\alpha} \end{cases}$$

μ desconocida

$$\begin{cases} \text{Se acepta } H_0 & \text{si } \frac{(n-1)S^2}{\sigma_0^2} \geqslant \chi^2_{n-1;1-\alpha} \\ \text{Se rechaza } H_0 & \text{si } \frac{(n-1)S^2}{\sigma_0^2} < \chi^2_{n-1;1-\alpha} \end{cases}$$

7.5. Contraste de hipótesis relativas a las varianzas de dos poblaciones normales independientes

En esta sección se aborda el problema de la comparación de las varianzas de dos poblaciones normales independientes $N(\mu_1, \sigma_1)$ y $N(\mu_2, \sigma_2)$, utilizando muestras aleatorias de ambas poblaciones $X_1, ..., X_{n_1}$ e $Y_1, ..., Y_{n_2}$.

$$\boxed{\begin{array}{l} H_0 : \sigma_1^2 = \sigma_2^2 \\ H_1 : \sigma_1^2 \neq \sigma_2^2 \end{array}}$$

μ_1 y μ_2 conocidas

Si las medias poblacionales son conocidas, el test óptimo es

$$\begin{cases} \text{Se acepta } H_0 \quad \text{si } \dfrac{\dfrac{\sum_{i=1}^{n_1}(X_i-\mu_1)^2/n_1}{n_2}}{\sum_{j=1}^{}(Y_j-\mu_2)^2/n_2} \in \left[F_{n_1,n_2;1-\frac{\sigma}{2}}, F_{n_1,n_2;\frac{\sigma}{2}}\right] \\ \text{Se rechaza } H_0 \quad \text{si } \dfrac{\dfrac{\sum_{i=1}^{n_1}(X_i-\mu_1)^2/n_1}{n_2}}{\sum_{j=1}^{}(Y_j-\mu_2)^2/n_2} \notin \left[F_{n_1,n_2;1-\frac{\sigma}{2}}, F_{n_1,n_2;\frac{\sigma}{2}}\right] \end{cases}$$

μ_1 y μ_2 desconocidas

El test óptimo es

$$\begin{cases} \text{Se acepta } H_0 \quad \text{si } \frac{S_1^2}{S_2^2} \in \left[F_{n_1-1,n_2-1;1-\frac{\alpha}{2}}, F_{n_1-1,n_2-1;\frac{\alpha}{2}}\right] \\ \text{Se rechaza } H_0 \quad \text{si } \frac{S_1^2}{S_2^2} \notin \left[F_{n_1-1,n_2-1;1-\frac{\alpha}{2}}, F_{n_1-1,n_2-1;\frac{\alpha}{2}}\right] \end{cases}$$

$$\boxed{\begin{array}{l} H_0 : \sigma_1^2 \leqslant \sigma_2^2 \\ H_1 : \sigma_1^2 > \sigma_2^2 \end{array}}$$

μ_1 y μ_2 conocidas

$$\begin{cases} \text{Se acepta } H_0 \quad \text{si } \dfrac{\dfrac{\sum_{i=1}^{n_1}(X_i-\mu_1)^2/n_1}{n_2}}{\sum_{j=1}^{}(Y_j-\mu_2)^2/n_2} \leqslant F_{n_1,n_2;\alpha} \\ \text{Se rechaza } H_0 \quad \text{si } \dfrac{\dfrac{\sum_{i=1}^{n_1}(X_i-\mu_1)^2/n_1}{n_2}}{\sum_{j=1}^{}(Y_j-\mu_2)^2/n_2} > F_{n_1,n_2;\alpha} \end{cases}$$

μ_1 y μ_2 desconocidas

$$\begin{cases} \text{Se acepta } H_0 \quad \text{si } \frac{S_1^2}{S_2^2} \leqslant F_{n_1-1,n_2-1;\alpha} \\ \text{Se rechaza } H_0 \quad \text{si } \frac{S_1^2}{S_2^2} > F_{n_1-1,n_2-1;\alpha} \end{cases}$$

$$\boxed{\begin{array}{l} H_0 : \sigma_1^2 \geqslant \sigma_2^2 \\ H_1 : \sigma_1^2 < \sigma_2^2 \end{array}}$$

μ_1 y μ_2 conocidas

$$\begin{cases} \text{Se acepta } H_0 \quad \text{si } \dfrac{\dfrac{\sum_{i=1}^{n_1}(X_i-\mu_1)^2/n_1}{n_2}}{\sum_{j=1}^{}(Y_j-\mu_2)^2/n_2} \geqslant F_{n_1,n_2;1-\alpha} \\ \text{Se rechaza } H_0 \quad \text{si } \dfrac{\dfrac{\sum_{i=1}^{n_1}(X_i-\mu_1)^2/n_1}{n_2}}{\sum_{j=1}^{}(Y_j-\mu_2)^2/n_2} < F_{n_1,n_2;1-\alpha} \end{cases}$$

μ_1 y μ_2 desconocidas

$$\begin{cases} \text{Se acepta } H_0 \quad \text{si } \frac{S_1^2}{S_2^2} \geqslant F_{n_1-1,n_2-1;1-\alpha} \\ \text{Se rechaza } H_0 \quad \text{si } \frac{S_1^2}{S_2^2} < F_{n_1-1,n_2-1;1-\alpha} \end{cases}$$

7.6. Contraste de hipótesis relativas a la diferencia de medias de dos poblaciones normales independientes

La situación que tenemos aquí planteada es la de dos poblaciones normales $N(\mu_1, \sigma_1)$ y $N(\mu_2, \sigma_2)$, de las que hemos extrído sendas muestar aleatorias independientes de tamaños n_1 y n_2 respectivamente, $X_1, ..., X_{n_1}$ e $Y_1, ..., Y_{n_2}$, representando, como siempre, por $\overline{x}_1$, S_1^2 y por $\overline{x}_2$, S_2^2 la media y cuasivarianza de la primera y segunda muestra respectivamente.

$$\boxed{\begin{array}{l} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 \neq \mu_2 \end{array}}$$

σ_1 y σ_2 conocidas

En este caso el test óptimo es

$$\begin{cases} \text{Se acepta } H_0 \quad \text{si } \frac{|\overline{x}_1-\overline{x}_2|}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \leqslant z_{\alpha/2} \\ \text{Se rechaza } H_0 \quad \text{si } \frac{|\overline{x}_1-\overline{x}_2|}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} > z_{\alpha/2} \end{cases}$$

σ_1 y σ_2 desconocidas. Muestras pequeñas

Si las muestras no son suficientemente grandes, con objeto de determinar el test óptimo, no sólo en el contraste bilateral, sino también en los unilaterales, habrá que distinguir los casos en que las varianzas poblacionales puedan considerarse iguales y aquellos en los que no puedan considerarse iguales.

(a) $\sigma_1 = \sigma_2$

Si las varianzas poblacionales se puede considerar iguales, entonces el test óptimo es

$$\begin{cases} \text{Se acepta } H_0 \text{ si} \\ \dfrac{|\overline{x}_1-\overline{x_2}|}{\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \leqslant t_{n_1+n_2-2;\alpha/2} \\ \text{Se rechaza } H_0 \text{ si} \\ \dfrac{|\overline{x}_1-\overline{x_2}|}{\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} > t_{n_1+n_2-2;\alpha/2} \end{cases}$$

(b) $\sigma_1 \neq \sigma_2$

En el caso de que las varianzas poblacionales no puedan considerarse iguales, el test óptimo es

$$\begin{cases} \text{Se acepta } H_0 \quad \text{si} \quad \dfrac{|\overline{x}_1-\overline{x}_2|}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}} \leqslant t_{f;\alpha/2} \\ \text{Se rechaza } H_0 \quad \text{si} \quad \dfrac{|\overline{x}_1-\overline{x}_2|}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}} > t_{f;\alpha/2} \end{cases}$$

$$\boxed{\begin{array}{l} H_0 : \mu_1 \leqslant \mu_2 \\ H_1 : \mu_1 > \mu_2 \end{array}}$$

σ_1 y σ_2 conocidas

La regla óptima es

$$\begin{cases} \text{Se acepta } H_0 \quad \text{si} \quad \dfrac{\overline{x}_1-\overline{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \leqslant z_\alpha \\ \text{Se rechaza } H_0 \quad \text{si} \quad \dfrac{\overline{x}_1-\overline{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} > z_\alpha \end{cases}$$

σ_1 y σ_2 desconocidas. Muestras pequeñas

De nuevo hay que distinguir si las varianzas puedes ser consideradas iguales o no.

(a)$\sigma_1 = \sigma_2$

En este caso la regla óptima es

$$\begin{cases} \text{Se acepta } H_0 \text{ si} \\ \dfrac{\overline{x}_1-\overline{x_2}}{\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \leqslant t_{n_1+n_2-2;\alpha} \\ \text{Se rechaza } H_0 \text{ si} \\ \dfrac{\overline{x}_1-\overline{x_2}}{\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} > t_{n_1+n_2-2;\alpha} \end{cases}$$

(b)$\sigma_1 \neq \sigma_2$

El test óptimo es

$$\begin{cases} \text{Se acepta } H_0 \quad \text{si} \quad \dfrac{\overline{x}_1-\overline{x}_2}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}} \leqslant t_{f;\alpha} \\ \text{Se rechaza } H_0 \quad \text{si} \quad \dfrac{\overline{x}_1-\overline{x}_2}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}} > t_{f;\alpha} \end{cases}$$

$$\boxed{\begin{array}{l} H_0 : \mu_1 \geqslant \mu_2 \\ H_1 : \mu_1 < \mu_2 \end{array}}$$

Realmente este apartado es innecesario, ya que se corresponde exactamente con el anterior, intercambiando los papeles de las dos poblaciones, debido a la simetría de las distribuciones normal y t de Student.

Podríamos, por tanto, prescindir de él llamando siempre población 1 a la de mayor media muestral, siendo entonces el contraste de interés el de la hipótesis nula $H_0 : \mu_1 \leqslant \mu_2$ frente a la alternativa $H_1 : \mu_1 > \mu_2$.

El simétrico no es que no tenga interés, es que su resultado es obvio: Si $\overline{x}_1 > \overline{x}_2$, la hipótesis nula $H_0 : \mu_1 \geqslant \mu_2$ se aceptará siempre frente a $H_1 : \mu_1 < \mu_2$, con tal que sea $\alpha < 0,5$, ya que la región crítica para el contraste de esas hipótesis es la cola izquierda de la distribución, incluida en el semieje de los números negativos por ser $\alpha < 0,5$, mientras que al ser $\overline{x}_1 > \overline{x}_2$, el estadístico del contraste será siempre positivo.

