

Apuntes de Bioestadística.

R. Alberich y A. Mir

Departamento de Matemáticas e
Informàtica
Universitat Illes Balears

14 de julio de 2010

1 El análisis estadístico en la Ciencia. Análisis de Datos

Apuntes de Bioestadística.

R. Alberich y A. Mir

Departamento de Matemáticas e
Informàtica
Universitat Illes Balears

14 de julio de 2010

1 El análisis estadístico en la Ciencia. Análisis de Datos

Algunas ideas sobre la Estadística y el Método Científico

- La ciencia normal avanza definiendo teorías que intentan explicar el mundo.
- Para ello la comunidad científica elabora teorías y paradigmas que intentan explicar hechos concretos.
- Cuando alguien realiza un nuevo descubrimiento lo envía a una revisión por pares de la comunidad científica.
- Si estos aceptan el descubrimiento pasa a engrosar el cuerpo del conocimiento científico.

¿Son las teorías científicas siempre ciertas?

- No, las teorías científicas son aceptadas mientras sean la mejor explicación del mundo.
- Cuando se descubre una anomalía en una de esas teorías se intenta elaborar otra que resuelva este problema.
- Por ejemplo de la teoría del éter a la física de Einsten.
- De las teorías creacionistas o la generación espontánea a la teoría de la evolución de Darwin.

Principios básicos de las teorías científicas

- Una hipótesis es científica si existe alguna manera para comprobar su veracidad.
- La rama de la filosofía que estudia el conocimiento científico es la epistemología.
- El filósofo Karl Popper (Viena 1902-1994) fundó la corriente epistemológica del falsacionismo.
- Según esta corriente constatar una teoría significa intentar refutarla con un contraejemplo.

- Dicho de otro modo una teoría de la que no exista forma de realizar experiencias para comprobarla no es científica.
- Es decir será ciencia normal si podemos plantear experimentos para comprobar si se cumplen las afirmaciones de la teoría.
- Otros autores que han profundizado en la filosofía de la ciencia son Thomas Kuhn y Imre Lakatos.

El papel de la estadística en el método científico

- La naturaleza tiene un comportamiento incierto.
- Esto quiere decir que si repetimos bajo aproximadamente las mismas condiciones un experimento se obtienen resultados similares pero no idénticos.
- La estadística puede analizar estos resultados y ver si las desviaciones de la teoría son razonables o no.
- Un experimento estadístico es un proceso que cumple:
 - ▶ Que tiene dos o más resultados posibles.
 - ▶ Del que conocemos todos los resultados posibles.
 - ▶ Del que no podemos predecir con certeza su resultado.
 - ▶ Del que podemos explicar sus resultados a largo plazo, es lo que se denomina principio de regularidad estadística.

Reproductibilidad

- Hoy en día la investigación depende de numerosos factores: colaboración con muchos investigadores, acceso a los datos, métodos analíticos, laboratorios, programas, instrumentos....
- La posibilidad de que las investigaciones sean reproducibles es particularmente importante en los estudios que pueden influir en la decisión de políticas como las ambientales, sanitarias...

La investigación reproducible

- Muchos estudios no pueden ser replicados: falta de tiempo, falta de recursos, son únicos.
- Las TIC han aumentado de forma exponencial el acceso a los datos, estos son más complejos y llegan a ser extremadamente multidimensionales.
- Existen bases de datos que puede unirse a otras todavía más grandes.
- El poder computacional crece de forma incesante y permite cada vez más sofisticados análisis.

¿Qué es la investigación reproducible?

- Los datos brutos (micro datos, raw data,...) están disponibles.
- El código para leer estos datos es accesible.
- El código de los programas está disponible
- La documentación (artículo) incluye el accesos a los datos y a los programas.
- La distribución de esta información se hace a través de métodos estándar.

Programación Literaria

- Fue Donal E. Knuth el que introduce el concepto de Programación Literaria en 1983(*Literate Programming*).
- Kunth crea el T_EX y la herramienta WEB. Que permiten hacer programación literaria.
- El entono R dispone de métodos de programación literaria para Open Office, HTML y L^AT_EX.
- Las librerías de R para estos fines son Sweave para L^AT_EX, Openweave para Open Office y R2HTML para HTML.
- En prácticas veremos como escribir un informe y mezclar en el código R de forma que el resultado sea el documento final.
- Recomendar el artículo “Reproducible Epidemilogic Research” de R D. Peng, F. Dominici and S. L. Zeger. American journal of Epidemiology (2008).

Análisis de Datos

- Partiremos de una serie de datos sobre un colectivo de individuos.
- Utilizaremos técnicas de **estadística descriptiva** para el análisis de estos datos.
- Estas técnicas consisten en una serie de medidas, gráficos y modelos descriptivos que resumen y exploran los datos.
- El objetivo de estas técnicas es obtener una comprensión básica de los datos y las relaciones existentes entre las distintas variables analizadas.
- Así pues el **análisis exploratorio** de datos es un conjunto de técnicas, la mayoría del ámbito de la estadística descriptiva, que sirven para resumir, graficar, explicar los datos. . .
- El **análisis confirmatorio** es el que se presenta cuando tenemos una hipótesis y realizamos un experimento para confirmarla. En este caso se utilizan técnicas de **inferencia estadística**.

Algunas ideas sobre la estructura de los datos

- Los datos suelen ser multidimensionales, en el sentido de que observamos k características sobre un conjunto de n individuos.
- Estos datos hay que recolectarlos de alguna forma. Podemos escribirlos con un lápiz en un papel o podemos guardarlos en algún formato electrónico.
- Los formatos de almacenamiento de datos en un ordenador son múltiples: texto simple (codificado en distintos formatos ASCII, isolatin, utf8...), hojas de cálculo (como open office o excel), bases de datos...
- Una de las formas básicas para almacenar datos es la tabla de datos (en R `data frame`).

- En una tabla de datos u hoja de datos cada columna expresa una variable, mientras que cada fila son los resultados de las observaciones de un individuo en concreto.
- Cada individuo tendrá un nombre que lo identifica (en R `row.name`) y cada variable tendrá también un nombre (en R `name`).
- Los datos de una columna tienen todos la misma naturaleza. Es decir están formadas por datos del mismo tipo.
- Las filas tendrán naturaleza heterogénea, pues pueden contener datos de distinto tipo: Especie del individuo, sexo, peso, edad...

Naturaleza de los datos

- Cuando realizamos observaciones sobre un individuo obtenemos distintos tipos de datos.
- El análisis de los datos es distinto según el tipo de dato.
- Se pueden hacer distintas clasificaciones de los tipos de datos. Una es la siguiente:

Tipos de datos

- Datos de tipo **atributo o cualitativos**: expresan una cualidad del individuo como por ejemplo el sexo, el DNI, la especie...
- Datos **ordinales**: Son datos similares a los atributos pero que admiten comparación ordinal. Por ejemplo niveles de calidad ambiental de un ecosistema: malo, regular, normal, bueno y muy bueno.
- Datos **cuantitativos**: Los datos cuantitativos son los que se refieren a medidas. Se dividen, en **discretos** y **continuos**.
- Datos o **series temporales**: Son los que se observan a lo largo del tiempo o el espacio. No se tratarán en este curso.

Datos de tipo atributo o cualitativo

- Los datos de este tipo corresponden a observaciones sobre las cualidades de un individuo.
- Por ejemplo: el color de ojos de un mamífero, el sexo, el tipo de dieta....
- Suelen codificarse con cadenas de caracteres pero también se pueden codificar con números a los que asignaremos etiquetas.
- Estos datos pueden ser iguales o distintos. No admiten otro tipo de comparación, como la ordinal.
- Tampoco pueden ser sometidos a operaciones como la suma o el producto.

Utilidad de los datos cualitativos

Aparte de su significado, los datos cualitativos tienen las siguientes utilidades:

- Permiten segmentar las observaciones en grupos de forma que podemos comparar otras variables y en caso de encontrar variaciones en su comportamiento podrían explicar.
- Por este motivo estas variables reciben también el nombre de factores, tratamientos...
- Los distintos valores que toma un factor o tratamiento recibe el nombre de niveles.

En principio parece que estos datos aportan menos información que los de tipo cualitativo pero, por ejemplo:

- Si el factor es el sexo de una especie podríamos investigar si el peso de los especímenes difieren entre los distintos niveles de sexo.
- Si el factor es un tratamiento contra la anemia consistente donde los niveles son 5 tipos de fármacos podríamos comparar el nivel de hemoglobina, hematocrito y el número de eritrocitos después de haber aplicado el tratamiento.

Notaciones básicas para datos cualitativos

- Las estadísticas básicas para este tipo de variables son sencillas.
- Supongamos que tenemos una variable cualitativa que dispone de los niveles l_1, l_2, \dots, l_k .
- Disponemos de n observaciones de esta variable a las que denotaremos por x_1, x_2, \dots, x_n .
- Cada una de estas observaciones x_j toma como valor uno de los niveles.

Estadísticas básicas para datos cualitativos

- La **frecuencia absoluta** del nivel l_j y la denotaremos por n_j , como el número de observaciones $x_i = l_j$.
- La **frecuencia relativa** del nivel l_j es $f_j = \frac{n_j}{n}$. Por lo tanto es el tanto por uno de observaciones que corresponden a ese nivel. EL tanto por ciento de observaciones del nivel l_j es $f_j \cdot 100 \%$.
- El valor del nivel (o valores) de mayor frecuencia, absoluta o relativa, recibe el nombre de **moda**.

Gráficos básicos para datos cualitativos

- Un **diagrama de barras** (*bar plot*) es un gráfico bidimensional donde para cada nivel se dibuja una barra, en general rectangular, cuya altura que es proporcional a la frecuencia absoluta o relativa.
- Un **diagrama circular** (*pie chart*) es un círculo dividido en k sectores circulares etiquetados por cada uno de los niveles, de forma que el ángulo (el área) de cada sector es proporcional a la frecuencia absoluta o relativa del nivel que la etiqueta.

Ejemplo

Se ha realizado un seguimiento de 20 personas de un geriátrico, uno de los datos que se recogieron fue su sexo. Los niveles de sexo son $l_1 = \text{Mujer}$, $l_2 = \text{Hombre}$ Los resultados obtenidos son:

x_i	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
sexo	Mujer	Mujer	Hombre	Mujer	Mujer	Mujer	Mujer	Mujer	Hombre	Mujer
x_i	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}	x_{19}	x_{20}
sexo	Hombre	Hombre	Mujer	Mujer	Hombre	Mujer	Mujer	Mujer	Mujer	Hombre

Ejemplo

Las frecuencias absolutas de la variable sexo se presentan en la siguiente tabla:

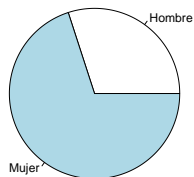
Cuadro: Frecuencias de la variable sexo.

Sexo	Frec. abs. n_j	Frec. rel. f_j	%
Hombre	6	0.3	30 %
Mujer	14	0.7	70 %
Total	20	1	100 %

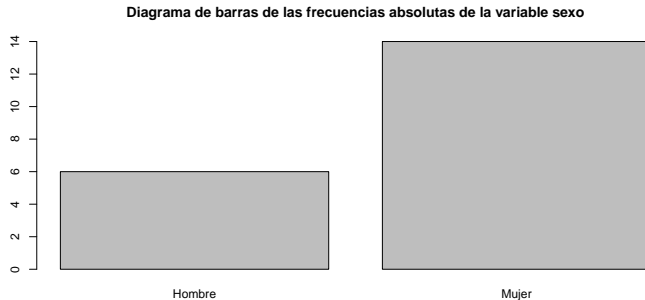
La moda de la variable sexo es el nivel Mujer.

Ejemplo: gráficos

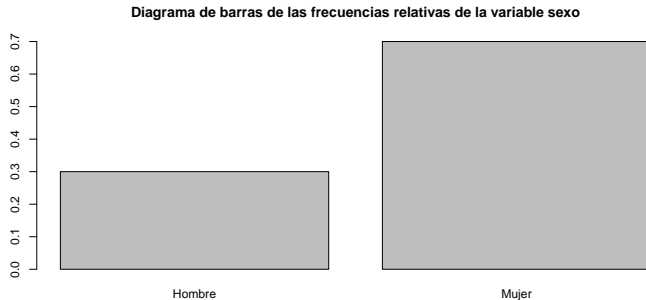
Diagrama circular de la variable sexo



Ejemplo: gráficos



Ejemplo: gráficos



Datos ordinales

- Los datos ordinales son variables de tipo cualitativo pero que tienen un orden.
- La escala *Likert* que se utiliza para saber la opinión de un grupo de personas sobre un tema determinado aporta datos ordinales.
- Por ejemplo, para saber el comportamiento **ético** de la empresa para la que trabajaban un grupo técnicos de impacto ambiental se les hizo la pregunta siguiente: ¿Cree usted que los técnicos de impacto ambiental son animados por sus empresas para que utilicen métodos que favorezcan la opinión del cliente que ha encargado el estudio?

Ejemplo: ética

Las posibles respuesta son una escala ordinal del tipo *Likert*:

Cuadro: Un ejemplo de datos ordinales

Nivel	Significado
1	Bastante en desacuerdo
2	Algo en desacuerdo
3	neutral
4	Algo de acuerdo
5	Bastante de acuerdo

Notaciones básicas para variables ordinales

- Consideremos una variable ordinal que toma los valores $l_1 < l_2 < \dots < l_k$.
- Consideremos las observaciones de estos niveles sobre n individuos x_1, x_2, \dots, x_n .
- Las definiciones de **frecuencias absolutas** (n_j), **frecuencias relativas** (f_j) y **moda** son las mismas que para datos cualitativos.
- La diferencia es que podemos definir la **frecuencia absoluta acumulada** del valor l_j como $N_j = \sum_{i=1}^j n_i$. Es decir N_j es el número de observaciones tales que $x_i < l_j$.

Estadísticos básicos variables ordinales

- La **frecuencia relativas acumulada** del valor l_j se definen como
$$F_j = \frac{N_j}{n} = \sum_{i=1}^j f_i$$
- Notemos que $N_k = n$, $\sum_{j=1}^k f_j = 1$ y $F_k = 1$.
- Si queremos obtener porcentajes basta multiplicar por 100 las frecuencias relativas o relativas acumuladas.

Ejemplo: ética (continuación)

Supongamos que hemos recogido las respuestas de 100 ($n = 100$) técnicos que elaboran informes de impacto ambiental.

Los resultados fueron:

4	4	2	1	3	3	4	1	1	3	5	2	5	2	1	2	3	2	1	1	2	2	4	2	1
1	1	2	4	5	3	4	2	4	4	3	1	3	3	2	1	5	4	1	2	2	3	3	3	1
4	3	5	1	5	1	2	5	5	2	4	5	1	4	3	1	1	4	3	3	4	4	1	2	1
3	4	1	4	2	2	4	1	3	5	3	3	3	2	2	3	3	3	2	4	1	1	4	3	2

Ejemplo: ética (continuación)

Construyamos ahora una tabla resumen con todas las frecuencias. En primer lugar calculamos las frecuencias absolutas (n_i) de cada nivel y a partir de esta el resto de las frecuencias. Notad que en este caso se pueden calcular las frecuencias acumuladas.

Valor	n_i	N_i	f_i	F_i
1	24	24	0.24	0.24
2	22	46	0.22	0.46
3	24	70	0.24	0.70
4	20	90	0.20	0.90
5	10	100	0.10	1.00
Total	100		1	

Cuadro: Tabla de frecuencia de los datos ordinales

Gráficos básicos para variables ordinales

- Como ya hemos dicho, los gráficos para frecuencias absolutas y relativas son los mismos que para variables cualitativas.
- Podemos dibujar también los gráficos de frecuencias acumuladas.
- En principio los diagramas circulares no son adecuados para frecuencias acumuladas.

Gráficos básicos para variables ordinales

- Los gráficos para frecuencias absolutas y relativas son los mismos que para variables cualitativas.
- Podemos dibujar también los gráficos de frecuencias acumuladas donde cada barra es proporcional a su frecuencia acumulada y los niveles se colocan, generalmente en orden ascendente.
- En principio los diagramas circulares no son adecuados para frecuencias acumuladas.

Diagrama de barras de las frecuencias absolutas de la variable ética

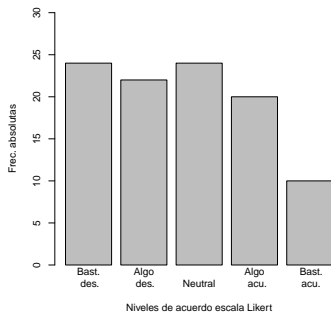


Diagrama de barras de las frecuencias absolutas acumuladas de la var. ética

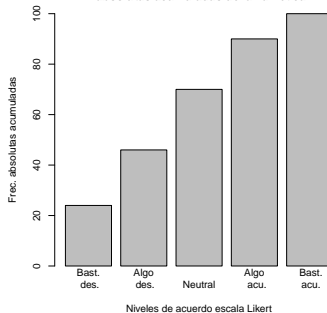


Figura: Diagrama de barras de datos ordinales frecuencias absolutas y absolutas acumuladas.

Diagrama de barras de las frecuencias relativas de la variable ética

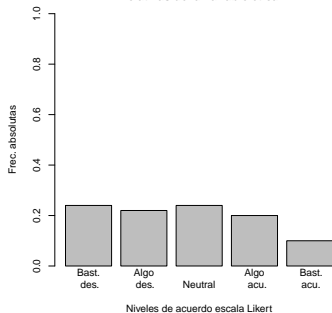


Diagrama de barras de las frecuencias relativas acumuladas de la var. ética

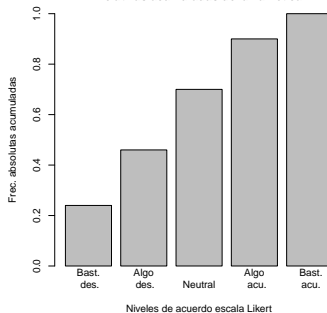


Figura: Diagrama de barras de datos ordinales frecuencias relativas y relativas acumuladas.

Datos cuantitativos

- Los datos cuantitativos son los que expresan cantidades que se representan por números.
- Por ejemplo los datos que cuentan cosas o que miden pesos, distancias, tiempos, concentraciones....
- Entre otras clasificaciones, los datos cuantitativos pueden ser **discretos** o **continuos**

Datos cuantitativos

- Los datos continuos son los que toman valores en intervalos de la recta real.
- Los datos discretos son los que toman un número finito o infinito, pero contable, de valores.
- Por ejemplo los resultados de lanzar un dado de parchís 10 veces y anotar el número de puntos observado son datos discretos.
- También lo son el número de nuevas crías de una manada, el número de aminoácidos de una proteína, ...
- Son datos continuos la edad, el peso, la estatura de un individuo, ...

Notación datos cuantitativos

La notación se diferencia de la de los datos cualitativos y ordinales en que no tenemos porqué disponer de todos los “niveles”.

- Sean x_1, \dots, x_n las observaciones cuantitativas sobre un conjunto de n individuos.
- Denotemos por X_1, \dots, X_k a los distintos valores que aparecen en las observaciones.
- Al ser los datos numéricos tienen orden y los etiquetaremos de menor a mayor, así que

$$X_1 < X_2 < \dots < X_k.$$

Principales estadísticos datos cuantitativos

- Llamaremos frecuencia absoluta de X_j y la denotaremos con n_j al número de datos de la muestra que son iguales a X_j .
- Llamaremos frecuencia absoluta acumulada de X_j y la denotaremos con N_j al número de datos de la muestra que son menores o iguales a X_j . Obviamente $N_j = \sum_{i=1}^j n_i$.
- Llamaremos N_j a la frecuencia absoluta acumulada del valor X_j al número de valores x_i tales que $x_i \leq X_j$.
- Llamaremos frecuencia relativa del valor X_j a $f_j = \frac{n_j}{n}$.
- Llamaremos frecuencia relativa acumulada del valor X_j a $F_j = \sum_{i=1}^j f_i = \frac{N_j}{n}$.

Propiedades

Las siguientes propiedades se deducen de forma sencilla:

- $n_j = N_j - N_{j-1}$, $f_j = F_j - F_{j-1}$.
- $\sum_{j=1}^k n_j = n$, $\sum_{i=1}^k f_i = 1$.
- $N_k = n$, $F_k = 1$,

Ejemplo

- Lanzamos un dado de parchís diez veces $n = 10$
- Los resultados son $x_1 = 1, x_2 = 2, x_3 = 1, x_4 = 4, x_5 = 5, x_6 = 6, x_7 = 3, x_8 = 5, x_9 = 6, x_{10} = 3$
- Los tipos de valores observados son
 $X_1 = 1, X_2 = 2, X_3 = 3, X_4 = 4, X_5 = 5, X_6 = 6$
- Podemos resumir todas las frecuencias en forma de tabla:

X_j	n_j	N_j	f_j	F_j
1	2	2	0.20	0.20
2	1	3	0.10	0.30
3	2	5	0.20	0.50
4	1	6	0.10	0.60
5	2	8	0.20	0.80
6	2	10	0.20	1.00

Presentación frecuencia datos cuantitativos

Una tabla genérica para representar las frecuencias es la que sigue:

X_j	n_j	N_j	f_j	F_j
X_1	n_1	N_1	f_1	F_1
X_2	n_2	N_2	f_2	F_2
\vdots	\vdots	\vdots	\vdots	\vdots
X_k	n_k	$N_k = n$	f_k	$F_k = 1$
Total \sum	n		1	

Medidas de tendencia central para datos cuantitativos

Las medidas de tendencia central o estadísticos de tendencia central son las que nos dan un valor representativo de todas las observaciones:

- **Moda:** es el valor (o valores) de máxima frecuencia.
- **Media aritmética:** $\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{\sum_{j=1}^k n_j \cdot X_j}{n} = \sum_{j=1}^k f_j \cdot X_j$.
- En lo que sigue cuando hablemos de “*la media de unos datos*” nos referiremos a la media aritmética salvo que indiquemos lo contrario.
- Hay otros tipos de medias como: la **media aritmética recortada**, la **media geométrica** y la **media armónica**.
- Al final de estas notas de estadística descriptiva daremos las fórmulas de estas medias.

- **Los cuantiles.** Dado un valor $0 < p < 1$ llamaremos cuantil de orden p y lo denotaremos por Q_p al valor mas pequeño cuya frecuencia relativa acumulada es mayor o igual a p .
- **La mediana, los cuartiles, los deciles, los percentiles**
 - ▶ **La mediana** es el valor más pequeño que deja su izquierda al menos la mitad de los datos, es decir es $Q_{0.5}$.
 - ▶ **Los cuartiles:** son los cuantiles $Q_{0.25}$, $Q_{0.5}$, $Q_{0.75}$ que reciben el nombre de primer cuartil, segundo cuartil (o mediana) y tercer cuartil respectivamente. Son los valores más pequeños que dejan a su izquierda al menos la cuarta parte de los datos, la mitad de los datos y las tres cuartas partes de los datos.
 - ▶ No existe un consenso para el cómputo de estos valores. En R la función `quantile` tiene hasta 9 maneras de calcular estos valores.

El cálculo de la mediana

- En este caso sí hay un consenso general para su cálculo.
- Supongamos que tenemos una muestra de datos: x_1, x_2, \dots, x_n .
- Denotaremos por $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ la muestra ordenada de menor a mayor.
- Calcularemos la mediana, $Q_{0.5}$, de la siguiente forma:

$$Q_{0.5} = \begin{cases} \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2} & \text{si } n \text{ es par} \\ x_{(\frac{n+1}{2})} & \text{si } n \text{ es impar} \end{cases}$$

El cálculo de la mediana

En este caso sí hay un consenso general para su cálculo. Supongamos que tenemos una muestra de datos: x_1, x_2, \dots, x_n

Denotaremos por $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ la muestra ordenada de menor a mayor.

Calcularemos la mediana, $Q_{0.5}$, de la siguiente forma:

$$Q_{0.5} = \begin{cases} \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2} & \text{si } n \text{ es par} \\ x_{(\frac{n+1}{2})} & \text{si } n \text{ es impar} \end{cases}$$

El cálculo de los cuartiles

- En este caso no hay un consenso general para su cálculo.
- La instrucción R `quantile` dispone de hasta 9 maneras de cálculo.
- Supongamos que tenemos una muestra de datos: x_1, x_2, \dots, x_n .
- Denotaremos por $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ la muestra ordenada de menor a mayor.
- Calcularemos el primer cuartil, $Q_{0.25}$, de la siguiente forma:

$$Q_{0.25} = \begin{cases} \text{la mediana de } x_{(1)}, x_{(2)}, \dots, x_{(\frac{n}{2})} & \text{si } n \text{ es par} \\ \text{la mediana de } x_{(1)}, x_{(2)}, \dots, x_{(\frac{n+1}{2})} & \text{si } n \text{ es impar} \end{cases}$$

- Mientras que para el tercer cuartil:

$$Q_{0.75} = \begin{cases} \text{la mediana de } x_{(\frac{n}{2})}, x_{(\frac{n}{2}+1)}, \dots, x_{(n)} & \text{si } n \text{ es par} \\ \text{la mediana de } x_{(\frac{n+1}{2})}, x_{(\frac{n+1}{2}+1)}, \dots, x_{(n)} & \text{si } n \text{ es impar} \end{cases}$$

Ejemplo

Consideremos el siguiente conjunto de datos:

$$x_1 = 6, x_2 = 3, x_3 = 2, x_4 = 1, x_5 = 5, x_6 = 6.$$

El número de datos es $n = 6$ y su media aritmética es

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{6 + 3 + 2 + 1 + 5 + 6}{6} = \frac{23}{6} = 3.8333.$$

Ejemplo

La tabla de frecuencias de estos datos es:

X_j	n_j	N_j	f_j	F_j
1	1	1	0.1667	0.1667
2	1	2	0.1667	0.3333
3	1	3	0.1667	0.5000
5	1	4	0.1667	0.6667
6	2	6	0.3333	1.0000
Total	$n = 6$		1	

Todos los cálculos se han redondeado hasta el cuarto decimal.

Para los cálculos de la media con frecuencias se suele ampliar esta tabla como sigue:

X_j	n_j	N_j	f_j	F_j	$n_j \cdot X_j$	$f_j \cdot X_j$
1	1	1	0.1667	0.1667	1.0000	0.1667
2	1	2	0.1667	0.3333	2.0000	0.3333
3	1	3	0.1667	0.5000	3.0000	0.5000
5	1	4	0.1667	0.6667	5.0000	0.8333
6	2	6	0.3333	1.0000	12.0000	2.0000
Total	$n = 6$		1		23	3.8333

Así, con los datos obtenidos de los totales de las columnas, tenemos que $\bar{x} = \frac{23}{6} = 3.8333$.

Calculemos ahora la mediana. Los datos ordenados de menor a mayor son:

$$x_{(1)} = 1, x_{(2)} = 2, x_{(3)} = 3, x_{(4)} = 5, x_{(5)} = 6, x_{(6)} = 6.$$

Como $n = 6$ es par tenemos que:

$$Q_{0.5} = \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2} = \frac{x_{(\frac{6}{2})} + x_{(\frac{6}{2}+1)}}{2} = \frac{x_{(3)} + x_{(4)}}{2} = \frac{3+5}{2} = 4.$$

Medidas de dispersión para datos cuantitativos

Las medidas de dispersión o estadísticos de dispersión, son los que nos dan un valor de los alejados que están entre sí los datos:

- **Rango:** Es la diferencia entre el máximo y el mínimo de las observaciones.
- **Rango intercuartílico:** Es la diferencia entre el tercer cuartil y el primer cuartil $Q_{0.75} - Q_{0.25}$.

- La **desviación cuadrática del dato** x_i respecto de la media es $(x_i - \bar{x})^2$.
- **La varianza:** La media de las desviaciones cuadráticas respecto de la media es la varianza.
- **Fórmulas de la varianza:**

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n} = \frac{\sum_{i=1}^n n_{ij} \cdot (X_{ij} - \bar{x})^2}{n} = \sum_{i=1}^n f_i \cdot (X_i - \bar{x})^2.$$
- **La desviación típica:** Es la raíz cuadrada positiva de la varianza $s = +\sqrt{s^2}$.

Principales estadísticos para datos cuantitativos

Medidas de dispersión, continuación:

- **La cuasivarianza:** Es una corrección de la varianza la denotamos y se calcula con $\tilde{s}^2 = \frac{n-1}{n} \cdot s^2$.
- **Fórmulas de la cuasi varianza:**

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1} = \frac{\sum_{i=1}^n n_i \cdot (X_i - \bar{x})^2}{n - 1}$$

- **La cuasi desviación típica:** Es la raíz cuadrada positiva de la cuasi varianza $\tilde{s} = +\sqrt{\tilde{s}^2}$.
- Hay otras medidas de dispersión como la desviación media o la desviación respecto de la mediana ...

Algunas propiedades de la varianza

- La varianza siempre es mayor o igual que cero: $s^2 \geq 0$.
- Si $s^2 = 0$ entonces todos los datos son iguales. Dicho de otro modo los datos son constantes y por lo tanto iguales a su media.
- De todas las desviaciones medias cuadráticas respecto de un punto, la varianza es la más pequeña. Más formalmente:

$$s^2 \leq \frac{\sum_{i=1}^n (x_i - M)^2}{n}$$

para cualquier valor M .

Algunas propiedades de la varianza

- $s^2 = \frac{\sum_{i=1}^n x_i^2}{n} - \bar{x}^2 = \frac{\sum_{j=1}^k n_j \cdot X_j^2}{n} - \bar{x}^2 = \sum_{j=1}^k f_j \cdot X_j^2 - \bar{x}^2.$
- $\tilde{s}^2 = \frac{\sum_{i=1}^n x_i^2}{n-1} - \frac{n}{n-1} \cdot (\bar{x}^2) = \frac{\sum_{i=1}^n x_i^2 - \frac{1}{n} \cdot (\sum_{i=1}^n x_i)^2}{n-1}.$
- $\tilde{s}^2 = \frac{\sum_{j=1}^k n_j \cdot X_j^2}{n-1} - \frac{n}{n-1} \cdot (\bar{x}^2) = \frac{\sum_{j=1}^k n_j \cdot X_j^2 - \frac{1}{n} \cdot (\sum_{i=1}^k n_j \cdot X_j)^2}{n-1}.$

Ejemplo

- Sigamos con los datos del ejemplo anterior:

$$x_1 = 6, x_2 = 3, x_3 = 2, x_4 = 1, x_5 = 5, x_6 = 6.$$

- El número de datos es $n = 6$ y su media aritmética es

$$\bar{x} = 3.8333.$$

- Queremos calcular la varianza, la desviación típica, la cuasivarianza y la cuasi desviación típica. Cuando sea necesario redondearemos hasta el cuarto decimal.

Ejemplo

Reconstruyamos parte de la tabla de frecuencia y la ampliaremos para nuestros cálculos:

X_j	n_j	$n_j \cdot X_j$	$n_j \cdot (X_j - \bar{x})^2$	$n_j \cdot X_j^2$
1	1	1	8.0276	1
2	1	2	3.3610	4
3	1	3	0.6944	9
5	1	5	1.3612	25
6	2	12	9.3892	72
Total	6	23	22.8333	111

En la última fila de tabla obtenemos los totales de las columnas.

Por lo tanto (redondeando siempre al cuarto decimal):

- Como $\sum_{j=1}^5 n_j = n = 6$, $\sum_{j=1}^5 n_j \cdot x_j = 23$ tenemos que $\bar{x} = \frac{23}{6} = 3.8333$
- Como $\sum_{j=1}^5 ni \cdot (X_j - \bar{x})^2 = 22.8333$ tenemos que $s^2 = \frac{22.8333}{6} = 3.8056$ y que $\tilde{s}^2 = \frac{22.8333}{5} = 4.5667$.
- Como $\sum_{j=1}^5 ni \cdot X_j^2 = 111$ tenemos que $s^2 = \frac{111}{6} - 3.8333^2 = 3.8058$ y $\tilde{s}^2 = \frac{111}{5} - \frac{6}{5}3.8333^2 = 4.5697$.
- Las discrepancias son debidas al redondeo.
- Se deja como ejercicio calcular la desviación típica y la cuasi desviación típica y también repetir estos cálculos con las fórmulas de las frecuencias relativas.

Agrupamiento de datos

- Los datos discretos son más fáciles de observar. Cuando un dato responde a un conteo suele ser fácil estimar de forma exacta el resultado. En otras ocasiones es más difícil como por ejemplo:
 - ▶ Cuando los datos son continuos y de una precisión elevada, como por ejemplo el peso, el tiempo
 - ▶ O en ocasiones, discretos con un número elevado de posibles valores (número de aminoácidos de una proteína).
- Se corre el riesgo de que las frecuencias resuman escasamente la muestra, es decir que las frecuencias absolutas de cada valor sean 1 o a lo más 2.
- Este conteo de frecuencias es poco útil para estudiar el comportamiento de los datos.
- En ambos casos se suele recurrir al conteo de datos por grupos o intervalos de valores a los que se denomina clases; es lo que se llama recuento de datos agrupados.

Los intervalos

- Consideremos el caso del peso en kilogramos de una persona. Cuando decimos yo peso 60 kilos ¿qué estoy diciendo en realidad?
- Si consideramos la edad y digo que tengo 21 años ¿qué estoy diciendo en realidad?
- En la variable continua “peso en kilos” tenemos que los valores se calculan hasta las unidades
- Si estamos tomando una medida en forma correcta decir que pesamos 60 Kg. debería ser equivalente a decir que pesamos 60 ± 0.5 Kg

- El error cometido será la mitad de la precisión del instrumento de medida.
- Lo mismo sucede con la edad; cometemos menos error si decimos que tenemos 18 años cuando tengamos 18 ± 0.5 años.
- Evidentemente en la vida real no se hace así. Si digo que tengo 19 años queremos decir que nuestra edad está en el intervalo $[19, 20)$.
- Si utilizamos esta forma de medir, el error de medida es la mitad de la precisión.

¿En qué consiste agrupar datos?

Supongamos que disponemos de un conjunto de datos x_1, x_2, \dots, x_n , del que conocemos su precisión. Queremos agrupar los datos por intervalos para contar las frecuencia por grupos, para ello:

- Necesitamos decidir el número de intervalos. Estos intervalos reciben el nombre de **clases**. Denotaremos su número por k .
- El número de intervalos puede ser determinado por el interesado. También se puede determinar con distintas formas como la de regla de Sturges, la regla de Scott o la regla de Freedman and Diaconis:

- Regla de Sturges: $k = \lceil 1 + \log_2(n) \rceil$.
- Regla de Scott: Determina primero la amplitud A de las clases $A = 3.5 \cdot s \cdot n^{-\frac{1}{3}}$ y ahora $k = \lceil \frac{(\text{máx}-\text{mín})}{A} \rceil$.
- Regla de Freedman and Diaconis: Determina primero la amplitud A de las clases $A = 2 \cdot (Q_3 - Q_2) \cdot n^{-\frac{1}{3}}$ y ahora $k = \lceil \frac{(\text{máx}-\text{mín})}{A} \rceil$.
- La función $\lceil x \rceil$ = al menor entero superior a x , es la llamada función “techo” (*ceiling*).

- Una vez determinado el valor de k , necesitamos determinar la amplitud de los intervalos A_i . La forma más sencilla es suponer que todos los intervalos son de igual amplitud A y calcular ésta, redondeándola por exceso a un valor de la precisión de la medida.
- Por ejemplo si la precisión es el 1 Kilo, $k=7$, y el rango de nuestros datos es 30, como $\frac{30}{7} = 4.2857$, el valor de $A = 5$.
- Elegir el extremo inferior y superior del intervalo de cada clase. Los denotaremos por $[L_1, L_2), [L_2, L_3), \dots, [L_k, L_{k+1})$.

- El extremo más pequeño L_1 se calcula como $L_1 = \text{mínimo} - \text{precisión}/2$, el siguiente será $L_2 = L_1 + A$, y así sucesivamente $L_j = L_{j-1} + A$.
- Determinar las marcas de clase de cada intervalo X_j . Como regla general se toma como marca el punto medio de cada intervalo $X_j = \frac{L_{j+1} + L_j}{2}$. En ocasiones se eligen otras marcas de clase para los intervalos de los extremos.