σ_1 y σ_2 conocidas

En este caso el test óptimo es

$$\begin{cases} \text{Se acepta } H_0 \quad \text{si} \quad \dfrac{\overline{x}_1-\overline{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \geqslant z_{1-\alpha} \\ \text{Se rechaza } H_0 \quad \text{si} \quad \dfrac{\overline{x}_1-\overline{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} < z_{1-\alpha} \end{cases}$$

σ_1 y σ_2 desconocidas. Muestras pequeñas

(a) $\sigma_1 = \sigma_2$

Si las varianzas poblacionales pueden suponerse iguales y las muestras no tienen ambas, tamaños suficientemente grandes, el test óptimo es

$$\begin{cases} \text{Se acepta } H_0 \text{ si} \\ \dfrac{\overline{x}_1-\overline{x_2}}{\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \geqslant t_{n_1+n_2-2;1-\alpha} \\ \text{Se rechaza } H_0 \text{ si} \\ \dfrac{\overline{x}_1-\overline{x_2}}{\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} < t_{n_1+n_2-2;1-\alpha} \end{cases}$$

(b)$\sigma_1 \neq \sigma_2$

Si las varianzas poblacionales son distintas, el test óptimo es

$$\begin{cases} \text{Se acepta } H_0 \quad \text{si} \quad \dfrac{\overline{x}_1-\overline{x}_2}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}} \geqslant t_{f;1-\alpha} \\ \text{Se rechaza } H_0 \quad \text{si} \quad \dfrac{\overline{x}_1-\overline{x}_2}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}} < t_{f;1-\alpha} \end{cases}$$

7.7. Contraste de hipótesis relativas a la diferencia de medias de dos poblaciones independientes no necesariamente normales. Muestras grandes

La situación que se estudia en esta sección es la de dos muestas aleatorias independientes $X_1, ..., X_{n_1}$ e $Y_1, ..., Y_{n_2}$, de tamaños similares y suficientemente grandes, digamos $n_1 \approx n_2$ y $n_1 + n_2 > 30$.

Precisamente por esta razón no se requiere normalidad en las distribuciones modelo.

$$\boxed{\begin{array}{l} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 \neq \mu_2 \end{array}}$$

σ_1 y σ_2 conocidas

En este caso el test óptimo es

$$\begin{cases} \text{Se acepta } H_0 & \text{si } \frac{|\overline{x}_1-\overline{x}_2|}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \leqslant z_{\alpha/2} \\ \text{Se rechaza } H_0 & \text{si } \frac{|\overline{x}_1-\overline{x}_2|}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} > z_{\alpha/2} \end{cases}$$

σ_1 y σ_2 desconocidas

Si las varianzas poblacionales no se suponen conocidas, el test óptimo es

$$\begin{cases} \text{Se acepta } H_0 & \text{si } \frac{|\overline{x}_1-\overline{x}_2|}{\sqrt{\frac{S_1^2}{n_1}+\frac{\S_2^2}{n_2}}} \leqslant z_{\alpha/2} \\ \text{Se rechaza } H_0 & \text{si } \frac{|\overline{x}_1-\overline{x}_2|}{\sqrt{\frac{S_1^2}{n_1}+\frac{\S_2^2}{n_2}}} > z_{\alpha/2} \end{cases}$$

$$\boxed{\begin{array}{l} H_0 : \mu_1 \leqslant \mu_2 \\ H_1 : \mu_1 > \mu_2 \end{array}}$$

σ_1 y σ_2 conocidas

Si las varianzas de las poblaciones son conocidas, el test óptimo es

$$\begin{cases} \text{Se acepta } H_0 & \text{si } \frac{\overline{x}_1-\overline{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \leqslant z_{\alpha} \\ \text{Se rechaza } H_0 & \text{si } \frac{\overline{x}_1-\overline{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} > z_{\alpha} \end{cases}$$

σ_1 y σ_2 desconocidas

En el caso de que se desconozcan las varianzas de las poblaciones, el test óptimo es

$$\begin{cases} \text{Se acepta } H_0 & \text{si } \frac{\overline{x}_1-\overline{x}_2}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}} \leqslant z_{\alpha} \\ \text{Se rechaza } H_0 & \text{si } \frac{\overline{x}_1-\overline{x}_2}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}} > z_{\alpha} \end{cases}$$

$$\boxed{\begin{array}{l} H_0 : \mu_1 \geqslant \mu_2 \\ H_1 : \mu_1 < \mu_2 \end{array}}$$

σ_1 y σ_2 conocidas

Si las varianzas poblacionales son conocidas, el test óptimo es

$$\begin{cases} \text{Se acepta } H_0 & \text{si } \frac{\overline{x}_1-\overline{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \geqslant z_{1-\alpha} \\ \text{Se rechaza } H_0 & \text{si } \frac{\overline{x}_1-\overline{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} < z_{1-\alpha} \end{cases}$$

σ_1 y σ_2 desconocidas

Si son desconocidas, el test a utilizar es

$$\begin{cases} \text{Se acepta } H_0 & \text{si } \frac{\overline{x}_1-\overline{x}_2}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}} \geqslant z_{1-\alpha} \\ \text{Se rechaza } H_0 & \text{si } \frac{\overline{x}_1-\overline{x}_2}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}} < z_{1-\alpha} \end{cases}$$

Poblaciones binomiales

En este apartado las observaciones $X_1, ..., X_{n_1}$ e $Y_1, ..., Y_{n_2}$ se suponen procedentes de poblaciones independientes $B(1, p_1)$ y $B(1, p_2)$ respectivamente, siendo los tamaños muestrales similares y grandes, digamos $n_1 \approx n_2$ y $n_1 + n_2 > 100$.

$$\boxed{\begin{array}{l} H_0 : p_1 = p_2 \\ H_1 : p_1 \neq p_2 \end{array}}$$

En este caso, el test óptimo es

$$\begin{cases} \text{Se acepta } H_0 & \text{si } \frac{|\hat{p}_1-\hat{p}_2|}{\sqrt{\frac{\overline{p}(1-\overline{p})}{n_1}+\frac{\overline{p}(1-\overline{p})}{n_2}}} \leqslant z_{\alpha/2} \\ \text{Se rechaza } H_0 & \text{si } \frac{|\hat{p}_1-\hat{p}_2|}{\sqrt{\frac{\overline{p}(1-\overline{p})}{n_1}+\frac{\overline{p}(1-\overline{p})}{n_2}}} > z_{\alpha/2} \end{cases}$$

siendo $\hat{p}_1 = x_1/n_1$ y $\hat{p}_2 = x_2/n_2$, respectivamente, las proporciones de la primera y segunda muestras, y $\overline{p} = (x_1 + x_2)/(n_1 + n_2)$.

$$\boxed{\begin{array}{l} H_0 : p_1 \leqslant p_2 \\ H_1 : p_1 > p_2 \end{array}}$$

El test óptimo es

$$\begin{cases} \text{Se acepta } H_0 & \text{si } \frac{\hat{p}_1-\hat{p}_2}{\sqrt{\frac{\overline{p}(1-\overline{p})}{n_1}+\frac{\overline{p}(1-\overline{p})}{n_2}}} \leqslant z_{\alpha} \\ \text{Se rechaza } H_0 & \text{si } \frac{\hat{p}_1-\hat{p}_2}{\sqrt{\frac{\overline{p}(1-\overline{p})}{n_1}+\frac{\overline{p}(1-\overline{p})}{n_2}}} > z_{\alpha} \end{cases}$$

$$\boxed{\begin{array}{l} H_0 : p_1 \geqslant p_2 \\ H_1 : p_1 < p_2 \end{array}}$$

Como ahora ya no necesitamos comparar las varianzas, no tenemos restricciones acerca de cuál debería ser la población 1.

Para realizar contrastes de interés se aconseja tomar como población 1 la de mayor proporción muestral, y considerar el contraste del apartado anterior.

Si se considera este apartado, el test óptimo que debe utilizarse es

$$\begin{cases} \text{Se acepta } H_0 & \text{si } \frac{\hat{p}_1-\hat{p}_2}{\sqrt{\frac{\overline{p}(1-\overline{p})}{n_1}+\frac{\overline{p}(1-\overline{p})}{n_2}}} \geqslant z_{1-\alpha} \\ \text{Se rechaza } H_0 & \text{si } \frac{\hat{p}_1-\hat{p}_2}{\sqrt{\frac{\overline{p}(1-\overline{p})}{n_1}+\frac{\overline{p}(1-\overline{p})}{n_2}}} < z_{1-\alpha} \end{cases}$$

7.8. Contrastes de hipótesis para datos apareados

Como siempre que tratamos esta situación, supondremos que tenemos n parejas de datos $(X_1, Y_1), ..., (X_n, Y_n)$ en donde las variables X_i e Y_i no pueden calificarse de independientes.

El tratamiento que se hace aquí para este tipo de datos es similar al que se hizo en las Secciones 5.10 y 6.8, el cual consistía en definir la variable unidimensional diferencia, $D_i = X_i - Y_i$, y trasladar los posibles contrastes sobre los parámetros de X e Y a los correspondientes de la variables unidimensional D, utilizando los tests óptimos ya estudiados en las Secciones 7.2, 7.3 y 7.4.

8. Contrastes no paramétricos

8.1. Introducción

En este capítulo estudiaremos algunos tests, denominados *no paramétricos*, que por un lado no requieren especificar un modelo para la variable en estudio y, por otro, contrastan hipótesis que no se refieren a los valores de la media, ni de la varianza.

8.2. Pruebas χ^2

En esta sección estudiaremos tres tests que tienen la peculiaridad de estar definidos en base a recuentos o frecuencias de k posibles clases o grupos en los que se clasifican los datos y no en valores concretos de las variables en análisis. Se trata del contraste de *bondad del ajuste*, mediante el cual analizamos si puede admitirse una determinada distribución como modelo probabilístico de nuestros datos; del contraste de *homogeneidad de varias muestras*, con el que analizamos si puede admitirse la igualdad de las poblaciones de donde se extrajeron los datos y, por último, del contraste de *independencia de caracteres*, con el que contrastamos la hipótesis nula de independencia de dos variables.

Además, los tres tienen un estadístico de contraste con distribución (aproximada) χ^2.

Como dijimos, los tres contrastes que veremos en esta sección están basados en el denominado *estadístico λ de Pearson*, el cual mide, para cada clase E_i, las discrepancias entre las frecuencias (relativas) observadas n_i/n y las esperadas de ser cierta la hipótesis nula H_0 de que las clases E_i tienen probabilidades p_i, determinadas éstas por la hipótesis nula de la distribución modelo supuesta, o por la hipótesis nula de homogeneidad de las muestras, o por la de independencia de las dos variables en análisis. Este estadístico,

$$\lambda = \sum_{i=1}^{k} \frac{n}{p_i} \left(\frac{n_i}{n} - p_i \right)^2$$

mide las discrepancias normalizadas. Si toma un valor grande, entonces deberemos rechazar la hipótesis nula.

En la formalización del contraste necesitaremos determinar la distribución en el muestreo, bajo H_0, del estadístico del contraste λ. Si el tamaño muestral n es suficientemente grande, digamos $n > 30$, el estadístico λ de Pearson se distribuye, aproximadamente, según una χ^2 con $k-1$ grados de libertad. Ello permitirá determnar los puntos críticos de los contrastes que veremos a continuación.