Ejemplo alergia (S. Milton pág 27)

Mucha gente manifiesta reacciones alérgicas sistémicas a las picaduras de insectos. Estas reacciones varían de paciente en paciente, no sólo en cuanto a la gravedad, sino también en el tiempo transcurrido hasta que se inicia la reacción. Los datos siguientes representan ese “tiempo de inicio hasta la reacción” en 40 pacientes que experimentaron una reacción sistémica a la picadura de abeja. Los datos están en minutos.

10.5 11.2 9.9 15.0 11.4 12.7 16.5 10.1 12.7 11.4 11.6 6.2 7.9 8.3 10.9 8.1
3.8 10.5 11.7 8.4 12.5 11.2 9.1 10.4 9.1 13.4 12.3 5.9 11.4 8.8 7.4 8.6 13.6
14.7 11.5 10.9 9.8 12.9 9.9

- El este caso la precisión de los datos es la décima de segundo.
- El número de datos es $n = 40$, la regla de Sturges consiste en tomar el entero superior a $1 + \log_2(n)$. En nuestro caso $1 + \log_2(40) = 6.3219$ por lo tanto $k = 7$ intervalos El mínimo y el máximo son 3.8 y 16.5 respectivamente. Así el rango es $16.5 - 3.8 = 12.7$.
- Tomaremos todos los intervalos de igual amplitud A . Como el rango dividido por k es $12.7/7 = 1.8143$. Redondeando la amplitud a la precisión por exceso obtenemos que $A = 1.9$.

- Así los extremos de las clases serán $L_1 = 3.8 - 0.05 = 3.75$,
 $L_2 = L_1 + 1.9 = 5.65$ y sucesivamente
 $L_3 = 7.55, L_4 = 9.45, L_5 = 11.35, L_6 = 13.25, L_7 = 15.15, L_8 = 17.05$.
- Las marcas de clase son $X_1 = 4.7, X_2 = 6.6, X_3 = 8.5, X_4 = 10.4, X_5 = 12.3, X_6 = 14.2, X_7 = 16.1$.

Ahora podemos disponer los datos en forma de tabla y calcular las frecuencias absolutas, y sus acumuladas para cada intervalo:

$[L_j, L_{j+1})$	X_j	n_j	N_j	f_j	F_j
3.75 5.65	4.70	1	1	0.03	0.03
5.65 7.55	6.60	3	4	0.07	0.10
7.55 9.45	8.50	8	12	0.20	0.30
9.45 11.35	10.40	11	23	0.28	0.58
11.35 13.25	12.30	12	35	0.30	0.88
13.25 15.15	14.20	4	39	0.10	0.97
15.15 17.05	16.10	1	40	0.03	1.00

Tabla de frecuencias de un conteo agrupado.

En general en un conteo de datos agrupados la tabla de frecuencias es:

intervalos	(Marca de clase) X_j	n_j	N_j	f_j	F_j
$[L_1, L_2)$	X_1	n_1	N_1	f_1	F_1
$[L_2, L_3)$	X_2	n_2	N_2	f_2	F_2
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
$[L_k, L_{k+1})$	X_k	n_k	N_k	f_k	F_k
Suma \sum		n		1	

Ejemplo árboles frutales

Consideremos los siguientes datos sobre el número de árboles afectados por la mosca de la fruta en 50 parcelas rústicas.

8	11	11	8	9	10	16	6	12	19
13	6	9	13	15	9	12	16	8	7
14	11	15	6	14	14	17	11	6	9
10	19	12	11	12	6	15	16	16	12
13	12	12	8	17	13	7	12	14	12

Ejemplo árboles frutales

La tabla de frecuencias agrupadas en y amplitud fija de los intervalos $A = 3$ es:

intervalos	X_j	n_j	N_j	f_j	F_j
[5.5, 8.5)	7	11	11	0.22	0.22
[8.5, 11.5)	10	11	22	0.22	0.44
[11.5, 14.5)	13	17	39	0.34	0.78
[14.5, 17.5)	16	9	48	0.18	0.96
[17.5, 20.5)	19	2	50	0.04	1.00

Los histogramas

- La descripción gráfica de los datos agrupados se hace mediante histogramas. En la fig. 2 tenemos un ejemplo de histograma.
- En este caso es el histograma de las frecuencias absolutas a la izquierda y a la derecha tenemos el gráfico de las frecuencias absolutas acumuladas.

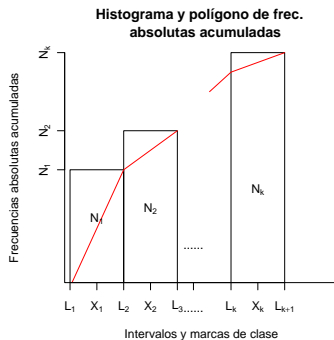
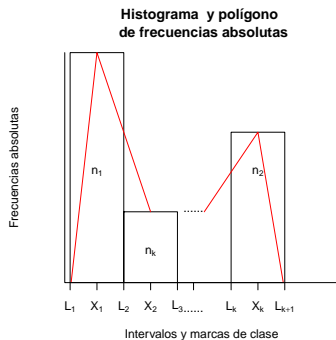


Figura: Frecuencias absolutas. Variables agrupadas

Los histogramas

- Las frecuencias absolutas n_j del gráfico de la izquierda representen las áreas de los rectángulos de la base $A_j = L_{j+1} - L_j$ la amplitud del intervalo de clase.
- Las frecuencias absolutas acumuladas N_j del gráfico de la derecha representen las alturas de los rectángulos de base $A_j = L_{j+1} - L_j$.
- La curva que une los pares ordenados (X_j, h_j) recibe el nombre se llama polígono de frecuencias absolutas (léase igual para relativas).
- El polígono de frecuencias absolutas acumuladas (de forma similar para relativas) es el formado por los puntos $(L_1, 0), (L_2, N_1), \dots, (L_{k+1}, N_k)$.

Ejemplo árboles frutales: histograma

Consideremos los 50 datos sobre árboles frutales. Tomamos intervalos de amplitud 3. El histograma de frecuencias absolutas con el correspondientes polígono de frecuencias acumuladas se muestra en la fig. 4.

Notemos que las alturas de los rectángulos se calculan teniendo en cuenta que la amplitud de los intervalos es 3:

$$h_1 = \frac{n_1}{3} = \frac{11}{3} = 3.6666, \quad h_2 = \frac{n_2}{3} = \frac{11}{3} = 3.666,$$

$$h_3 = \frac{n_3}{3} = \frac{17}{3} = 5.6666, \quad h_4 = \frac{n_4}{3} = \frac{9}{3} = 3,$$

$$h_5 = \frac{n_5}{3} = \frac{2}{3} = 0.666.$$

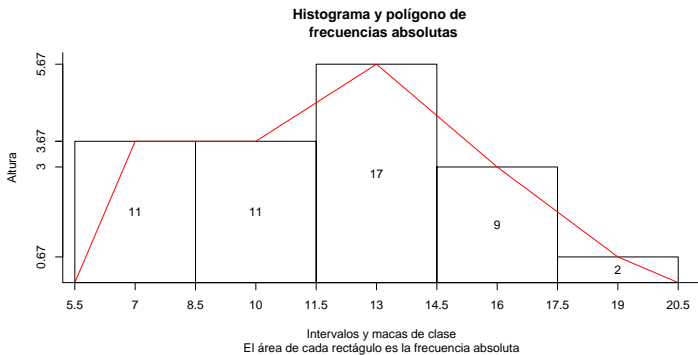


Figura: Histograma de frecuencias absolutas de los árboles frutales.

Distintos algoritmos de agrupamiento

- La instrucción básica para agrupar datos en R es `cut(datos,breaks,right=F)`.
- Donde `datos` es el vector de datos.
- El parámetro `breaks` (cortes) es el vector de límites de los intervalos de clase.
- Y `right=F` es un parámetro lógico puesto a `FALSE`, para indicar que los intervalos son abiertos a derecha.

El código siguiente hace el recuento de frecuencias absolutas, tal y como lo hemos hecho en el ejemplo anterior:

```
> options(width = 60)
> frutas <- c(8, 11, 11, 8, 9, 10, 16, 6, 12, 19,
+           13, 6, 9, 13, 15, 9, 12, 16, 8, 7, 14, 11,
+           15, 6, 14, 14, 17, 11, 6, 9, 10, 19, 12, 11,
+           12, 6, 15, 16, 16, 12, 13, 12, 12, 8, 17,
+           13, 7, 12, 14, 12)
> L <- c(5.5, 8.5, 11.5, 14.5, 17.5, 20.5)
> table(cut(frutas, breaks = L))

(5.5,8.5]  (8.5,11.5]  (11.5,14.5]  (14.5,17.5]  (17.5,20.5]
          11          11          17           9           2
```

Pero si pedimos breaks=5 el recuento de frecuencias absolutas agrupadas puede ser distinto:

```
> options(width = 60)
> frutas <- c(8, 11, 11, 8, 9, 10, 16, 6, 12, 19,
+ 13, 6, 9, 13, 15, 9, 12, 16, 8, 7, 14, 11,
+ 15, 6, 14, 14, 17, 11, 6, 9, 10, 19, 12, 11,
+ 12, 6, 15, 16, 16, 12, 13, 12, 12, 8, 17,
+ 13, 7, 12, 14, 12)
> table(cut(frutas, breaks = 5))
```

(5.99,8.59]	(8.59,11.2]	(11.2,13.8]	(13.8,16.4]	(16.4,19]
11	11	13	11	4

Lo mismo sucede si no modificamos las opciones del histograma:

```
> options(width = 60)
> frutas <- c(8, 11, 11, 8, 9, 10, 16, 6, 12, 19,
+           13, 6, 9, 13, 15, 9, 12, 16, 8, 7, 14, 11,
+           15, 6, 14, 14, 17, 11, 6, 9, 10, 19, 12, 11,
+           12, 6, 15, 16, 16, 12, 13, 12, 12, 8, 17,
+           13, 7, 12, 14, 12)
> hist(frutas)
```

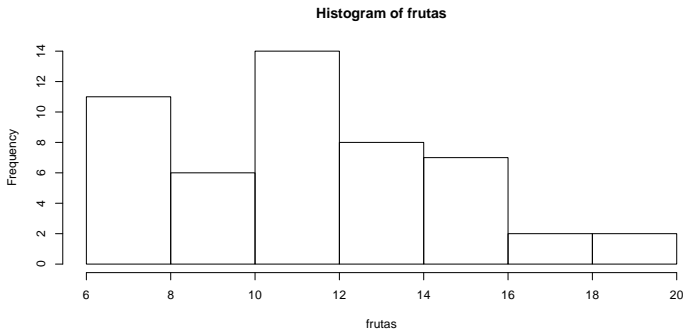


Figura: Otro histograma de las frecuencias absolutas de los árboles frutales

Mismos nombres distintos resultados

- Hemos visto que los algoritmos de agrupamiento y la forma de dibujar histogramas pueden ser distintas.
- Esto es debido a que hay varias maneras de agrupar datos.
- Respecto a los histogramas, se suelen confundir con los diagramas de barras.
- De hecho el concepto de histograma, como pone la ayuda de R para la función `hist` puede diferir según el idioma.
- Si el objetivo del histograma es averiguar el perfil de las frecuencias, debemos usar un histograma en el que las áreas representen a las frecuencias.
- Si todos los intervalos de clase son de la misma amplitud el histograma de áreas o de alturas es similar. En caso contrario pueden dar severas distorsiones.

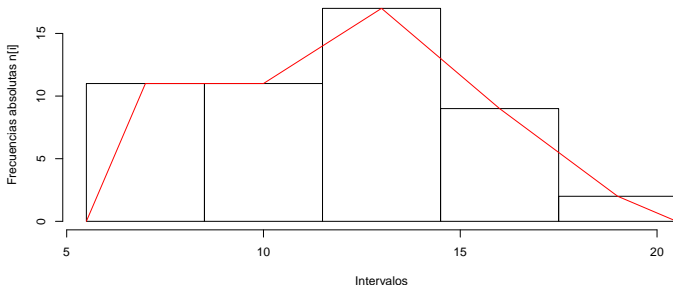

```

> options(width = 55)
> frutas <- c(8, 11, 11, 8, 9, 10, 16, 6, 12,
+           19, 13, 6, 9, 13, 15, 9, 12, 16, 8, 7,
+           14, 11, 15, 6, 14, 14, 17, 11, 6, 9, 10,
+           19, 12, 11, 12, 6, 15, 16, 16, 12, 13,
+           12, 12, 8, 17, 13, 7, 12, 14, 12)
> L <- c(5.5, 8.5, 11.5, 14.5, 17.5, 20.5)
> h <- hist(frutas, breaks = L, main = "Hist. freq. abs.",
+           sub = "Este histograma es la salida estándar de R",
+           xlab = "Intervalos", ylab = paste("Frec. abs.",
+           expression(n[i]))))
> lines(c(min(h$breaks), h$mids, max(h$breaks)),
+       c(0, h$counts, 0), type = "l", col = "red")

```

Produce la fig. 13.

Histograma y polígono de frecuencias absolutas



Este histograma es la salida estándar de R

Figura: Histograma (real) de los datos de árboles frutales

Como se ve dibuja el histograma de una forma diferente (ejercicio comentar).

Datos brutos, datos agregados.

- Se habla de datos brutos, micro datos,... cuando disponemos de los datos originales sin ningún tratamiento informático ni estadístico.
- Se habla de datos agregados, elaborados,... cuando los datos han sufrido algún proceso, como el de agrupamiento.
- Como responsables de tratar los datos podemos disponer de los datos brutos, agrupados o de ambos.
- Los estadísticos básicos de los datos se calculan en cada situación de forma diferente y en ocasiones de ¡varias formas!

Principales estadísticos para datos agrupados

Ya vimos como se calculaban los principales estadísticos para datos brutos (sin agrupar). Veamos cómo se calculan para datos agrupados:

- Los estadísticos \bar{x} , s^2 , \tilde{s}^2 , s , \tilde{s} , se calculan con las mismas fórmulas para frecuencias absolutas y relativas que para datos brutos.
- La diferencia es que para datos brutos los valores X_j son los valores que se observan en los datos y para datos agrupados los valores X_j son las marcas de clase.

- Los estadísticos básicos pueden tener distintos valores cuando se calculan desde datos brutos o desde datos agrupados. Esto es debido a que el tratamiento agrupado puede tener menos precisión que el de los datos brutos.
- Esto no es en general un demérito para agrupar los datos. Es muy posible que los datos agrupados den más significado a la interpretación de los datos brutos.
- Cuando tenemos intervalos de clase se dice **intervalo modal** al que alcanza mayor frecuencia absoluta o relativa.
- Existen fórmulas de aproximación de la moda para datos agrupados. No las veremos pues creemos que carecen de interés en este curso.

Principales estadísticos para datos agrupados

- Respecto a los gráficos ya hemos hablado de los histogramas.
- Comentar sobre los histogramas que no es en general admisible que encontremos clases con frecuencia nula.
- Salvo comportamientos patológicos, como la existencia de dos poblaciones muy diferentes, es extraño encontrarse clases vacías. Esto suele pasar si el número de datos es pequeño y su varianza es grande.
- En principio es aconsejable unir las clases vacías a una de sus clases contiguas.
- Respecto a los cuantiles para datos agrupados sí hay una manera (tradicionalmente estándar) de cálculo que expondremos más adelante.

Ejemplo: Estadísticos de datos brutos y agrupados

Para ilustrar las diferencias entre los estadísticos para datos brutos y agrupados consideremos los siguientes datos:

10	5	2	7	9	5	7	6	5	9
12	2	6	6	9	12	6	6	6	4
9	7	12	11						

La media aritmética de los datos anteriores sin agrupar en intervalos es:

$$\bar{x} = \frac{10 + 5 + 2 + \cdots + 12 + 11}{24} = \frac{173}{24} = 7.20833$$

Si los agrupamos en intervalos de amplitud 3, la media será (hacemos primero la correspondiente tabla de frecuencias)

intervalos	X_j	n_j	$n_j X_j$
[1.5, 4.5)	3	3	9
[4.5, 7.5)	6	12	72
[7.5, 10.5)	9	5	45
[10.5, 13.5)	12	4	48
Suma		24	174

$$\bar{x} = \frac{174}{24} = 7.25$$

Notemos que los valores difieren del de los datos brutos ya que el agrupamiento provoca una pérdida de precisión.

Ejemplo estadísticos datos agrupados

Consideremos la siguiente distribución de frecuencias

intervalos	X_j	n_j	$n_j X_j$
[9.5, 29.5)	19.5	38	741.0
[29.5, 49.5)	39.5	18	711.0
[49.5, 69.5)	59.5	31	1844.5
[69.5, 89.5)	79.5	20	1590.0
Sumas		107	4886.5

- Vamos a calcular la varianza, primero necesitamos calcular la media
 $\bar{x} = \frac{4886.5}{107} = 45.6682$.
- Para calcular la varianza hemos de añadir dos columnas a la tabla anterior

X_j	X_j^2	$n_j X_j^2$
19.5	380.25	14449.50
39.5	1560.25	28084.50
59.5	3540.25	109747.75
79.5	6320.25	126405.00
Suma		278686.75

La varianza y la desviación típica valen

$$s_X^2 = \frac{278686.75}{107} - 45.6682^2 = 518.962 \quad s_X = \sqrt{518.962} = 22.7807.$$

Cuantiles para datos agrupados

- Veamos alguna manera de cálculo aproximado de la mediana ($Q_{0.5}$) a partir de las frecuencias de los datos agrupados.
- Necesitaremos las columnas de frecuencias absolutas y la de frecuencias absolutas acumuladas para los datos agrupados.

En general una tabla de frecuencias agrupadas es:

intervalos	X_j	n_j	N_j
$[L_1, L_2)$	X_1	n_1	N_1
$[L_2, L_3)$	X_2	n_2	N_2
\vdots	\vdots	\vdots	\vdots
$[L_I, L_{I+1})$	X_I	n_I	N_I
Σ		n	

- Llamaremos intervalo crítico para la mediana al primer intervalo en el que su frecuencia absoluta acumulada supere o iguale a $\frac{n}{2}$.
- Denotemos por $[L_c, L_{c+1})$ el intervalo crítico.
- Sea N_{c-1} la frecuencia absoluta acumulada del intervalo anterior al crítico.
- En el caso en que el intervalo crítico sea el primero, $N_{c-1} = 0$.
- Sea n_c la frecuencia absoluta del intervalo crítico.
- Sea $A_c = L_{c+1} - L_c$ la amplitud del intervalo crítico.
- Una aproximación para la **mediana** la da la siguiente fórmula:

$$Q_{0.5} = L_c + A_c \frac{\left(\frac{n}{2} - N_{c-1}\right)}{n_c}.$$

Cuantiles para datos agrupados

- La justificación de la fórmula anterior es la siguiente.
- Si representásemos las frecuencias absolutas acumuladas entre los extremos de los intervalos, la mediana sería la antiimagen de $\frac{n}{2}$ en el intervalo crítico haciendo una interpolación por lineal (ver figura 7).

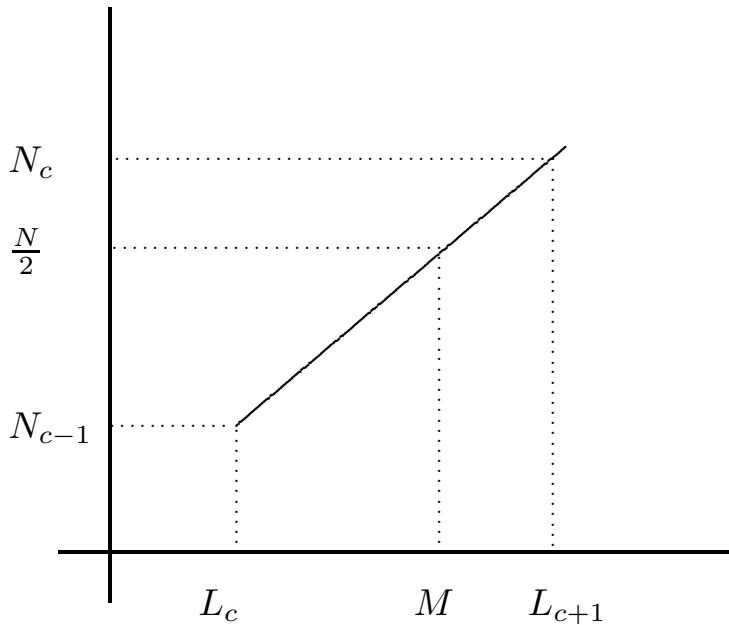


Figura: Interpretación geométrica de la Mediana

- Los cuantiles son una generalización de la mediana. La mediana es el cuantil 0.5 ya que deja el 50 % de las observaciones a su izquierda.
- En general el **cuantil** p es aquel valor que deja el $p \cdot 100\%$ de las observaciones a su izquierda. Por lo tanto el cuantil p es Q_p . El cálculo, dada la distribución de frecuencias es semejante al cálculo de la mediana.
- Definimos el intervalo crítico en este caso como el primer intervalo del que su frecuencia absoluta acumulada supera o iguala a $n \cdot p$.

- Sean entonces, $[L_c, L_{c+1})$ el intervalo crítico, N_{c-1} la frecuencia absoluta acumulada del intervalo anterior al crítico y n_c la frecuencia absoluta del intervalo crítico.
- Si denotamos por A_c a la amplitud del intervalo crítico, la fórmula para calcular el **cuantil** p es: $Q_p = L_c + A_c \frac{(n \cdot p - N_{c-1})}{n_c}$.

Ejemplo del cálculo de la mediana.

Calculemos la mediana, sin agrupar, de los siguientes datos:

14, 15, 16, 18, 18, 18, 18, 19, 20, 20, 22.

El tamaño de la muestra es $n = 11$ observaciones y ya están ordenadas. El lugar central, es el que ocupa el sexto puesto, el valor que ocupa este lugar es el 18, por lo tanto, la mediana es 18.

En la siguiente muestra tenemos un número par de datos:

24, 25, 26, 26, 27, 27, 27, 29.

El tamaño muestral es $n = 8$ observaciones que ya están ordenadas. El lugar central estará entre el cuarto y el quinto puesto. Los datos que ocupan estos lugares son el 26 y el 27. Por lo tanto la mediana vale

$$Q_{0.5} = \frac{26 + 27}{2} = 26.5.$$

Ejemplo

Consideremos la siguiente distribución de frecuencias:

intervalos	X_j	n_j	N_j
[1.5, 4.5)	3	3	3
[4.5, 7.5)	6	12	15
[7.5, 10.5)	9	5	20
[10.5, 13.5)	12	4	24

Tenemos que $n = 24$ y que $\frac{n}{2} = 12$. El intervalo crítico es: [4.5, 7.5)

La mediana valdrá entonces:

$$Q_{0.5} = 4.5 + 3 \frac{(12 - 3)}{12} = 6.75.$$

Cuantil 0.25: 25 % $\Rightarrow n \cdot p = 6$. Intervalo crítico: [4.5, 7.5).

$$Q_{0.25} = 4.5 + 3 \frac{(6 - 3)}{12} = 5.25$$

Cuantil 0.75: 75 % $\Rightarrow n \cdot p = 18$. Intervalo crítico: [7.5, 10.5).

$$Q_{0.75} = 7.5 + 3 \frac{(18 - 15)}{5} = 9.3$$

- Los cuartiles que dividen a la población en cuartos son llamados cuartiles, así el primer cuartil $Q_{0.25}$ deja a su izquierda el 25 % de las observaciones, el segundo cuartil $Q_{0.5}$ es la mediana y el tercer cuartil $Q_{0.75}$ deja a su izquierda el 75 % de las observaciones.
- También se habla de los deciles que son los estadísticos que dividen a la población en décimas partes: $Q_{0.1}, Q_{0.2}, \dots$
- Los percentiles son los que dividen la muestra en centésimas partes.

Cambios lineales

- Supongamos que tenemos una serie de datos x_1, x_2, \dots, x_n de una variable X .
- Les vamos a realizar una transformación lineal $Y = a \cdot X + b$.
- Obtendremos la serie de datos
 $y_1 = a \cdot x_1 + b, \quad y_2 = a \cdot x_2 + b, \dots, y_n = a \cdot x_n + b$.
- La pregunta es ¿cómo afecta esta operación a los estadísticos básicos?
- Denotaremos por \bar{x} , s_X^2 , a la media y la varianza de los datos X y denotaremos por \bar{y} y s_Y^2 a la media y la varianza de los datos de la variable Y .

Interpretación cambios lineales

- Si $b > 0$ los datos se desplazan su origen en una cantidad b a su derecha.
- Si $b < 0$ los datos se desplazan su origen en una cantidad b a su izquierda.
- Por esto motivo sumar una cantidad b recibe el nombre de **cambio de origen**.
- Si $a \geq 1$ las unidades de los datos aumenta su escala en esa proporción.
- Si $0 < a < 1$ los datos reducen su escala es esa proporción.
- Si $a < 0$ la interpretación es similar salvo porque los datos sufren también un cambio de orientación.
- Es por esos motivos por lo que multiplicar los datos por una cantidad a recibe el nombre de **cambio de escala**.

Se cumplen las siguientes propiedades:

- La media queda afectada de la misma forma que los datos por el cambio lineal.
- Es decir $\bar{y} = a \cdot \bar{x} + b$.
- La varianza es independiente respecto a cambios de origen y que queda afectada por el cuadrado de los cambios de escala.
- Es decir $s_Y^2 = a^2 s_X^2$.
- Para las desviaciones típicas tendremos

$$s_Y = |a| s_X.$$

Puntuaciones típicas

- Consideremos los datos de una variable X , x_1, x_2, \dots, x_n de los que conocemos \bar{x} y s_X . Consideremos el cambio lineal, o de escala y origen, $Z = \frac{X - \bar{x}}{s_X}$.
- Las puntuaciones $z_1 = \frac{x_1 - \bar{x}}{s_X}$, $z_2 = \frac{x_2 - \bar{x}}{s_X}$, \dots , $z_n = \frac{x_n - \bar{x}}{s_X}$ reciben el nombre de puntuaciones típicas o tipificadas o estándar de los datos X .
- Estas puntuaciones cumplen que $\bar{z} = 0$ y $s_z = 1$. Es decir transformamos los datos a unos datos que tienen media cero y varianza 1.
- Las puntuaciones típicas, entre otras cosas, son útiles para comparar distribuciones de frecuencias de dos o más variables medidas en distintas unidades.

Ejemplo de datos agrupados

Consideremos la siguiente distribución de frecuencias

intervalos	X_j	n_j	$n_j X_j$
[9.5, 29.5)	19.5	38	741.0
[29.5, 49.5)	39.5	18	711.0
[49.5, 69.5)	59.5	31	1844.5
[69.5, 89.5)	79.5	20	1590.0
Suma		107	4886.5

- Vamos a calcular la varianza, primero necesitamos calcular la media
 $\bar{x} = \frac{4886.5}{107} = 45.6682$.
- Para calcular la varianza hemos de añadir dos columnas a la tabla anterior

X_j	X_j^2	$n_j X_j^2$
19.5	380.25	14449.50
39.5	1560.25	28084.50
59.5	3540.25	109747.75
79.5	6320.25	126405.00
Suma		278686.75

La varianza y la desviación típica valen

$$s_X^2 = \frac{278686.75}{107} - 45.6682^2 = 518.962 \quad s_X = \sqrt{518.962} = 22.7807.$$

Ejemplo de datos agrupados

- El coeficiente de variación se define como el cociente entre la desviación típica y la media aritmética.
- Se utiliza para variables en las que la media represente a la magnitud de los datos. Esto sucede cuando por ejemplo todos son positivos.
- La notación y fórmula para el cálculo es $CV = \frac{s}{\bar{x}}$.

- El coeficiente de variación es independiente del cambio de escala.
- Más concretamente, consideremos el cambio lineal de la variable X $Y = aX$, con $a > 0$, el coeficiente de variación de la variable Y es el mismo que el de la variable X :

$$CV_Y = CV_X.$$

- El coeficiente de variación será útil para comparar la dispersión de distribuciones de frecuencias de dos o más variables en diferentes escalas.

Ejemplo

Consideremos la siguiente distribución de frecuencias:

intervalos	X_j	n_j	$n_j X_j$
[9.5, 29.5)	19.5	38	741.0
[29.5, 49.5)	39.5	18	711.0
[49.5, 69.5)	59.5	31	1844.5
[69.5, 89.5)	79.5	20	1590.0
Sumas		107	4886.5

Ejemplo

- La media y la desviación típica son $\bar{x} = 45.6682$, $s_x = 22.7807$.
- Por lo tanto el coeficiente de variación es $CV = \frac{s}{\bar{x}} = \frac{22.6807}{45.6682} = 0.4988$.

Perfil de una distribución

- El perfil de una distribución viene determinado por alguno de sus polígonos de frecuencias.
- Es mejor utilizar las frecuencias relativas ya que no dependen del tamaño de la muestra.
- La idea es encontrar la curva la que tiende el polígono de frecuencias cuando la muestra se hace grande, que en definitiva sería la curva de frecuencias de toda la población.

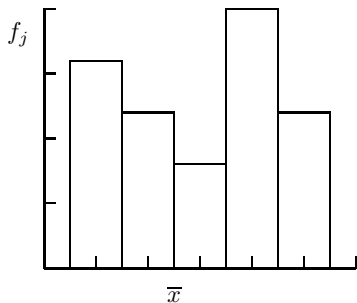
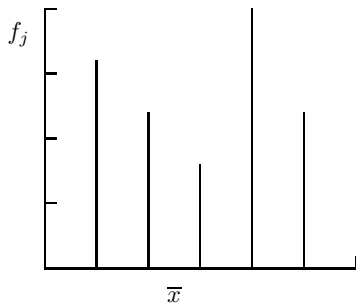


Figura: Diagrama de barras e histograma de las frecuencias relativas

El perfil de una distribución

- Una curva continua en forma de campana llamada curva de Gauss. Puede servir como un modelo matemático ideal para comparar el perfil de cualquier distribución.
- Esta curva corresponde a la gráfica de la función

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

- Donde μ se aproxima por \bar{x} y σ por s .
- Su representación gráfica es la de la figura 9, gaussiana o campana de gauss, para el caso (estándar) en el que $\mu = 0$ y $\sigma = 1$.

La campana de gauss

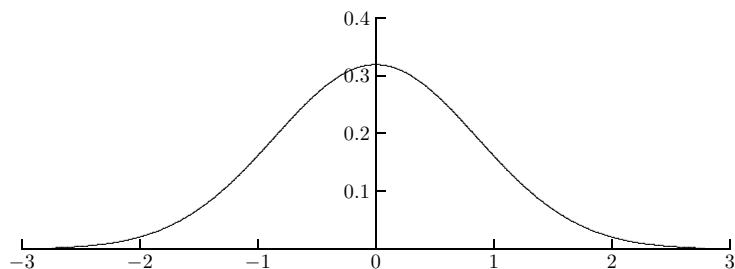


Figura: Curva normal o campana de Gauss

Propiedades de la curva normal

Las propiedades más importantes de la curva normal son:

- a) Está definida para cualquier real y es siempre positiva.
- b) El área comprendida entre la curva y el eje de abscisas vale siempre 1 para cualquier valor de μ y $\sigma > 0$.
- c) Es simétrica respecto a la recta vertical $X = \mu$ y en este punto tiene un máximo absoluto que vale $\frac{1}{\sqrt{2\pi}\sigma}$.
- d) Tiene dos puntos de inflexión en $x = \mu \pm \sigma$.
- e) El eje de abscisas es una asíntota de la curva.

La normal y las medidas de simetría y apuntamiento

- Las medidas de simetría y apuntamiento se suelen referir a la correspondiente distribución normal.
- Es decir aquella en la que los parámetros se estiman por $\mu = \bar{x}$ y $\sigma = s$. (o por la cuasivarianza).
- Se entiende, entonces, que la distribución normal es simétrica y es perfecta respecto al apuntamiento.
- Es decir, que no es ni apuntada ni chata.

Índice de simetría

- Para ver si una distribución es simétrica o asimétrica por la derecha o por la izquierda se toma como índice de simetría (*skewness*): $g_1 = \frac{m_3}{s^3}$,
- Donde m_3 es el momento central de tercer orden y se calcula de la siguiente forma: $m_3 = \frac{1}{n} \sum_{j=1}^k n_j (X_j - \bar{x})^3$,
- Por supuesto, s es la desviación típica.

Índice de simetría

Interpretación del índice de simetría:

- Si $g_1 > 0$, la distribución es asimétrica por la derecha o asimetría positiva.
- Si $g_1 = 0$, la distribución es simétrica o el índice no decide.
- Si $g_1 < 0$, la distribución es asimétrica por la izquierda o asimetría negativa.

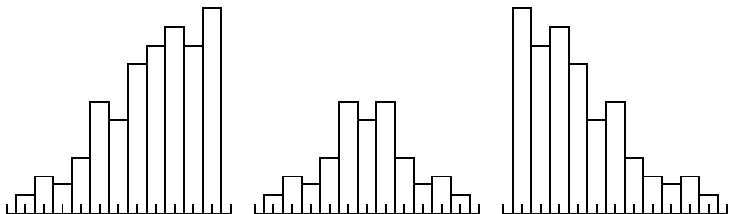


Figura: Histogramas, de izquierda a derecha, para $g_1 > 0$, $g_1 \approx 0$ y $g_1 < 0$.

Ejemplo simetría

Consideremos la siguiente distribución de frecuencias:

intervalos	X_j	n_j	$n_j X_j$	$n_j X_j^2$
[14.5, 19.5)	17	4	68	1156
[19.5, 24.5)	22	6	132	2904
[24.5, 29.5)	27	8	216	5832
[29.5, 34.5)	32	11	352	11264
[34.5, 39.5)	37	35	1295	47915
[39.5, 44.5)	42	100	4200	176400
[44.5, 49.5)	47	218	10246	481562
Suma		382	16509	727033

Ejemplo simetría

- La media y la varianza valen:

$$\bar{x} = \frac{16509}{382} = 43.2173, \quad s_X^2 = \frac{727033}{382} - \left(\frac{16509}{382}\right)^2 = 35.49$$

- Calculemos el coeficiente de simetría g_1 . Para hacerlo, hemos de añadir una columna más a la tabla anterior.

Ejemplo simetría

X_j	n_j	$n_j(X_j - \bar{x})^3$
17	4	-72081.33
22	6	-57308.66
27	8	-34121.16
32	11	-15525.84
37	35	-8411.41
42	100	-180.37
47	218	11799.67
Sumas	382	-175829.09

Ejemplo simetría

- El momento de tercer orden vale $m_3 = \frac{-175829.09}{382} = -460.285$.
- Calculemos el índice de simetría $g_1 = \frac{m_3}{s^3} = \frac{-460.285}{(\sqrt{35.49})^3} = -2.18$.
- Por lo tanto podemos decir que se trata de una distribución asimétrica por la izquierda o negativa.

Medidas de apuntamiento para datos numéricos

- Las medidas de apuntamiento nos miden si el perfil de una distribución muestral está muy apuntado o no en comparación con un perfil ideal.
- EL modelo ideal es el de la campana de gauss asociada.
- Para estudiar el apuntamiento se utiliza un índice basado en el momento de cuarto orden, que recibe el nombre de coeficiente de apuntamiento o curtosis (*kurtosis*.)

- La fórmula de la curtosis es

$$g_2 = \frac{m_4}{s^4} - 3,$$

- Donde m_4 es el llamado momento central de cuarto orden y se calcula de la siguiente forma:

$$m_4 = \frac{1}{n} \sum_{j=1}^k n_j (X_j - \bar{x})^4,$$

- Donde s es la desviación típica.

Medidas de apuntamiento para datos numéricos

Tenemos, pues que:

- Si $g_2 > 0$, la distribución es puntiaguda o leptocúrtica.
- Si $g_2 = 0$, la distribución es similar a la normal o mesocúrtica.
- Si $g_2 < 0$, la distribución es achatada o platicúrtica.

Medidas de apuntamiento para datos numéricos

La siguiente figura ilustra las diferencias entre los valores de apuntamiento.