El estadístico λ de Pearson, después de simplificar, puede calcularse como

$$\lambda = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i} = \sum_{i=1}^{k} \left(\frac{n_i^2}{np_i} \right) - n.$$

8.2.1. Pruebas χ^2 con R

Para ejecutar con R los tres tipos de contrastes χ^2 estudiados la función a utilizar será

```
chisq.test(x,correct=TRUE,p)
```

en donde incluiremos en el primer argumento `x` el vector de observaciones (frecuencias absolutas) en el test de bondad del ajuste, o la tabla de doble entrada en el caso de los otros dos tests.

El segundo argumento es opcional y permite utilizar la *corrección de Yates* aunque sólo en el caso de tablas de contingencia 2×2. Esta es la opción que se toma por defecto; para no utilizarla deberemos ejecutra `correct=F`.

El tercer argumento también es opcional y es utilizado, únicamente, en los tests de bondad del ajuste para indicar el vector de probabilidades teóricas que comparamos con las observadas, es decir, las que establecemos en la hipótesis nula si es que el modelo teórico es distinto del *uniforme* que asigna igual probabilidad a todas las clases consideradas ya quem en ese caso, podemos prescindir de esta opción al ser la que se toma por defecto.

8.2.2. Contraste de bondad del ajuste

Este contraste tiene por objeto averiguar si puede admitirse o no la hipótesis nula H_0 de seguir la variable aleatora en observación una determinada distribución modelo F_0, expresando dicha hipótesis nula de la forma $H_0 : F(x) = F_0(x)\ \forall x$, o más brevemente, $H_0 : X \rightsquigarrow F_0$.

Contraste de hipótesis

Sea $(\Omega, \mathcal{A}, P_0)$ un espacio probabilístico (con función de distribución asociada $F_0(x)$), en el que se considera una partición de Ω formada por k sucesos de $\mathcal{A}$, $E_1, E_2, ..., E_k$, tales que $p_i = P_0(E_i) > 0 \ \ i = 1, ..., k$ y $\sum_{i=1}^{k} p_i = 1$.

Si realizado un experimento aleatorio se obtuvieron las frecuencias absolutas $n_1, ..., n_k$ para las clases $E_1, ..., E_k$, y por

$$\lambda = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i} = \sum_{i=1}^{k} \left(\frac{n_i^2}{np_i} \right) - n$$

representamos el estadístico de Pearson, para contrastar a nivel α la hipótesis nula $H_0 : X \rightsquigarrow F_0$, frente a la alternativa de que los datos no se ajustan al modelo F_0, el test óptimo a utilizar es

$$\begin{cases} \text{Se acepta } \ H_0 & \text{si } \ \lambda < \chi^2_{k-1;\alpha} \\ \text{Se rechaza } \ H_0 & \text{si } \ \lambda \geqslant \chi^2_{k-1;\alpha} \end{cases}$$

teniendo perfecto sentido, tanto en este como en todos los contrastes que veremos, el cálculo e interpretación de su p-valor.

Observación 1

En la definición del contraste anterior suponíamos que la distribución modelo a contrastar estaba completamente especificada.

En muchas ocasiones el que admitamos como razonable un determinado tipo de distribución para los datos, no implica necesariamente conocer unos valores para los parámetros de esta distribución.

Observación 2

Para que la aproximación χ^2 a la distribución del estadístico λ sea aceptable, no sólo es necesario que el tamaño muestral sea grande, sino que además las frecuencias esperadas no sean demasiado pequeñas, digamos $n \cdot p_i \geqslant 5$.

Si esto no es así, deberemos agrupar clases E_i contiguas hasta que se tenga esta acotación, reduciendo en igual medida los grados de libertad de la distribución límite χ^2.

Y esto supuesto que no estimemos parámetros de la distribución modelos a partir de la muestra, ya que de ser así, aún deberíamos reducir más los grados de libertad, de acuerdo con la observación anterior.

Si las frecuencias esperadas son pequeñas y no queremos agrupar clases contiguas, podemos corregir el estadístico de Pearson mediante la denominada *corrección de Yates*, y utilizar como estadístico de contraste

$$\lambda_c = \sum_{i=1}^{k} \frac{(|n_i - np_i| - 0{,}5)^2}{np_i}$$

el cual seguirá teniendo una distribución χ^2_{k-1} grados de libertad (menos los parámetros que estimemos a partir de la muestra).

El utilizar la corrección de Yates generalmente conduce a tests más conservadores, es decir, tendentes a aceptar en muchos más casos la hipótesis nula de los que se hubiera aceptado de haber utilizado el estadístico λ de Pearson. Este hecho hace desaconsejar su uso, salvo en aquellas situaciones en las que una reducción del número de clases condujese a un χ^2 *sin grados de libertad*.

Observación 3

Digamos también que, siempre que sea posible, debemos elegir los sucesos E_k de forma que sea $p_i = 1/k$, consiguiendo de esta manera una mejor aproximación de la distribución de λ a la χ^2.

8.2.3. Contraste de homogeneidad de varias muestras

Este contraste tiene por objeto averiguar si existen o no diferencias significativas entre r poblaciones, de las que se han extraído sendas muestras aleatorias simples.

Es decir, es un contraste semejante, en cuanto a propósitos, a los contrastes de análisis de la varianza, aunque con la diferencia de que ahora los datos son frecuencias o recuentos del número de individuos pertenecientes a cada una de las clases en las que se han dividido las poblaciones, y no los valores de una variable observable.

En general, tendremos s clases en las que se han dividido las r poblaciones, estando clasificadas las r muestras aleatorias extraídas (una de cada población) en una tabla de frecuencias absolutas de la forma

	Clases				
Muestras	C_1	C_2	...	C_s	Totales
M_1	n_{11}	n_{12}	...	n_{1s}	n_1
M_2	n_{21}	n_{22}	...	n_{2s}	n_2
...	...	...	...	...	...
M_r	n_{r1}	n_{r2}	...	n_{rs}	n_r
Totales	m_1	m_2	...	m_s	n

en donde n_{ij} es el número de individuos de la muestra i-ésima que pertenecen a la clase j-ésima, $n_i = \sum_{j=1}^{s} n_{ij}$ el tamaño de la muestra i-ésima, $m_j = \sum_{i=1}^{r} n_{ij}$ la frecuencia absoluta marginal de la clase C_j y $n = \sum_{i=1}^{r} n_i = \sum_{j=1}^{s} m_j = \sum_{i=1}^{r} \sum_{j=1}^{s} n_{ij}$ el tamaño muestral.

El propósito de este test es contrastar la hipótesis nula H_0: *las r poblaciones son homogéneas*, frente a la alternativa de no serlo.

Denominando p_j a la probabilidad teórica de la clase C_j, $j = 1, ..., s$, podemos aplicar a la muestra M_i el estadístico λ de Pearson, obteniendo que

$$\sum_{j=1}^{s} \frac{(n_{ij} - n_i p_j)^2}{n_i p_j} \approx \chi^2_{s-1}$$

Si es cierta la hipótesis nula de igualdad de las r poblaciones, la probabilidad de la clase C_j seguirá siendo p_j en cada una de las r muestras y, como además éstas son independientes, la suma de los estadísticos de Pearson utilizados en cada una de las muestras tendrá también una distribución χ^2 (aproximadamente), de grados de libertad la suma de los grados de libertad de cada una de las χ^2. Por tanto, será

$$\sum_{i=1}^{r} \sum j = 1^{s} \frac{(n_{ij} - n_i p_j)^2}{n_i p_j} \approx \chi^2_{r(s-1)}$$

De forma casi generalizada, las probabilidades p_j serán desconocidas, por lo que será necesario estimarlas mediante los estimadores de máxima verosimilitud, $\hat{p}_j = m_j/n$, teniendo que restar $s-1$ grados de libertad a la $\chi^2_{r(s-1)}$ (una e las p_j no hay que estimarla puesto que la suma de todas ellas debe ser 1) quedando en definitiva que

$$\sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(n_{ij} - n_i m_j/n)^2}{n_i m_j/n} \approx \chi^2_{(r-1)(s-1)}$$

al ser $r(s-1) - (s-1) = (s-1)(r-1)$.

Contraste de hipótesis

Supongamos n datos como los de la tabla anterior. Para contrastar, a nivel α, la hipótesis nula H_0: *son homogéneas las r poblaciones*, de las que se extraen las muestras $M_1, ..., M_r$, frente a la alternativa de *no homogeneidad de las r poblaciones*, y si es

$$\lambda = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(n_{ij} - n_i m_j/n)^2}{n_i m_j/n}$$

entonces el contraste óptimo a utilizar consiste en

$$\begin{cases} \text{Se acepta } H_0 & \text{si } \lambda < \chi^2_{(r-1)(s-1);\alpha} \\ \text{Se rechaza } H_0 & \text{si } \lambda \geqslant \chi^2_{(r-1)(s-1);\alpha} \end{cases}$$

Observación 4

Con objeto de obtener una rápida convergencia hacia la χ^2, las frecuencias esperadas deberán ser suficientemente grandes, digamos $n_i m_j/n \geqslant 5$.

Si esto no se cumple, deberemos agrupar clases contiguas, reduciendo adecuadamente los grados de libertad, o de forma alternativa utilizar el estadístico corregido

$$\lambda_c = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{|(n_{ij} - n_i m_j/n| - 0{,}5)^2}{n_i m_j/n}$$

manteniendo, en este caso, los mismos grados de libertad.

Como ocurría antes, la corrección de Yates conducirá, en general, a tests más conservadores.

8.2.4. Contraste de independencia de caracteres

El último contraste de la χ^2 que vamos a estudiar analiza la posbile independencia entre dos caracteres observados en los individuos de una población.

En general tendremos dos caracteres, A con a modalidades y B con b modalidades, estando los n individuos de la muestra clasificados en una *tabla de doble entrada* o de *contingencia* de la forma

B / A	1	2	...	b	
1	n_{11}	n_{12}	...	n_{1b}	$n_{1.}$
2	n_{21}	n_{22}	...	n_{2b}	$n_{2.}$
...	...	...	...	...	...
a	n_{a1}	n_{a2}	...	n_{ab}	$n_{a.}$
	$n_{,1}$	$n_{,2}$	...	$n_{.b}$	n

en donde n_{ij} es el número de individuos de la muestra de tamaño n que presentan a la vez la modalidad i-ésima del carácter A y la j-ésima del carácter B.

Las hipótesis a contrastar son, H_0: *los caracteres A y B son independientes*, frente a la alternativa, H_1: *A y B no son independientes*.

Llamando p_i a la probabilidad (marginal) de obtener un individuo de la población que presente la modalidad i-ésima del carácter A, y q_j la probabilidad (marginal) de obtener un individuo de la población que presente la modalidad j-ésima del carácter B, si la hipótesis nula fuese correcta, la probabilidad p_{ij} de obtener un individuo de la población que presente a la vez la modalidad i-ésima del carácter A y j-ésima del carácter B, sería $p_i \cdot q_j$, con lo que, en la muestra de tamaño n, cabría esperar que $n \cdot p_i \cdot q_j$ presenten a la vez ambas modalidades.