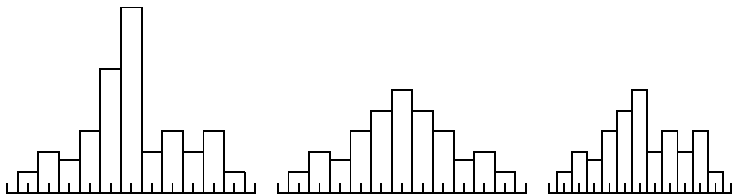


Figura: Histogramas de los tres tipos de apuntamiento

Ejemplo apuntamiento

Consideremos la siguiente distribución de frecuencias:

intervalos	X_j	n_j	$n_j X_j$	$n_j X_j^2$
[14.5, 19.5)	17	4	68	1156
[19.5, 24.5)	22	6	132	2904
[24.5, 29.5)	27	8	216	5832
[29.5, 34.5)	32	11	352	11264
[34.5, 39.5)	37	35	1295	47915
[39.5, 44.5)	42	100	4200	176400
[44.5, 49.5)	47	218	10246	481562
Sumas		382	16509	727033

Ejemplo apuntamiento

- La media y la varianza valen

$$\bar{x} = \frac{16509}{382} = 43.22, \quad s_X^2 = \frac{727033}{382} - \left(\frac{16509}{382}\right)^2 = 35.49.$$

- Calculemos coeficiente de apuntamiento g_2 . Hemos de añadir una columna a la tabla:

intervalos	X_j	n_j	$n_j(X_j - \bar{x})^4$
[14.5, 19.5)	17	4	1889776.11
[19.5, 24.5)	22	6	1215933.65
[24.5, 29.5)	27	8	553352.38
[29.5, 34.5)	32	11	174157.64
[34.5, 39.5)	37	35	52296.07
[39.5, 44.5)	42	100	219.56
[44.5, 49.5)	47	218	44634.88
Sumas		382	3930370.29

Ejemplo apuntamiento

- El momento de cuarto orden vale $m_4 = \frac{3930370.29}{382} = 10288.93$.
- Calculemos el índice de apuntamiento
$$g_2 = \frac{m_4}{s^4} - 3 = \frac{10288.93}{35.49^2} - 3 = 5.17.$$
- Por lo tanto se trata de una distribución puntiaguda o leptocúrtica.

El índice de apuntamiento es independiente respecto cambios lineales de la forma $Y = aX + b$, es decir:

$$g_2(X) = g_2(Y).$$

El índice g_2 no queda afectado por cambios de origen ni de escala.

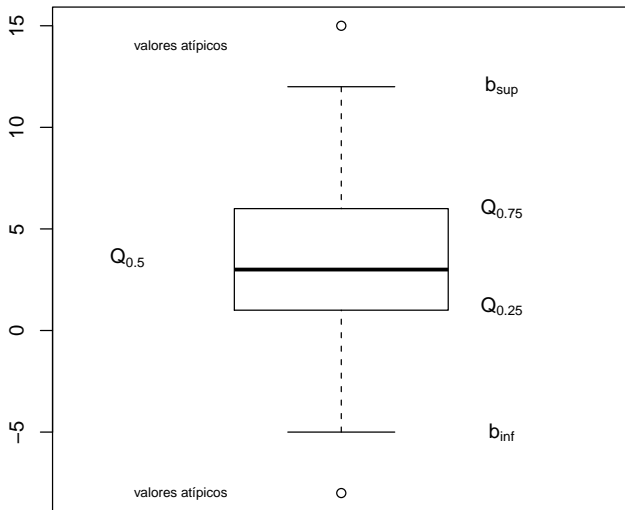
El coeficiente de variación es independiente del cambio de escala. Más concretamente, si hacemos el cambio lineal de la variable X $Y = aX$, con $a > 0$, el coeficiente de variación de la variable Y es el mismo que el de la variable X :

$$CV_Y = CV_X.$$

El coeficiente de variación será útil para comparar la dispersión de distribuciones medidas en diferentes escalas.

El diagrama de caja

- EL diagrama de caja es un gráfico que resume algunos estadísticos de una serie de datos en un gráfico.
- Este gráfico está formado por cinco números que en orden son:
 $b_{inf}, Q_{0.25}, Q_{0.5}, Q_{0.75}, b_{sup}$
- Los valores b_{inf}, b_{sup} determinan los extremos o bigotes del dibujo. Si denotamos por m y M el mínimo y el máximo de los datos, se calculan de la siguiente forma: $b_{inf} = \min\{m, Q_{0.25} - 1.5 \cdot (Q_{0.75} - Q_{0.25})\}$
 $b_{sup} = \max\{M, Q_{0.75} + 1.5 \cdot (Q_{0.75} - Q_{0.25})\}$
- La fig. 3 explica como se dibuja el diagrama de caja (*box plot*)



Variables multidimensionales

- Hasta ahora sólo hemos estudiado una variable, es evidente que en la realidad interesa el comportamiento conjunto de dos o más variables.
- En cualquier disciplina técnica o científica, economía, ciencias de la computación, bioinformática, telecomunicaciones, . . . son muy utilizados los conceptos de asociación, independencia y otros, entre dos o más variables.

Variables multidimensionales

- Para introducirlos estudiaremos el caso más sencillo; el de las variables estadísticas bidimensionales. En lo que respecta a esta sección cada individuo de la población tiene asociado más de un valor o cualidad observada.
- Por ejemplo peso y altura de un grupo de personas, peso y sexo, altura y nivel de estudios,

Variables multidimensionales

- Por ejemplo si estudiamos el peso (p) y la altura (h) de una población una muestra genérica de tamaño n tendría el siguiente aspecto:

$$(p_1, h_1), (p_2, h_2), \dots, (p_n, h_n),$$

- Donde (p_i, h_i) es el peso y la estatura correspondientes a la observación i -ésima.
- Otro ejemplo sería el estudio de la relación entre los turistas llegados a nuestra isla y el año de llegada.
- Los datos serían $(t_1, n_1), (t_2, n_2), \dots, (t_N, n_N)$ donde t_i es el año i -ésimo y n_i = número de turistas llegados ese año.

Descripción numérica: caso bidimensional

- Supongamos que tenemos (X, Y) un par de variables que se pueden medir conjuntamente en un individuo de la población que se desea estudiar.
- Sean $\{X_1, X_2, \dots, X_I\}$ los valores que han tomado los datos de X y $\{Y_1, Y_2, \dots, Y_J\}$ los de Y .
- El conjunto de valores que puede tomar la variable conjunta (X, Y) son: $\{(X_1, Y_1), \dots, (X_1, Y_J), (X_2, Y_1) \dots, (X_2, Y_J), \dots, (X_I, Y_1), \dots, (X_I, Y_J)\} ..$

- Sean n_{ij} la frecuencia absoluta correspondiente al valor (X_i, Y_j) , o sea, es el nombre de individuos de la muestra que tienen la variable X igual a X_i y la variable Y igual a Y_j .
- Toda esta información se puede resumir en la siguiente tabla de frecuencias absolutas o tabla de contingencia:

$X \backslash Y$	Y_1	Y_2	\dots	Y_j	\dots	Y_J	$n_{i\bullet}$
X_1	n_{11}	n_{12}	\dots	n_{1j}	\dots	n_{1J}	$n_{1\bullet}$
X_2	n_{21}	n_{22}	\dots	n_{2j}	\dots	n_{2J}	$n_{2\bullet}$
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
X_i	n_{i1}	n_{i2}	\dots	n_{ij}	\dots	n_{iJ}	$n_{i\bullet}$
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
X_I	n_{I1}	n_{I2}	\dots	n_{Ij}	\dots	n_{IJ}	$n_{I\bullet}$
$n_{\bullet j}$	$n_{\bullet 1}$	$n_{\bullet 2}$	\dots	$n_{\bullet j}$	\dots	$n_{\bullet J}$	$N = n_{\bullet\bullet}$

- En la tabla anterior, los valores de $n_{i\bullet}$ representan el número de individuos con $X = X_i$
- Mientras que $n_{\bullet j}$ el nombre de individuos con $Y = Y_j$ y n es el nombre total de individuos.

Ejemplo

Consideremos la siguiente muestra de tamaño 12 de dos características conjuntas; la edad y peso de unas personas:

(20, 75)	(20, 75)	(30, 75)	(40, 85)
(30, 65)	(20, 75)	(40, 85)	(30, 65)
(20, 65)	(40, 75)	(30, 65)	(20, 75)

La variable X es “edad” y toma los valores $\{20, 30, 40\}$ y la variable Y es “peso” y toma los valores $\{65, 75, 85\}$. La tabla de frecuencias será:

$X \backslash Y$	65	75	85	
20	1	4	0	5
30	3	1	0	4
40	0	1	2	3
	4	6	2	12

En el caso en que las variables X y Y estén su tabla de frecuencias conjunta o de contingencia es:

$X \setminus Y$	intervalos	$[L'_0, L'_1) \cdots [L'_{j-1}, L'_j) \cdots [L'_{J-1}, L'_J)$					
intervalos	M. Clase	Y_1	\cdots	Y_j	\cdots	Y_J	$n_{i\bullet}$
$[L_0, L_1)$	X_1	n_{11}	\cdots	n_{1j}	\cdots	n_{1J}	$n_{1\bullet}$
$[L_1, L_2)$	X_2	n_{21}	\cdots	n_{2j}	\cdots	n_{2J}	$n_{2\bullet}$
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
$[L_{i-1}, L_i)$	X_i	n_{i1}	\cdots	n_{ij}	\cdots	n_{iJ}	$n_{i\bullet}$
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
$[L_{I-1}, L_I)$	X_I	n_{I1}	\cdots	n_{Ij}	\cdots	n_{IJ}	$n_{I\bullet}$
	$n_{\bullet j}$	$n_{\bullet 1}$	\cdots	$n_{\bullet j}$	\cdots	$n_{\bullet J}$	n

En la tabla anterior las X_i son las marcas de clase correspondientes a los intervalos de la variable X y las Y_j son las marcas de clase correspondientes a los intervalos de la variable Y .

Ejemplo

Consideremos la siguiente tabla que nos da el peso y la estatura de 15 individuos:

Individuo	X=peso	Y=estatura
1	65	1.6
2	62	1.6
3	71	1.6
4	72	1.7
5	75	1.8
6	80	1.6
7	74	1.6
8	77	1.7
9	81	1.8
10	90	1.8
11	89	1.7
12	83	1.8
13	82	1.8
14	81	1.7
15	71	1.7

- Tomamos intervalos de amplitud 10 para la variable X =peso. Así los intervalos para X empiezan en el límite real del mínimo peso 62:

$[61.5, 71.5), [71.5, 81.5), [81.5, 91.5).$

- Tomamos intervalos de amplitud 0.1 para la variable Y =talla. Los intervalos para Y empiezan en el límite real de la mínimo altura 1.6.

$[1.55, 1.65), [1.65, 1.75), [1.75, 1.85).$

La tabla de frecuencias agrupadas conjunta es:

$X \backslash Y$	intervalos	[1.55, 1.65)	[1.65, 1.75)	[1.75, 1.85)	
intervalos	M. Clase	1.6	1.7	1.8	$n_{i\bullet}$
[61.5, 71.5)	66.5	3	1	0	4
[71.5, 81.5)	76.5	2	3	2	7
[81.5, 91.5)	86.5	0	1	3	4
	$n_{\bullet j}$	5	5	5	15

Distribuciones marginales

- A la distribución unidimensional de la variable X la llamaremos distribución marginal de X y es la que toma los valores $\{X_1, X_2, \dots, X_I\}$,
- La que la frecuencia absoluta correspondiente a X_i es $n_{i\bullet} = \sum_{j=1}^J n_{ij}$.
- Es decir, la frecuencia absoluta del valor X_i es el número total de individuos que tienen la variable $X = X_i$.
- De la misma forma, la distribución marginal de Y es aquella variable unidimensional que toma los valores $\{Y_1, Y_2, \dots, Y_J\}$.
- La frecuencia absoluta correspondiente al valor Y_j vale $n_{\bullet j} = \sum_{i=1}^I n_{ij}$, lo que corresponde al número total de individuos observados que tienen la variable $Y = Y_j$.

Las tablas de frecuencias correspondientes a las distribuciones marginales son :

Distribución
marginal de la
variable X

X_i	$n_{i\bullet}$
X_1	$n_{1\bullet}$
X_2	$n_{2\bullet}$
\vdots	\vdots
X_i	$n_{i\bullet}$
\vdots	\vdots
X_I	$n_{I\bullet}$
	n

Distribución
marginal de la
variable Y

Y_j	$n_{\bullet j}$
Y_1	$n_{\bullet 1}$
Y_2	$n_{\bullet 1}$
\vdots	\vdots
Y_i	$n_{\bullet j}$
\vdots	\vdots
Y_I	$n_{\bullet J}$
	n

Ejemplo

Consideremos una distribución conjunta (X, Y) con tabla de frecuencias:

$X \backslash Y$	65	75	85	
20	1	4	0	5
30	3	1	0	4
40	0	1	2	3
	4	6	2	12

Las distribuciones marginales de X e Y son:

Distribución marginal de X		Distribución marginal de Y	
X_i	$n_{i\bullet}$	Y_j	$n_{\bullet j}$
20	5	65	4
30	4	75	6
40	3	85	2
12		12	

Ejemplo

Consideremos una distribución conjunta (X, Y) en este caso de valores agrupados con tabla de frecuencias:

$X \backslash Y$	intervalos	$[1.55, 1.65)$	$[1.65, 1.75)$	$[1.75, 1.85)$	
intervalos	M. Clase	1.6	1.7	1.8	$n_{i\bullet}$
$[61.5, 71.5)$	66.5	3	1	0	4
$[71.5, 81.5)$	76.5	2	3	2	7
$[81.5, 91.5)$	86.5	0	1	3	4
	$n_{\bullet j}$	5	5	5	15

Les distribuciones marginales de X e Y son:

Distribución marginal de X			Distribución marginal de Y		
Intervalo	X_i	$n_{i\bullet}$	Intervalo	Y_j	$n_{\bullet j}$
[61.5, 71.5)	66.5	4	[1.55, 1.65)	1.6	5
[71.5, 81.5)	76.5	7	[1.65, 1.75)	1.7	5
[81.5, 91.5)	86.5	4	[1.75, 1.85)	1.8	5
		15			15

Distribuciones condicionadas

- Consideremos una distribución conjunta de variables (X, Y) donde X toma valores $\{X_1, X_2, \dots, X_I\}$, e Y toma valores $\{Y_1, Y_2, \dots, Y_J\}$
- Las frecuencias conjuntas son n_{ij} .
- Consideremos un valor concreto de la variable Y , Y_j .
- Definimos la distribución condicionada de X respecto al valor Y_j de Y y lo denotaremos por $X/Y = Y_j$ como aquella distribución unidimensional que toma los mismo valores que X , es decir, $\{X_1, X_2, \dots, X_I\}$, y tal que la frecuencia absoluta del valor X_i (a la que denotaremos por $n_{i/j}$) se define como el número de individuos observados que tienen $X = X_i$ e $Y = Y_j$.
- De la misma manera, podemos considerar un valor concreto de la variable X , X_i .

Distribuciones condicionadas

- Definimos distribución condicionada de Y respecto del valor X_i y la denotaremos por $Y/X = X_i$ como aquella distribución unidimensional que toma los mismo valores que Y , $\{Y_1, Y_2, \dots, Y_J\}$,
- De forma que la frecuencia absoluta del valor Y_j (a la que denotaremos por $n_{j/i}$) se define como el número de individuos observados que tienen la $Y = Y_j$ y la $X = X_i$.
- Observemos que existen tantas distribuciones condicionadas $X/Y = Y_j$ como valores distintos toma Y y que existen tantas condicionales $Y/X = X_i$ como valores distintos toma X .

Ejemplo

Consideremos una distribución conjunta (X, Y) sin agrupar, con tabla de frecuencias:

$X \backslash Y$	65	75	85	
20	1	4	0	5
30	3	1	0	4
40	0	1	2	3
	4	6	2	12

Fijemos $Y = 75$. La tabla de frecuencias de la distribución $X/Y = 75$ es:

$X_i/Y = 75$	$n_{i/75}$
20	4
30	1
40	1
	6

Fijemos por ejemplo $X = 30$. La tabla de frecuencias de la distribución $Y/X = 30$ es:

$Y_j/X = 30$	$n_{j/30}$
65	3
75	1
85	0
	4

Ejemplo

Consideremos una distribución conjunta (X, Y) , caso agrupado y con tabla de contingencia:

$X \backslash Y$	intervalos	$[1.55, 1.65)$	$[1.65, 1.75)$	$[1.75, 1.85)$	
intervalos	M. Clase	1.6	1.7	1.8	$n_{i\bullet}$
$[61.5, 71.5)$	66.5	3	1	0	4
$[71.5, 81.5)$	76.5	2	3	2	7
$[81.5, 91.5)$	86.5	0	1	3	4
	$n_{\bullet j}$	5	5	5	15

Ejemplo

Fijemos por ejemplo $Y = 1.6$. La tabla de frecuencias de $X/Y = 1.6$ es:

Intervalo	X_i	$n_{i/1.6}$
[61.5, 71.5)	66.5	3
[71.5, 81.5)	76.5	2
[81.5, 91.5)	86.5	0
		5

Fijemos por ejemplo $X = 86.5$. La tabla de frecuencias de $Y/X = 86.5$ es:

Intervalo	Y_j	$n_{j/86.5}$
$[1.55, 1.65)$	1.6	0
$[1.65, 1.75)$	1.7	1
$[1.75, 1.85)$	1.8	3
		4

Estadísticos descriptivos bidimensionales

- Consideremos una distribución conjunta de las variables (X, Y) donde X toma valores $\{X_1, X_2, \dots, X_I\}$, mientras que Y toma los valores $\{Y_1, Y_2, \dots, Y_J\}$ con la correspondiente tabla de frecuencias conjunta n_{ij} .
- Vamos a estudiar los estadísticos de tendencia central y de dispersión. Los estadísticos de tendencia central la media de X (\bar{x}) y la media de Y (\bar{y}).

- Se calculan de la siguiente forma $\bar{x} = \frac{\sum_{i=1}^I n_{i\bullet} X_i}{n}$, $\bar{y} = \frac{\sum_{j=1}^J n_{\bullet j} Y_j}{n}$.
- Los estadísticos de dispersión son la varianza de X (s_X^2), la varianza de Y (s_Y^2) y la **covarianza** de X e Y s_{XY} , que mide la variación conjunta.

• Las fórmulas respectivas son:

$$\blacktriangleright s_X^2 = \frac{1}{n} \sum_{i=1}^I n_{i\bullet} (X_i - \bar{x})^2 = \frac{1}{n} \sum_{i=1}^I n_{i\bullet} X_i^2 - \bar{x}^2.$$

$$\blacktriangleright s_Y^2 = \frac{1}{n} \sum_{j=1}^J n_{\bullet j} (Y_j - \bar{y})^2 = \frac{1}{n} \sum_{j=1}^J n_{\bullet j} Y_j^2 - \bar{y}^2.$$

$$\blacktriangleright s_{XY} = \frac{\sum_{i=1}^I \sum_{j=1}^J n_{ij} (X_i - \bar{x})(Y_j - \bar{y})}{n} = \frac{\sum_{i=1}^I \sum_{j=1}^J n_{ij} X_i Y_j}{n} - \bar{x} \cdot \bar{y}.$$

► También tenemos las cuasivarianzas $\tilde{s}_X^2 = \frac{n}{n-1} \cdot s_X^2$, $\tilde{s}_Y^2 = \frac{n}{n-1} \cdot s_Y^2$ y la **cuasicovarianza** $\tilde{s}_{XY} = \frac{n}{n-1} \cdot s_{XY}$.

► Una propiedad evidente es que $s_{XY} = s_{YX}$.

Ejemplo

Consideremos la siguiente distribución de las variables (X, Y) :

$X \backslash Y$	65	75	85	
20	1	4	0	5
30	3	1	0	4
40	0	1	2	3
	4	6	2	12

Los momentos de primer orden son:

- $\bar{X} = \frac{n_{1\bullet}X_1 + n_{2\bullet}X_2 + n_{3\bullet}X_3}{n} = \frac{5 \cdot 20 + 4 \cdot 30 + 3 \cdot 40}{12} = 28.333.$
- $\bar{Y} = \frac{n_{\bullet 1}Y_1 + n_{\bullet 2}Y_2 + n_{\bullet 3}Y_3}{n} = \frac{4 \cdot 65 + 6 \cdot 75 + 2 \cdot 85}{12} = 73.333.$

Ejemplo

Las medidas de dispersión son

- $s_X^2 = \frac{n_{1\bullet}X_1^2 + n_{2\bullet}X_2^2 + n_{3\bullet}X_3^2}{n} - \bar{X}^2 = \frac{5 \cdot 20^2 + 4 \cdot 20^2 + 3 \cdot 40^2}{12} - 28.333^2 = 63.888.$
- $s_Y^2 = \frac{n_{\bullet 1}Y_1^2 + n_{\bullet 2}Y_2^2 + n_{\bullet 3}Y_3^2}{n} - \bar{Y}^2 = \frac{4 \cdot 65^2 + 6 \cdot 75^2 + 2 \cdot 85^2}{12} - 73.333^2 = 47.222.$
- La covarianza es

$$\begin{aligned}s_{XY} &= \frac{1}{n} (n_{11}X_1Y_1 + n_{12}X_1Y_2 + n_{13}X_1Y_3 + n_{21}X_2Y_1 + n_{22}X_2Y_2 \\ &\quad + n_{23}X_2Y_3 + n_{31}X_3Y_1 + n_{32}X_3Y_2 + n_{33}X_3Y_3) - \bar{X} \cdot \bar{Y} \\ &= \frac{1}{12} (1 \cdot 20 \cdot 65 + 4 \cdot 20 \cdot 75 + 0 \cdot 20 \cdot 85 + 3 \cdot 30 \cdot 65 \\ &\quad + 1 \cdot 30 \cdot 75 + 0 \cdot 30 \cdot 85 + 0 \cdot 40 \cdot 65 + 1 \cdot 40 \cdot 75 + \\ &\quad 2 \cdot 40 \cdot 85) - 28.333 \cdot 73.333 = 22.222.\end{aligned}$$

Ejemplo

Consideremos una distribución conjunta (X, Y) (datos agrupados) con tabla de frecuencias:

X/Y	intervalos	$[1.55, 1.65)$	$[1.65, 1.75)$	$[1.75, 1.85)$	
intervalos	M. Clase	1.6	1.7	1.8	$n_{i\bullet}$
$[61.5, 71.5)$	66.5	3	1	0	4
$[71.5, 81.5)$	76.5	2	3	2	7
$[81.5, 91.5)$	86.5	0	1	3	4
	$n_{\bullet j}$	5	5	5	15

- Las medias son $\bar{x} = 76.5$, $\bar{y} = 1.7$.
- Las varianzas son $s_X^2 = 53.333$, $s_Y^2 = 0.006$, $s_{XY} = 0.4$.

Independencia e incorrelación

- Vamos a introducir dos conceptos nuevos: el de independencia y el de incorrelación.
- El concepto de independencia formaliza la idea conocer el valor de la variable X no aporta información alguna sobre el valor de Y y viceversa.
- Dada una variable bidimensional (X, Y) con tabla de frecuencias conjunta n_{ij} , diremos que X e Y son independientes si:

$$\frac{n_{ij}}{n} = \frac{n_{i\bullet}}{n} \frac{n_{\bullet j}}{n}, \text{ para todo } i \text{ y para todo } j.$$

- En el caso en que la relación anterior falle para un i y un j diremos que las dos variable no son independientes.

Ejemplo

En este ejemplo las variables X e Y no son independientes:

$X \backslash Y$	65	75	85	
20	1	4	0	5
30	3	1	0	4
40	0	1	2	3
	4	6	2	12

ya que por ejemplo

$$\frac{n_{11}}{n} \neq \frac{n_{1\bullet}}{n} \frac{n_{\bullet 1}}{n}, \quad \frac{1}{12} \neq \frac{5}{12} \cdot \frac{4}{12}.$$

En cambio en este otro caso, sí son independientes:

$X \backslash Y$	65	75	85	
20	3	2	1	6
30	6	4	2	12
40	6	4	2	12
	15	10	5	30

Dejamos al lector la comprobación como ejercicio.

Correlación

- El concepto de incorrelación formaliza la idea de relación lineal en el sentido de que las variables crecen de forma lineal conjuntamente (relación directa) o bien si una crece, la otra decrece (relación inversa).
- Dada una variable bidimensional (X, Y) con tabla de frecuencias conjunta n_{ij} , diremos que X e Y son incorreladas si su covarianza $s_{XY} = 0$.
- La relación que existe entre los dos conceptos introducidos, el de independencia y el de incorrelación viene dada por la siguiente propiedad:

Relación entre independencia y correlación

Theorem

Si las variables X e Y son independientes entonces son incorreladas.

- El recíproco del teorema anterior no es cierto en general. Podemos decir que independencia implica incorrelación pero lo contrario no es cierto en general.

Demostración del teorema:

- Si X es independiente de Y tenemos que

$$\frac{n_{ij}}{n} = \frac{n_{i\bullet}}{n} \frac{n_{\bullet j}}{n}.$$

- Por lo tanto:

$$\begin{aligned} s_{XY} &= \frac{1}{n} \sum_{i=1}^I \sum_{j=1}^J n_{ij} (x_i - \bar{x}) (y_j - \bar{y}) \\ &= \sum_{i=1}^I (x_i - \bar{x}) \frac{n_{i\bullet}}{n} \sum_{j=1}^J (y_j - \bar{y}) \frac{n_{\bullet j}}{n} \\ &= 0 \cdot 0 = 0, \end{aligned}$$

- Ahora teniendo en cuenta que si X es una variable unidimensional con valores $\{x_1, x_2, \dots, x_I\}$, con las correspondientes frecuencias absolutas $\{n_1, n_2, \dots, n_I\}$, tenemos:

$$\sum_{i=1}^I n_i (x_i - \bar{x}) = \sum_{i=1}^I n_i x_i - n\bar{x} = 0.$$

Introducción a las medidas de asociación

- En esta sección estudiaremos si existe algún tipo de relación entre dos variables X e Y .
- Hasta ahora sabemos cuando dos variables son independientes o no.
- En caso de que no se sean independientes, nos interesará medir el grado de dependencia que tienen, es decir, si son “muy dependientes o no”.
- Para medir la dependencia utilizaremos una serie de coeficientes como son el **coeficiente de contingencia de Pearson** y en el caso de variables con sólo dos valores el **coeficiente de contingencia de Yule**.

Coeficiente de contingencia o de correlación de Pearson

- Consideremos una variable bidimensional (X, Y) con tabla de frecuencias conjunta n_{ij} ,
- definimos el coeficiente de correlación de Pearson como:
$$C_P = \sqrt{\frac{\chi^2}{n + \chi^2}}.$$
- Donde χ^2 es el llamado estadístico de Pearson, que se define como:

$$\chi^2 = \sum_{i=1}^I \sum_{j=1}^J \frac{\left(n_{ij} - \frac{n_{i\bullet} n_{\bullet j}}{n}\right)^2}{\frac{n_{i\bullet} n_{\bullet j}}{n}}.$$

- Desde este punto de vista a las n_{ij} se las denomina como frecuencias empíricas u observadas.
- Mientras que a las $\frac{n_{i\bullet} \cdot n_{\bullet j}}{n}$ se las denomina frecuencias teóricas; o sea, las frecuencias conjuntas que tendrían las variables (X, Y) si fueran independientes.
- Por lo tanto, cuando más cerca estén las frecuencias empíricas de las teóricas, más pequeño será el estadístico de Pearson χ^2 y el coeficiente de contingencia de Pearson C_P .

Ejemplo

Consideremos la siguiente distribución conjunta:

X/Y	65	75	85	
20	1	4	0	5
30	3	1	0	4
40	0	1	2	3
	4	6	2	12

La tabla anterior es la de frecuencias empíricas.

Ahora construimos la tabla de frecuencias teóricas:

X/Y	65	75	85	
20	1.66	2.5	0.83	5
30	1.33	2	0.66	4
40	1	1.5	0.5	3
	4	6	2	12

Como podemos observar, las frecuencias teóricas no coinciden con las empíricas. Por lo tanto, deducimos que las dos variables X e Y no son independientes.

Vamos a calcular ahora el coeficiente de contingencia de Pearson C_P . En primer lugar hemos de calcular el estadístico χ^2 :

$$\begin{aligned}\chi^2 &= \frac{(1-1.66)^2}{1.66} + \frac{(4-2.5)^2}{2.5} + \frac{(0-0.83)^2}{0.83} + \frac{(3-1.33)^2}{1.33} + \frac{(1-2)^2}{2} + \\ &\quad \frac{(0-0.66)^2}{0.66} + \frac{(0-1)^2}{1} + \frac{(1-1.5)^2}{1.5} + \frac{(2-0.5)^2}{0.5} \\ &= 10.916\end{aligned}$$

Por último calculamos el coeficiente de contingencia de Pearson C_P :

$$C_P = \sqrt{\frac{10.916}{12 + 10.916}} = 0.690$$

Propiedades de C_p

El coeficiente de asociación C_p cumple las siguientes propiedades:

- 1) El valor de C_p es mayor o igual que 0 y menor que 1. En el caso en que X e Y sean independientes, las frecuencias empíricas y teóricas coinciden y $C_p = 0$.
- 2) Cuanto más dependientes son las variables X e Y , C_p se aproxima más a 1.

Por lo tanto si C_p es pequeño podemos decir que el grado de dependencia es bajo, mientras que si C_p aumenta es alto.

Tablas 2×2

- Para el caso más trivial en el que tengamos una tabla 2×2 , es decir $I = J = 2$, es decir cuando las variables sólo toman dos valores cada una.
- Se utiliza otro coeficiente, el coeficiente de contingencia de Yule. Se define así:

$$C_{\gamma} = \frac{n_{11}n_{22} - n_{12}n_{21}}{n_{11}n_{22} + n_{12}n_{21}}.$$

Recordamos al lector que la tabla de frecuencias tendrá el siguiente aspecto:

X/Y	Y_1	Y_2	
X_1	n_{11}	n_{12}	$n_{1\bullet}$
X_2	n_{21}	n_{22}	$n_{2\bullet}$
	$n_{\bullet 1}$	$n_{\bullet 2}$	n

Ejemplo

Calculemos el coeficiente de contingencia de Yule de la siguiente distribución conjunta:

X/Y	Y_1	Y_2	
X_1	2	8	10
X_2	3	7	10
	5	15	20

$$C_\gamma = \frac{2 \cdot 7 - 3 \cdot 8}{2 \cdot 7 + 3 \cdot 8} = -0.263158$$

El coeficiente de contingencia de Yule siempre está entre -1 y 1 . En caso de independencia entre las variables, se tiene que $C_\gamma = 0$.

Correlación Lineal

El problema que nos planteamos en este punto es saber si existe una relación lineal entre X e Y , es decir, si existen dos valores numéricos a y b tales que:

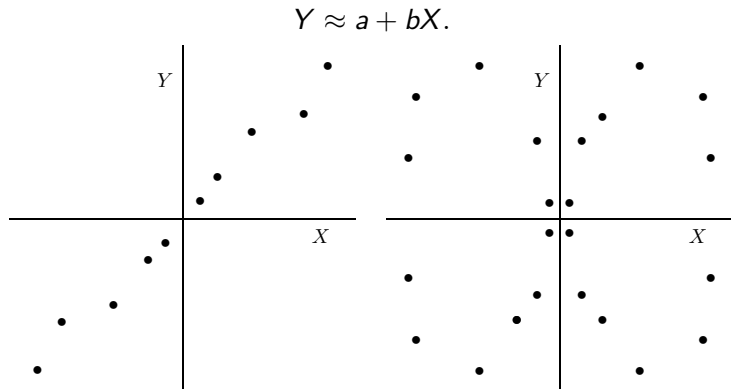


Figura: Nube de puntos con relación lineal y sin relación lineal

- La relación lineal no tiene por que ser perfecta. Lo que nos interesa es medir esa relación lineal.
- En el gráfico de la izquierda de la figura 1 se vislumbra una relación lineal mayor que en el de la derecha ya que podemos encontrar una recta que aproxime mejor Y en función de X .
- Vamos a introducir un coeficiente que mide la relación lineal entre dos variables. Este coeficiente es el coeficiente de correlación lineal de Pearson r_{XY} y se define de la manera siguiente

$$r_{XY} = \frac{s_{XY}}{\sqrt{s_X^2 s_Y^2}} = \frac{s_{XY}}{s_X s_Y}.$$

Propiedades del coeficiente de correlación lineal

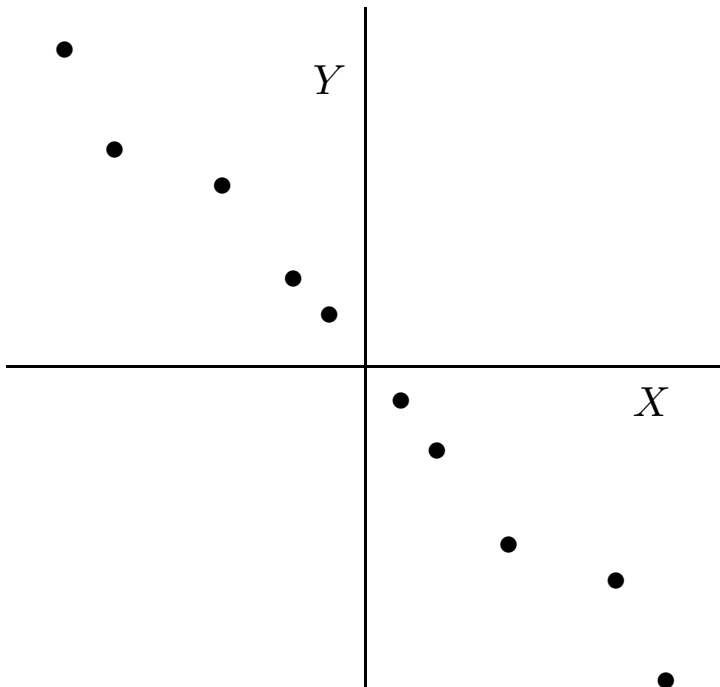
1) $-1 \leq r_{XY} \leq 1$

2) $r_{XY} = r_{YX}$

interpretación del coeficiente de correlación lineal

Interpretación de r_{XY} :

- $r_{XY} > 0$ y a medida que se aproxima a 1, aumenta la relación lineal positiva entre las dos variables X e Y ; lo que quiere decir que si X crece, la variable Y también y si X decrece, Y también, Obsérvese la parte izquierda de la figura 1 como ejemplo de este caso.
- $r_{XY} < 0$ y su valor está muy cerca de -1 quiere decir que hay una buena relación lineal negativa entre las dos variables X e Y ; lo que significa que si la variable X crece, la variable Y decrece o viceversa. Como ejemplo ver el gráfico de la derecha de la figura anterior.



- Si $r_{XY} = 0$ o es pequeño, quiere decir que no hay ningún tipo de relación lineal entre las variables X e Y .
- Si $r_{XY} = \pm 1$, hay relación lineal exacta entre X e Y , o sea, existen dos números reales a y b tales que $Y = a + bX$.

Ejemplo

Consideremos la siguiente distribución conjunta de la variable (X, Y) :
Los valores de s_X^2 , s_Y^2 y de s_{XY} son:

$$s_X^2 = 63.888, \quad s_Y^2 = 47.222, \quad s_{XY} = 22.222$$

El coeficiente de correlación lineal vale:

$$r_{XY} = \frac{22.222}{\sqrt{63.888 \cdot 47.222}} = 0.405$$

Correlación ordinal

- Vamos a estudiar ahora la relación que existe entre dos ordenaciones dadas por una muestra de datos bidimensionales.
- Los estadísticos que miden este tipo de relaciones reciben el nombre de coeficiente de correlación ordinal y nos darán medidas de la similitud de las dos ordenaciones a lo que se suele llamar concordancia.
- Más concretamente, consideremos un conjunto de individuos y los ordenamos según dos criterios.

- Tendremos así dos ordenaciones de los individuos.
- Estas ordenaciones las podemos disponer como si se tratara de una estadística bidimensional, donde la primera componente de la observación de un individuo correspondería al número de orden del primer criterio de ordenación y la segunda componente al otro.