La comparación de las frecuencias observadas, n_{ij}, con las esperadas, $n \cdot p_i \cdot q_j$, para cada una de las $k = a \cdot b$ clases se hará a través del estadístico λ de Pearson.

En efecto, el estadístico

$$\lambda = \sum_{i=1}^{a}\sum_{j=1}^{b} \frac{(n_{ij} - n p_i q_j)^2}{n_i p_i q_j}$$

seguirá, aproximadamente, una distribución χ^2 con $ab-1$ grados de libertad.

Las probabilidades p_i y q_j serán habitualmente desconocidas, por lo que deberemos estimarlas utilizando sus estimadores máximo-verosímiles, las frecuencias relativas, $\hat{p}_i = n_{i.}/n$ y $\hat{q}_j = n_{.j}/n$, quedando el estadístio de Pearson a utilizar habitualmente de la forma

$$\lambda = \sum_{i=1}^{a}\sum_{j=1}^{b} \frac{(n_{ij} - n_{i.} n_{.j}/n)^2}{n_{i.} n_{.j}/n}$$

el cual seguirá, aproximadamente, una distribución χ^2 con $ab-1-(a-1)-(b-1) = (a-1)(b-1)$ grados de libertad.

Contraste de hipótesis

Así pues, para contrastar, a nivel α, la hipótesis nula H_0: *los caracteres A y B no son independientes*, el contraste a utilizar es

$$\begin{cases} \text{Se acepta } H_0 & \text{si } \lambda < \chi^2_{(a-1)(b-1);\alpha} \\ \text{Se rechaza } H_0 & \text{si } \lambda \geqslant \chi^2_{(a-1)(b-1);\alpha} \end{cases}$$

siendo

$$\lambda = \sum_{i=1}^{a}\sum_{j=1}^{b} \frac{(n_{ij} - n_{i.} n_{.j}/n)^2}{n_{i.} n_{.j}/n}$$

Observación 5

De nuevo las frecuencias esperadas no deben ser muy pequeñas, digamos $n_{i.} n_{.j}/n \geqslant 5$, debiendo agruparse las clases contiguas en caso contrario, o utlizar la corrección de Yates.

Por último queremos hacer notar que, matemáticamente, el contraste de homogeneidad de varias muestras y el de independencia de caracteres has resultado técnicamente idénticos.

Es decir, que al expresarse los datos de ambos en una tabla de contingencia el estadístico de Pearson *se calcula* de la misma manera.

Existe, no obstante, una diferencia fundamental entre ambos: en el contraste de homogeneidad, los totales marginales n_i son los que fija el investigador, el cual decide, por tanto, cuántos individuos deben elegirse de cada población. Por otro lado, en el contraste de independencia, lo que fija el experimentador es n, quedando los totales marginales $n_{i.}$ fuera de control del investigador.

8.3. Test relativos a una muestra y datos apareados

La hipótesis nula que aquí contrastamos hará referencia a la mediana de la población de donde se extrajeron los datos o, si son datos apareados, a la mediana de la diferencia de las variables (que puede ser distinta a la diferencia de las medianas). Es decir, la hipótesis nula será $H_0 : M = M_0$ que se contrastará frente a la hipótesis alternativa $H_1 : M \neq M_0$.

8.3.1. El contraste de los signos

El *test de los signos* no necesita suponer una distribución modelo específica para la variable aleatoria en estudio; sólo se exige que la distribución modelo sea de tipo continuo, al menos en un entorno de la mediana poblacional

M. Pero además, este test es tan genérico que no requiere ni de los valores de las observaciones, sólo de sus *rangos*, es decir, para ser aplicado sólo necesita de las ordenaciones de la observaciones, y no los valores númericos de éstas.

Para su definición deberemos distinguir los casos en los que las hipótesis a contrastar sean bilaterales o sean unilaterales.

$$\boxed{\begin{array}{c} H_0 : M = M_0 \\ H_1 : M \neq M_0 \end{array}}$$

Dada una muestra aleatoria simple de la población, $X_1, ..., X_n$, si la hipótesis nula $H_0 : M = M_0$ es cierta, aproximadamente la mitad de las observaciones serán menores que M_0 y la otra mitas mayores, ya que la mediana poblacional se define como aquel valor M tal que

$$P\{X < M\} = P\{X > M\} = 0{,}5.$$

Por tanto, si consideramos como estadístico del contraste el *número T de observaciones mayores que* M_0, o equivalentemente, el *número de signos positivos* de entre todas las diferencias $X_i - M_0$, $i = 1, ..., n$, el observar un valor de T muy grande o muy pequeño tenderá a desacreditar la hipótesis nula en favor de la alternativa.

A pesar de ser éste un contraste no paramétrico, con objeto de determinar los puntos críticos del contraste necesitamos conocer la distribución en el muestreo del estadístico del test, T, con objeto de poder precisar lo que se entiende por *muy grande* o *muy pequeño*.

Afortunadamente, la distribución de T bajo la hipótesis nula no depende del modelo de X. Por esta razón, a este tipo de contrastes se les suele denominar de *distribución libre*, además de no paramétricos. Si llamamos *éxito* al suceso en el que $X_i > M_0$ y *fracaso* al suceso $X_i < M_0$, T será el número de éxitos en n pruebas de Bernoulli, con lo que su distribución, si es cierta la hipótesis nula, será binomial $B(n, 0{,}5)$. (Sección 4.4.1)

Contraste de hipótesis

Si el valor de T es muy grande o muy pequeño, rechazaremos la hipótesis nula $H_0 : M = M_0$, aceptando en consecuencia la alternativa $H_1 : M \neq M_0$. En concreto, fijado un nivel de significación α

$$\begin{cases} \text{Se acepta } H_0 & \text{si } t_{1-\alpha/2} < T < t_{\alpha/2} \\ \text{Se rechaza } H_0 & \text{si } T \leqslant t_{1-\alpha/2} \text{ ó } T \geqslant t_{\alpha/2} \end{cases}$$

en donde t_β es el valor de una binomial $B(n, 0{,}5)$ que deja a la derecha una área de probabilidad β, es decir, tal que $P\{W \geqslant t_\beta\} = \beta$ con $W \rightsquigarrow B(n, 0{,}5)$.

Como por las propiedades de la distribución binomial, es $t_{1-\alpha/2} = n - t_{\alpha/2}$, podemos expresar todo el test en función del punto crítico $t_{\alpha/2}$ en la forma

$$\begin{cases} \text{Se acepta } H_0 & \text{si } n - t_{\alpha/2} < T < t_{\alpha/2} \\ \text{Se rechaza } H_0 & \text{si } T \leqslant n - t_{\alpha/2} \text{ ó } T \geqslant t_{\alpha/2} \end{cases}$$

en donde $t_{\alpha/2}$ es tal que

$$\sum_{t=t_{\alpha/2}}^{n} \binom{n}{t} (0{,}5)^n = \frac{\alpha}{2}.$$

Como la distribución binomial es de tipo discreto, es posible que no exista ningún valor $t_{\alpha/2}$ que cumpla la relación anterior. Por tanto, el $t_{\alpha/2}$ que se toma es el menor número entero tal que

$$\sum_{t=t_{\alpha/2}}^{n} \binom{n}{t} (0{,}5)^n \leqslant \frac{\alpha}{2}.$$

$$\boxed{\begin{array}{c} H_0 : M \leqslant M_0 \\ H_1 M > M_0 \end{array}}$$

En este caso, si el número de *signos positivos* es grande, rechazaremos H_0 en favor de H_1. En concreto, fijado un nivel de significación α

$$\begin{cases} \text{Se acepta } H_0 & \text{si } T < t_\alpha \\ \text{Se rechaza } H_0 & \text{si } T \geqslant t_\alpha \end{cases}$$

en donde t_α es el valor de una binomial $B(n, 0{,}5)$ que deja a la derecha una área de probabilidad α, es decir, tal que $P\{W \geqslant t_\alpha\} = \alpha$ con $W \rightsquigarrow B(n, 0{,}5)$.

De nuevo puede ocurrir que α no sea accesible, por lo que t_α se toma como el menor número entero tal que

$$\sum_{t=t_\alpha}^{n} \binom{n}{t} (0{,}5)^n \leqslant \alpha.$$

$$\boxed{\begin{array}{c} H_0 : M \geqslant M_0 \\ H_1 : M < M_0 \end{array}}$$

Ahora, un número pequeño de *signos positivos*, es decir un valor de T pequeño, desacreditará la hipótesis nula en favor de la alternativa. Por tanto, fijado un nivel de significación α

$$\begin{cases} \text{Se acepta } H_0 & \text{si } T > n - t_\alpha \\ \text{Se rechaza } H_0 & \text{si } T \leqslant n - t_\alpha \end{cases}$$

siendo de nuevo t_α el menor número entero tal que

$$\sum_{t=t_\alpha} Pn \binom{n}{t} (0{,}5)^n \leqslant \alpha.$$

Muestras grandes

Cuando n es grande, digamos $n \geqslant 12$, podremos determinar los puntos críticos del test de los signos por la distribución normal, habiéndose añadido un factor de corrección por estar aproximando una distribución discreta por una de tipo continuo como es la normal.

En concreto, si fijado un nivel de significación α, es como siempre z_α el valor de la abscisa de una normal $N(0, 1)$ que deja a la derecha una área de probabilidad α, el punto crítico t_α en los contrastes anteriores es

$$t_\alpha = 0{,}5(z_\alpha \sqrt{n} + n + 1).$$

Obviamente, para el contraste bilateral deberá cambiarse α por $\alpha/2$ en la fórmula anterior.

El problema de los empates

Aunque teóricamente no deberían observarse valores iguales a la mediana a contrastar M_0, al haberse supuesto la distribución continua en la vecindad de la mediana, de hecho se producen.

Existen varias alternativas para solucionar este problema. La primera y más razonable, es medir con mayor precisión cerca de M_0 de forma que podamos discriminar si el dato es menor o mayor que M_0, para poder decidir el signo aportado por el valor muestral.

Si los datos ya vienen dados, lo más aconsejable es ignorar las diferencias cero, disminuyendo, en consecuencia, el tamaño de la muestra.

Datos apareados

Los resultados vistos hasta ahora pueden aplicarse al caso de datos apareados $(X_1, Y_1), ..., (X_n, Y_n)$ definiendo la variable diferencia $D = X - Y$.

Los contrastes que se hagan harán referencia a la mediana de la variable diferencia pero no necesariamente a la diferencia de las medianas. Ambas cantidades coincidirán cuando las poblaciones X e Y sean simétricas con el mismo centro de simetría y además la población diferencia D también sea simétrica.

8.3.2. El contraste de los rangos signados de Wilcoxon

El contraste de los signos anterior tiene la gran ventaja de su sencillez, pero el inconveniente de utilizar solamente el signo suministrado por cada observación, es decir, si es $X_i - M_0 > 0$, o si es $X_i - M_0 < 0$, sin considerar la magnitud de dicha diferencia.

El *contraste de rangos signados de Wilcoxon* recoge esta información, aunque requiere a cambio, que la distribución modelo sea continua y simétrica.