Ejemplo

Por ejemplo consideremos las observaciones en 5 humanos de su peso X en Kg. y estatura Y en metros:

Individuo i	(X_i, Y_i)	Orden X	Orden Y
Individuo 1	(80, 1.75)	3	2
Individuo 2	(75, 1.92)	2	4
Individuo 3	(85, 1.67)	4	1
Individuo 4	(66, 1.80)	1	3
Individuo 5	(90, 2.00)	5	5

Si ordenamos los individuos en orden ascendente (de menor a mayor) según el peso quedan así:

Rango	Peso	1	2	3	4	5
Individuo		4	2	1	3	5

mientras que si los ordenamos en orden ascendente según su altura:

Rango	1	2	3	4	5
Individuo	3	1	4	2	5

- Tenemos así dos ordenaciones de números ordinales enteros que reciben el nombre de rangos
- El cálculo de rangos se complica en el caso de empates, es decir cuando hay valores repetidos en las series de datos. En estos casos se puede romper el empate de varias maneras.
- En general podemos escribir:

$$\begin{aligned} \text{para } X &\rightarrow \{r_{x_1}, r_{x_2}, r_{x_3}, \dots, r_{x_n}\}, \\ \text{para } Y &\rightarrow \{r_{y_1}, r_{y_2}, r_{y_3}, \dots, r_{y_n}\}, \end{aligned}$$

- Donde los valores de r_{x_i} y r_{y_i} dan el lugar que ocupa el valor x_i o el y_i en cada una de las muestras ordenadas.
- Estos valores están comprendidos entre 1 y n , luego son dos permutaciones de orden n .
- Las diferencias entre las ordenaciones son $d_i = r_{x_i} - r_{y_i}; i = 1, 2, \dots, n$.
- El coeficiente de correlación ordinal o por rangos de Spearman queda definido por:

$$r_S = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)}.$$

- De hecho, r_S no es más que el coeficiente de correlación lineal introducido en la sección anterior aplicado a los rangos.

Propiedades de r_S :

El coeficiente de correlación lineal r_S cumple las propiedades siguientes:

- Si $r_S = 1$, las dos ordenaciones coinciden; o sea, $r_{x_i} = r_{y_i}$ para cualquier i entre 1 y n .
- Si $r_S = -1$, la ordenación de Y es exactamente la opuesta a la de X , es decir, $r_{x_i} = r_{y_{n-i+1}}$ para cualquier i entre 1 y n .
- El coeficiente r_S está siempre comprendido entre -1 y 1. Si $r_S > 0$, podemos decir que las dos ordenaciones son del mismo sentido y si $r_S < 0$, las dos ordenaciones son de sentidos opuestos.

Ejemplo

Consideremos la muestra anterior de pesos y estaturas de 5 individuos:

Individuo i	(X_i, Y_i)	r_{X_i}	r_{Y_i}	d_i^2
Individuo 1	(80, 1.75)	3	2	1
Individuo 2	(75, 1.92)	2	4	4
Individuo 3	(85, 1.67)	4	1	9
Individuo 4	(66, 1.80)	1	3	4
Individuo 5	(90, 2.00)	5	5	0
Σ				18

Luego tenemos que :

$$r_s = 1 - \frac{6 \cdot 18}{5 \cdot (25 - 1)} = 1 - \frac{108}{120} = 0.1$$

Medias armónica y geométrica

- Las medias armónica y geométrica no son de gran utilidad salvo en problemas concretos. Se calculan de la siguiente forma:
- Media Armónica: $M_h = \frac{n}{\sum_{i=1}^n \frac{1}{x_i}} = \frac{n}{\sum_{j=1}^I \frac{n_j}{\bar{x}_j}}$.
- Media geométrica: $M_g = \sqrt[n]{\prod_{i=1}^n x_j} = \sqrt[n]{\prod_{j=1}^I x_j^{n_j}}$.
- Estas medias tienen restricciones sobre los datos, no pueden tener datos nulos, y en general se utilizan para datos positivos.

Media general de orden n

- Definimos la media general $M_{(m)}$ de orden m como

$$M_{(m)} = \left(\frac{\sum_{i=1}^n n_i x_i^m}{n} \right)^{\frac{1}{m}} = \left(\frac{\sum_{j=1}^J n_j X_j^m}{n} \right)^{\frac{1}{m}}.$$

- Se cumple que $M_{(-1)} = M_h$; $M_{(0)} = M_g$; $M_{(1)} = \bar{x}$.
- Además se cumple que $M_{(m)}$ es una función creciente en m y por lo tanto : $M_h \leq M_g \leq \bar{x}$.

Moda

- Hay algunos algoritmos para aproximar la moda para datos agrupados.
- Si los intervalos tienen la misma amplitud, podemos aproximar la moda de la siguiente manera-
- En primer lugar localizamos el intervalo con frecuencia absoluta más alta.
- Sean $[L_c, L_{c+1})$ los extremos del intervalo con frecuencia absoluta máxima.

- Para calcular la moda podemos utilizar la siguiente fórmula, en la que suponemos que todos los intervalos tienen la misma amplitud A (en caso contrario se utilizan otras aproximaciones):

$$M_o = L_c + A \frac{n_{c+1}}{(n_{c-1} + n_{c+1})}.$$

- Donde:

- ▶ A : amplitud de los intervalos
- ▶ n_{c-1} : frecuencia absoluta del intervalo anterior al de frecuencia máxima.
- ▶ n_{c+1} : frecuencia absoluta del intervalo posterior al de frecuencia máxima.

Ejemplo

Consideremos la siguiente distribución de frecuencias:

intervalos	X_j	n_j	N_j
[1.5, 4.5)	3	3	3
[4.5, 7.5)	6	12	15
[7.5, 10.5)	9	5	20
[10.5, 13.5)	12	4	24

El intervalo con la frecuencia absoluta mas alta es el $[4.5, 7.5)$. Por lo tanto, la moda vale:

$$M_0 = 4.5 + 3 \frac{5}{(3 + 5)} = 6.375.$$

Desviación media

- La desviación media es un índice de dispersión respecto a la mediana o a la media. Queda definido por: $D_M = \frac{1}{n} \sum_{j=1}^J n_j |X_j - M|$.
- Donde M es la mediana o la media aritmética.
- La propiedad fundamental de la desviación media respecto a la mediana es que minimiza las desviaciones en valor absoluto respecto de un punto cualquiera X_0 . Es decir $\min_{X_0} \frac{1}{n} \sum_{j=1}^J n_j |X_j - X_0| = D_M$.

- Otra medida de dispersión es el recorrido. Se define como la diferencia entre el valor máximo y mínimo de los valores observados.
- Consideremos la siguiente distribución de frecuencias:

intervalos	X_j	n_j	$n_j X_j$
[0.5, 15.5)	8	4	32
[15.5, 30.5)	23	4	92
[30.5, 45.5)	38	2	76
Sumas		10	200

- Vamos a calcular la desviación media:

- Calculamos la media $\bar{x} = \frac{200}{10} = 20$.
- Añadimos dos columnas más a la tabla de frecuencias:

X_j	$ X_j - \bar{x} $	$n_j X_j - \bar{x} $
8	12	48
23	3	12
38	18	36
Sumas		96

- La desviación media es $D_M = \frac{96}{10} = 9.6$.

También se utiliza el recorrido intercuartílico que es $Q_{0.75} - Q_{0.25}$; la diferencia entre el tercer y primer cuartil. También se pueden calcular recorridos con deciles, percentiles y cuantiles en general.

Apuntes de Bioestadística.

R. Alberich y A. Mir

Departamento de Matemáticas e
Informàtica
Universitat Illes Balears

14 de julio de 2010

2 Sobre la estadística y el método científico

- La reproductibilidad en las investigaciones científicas
- Análisis Exploratorio de Datos

3 Los datos

- Datos de tipo atributo o cualitativo
- Datos ordinales
- Datos cuantitativos

4 Datos agrupados

- Descripción gráfica datos agrupados
- Cambios de escala y de origen

5 Más estadísticos

- Coeficiente de variación
- Medidas de simetría
- Medidas de apuntamiento
- Cambios de escala y origen

6 Variables multidimensionales

- Descripción numérica: caso bidimensional
- Distribuciones marginales
- Distribuciones condicionadas
- Estadísticos descriptivos bidimensionales

Variables Aleatorias

- Consideremos un experimento aleatorio y con espacio muestral Ω .
- Una **variable aleatoria** (v.a) es una aplicación (que cumple algunas propiedades adicionales)

$$X : \Omega \rightarrow \mathbb{R}.$$

- Es decir es una aplicación que transforma sucesos elementales del espacio muestral en un número real.

Variables aleatorias discretas y continuas

- Llamaremos **dominio** de una variable aleatoria X al conjunto $D_X = X(\Omega)$. Es decir el dominio de una v.a. es el conjunto de valores reales que asume.
- Diremos que una v.a. X es **discreta**, o más exactamente que tiene dominio discreto, si el conjunto D_X es finito o numerable. Lo denotaremos por $D_X = \{x_1, \dots, x_n, \dots\}$ (normalmente $x_1 < x_2 < \dots < x_n \dots$) si es infinito y por $D_X = \{x_1, \dots, x_n\}$ si es finito.
- Diremos que una variable es continua, o más exactamente que tiene dominio continuo, si D_X es uno o varios intervalos de la recta real.

Sucesos con variables aleatorias

Una vez tenemos una variable aleatoria podemos definir diferentes sucesos, estos suelen tener una notación peculiar. Veamos unos ejemplos:

- El suceso $\{X = x\}$, habitualmente se representa sin llaves: $P(X = x)$.
- El suceso $\{X \leq x\} = \{X \in (-\infty, x]\}$ habitualmente se representa sin llaves: $P(X \leq x)$.
- El suceso $\{a < X \leq b\} = \{X \in (a, b]\}$ habitualmente se representa sin llaves: $P(a < X \leq b)$.
- Y otros sucesos que se definen de forma similar.
- Para la operación unión de sucesos utilizaremos el símbolo convencional \cup . Por ejemplo el suceso que $X > 3$ o $X < -1$ lo escribiremos como $P(\{X > 3\} \cup \{X < -1\})$.
- En cambio para la intersección se suelen poner comas: $P(1 < X, X \leq 3)$ es $P(\{1 < X\} \cap \{X \leq 3\})$.

Función de distribución

- **Llamaremos función de distribución (acumulada) o función de probabilidad acumulada** de una v.a. X a la función $F : \mathbb{R} \rightarrow [0, 1]$ definida por

$$F(x) = P(X \in (-\infty, x]) = P(X \leq x)$$

Propiedades

- Sea F una función de distribución acumulada de la v.a. X entonces:
 - a) F es creciente.
 - b) F es continua por la derecha.
 - c) $\lim_{x \rightarrow \infty} F(x) = 1$; $\lim_{x \rightarrow -\infty} F(x) = 0$.
 - d) $0 \leq F(x) \leq 1$.
 - e) Toda función F verificando a), b) y c) es función de distribución de alguna v.a. X .

Propiedad

Sea F una función de distribución de la v.a. X entonces dados $a, b \in \mathbb{R}$ con $a < b$

- $P(X \geq a) = 1 - P(X \leq a) = 1 - F(a).$
- $P(X < a) = P(X \leq a) - P(X = a) = F(a) - P(X = a).$
- $P(a < X \leq b) = P(X \leq b) - P(X \leq a) = F(b) - F(a).$

Más propiedades

Si denotamos por $F(x_0^-) = \lim_{x \rightarrow x_0^-} F(x)$.

- a) $P(X = x) = P(X \leq x) - P(X < x) = F(x) - F(x^-)$.
- b) $P(X < a) = F(a^-)$.
- c) $P(X \geq a) = 1 - P(X < a) = 1 - F(a^-)$.
- d) $P(X > a) = 1 - F(a)$.
- e) $P(a \leq X \leq b) = F(b) - F(a^-)$.
- f) $P(a < X < b) = P(X < b) - P(X \leq a) = F(b^-) - F(a)$.
- g) $P(a \leq X < b) = P(X < b) - P(X < a) = F(b^-) - F(a^-)$.

Cuantiles de una distribución

- Llamaremos **cuantil** de orden p de la v.a. X al valor x_p , si existe, tal que $P(X \leq x_p) = F(x_p) = p$.
- En caso de que no exista ese valor, suele pasar con variables discretas, se toma como **cuantil** el valor más pequeño tal que $P(X \leq x_p) = F(x_p) \geq p$.
- Los cuantiles de las distribuciones son las versiones poblacionales de los cuantiles muestrales.
- En este tema, y en los que siguen, a medida que sea necesario iremos exponiendo el cálculo de los cuantiles de las distribuciones de probabilidad que necesitemos.

Función de probabilidad de una v.a. discreta

Sea X una v.a. discreta con $D_X = \{x_1, x_2, \dots, x_n, \dots\}$. Llamaremos ley de probabilidad, función de probabilidad o función de cuantía de la v.a. X a la función $f : \mathbb{R} \rightarrow \mathbb{R}$ definida por

$$f(x) = P(X = x),$$

para todo $x \in \mathbb{R}$.

Propiedad Sea X una v.a. discreta con función de probabilidad f entonces:

a) $\sum_{x \in D_X} f(x) = 1; f(x) = 0$ si $x \notin D_X$.

b)

$$P(X \leq a) = \sum_{\substack{x \in X(\Omega) \\ x \leq a}} f(x)$$

Esperanza y varianza para variables aleatorias discretas

Sea X una v.a. discreta con función de probabilidad f y dominio D_X

- Se define el valor esperado, esperanza o media como

$$E(X) = \sum_{x \in D_X} x \cdot f(x).$$

- Si $H : \mathbb{R} \rightarrow \mathbb{R}$ es una función se define:

$$E(H(X)) = \sum_{x \in D_X} H(x) \cdot f(x).$$

- Se define la varianza por

$$\text{Var}(X) = E((X - E(X))^2) = E(X^2) - E(X)^2.$$

- La desviación típica es $\sqrt{\text{Var}(X)}$.
- Habitualmente se utilizan las letras griegas μ , σ^2 y σ para representar la esperanza, la varianza y la desviación típica.

Propiedades

- a) $E(cte.) = cte$
- b) $E(aX + b) = aE(X) + b$
- c) Si $a < X < b$ entonces $a < E(X) < b$
- d) Si X es una v.a. no negativa entonces $E(X) \geq 0$.
- e) Si $g(X) \leq h(X)$ entonces $E(g(X)) \leq E(h(X))$
- f) $Var(aX + b) = a^2 Var(x)$.

Ejemplo

La tabla siguiente muestra la función de probabilidad asociada a la variable aleatoria X “número de batidos de ala por segundo en individuos de una especie de mariposa”

x	6	7	8	9	10
$f(x)$	0.05	0.1	0.6	0.15	?

¿Cuál es el valor de la entrada que falta?

$$\begin{aligned}1 &= f(6) + f(7) + f(8) + f(9) + f(10) \\&= 0.05 + 0.1 + 0.6 + 0.15 + f(10) \\&= 0.9 + f(10) \Rightarrow f(10) = 0.1\end{aligned}$$

Ejemplo

x	6	7	8	9	10
$f(x)$	0.05	0.1	0.6	0.15	0.1

¿Cual es la función de distribución acumulativa?

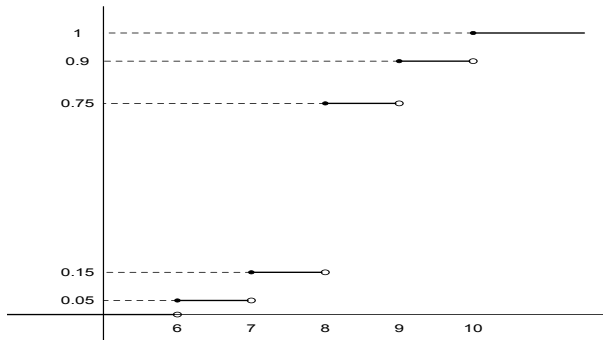
Se producen saltos en 6, 7, 8, 9, 10:

$$F(x) = \begin{cases} 0 & \text{si } x < 6 \\ 0.05 & \text{si } 6 \leq x < 7 \\ 0.15 & \text{si } 7 \leq x < 8 \\ 0.75 & \text{si } 8 \leq x < 9 \\ 0.9 & \text{si } 9 \leq x < 10 \\ 1 & \text{si } 10 \leq x \end{cases}$$

Ejemplo

x	6	7	8	9	10
$f(x)$	0.05	0.1	0.6	0.15	0.1

¿Cual es la distribución acumulada?

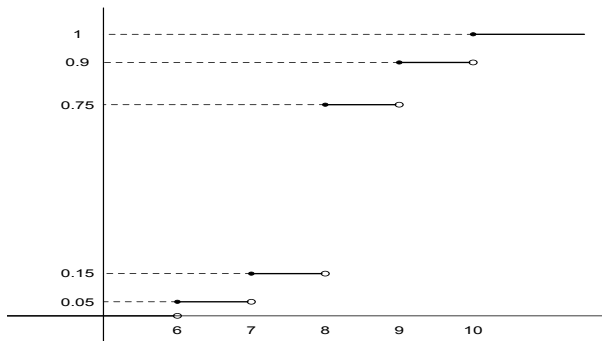


Ejemplo

Si nos dan la distribución $F : S \rightarrow [0, 1]$, con

$$D_X = \{6, 7, 8, 9, 10\},$$

¿cómo calcularías la función de probabilidad?



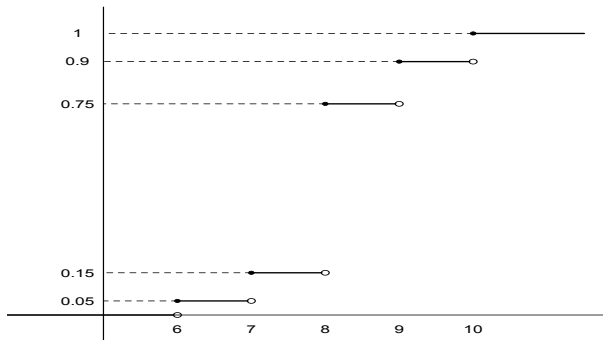
Ejemplo

$$D_X = \{6, 7, 8, 9, 10\}$$

$$f(6) = 0.05, \quad f(7) = 0.15 - 0.05 = 0.1$$

$$f(8) = 0.75 - 0.15 = 0.6,$$

$$f(9) = 0.9 - 0.75 = 0.15, \quad f(10) = 1 - 0.9 = 0.1$$



Ejemplo

La tabla siguiente nos da la función de probabilidad asociada a la variable aleatoria X “número de batidos de ala por segundo en individuos de una determinada especie de mariposas”

x	6	7	8	9	10
$f(x)$	0.05	0.1	0.6	0.15	0.1

¿Cuál es el valor esperado del número de batidas?

$$E(X) = 6 \cdot f(6) + 7 \cdot f(7) + 8 \cdot f(8) + 9 \cdot f(9) + 10 \cdot f(10) = 8.15$$

Ejemplo

La tabla siguiente nos da la función de probabilidad asociada a la variable aleatoria X “número de batidos de ala por segundo en individuos de una determinada especie de mariposas”

x	6	7	8	9	10
$f(x)$	0.05	0.1	0.6	0.15	0.1

¿Cuál es el valor esperado de X^2 ?

$$\begin{aligned} E(X^2) &= 6^2 \cdot f(6) + 7^2 \cdot f(7) + 8^2 \cdot f(8) + 9^2 \cdot f(9) \\ &\quad + 10^2 \cdot f(10) = 67.25 \end{aligned}$$

¡Alerta! $E(X^2) \neq E(X)^2$ ($67.25 \neq 8.15^2 = 66.4225$)

Ejemplo

La tabla siguiente nos da la función de probabilidad asociada a la variable aleatoria X “número de batidos de ala por segundo en individuos de una determinada especie de mariposas”

x	6	7	8	9	10
$f(x)$	0.05	0.1	0.6	0.15	0.1

¿Cuánto valen la varianza y la desviación típica de X ? ($E(X) = 8.15$)

$$\begin{aligned} \text{Var}(X) &= (6 - 8.15)^2 \cdot f(6) + (7 - 8.15)^2 \cdot f(7) \\ &\quad + (8 - 8.15)^2 \cdot f(8) + (9 - 8.15)^2 \cdot f(9) \\ &\quad + (10 - 8.15)^2 \cdot f(10) = 0.8275 \end{aligned}$$

$$\text{Var}(X) = E(X^2) - E(X)^2 = 67.25 - 8.15^2 = 0.8275$$

$$\sigma(X) = \sqrt{\text{Var}(X)} = \sqrt{0.8275} \approx 0.90967$$

Distribución Bernoulli

Consideremos un experimento con dos resultados posibles éxito (E) y fracaso (F). Sea $\Omega = \{E, F\}$ el espacio muestral asociado al experimento. De forma que sabemos que $P(E) = p$ y $P(F) = 1 - p = q$ con $0 < p < 1$. Consideremos la aplicación $X : \Omega = \{E, F\} \rightarrow \mathbb{R}$ definida por $X(E) = 1$, $X(F) = 0$ entonces su función de probabilidad es

$$f(x) = \begin{cases} q & \text{si } x = 0 \\ p & \text{si } x = 1 \\ 0 & \text{en cualquier otro caso} \end{cases}$$

Bajo estas condiciones diremos que X sigue una distribución de probabilidad **Bernoulli** de parámetro p y lo denotaremos por $Ber(p)$ o $B(1, p)$. A los experimentos de este tipo (éxito/fracaso) se les denomina experimentos Bernoulli.

Distribución Binomial

Supongamos que repetimos n veces de forma independiente un experimento Bernoulli de parámetro p . Entonces Ω estará formado por cadenas de E 's y F 's de longitud n .

Sea $X : \Omega \rightarrow \mathbb{R}$ definida por $X(\omega)$ = número de éxitos en ω . Entonces

$f(k) = P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$ si $k = 0, 1, \dots, n$ siendo nula en el resto de casos.

Entonces diremos que la v.a. sigue una **ley de probabilidad binomial** con parámetros n y p y lo denotaremos por $B(n, p)$. (Nota: $B(1, p) = \text{Ber}(p)$)

Resumen Binomial

Sea X una v.a. con distribución $B(n, p)$.

Valores admisibles.	$P_X(x) = P(X = x) =$	$F(x) =$ $P(X \leq x) =$	$E(X)$	$Var(X)$
$D_X = \{0, 1, \dots, n\}$	$\begin{cases} \binom{n}{x} p^x q^{n-x} & \text{si } x = 0, 1, \dots, n \\ 0 & \text{en otro caso} \end{cases}$	Tabulada	np	npq

R y las distribuciones de probabilidad

R conoce las distribuciones de probabilidad más importantes, por ejemplo la binomial (`binom`).

Dada una distribución

- Añadiendo el prefijo `d`, obtenemos la correspondiente función de probabilidad (o de densidad): por ej., de la binomial, `dbinom`
- Añadiendo el prefijo `p`, obtenemos la correspondiente distribución acumulativa: por ej., de la binomial, `pbinom`
- Añadiendo el prefijo `q`, obtenemos el cuantil de la correspondiente distribución acumulativa: por ej., de la binomial, `qbinom`
- Añadiendo el prefijo `r`, obtenemos muestras aleatorias de esa distribución: por ej., de la binomial, `rbinom`

Con R

Por ejemplo, si estamos usando una $B(20, 0.3)$

```
> dbinom(5, 20, 0.3)
```

```
[1] 0.1788631
```

```
> pbinom(5, 20, 0.3)
```

```
[1] 0.4163708
```

```
> qbinom(0.5, 10, 0.3)
```

```
[1] 3
```

```
> rbinom(10, 5, 0.3)
```

```
[1] 0 0 0 4 1 3 0 1 5 1
```

Nos dan respectivamente:

- `dbinom(5,20,0.3)` nos da el valor de la función densidad en el 5, $f(5)$.
- `pbinom(5,20,0.3)` nos da el valor de la distribución acumulada en 5, $F(5)$.
- `qbinom(0.5,10,0.3)` nos da el $Q_{0.5}$ la mediana para la distribución acumulada.
- `rbinom(10,5,0.3)` nos da diez números aleatorios de una $B(5, 0.3)$.

Ejemplo

En una cierta enfermedad, la probabilidad que un hijo de madre enferma enferme es 0.5006. Una mujer enferma tiene 4 hijos.

¿Cuál es la probabilidad de que tenga exactamente un hijo enfermo?

X = número de hijos enfermos de la madre enferma. una distribución $B(4, 0.5006)$. Por lo tanto

$$P(X = 1) = f(1) = \binom{4}{1} 0.5006^1 \cdot 0.4994^3 = \dots$$

Ejemplo

$$f(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

En una cierta enfermedad, la probabilidad que un hijo de madre enferma enferme es 0.5006. Una mujer enferma tiene 4 hijos.

¿Cuál es la probabilidad de que tenga al menos un hijo enfermo?

X = número de hijos enfermos de la madre enferma.

Que sigue una ley de distribución $B(4, 0.5006)$. Por lo tanto

$$\begin{aligned} P(X \geq 1) &= P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) \\ &= f(1) + f(2) + f(3) + f(4) \\ &= \binom{4}{1} 0.5006^1 \cdot 0.4994^3 + \binom{4}{2} 0.5006^2 \cdot 0.4994^2 \\ &\quad + \binom{4}{3} 0.5006^3 \cdot 0.4994^1 + \binom{4}{4} 0.5006^4 \cdot 0.4994^0 = \dots \end{aligned}$$

Ejemplo

$$f(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

X = número de hijos enfermos de la madre enferma.

¿Cuál es la probabilidad de que tenga al menos un hijo enfermo?

X = número de hijos enfermos de la madre enferma.

Su distribución de probabilidad es $B(4, 0.5006)$. En esta ocasión la calcularemos de otra manera

$$\begin{aligned} P(X \geq 1) &= 1 - P(X = 0) = 1 - f(0) \\ &= 1 - \binom{4}{0} 0.5006^0 \cdot 0.4994^4 = \dots \end{aligned}$$

Ejemplo

$$f(k) = \binom{n}{k} p^k (1-p)^{n-k}$$

X = número de hijos enfermos de la madre enferma.

¿Cuál es la media y la varianza del número de hijos enfermos?

Su distribución de probabilidad es $B(4, 0.5006)$. Por lo tanto $n = 4$ y $p = 0.5006$. Por lo tanto la esperanza es

$$E(X) = np = 4 \cdot 0.5006,$$

la varianza es,

$$\text{Var}(X) = npq = 4 \cdot 0.5006 \cdot (1 - 0.5006).$$

y la desviación típica es $\sqrt{4 \cdot 0.5006 \cdot (1 - 0.5006)}$.

Uso de tablas para el cálculo de la distribución binomial

- Tradicionalmente, hace no muchos años, para el cálculo de las probabilidades de la distribución binomial se utilizaban tablas.
- Algunas de estas tablas las podéis encontrar en la carpeta de material adicional en el sitio de la asignatura en campus extens.
- Estas tablas han sido elaboradas con R.
- Como ejercicio repetir los cálculos anteriores con estas tablas.
- **Atención:** Son estas tablas las que utilizaréis en los controles escritos.

Distribución Geométrica

- Consideremos una experiencia consistente en repetir un experimento Bernouilli, de parámetro p , de forma independiente hasta obtener el primer éxito.
- Sea X la v.a. que cuenta el número de fracasos necesarios para obtener el primer éxito.
- Entonces $f(x) = P(X = x) = p(1 - p)^x$ si $x = 0, 1, 2, \dots$ siendo nula en el resto de casos.
- Una v.a. de este tipo diremos que sigue una distribución geométrica de parámetro p y lo denotaremos por $Ge(p)$.

Resumen Geometría

Sea X una v.a. $Ge(p)$.

X = número de fracasos para conseguir el primer éxito.

Valores admisibles.	$P_X(x) = P(X = x) =$	$F(x) = P(X \leq x) =$	$E(X)$	$Var(X)$
$D_X = \{0, 1, \dots\}$	$\begin{cases} q^k p & \text{si } k = 0, 1, 2, \dots \\ 0 & \text{en otro caso} \end{cases}$	$\begin{cases} 0 & \text{si } x < 0 \\ 1 - q^{k+1} & \text{si } \begin{cases} k \leq x < k+1 \\ \text{para } k = 0, 1, 2, \dots \end{cases} \end{cases}$	$\frac{q}{p}$	$\frac{q}{p^2}$

Propiedad de la carencia de memoria

Sea X una v.a. discreta, con dominio $D_X = \{0, 1, 2, 3, \dots\}$. Entonces X sigue una ley $Ge(p)$ sii $P(X \geq k + j | X > j) = P(X \geq k)$ para todo $k, j = 1, 2, 3, \dots$ y $P(X = 0) = p$

Ejemplo

- Supongamos que repetimos sucesivamente varias veces cultivos de unas pruebas analíticas, por ejemplo de orina humana.
- En ocasiones el cultivo se puede contaminar y se invalida la prueba.
- El motivo de la contaminación puede ser diverso. Se estima que sucede en 1 de cada 10 casos y es fácilmente detectable por la abundancia de tipos de microorganismos presentes de forma natural en el individuo que aparecen en el cultivo.
- Consideremos la variable X que nos da el número de análisis consecutivos sin error.

Suponiendo independencia entre cada análisis modelizar la variable X mediante una ley geométrica. Calcular $P(X > 5)$, $P(X = 2)$, $P(X \leq 3)$ y la esperanza, varianza y desviación típica de X . Interpretar los resultados.

Ejemplo: el problema de las llaves

- Llegamos a casa de noche después después de una fiesta.
- Para abrir la puerta de nuestra casa, que sólo tiene una cerradura, disponemos de 5 llaves.
- Tenemos la mala suerte de que se ha estropeado la luz en la puerta de nuestra casa, de forma que estamos totalmente a oscuras.
- Nos disponemos a abrir la puerta eligiendo al azar una llave.
- Como estamos “cansados” cada vez “olvidamos” (no tenemos memoria) la llave utilizada.

Ejemplo: el problema de las llaves

Sea Y la variable que nos da el número de intentos necesarios para abrir la puerta. Notemos que en este caso se incluye el intento exitoso. Modelizar $Y = X + 1$ donde X sea una variable geométrica. Calcular $P(Y > 3)$, $P(Y \leq 2)$ y $P(X = 5)$. Calcular la esperanza, varianza y desviación típica de Y .

Responder a la siguiente pregunta ¿si ya he hecho 10 intentos fallidos, cuál es la probabilidad de que tenga que hacer 10 más hasta abrir la puerta?

Cálculo probabilidades de la distribución Geométrica con R

Con R, es geom, dado p

- $f(k) = \text{dgeom}(k, p)$
- $F(k) = \text{pgeom}(k, p)$

```
> dgeom(0, 0.4)
```

```
[1] 0.4
```

```
> dgeom(5, 0.4)
```

```
[1] 0.031104
```

```
> pgeom(0, 0.4)
```

```
[1] 0.4
```

```
> pgeom(5, 0.4)
```

```
[1] 0.953344
```

```
> pgeom(30, 0.4)
```

```
[1] 0.9999999
```

Ejemplo

Leemos de forma equiprobable bases de una secuencia genómica en la que cada base aparece con la misma frecuencia, hasta que encontremos una A
¿Cuál es la probabilidad de que paremos antes de leer 4 bases?

Y = número de bases que hemos de leer para leer una A

La distribución es $Y = X + 1$ donde X es una $G(0.25)$

$$P(Y < 4) = P(X + 1 < 4) = P(X < 3) = P(X \leq 2) = F(2) = 1 - (1 - 0.25)^{2+1} = \dots$$

Ejemplo

¿Cuál es el número esperado de veces que tendríamos que tirar un dado al aire antes de que salga un 6?

X = número de veces que hemos de tirar un dado antes de que salga un 6.

La distribución de X es $G(1/6)$

$$E(X) = \frac{1}{1/6} = 6$$

Binomial negativa(**OPCIONAL**)

Bajo las mismas condiciones que en el caso anterior repetimos el experimento hasta obtener el r -ésimo éxito. Sea X la v.a que cuenta el número de repeticiones del experimento hasta el r -ésimo éxito. Entonces

$$f(x) = P(X = x) = \binom{x-1}{r-1} (1-p)^{x-r} p^r \text{ si } x = r, r+1, \dots,$$

siendo cero en el resto de casos. Se denota por $BN(p, r)$. Notar que $BN(p, 1) = Ge(p)$.

Distribución de Poisson

Diremos que una v.a. discreta X con $X(\Omega) = \mathbb{N}$ tiene distribución de Poisson con parámetro $\lambda > 0$, y lo denotaremos por $Po(\lambda)$ si su función de probabilidad es:

$$f(x) = P(X = x) = \frac{\lambda^x}{x!} e^{-\lambda} \text{ si } x = 0, 1, \dots$$

siendo cero en el resto de casos.

La distribución Poisson como límite de una binomial

- La distribución Poisson aparece en el conteo de determinados eventos que se producen en un intervalo de tiempo o en el espacio.
- Supongamos que nuestra variable de interés es $X =$ número de eventos en el intervalo de tiempo $(0, t]$ (por ejemplo el número de especímenes que caen en una trampa para insectos en una hora) y que cumpla las siguientes condiciones:

Condiciones de la distribución Poisson

- a) El número promedio de eventos en el intervalo $(0, t]$ es $\lambda > 0$
- b) Es posible dividir el intervalo de tiempo en un gran número de subintervalos n de forma que:
 - ▶ La probabilidad de que se produzcan dos o más eventos en un subintervalo es despreciable.
 - ▶ La ocurrencia de eventos en un subintervalo es independiente de los demás.
 - ▶ La probabilidad de que un evento ocurra en un subintervalo es $p = \lambda/n$

- Bajo estas condiciones podemos considerar que el número de eventos en el intervalo $(0, t]$ será el número de “éxitos” en n repeticiones independientes de un proceso Bernoulli de parámetro p
- Entonces si $n \rightarrow \infty$ y pn se mantiene igual a λ resulta que la función de probabilidad de X se puede poner como:

$$f(k) = \lim_{n \rightarrow \infty} \binom{n}{k} p^k q^{n-k} = \frac{\lambda^k}{k!} e^{-\lambda}$$

Resumen distribución Poisson

Valores admisibles.	$P_X(x) = P(X = x) =$	$F(x) = P(X \leq x) =$	$E(X)$	$Var(X)$
$D_X = \{0, 1, \dots\}$	$\begin{cases} \frac{\lambda^x}{x!} e^{-\lambda} & \text{si } x = 0, 1, \dots \\ 0 & \text{en otro caso} \end{cases}$	Tabulada.	λ	λ

Procesos Poisson

Si tenemos un experimento *tipo* Poisson con λ igual al promedio de eventos en una unidad de tiempo (u.t.) entonces si t es una cantidad de tiempo en u.t., la v.a. X_t =numero de eventos en el intervalo $(0, t]$ es una $Po(\lambda \cdot t)$.

Al conjunto de variables X_t se les denomina proceso de Poisson.

Distribución Poisson con R

La función es pois:

- $f(k) = \text{dpois}(k, \text{lambda})$.
- $F(k) = \text{ppois}(k, \text{lambda})$
- Para el cuantil p $\text{qpois}(p, \text{lambda})$.
- Para la generación aleatoria es $\text{rpois}(n, \text{lambda})$.

Unos ejemplos:

```
> dpois(5, 20)
```

```
[1] 5.49641e-05
```

```
> dpois(25, 20)
```

```
[1] 0.04458765
```

```
> ppois(25, 20)
```

```
[1] 0.887815
```

```
> qpois(0.75, 20)
```

```
[1] 23
```

```
> rpois(10, 20)
```

```
[1] 23 18 26 23 23 24 21 23 17 17
```

Ejemplo

Recordemos que la distribución Poisson es $f(k) = e^{-\lambda} \cdot \frac{\lambda^k}{k!}$

- En un cierto campo, encontramos un promedio de 3 ejemplares de una cierta planta invasora por m^2 .
- Sea X la variable que nos da el número de ejemplares de esta planta por m^2 .

Modelizar X con una ley Poisson. ¿Cuál es la probabilidad de que, en un región determinada de 1 m^2 hallemos como a mínimo 3 ejemplares de esta planta invasora?

La distribución de X es una $\text{Po}(3)$. Calculemos la probabilidad pedida:

$$\begin{aligned} P(X \geq 3) &= 1 - (P(X = 0) + P(X = 1) + P(X = 2)) = \\ 1 - f(0) - f(1) - f(2) &= 1 - e^{-3}(1 + 3 + \frac{3^2}{2}) = \dots \end{aligned}$$

Ejemplo

- Supongamos que disponemos de una trampa para insectos. El número promedio (λ) de insectos capturados por minuto es 2.
- Sea X_t la variable que nos da el número de insectos capturados en t minutos.

Modelizar X_t mediante una ley Poisson. Calcular $P(X_5 < 8)$ y $P(X_4 = 3)$. Calcular la esperanza, varianza y desviación típica de X_6 .

Aproximación de la distribución binomial por la Poisson:

Bajo el punto de vista anterior y si p es pequeño y n suficientemente grande (existen distintos criterios por ejemplo $n > 20$ ó 30 y $p \leq 0.1$) podemos aproximar una $B(n, p)$ por una $Po(np)$

Ejercicio: comprobadlo con R.

Distribución Hipergeométrica:

- Es la que modeliza el número de bolas blancas extraídas de una urna sin reposición.
- Consideremos una urna que contiene N bolas de las que N_1 son blancas y las restantes N_2 no. Obviamente $N = N_1 + N_2$.
- Extraemos n bolas de la urna sin reemplazarlas.
- Sea X la v.a. que cuenta el número de bolas blancas extraídas.
- Entonces

$$f(x) = P(X = x) = \frac{\binom{N_1}{x} \binom{N_2}{n-x}}{\binom{N}{n}}$$

si $\max\{0, n - N_2\} \leq x \leq \min\{n, N_1\}$ con $x \in \mathbb{N}$ y cero en el resto de casos. Denotaremos que una v.a. tiene distribución hipergrométrica por $H(N_1, N_2, n)$.

Resumen distribución hipergeométrica

Valores admisibles.	$P_X(x) = P(X = x) =$	$F(x) =$ $P(X \leq x) =$	$E(X)$	$Var(X)$
$D_X =$ $\{x \in \mathbb{N} \mid \max\{0, n - N_2\} \leq x$ $x \leq \min\{n, N_1\}\}$	$\begin{cases} \frac{\binom{N_1}{x} \binom{N_2}{n-x}}{\binom{N}{n}} & \text{si } x \in D_X \\ 0 & \text{en otro caso} \end{cases}$	No tiene expresión.	$\frac{nN_1}{N}$	$n \frac{N_1}{N} \left(1 - \frac{N_1}{N}\right) \frac{N-n}{N-1}$

La función es `hyper`:

- $f(k) = \text{dhyper}(k, N1, N2, n)$
- $F(k) = \text{phyper}(k, N1, N2, n)$
- Para cuantiles de orden p es `qhiper(p, N1, N2, n)`.
- Para la generación de 10 números aleatorios `rhiper(10, N1, N2, n)`.

```
> dhyper(5, 20, 30, 8)
```

```
[1] 0.1172448
```

```
> phyper(5, 20, 30, 8)
```

```
[1] 0.9640288
```

```
> qhyper(0.5, 20, 30, 8)
```

```
[1] 3
```

```
> rhyper(10, 20, 30, 8)
```

```
[1] 4 3 3 3 4 6 4 1 3 3
```


Ejemplo

En un lago con 500 peces, hay 20 marcados. Si capturamos 15 ¿cuál es la probabilidad de que capturemos alguno marcado?

X = número de peces marcados entre los 15 capturados.

La distribución de X es $H(15, 20, 480)$

$$\begin{aligned}P(X \geq 1) &= 1 - P(X = 0) = 1 - f(0) \\&= 1 - \frac{\binom{20}{0} \cdot \binom{480}{15}}{\binom{500}{15}} = \dots\end{aligned}$$

Ejemplo

En un lago con 500 peces, hay 20 marcados. Si capturamos 15 ¿cuál es el número esperado de peces marcados en nuestra captura?

X = número de peces marcados entre los 15 capturados.

La distribución es $H(20, 480, 15)$

$$E(X) = \frac{15 \cdot 20}{500}$$

Variables aleatorias continuas

- Diremos que una v.a. X es continua si su función de distribución es continua.
- Notemos que en este caso $F_X(x^-) = F_X(x)$ y entonces $P(X = x) = F_X(x) - F_X(x^-) = 0$ para todo $x \in \mathbb{R}$
- Una función $f : \mathbb{R} \rightarrow \mathbb{R}$ es una **función de densidad** o simplemente **densidad** si cumple las siguientes condiciones:
 - a) $f(x) \geq 0$ para todo $x \in \mathbb{R}$
 - b) f es integrable y $\int_{-\infty}^{+\infty} f(x)dx = 1$

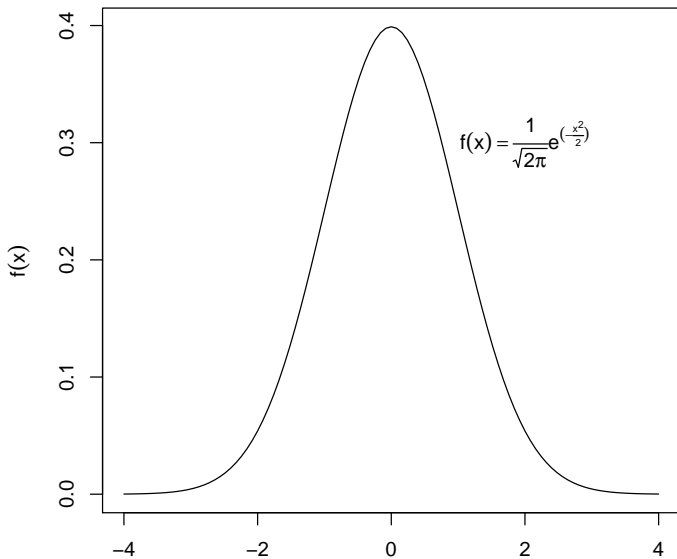
Interpretación de la función de densidad

El área comprendida entre la curva formada por la función de densidad y el eje horizontal es 1.

El siguiente código de R nos dibuja esta curva para el caso de la densidad gaussiana estándar:

```
> options(width = 60)
> curve((1/sqrt(2 * pi)) * exp(-(1/2) * x^2), from = -4,
+       to = 4, ylab = expression(f(x)))
> text(c(-3.5, 0.3), expression(f(x) == frac(1,
+       sqrt(2 * pi)) * e^(-frac(x^2, 2))))
```

Veamos el gráfico....



Variables absolutamente continuas

- Diremos que una v.a. tiene una ley **absolutamente continua** si su función de distribución se puede escribir como

$$F(x) = \int_{-\infty}^x f_X(t) dt$$

para todo $x \in \mathbb{R}$.

- Donde f es una función de densidad a la que llamaremos función de densidad de la v.a. X
- Cometeremos el abuso de decir **continua** una variable **absolutamente continua**.

Propiedad

Si X es una v.a. tiene una función de distribución F absolutamente continua se cumple que:

- a) F_X es continua.
- b) Su función de densidad es $f(x) = F'_X(x) = \frac{d}{dx} F_X(x)$.
- c) $D_X = \{x | f(x) > 0\}$. Además $1 = \int_{-\infty}^{\infty} f(x) dx = \int_{D_X} f(x) dx$.
- d) Si A es un intervalo real, entonces $P(X \in A) = \int_A f(x) dx$.

Por ejemplo si $A = (a, b]$ entonces

$$P(X \in (a, b]) = P(a < X \leq b) = \int_a^b f(x) dx.$$

- e) Esta propiedad es similar para otro tipo de intervalos.

Ejemplo

- Supongamos que elegimos con “equiprobabilidad” un número al azar del intervalo real $(0, 1)$.
- Sea X la v.a. que nos da ese número.
- Dado $0 < x < 1$ tenemos que

$$P(X \leq x) = \frac{\text{longitud casos favorables}}{\text{longitud casos posibles}} = \frac{x - 0}{1 - 0} = x.$$

- Por lo tanto la función de distribución es

$$F(x) = \begin{cases} 0 & \text{si } x \leq 0 \\ x & \text{si } 0 < x < 1 \\ 1 & \text{si } x \geq 1 \end{cases}$$

- La v.a. X es absolutamente continua. Su función de densidad es

$$F'(x) = f(x) = \begin{cases} 1 & \text{si } 0 < x < 1 \\ 0 & \text{en el resto de casos} \end{cases}$$

- Ahora tenemos que dado $0 < x < 1$

$$F(x) = \int_{-\infty}^x 1 dt = \int_{-\infty}^0 1 dt + \int_0^x t dt = t \Big|_{t=0}^{t=x} = x - 0 = x.$$

- Efectivamente la integral de la función de densidad es la función de distribución.

Ejemplo

Sea X una variable aleatoria con dominio $D_X = (1, 2)$. Su función de densidad es:

$$f_X(x) = \begin{cases} k \cdot x^2 + \frac{1}{3} & \text{si } 0 < x < 1 \\ 0 & \text{en otro caso} \end{cases}$$

Donde k es una constante real ¿Cuál es el valor de k ?

$$1 = \int_0^1 (kx^2 + \frac{1}{3}) dx = \left(k \frac{x^3}{3} + \frac{1}{3}x \right) \Big|_{x=0}^{x=1} = \left(k \frac{1}{3} + \frac{1}{3} \right) - (0 + 0) = \frac{k+1}{3}$$

Resolviendo la ecuación $1 = \frac{k+1}{3}$ obtenemos que $k = 2$.

Gráfica de la función de densidad

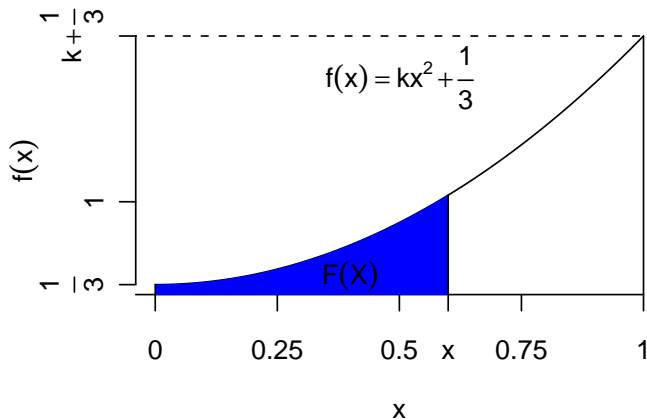


Figura: Gráfica de una densidad y su relación con la función de distribución.

- En resumen, la función de densidad de la v.a. X es

$$f_X(x) = \begin{cases} 2x^2 + \frac{1}{3} & \text{si } 0 < x < 1 \\ 0 & \text{en otro caso} \end{cases}$$

- Dado $0 < x < 1$, la función de distribución es,

$$F(x) = \int_0^x (2t^2 + \frac{1}{3}) dt = \left(2\frac{t^3}{3} + \frac{1}{3}t \right) \Big|_{t=0}^{t=x} = \left(2\frac{x^3}{3} + \frac{1}{3}x \right) = \frac{2x^3+x}{3}.$$

Esperanza de una variable aleatoria continua

Sea X una v.a. continua con función de densidad $f(x)$.

- Llamaremos **esperanza, media o valor esperado** de la v.a. X a

$$E(X) = \int_{-\infty}^{\infty} x \cdot f_X(x) dx$$

- Si el dominio de X es $D_X = (a, b)$ entonces

$$E(X) = \int_a^b x \cdot f_X(x) dx$$

Ejemplo

Consideremos la v.a. X con función de densidad:

$$f(x) = \begin{cases} 2x^2 + \frac{1}{3} & \text{si } 0 < x < 1 \\ 0 & \text{en el resto de casos} \end{cases}$$

Su valor esperado es

$$\begin{aligned} E(X) &= \int_0^1 x \left(2x^2 + \frac{1}{3} \right) dx = \int_0^1 \left(2x^3 + \frac{1}{3}x \right) dx = \left(2\frac{x^4}{4} + \frac{1}{3}\frac{x^2}{2} \right) \Big|_{x=0}^{x=1} \\ &= \frac{1}{2} + \frac{1}{6} - (0 + 0) = \frac{2}{3}. \end{aligned}$$

Esperanza de una función de una v.a.

- Sea X una v.a. continua con densidad $f(x)$ y sea $Y = g(X)$ una v.a. continua función de X . Entonces se define la **esperanza de la función** como :

$$E(Y) = E(g(X)) = \int_{-\infty}^{+\infty} g(x)f(x)dx.$$

- Si el dominio de X es $D_X = (a, b)$ entonces:

$$E(Y) = E(g(X)) = \int_a^b g(x)f(x)dx.$$

Ejemplo

Consideremos la v.a. X con función de densidad:

$$f(x) = \begin{cases} 2x^2 + \frac{1}{3} & \text{si } 0 < x < 1 \\ 0 & \text{en el resto de casos} \end{cases}$$

El valor esperado de la función $Y = X^2$ es:

$$\begin{aligned} E(Y) &= E(X^2) = \int_0^1 x^2 \left(2x^2 + \frac{1}{3}\right) dx = \int_0^1 \left(2x^4 + \frac{1}{3}x^2\right) dx \\ &= \left(2\frac{x^5}{5} + \frac{1}{3}\frac{x^3}{3}\right) \Big|_{x=0}^{x=1} = \frac{2}{5} + \frac{1}{9} - (0 + 0) = \frac{23}{45}. \end{aligned}$$

Varianza de una v.a. continua

Al igual que en el caso discreto, se define la **varianza** de una variable aleatoria continua como:

$$\text{Var}(X) = E(X - E(X))^2$$

Se cumple que

$$\text{Var}(X) = E(X^2) - E(X)^2$$

Ejemplo

Consideremos la v.a. X con función de densidad:

$$f(x) = \begin{cases} 2x^2 + \frac{1}{3} & \text{si } 0 < x < 1 \\ 0 & \text{en el resto de casos} \end{cases}$$

Hemos visto anteriormente que $E(X) = \frac{2}{3}$ y $E(X^2) = \frac{23}{45}$. Por lo tanto su varianza es:

$$\text{Var}(X) = E(X^2) - E(X)^2 = \frac{23}{45} - \left(\frac{2}{3}\right)^2 = \frac{1}{15}.$$

Notaciones esperanzas y varianzas

- De forma frecuente se denota por μ y σ^2 a la esperanza y varianza poblacionales de una distribución.
- La raíz cuadrada positiva de la varianza se denota por σ y recibe el nombre de desviación típica poblacional.

Propiedades de la esperanza y la varianza

- a) $E(cte.) = cte.$
- b) $E(aX + b) = aE(X) + b.$
- c) $E\left(\sum_{k=1}^n g_k(X)\right) = \sum_{k=1}^n E(g_k(X)).$
- d) Si $a < X < b$ entonces $a < E(X) < b.$
- e) Si X es una v.a. no negativa entonces $E(X) \geq 0.$
- f) Si $g(X) \leq h(X)$ entonces $E(g(X)) \leq E(h(X)).$
- g) $Var(aX + b) = a^2 Var(X)$ donde a, b son ctes. reales.
- h) $Var(cte.) = 0$

Algunas distribuciones continuas notables

- En esta sección describiremos tres modelos de variables continuas.
- Concretamente son: el modelo **uniforme**, el **exponencial** y el **gaussiano** o **normal**.
- A lo largo del curso, en la medida que sea necesario, expondremos otros modelos de distribuciones continuas como: la t de **Student**, la **ji-cuadrado**, la F de **Fisher**....
- La función de distribución de estas distribuciones puede que no tenga una fórmula explícita.
- En caso de no tener fórmulas explícitas se calcularán mediante tablas y usando R.

Distribución uniforme

Una v.a. continua X diremos que tiene una **distribución uniforme** sobre el intervalo real (a, b) , $(a < b)$, si su función de densidad es

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{si } a < x < b \\ 0 & \text{en cualquier otro caso} \end{cases}$$

(como ejercicio comprobar que el área comprendida entre f_X y el eje de abscisas (eje horizontal o eje X) vale 1.)

Si X es una v.a. uniforme en el intervalo (a, b) , escribiremos $X \equiv U(a, b)$.

Su función de distribución es:

$$F_X(x) = \begin{cases} 0 & \text{si } x \leq a \\ \frac{x-a}{b-a} & \text{si } a < x < b \\ 1 & \text{si } b \leq x \end{cases}$$

Efectivamente:

- Si $x \leq a$ entonces $F_X(x) = \int_{-\infty}^x f(t)dt = \int_{-\infty}^x 0dt = 0$
- Si $a < x < b$ entonces $F_X(x) = \int_{-\infty}^x f(t)dt = \int_{-\infty}^a 0dt + \int_a^x \frac{1}{b-a} dt = \frac{t}{b-a} \Big|_a^x = \frac{x}{b-a} - \frac{a}{b-a} = \frac{x-a}{b-a}$
- Por último si $x \geq b$ entonces $F_X(x) = \int_{-\infty}^x f(t)dt = 1$ (ejercicio).

Esperanza y varianza para una v.a. X con distribución $U(a, b)$

$$E(X) = \int_{-\infty}^{+\infty} x f_X(x) dx = \int_a^b x \frac{1}{b-a} dx = \frac{x^2}{2(b-a)} \Big|_a^b = \frac{b+a}{2},$$

$$\begin{aligned} E(X^2) &= \int_{-\infty}^{+\infty} x^2 f_X(x) dx = \int_a^b x^2 \frac{1}{b-a} dx = \frac{x^3}{3(b-a)} \Big|_a^b = \frac{b^3-a^3}{3(b-a)} \\ &= \frac{b^2+ab+a^2}{3}, \end{aligned}$$

$$Var(X) = E(X^2) - (E(X))^2 = \frac{b^2+ab+a^2}{3} - \left(\frac{b+a}{2}\right)^2 = \frac{(b-a)^2}{12}.$$

Gráfica de la distribución $U(-1, 2)$

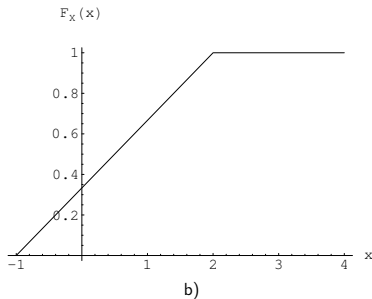
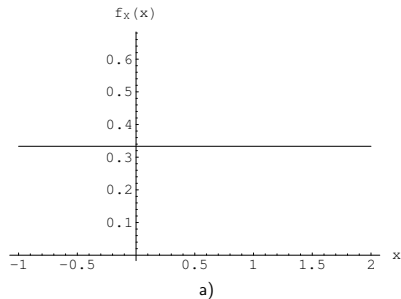


Figura: Gráficas de la función de densidad (a) y de la función de distribución (b) de una v.a. $U(-1, 2)$.

Valores admisibles.	$f_X(x)$	$F_X(x) = P(X \leq x) =$	$E(X)$	$Var(X)$
$D_X = (a, b)$	$\begin{cases} \frac{1}{b-a} & \text{si } a < x < b \\ 0 & \text{en cualquier otro caso} \end{cases}$	$\begin{cases} 0 & \text{si } x \leq a \\ \frac{x-a}{b-a} & \text{si } a \leq x \leq b \\ 1 & \text{si } b \leq x \end{cases}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$

Modelo exponencial

- Supongamos que tenemos un proceso Poisson con parámetro λ en una unidad de tiempo.
- Entonces dadas t unidades de tiempo tenemos que $X_t =$ número de eventos en el intervalo de tiempo $(0, t]$ es una $Po(\lambda \cdot t)$.
- Consideremos la v.a. $T =$ tiempo transcurrido entre dos eventos Poisson consecutivos.
- Dado $t > 0$, tenemos que
$$P(T > t) = P(\text{Cero eventos en el intervalo}(0, t]) = P(X_t = 0) = \frac{(\lambda t)^0}{0!} e^{-\lambda t} = e^{-\lambda t}.$$

- Tomando complementarios, la función de distribución de T será

$$F_T(t) = P(T \leq t) = \begin{cases} 0 & \text{si } t \leq 0 \\ 1 - P(T > t) = 1 - e^{-\lambda t} & \text{si } t > 0 \end{cases}$$

- Por lo tanto

$$f_T(t) = \begin{cases} \lambda e^{-\lambda t} & \text{si } t > 0 \\ 0 & \text{si } t \leq 0 \end{cases}$$

- Ésta es la distribución exponencial y la denotaremos por $Exp(\lambda)$

Propiedad de la falta de memoria

Sea X una v.a. $Exp(\lambda)$ entonces

$$P(X > s + t | X > s) = P(X > t) \text{ para todo } s, t \in \mathbb{R}$$

Toda v.a. absolutamente continua, que tome valores positivos y que verifique la propiedad de la falta de memoria es una v.a. exponencial.

Sea $X \equiv \text{Exp}(\lambda)$.

Valores admisibles.	$f_X(x)$	$F_X(x) = P(X \leq x) =$	$E(X)$	$\text{Var}(X)$
$D_X = (0, +\infty)$	$\begin{cases} \lambda e^{-\lambda x} & \text{si } x > 0 \\ 0 & \text{si } x \leq 0 \end{cases}$	$F_X(x) = P(X \leq x) = \begin{cases} 0 & \text{si } x \leq 0 \\ 1 - e^{-\lambda x} & \text{si } x > 0 \end{cases}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$

Distribución normal o Gaussiana

Diremos que una v.a. X sigue una **ley normal** o **gaussiana** de parámetros μ y σ y lo denotaremos por $N(\mu, \sigma)$ si tiene por función de densidad:

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \text{ para todo } x \in \mathbb{R}$$

La gráfica de esta función es la conocida campana de Gauss, que ya hemos dibujado dos veces. La v.a. normal con $\mu = 0$ y $\sigma = 1$ recibe el nombre de normal estándar.

La normal estándar se suele denotar por la letra Z y su función de distribución F_Z .

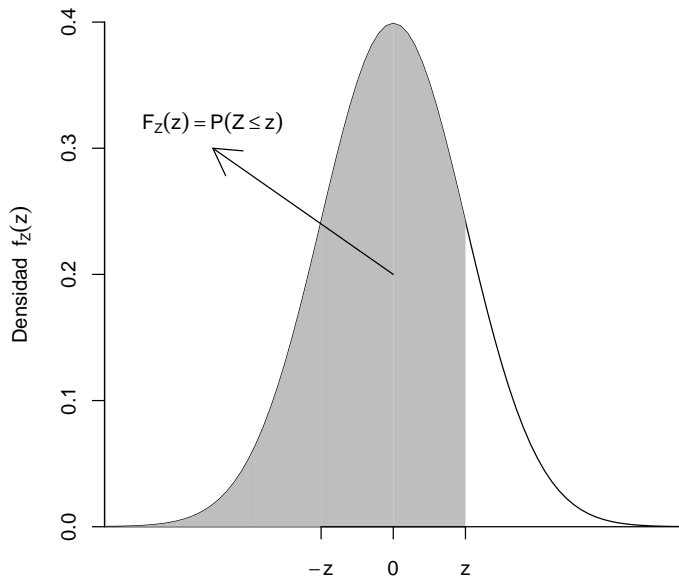
Propiedades

Sea X una v.a. $N(\mu, \sigma^2)$ y sea f_X su función de densidad. Entonces:

- a) Evidentemente f_X verifica todas las propiedades de las funciones de densidad.
- b) $f_X(\mu - x) = f_X(\mu + x)$ es simétrica respecto de la recta $x = \mu$
- c) f_X alcanza el máximo en $x = \mu$
- d) Si F_X la función de distribución de X entonces
 $F_X(\mu + x) = 1 - F_X(\mu - x)$. En particular si Z es una $N(0, 1)$ entonces
 $F_Z(-x) = 1 - F_Z(x)$
- e) $Z = \frac{X - \mu}{\sigma}$ es una v.a. $N(0, 1)$ y $X = \sigma Z + \mu$ es una $N(\mu, \sigma^2)$ donde Z es la normal estándar.

Importancia y propiedades de la distribución normal

- La variable aleatoria normal es una de las más importantes de la estadística.
- La justificación es que la distribución de muchos estadísticos como la media o la proporción muestral se aproximan a una distribución normal cuando los tamaños de las muestras son grandes.
- Como ya hemos visto en distribuciones continuas las probabilidades se calculan como áreas que se forman debajo de una curva, llamada curva de densidad, y el eje horizontal.
- Como en todas las distribuciones continuas la probabilidad $P(X \leq x)$ es el área comprendida entre el eje horizontal la curva normal y la vertical que corta al eje horizontal en x , ver la fig. 18.



- La curva normal toma distintas formas en función del valor de sus parámetros.
- Consideremos dos curvas normales con medias $\mu_1 = \mu_2$ respectivamente y varianzas $\sigma_1^2 < \sigma_2^2$.
- Las dos curvas estarán centradas en la media, pero la segunda será más achatada, pues tiene varianza mayor y los valores se alejan más del valor medio.
- Este caso se ilustra en la fig. 19

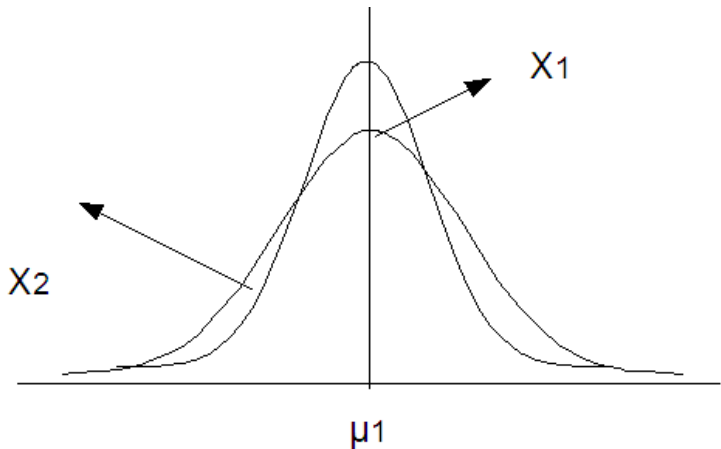


Figura: Dos curvas normales X_1 y X_2 con $\mu_1 = \mu_2$ y $\sigma_1^2 < \sigma_2^2$.

En el caso en que las varianzas sean iguales y las medias distintas, las curvas normales tienen la misma forma pero cada una de ellas está centrada en cada una de sus medias. El efecto puede verse en la fig. 20

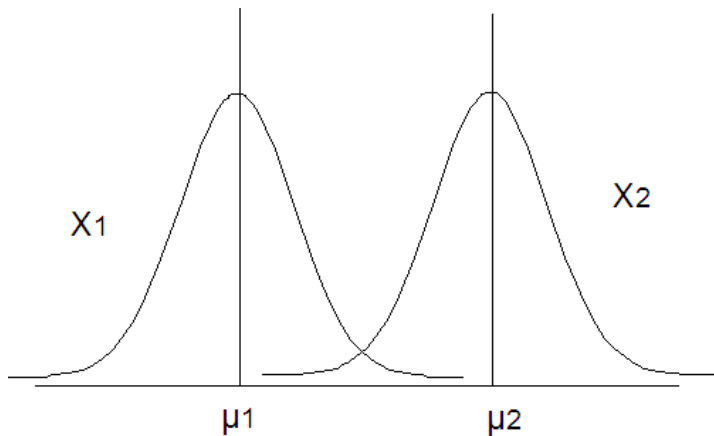


Figura: Dos curvas normales X_1 y X_2 con $\mu_1 < \mu_2$ y $\sigma_1^2 = \sigma_2^2$.

Por último si las medias son distintas y las varianzas también se obtienen curvas como las de la fig. 21

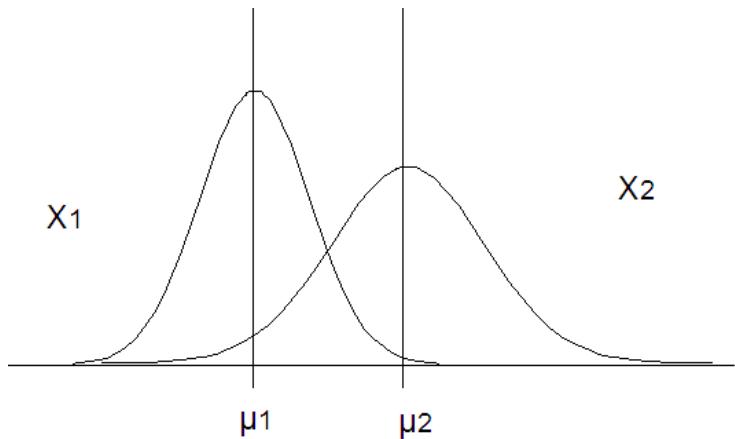


Figura: Dos curvas normales X_1 y X_2 con $\mu_1 < \mu_2$ y $\sigma_1^2 < \sigma_2^2$.

Estandarización de la variable normal

- La estandarización o tipificación de una variable ya fue comentada en el primer tema.
- Consiste en transformar una variable o una serie de datos a unos datos típicos o estándares en el sentido que todos tengan esperanza o media 0 y varianza poblacional o muestral igual a 1.
- Para conseguir este efecto es suficiente restar a la variable la media y dividir el resultado por la desviación típica. Estas ideas ya se vieron en el módulo anterior.

- En general supongamos que tenemos un variable X con valor esperado μ y desviación típica σ . Entonces sus puntuaciones estándar (*z-scores*) se suele denotar por la letra Z se calculan como $Z = \frac{X-\mu}{\sigma}$. para la versión poblacional
- Mientras que para la versión muestral, como ya vimos, es $= \frac{x-\bar{x}}{s}$

- La variable o serie de datos Z tiene esperanza o media 0 y varianza poblacional o muestral 1.
- En este sentido es una variable de puntuaciones estándar y puede servirnos para comparar las puntuaciones de dos variables o series de datos
- Este es el motivo por el que recibe el nombre de estándar, pues reduce los datos o variables a “*unidades estándar*”.
- Existen otras formas de reducir variables o series a puntuaciones que sean comparables.

- Para el caso de la variable aleatoria normal tenemos una propiedad importante.
- Es suficiente conocer todas las probabilidades de una variable aleatoria Z normal estándar o típica es decir con $\mu = 0$ y $\sigma = 1$ para conocer las probabilidades de cualquier variable normal de media μ y desviación típica σ .
- La igualdad que relaciona estas dos probabilidades es

$$P(X \leq x) = P(Z \leq \frac{x - \mu}{\sigma}).$$

Donde Z es una variable que sigue la ley normal estándar, es decir una normal con $\mu = 0$ y $\sigma = 1$.

- Lo que nos dice esta propiedad es que cuando X es una variable aleatoria con distribución normal de parámetros μ y σ^2 , la variable $Z = \frac{X - \mu}{\sigma}$ sigue una distribución normal estándar.
- A la función que nos da las probabilidades de la variable normal estándar Z la denotaremos por $F_Z(z) = P(Z \leq z)$.

Es sencillo comprender mirando la fig. 22 que la función de distribución de la variable normal estándar cumple la propiedad: $F_Z(-z) = 1 - F_Z(z)$.

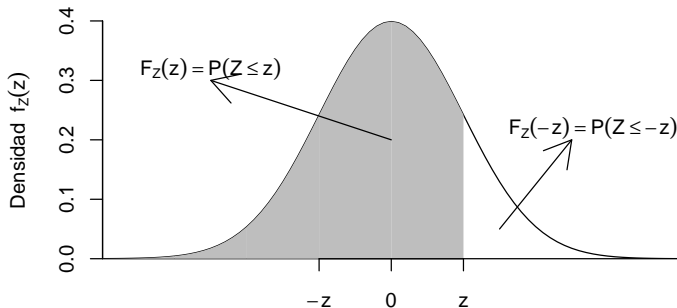


Figura: Justificación gráfica de la propiedad de la normal estándar:
 $F_Z(-z) = 1 - F_Z(z)$.

- Las tablas para el cálculo de la normal están disponibles en “<http://bioinfo.uib.es/~recerca/mates2/tablasDistribuciones/>”.
- Estas tablas contienen los valores de la distribución normal estándar. Es decir los valores de $F_Z(z) = P(Z \leq z)$ para una variable Z normal estándar.
- La primera tabula los valores negativos y la segunda, los positivos. Por ejemplo $F_Z(1.78) = 0.9625$, es un resultado que se encuentra en la segunda tabla. Si queremos saber el valor de $F_Z(-1.78) = 1 - F_Z(1.78) = 0.0375$ donde hemos utilizado la propiedad $F_Z(-z) = 1 - F_Z(z)$.
- La entrada horizontal de la tabla es el segundo decimal del número que busquemos, mientras que la primera columna contiene las unidades y el primer decimal.
- Si queremos calcular la probabilidad de que Z esté comprendida entre -2 y 2
- $P(-2 \leq Z \leq 2) = F_Z(2) - F_Z(-2) = 0.9772 - 0.0228 = 0.9544$.

- La explicación es la siguiente: la probabilidad de que Z esté entre -2 y 2 es el área comprendida bajo la curva normal, el eje horizontal y las verticales que pasa por -2 y 2 .
- Este área es el área encerrada bajo la curva y menor o igual que 2 menos el área encerrada bajo la curva y menor que -2 .
- En general se tiene que

$$P(a \leq Z \leq b) = P(Z \leq a) - P(Z \leq b) = F_Z(a) - F_Z(b).$$
- En particular dado $\delta > 0$ entonces

$$P(-\delta \leq Z \leq \delta) = F_Z(\delta) - F_Z(-\delta).$$
- **Nota:** En el caso de variables continua los \leq se pueden cambiar por $<$ (uno todos o ninguno) y las probabilidades no varían. Es decir

$$P(Z \leq a) = P(Z < a).$$

Ejemplo

Sea Z una variable normal estándar. Utilizando las tablas de la distribución normal estándar calcular las siguientes probabilidades:

- a) $P(0 < Z < 3.99) = F_Z(3.99) - F_Z(0) \approx 1 - 0.5 = 0.5.$
- b) $P(-3.99 \leq Z \leq 3.99) = F_Z(3.99) - F_Z(-3.99) \approx 1.$
- c) $P(-3 \leq Z \leq 3) = F_Z(3) - F_Z(-3) = 0.9987 - 0.0013 = 0.9974.$
- d) $P(Z \leq -2) = F_Z(-2) = 0.0228.$
- e) $P(Z \leq 2) = F_Z(2) = 0.9772.$
- f) $P(Z \geq 2) = 1 - P(Z < 2) = 1 - F_Z(2) = 0.0228.$
- g) $P(Z \geq -2) = 1 - P(Z < -2) = 1 - F_Z(-2) = 1 - (1 - F_Z(2)) = F_Z(2).$
- h) Dado $\delta > 0$,
$$P(-\delta \leq Z \leq \delta) = F_Z(\delta) - F_Z(-\delta) = F_Z(\delta) - (1 - F_Z(\delta)) = 2 \cdot F_Z(\delta) - 1.$$
- i) Utilizando la igualdad anterior
$$P(-2 \leq Z \leq 2) = 2 \cdot F_Z(2) - 1 = 2 \cdot 0.9772 - 1 = 0.9544.$$

- No sólo podemos utilizar la tabla para calcular la probabilidad en un valor determinado.
- También la podemos utilizar para calcular los cuantiles de Z que tienen la misma definición que vimos en estadística descriptiva.
- Por ejemplo si quiero calcular el valor z tal que $P(Z \leq z) = 0.9099$ buscamos dentro de la tabla el valor 0.9099, o el más cercano, en este caso es $F_Z(1.34) = 0.9099$ por lo tanto el valor buscado es $z = 1.34$.

Ejemplo

Calcular los valores que se piden para una variable Z normal estándar.

- a) Calcular el valor z tal que $P(Z \leq z) = 0.504$. Mirando las tablas se obtiene que $F_Z(0.01) = 0.504$ por lo tanto el valor pedido es $z = 0.01$.
- b) Calcular el valor z tal que $P(Z > z) = 0.9633$. Primero hacemos el complementario $P(Z > z) = 1 - P(Z < z) = 1 - F_Z(z) = 0.9633$. Ahora despejando obtenemos que el valor buscado cumple que $F_Z(z) = 1 - 0.9633 = 0.0367$. Mirando dentro de las tablas se obtiene que $F_Z(-1.79) = 0.0367$. En definitiva el valor buscado es $z = -1.79$

Cálculo de probabilidad de una v.a. $N(\mu, \sigma)$.

- Sólo nos queda ver como calculamos las probabilidades de una variable aleatoria normal de media μ y varianza σ^2 .
- Recordemos la relación básica que dice que la variable tipificada de una normal sigue una ley normal estándar. Si X es una normal de media $\mu = 1$ y varianza $\sigma^2 = 4$ se tiene que

$$Z = \frac{X - \mu}{\sigma} = \frac{X - 1}{2}$$

sigue una ley normal estándar.

- Por lo tanto $P(X \leq x) = F_Z\left(\frac{x-\mu}{\sigma}\right)$.