Aunque no haremos referencia a ello, este test se puede utilizar en el caso de datos apareados análogamente a como ocurría con el tests de los signos.

$$\boxed{\begin{array}{l} H_0 : M = M_0 \\ H_1 : M \neq M_0 \end{array}}$$

Sea $X_1, ..., X_n$ una muestra aleatoria de la variable en observación X y $D_i = X_i - M_0$ las diferencias de la muestra con la mediana a contrastar M_0.

Si ordenamos sus valores absolutos $|D_1|, ..., |D_n|$ asignando a cada uno su rango $r(|D_i|)$, es decir, al menor $|D_i|$ el valor 1 y así hasta el último al que asignamos el valor n, el test de Wilcoxon utiliza como estadístico de contraste, T^+, la *suma de los rangos de las diferencias positivas*, es decir, los *rangos signados*. Analíticamente,

$$T^+ = \sum_{i=1}^{n} z_i r(|D_i|)$$

con

$$z_i = \begin{cases} 1 & \text{si } D_i > 0 \\ 0 & \text{si } D_i < 0. \end{cases}$$

Con objeto de determinar los puntos críticos del test deberemos conocer la distribución en el muestreo de T^+.

Los valores extremos de T^+ son 0 (todas las diferencias negativas) y $n(n+1)/2$ (todas las diferencias positivas).

La determinación de la funció nde masa de T^+ resulta complicada por lo que en la *Tabla 13* de ADD aparecen los puntos críticos para tamaños muestrales pequeños.

Contraste de hipótesis

Valores muy grandes o muy pequeños de T^+ desacreditarán la hipótesis nula $H_0 : M = M_0$ en favor de la alternativa $H_1 : M \neq M_0$, con lo que fijado un nivel de significación α,

$$\begin{cases} \text{Se acepta } H_0 & \text{si } \frac{n(n+1)}{2} - t_{\alpha/2} < T^+ < t_{\alpha/2} \\ \text{Se rechaza } H_0 & \text{si } T^+ \leqslant \frac{n(n+1)}{2} - t_{\alpha/2} \text{ ó } T^+ \geqslant t_{\alpha/2} \end{cases}$$

en donde $t_{\alpha/2}$ es tal que $P\{T^+ \geqslant_{\alpha/2}\} \leqslant \alpha/2$.

Como la distribución binomial es discreta, es posible que no exista ningún valor $t_{\alpha/2}$ que cumpla la relación anterior. Por tanto, el $t_{\alpha/2}$ que se toma es el menor número entero tal que

$$P\{T^+ \geqslant t_{\alpha/2}\} \leqslant \frac{\alpha}{2}.$$

Contraste de los rangos signados de Wilcoxon con R

El test de los rangos signados de Wilcoxon se ejecuta con la función `wilcox.test()`, que será la misma que utilizaremos para el contraste de Wilcoxon-Mann-Whitney más adelante,

```
wilcox.test(x,alternative="two.sided",mu=0,
exact=T,correct=T)
```

en donde incluiremos en el primer argumento `x` el vector de observaciones. Con el argumento `alternative` podemos elegir el tipo de test que vamos a ejecutar, bilateral (que es el que se utiliza por defecto) , `less` o `greater` si la hipótesis alternativa que queremos contrastar es, respectivamente, menor o mayor. Con `mu` podemos señalar el valor de la hipótesis a contrastar, eligiendo la función el valor 0 por defecto. Con `exact` indicamos si queremos que R calcule el valor exacto de la distribución del estadístico T^+ de Wilcoxon (opción tomada por defecto) o que calcule el valor aproximado del p-valor para muestras grandes cuya expresión daremos más abajo, ejecutando `exact=F`. Finalmente, con `correct` indicamos si queremos utilizar la corrección de continuidad.

$$\boxed{\begin{array}{l} H_0 : M \leqslant M_0 \\ H_1 : M > M_0 \end{array}}$$

En este caso, fijado un nivel de significación α

$$\begin{cases} \text{Se acepta } H_0 & \text{si } T^+ < t_\alpha \\ \text{Se rechaza } H_0 & \text{si } T^+ \geqslant t_\alpha \end{cases}$$

en donde de nuevo t_α es el menor número entero tal que

$$P\{T^+ \geqslant t_\alpha\} \leqslant \alpha.$$

$$\boxed{\begin{array}{l} H_0 : M \geqslant M_0 \\ H_1 : M < M_0 \end{array}}$$

Para este último contraste unilateral, fijado un nivel de significación α

$$\begin{cases} \text{Se acepta } H_0 & \text{si } T^+ > \frac{n(n+1)}{2} - t_\alpha \\ \text{Se rechaza } H_0 & \text{si } T^+ \leqslant \frac{n(n+1)}{2} - t_\alpha \end{cases}$$

siendo de nuevo t_α el menor número entero tal que

$$P\{T^+ \geqslant t_\alpha\} \leqslant \alpha.$$

Muestras grandes

La distribución del estadístico T^+ se puede aproximar por una $N(0,1)$ aplicando el teorema central del límite, cuando el tamaño muestral es suficientemente grande, digamos $n > 15$. En ese caso es

$$\frac{4T^+ - n(n+1)}{\sqrt{2n(n+1)(2n+1)/3}} \approx N(0,1)$$

con lo que, despejando, el punto crítico (para el contraste unilateral) quedaría

$$t_\alpha = \frac{n(n+1)}{4} + \frac{1}{4} z_\alpha \sqrt{\frac{2n(n+1)(2n+1)}{3}}.$$

Si la muestra no es muy grande podría utilizarse una corrección de continuidad, restando 0,5 al valor absoluto del numerador de la distribución T^+.

El problema de los empates y el de las diferencias iguales

Aunque teóricamente no deberían observarse ni valores iguales a la mediana a contrastar M_0, ni diferencias D_i iguales, cuestión esta última que produce problemas al asignar los rangos, de hecho se obtendrán.

Respecto a los empates, la solución que se propone es la misma que en el test de los signos ignorarlos disminuyendo el tamaño de la muestra.

Por otro lado, si dos o más diferencias absolutas son iguales, $|D_i| = |D_j|$ para al menos un $i \neq j$, se propone tomar como rango común a todas las diferencias iguales, la media aritmética de los rangos que tendrían si fueran distinguibles, aunque conservando cada D_i su signo.

8.4. Tests relativos a dos muestras independientes

En estaa sección estudiaremos dos contrastes no paramétricos para contrastar la hipótesis nula de igualdad de dos poblaciones independientes, expresada ésta mediante la igualdad de sus medianas poblacionales, $H_0 : M_X = M_Y$.

8.4.1. El contraste de Wilcoxon-Mann-Whitney

Este test requiere que las distribuciones poblacionales F y G sean continuas.

$$\boxed{\begin{array}{l} H_0 : M_X = M_Y \\ H_1 : M_X \neq M_Y \end{array}}$$

Sea $X_1, ..., X_m$ una muestra aleatoria simple de tamaño m de la primera población e $Y_1, ..., Y_n$ una de tamaño n de la segunda.

La idea del contraste consiste en medir las magnitudes de los valores Y_i en relación con los X_i, es decir, las posiciones de los Y_i en la muestra conjunta de las X_i e Y_i. Si observamos que la mayoría de los Y_i están hacia el principio o hacia el final de la muestra conjunta, deberemos rechazar la hipótesis nula de igualdad de ambas poblaciones.

En concreto, si llamamos

$$D_{ij} = \begin{cases} 1 & Y_j < X_i \\ 0 & Y_j \geqslant X_i \end{cases}$$

$\forall i = 1, ..., m$ y $j = 1, ..., n$, el estadístico U en el que está basado el contraste es

$$U = \sum_{i=1}^{m} \sum_{j=1}^{n} D_{ij}$$

es decir, el número de Y_j que preceden estrictamente a cada X_i.

La distribución exacta en el muestreo de U es complicada, apareciendo en la *Tabla 14* del ADD los puntos críticos en el caso de tamaños muestrales pequeños. Además veremos que cuando m y n sean mayores que 5, la aproximación normal es adecuada. Apuntemos, no obstante, que los valores de U están entre 0 y $m \cdot n$, así como que la distribución de U es simétrica respecto a su media $m \cdot n/2$.

Contraste de hipótesis

Valores muy grandes o muy pequeños de U desacreditarán la hipótesis nula de igualdad de ambas poblaciones. Así pues, fijado un nivel de significación α,

$$\begin{cases} \text{Se acepta } H_0 & \text{si } m \cdot n - u_{m,n;\alpha/2} < U < u_{m,n;\alpha/2} \\ \text{Se rechaza } H_0 & \text{si } U \leqslant m \cdot n - u_{m,n;\alpha/2} \text{ ó } U \geqslant u_{m,n;\alpha/2} \end{cases}$$

en donde $u_{m,n;\alpha/2}$ es el menor número entero tal que

$$P\{U \geqslant u_{m,n;\alpha/2}\} \leqslant \frac{\alpha}{2}.$$

$$\boxed{\begin{array}{l} H_0 : M_X \leqslant M_Y \\ H_1 : M_X > M_Y \end{array}}$$

En este caso, la existencia de muchas Y_i que preceden a X_i hará que U tome valores altos, situación que parece confirmar la hipótesis alternativa.

Por tanto, fijado un nivel de significación α

$$\begin{cases} \text{Se acepta } H_0 & \text{si } U < u_{m,n;\alpha} \\ \text{Se rechaza } H_0 & \text{si } U \geqslant u_{m,n;\alpha} \end{cases}$$

en donde de nuevo $u_{m,n;\alpha}$ es el menos número entero tal que

$$P\{U \geqslant u_{m,n;\alpha}\} \leqslant \alpha.$$

$$\boxed{\begin{array}{l} H_0 : M_X \geqslant M_Y \\ H_1 : M_X < M_Y \end{array}}$$

En este último contraste unilateral rechazaremos H_0 cuando U tome valores pequeños. Es decir, fijado un nivel de significación α

$$\begin{cases} \text{Se acepta } H_0 & \text{si } U > m \cdot n - u_{m,n;\alpha} \\ \text{Se rechaza } H_0 & \text{si } U \leqslant m \cdot n - u_{m,n;\alpha} \end{cases}$$

siendo de nuevo $u_{m,n;\alpha}$ el menor número entero tal que

$$P\{U \geqslant u_{m,n;\alpha}\} \leqslant \alpha.$$

Muestras grandes

La distribución de U se aproximará por una normal en cuanto el tamaño de las muestras sea parecido y lo suficientemente grande ($m \approx n$, y además, $m, n > 5$), pudiendo usarse, como de costumbre, una corrección de continuidad (restar 0,5 al valor absoluto del numerador), si los tamaños muestrales son cercanos a 5.

En concreto se tiene que

$$\frac{U - \frac{mn}{2}}{\sqrt{mn(m+n+1)/12}} \approx N(0,1)$$

con lo que el punto crítico $u_{m,n;\alpha}$ será

$$u_{m,n;\alpha} = \frac{mn}{2} + z_\alpha \sqrt{\frac{mn(n+m+1)}{12}}.$$

8.4.2. El contraste de la Mediana

El *contraste de la Mediana* es un contraste en el que, de nuevo, las hipótesis hacen referencia a las medianas poblacionales, M_X y M_Y.

$$\boxed{\begin{array}{l} H_0 : M_X = M_Y \\ H_1 : M_X \neq M_Y \end{array}}$$

Sean $X_1, ..., X_m$ e $Y_1, ..., Y_n$ muestras aleatorias de las dos poblaciones en consideración. Si la hipótesis nula es cierta, entonces ambas muestras procederán de poblaciones con la misma mediana, por lo que, en la muestra combinada, de tamaño $m+n$ y de media muestral M_s cabría esperar que la mitad de las observaciones fueran menores que M_s y la otra mitad mayores.

Por lo tanto, si consideramos como estadístico del test, A = *número de observaciones* x_i *menores o iguales que* M_s, valores muy grandes o muy pequeños soyos desacreditarán la hipótesis nula.

Desgraciadamente la distribución de A resulta complicada y como en cuanto los tamaños muestrales sean moderadamente grandes, digamos $m > 10$ y $n > 10$, se puede utilizar una distribución χ^2 en la determinación de los puntos críticos, omitiremos el caso de muestras pequeñas.

Así pues, supongamos que los tamaños muestrales son suficientemente grandes. En ese caso, si expresamos los datos como en la siguiente tabla

	Valores menores o iguales que M_s	Valores mayores que M_s	Total muestral
$X_1, ..., X_m$	a	$m-a$	m
$Y_1, ..., Y_n$	b	$n-b$	n
	$a+b$	$m+n-a-b$	$m+n$

podemos aplicar la técnica de la χ^2 estudiada en la Sección 8.2.3 y utilizar como estadístico del contraste el λ de Pearson, que para el caso particular de la tabla anterior queda, después de simplificar,

$$\lambda = \frac{(m+n)(an-bm)^2}{mn(a+b)(m+n-a-b)}.$$

Contraste de hipótesis

Por tanto, fijado un nivel de significación α ,

$$\begin{cases} \text{Se acepta } H_0 & \text{si } \lambda < \chi^2_{1;\alpha} \\ \text{Se rechaza } H_0 & \text{si } \lambda \geqslant \chi^2_{1;\alpha} \end{cases}$$

en donde por $\chi^2_{1;\alpha}$ representamos, como de costumbre, el valor de una abscisa de una χ^2_1 que deja a la derecha un área de probabilidad α.

9. Análisis de la Varianza

9.1. Introducción

En este capítulo expondremos las técnicas denominadas *Análisis de varianza* las cuales permiten comparar las medias de más de dos poblaciones. Las suposiciones que estas técnicas requieren son, básicamente, la normalidad de las poblaciones a comparar, el que tengan la misma varianza (suposición de *homocedasticidad*) y el que sean independientes.

La técnica del análisis de la varianza se basa en dividir la variabilidad total existente en un conjunto de datos, en diversas *fuentes de variación*, analizando, mediante un contraste de hipótesis, si la aportación relativa de cada una de estas fuentes de variación a la variación total es significativa o no.

9.2. Análisis de la varianza para un factor: Diseño Completamente Aleatorizado

En este capítulo analizaremos situaciones en las que hay *un factor* en estudio, el cual actúa a r *niveles*.

En estos casos en los que sólo se considera un factor, a los niveles se les suele llamar *tratamientos*.

Si denotamos por $\mu_1, ..., \mu_r$ los efectos medios de los tratamientos, el interés del investigador se centra en contrastar la hipótesis nula de igualdad de dichos efectos medios, $H_0 : \mu_1 = \mu_2 = ... = \mu_r$ frente a la alternativa de no ser iguales todos estos efectos medios, H_1:*no todos son iguales*, en base a observar los valores de cada uno de los tratamientos en individuos elegidos al azar.

Concretamente, si designamos por X_{ij} el valor o respuesta observado en el individuo j-ésimo, $j = 1, ..., n_i$ sometido al tratamiento i-ésimo, $i = 1, ..., r$, los $n = \sum_{i=1}^{r} n_i$ datos correspondientes a las r muestras aleatorias, de tamaños $n_1, ..., n_r$ pueden representarse en la forma

Tratamiento	Observaciones				Totales	Medias muestrales
1	x_{11}	x_{12}	...	x_{1n_1}	T_1	$\overline{x}_1$
2	x_{21}	x_{22}	...	x_{2n_2}	T_2	$\overline{x}_2$
...	...	...	...	...	...	...
r	x_{r1}	x_{r2}	...	x_{rn_r}	T_r	$\overline{x}_r$
					T	

en donde es $T_i = \sum_{j=1}^{n_i} x_{ij}$, $T = \sum_{i=1}^{r} T_i$ y $\overline{x}_i = T_i/n_i$.

Para contrastar las hipótesis $H_0 : \mu_1 = ... = \mu_r$ frente $H_| :$*alguna es distinta* serán necesarias las siguientes suposiciones:

a) La i-ésima población o tratamiento X_i se distribuye según una $N(\mu_i, \sigma)$ $i = 1, ..., r$.

b) Las r poblaciones o tratamientos son independientes entre sí.

c) La muestra de tamaño (prefijado) n_i de la población i-ésima es aleatoria simple.

Obsérvese que la suposición a) lleva implícita no sólo la normalidad sino también la *homocedasticidad*, es decir, el que todos los tratamientos tengan igual varianza.

Modelo del diseño

De forma trivial puede escribirse que

$$x_{ij} = \mu + (\mu_i - \mu) + (x_{ij} - \mu_i)$$

y llamando $\alpha_i = \mu_i - \mu$ y $e_{ij} = x_{ij} - \mu_i$, la expresión anterior puede escribirse de la forma

$$x_{ij} = \mu + \alpha_i + e_{ij}$$

Las expresiones anteriores son válidas para cualquier constante μ, pero aquí tomaremos

$$\mu = \frac{\sum_{i=1}^{r} n_i \mu_i}{n}$$

es decir, una media ponderaa de los efectos medios de los tratamientos.

Si es cierto H_0, el efecto medio común será precisamente μ. Por esta razón, α_i puede interpretarse como el *efecto del tratamiento i-ésimo*. Cuando más se distancie μ_i del efecto medio común μ, mayor sera α_i.

Con esta notación, las hipótesis a contrastar se expresan de la forma $H_0 : \alpha_i = 0 \ \ \forall i = 1, ..., r$ frente a $H_1 :$*no todas las* $\alpha_i = 0$.

Como x_{ij} es un valor muestral obtenido por la variable X_i, la cual tiene media μ_i, la diferencia e_{ij} puede interpretarse como el *error*, debido al azar, que se produce en todo el muestreo, el cual hace que no todas las observaciones muestrales sean iguales a su media.

Por tanto, la ecuación de x_{ij} anterior puede interpretarse diciendo que cada dato observado x_{ij} es el resultado de un efecto común, μ, más el efecto propio del tratamiento i-ésimo de donde procede el dato, α_i, más un término de error, e_{ij}, fruto del muestreo aleatorio efectuado dentro de la población i-ésima.

Fuentes de variación

Si llamamos $\overline{\overline{x}}$ a la media de la muestra global,

$$\overline{\overline{x}} = \frac{1}{n} \sum_{i=1}^{r} \sum_{j=1}^{n_i} x_{ij} = \frac{1}{n} \sum_{i=1}^{r} n_i \overline{x}_i = \frac{T}{n}$$

mediante sencillas operaciones algebraicas puede comprobarse la igualdad

$$\sum_{i=1}^{r} \sum_{j=1}^{n_i} (x_{ij} - \overline{\overline{x}})^2 = \sum_{i=1}^{r} \sum_{j=1}^{n_i} (\overline{x}_i - \overline{\overline{x}})^2 + \sum_{i=1}^{r} \sum_{j=1}^{n_i} (x_{ij} - \overline{x}_i)^2. \tag{9.2}$$

El miembro de la izquierda se denomina *suma total de cuadrados*, se representa por SST y se calcula por la expresión

$$SST = \sum_{i=1}^{r} \sum_{j=1}^{n_i} x_{ij}^2 - \frac{T^2}{n}$$

La interpretación de SST se deduce de su denominación y es clara. Si todos los datos x_{ij} fueran iguales, $\overline{\overline{x}}$ también sería igual a este valor común, siendo la variación total existente en los datos igual a cero. En ese caso, SST también sería cero.

En otros casos, SST recoge la dispersión existente en los datos de la última tabla. Cuanto mayor sea la dispersión, mayor será SST.

SST no tiene en cuenta de dónde procede la dispersión existente en los datos, si de las filas o de las columnas, y este punto es muy importante, ya que si las filas de la tabla están formadas por números idénticos, es decir, es $x_{ij} = c_i$, será $\overline{x}_i = c_i$ no existiendo variación dentro del tratamiento i-ésimo, procediendo toda la dispersión de las diferencias entre filas. Pero si, aunque las filas no sean constantes, sus efectos medios muestrales lo son, será $\overline{x}_i = k = \overline{\overline{x}} \ \ \forall i$, y el primer miembro de la derecha en la expresión (9.2), denominado *suma de cuadrados debida a los tratamientos* SST_i, será cero.

Éste se calcula por la expresión

$$SST_i = \sum_{i=1}^{r} \frac{T_i^2}{n_i} - \frac{T^2}{n}$$

Por último, el tercer miembro de (9.2) se denomina *suma residual de cuadrados*, se representa por SSE y se corresponde con aquella parte de la variación total *no explicada* por lo tratamientos.

Se calcula por diferencia de las otras dos sumas de cuadrados,

$$SSE = SST - SST_i.$$

La razón de haber descompuesto la suma total de cuadrados de la manera anterior está motivada porque si H_0 es cierta, las medias muestrales $\overline{x}_i$ tenderán a ser iguales y, por tanto, iguales a $\overline{\overline{x}}$, siendo la suma de cuadrados SST_i cercana a cero o, con más precisión, pequeña en relación a SSE.

Por tanto, si H_0 es cierta, el cociente SST_i/SSE tenderá a ser pequeño, mientras que valores grandes de este cociente tenderán a desacreditar la hipótesis nula. Ése será, salvo constantes, nuestro estadístio de contraste.

Para formalizar el test óptimo y poder calcular los puntos críticos, necesitamos determinar su distribución en el muestreo.

Teorema 9.1

(I) $SSE/\sigma^2 \rightsquigarrow \chi^2_{n-r}$.

(II) Si H_0 es cierta, entonces $SST_i/\sigma^2 \rightsquigarrow \chi^2_{r-1}$.

(III) SSE y SST_i son independientes.

Como conclusión se tiene que, si H_0 es cierta, el estadístico

$$F = \frac{\frac{SST_i}{\sigma^2}\frac{1}{r-1}}{\frac{SSE}{\sigma^2}\frac{1}{n-r}} = \frac{SST_i/(r-1)}{SSE/(n-r)}$$

seguirá una distribución F de Snedecor con $(r-1, n-r)$ grados de libertad por ser el cociente de dos χ^2 independientes divididas por sus grados de libertad.