Ejemplo

a) Por ejemplo

$$P(X \leq 2) = P\left(\frac{X-1}{2} \leq \frac{2-1}{2}\right) = P\left(Z \leq \frac{1}{2}\right) = F_Z(0.5) = 0.6915.$$

b) Si queremos calcular el cuantil 0.6915 de la variable X será aquel valor X tal que $P(X \leq x) = 0.6915$. Tipificamos la variable X

$$P(X \leq x) = P\left(\frac{X-1}{2} \leq \frac{x-1}{2}\right) = P\left(Z \leq \frac{x-1}{2}\right) = F_Z\left(\frac{x-1}{2}\right) = 0.6915$$

mirando en las tabla de la normal estándar resulta que

$F_Z(0.5) = 0.6915$, entonces

$$\frac{x-1}{2} = 0.5,$$

despejando x de la ecuación anterior se obtiene que $x = 2 \cdot 0.5 + 1 = 2$.

Resumen propiedades de la normal

Resumiendo podemos utilizar las siguientes propiedades, $X \equiv N(\mu, \sigma)$

- Z es su variable tipificada, es decir, $Z = \frac{X - \mu}{\sigma} \equiv N(0, 1)$ entonces:

$$P(X \leq x) = P\left(\frac{X - \mu}{\sigma} \leq \frac{x - \mu}{\sigma}\right) = F_Z\left(\frac{x - \mu}{\sigma}\right)$$

- Cuando tengamos un intervalo

$$\begin{aligned} P(a < X < b) &= P\left(\frac{a - \mu}{\sigma} < \frac{X - \mu}{\sigma} < \frac{b - \mu}{\sigma}\right) = \\ &= P\left(\frac{a - \mu}{\sigma} < Z < \frac{b - \mu}{\sigma}\right) = F_Z\left(\frac{b - \mu}{\sigma}\right) - F_Z\left(\frac{a - \mu}{\sigma}\right) \end{aligned}$$

- Si $\delta > 0$ $P(\mu - \delta \leq X \leq \mu + \delta) = 2F_Z\left(\frac{\delta}{\sigma}\right) - 1$

Ejemplo

Sea X una normal con media 2 y varianza 4, entonces

- a) $P(1 < X < 2) = P(\frac{1-2}{2} < \frac{X-2}{2} < \frac{2-2}{2}) = P(\frac{-1}{2} < Z < 0) = F_Z(0) - F_Z(-0.5) = \frac{1}{2} - 1 + F_Z(0.5).$
- b) $P(X > 3) = P(\frac{X-2}{2} > \frac{3-2}{2}) = P(Z > 0.5) = 1 - F_Z(0.5).$

Aproximación de una Binomial por una distribución normal

- Bajo determinadas condiciones la distribución normal puede aproximar a la distribución binomial.
- Sea X una v.a. con distribución $B(n, p)$ entonces $E(X) = np$ y $Var(X) = npq$.
- Sea $Z = \frac{X - E(X)}{\sqrt{Var(X)}} = \frac{X - np}{\sqrt{np(1-p)}}$.
- Si n es grande y p no está muy cercano a 0 o a 1 la distribución de Z se aproxima a una normal estándar.
- La aproximación se realiza de la siguiente manera
$$P(X = k) \approx P\left(\frac{k - 0.5 - np}{\sqrt{npq}} \leq Z \leq \frac{k + 0.5 - np}{\sqrt{npq}}\right)$$
- Mediante un razonamiento similar :

$$P(X \leq k) \approx P\left(Z \leq \frac{k + 0.5 - np}{\sqrt{npq}}\right)$$

- Y también $P(a \leq X \leq b) \approx P\left(\frac{a - 0.5 - np}{\sqrt{npq}} \leq Z \leq \frac{b + 0.5 - np}{\sqrt{npq}}\right)$
- Donde, en todos los casos Z se toma como una normal estándar.

Corrección de continuidad

- El motivo de sumar o restar 0.5 en las aproximaciones es corregir el efecto que tienen aproximar una v.a. discreta por una continua.
- Esta operación recibe el nombre de corrección de continuidad de Fisher.
- Gráficamente el área que hacemos corresponder a la probabilidad de cada valor entero k en una binomial corresponde a la comprendida entre la curva normal y el segmento centrado en k de amplitud 1.

Ejemplo

X = número de caras en 100 lanzamientos de una moneda. $P(\text{cara}) = \frac{1}{2}$.

Calcular:

a) $P(40 \leq X \leq 49)$

b) $P(X = 37)$

c) $P(X \leq 50)$

Solución:

$$E(X) = 50 = \mu_X, \text{Var}(X) = 25 \text{ y } \sigma_X = 5$$

$Z = \frac{X-50}{5}$ se aproxima a normal estándar

Entonces....

a)

$$\begin{aligned}
 P(40 \leq X \leq 49) &\approx P\left(\frac{40-0.5-50}{5} \leq Z \leq \frac{49+0.5-50}{5}\right) \\
 &= P\left(-\frac{10.5}{5} \leq Z \leq -\frac{0.5}{5}\right) = F_Z\left(-\frac{0.5}{5}\right) - F_Z\left(-\frac{10.5}{5}\right) \\
 &= 1 - F_Z\left(\frac{0.5}{5}\right) - 1 + F_Z\left(\frac{10.5}{5}\right) = F_Z\left(\frac{10.5}{5}\right) - F_Z\left(\frac{0.5}{5}\right) \\
 &= F_Z(2.1) - F_Z(0.1) = 0.9821 - 0.5398 = 0.4423.
 \end{aligned}$$

La probabilidad exacta da 0.442605. ¡¡La aproximación es bastante buena!!

b)

$$\begin{aligned}
 P(X = 37) &= P(37 \leq X \leq 37) \approx P\left(\frac{37-0.5-50}{5} \leq Z \leq \frac{37+0.5-50}{5}\right) \\
 &= P\left(-\frac{13.5}{5} \leq Z \leq -\frac{12.5}{5}\right) = F_Z\left(\frac{13.5}{5}\right) - F_Z\left(\frac{12.5}{5}\right) \\
 &= F_Z(2.7) - F_Z(2.5) = 0.9965 - 0.9938 = 0.0027.
 \end{aligned}$$

La probabilidad exacta da 0.0026979. ¡¡La aproximación es bastante buena!!

c) $P(X \leq 50) \approx P\left(Z \leq \frac{50+0.5-50}{5}\right) = P(Z \leq 0.1) = F_Z(0.1) = 0.5398$

La probabilidad exacta calculada con un programa adecuado da 0.539795 ¡¡La aproximación es bastante buena!!

Aproximación de una Poisson por una distribución normal

- De forma similar a la aproximación de una binomial por una normal podemos aproximar la probabilidad de una v.a. Poisson por una normal.
- Tendremos que aplicar también la corrección de continuidad.
- Si $X \equiv Po(\lambda)$ y λ es grande, entonces podemos usar estas aproximaciones:

- ▶ $P(X = k) \approx P\left(\frac{k-0.5-\lambda}{\sqrt{\lambda}} \leq Z \leq \frac{k+0.5-\lambda}{\sqrt{\lambda}}\right)$
- ▶ $P(X \leq k) \approx P\left(Z \leq \frac{k+0.5-\lambda}{\sqrt{\lambda}}\right)$
- ▶ $P(a \leq X \leq b) \approx P\left(\frac{a-0.5-\lambda}{\sqrt{\lambda}} \leq Z \leq \frac{b+0.5-\lambda}{\sqrt{\lambda}}\right)$

Ejemplo

Sea X =número de trabajos que llegan a un centro de cálculo en un lapso de 60 minutos.

Supongamos que X sigue una ley Poisson y que el número medio de trabajos que llegan por minuto sea 0.2. Entonces $E(X) = 0.2 \cdot 60 = 12$ por lo tanto X es una $Po(12)$ es decir $\lambda = 12$ y por lo tanto $\mu_X = 12$ y $\sigma_X^2 = 12$.

Si queremos calcular

$$P(Y \leq 10) \approx P(Z \leq \frac{10+0.5-12}{\sqrt{12}}) = P(Z \leq -0.4330127)$$

$$F_Z(-0.4330127) \approx 1 - F_Z(0.43) = 1 - 0.6664 = 0.3336$$

La probabilidad exacta¹ da 0.347229. La aproximación es buena.

¹Con R es `ppois(10,12)`

Conclusión

- Si X es una $B(n, p)$ entonces $E(X) = np$ y $Var(X) = npq$
- Si X es una $Po(\lambda)$ entonces $E(X) = Var(X) = \lambda$
- Si X es una $Ge(p)$ con $X(\Omega) = \{1, 2, 3, \dots\}$ entonces $E(X) = \frac{1}{p}$ y $Var(X) = \frac{q}{p^2}$
- Si X es una $Ge(p)$ con $X(\Omega) = \{0, 1, 2, 3, \dots\}$ entonces $E(X) = \frac{q}{p}$ y $Var(X) = \frac{q}{p^2}$
- Si X es una $U(a, b)$ entonces $E(X) = \frac{a+b}{2}$ y $Var(X) = \frac{(b-a)^2}{12}$
- Si X es una $Exp(\lambda)$ $E(X) = \frac{1}{\lambda}$ y $Var(X) = \frac{1}{\lambda^2}$
- Si X es una $N(\mu, \sigma^2)$ $E(X) = \mu$ y $Var(X) = \sigma^2$

Apuntes de Bioestadística.

R. Alberich y A. Mir

Departamento de Matemáticas e
Informàtica
Universitat Illes Balears

14 de julio de 2010

10 Variables aleatorias

- Funciones de distribución
- Cuantiles

11 Variables aleatorias discretas

- Función probabilidad variables discretas
- Valor esperado y varianza para variables discretas

12 Algunas distribuciones de probabilidad discretas

- Distribución Bernoulli
- Distribución Binomial
- Distribución Geométrica
- Distribución Binomial negativa
- Distribución Poisson
 - Aproximación binomial por Poisson
- Distribución Hipergeométrica

13 Variables aleatorias continuas

- Variables absolutamente continuas
- Esperanza y varianza para variables aleatorias continuas

14 Algunos modelos de distribuciones continuas

- Distribución uniforme en el intervalo (a,b)
 - Esperanza y varianza para $U(a, b)$

- En esta tema sentaremos las bases del muestreo estadístico.
- Estudiaremos las distribuciones de algunos estadísticos muestrales; como la media aritmética, la proporción y la varianza.

Conceptos básicos

- Ya hemos estudiado algunos de los conceptos básicos sobre muestras
- Ahora los repasaremos y ampliaremos.
- Recordemos:
 - ▶ **Población:** Conjunto de individuos con una característica observable común.
 - ▶ **Muestra:** Subconjunto de la población del que se espera que la represente.

Análisis exploratorio, análisis confirmatorio

- Si tenemos información o datos que estudian un determinado fenómeno, podemos realizar un **análisis exploratorio**. Es decir, analizaremos, resumiremos e intentaremos interpretar los datos.
- Otra situación puede ser contrastar o testear una hipótesis sobre el comportamiento de unos determinados datos.
- Para confirmar dicha hipótesis, diseñaremos un estudio estadístico. A partir de dicho estudio, confirmaremos o refutaremos la hipótesis.
- Dicho estudio estadístico consistirá en el diseño de un experimento que incluye una recogida de datos.
- Dicho estudio estadístico se denomina **análisis de datos confirmatorio**, con el que queremos reforzar y aportar evidencias sobre la veracidad de nuestra hipótesis.

La estadística inferencial

- En los estudios de tipo **exploratorio**, la estadística descriptiva es una de las herramientas principales.
- En los estudios de tipo **confirmatorio** lo es la **estadística inferencial**
- El objetivo de la **estadística inferencial** es obtener información sobre el conjunto de la población a partir de un subconjunto representativo de ella llamado muestra.
- **Inferir información** de una muestra es contestar preguntas sobre el total de la población a partir del estudio de una muestra representativa de la misma.

Pasos en un estudio inferencial

Los siguientes son unos pasos habituales en un estudio inferencial:

- ¿Qué información se necesita?
- ¿Cuál es la información relevante? ¿Se dispone de acceso a todos los individuos de la población?
- ¿Cómo seleccionamos los individuos de la muestra?
- ¿Qué método emplearemos para obtener la información de los individuos de la muestra?
- ¿Qué herramientas utilizaremos para hacer inferencias?
- ¿Qué conclusiones podemos obtener?
- Si las conclusiones son fiables y suficientes, redactar informe; en caso contrario, se vuelve a empezar.

- El objetivo de las técnicas de muestreo es encontrar métodos para seleccionar muestras representativas de la población.
- Las técnicas básicas son: el muestreo aleatorio simple, el muestreo aleatorio estratificado, el muestreo sistemático y el muestreo polietápico. Cada una de estas técnicas proporciona una muestra representativa de la población.
- Describiremos de forma breve estas técnicas.

Muestreo aleatorio probabilístico

- El muestreo aleatorio consiste en seleccionar muestra de la población con igual probabilidad.
- Lo que equivale a que cualquier conjunto de individuos tiene la misma probabilidad de ser seleccionado.
- Pensemos en una urna con 100 bolas de colores. Hay dos maneras de obtener una muestra de 10 bolas.
- Una podría ser sacar una bola de la urna, observar su color y devolverla a la urna.
- Es decir vamos obteniendo individuos y los volvemos a poner en la urna.
- Este tipo de muestreo recibe el nombre de **muestreo con reposición** o **muestreo aleatorio simple**.

Muestreo aleatorio probabilístico

- Otra forma sería repetir la experiencia anterior pero no devolver las bolas a la urna.
- En este caso también sucede que cualquier selección de los 10 individuos es equiprobable. En este caso se habla de **muestreo aleatorio sin reposición**.
- Cuando el tamaño de la población sea muy grande en relación a la muestra y por lo tanto la probabilidad de que dos individuos se repitan sea muy pequeña, el muestreo aleatorio con reemplazo y el muestreo aleatorio sin reemplazo serán aproximadamente equivalentes.
- De todas maneras, si el tamaño de la población es pequeño, se suelen aplicar estadísticos corregidos por el efecto de población finita.

Muestreo Aleatorio Estratificado

- Se utiliza en el caso en que la población esté dividida en grupos o estratos y que éstos sean de interés para la variable de estudio.
- Se toman muestras donde cada grupo esté representado en función de su tamaño.
- Por ejemplo los estratos podrían ser los grupos de edad; o en las Islas Baleares, los estratos podrían ser las islas en proporción a su número de habitantes; o en una provincia, los estratos podrían ser los municipios también en función de su número de habitantes, o los estratos se podrían diseñar según el nivel educativo de sus habitantes, etc.
- En estos casos Se determina el tamaño de la muestra en cada estrato y luego se toma una muestra aleatoria simple en ese bloque.

Muestreo por conglomerados

- El proceso de obtener una muestra aleatoria en algunos casos es caro.
- Por ejemplo, si el estudio se realiza sobre conjuntos de personas, tener una lista completa de dichas personas puede ser muy costoso.
Imaginemos que queremos saber los hábitos de alimentación que tienen los estudiantes de Primaria de Baleares.
- Para ello, previo permiso de la autoridad responsable, queremos seleccionar una muestra representativa de los escolares de Baleares.
- En vez de hallar una muestra representativa de todos los estudiantes de Primaria, elegimos al azar primeramente un conjunto de colegios a los que llamamos conglomerados.
- Seguidamente, dentro de cada colegio (conglomerado) elegimos al azar un conjunto de estudiantes. Pensemos que es mucho más sencillo poseer una lista completa de estudiantes de una serie de colegios que poseer una lista completa de todos los estudiantes.

- Cuando no se da algún tipo de aleatoriedad en la selección de la muestra se habla de **muestreo no probabilístico**.
- Suele ser frecuente este tipo de muestreo. En muchos casos, nos tenemos que conformar con la información disponible o la obtenida voluntariamente.
- Existen **otros tipos** de muestreo que suelen ser combinaciones de las técnicas anteriores y otro tipos de técnicas.
- En cualquier caso, lo importante es que el estudio estadístico que se realiza “a posteriori” es diferente según el muestreo realizado.

Diferentes tipos de muestreo, diferentes estadísticos

- Una vez realizado el muestreo y obtenidos los datos (“raw data”), hemos de explicar cómo obtener estadísticos a partir de dichos datos como pueden ser proporciones, medias, varianzas, etc.
- La forma de obtener dichos estadísticos se realiza mediante los llamados estimadores. O sea, un estimador es simplemente una fórmula a aplicar a los datos del muestreo para obtener el valor de dicho estadístico.
- El ejemplo más conocido es la media aritmética: sean x_1, \dots, x_n los datos del muestreo. El estimador que nos da la media aritmética es:

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

- En este curso estudiaremos técnicas de estimación para el caso de muestreo aleatorio simple.

Muestreo aleatorio simple

- Estudiemos un poco más en detalle el caso del muestreo aleatorio simple con o sin reposición.
- La idea es que queremos seleccionar una muestra de tamaño n (es decir formada por n individuos) de una población de tamaño N .
- Obtendremos una muestra aleatoria simple (m.a.s.) cuando todas las muestras posibles de n individuos tengan la misma probabilidad de ser elegidas.
- El tener una m.a.s de una población junto con un tamaño muestral adecuado nos asegurará la suficiente representatividad de la muestra.

Observaciones sobre el muestreo aleatorio simple

Hagamos algunas observaciones:

- El proceso mismo del muestreo aleatorio simple es complejo.
- Una forma sencilla es numerar, si es posible a todos los individuos de la población y sortearlos eligiendo números como si se tratase de una lotería.
- Por ejemplo con una tabla de números aleatorios o con algún generador de números aleatorios.² Tanto los paquetes estadísticos como R o las hojas de cálculo como Open Office tienen generadores de números aleatorios.

²En realidad los números aleatorios generados por diversos tipos de algoritmos son pseudoaleatorios; son números que superan determinados test de aleatoriedad

Inferencias

- Una vez definido nuestro valor de interés y el estimador que lo estima o calcula, necesitamos estudiar dicho estimador.
- Por ejemplo, supongamos que tenemos una muestra aleatoria simple de una población y deseamos obtener información sobre la media o la varianza poblacionales.
- Para obtener dicha información sobre la media o la varianza, necesitamos definir los estimadores para calcular dichos valores a partir de los valores de la muestra. Dichos estimadores se denominan **estadísticos**.
- Un **estadístico** es una función que depende de la muestra.
- Pensemos por ejemplo en la media aritmética, proporción muestral, etc.

Distribución muestral de un estadístico

- Desde el punto de vista teórico, una m.a.s. es un conjunto de n variables aleatorias independientes e idénticamente distribuidas según la distribución de la variable aleatoria X , que representa la variable de estudio. Escrito de forma matemática: X_1, \dots, X_n .
- La muestra aleatoria serían unos determinados valores que cogen dichas variables aleatorias: x_1, \dots, x_n .
- Por tanto, al ser los estadísticos funciones de la muestra, serán funciones de las variables X_1, \dots, X_n : $T = f(X_1, \dots, X_n)$, donde T representa el estadístico y f la función a considerar.
- En el caso de la media aritmética, tenemos que
$$f(X_1, \dots, X_n) = \frac{X_1 + \dots + X_n}{n}.$$
- Un estadístico T será por tanto una variable aleatoria, desde el punto de vista teórico.
- La **distribución muestral o distribución en el muestreo** de un estadístico T será la distribución de probabilidad de la variable aleatoria T .

Ejemplo

- Supongamos que queremos estimar cuál es número medio de pruebas de embarazo defectuosas, de una determinada marca, que hay en cada caja de 10 unidades cada una.
- Para ello tomamos una muestra aleatoria simple de cuatro cajas X_1, X_2, X_3, X_4 .
- Se comprueba si son correctas o no y obtenemos los siguientes resultados:

primera caja	:	x_1	1	defectuosa
segunda caja	:	x_2	2	defectuosas
tercera caja	:	x_3	0	defectuosa
cuarta caja	:	x_4	1	defectuosas

Definimos el estadístico media aritmética de pruebas defectuosas como:

$$\bar{X} = f(X_1, X_2, X_3, X_4) = \frac{X_1 + X_2 + X_3 + X_4}{4}$$

En este caso $\bar{X} = 1$.

Supongamos que tomamos repetidas muestras de tamaño 4, los resultados son:

M. 1	M. 2	M. 3	M. 4	M. 5	M. 6	M. 7	M. 8	M. 9	M. 10
0	1	3	0	0	1	0	0	0	1
1	1	1	0	1	1	1	0	0	2
0	1	2	1	0	0	1	2	0	1
1	1	2	2	1	3	0	0	1	1

M. 11	M. 12	M. 13	M. 14	M. 15	M. 16	M. 17	M. 18	M. 19	M. 20
0	0	1	2	0	2	1	2	1	1
1	0	1	0	1	1	2	0	0	1
1	0	2	0	1	1	0	1	1	0
3	3	1	0	0	2	1	0	1	1

Las medias aritméticas de cada muestra son:

0.50	1.00	2.00	0.75	0.50
1.25	0.50	0.50	0.25	1.25
1.25	0.75	1.25	0.50	0.50
1.50	1.00	0.75	0.75	0.75

Entonces:

$$P_{\bar{X}}(0.25)) = P(\bar{X} = 0.25) = \frac{1}{20} = 0.05$$

$$P_{\bar{X}}(0.50)) = P(\bar{X} = 0.50) = \frac{6}{20} = 0.30$$

$$P_{\bar{X}}(0.75)) = P(\bar{X} = 0.75) = \frac{5}{20} = 0.25$$

$$P_{\bar{X}}(1)) = P(\bar{X} = 1) = \frac{2}{2} = 0.10$$

$$P_{\bar{X}}(1.25)) = P(\bar{X} = 1.25) = \frac{4}{20} = 0.20$$

$$P_{\bar{X}}(1.50)) = P(\bar{X} = 1.5) = \frac{1}{20} = 0.05$$

$$P_{\bar{X}}(2)) = P(\bar{X} = 2) = \frac{1}{20} = 0.05$$

Esta sería una aproximación a la distribución muestral del estadístico \bar{X} a partir de los datos de varias muestras.

Distribución de la media muestral

- La distribución del estadístico puede seguir un modelo preestablecido si se cumplen varias condiciones.
- Por ejemplo, supongamos que hemos tomado una muestra aleatoria simple de n observaciones de una v.a. X en una población de media μ_X y desviación típica σ_X .
- Representemos por X_1, X_2, \dots, X_n n observaciones independientes que forman una muestra aleatoria simple de ésta población.
- Cada una de las observaciones de la población son así mismo variables aleatorias con la misma distribución, esperanza y varianza que la población.

Estadístico media muestral

- Llamaremos **media aritmética o media muestral** de la muestra X_1, \dots, X_n a

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

- Bajo estas condiciones se cumple que:
 - a) $E(\bar{X}) = \mu_X$.
 - b) Es decir, el valor esperando de la media aritmética de la muestra es la media poblacional.
 - c) Por lo tanto el estadístico media muestral **estima** la media poblacional.
 - d) Dicho de otra forma, la esperanza de la distribución muestral de la media aritmética es la media poblacional.

- Que el valor esperado sea μ_X , no quiere decir que \bar{X} sea exactamente μ_X .
- Estudiemos la varianza de \bar{X} . Si X_1, \dots, X_n son independientes se cumple que:
 - a) $Var(\bar{X}) = \frac{1}{n}\sigma_X^2$.
 - b) Luego si n es suficientemente grande (o cuando $n \rightarrow \infty$) la varianza tenderá a estar muy próxima a cero.

Ejemplo

- No siempre tendremos independencia entre X_1, \dots, X_n .
- Por ejemplo en el caso en el que queramos averiguar cuántos votos afirmativos hay en una urna con 10 votos.
- Tenemos dos opciones para realizar la muestra aleatoria simple:
 - a) Tomar un voto al azar anotar su resultado y devolverlo a la urna, repetir el proceso 3 veces más. En este caso es un muestreo con reemplazamiento.
 - b) Tomar sucesivamente 4 votos de la urna sin reemplazarlos. En este caso es un muestreo sin reemplazamiento.

- En ambos casos la muestra obtenida es una muestra aleatoria pues todos los subconjuntos de individuos tienen igual probabilidad de ser elegidos.
- Pero en el primer caso tenemos independencia entre cada una de las observaciones mientras que en el segundo esto no es así.

- En la práctica se elige casi siempre el muestreo consistente en observar n individuos distintos.
- Si además, n es pequeño con respecto al tamaño de la población N , podemos suponer que las variables son prácticamente independientes.
- En caso contrario, tenemos que corregir la varianza de \bar{X} multiplicándola por lo que se denomina **factor de población finita** y tendremos que

$$\sigma_{\bar{X}}^2 = \text{Var}(\bar{X}) = \frac{1}{n} \sigma_X^2 \frac{N - n}{N - 1}$$

Tipificación de la media muestral. Teorema del Límite Central

- Frecuentemente utilizaremos la **expresión tipificada de la media muestral**:

$$Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu_X}{\frac{\sigma_X}{\sqrt{n}}}$$

- Además para tamaños muestrales grandes se cumple el importantísimo **Teorema del Límite Central** que afirma que la distribución de Z es aproximadamente una normal estándar.
- Este resultado es cierto **sea cual sea la distribución de la variable muestreada X** .
- Deducimos por tanto, usando las propiedades de la distribución normal, que la distribución de \bar{X} será aproximadamente normal estándar si n es suficientemente grande.

Propiedades de la media muestral

Sea X la variable aleatoria de interés que queremos observar en una cierta población. Supongamos que la esperanza poblacional es $E(X) = \mu_X$ y su varianza $Var(X) = \sigma_X^2$. Sea X_1, \dots, X_n una muestra aleatoria simple de dicha población:

Entonces se cumplen las propiedades siguientes:

- $\mu_{\bar{X}} = E(\bar{X}) = \mu_X$; el valor esperado de la media es la media poblacional.
- La varianza y la desviación típica de \bar{X} se pueden obtener con las siguientes fórmulas $\sigma_{\bar{X}}^2 = \frac{1}{n}\sigma_X^2$, \bar{X} es $\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}}$ que también recibe el nombre de **error estándar** de \bar{X} .
- Para poblaciones finitas de tamaño N si n es pequeño respecto de N hemos de aplicar el factor de corrección de población finita en el cálculo de la varianza y del error estándar de \bar{X} :

$$\sigma_{\bar{X}}^2 = \frac{1}{n}\sigma_X^2 \frac{N-n}{N-1}, \quad \sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

- Si la distribución de la población (X) es normal entonces la variable aleatoria:

$$Z = \frac{\bar{X} - \mu_X}{\frac{\sigma_X}{\sqrt{n}}}$$

es una normal estándar. O lo que es lo mismo \bar{X} es una normal con media μ_X y desviación típica $\sigma_{\bar{X}}$

- **Teorema del Límite Central:** Si la distribución de la población no es normal pero el tamaño muestral es suficientemente grande entonces por el T.L.C. la distribución de Z también se aproxima a una normal estándar y por lo tanto \bar{X} se aproxima a una normal con media μ_X y desviación típica $\sigma_{\bar{X}}$

Ejemplo

La altura media de un determinado arbusto se sabe que tiene una media poblacional 115 centímetros y una desviación típica poblacional de 25. Se toma una muestra aleatoria de 100 arbustos de esta especie.

- a) ¿Cuál es la probabilidad de que la media muestral de las de las alturas sea menor que 110 cm.?
- b) ¿Cuál es la probabilidad de que la media muestral de las alturas esté entre 113 cm. y 117 cm.?
- c) ¿Cuál es la probabilidad de que la media muestral de las alturas esté entre 114 cm.y 116 cm.?
- d) Sin hacer cálculos, razonar en cuál de los siguientes rangos resulta más probable que se encuentre la media muestral de las alturas.

113 cm.- 115 cm.

114 cm.- 116 cm.

115 cm.- 117 cm.

116 cm.- 118 cm.

- Supongamos que el número de individuos de la población de arbustos muy grande en relación al tamaño muestral $n = 100$.
- Sea X es la v.a. altura de un arbusto en cm.
- Con los datos del enunciado tenemos que $\mu_X = E(X) = 115$. y $\sigma_X = 25$.
- Sea X_1, \dots, X_{100} la muestra aleatoria simple de alturas.
- Tenemos que $\mu_{\bar{X}} = \mu_X = 115$ y $\sigma_{\bar{X}} = \frac{\sigma_X}{\sqrt{n}} = \frac{25}{\sqrt{100}} = 2.5$
- Además aplicando el T.L.C. se tiene que $Z = \frac{\bar{X} - \mu_X}{\frac{\sigma_X}{\sqrt{n}}} = \frac{\bar{X} - 115}{2.5}$ sigue aproximadamente una distribución normal estándar.

Solución:

- a) $P(\bar{X} \leq 110) = P(Z \leq \frac{110-115}{2.5}) =$
 $P(Z \leq -2) = F_Z(-2) = 1 - F_Z(2) = 1 - 0.9772 = 0.0228$
- b) $P(113 \leq \bar{X} \leq 117) = P(\frac{113-115}{2.5} \leq Z \leq \frac{117-115}{2.5}) =$
 $F_Z(0.8) - F_Z(-0.8) = 2F_Z(0.8) - 1 = 2(0.7881) - 1 = 0.5762$
- c) $P(114 \leq \bar{X} \leq 116) = P(\frac{114-115}{2.5} \leq Z \leq \frac{116-115}{2.5}) =$
 $F_Z(0.4) - F_Z(-0.4) = 2F_Z(0.4) - 1 = 2(0.6554) - 1 = 0.3108$
- d) La media aritmética de los precios \bar{X} sigue aproximadamente una distribución normal. Gráficamente el intervalo de mayor probabilidad será el que mayor área cubra bajo la curva normal (centrada en 115) y ese intervalo es 114 cm.-116 cm.

El estadístico proporción muestral

- La proporción muestral de un evento en una población vendrá generalmente asociada a una variable binomial. Veamos por qué con un ejemplo.
- Si tomamos una muestra de tamaño n , determinar el porcentaje de personas que recicla las basuras.
- Sea X_i la variable aleatoria que vale 0 si la persona i -ésima no recicla y vale 1 si la persona i -ésima recicla.
- Tendremos que $S = \sum_i^n X_i$ nos dará el número total de personas que reciclan.
- Llamamos p a la probabilidad de que una persona elegida al azar recicle. Nuestro objetivo es estimar p .
- Suponiendo independencia y que el valor de p no cambia para las personas, tendremos que la distribución de cada variable X_i será de Bernoulli de parámetro p y, por tanto, la distribución de S será binomial de parámetros n y p .

- ¿Será realmente binomial? Notemos que en la muestra no preguntaremos dos veces al mismo individuo, (el muestreo es sin reposición). Luego las observaciones no son exactamente independientes, pero si el tamaño de la población es grande respecto a la muestra podemos considerarlas así, ya que la probabilidad de respuesta afirmativa no cambia. (es despreciable el cambio).

Cálculo de la proporción muestral

- Sea S el número de éxitos en una muestra binomial de n observaciones, con probabilidad de éxito p .
- Entonces la proporción de éxitos en la muestra es:

$$\hat{p}_X = \frac{S}{n},$$

que recibe el nombre de **proporción muestral**.

Propiedades de la proporción muestral

Sea \hat{p}_X la proporción de éxitos en una muestra aleatoria de n observaciones. Entonces:

- $E(\hat{p}_X) = p$
- La distribución muestral de \hat{p}_X tiene varianza $\sigma_{\hat{p}_X}^2 = \frac{p(1-p)}{n}$ y por lo tanto su desviación típica es $\sigma_{\hat{p}_X} = \sqrt{\frac{p(1-p)}{n}}$ que recibe el nombre de **error estándar de la proporción muestral**.
- Si n es pequeño en relación al tamaño de la población N tenemos que aplicar el factor de corrección de población finita y entonces el error estándar de \hat{p}_X es

$$\sigma_{\hat{p}_X} = \sqrt{\frac{p(1-p)}{n}} \sqrt{\frac{N-n}{N-1}}.$$

Distribución de la proporción muestral

- Si el tamaño muestral es grande (por ejemplo $n > 30$ o mejor $n > 40$) entonces la variable

$$Z = \frac{\hat{p}_X - p}{\sigma_{\hat{p}_X}},$$

por el T.L.C., se distribuye aproximadamente como una normal estándar o lo que es lo mismo \hat{p}_X se distribuye aproximadamente como una normal con esperanza p_X y desviación típica $\sigma_{\hat{p}_X}$.

- **Observación** Notemos que si n crece el error estándar disminuye y entonces \hat{p} estará más cerca del valor real p .

Ejemplo: Se estima que el 20 % de los ciudadanos separa el papel y el cartón cuando tira su basura. Cierta día utilizaron $n = 180$ personas uno de los puntos de recogida de basuras. Consideremos esta personas como una muestra aleatoria de todos los ciudadanos.

- a) ¿Cuál será la media de la proporción muestral de personas que reciclan papel?
- b) ¿Cuál es la varianza de la proporción muestral?
- c) ¿Cuál es el error estándar de la proporción muestral?
- d) ¿Cuál es la probabilidad de que la proporción muestral sea mayor que 0.15?

Solución: El tamaño de la muestra es pequeño en relación al número total de ciudadanos (si la ciudad es grande). Tenemos que $p = 0.2$ (probabilidad de éxito en la venta).

- $E(\hat{p}_X) = p = 0.2$
- $\sigma_{\hat{p}_X}^2 = \frac{p(1-p)}{n} = \frac{0.2(1-0.2)}{180} = 0.0009$
- $\sigma_{\hat{p}_X} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{0.0009} = 0.03$
- Como n es grande entonces $Z = \frac{\hat{p}_X - p}{\sigma_{\hat{p}_X}} = \frac{\hat{p}_X - 0.2}{0.03}$ sigue aproximadamente una distribución normal estándar, entonces:
$$P(\hat{p}_X > 0.15) = 1 - P(\hat{p}_X \leq 0.15) = 1 - P(Z \leq \frac{0.15 - 0.2}{0.03}) = 1 - F_Z(-1.67) = F_Z(1.67) = 0.9525$$

La varianza muestral

- Sea X_1, \dots, X_n una muestra aleatoria simple de una población (X) con $E(X) = \mu_X$ y $Var(X) = \sigma_X^2$.
- Llamaremos **varianza muestral** al estadístico:

$$\tilde{S}_X^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}.$$

- $\tilde{S}_X = +\sqrt{\tilde{S}_X^2}$ recibe el nombre de desviación típica muestral.
- Denotaremos por $S_X^2 = \frac{n-1}{n} \tilde{S}_X^2$ y $S_X = +\sqrt{S_X^2}$.

Propiedades de la varianza muestral

- $S_X^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n} = \left(\frac{\sum_{i=1}^n X_i^2}{n} - \bar{X}^2 \right)$
- $E(S_X^2) = \frac{n-1}{n} \sigma_X^2$ (en el caso en que las variables X_i sean normales)
- $\tilde{S}_X^2 = \frac{n}{n-1} \left(\frac{\sum_{i=1}^n X_i^2}{n} - \bar{X}^2 \right)$
- $E(\tilde{S}_X^2) = \sigma_X^2$ (en el caso en que las variables X_i sean normales)

Distribución de la varianza muestral

Con las notaciones anteriores tenemos que:

- $E(\tilde{S}_X^2) = \sigma_X^2$ (en el caso en que las variables X_i sean normales)
- Si la distribución de la población es normal entonces la variable $\frac{(n-1)\tilde{S}_X^2}{\sigma_X^2}$ se distribuye según una ley conocida denominada χ_{n-1}^2 , que explicamos a continuación.

La distribución χ_n^2 (chi-cuadrado con n g.l.)

- Supongamos que X_1, X_2, \dots, X_n son n v.a. independientes y que $X_i \equiv N(0, 1)$
- Entonces:

$$X = X_1^2 + X_2^2 + \dots + X_n^2$$

es una v.a. que tiene distribución ji-cuadrado con n grados de libertad a la que denotaremos por χ_n^2 .

- La función de densidad de una χ_n^2 es :

$$f(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{(n/2)-1} e^{-x/2}$$

con $x \geq 0$ y $\Gamma(n/2) = \int_0^{+\infty} u^{(n/2)-1} e^{-u} du$ la llamada función gamma.

- Esta función de distribución está tabulada. También disponemos funciones de R que la calculan.