Contraste de hipótesis

Como antes dijimos, si H_0 es falsa, el estadístico F tendrá que ser grande por lo que, en ese caso, deberemos rechazar la hipótesis nula.

En concreto, la estadística matemática propone como test óptimo de nivel α para contrastar $\begin{cases} H_0 : \mu_1 = ... = \mu_r \\ H_1 : \textit{alguna distinta} \end{cases}$ cuando se verifican las suposiciones (i), (ii) y (iii), el siguiente

$$\begin{cases} \text{Se acepta } H_0 & \text{si } F < F_{(r-1,n-r);\alpha} \\ \text{Se rechaza } H_0 & \text{si } F \geqslant F_{(r-1,n-r);\alpha} \end{cases}$$

Tabla de análisis de la varianza

Los resultados anteriores se resumen en una tabla la cual suele denominarse ANOVA, apareciendo una reproducción de la misma en ADD.

Estimador de la varianza

Una consecuencia directa que también se obtiene del análisis de la varianza es la estiamción de la varianza poblacional común σ^2. En concreto, por las propiedades de la distribución χ^2 de Pearson, un estimador de σ^2 con buenas propiedades estadísticas es

$$\hat{\sigma^2} = S^2 = \frac{SSE}{n-r}.$$

9.3. Análisis de la varianza con R

La función de R que vamos a utilizar para ejecutar el análisis de la varianza será

```
aov(modelo,datos)
```

incluyendo en le argumento `modelo` es "modelo lineal"mediante el cual expresamos la variable dependiente cuantitativa observada, en función del factor que define las poblaciones a comparar. En `datos` incluiremos las observaciones que tendrán que venir expresadas en formato *data frame*.

9.4. Análisis de las condiciones

Las poblaciones a comparar deben seguir un modelo normal y además debe verificarse la suposición de homocedasticidad, es decir, que todas ellas deben tener la misma varianza.

El análisis de normalidad de unos datos se puede efectuar gráficamente con ayuda del denominado *gráfico de normalidad* que consiste en representar en el eje de abscisas los cuantiles de la normal estándar y en el eje de ordenadas los cuantiles de la muestra; si estos pares de puntos están más o menos en la diagonal del gráfico, se tendrá que los cuantiles muestrales serán similares a los de la $N(0,1)$ y podremos concluir con la normalidad de los datos. Este gráfico se puede obtener fácilmente con R gracias a la función `qqnorm`.

El análisis de la homocedasticidad se puede hacer gráficamente mediante un *gráfico de cajas*, obtenido con la función `boxplot` y también con un test que incluimos por completar la cuestión aunque no analizamos con detalle, denominado test de Barlett y que contrasta la hipótesis nula de igualdad de las varianzas; se ejecuta con la función de R, `barlett.test`.

R tiene un función que puede utilizarse para cuando no puede admitirse la igualdad de las varianzas, la cual ejecuta un test similar a la aproximación de Welch en la comparación de dos poblaciones independientes. Se trata de la función `oneway.test`.

9.5. Comparaciones múltiples

En muchas ocasiones, rechazaremos las hipótesis nulas de igualdad de los efectos medios de las poblaciones a comparar, pudiendo hacer *comparaciones múltiples* entre los diversos tratamientos sobre los que hemos rechazado la igualdad común de todos ellos, con la idea de formar grupos de tratamientos equivalentes.

La primera idea que se le ocurrirá al lector es la de hacer tests de comparación de dos poblaciones, de nivel α, formando grupos de dos tratamientos. Este método es erróneo porque, en ese caso, el nivel de significación global ya no sería α.

Los tests denominados *comparaciones múltiples*, que expondremos a continuación en este apartado, sí tienen en cuenta este problema.

Estos tests sólo son válidos para el caso que aquí nos ocupa de un análisis de la varianza para un factor y un diseño completamente aleatorizado.

Se requiere también que el tamaño muestral de cada tratamiento sea el mismo, es decir que sea n_i constante.

Contraste de la mínima diferencia significativa (LSD)

Este contraste propone calcular la mínima diferencia significativa, definida como

$$LSD = t_{n-r;\alpha/2}\sqrt{\frac{2SSE/(n-r)}{n/r}}$$

y concluir diciendo que existe diferencia significativa, a nivel α, entre dos medias poblacionales μ_i y μ_j cuando y sólo cuando sea $|\overline{x}_i - \overline{x}_j| \geqslant LSD$.

Es decir,

$$\begin{cases} \text{Se acepta } \mu_i = \mu_j & \text{si } |\overline{x}_i - \overline{x}_j| < LSD \\ \text{Se rechaza } \mu_i \neq \mu_j & \text{si } |\overline{x}_i - \overline{x}_j| \geqslant LSD \end{cases}$$

Contraste de Tukey para una diferencia francamente significativa (HSD)

Este contraste se basa en calcular el valor HSD, definido por

$$HSD = q_{r,n-r;\alpha}\sqrt{\frac{SSE/(n-r)}{n/r}}$$

y declarar significativa cualquier diferencia que exceda de dicho valor.

El valor del punto crítico $q_{r,n-r;\alpha}$ se obtiene en unas tablas del *Recorrido Studentizado*, como *Tabla 7* de ADD.

9.6. Comparaciones múltiples con R

Con R sólo haremos comparaciones múltiples utilizando el *contraste de Tukey HSD* mediante la función

```
TukeyHSD(x,conf.level=0.95)
```

cuyo primer argumento `x` debe ser un objeto creado con la función `aov`. El segundo es el $1-$ el nivel de significación (coeficiente de confianza del intervalo de confianza/región de aceptación) de los tests donde la hipótesis nula es la igualdad de las medias de las poblaciones comparadas.

10. Regresión lineal y correlación

10.1. Introducción

El propósito del análisis de regresión y correlación es el estudio de la relación existente entre dos variables aleatorias, una denominada *independiente* o *covariable*, bajo el control del experimentador, habitualmente representada por X y con valores en el eje de abscias, y otra denominada *dependiente*, habitualmente representada por Y y con valores en el eje de ordenadas.

El *análisis de la regresión* se ocupa de estudiar la *forma* de la relación existente entre dos o más variables aleatorias, mientras que el *análisis de la correlación* investiga el grado o *fuerza* de dicha relación.

Esta relación lineal existente entre dos variables aleatorias se denomina *regresión lineal simple*, mientras que cuando se consideran más de dos covariables se hablará de *regresión lineal múltiple*.

10.2. Modelo de la regresión lineal simple

La situación general que se plantea para la regresión lineal simple es la de dos variables aleatorias, X e Y, estando interesados en inferir la existencia o no de una relación lineal entre ambas, de la forma

$$Y = \beta_0 + \beta_1 X + e$$

interpretada ésta en el sentido de que, fijados unos valores x_i de la variable X, obtendremos valores

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

de la variable Y, los cuales no llegan a estar sobre la recta $y_t = \beta_0 + \beta_1 x$ debido al error de muestreo e_i. Los parámetros β_0 y β_1 se denominan *coeficientes de regresión*.

El modelo de regresión lineal supone que los errores e_i son independientes y con distribución $N(0,\sigma)$; es decir, que dado un valor x de la variable aleatoria X, la distribución condicionada Y/x es normal $N(\mu_{y/x},\sigma)$, con $\mu_{y/x} = E[Y/x] = \beta_0 + \beta_1 x$, siendo σ^2 la varianza común a todas las distribuciones condicionadas (hipótesis de homocedasticidad), y que las distribuciones condicionadas por distintos x son independientes entre sí.

10.2.1. Interpretación de los coeficientes de regresión

En un modelos de regresión lineal se supone que la media de Y/x es de la forma

$$\mu_{y/x} = E[Y/x] = \beta_0 + \beta_1 x$$

por lo que el estimador $\hat{\beta}_1$ de la pendiente de la recta de regresión ajustada

$$y_t = \hat{\beta}_0 + \hat{\beta}_1 x$$

se interpreta como el cambio en promedio de la variable Y por el incremento en una unidad de la variable X. La estimación $\hat{\beta}_0$, es decir, la estimación de la ordenada en el origen, se interpreta como el valor promedio cuando la covariable es igual a cero. De hecho, esta interpretación es la evidente ya que la función anterior y_t es una función (una recta) de la variable x, digamos $h(x)$ que, por las observaciones del comienzo de este apartado, podemos denominar función valor promedio de Y. Si $x = 0$, entonces es $h(0) = \hat{\beta}_0$ por lo que se interpreta $\hat{\beta}_0$ como el valor promedio cuando la covariable X es igual a cero. De la misma manera, la interpretación de $\hat{\beta}_1$ es la habitual para la derivada de una función ya que, en este caso, la derivada de la función valor promedio de Y, es $h'(x) = \hat{\beta}_1$.

10.3. Contraste de la Regresión Lineal Simple

La recta de regresión lineal siempre se puede determinar y en unos casos explicará bien a la variable dependiente en función de la independiente y en otros casos, no lo hará.

10.3.1. Análisis de la variación explicada frente a la no explicada por la recta de regresión

Entre dos funciones que ajustamos por mínimos cuadrados a una nube de puntos, deberíamos elegir aquella para la cual se obtuviera una menor varianza residual. Si sólo consideramos una función, el ajuste por ésta se puede considerar bueno, si el coeficiente de determinación (relacionado con la varianza residual) era cercano a 1.

En esta sección vamos a precisar estas ideas mediante tests de hipótesis para contrastar la regresión lineal.

Si llamamos *suma total de cuadrados SST* a

$$SST = \sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}$$

la cual representa la dispersión de los datos y_i entorno a su media muestral $\overline{y}$, se puede demostrar fácilmente que es

$$\sum_{i=1}^{n}(y_i-\overline{y})^2=\sum_{i=1}^{n}(y_{t_i}-\overline{y})^2+\sum_{i=1}^{n}(y_i-y_{t_i})^2.$$

El primer miembro de la derecha, denominado *suma de cuadrados debida a la regresión lineal, SSEX* representa la parte de la suma total de cuadrados explicada por la recta de mínimos cuadrados $y_t=\beta_0+\beta_1 x$, suma de cuadrados que se calcula por la expresión

$$SSEX=\hat{\beta}_1^2\left(\sum_{i=1}^{n}x_i^2-\frac{\left(\sum_{i=1}^{n}x_i\right)^2}{n}\right).$$

La variación restante, es decir, la *suma de cuadrados no explicada* por la recta de mínimos cuadrados, será

$$SSNEX=\sum_{i=1}^{n}(y_i-y_{t_i})^2=\sum_{i=1}^{n}r_i^2=SST-SSEX$$

igual, por otra parte, a n veces la varianza residual.