Gráfica de la función de densidad ji-cuadrado

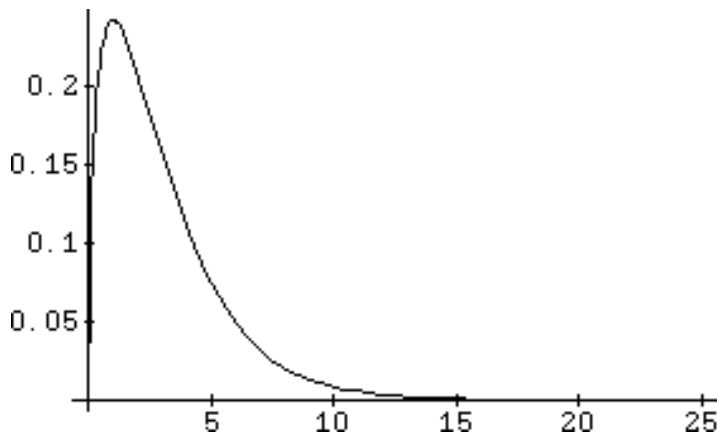


Figura: Gráfica de la función de densidad de una χ^2

Ejemplo:

El aumento del peso diario de un pollo de una granja sigue una distribución normal con desviación típica 1.7. Se toma una muestra de 12 pollos.

- a) Hallar la probabilidad de que la desviación típica muestral sea menor que 2.5.
- b) Hallar la probabilidad de que la desviación típica muestral sea mayor que 1.

Solución Sea X = el aumento del peso diario de un pollo. Sabemos que $\sigma_X^2 = (1.7)^2$ además como la distribución de la población es normal y $n = 12$ tenemos que $\frac{(n-1)\tilde{S}_X^2}{\sigma_X^2}$ sigue una distribución χ_{11}^2 .

- $P(\tilde{S}_X < 2.5) = P(\tilde{S}_X^2 < (2.5)^2) = P\left(\frac{(12-1)\tilde{S}_X^2}{(1.7)^2} < \frac{(12-1)(2.5)^2}{(1.7)^2}\right) = P(\chi_{11}^2 < 23.7889) \approx P(\chi_{11}^2 < 24.725) = 0.99.$
- $P(\tilde{S}_X > 1) = P(\tilde{S}_X^2 > 1) = P\left(\frac{(12-1)\tilde{S}_X^2}{1.7^2} > \frac{(12-1)1}{1.7^2}\right) = P(\chi_{11}^2 > 3.80623) \approx 1 - P(\chi_{11}^2 < 3.816) = 1 - 0.025 = 0.975$

Apuntes de Bioestadística.

R. Alberich y A. Mir

Departamento de Matemáticas e
Informàtica
Universitat Illes Balears

14 de julio de 2010

16 Muestreo Estadístico

17 Conceptos básicos

- Pasos en un estudio inferencial
- Diseño de experimentos, técnicas de muestreo
 - Muestreo aleatorio simple
 - Estadísticos y distribuciones muestrales
 - Distribución muestral de un estadístico
- Distribución de la media muestral
- Distribución de una proporción muestral
 - Distribución en el muestreo de \hat{p}_x
- Distribución muestral de la varianza muestral
 - Distribución en el muestreo de la varianza muestral

18 Inferencia estadística: estimación de parámetros y contraste de hipótesis.

Muestra aleatoria simple

- Supongamos que tenemos una población cuyo característica a estudiar de la misma viene dada por la variable aleatoria X .
- Diremos muestra aleatoria simple de tamaño n de la población anterior X a un conjunto de n variables aleatorias X_1, \dots, X_n independientes e idénticamente distribuidas todas con la misma distribución que la variable X .
- En la práctica, lo que tendremos serán unos valores determinados de la muestra, que llamaremos x_1, \dots, x_n .

Parámetro

- La distribución de la variable aleatoria X objeto de nuestro estudio puede depender de un parámetro θ o de varios.
- Por ejemplo, si X es binomial, los parámetros serán n y p ; si X es Poisson, el parámetro será λ ; si X es geométrica, el parámetro será p y si X es normal los parámetros serán μ y σ .
- El objetivo de la estadística inferencial es obtener información de dichos parámetros, en general desconocidos de la variable X .
- Dicha información se puede obtener de tres formas:
 - estimación puntual** Hallamos un valor aproximado del parámetro.
 - estimación por intervalo** Hallamos un intervalo donde el parámetro tiene una probabilidad “alta” de estar dentro de dicho intervalo.
 - contraste de hipótesis** Establecemos dos hipótesis para testear valores concretos del parámetro.

- Sean X_1, \dots, X_n n v.a. iid que forman una m.a.s. de una población.
- Un **estadístico** es una variable aleatoria que es función de la muestra.
- Un **estimador puntual** de un parámetro θ es un estadístico que da como resultado un único valor del que se espera que se aproxime a θ .
- Una **realización del estimador** $T(x_1, \dots, x_n) = \hat{\theta}$ en una muestra se llama **estimación puntual de parámetro**.

Estimadores básicos

Consideremos una m.a.s. X_1, \dots, X_n y una realización de la misma x_1, \dots, x_n los principales estimadores de los parámetros poblacionales que hemos visto son:

Parámetro Poblacional	Estimador(θ)	Estimación($\hat{\theta}$)
μ_X	$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$	$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$
σ_X	$\tilde{S}_X = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$	$\tilde{s}_X = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$
p	$\hat{p}_X = \frac{\sum_{i=1}^n X_i}{n}$	$\frac{\sum_{i=1}^n x_i}{n}$

Ejemplo

Consideremos una m.a.s. X_1, X_2, X_3, X_4, X_5 del lanzamiento de un dado ($n = 5$).

Una realización de esta muestra es $x_1 = 2, x_2 = 3, x_3 = 3, x_4 = 5, x_5 = 6$. Sabemos que, si el dado es perfecto, $\mu = 3.5$; el estadístico de esta muestra es

$$\bar{X} = \frac{X_1 + X_2 + X_3 + X_4 + X_5}{5}$$

y una estimación es

$$\bar{x} = \frac{x_1 + x_2 + x_3 + x_4 + x_5}{5} = \frac{2 + 3 + 3 + 5 + 6}{5} = \frac{19}{5} = 3.8$$

Ejemplo

Si queremos estimar la proporción de veces que sale 3 es $p_3 = \frac{1}{6}$
el estadístico es

$$\hat{p}_3 = \frac{\text{frec. de 3 en la muestra}}{5}$$

y una realización será $\frac{2}{5}$.

- ¿Qué estimador es mejor?
- Para decidirlo definiremos diversas propiedades de los estimadores.
- La más inmediata es pedirles que su valor esperado sea el valor del parámetro que estima.
- Dado $\hat{\theta}$ un estimador de un parámetro poblacional θ . Diremos que $\hat{\theta}$ es **insesgado** si $E(\hat{\theta}) = \theta$.
- En este caso la estimación puntual se dice que es insesgada.

Ejemplo

En el ejemplo del dado y p para cualquier muestra de tamaño n ,

$$X_1, \dots, X_n.$$

Se tiene que :

$$E(\bar{X}) = \mu_X$$

por lo tanto \bar{X} es un estimador insesgado de μ_X .

Algunos estimadores insesgados notables

Dada una m.a.s. La media, varianza y proporción muestrales son estimadores insesgados de sus correspondientes parámetros poblacionales. Es decir:

- $E(\bar{X}) = \mu.$
- $E(\hat{p}) = p.$
- $E(\tilde{s}^2) = \sigma^2.$

El sesgo de un estimador

Sea $\hat{\theta}$ un estimador puntual de un parámetro poblacional θ , llamaremos **sesgo** de $\hat{\theta}$ a:

$$\text{Sesgo}(\hat{\theta}) = E(\hat{\theta}) - \theta$$

Observación Diremos que un estimador es insesgado si y sólo si tiene sesgo cero.

La varianza y el error estándar de un estimador

- Una propiedad buena para un estimador es la carencia de sesgo.
- Pero podría suceder que tuviera una gran variabilidad.
- Entonces, aunque su valor central sea el verdadero valor del parámetro que se estima, una realización del estadístico podría estar lejos del verdadero valor del parámetro.
- Parece pues interesante emplear aquellos estimadores que tengan varianza más pequeña.
- A la desviación típica, es decir la raíz cuadrada de la varianza, de un estimador la denominaremos **error estándar** del estimador.

Eficiencia de un estimador

- Sean $\hat{\theta}_1$ y $\hat{\theta}_2$ dos estimadores de un parámetro poblacional θ obtenidos de la misma muestra.
- Diremos que $\hat{\theta}_1$ es más eficiente que $\hat{\theta}_2$ si $Var(\hat{\theta}_1) < Var(\hat{\theta}_2)$.
- O lo que es lo mismo si el error estándar de $\hat{\theta}_1$ es más pequeño que el error estándar de $\hat{\theta}_2$;

$$\sqrt{Var(\hat{\theta}_1)} < \sqrt{Var(\hat{\theta}_2)}.$$

Ejemplo:

- Sea $x_{(1)}, \dots, x_{(n)}$ la realización ordenada de menor a mayor de una muestra de tamaño n .
- Se define la mediana muestral como
$$Q_{0.5} = \begin{cases} x_{(\frac{n+1}{2})} & \text{si } n \text{ es impar} \\ \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2} & \text{si } n \text{ es par} \end{cases}$$
- Como vimos la mediana es también un valor de tendencia central, pero ¿es un buen estimador de μ ?
- Se puede demostrar que cuando la población tiene distribución normal con media μ y varianza σ_X^2 entonces $E(Me) = \mu$ y
$$Var(Me) = \frac{\pi}{2} \frac{\sigma_X^2}{n} \approx \frac{1.57 \sigma_X^2}{n}$$
- Luego si la muestra es de una población normal \bar{X} es más eficiente (un 57 % m de menos de varianza) que la Mediana.

Estimador más eficiente

Diremos que un estimador insesgado $\hat{\theta}$ del parámetro θ es el **estimador más eficiente** si no existe ningún otro estimador insesgado que tenga menor varianza que él (también se le denomina estimador insesgado de varianza mínima).

Algunos estimadores más eficientes

- Si la población es normal la media muestral es el estimador insesgado más eficiente de la media poblacional.
- Si la población es normal la varianza muestral es el estimador insesgado más eficiente de la varianza poblacional.
- Si la población es binomial la proporción muestral es el estimador insesgado más eficiente de la proporción poblacional.

Métodos para calcular estimadores. (Opcional)

Existen muchos métodos para el encontrar estimadores:

- Método de los momentos. Momento central de orden r
$$m_r = \frac{\sum_{i=1}^n (X_i - \bar{X})^r}{n}$$
- El de menor error cuadrático medio
$$E((\hat{\theta} - \theta)^2)$$
- Convergencia en probabilidad
$$P(|\hat{\theta}_n - \theta| < \epsilon) \rightarrow 1$$
- Estimadores máximo verosímiles.
- Otras técnicas, estimación robusta, remuestreo....

Función de verosimilitud

- Sea X una v.a. tal que su distribución (densidad o función de probabilidad) depende de un parámetro desconocido λ .
- En el caso discreto $P_X(x; \lambda)$ y en el continuo $f_X(x; \lambda)$.
- Sea X_1, \dots, X_n una m.a.s. de X (es decir son n v.a. iid como X).
- Sean x_1, x_2, \dots, x_n una realización de la muestra.
- Entonces la función de verosimilitud de la muestra es:
 - a) En el caso discreto $L(\lambda) = P_X(x_1; \lambda) \cdots P_X(x_n; \lambda)$
 - b) En el caso continuo $L(\lambda) = f_X(x_1; \lambda) \cdots f_X(x_n; \lambda)$

Estimador máximo verosímil

- Dada una función de verosimilitud $L(\lambda)$ de una muestra.
- Sea $\hat{\lambda} = g(x_1, \dots, x_n)$ el punto donde se alcanza en máximo de $L(\lambda)$ para la realización de la muestra x_1, \dots, x_n .
- Es decir $L(\hat{\lambda}) = \max_{\lambda} L(\lambda)$.
- El valor $\hat{\lambda}$ recibe el nombre de estimador máximo verosímil.
- Es decir **el estimador máximo verosímil es el valor del parámetro es el que máxima la probabilidad (densidad) de la muestra.**

El logaritmo de la función de verosimilitud

- En ocasiones es conveniente trabajar con el logaritmo de la función de verosimilitud.
- Ya que, al ser la función log creciente, el máximo de $\log(L(\lambda))$ y $L(\lambda)$ es el mismo y este último suele ser más fácil de calcular.

Ejemplo

- Sea X_1, \dots, X_n una muestra con observaciones independientes, de una población Bernouilli.
- Por ejemplo se analiza el genoma a 100 personas para saber si tienen una forma de un determinado alelo de un gen.
- Se anota un 1 si tienen ese alelo y cero en cualquier otro caso.
- Sea p la proporción poblacional de personas tienen ese alelo.
- Entonces

$$P(X_i = 1) = p \text{ y } P(X_i = 0) = 1 - p = q,$$

o lo que es lo mismo

$$P(X = x_i) = p^{x_i} q^{1-x_i} \text{ si } x_i = 0, 1$$

Ejemplo

- Como las observaciones son independientes. la función de verosimilitud es:

$$\begin{aligned}L(p) &= P_{X_1, \dots, X_n}(x_1, \dots, x_n) = P(X_1 = x_1, \dots, X_n = x_n) \\&= P(X_1 = x_1) \cdots P(X_n = x_n) = p^{x_1} q^{1-x_1} \cdots p^{x_n} q^{1-x_n} \\&= p^{\sum_{i=1}^n x_i} q^{\sum_{i=1}^n (1-x_i)} \\&= p^{\sum_{i=1}^n x_i} q^{n - \sum_{i=1}^n x_i}\end{aligned}$$

- Entonces el valor de p que hace máxima esta probabilidad es el más verosímil o el de máxima verosimilitud de esta muestra.
- El problema se reduce a estudiar qué valor de p maximiza

$$p^{\sum_{i=1}^n x_i} q^{n - \sum_{i=1}^n x_i} = p^{\sum_{i=1}^n x_i} (1 - p)^{n - \sum_{i=1}^n x_i}$$

Ejemplo

- Tomando logaritmos ...

$$\log \left(p^{\sum_{i=1}^n x_i} (1-p)^{n-\sum_{i=1}^n x_i} \right) = \sum_{i=1}^n x_i \log p + \left(n - \sum_{i=1}^n x_i \right) \log(1-p)$$

- Derivando respecto de p

$$\left(\sum_{i=1}^n x_i \right) \frac{1}{p} - \left(n - \sum_{i=1}^n x_i \right) \frac{1}{1-p} = 0$$

Ejemplo

- Despejando

$$(1 - p) \sum_{i=1}^n x_i - p(n - \sum_{i=1}^n x_i) = 0$$

- Por lo tanto

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

- Luego el estimador máximo verosímil de p es la proporción muestral \hat{p} , que es el que maximiza la función de verosimilitud $L(\hat{p}) \geq L(p)$.

Estimación por intervalos

- Una estimación por intervalos de un parámetro poblacional es una regla para determinar un rango o un intervalo donde, con cierta probabilidad, se encuentre el verdadero valor del parámetro.
- La estimación correspondiente se llama estimación por intervalo.
- Más formalmente, sea θ un parámetro, el intervalo (A, B) es un intervalo de confianza del $(1 - \alpha)100\%$ para el parámetro θ si

$$P(A < \theta < B) = 1 - \alpha.$$

- El valor $1 - \alpha$ recibe el nombre de **nivel de confianza**
- El valor $0 < \alpha < 1$ es la “cola” de probabilidad sobrante que normalmente se reparte por igual ($\alpha/2$) a cada lado del intervalo.
- Es frecuente que el nivel de confianza se den en tanto por ciento.

Intervalo de confianza para la media de una población normal: varianza poblacional conocida

En lo que sigue expondremos distintas maneras de calcular o aproximar intervalos de confianza para distintos parámetros.

- Sea X_1, \dots, X_n una m.a.s. de una v.a. X con distribución normal y $Var(X) = \sigma^2$ conocida.
- Busquemos un intervalo de confianza al *nivel de confianza* del 97.5 % para la media poblacional μ .
- Sabemos que, bajo estas condiciones, la variable $Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$ sigue una distribución normal estándar pues es una transformación lineal de una combinación lineal de variables normales e independientes..

Ejemplo

Comencemos calculando un intervalo centrado en 0 para esta Z que tenga probabilidad 0.975.

- $0.975 = P(-\delta < Z < \delta) = F_Z(\delta) - F_Z(-\delta) = 2F_Z(\delta) - 1$
- Entonces
$$F_Z(\delta) = \frac{1.975}{2} = 0.9875$$
- Consultando las tablas de la distribución normal estándar, entonces $F_Z(2.24) = 0.9875$ y por lo tanto $\delta = 2.24$

Ejemplo

- Luego $P(-2.24 < Z < 2.24) = 0.975$
- En resumen, hemos obtenido lo siguiente

$$0.975 = P\left(-2.24 < \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} < 2.24\right) =$$

- Por lo tanto

$$P\left(\bar{X} - 2.24 \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 2.24 \frac{\sigma}{\sqrt{n}}\right) = 0.975$$

Ejemplo

- Hemos encontrado un intervalo de confianza para μ .
- La probabilidad de que μ se encuentre en el intervalo

$$\left(\bar{X} - 2.24 \frac{\sigma}{\sqrt{n}}, \bar{X} + 2.24 \frac{\sigma}{\sqrt{n}} \right)$$

es 0.975.

- Luego es un intervalo de confianza con nivel de confianza 97.5 %
- Es decir en 97.5 de cada 100 ocasiones, en que tomemos una muestra de tamaño n y bajo estas condiciones, el verdadero valor de μ se encontrará en ese intervalo.

Ejemplo

- Supongamos que tenemos una muestra con $n = 16$ de una v.a. normal de forma que $\bar{x} = 20$, y la desviación típica poblacional es conocida $\sigma = 4$.
- Entonces un intervalo de confianza al 97.5 % para μ es:

$$\left(20 - \frac{(2.24)4}{\sqrt{16}}, 20 + \frac{(2.24)4}{\sqrt{16}} \right)$$

- La probabilidad con que el verdadero valor del parámetro μ se encuentra en el intervalo $(17.76, 22.24)$ es 0.975.
- O lo que es lo mismo $P(17.76 < \mu < 22.24) = 0.975$

Interpretación del intervalo de confianza

- En el 97.5 % de la muestras de tamaño 16 el verdadero valor del parámetro μ se encontrará dentro del intervalo correspondiente.

Fórmula general

- En general si tenemos una m.a.s. X_1, \dots, X_n de una población normal (representado por la v.a. X) con distribución normal de media μ y varianza conocida σ^2 .
- El intervalo de confianza para μ al nivel de confianza $(1 - \alpha) \cdot 100\%$ es

$$\begin{aligned}1 - \alpha &= P(z_{\alpha/2} < Z < z_{1-\alpha/2}) \\&= P(z_{\alpha/2} < \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} < z_{1-\alpha/2}) \\&= P(z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \bar{X} - \mu < z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}) \\&= P(\bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}})\end{aligned}$$

Resumen: Intervalo de confianza para μ : σ^2 conocida.

Condiciones:

- a) Población Normal con media μ y varianza σ^2 conocida
- b) Muestra aleatoria de tamaño n

Entonces el intervalo de confianza del $100(1 - \alpha)\%$ para μ es:

$$\left(\bar{X} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right)$$

- Donde $z_{\frac{\alpha}{2}}$ es el cuantil $\frac{\alpha}{2}$, es decir $P(Z \leq z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$, cuando Z tiene distribución normal estándar.
- $z_{1-\frac{\alpha}{2}}$ es el cuantil $1 - \frac{\alpha}{2}$, es decir $P(Z \leq z_{1-\frac{\alpha}{2}}) = 1 - \frac{\alpha}{2}$, cuando Z tiene distribución normal estándar.
- Notemos que $z_{\frac{\alpha}{2}} = -z_{1-\frac{\alpha}{2}}$

Ejemplo Tenemos un aparato para medir volúmenes de líquido. Para saber si está bien calibrado se toman 10 muestras consistentes en rellenar un recipiente, especialmente calibrado, de un litro. Se comprueban las mediciones obteniéndose los resultados de la siguiente tabla:

Volumen en litros	Frec. Absoluta	Volumen \times Frec. Absoluta
1.000	1	1.000
1.002	2	2.004
1.004	1	1.004
1.006	2	2.012
1.008	1	1.008
1.010	2	2.020
1.012	1	1.012
Total	10	10.06

Ejemplo

Supongamos que el volumen de líquido sigue una distribución normal con varianza poblacional conocida $\sigma^2 = 4$ calcular un intervalo de confianza al 90 % para la media del volumen.

Solución: Tenemos las siguientes condiciones:

- Población de volúmenes normal varianza $\sigma^2 = 4$ conocida
- Muestra aleatoria de tamaño $n = 10$

- Podemos aplicar la formula anterior, para $1 - \alpha = 0.9$.
- Entonces se tiene que $\alpha = 0.1$, $\frac{\alpha}{2} = 0.05$ y $1 - \frac{\alpha}{2} = 0.95$
- Calculamos la media aritmética de las observaciones
 $\bar{x} = \frac{10.06}{10} = 1.006$,
- Entonces el intervalo es

$$\left(1.006 + z_{0.05} \frac{2}{\sqrt{10}}, 1.006 + z_{1-0.05} \frac{2}{\sqrt{10}} \right).$$

Ejemplo

- Consultando las tablas de la normal $P(Z \leq 1.65) = 0.9505 \approx 0.95$ entonces $z_{0.95} = 1.65$, y $z_{0.05} = -1.65$
- Sustituyendo obtenemos que

$$z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} = 1.65 \frac{2}{\sqrt{10}} = 1.0435$$

$$z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} = -1.65 \frac{2}{\sqrt{10}} = -1.0435$$

- Por lo que el intervalo de confianza del 90 % para la media del volumen es : $(1.006 - 1.0435, 1.006 + 1.0435) = (-0.0375, 2.0495)$
- Lo que quiere decir que en el 90 % de la ocasiones en que tomemos una muestra de tamaño 10 el volumen medio estará comprendido entre -1.081 y 3.093 .
- Como se ve en este caso hay un abuso de la suposición de normalidad en la distribución del volumen. Y que el intervalo admite valores negativos.

Amplitud del intervalo de confianza

- Como de todos es conocido la amplitud (longitud) de un intervalo es la diferencia entre sus extremos superior e inferior.

- En el ejemplo anterior la amplitud A es

$$A = \bar{X} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} - \left(\bar{X} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right) = z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} = 2z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

- El *error* máximo, al nivel $(1 - \alpha)$, que cometemos al estimar μ por \bar{X} será la mitad de la amplitud del intervalo de confianza $z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$
- Si queremos calcular el tamaño n de la muestra para asegurarnos que el intervalo de confianza para μ al nivel $(1 - \alpha)$ tiene amplitud prefijada A (o un error $\frac{A}{2}$) se puede despejar así:

$$n = \left(2z_{1-\frac{\alpha}{2}} \frac{\sigma}{A} \right)^2$$

Observaciones:

- El intervalo está centrado en \bar{X} .
- Para n y $1 - \alpha$ fijos si la varianza poblacional aumenta entonces A aumenta.
- Para una varianza poblacional conocida y $1 - \alpha$ fijos si n aumenta entonces A disminuye.
- Para una varianza poblacional conocida y n fijos si $1 - \alpha$ aumenta entonces A aumenta.

Intervalo de confianza para la media poblacional: tamaños muestrales grandes

Condiciones:

- Población con media μ y varianza σ^2 conocida o si no se estima por \tilde{S}^2
- Muestra aleatoria de tamaño n grande (criterio $n \geq 30$)

Entonces el intervalo de confianza del $100(1 - \alpha)\%$ para μ es:

$$\left(\bar{X} + z_{\frac{\alpha}{2}} \frac{\tilde{S}}{\sqrt{n}}, \bar{X} + z_{1-\frac{\alpha}{2}} \frac{\tilde{S}}{\sqrt{n}} \right)$$

En caso de que σ sea conocida pondremos σ en lugar de \tilde{S}

Ejemplo:

- Se tomó una muestra de 147 expertos en informes de impacto ambiental y se les pidió que calificasen en una escala de 1 (totalmente en desacuerdo) a 10 (totalmente de acuerdo) la siguiente afirmación: “A veces utilizo técnicas de investigación que garantizan la obtención de los resultados que mi cliente o jefe desea”.
- La calificación media de la muestra fue 6.06 y la desviación típica muestral fue 1.43. Se pide calcular un intervalo de confianza al 90 % para la media de las puntuaciones.

Solución:

- El enunciado no nos asegura que la población sea normal pero como el tamaño de la población es grande podemos aplicar el resultado anterior.
- Tenemos $n = 147$, $\tilde{S} = 1.43$, $1 - \alpha = 0.9$ entonces $\frac{\alpha}{2} = 0.05$ y por lo tanto $z_{1-0.05} \approx 1.65$
- El intervalo para la media poblacional de las puntuaciones al nivel de confianza del 90 % es

$$\left(6.06 - 1.65 \frac{1.43}{\sqrt{147}}, 6.06 + 1.65 \frac{1.43}{\sqrt{147}} \right) = (5.8654, 6.2546)$$

Distribución t de Student

- Si queremos calcular un intervalo de confianza para μ en una población normal con varianza poblacional desconocida necesitamos una nueva distribución: la t de Student.
- Dada una muestra de n observaciones con media muestral \bar{X} y desviación típica muestral \tilde{S}_X procedente de una población normal con media μ la variable aleatoria:

$$t = \frac{\bar{X} - \mu}{\frac{\tilde{S}_X}{\sqrt{n}}}$$

sigue una distribución t de Student con $n - 1$ grados de libertad.

Propiedad

- La distribución t de Student es similar a la normal si el número de grados de libertad es grande. Su función de densidad es simétrica respecto al origen como la de la normal estándar.
- Es decir si t_ν es una v.a. que sigue la distribución t de Student con ν g.l. entonces:

$$P(t_\nu \leq -t) = 1 - P(t_\nu \leq t)$$

- Sea t_ν una v.a. que sigue una distribución t de Student con ν g.l. Denotaremos por $t_{\nu,\alpha}$ al valor para el que se verifica que:

$$P(t_\nu \leq t_{\nu,\alpha}) = \alpha.$$

- Luego $t_{\nu,\alpha}$ es el α cuantil de una t de Student con ν g.l. y $t_{\nu,\alpha} = -t_{\nu,1-\alpha}$.

Intervalo de confianza para la media de una población normal: varianza poblacional desconocida

Condiciones:

- Muestra aleatoria de n observaciones independientes.
- Población normal varianza desconocida

Entonces si \bar{X} y \tilde{S}_X son respectivamente la media y la desviación típica muestrales un intervalo de confianza al nivel $(1 - \alpha)100\%$ para la media de la población μ es:

$$\left(\bar{X} + t_{n-1, \frac{\alpha}{2}} \frac{\tilde{S}_X}{\sqrt{n}}, \bar{X} + t_{n-1, 1-\frac{\alpha}{2}} \frac{\tilde{S}_X}{\sqrt{n}} \right)$$

Siendo $t_{n-1, \frac{\alpha}{2}}$ y $t_{n-1, 1-\frac{\alpha}{2}}$ los cuantiles de una v.a. t_{n-1} con distribución t de Student con $n-1$ g.l., respectivamente.

Ejercicio

Demostrar que la probabilidad con que μ se encuentra en el intervalo anterior es $1 - \alpha$.

Ejemplo:

La empresa rayosX-print ofrece una impresora de altísima calidad para la impresión de radiografías. En su publicidad afirma que sus *cartuchos* imprimirán un promedio de 500 radiografías*; donde el asterisco remite a una nota a pie de página donde afirma que: “ Datos técnicos: Muestra mensual de tamaño $n = 25$ población supuesta normal nivel de confianza del 90 %”.

Una organización de radiólogos desea comprobar estas afirmaciones y toma también una muestra al azar de tamaño $n = 25$ obteniendo como media $\bar{x} = 518$ páginas y una desviación estándar $\tilde{S}_X = 40$. Comprobar que con esta muestra la media poblacional que afirma el fabricante cae dentro del intervalo de confianza del 90 %

Ejemplo

Solución: El problema se reduce a calcular, bajo las condiciones que afirma el fabricante el intervalo de confianza para μ con $\alpha = 0.1$.

Mirando en las tablas de la t de Student para $n - 1 = 24$ g.l. tenemos que

$$t_{n-1, 1 - \frac{\alpha}{2}} = t_{24, 1-0.05} = 1.71$$

El intervalo para la media al 90 % es

$$\left(518 - 1.71 \frac{40}{\sqrt{25}}, 518 + 1.71 \frac{40}{\sqrt{25}} \right) = (504.32, 531.68).$$

Es este caso la afirmación del fabricante queda contradicha por la muestra pues 500 cae fuera del intervalo. En cualquier caso se equivoca a favor del consumidor.

Intervalos de confianza para una proporción: Ejemplo

El procedimiento es similar al caso de las medias. Comencemos con un ejemplo.

- En una muestra aleatoria de 500 familias con niños en edad escolar se encontró que 340 introducen fruta de forma diaria en la dieta de sus hijos.
- Encontrar un intervalo de confianza del 95 % para la proporción actual de familias de esta ciudad con niños en edad escolar que incorporan fruta fresca de forma diaria en la dieta de sus hijos.
- Tenemos una población binomial donde los éxitos son las familias que aportan fruta de forma diaria a la dieta de sus hijos.
- Sea X el número de familias con hijos en edad escolar que aportan diariamente fruta a su dieta en una muestra aleatoria de tamaño n .

Ejemplo

- Entonces X sigue una distribución binomial con n repeticiones y probabilidad de éxito p (proporción poblacional de familias que aportan fruta a la dieta).
- Si llamamos $\hat{p}_X = \frac{X}{n}$ a la proporción muestral, sabemos que $Z = \frac{\hat{p}_X - p}{\sqrt{\frac{p(1-p)}{n}}}$ sigue aproximadamente una distribución normal estándar.
- Pero como es evidente no conocemos p así que no tenemos más remedio que aproximar el denominador

$$\sqrt{\frac{p(1-p)}{n}} \approx \sqrt{\frac{\hat{p}_X(1-\hat{p}_X)}{n}}$$

- Si la muestra es grande $Z = \frac{\hat{p}_X - p}{\sqrt{\frac{\hat{p}_X(1-\hat{p}_X)}{n}}}$ seguirá siendo aproximadamente normal estándar.

Intervalos de confianza para la proporción poblacional:(muestras grandes)

Condiciones:

- Una muestra aleatoria de tamaño n grande.
- Población Bernouilli con proporción de éxitos p (desconocida)

Bajo estas condiciones y si \hat{p}_X es la proporción de éxitos en la muestra, un intervalo de confianza al nivel $(1 - \alpha)100\%$ es

$$\left(\hat{p}_X + z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_X(1 - \hat{p}_X)}{n}}, \hat{p}_X + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_X(1 - \hat{p}_X)}{n}} \right)$$

Criterio: los intervalos de confianza anteriores son fiables si $n \geq 40$.

Observaciones

- El intervalo de confianza anterior está centrado en la proporción muestral.
- Cuando n crece se reduce la amplitud del intervalo de confianza.
- La amplitud del intervalo de confianza es $A = 2z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_X(1-\hat{p}_X)}{n}}$
- De la fórmula anterior no podemos determinar el tamaño de la muestral sin conocer \hat{p}_X así que nos podremos en el caso peor: El máximo de

$$\sqrt{\frac{\hat{p}_X(1-\hat{p}_X)}{n}}$$

se alcanza en $\hat{p}_X = 0.5$ y en este caso

$$\sqrt{\frac{0.5(1-0.5)}{n}}$$

por lo tanto en el peor de los casos.

$$n = \frac{0.25z_{1-\frac{\alpha}{2}}^2}{(A/2)^2}.$$

Observación

Por esto en las especificaciones o detalles técnicos de las encuestas se suele leer, por ejemplo:

“Universo población Balear mayor de 18 años. Encuesta telefónica, selección aleatoria, de tamaño mil, error en las proporciones $\pm 3\%$ con una confianza del 95 % supuesto que $p = q = \frac{1}{2}$ ”

Intervalo de confianza para la varianza de una población normal

- Recordemos que si tenemos una población normal con varianza σ^2 y una muestra aleatoria de tamaño n de esta población con varianza muestral \tilde{S}_X^2 entonces el estadístico

$$\chi_{n-1}^2 = \frac{(n-1)S_X^2}{\sigma^2}$$

sigue una distribución χ^2 con $n-1$ g.l.

- Notación** Si χ_ν^2 es una v.a. que tiene distribución χ^2 con ν g.l. denotaremos por $\chi_{\nu,\alpha}^2$ al valor que verifica:

$$P(\chi_\nu^2 \leq \chi_{\nu,\alpha}^2) = \alpha$$

- Es decir el cuantil α de una v.a. con distribución χ_ν^2 . Estos valores están tabulados para distintos g.l. en la tabla de la distribución χ^2 .

Ejemplo

- Sea χ_{10}^2 una v.a. que tiene distribución χ^2 con 10 g.l
- Entonces $\chi_{10,0.995}^2 = 25.19$ y $\chi_{10,0.005}^2 = 2.16$, es decir

$$P(\chi_{10}^2 \leq 25.19) = 0.995 \text{ y } P(\chi_{10}^2 \leq 2.16) = 0.005$$

- Además tendremos que

$$\begin{aligned} P(2.16 \leq \chi_{10}^2 \leq 25.19) &= P(\chi_{10}^2 \leq 25.19) - P(\chi_{10}^2 \leq 2.16) \\ &= 0.995 - 0.005 = 0.99 \end{aligned}$$

En general

- En general dado α entre 0 y 1 tendremos que

$$1 - \alpha = P(\chi_{\nu, \frac{\alpha}{2}}^2 \leq \chi_{\nu}^2 \leq \chi_{\nu, 1 - \frac{\alpha}{2}}^2)$$

- Si tenemos una muestra de tamaño n de una población normal con desviación típica muestral \tilde{S}_X^2 , dado un nivel de confianza $1 - \alpha$ tendremos que $\chi_{n-1}^2 = \frac{(n-1)\tilde{S}_X^2}{\sigma^2}$.

En general

- Entonces:

$$\begin{aligned}1 - \alpha &= P(\chi_{n-1, \frac{\alpha}{2}}^2 \leq \chi_{n-1}^2 \leq \chi_{n-1, 1-\frac{\alpha}{2}}^2) \\&= P(\chi_{n-1, \frac{\alpha}{2}}^2 \leq \frac{(n-1)S_X^2}{\sigma^2} \leq \chi_{n-1, 1-\frac{\alpha}{2}}^2) \\&= P\left(\frac{(n-1)\tilde{S}_X^2}{\chi_{n-1, 1-\frac{\alpha}{2}}^2} \leq \sigma^2 \leq \frac{(n-1)\tilde{S}_X^2}{\chi_{n-1, \frac{\alpha}{2}}^2}\right)\end{aligned}$$

- Luego, bajo estas condiciones, un intervalo de confianza para la varianza poblacional del $(1 - \alpha)100\%$ es

$$\left(\frac{(n-1)\tilde{S}_X^2}{\chi_{n-1, 1-\frac{\alpha}{2}}^2}, \frac{(n-1)\tilde{S}_X^2}{\chi_{n-1, \frac{\alpha}{2}}^2} \right).$$

Intervalo de confianza para la varianza de una población normal

Condiciones

- Población normal
- Muestra aleatoria de tamaño n con varianza muestral S_X^2

Entonces un intervalo de confianza del $(1 - \alpha)100\%$ es

$$\left(\frac{(n-1)\tilde{S}_X^2}{\chi_{n-1, 1-\frac{\alpha}{2}}^2}, \frac{(n-1)\tilde{S}_X^2}{\chi_{n-1, \frac{\alpha}{2}}^2} \right)$$

- Donde $\chi_{n-1, \frac{\alpha}{2}}^2$ es el valor que verifica

$$P(\chi_{n-1}^2 < \chi_{n-1, \frac{\alpha}{2}}^2) = \frac{\alpha}{2}$$

- Mientras que

$$\chi_{n-1, 1-\frac{\alpha}{2}}^2$$

es el valor tal que

$$P(\chi_{n-1}^2 \leq \chi_{n-1, 1-\frac{\alpha}{2}}^2) = 1 - \frac{\alpha}{2}$$

- Donde χ_{n-1}^2 es una v.a. que sigue una distribución χ^2 con $n - 1$ g.l.
- **Observación:** El intervalo de confianza para σ^2 no está centrado en \tilde{S}_X^2 .