Si la hipótesis nula es $H_0: X\ e\ Y$ *no están relacionadas linealmente*, y la alternativa $H_1: X\ e\ Y$ *están relacionadas linealmente*, el contraste a construir parece claro:

Si la variación explicada por la recta de mínimos cuadrados $SSEX$ es grande con respecto a la variación residual $SSNEX$, deberemos rechazar H_0; en otro caso aceptarla. Salvo constantes necesarias para obtener una distribución en el muestreo conocida, el estadístico del test a considerar será por tanto, $SSEX/SSNEX$.

Teorema 10.1

En las condiciones de normalidad antes especificadas, se tiene que

(I) $SSNEX/\sigma^2 \rightsquigarrow \chi^2_{n-2}$.

(II) Si H_0 es cierta, entonces $SSEX/\sigma^2 \rightsquigarrow \chi^2_1$.

(III) $SSEX$ y $SSNEX$ son independientes.

Como conclusión se tiene que, si H_0 es cierta, el estadístico

$$F=\frac{\frac{SSEX}{\sigma^2}}{\frac{SSNEX}{\sigma^2}\frac{1}{n-2}}=\frac{SSEX}{SSNEX/(n-2)}$$

seguirá una distribución F de Snedecor con $(1,n-2)$ grados de libertad por ser el cociente de dos χ^2 independientes divididas por sus grados de libertad.

Contraste de hipótesis

Por lo que dijimos antes, si H_0 es falsa, el estadístico F tenderá a tomar valores grandes, rechazando en ese caso H_0. Por tanto, el test óptimo de nivel α para contrastar $\begin{cases} H_0: X\ e\ Y\ \textit{no están relacionadas linealmente} \\ H_1: X\ e\ Y\ \textit{están relacionadas linealmente} \end{cases}$ es el siguiente

$$\begin{cases} \text{Se acepta } H_0 & \text{si } F<F_{1,n-2;\alpha} \\ \text{Se rechaza } H_0 & \text{si } F\geqslant F_{1,n-2;\alpha} \end{cases}$$

teniendo perfecto sentido el cálculo e interpretación del p-valor del test.

Tabla de análisis de la varianza

Los resultados anteriores se resumen en una tabla denominada de Análisis de la Varianza (ANOVA) para la regresión lineal simple, una reproducción de la cual aparece en ADD.

Estimación de la varianza común σ^2

La media de una distribución χ^2 es igual a sus grados de libertad, por lo que del apartado (i) del Teorema 10.1 anterior, será $E[SSNEX/\sigma^2]=n-2$, o bien, $E[SSNEX/(n-2)]=\sigma^2$, con lo que

$$\hat{\sigma^2}=\frac{SSNEX}{n-2}$$

será un estimador insesgado, es decir, un buen estimador, de la varianza común σ^2. Además, observemos que este valor lo obtenemos como cuadrado medio de la suma de cuadrados residual en la tabla ANOVA anteriormente mencionada.

10.3.2. Contraste de hipótesis para β_1

Una forma alternativa al análisis de la varianza anterior, para analizar si puede considerarse válida la recta de regresión determinada, es contrastar si se puede aceptar que es cero o no el parámetro β_1 de la ecuación de regresión lineal entre ambas variables.

Si se rechaza la hipótesis nula $H_0:\beta_1=0$ y se acepta la alternativa $H_1:\beta_1\neq 0$ la regresión lineal dada por la recta de regresión será aceptable, o en terminología de tests de hipótesis, existe una relación lineal significativa, ya que de hecho, el test ha resultado significativo.

El contraste que veremos, se basa en que la distribución en el muestreo de los estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$ es normal de parámetros

$$\hat{\beta}_0 \rightsquigarrow N\left(\beta_0,\sqrt{\frac{\sigma^2\sum_{i=1}^{n}x_i^2}{n\sum_{i=1}^{n}(x_i-\overline{x})^2}}\right)$$

$$\hat{\beta}_1 \rightsquigarrow N\left(\beta_1,\sqrt{\frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}}\right)$$

siendo σ^2 la varianza de la distribución condicionada Y/x.

Al conocer las distribuciones en el muestreo de los estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$, podremos determinar intervalos de confianza y tests de hipótesis para β_0 y, especialmente, para β_1.

Si denominamos

$$S_b^2=\frac{\hat{sigma}^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}=\frac{SSNEX/(n-2)}{\sum_{i}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2/n}=\frac{SSNEX/(n-2)}{SSEX/\hat{\beta}_1^2}$$

al ser independientes

$$\frac{\hat{\beta}_1-\beta_1}{\sigma}\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}\rightsquigarrow N(0,1)\quad \text{y}\quad \frac{SSNEX}{\sigma^2}\rightsquigarrow \chi^2_{n-2}$$

el estadístico

$$\frac{\frac{\hat{\beta}_1-\beta_1}{\sigma}\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}}{\sqrt{\frac{SSNEX}{(n-2)\sigma^2}}}=\frac{\hat{\beta}_1-\beta_1}{S_b}$$

seguirá una distribución t de Student con $n-2$ grados de libertad, por lo que si queremos contrastar $H_0:\beta_1=0$ frente a $H_1:\beta_1\neq 0$,

$$\begin{cases}\text{Se acepta } H_0 & \text{si } |t|<t_{n-2;\alpha/2}\\ \text{Se rechaza } H_0 & \text{si } |t|\geqslant t_{n-2;\alpha/2}\end{cases}$$

siendo el estadístico de contraste

$$t=\frac{\hat{\beta}_1}{S_b}=\sqrt{\frac{SSEX(n-2)}{SSNEX}}.$$

10.4. Regresión lineal con R

Primero determinamos la recta en el caso de regresión lineal simple mediante la función `lm()`. Después, analizamos cuáles de las covariables X_i son significativas a la hora de predecir la variable dependiente Y, mediante la función `summary()`.

Es posible crear la tabla ANOVA aplicando la función `anova()` al objeto creado con la función `lm()`.

No cabe duda que es más interesante la vía recién estudiada mediante la cual contrastamos la significación de cada covariable que el análisis de todas a la vez.

Una vez determinada la recta de mínimos cuadrados, es posible predecir un valor de Y para un $X=x_i$ dado, simplemente sustituyendo dicho x_i en la ecuación de la recta de ajuste y determinando y_{t_i}. Esta sería una estimación por punto aunque también se podría determinar un intervalo de confianza para una predicción.

10.5. Correlación lineal

Hasta ahora hemos visto si el tipo de relación existente entre dos variables aleatorias era o no lineal. Ahora contrastaremos el grado o fuerza de esta relación.

Con más precisión, si los datos $(x_1,y_1),...,(x_n,y_n)$ son valores tomados por una variable aleatoria bidimensional (X,Y), la cual suponemos con distribución normal bivariante de medias $\mu_1=E[X]$, $\mu_2=E[Y]$, varianzas $\sigma_1^2=V(X),\sigma_2^2=V(Y)$ y coeficiente de correlación ρ, el propósito de esta sección es el hacer inferencias acerca del coeficiente de correlación poblacional ρ.

10.5.1. Estimación por punto de ρ

Vimos en el capítulo 5 que un buen estimador puntual del coeficiente de correlación poblacional ρ es el coeficiente de correlación muestral, r, definido en la sección 2.4.3 como coeficiente de correlación lineal de Pearson.

Pues bien, si se determinó la recta de mínimos cuadrados y, además, la tabla de análisis de la varianza para la regresión lineal, es más sencillo calcular r como la raíz cuadrada de

$$r^2=\frac{SSEX}{SST}=\frac{\hat{\beta}_1^2\left(\sum_{i=1}^{n}X_i^2-(\sum_{i=1}^{n}X_i)^2/n\right)}{\sum_{i=1}^{n}Y_i^2-(\sum_{i=1}^{n}Y_i)^2/n}$$

aunque hay que analizar por separado el signo de r.

La estadística matemática nos sugiere también otro estimador para ρ^2, muy parecido al anterior y que usaremos menos que la estiamción por punto, que, al igual que r^2, se puede obtener de la tabla ANOVA; se trata de

$$\hat{\rho^2}=1-\frac{SSNEX/(n-2)}{SST/(n-1)}.$$

10.5.2. Contraste de hipótesis sobre ρ

En este apartado vamos a explicar cómo ejecutar el test sobre la hipótesis nula $\rho=0$.

Contraste de $H_0:\rho=0$ frente a $H_1:\rho\neq 0$

Se puede demostrar que cuando H_0 es cierta, el estadístico

$$t=r\sqrt{\frac{n-2}{1-r^2}}$$

sigue una distribución t de Student con $n-2$ grados de libertad, con lo que, fijado un nivel de significación α, se define el siguiente test

$$\begin{cases}\text{Se acepta } H_0 & \text{si } |t|<t_{n-2;\alpha/2}\\ \text{Se rechaza } H_0 & \text{si } |t|\geqslant t_{n-2;\alpha/2}\end{cases}$$

10.6. Modelo de la regresión lineal múltiple

El modelo de la regresión lineal múltiple supone una relación del tipo

$$Y=\beta_0+\beta_1X_1+...+\beta_kX_k+e$$

entre la variable dependiente Y y las k independientes $X_1,...,X_k$.

Al igual que hacíamos en el con la regresión lineal simple, estimaremos los denominados *coeficientes de regresión* $\beta_0,\beta_1,...,\beta_k$, con objeto de determinar el mejor *hiperplano de regresión muestral* de entre todos los de la forma

$$y_t=\hat{\beta}_0+\hat{\beta}_1x_1+...+\hat{\beta}_kx_k$$

Para ello, deberemos tomar una muestra de tamaño n, la cual consistirá en una matriz de datos de la forma

$$\left.\begin{array}{ccc|c}x_{11} & ... & x_{k1} & y_1\\ x_{12} & ... & x_{k2} & y_2\\ ... & ... & ... & ...\\ x_{1n} & ... & x_{kn} & y_n\end{array}\right.$$

suponiendo que las distribuciones condicionadas por distintos $(x_1,...,x_k)$ de $Y/X_1=x_1,...,X_k=x_k$ son normales

de varianza constante σ^2, e independientes entre sí unas de otras.

Como hacíamos en el caso de la regresión lineal simple, los estimadores $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k$, de los coeficientes de regresión serán los de mínimos cuadrados, es decir, aquellos que hagan mínima la suma de cuadrados

$$\sum_{j=1}^{n} e_j^2 = \sum_{j=1}^{n} \left(y_j - \hat{\beta}_0 - \hat{\beta}_1 x_{1j} - ... - \hat{\beta}_k x_{kj} \right)^2$$

los cuales resultan ser las soluciones en $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k$, de un sistema de ecuaciones denominado *sistema de ecuaciones normales*.

10.6.1. Contraste de la regresión lineal múltiple

Se puede contrastar la adecuación global del modelo o igualdad a cero de los coeficientes de regresión. Antes vimos que estos dos test eran equivalentes pues sólo había una covariable independiente; aquí no y, desde luego, son mucho más interesantes los segundos porque permitirán decidir cuáles covariables X_i son significativas y cuales no, en la explicación de la variable dependiente Y, de manera que se pueden descartar algunas de estas covariables independientes no significativas, antes de terminar la ecuación del hiperplano de regresión a utilizar en las predicciones. Haremos este análisis con la función `summary()`.