Ejemplo

- Un índice de calidad de un reactivo químico es el tiempo que tarda en actuar.
- El estándar es que el tiempo no debe ser superior a los 30 segundos.
- Se supone que la distribución del tiempo de actuación del reactivo es aproximadamente normal. Se realizan 30 pruebas, que forman una muestra aleatoria, en las que se mide el tiempo de actuación del reactivo.
- Los tiempos fueron:
12, 13, 13, 14, 14, 14, 15, 15, 16, 17, 17, 18, 18, 19, 19, 25, 25, 26, 27, 30, 33, 34, 35, 40, 40, 51, 51, 58, 59, 83
- Se pide calcular un intervalo de confianza para la varianza al nivel 95 %.

Solución

- Sea X el tiempo de reacción. Haciendo los cálculos tenemos que (redondeando al segundo decimal):
 $\bar{X} = 28.37$ y $\tilde{s}_X = 17.37$
- Como $1 - \alpha = 0.95$ tenemos que $\frac{\alpha}{2} = 0.025$, entonces mirando en las tablas de la χ^2 (y redondeando también al segundo decimal)

$$\chi_{n-1, 1-\frac{\alpha}{2}}^2 = \chi_{29, 0.975}^2 = 45.72 \text{ y } \chi_{n-1, \frac{\alpha}{2}}^2 = \chi_{29, 0.025}^2 = 16.05.$$

- Por lo tanto un intervalo de confianza del 95 % para σ^2 es

$$\left(\frac{(30-1)(17.37)^2}{45.72}, \frac{(30-1)(17.37)^2}{16.05} \right) = (191.38, 545.16)$$

- Es decir $P(191.28 \leq \sigma^2 \leq 545.16) = 0.95$.

Parte V

Inferencia estadística. Contraste de hipótesis.

Introducción

- Hemos visto como puede estimarse un parámetro a partir de los datos contenidos en una muestra. Puede encontrarse una estimación puntual o bien una estimación por intervalo.
- Sin embargo muchos problemas en biología, bioquímica, o en cualquier ciencia o en economía o en la administración, requieren tomar una *decisión* es decir se debe aceptar o rechazar alguna afirmación sobre, por ejemplo, sobre el valor de un parámetro.

Introducción

- Esta afirmación recibe el nombre de *hipótesis* y el método estadístico de toma de decisión sobre la hipótesis recibe el nombre de prueba (o contraste) de hipótesis.
- Éste es uno de los aspectos más útiles de la inferencia estadística puesto que muchos problemas de toma de decisiones pueden plantearse en términos de contraste de hipótesis.

Ejemplo.

- a) Los responsables sanitarios del gobierno han determinado que el número de bacterias por centímetro cúbico de agua debe ser inferior o igual a 70 donde 70 es el máximo nivel aceptable para las aguas en las que se practica la recogida de almejas. La decisión se podría basar en varias muestras de agua.
- b) Un hospital recibe una partida de productos farmacéuticos. El encargado tiene orden de aceptar los envíos que contengan menos de un 5 % de unidades defectuosas. La decisión del encargado se podría basar en una muestra aleatoria de la partida.
- c) El promedio de proteínas en sangre en un adulto sano es de 7.25 gr/dL. En un análisis de sangre el técnico debe decir si la media del paciente es igual o distinta de ese valor.
- d) Una fábrica de abonos con nitratos afirma que su uso producirá un aumento de masa (medida en Kg. de materia seca) por hectárea y año. Los agricultores quieren tener garantías de que esto es cierto.

Hipótesis estadística

- Una hipótesis estadística es una afirmación que se realiza sobre los parámetros o las distribuciones de una o más poblaciones
- Las hipótesis estadística se contrastan una contra otra. Habitualmente las denominaremos **Hipótesis nula** H_0 e **hipótesis alternativa** H_1 .

Ejemplo

- Un fabricante de sobrasada asegura en su etiqueta que sus piezas pesan 200 gr.
- Un fabricante de la competencia sospecha que el peso es inferior al que figura en la etiqueta para ello toma una muestra aleatoria de sobrasadas y las pesa.
- Sea μ el contenido medio en gramos de la población de sobrasadas.

Ejemplo

- El contraste de interés para desde el punto de vista económico del fabricante es:

$$\begin{cases} H_0 : \mu = 200 \\ H_1 : \mu > 200 \end{cases}$$

No le interesa regalar gramos de sobrasada...

- Al consumidor sólo le interesa contrastar

$$\begin{cases} H_0 : \mu = 200 \\ H_1 : \mu < 200 \end{cases} ,$$

pues sólo quiere decidir si el peso es inferior al declarado.

- Pero si es el encargado del control de la producción le interesará contrastar

$$\begin{cases} H_0 : \mu = 200 \\ H_1 : \mu \neq 200 \end{cases}$$

Pues no debe engañar al consumidor pero tampoco quiere darle más peso gratis.

Tipos de hipótesis sobre parámetros

- $H : \theta = \theta_0$ **hipótesis simple** (en caso contrario compuesta).
- $H : \theta > \theta_0$ o $H : \theta < \theta_0$ **hipótesis unilateral**.
- $H : \theta \neq \theta_0$ **hipótesis bilateral**.

Resumiendo: Un contraste de hipótesis consiste en plantear una **hipótesis nula** y una **alternativa**.

$$\begin{cases} H_0 : \text{hipótesis nula} \\ H_1 : \text{hipótesis alternativa} \end{cases}$$

y generar un **regla de decisión** para **aceptar** la hipótesis nula o **rechazarla** en favor de la alternativa a partir de la información contenida en una muestra.

Ejemplo.

Supongamos que queremos decidir si una moneda está bien balanceada. Para ello lanzamos la moneda 100 veces obteniéndose X caras. Sea p la probabilidad de cara en esta moneda, queremos contrastar:

$$\begin{cases} H_0 : p = 0.5 \\ H_1 : p \neq 0.5 \end{cases}$$

Una regla podría ser aceptar H_0 contra H_1 si X no es muy distinto de 50 por ejemplo si $48 \leq X \leq 52$.

En lo que sigue definiremos los elementos necesarios para estudiar qué reglas (regiones) de rechazo son las más adecuadas para distintos tipos de contrastes.

Tipos de Error en un contraste

Cuando realizamos un contraste de hipótesis pueden darse las situaciones que detallamos en la tabla siguiente:

Decisión	Estados de la naturaleza	
	H_0 cierta	H_0 falsa
Aceptar H_0	Dec. correcta Prob= $1 - \alpha$	Error tipo II Prob= β
Rechazar H_0	Error tipo I Prob= α	Dec. correcta Prob = $1 - \beta$

Probabilidades de los Errores de Tipo I y II

- La probabilidad de Error Tipo I es
 $P(\text{Error Tipo I}) = P(\text{Rechazar } H_0 / H_0 \text{ cierta}) = \alpha$ y recibe el nombre de **nivel de significación** del contraste.
- La probabilidad de Error Tipo II es
 $P(\text{Error Tipo II}) = P(\text{Aceptar } H_0 / H_0 \text{ falsa}) = \beta$ el valor $1 - \beta$ recibe el nombre de **potencia** del contraste.

- En ocasiones daremos los niveles de significación y la potencia en tantos por cien, así un nivel de significación del 5 % implica que $\alpha = 0.05$
- Lo ideal es encontrar aquella regla de rechazo de H_0 que tenga menor probabilidad de Error Tipo I α .
- Pero que también tenga menor probabilidad de Error Tipo II β o lo que es lo mismo mayor potencia $1 - \beta$.

- Lo que sucede es que si modificamos la regla de rechazo para que disminuya α entonces aumentamos β .
- Buscaremos reglas de decisión que para un α fijo nos den un β lo más pequeño posible.
- Lo que se hace normalmente es fijar α y esto nos da la región crítica y luego, si es posible, controlar el tamaño de la muestra n para obtener la mayor potencia y por lo tanto el menor Error de Tipo II al menor coste.
- Para bajar el Error de Tipo II se aumenta el tamaño de la muestra. Excede el tiempo de este curso el cálculo del tamaño de la muestra para fijar el valor del Error de tipo II. Consultad la bibliografía.
- En resumen: Si el investigador fija un nivel de significación y un tamaño n , obtiene una regla de decisión que fija un Error de Tipo II.

Terminología

Resumamos los conceptos vistos hasta ahora:

- Hipótesis nula H_0 : Es la hipótesis que se desea aceptar si no hay prueba de que es falsa.
- Hipótesis Alternativa H_1 : Es la hipótesis frente a la que se contrasta la hipótesis nula y que se acepta si se rechaza la nula.
- Hipótesis simple: Es la que especifica un sólo valor para el parámetro a contrastar.
- Hipótesis compuesta: Es la que especifica un rango de valores para el parámetro a contrastar.
- Alternativa unilateral: Es una H_1 compuesta formada por un semi intervalo es decir $\theta > \theta_0$ o $\theta < \theta_0$.
Alternativa bilateral: Es aquella H_1 compuesta que es el complementario de una H_0 simple.

Terminología

- Decisión de un contraste de hipótesis: puede ser aceptar o rechazar la hipótesis nula lo que se hace en función de una regla de decisión que recoge la información de una muestra.
- Error de Tipo I: Se comete cuando se rechaza H_0 siendo cierta. Su probabilidad se denota por α .
Error Tipo II: Se comete cuando se acepta una H_0 falsa. Su probabilidad se denota por β .
- Nivel de significación α : Es la probabilidad de cometer un Error Tipo I, es decir, $\alpha = P(\text{Error Tipo I}) = P(\text{Rechazar } H_0 / H_0 \text{ cierta})$
- Potencia de un contraste: Es la probabilidad de rechazar una hipótesis nula que es falsa. Entonces la potencia es
$$P(\text{Rechazar } H_0 / H_0 \text{ es falsa}) = 1 - P(\text{Aceptar } H_0 / H_0 \text{ es falsa}) = 1 - P(\text{Error Tipo II}) = 1 - \beta$$

¿Inocente o culpable?

- La decisión de aceptar o rechazar una hipótesis nula se asemeja al concepto de declarar a un acusado en juicio inocente o culpable.
- El acusado es la hipótesis nula H_0 .
- Las pruebas son los elementos de la muestra.
- Si el jurado no encuentra suficientes las pruebas tiene que declarar inocente al acusado (Aceptar H_0).
- Sólo en el caso en que las pruebas sean lo suficientemente incriminatorias condenará al culpable y se aceptará la hipótesis alternativa.
- El jurado siempre corre el riesgo de declarar culpable a un inocente cometiendo un Error de Tipo I,
- O de declarar inocente a un culpable cometiendo un Error de Tipo II.
- Desde este punto de vista es más conveniente controlar el Error de Tipo I pues es mejor declarar inocente a un culpable que culpable a un inocente.

Ejemplo de un contraste de hipótesis para la media de una distribución normal: varianza poblacional conocida

- En lo que sigue, comenzando por esta sección, daremos distintos contrastes de hipótesis para la media de una población.
- Para contrastar las hipótesis dispondremos de una m.a.s. de n observaciones X_1, \dots, X_n . En este caso procedentes de una distribución normal con media μ y varianza σ^2 .
- Supondremos que la varianza es conocida.
- Consideremos el contraste:

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu > \mu_0 \end{cases}$$

- La regla de rechazo se basará en observar si la media aritmética \bar{X} es suficientemente mayor que valor μ_0 . Si es así rechazaremos la hipótesis nula.
- Como sabemos que bajo estas condiciones y si H_0 (es decir $\mu = \mu_0$) es cierta

$$Z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

sigue una distribución normal estándar.

- Rechazar H_0 si \bar{X} es muy alta es equivalente a obtener un valor alto del estadístico de contraste

$$Z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}.$$

- Entonces la regla consiste en rechazar H_0 si Z es mayor que un cierto umbral.
- Sabemos que $\alpha = P(\text{Rechazar } H_0 / H_0 \text{ cierta}) = P(Z > \text{umbral} / \mu = \mu_0) = P(Z > z_{1-\alpha})$ cuando Z es una normal estándar.
- Luego para que el nivel de significación del contraste sea α la regla de rechazo viene dada por la región crítica

$$Z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} > z_{1-\alpha}$$

- Estadístico de contraste: es el que nos permite definir una regla de rechazo de H_0 .
- Región crítica o región de rechazo: es aquel rango de valores tales que si el estadístico de contraste está entre ellos se rechaza H_0 .
- Región de aceptación: Es el complementario de la región crítica.

Condiciones:

- Población normal de media μ y varianza σ^2 conocida

Un contraste al nivel de significación α para las hipótesis:

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu > \mu_0 \end{cases}$$

Tiene por regla de decisión:

Rechazar H_0 si

$$Z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} > z_{1-\alpha}.$$

Valor crítico o p -valor.

- Llamaremos valor crítico o p -valor (p -value) al error tipo I máximo que cometeríamos con el valor obtenido del estadístico de contraste.
- Al realizar el contraste, si establecemos un valor de significación α menor que el p -valor, aceptaríamos la hipótesis nula H_0 y en caso de que sea mayor que el p -valor, la rechazaríamos.
- Por ejemplo si el p -valor es 0.05, significaría que el error tipo I valdría como máximo 0.05 para poder aceptar la hipótesis nula.
- Por tanto, p -valores grandes significa que tenemos mucho margen para aceptar la hipótesis nula y de aquí que se “sospeche” que ésta es cierta.
- Por tanto, si el p -valor es grande (mayor que 0.1) aceptaremos la hipótesis nula y en caso contrario, la rechazaremos.
- p -valores entre 0.05 y 0.1 representan valores moderadamente grandes y que requieren estudios posteriores de cara a tomar una decisión.

El Método de los seis pasos

- 1) Establecer la hipótesis nula H_0 , por ejemplo $\theta = \theta_0$
- 2) Establecer la hipótesis alternativa H_1 que podrá ser $\theta > \theta_0$, $\theta < \theta_0$ o $\theta \neq \theta_0$.
- 3) Seleccionar un nivel de significación α
- 4) Seleccionar el estadístico apropiado para la prueba y establecer la región crítica o región de rechazo. Si la decisión se basa en un p -valor, como veremos, no es necesario calcular la región crítica.
- 5) Calcular el valor del estadístico de contraste a partir de los datos muestrales.
- 6) Decidir: rechazar H_0 si el valor del estadístico de contraste cae dentro de la región crítica o si el p -valor es menor o igual que el nivel de significación prefijado α ; en caso contrario no rechazar H_0 .

Ejemplo.

- Una muestra aleatoria de 100 muertes registradas en un cierto país durante 1998 dio una vida promedio de 71.8 años.
- Suponiendo que la desviación típica poblacional es de 8.9 años, decidir si la vida promedio es, hoy en día, mayor que 70 años.
- Utilizad un nivel de significación (α) del 0.05 y suponer que la duración de la vida se distribuye aproximadamente normal.

Solución:

Sigamos los seis pasos:

- 1) $H_0 : \mu = 70 \text{ años.} (\mu_0 = 70)$
- 2) $H_1 : \mu > 70 \text{ años.}$
- 3) $\alpha = 0.05$
- 4) Bajo estas condiciones, población normal, $\sigma^2 = 8.9^2$ conocida y una muestra de tamaño $n = 100$ la región crítica para estas hipótesis es:

$$Z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} > z_{1-0.05} = 1.64$$

- 5) Cálculo del estadístico de contraste: $\bar{x} = 71.8$ años, $\sigma = 8.9$ años. Entonces el estadístico de contraste es:

$$Z = \frac{71.8 - 70}{\frac{8.9}{\sqrt{100}}} = 2.02$$

- 6) Decisión: Como $Z = 2.02 > 1.64$ resulta que el valor del estadístico de contraste cae dentro de la región crítica, luego a partir de esta muestra no podemos aceptar (H_0) que la vida promedio es de 70 años contra que es mayor de 70 años (H_1) al nivel de significación $\alpha = 0.05$

Ejemplo

En el ejemplo anterior calcular el p -valor e interpretarlo.

- Para calcular el p valor tenemos que buscar aquel nivel de significación α más pequeño para el que se rechaza la hipótesis nula.
- Para ello igualamos el valor del estadístico de contraste $Z = 2.02$ al umbral de la región de rechazo, es decir:

$$2.02 = z_{1-\alpha}$$

- Consultando las tablas de la distribución normal estándar obtenemos que $1 - \alpha = 0.9783$ luego $\alpha = 0.0217$.

Interpretación del p -valor

La interpretación de este valor es la siguiente:

- Rechazaremos (H_0) que la vida promedio es de 70 años contra que es mayor de 70 años (H_1) para todos los niveles de significación $\alpha > 0.0217$.
- Es decir la evidencia es más grande que el nivel de significación del ejemplo anterior.
- Y por lo tanto como el p -valor es 0.0217 no podemos aceptar H_0 para el nivel de significación $\alpha = 0.05$.

Ejemplo

Si en el contraste anterior utilizamos las hipótesis:

$$\begin{cases} H_0 : \mu \leq \mu_0 \\ H_1 : \mu > \mu_0 \end{cases}$$

Todavía tendríamos más evidencia para rechazar la hipótesis nula.

Entonces la región de contraste es la misma que en el caso $H_0 : \mu = \mu_0$

Reglas de decisión para contraste de la media de una población normal: varianza poblacional conocida

Condiciones:

- Una muestra aleatoria simple de una población normal de media μ y varianza σ^2 conocida.

Un contraste al nivel de significación α para las hipótesis:

- $$\begin{cases} H_0 : \mu = \mu_0 & (\text{o } H_0 : \mu \leq \mu_0) \\ H_1 : \mu > \mu_0 \end{cases}$$

Tiene por regla de decisión:

Rechazar H_0 si

$$Z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} > z_{1-\alpha}.$$

$$\begin{cases} H_0 : \mu = \mu_0 & (\text{o } H_0 : \mu \geq \mu_0) \\ H_1 : \mu < \mu_0 \end{cases}$$

Tiene por regla de decisión:

Rechazar H_0 si

$$Z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} < z_{\alpha}.$$

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases}$$

Tiene por regla de decisión:

Rechazar H_0 si

$$Z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} > z_{1-\frac{\alpha}{2}} \text{ o } Z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} < z_{\frac{\alpha}{2}}.$$

- Si no conocemos la distribución de la población o bien no es normal pero tenemos un tamaño muestral grande, podemos prescindir de la condición de normalidad de la población y aplicar las mismas reglas de rechazo para la hipótesis nula que en el caso anterior.
- Además si σ^2 es desconocida se puede sustituir por la desviación típica muestral \tilde{S}^2 . Criterio: si $n \geq 30$ podemos aplicar esta aproximación.

Ejemplo.

- Una organización ecologista afirma que el peso de medio de los individuos adultos de una especie marina ha disminuido drásticamente.
- Se sabe por los datos históricos que el peso medio poblacional es μ es 460 gr.
- Una muestra aleatoria de 36 individuos de esta especie tiene una media muestral de 420 gr.y una desviación típica muestral de 0.303.
- ¿Podemos afirmar, con un nivel de significación del 5 % y con estos datos que el peso medio es inferior a 460 gr.?

Solución

- Nadie nos asegura que la población es normal.
- Resulta que σ es desconocida.
- Pero como $n = 36$ podemos utilizar las regiones de rechazo anteriores sustituyendo σ por \tilde{s} .

Sigamos los seis pasos:

Solución

- 1) $H_0 : \mu = 460 \text{ gr. } (\mu_0 = 460)$
- 2) $H_1 : \mu < 460 \text{ gr.}$
- 3) $\alpha = 0.05$
- 4) Bajo estas condiciones, como $n \geq 30$, σ es desconocida pero la aproximamos por $\sigma \approx \tilde{s} = 11.9$. Entonces la región crítica para estas hipótesis es:

$$Z = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} < z_{0.05} = -1.64$$

- 5) Cálculo del estadístico de contraste: $\bar{x} = 420$, $\tilde{s} = 0.303$ entonces:

$$Z = \frac{420 - 460}{\frac{0.303}{\sqrt{36}}} = -2.02$$

- 6) Decisión: Como $Z = -2.02 < -1.64$ resulta que el valor del estadístico de contraste cae en la región crítica, luego a partir de esta muestra rechazamos (H_0) que el peso medio es de 460 gr. contra que es menor de 460 gr. (H_1) al nivel de significación $\alpha = 0.05$.

Conclusión: El peso medio de esta especie este año es significativamente inferior a 460 gr., por lo que podríamos aceptar los resultados de la asociación ecologista con esta muestra y a este nivel de significación.

Reglas de decisión para el contraste de una media: Tamaños muestrales grandes

Son las misma que las del caso de σ conocida pero estimando ésta por \tilde{S} .

- En el caso que tengamos una población normal, desconozcamos la varianza y no tengamos un tamaño muestral n grande utilizaremos el estadístico

$$t_{n-1} = \frac{\bar{X} - \mu_0}{\frac{\tilde{S}}{\sqrt{n}}}$$

- Este estadístico sigue una distribución t de Student con $n - 1$ g.l.
- Las regiones críticas serán similares a las de muestras grandes pero sustituyendo los valores de la normal estándar por los correspondientes valores de la t_{n-1} .

Reglas de decisión para el contraste de una media de una distribución normal: varianza poblacional desconocida

Condiciones:

- Muestra aleatoria de n observaciones población normal con media μ y varianza desconocida.

Entonces una contraste al nivel de significación α para las hipótesis:

•

$$\begin{cases} H_0 : \mu = \mu_0 \text{ o } H_0 : \mu \leq \mu_0 \\ H_1 : \mu > \mu_0 \end{cases}$$

Tiene por regla de decisión:

Rechazar H_0 si

$$t_{n-1} = \frac{\bar{X} - \mu_0}{\frac{\tilde{S}}{\sqrt{n}}} > t_{n-1, 1-\alpha}.$$

$$\begin{cases} H_0 : \mu = \mu_0 \text{ o } H_0 : \mu \geq \mu_0 \\ H_1 : \mu < \mu_0 \end{cases}$$

Tiene por regla de decisión:

Rechazar H_0 si

$$t_{n-1} = \frac{\bar{x} - \mu_0}{\frac{\tilde{s}}{\sqrt{n}}} < t_{n-1, \alpha}.$$

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases}$$

Tiene por regla de decisión:

Rechazar H_0 si

$$t_{n-1} = \frac{\bar{x} - \mu_0}{\frac{\tilde{s}}{\sqrt{n}}} > t_{n-1, 1-\frac{\alpha}{2}} \text{ o } t_{n-1} = \frac{\bar{x} - \mu_0}{\frac{\tilde{s}}{\sqrt{n}}} < t_{n-1, \frac{\alpha}{2}}.$$

Ejemplo.

- Se espera que el nivel de colesterol en plasma de unos enfermos bajo tratamiento se distribuya normalmente con media 220 mg/dL.
- Se toma una muestra de 9 enfermos, obteniéndose los siguientes resultados:

203, 229, 215, 220, 223, 233, 208, 228, 209

```
> options(width = 60)
> colesterol <- c(203, 229, 215, 220, 223, 233,
+               208, 228, 209)
```

- Contrastar la hipótesis de que esta muestra proviene de una población con media 220 mg./dL al nivel de significación del 10%.

Solución

Calculemos los parámetros de la muestra:

```
> media <- round(mean(colesterol), 4)
```

```
> media
```

```
[1] 218.6667
```

```
> n <- length(colesterol)
```

```
> n
```

```
[1] 9
```

```
> suma.cuadrados <- sum(colesterol^2)
```

```
> suma.cuadrados
```

```
[1] 431222
```

```
> varianza.muestral <- round((n/(n - 1)) * (suma.cuadrados/n -  
+      media^2), 4)
```

```
> varianza.muestral
```

```
[1] 110.7336
```

Solución

- La media muestral es

$$\bar{x} = \frac{203+229+215+220+223+233+208+228+209}{9} = 218.6667.$$

- La varianza muestral es

$$\begin{aligned}\tilde{s}^2 &= \frac{9}{8} (203^2 + 229^2 + \dots + (209)^2) - 218.6667^2 = \\ &= \frac{9}{8} \left(\frac{431222}{9} - 47815.1257 \right) = 110.7336\end{aligned}$$

- La población es normal y como σ es desconocida y n pequeño tendremos que utilizar como estadístico de contraste la t de Student.

Solución

- 1) $H_0 : \mu = 220$
- 2) $H_1 : \mu \neq 220$
- 3) $\alpha = 0.1; \frac{\alpha}{2} = 0.05.$
- 4) Bajo estas condiciones, población normal, σ desconocida y una muestra de tamaño $n = 9$ pequeño, la región crítica para estas hipótesis es:
 $t_8 > t_{8,1-0.05} = 1.86$ o $t_8 \leq t_{8,0.05} = -1.86$

Solución

- 5) Cálculo del estadístico de contraste: $\bar{x} = 218.67$,
 $\tilde{s} = \sqrt{110.75} = 10.52$ entonces: $t_{n-1} = \frac{218.67 - 220}{\frac{10.52}{\sqrt{9}}} = -0.38$
- 6) Decisión: Como $t_8 = -0.38 \not> 1.86$ $t_8 = -0.38 \not< -1.86$ resulta que el valor del estadístico de contraste cae fuera de la región crítica. Con esta muestra no podemos rechazar (H_0) que el nivel medio de colesterol en mg./dL en plasma sea igual a 220 contra que es distinto, con un nivel de significación del 10 %.

Contraste para la varianza de una población normal.

- Basaremos los contrastes para la varianza de una población normal en el estadístico muestral \tilde{S}^2 .
- Más concretamente en el estadístico $\chi_{n-1}^2 = \frac{(n-1)\tilde{S}^2}{\sigma^2}$, del que sabemos que sigue una distribución χ^2 con $n - 1$ g.l. si la población es normal.
- Claro que no conocemos el valor de σ^2 pero bajo la hipótesis nula $H_0 : \sigma = \sigma_0$ tendremos que $\chi_{n-1}^2 = \frac{(n-1)\tilde{S}^2}{\sigma_0^2}$ tendrá también una distribución χ^2 con $n - 1$ g.l.
Las condiciones del test se reumen a continuación:

Resumen reglas de decisión para el contraste de la varianza de una población normal

Condiciones³:

- Muestra aleatoria de n observaciones de una población normal.

Entonces una contraste al nivel de significación α para las hipótesis:

•

$$\begin{cases} H_0 : \sigma = \sigma_0 \text{ o } H_0 : \sigma \leq \sigma_0 \\ H_1 : \sigma > \sigma_0 \end{cases}$$

Tiene por regla de decisión:

Rechazar H_0 si

$$\chi_{n-1}^2 = \frac{(n-1)\tilde{s}^2}{\sigma_0^2} > \chi_{n-1,1-\alpha}^2.$$

³En realizan los contratos son $\frac{\sigma}{\sigma_0} = 1$ pero para simplificar se ponen como $\sigma = \sigma_0$.

$$\begin{cases} H_0 : \sigma = \sigma_0 \text{ o } H_0 : \sigma \geq \sigma_0 \\ H_1 : \sigma < \sigma_0 \end{cases}$$

Tiene por regla de decisión:

Rechazar H_0 si

$$\chi_{n-1}^2 = \frac{(n-1)\tilde{s}^2}{\sigma_0^2} < \chi_{n-1,\alpha}^2.$$

$$\begin{cases} H_0 : \sigma = \sigma_0 \\ H_1 : \sigma \neq \sigma_0 \end{cases}$$

Tiene por regla de decisión:

Rechazar H_0 si

$$\chi_{n-1}^2 = \frac{(n-1)\tilde{s}^2}{\sigma_0^2} > \chi_{n-1,1-\frac{\alpha}{2}}^2 \text{ o } \chi_{n-1}^2 = \frac{(n-1)\tilde{s}^2}{\sigma_0^2} < \chi_{n-1,\frac{\alpha}{2}}^2.$$

Ejemplo.

- Se han medido los siguientes valores en miles de personas para la audiencia de un programa de radio en $n = 10$ días:

521, 742, 593, 635, 788, 717, 606, 639, 666, 624.

- Contrastar que la varianza de la audiencia es 6400 al nivel de significación del 5 %,suponiendo que la población sea normal.

Solución

- 1) $H_0 : \sigma^2 = 6400$.
- 2) $H_1 : \sigma^2 \neq 6400$; ya que no se especifica qué alternativa se pide.
- 3) Nivel de significación $\alpha = 0.05$.
- 4) Bajo estas condiciones podemos utilizar como región crítica
$$\chi_9^2 = \frac{(9)\tilde{S}^2}{6400} > \chi_{9,1-0.025}^2 = 19.02 \text{ o } \chi_9^2 = \frac{(9)\tilde{S}^2}{6400} < \chi_{9,0.025}^2 = 2.70$$

- 5) $\bar{x} = 653.10$ y $\tilde{s}^2 = 6111.66$ entonces $\chi_9^2 = \frac{(9)6111.66}{6400} = 8.59452$
- 6) Como $\chi_9^2 = 8.59452 \not\geq 19.02$ y $\chi_9^2 = 8.66514 \not\leq 2.70$ resulta que el estadístico de contraste no cae dentro de la región crítica, luego no podemos rechazar H_0 contra H_1 al nivel de significación $\alpha = 0.05$.

Contrastes para la proporción muestral: muestras grandes

- Si denotamos por p la proporción poblacional y por \hat{p} la proporción muestral hemos visto que

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

sigue, por el T.L.C, aproximadamente una distribución normal.

- Como es lógico no conocemos la proporción muestral, pero si suponemos que la muestra es grande podríamos aproximar por

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

- Entonces

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$$

- Si $H_0 : p = p_0$ es cierta tenemos que $\sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{p_0(1-p_0)}{n}}$:

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

sigue teniendo aproximadamente una distribución normal estándar si n es grande.

- De forma similar al contraste de la media podemos definir las siguientes regiones críticas al nivel de significación α para las distintas hipótesis alternativas.

Reglas de decisión para el contraste de una proporción muestral: tamaño muestral grande

Condiciones:

- Muestra aleatoria simple de tamaño grande n
- Procedente de una población con proporción poblacional de la característica de interés p , y proporción muestral de la misma \hat{p}

Entonces una contraste al nivel de significación α para las hipótesis:

- $$\begin{cases} H_0 : p = p_0 & (\text{o } H_0 : p \leq p_0) \\ H_1 : p > p_0 \end{cases}$$

Tiene por regla de decisión:

Rechazar H_0 si

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} > z_{1-\alpha}$$

$$\begin{cases} H_0 : p = p_0 & (\text{o } H_0 : p \geq p_0) \\ H_1 : p < p_0 \end{cases}$$

Tiene por regla de decisión:

Rechazar H_0 si

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} < z_\alpha.$$

$$\begin{cases} H_0 : p = p_0 \\ H_1 : p \neq p_0 \end{cases}$$

Tiene por regla de decisión:

Rechazar H_0 si

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} > z_{1-\frac{\alpha}{2}} \text{ o } \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} < z_{\frac{\alpha}{2}}.$$

Ejemplo.

- Una asociación ganaderos afirma que en las matanzas caseras en Baleares el 70 % de los cerdos han sido analizados de triquinosis.
- En una investigación se obtiene en muestra aleatoria de 100 matanzas resultando que en 53 se han realizado estos análisis.
- ¿Estaríamos de acuerdo con la afirmación de los ganadero?
- Utilizar un nivel de significación $\alpha = 0.01$. Calcular el p -valor del contraste e interpretarlo.

Solución

Seguiremos los seis pasos:

1) $H_0 : p = 0.7$

2) $H_1 : p \neq 0.7$

3) $\alpha = 0.01$ luego $\frac{\alpha}{2} = 0.005$

4) Región crítica

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} > z_{1-0.005} = 2.57 \text{ o } \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} < z_{0.005} = -2.57$$

Solución

$$5) \hat{p} = \frac{53}{100} = 0.53 \quad Z = \frac{0.53 - 0.7}{\sqrt{\frac{0.7 \cdot 0.3}{100}}} = -3.7$$

- 6) $Z = -3.7 \not> 2.57$ pero $Z = -3.7 < -2.57$ luego el estadístico de contraste está en la región crítica por lo tanto no podemos aceptar la afirmación de la asociación al nivel de significación $\alpha = 0.01$.

Solución

- Calculemos el p valor. Lo haremos por el lado izquierdo que es por donde antes alcanzaremos el valor $Z = -3.7$ como tenemos que $z_{0.0001} = -3.7$ entonces el p -valor es $\frac{\alpha}{2} = 0.0001$
- Por lo tanto el p -valor es $\alpha = 2 \cdot 0.0001 = 0.0002$.
- Es decir, la afirmación de la asociación dista mucho de ser cierta.

Introducción

- Estudiaremos los principios más básicos del diseño de experimentos.
- Cuando comparamos dos parámetros entre dos poblaciones de individuos los diseños básicos son el de **muestras independientes** y el de muestras repetidas sobre los mismos individuos o **muestras emparejadas**.
- En algunos de estos casos veremos ambos tipos de contrastes.
- Por lo demás los contrastes funcionan de forma similar a los de un sólo parámetro.

Contrastes de dos parámetros muestras independientes

- Comenzaremos los contrastes de dos parámetros con el caso en que tengamos dos muestras aleatoria de una misma misma variable e independientes entre si.
- Que las muestras son independientes quiere decir que la selección de los individuos de cada población a observar es independiente.
- Por lo tanto se presume que si hay diferencias entre los parámetros deben deberse a que las poblaciones tienen alguna característica distinta (o factor que los diferencia).

Contrastes de dos parámetros muestras independientes

- Supongamos pues, que tenemos dos muestras aleatorias independientes de tamaños n_1 y n_2 y medias μ_1 y μ_2 respectivamente.
- Así tendremos una muestra será $x_{11}, x_{12}, \dots, x_{1n_1}$ y la otra $x_{21}, x_{22}, \dots, x_{2n_2}$. Denotaremos por \bar{X}_1 y \bar{X}_2 son las medias aritméticas de cada muestra.
- La hipótesis nula que se contrasta es: $H_0 : \mu_1 - \mu_2 = 0$ aunque se suele escribir como $H_0 : \mu_1 = \mu_2$.

Contrastes de dos parámetros muestras independientes

- Podemos aplicar este tipo de contrastes para poblaciones que tengan distribución normal.
- También podemos utilizarlo cuando n_1 y n_2 son suficientemente grandes y podemos aproximar la distribución de las medias por el Teorema del Límite Central.
- El test a aplicar tiene dos casos:
 - ▶ Que las varianzas σ_1^2 y σ_2^2 respectivamente sean conocidas.
 - ▶ O bien que sean desconocidas. En este caso hay dos variantes del test:
 - ★ Que aceptemos que las varianzas son iguales
 - ★ O bien el caso contrario que aceptemos que las varianzas son distintas.

Contrastes de dos parámetros muestras independientes

- La diferencia entre varianzas desconocidas iguales o distintas radica en utilizar una fórmula distinta para estimar la varianza muestral.
- En el caso en que sean iguales podemos aprovechar las dos muestras para estimar la varianza de la población.
- Los estadísticos de contraste, las regiones críticas se encuentra en las tablas de resúmenes de los contrastes.
- Estas tablas estan en la carpeta de MAterial Adicional del espacio de la asignatura en Campus Extens o en <http://bioinfo.uib.es/recerca/mates2/ContrasteHipotesis/TablaContrastesdeHipotesis.pdf>
- En las mismas tablas también podemos encontrar distintos intervalos de confianza para estimaciones de una y dos muestras.

Ejemplo

- Queremos estudiar los tiempos de ejecución de un algoritmo de alineamiento de proteínas.
- Para ello disponemos de dos muestras independientes de pares de proteínas de tamaños muestrales $n_1 = n_2 = 20$.
- Supongamos que los tiempos siguen aproximadamente una distribución normal y que las desviaciones típicas son conocidas $\sigma_1 = 1$ y $\sigma_2 = 2$.

Ejemplo

- Los resultados de la muestra del primer algoritmo sobre la muestra 1 son:
10.54, 10.73, 9.11, 8.07, 10.56, 9.87, 9.52, 8.34, 9.83, 8.11, 11.14, 10.11, 7.6, 11.13, 10.95, 9.48, 9.31, 11.82, 10.93, 8.63
- Mientras que los resultados de la muestra del segundo algoritmo para la muestra 2 son:
11.93, 14.24, 11.06, 11.2, 12.31, 12.92, 13.61, 15.03, 13.06, 11.47, 12.8, 14.55, 10.46, 15.21, 12.58, 11.76, 9.63, 10.81, 12.54, 10.53

Ejemplo

- Se tiene que las medias en cada una de las poblaciones son $\bar{x}_1 = 9.789$ y $\bar{x}_2 = 12.385$.
- Para la estimación de la varianza muestral, en este caso, se utiliza la fórmula $\tilde{S} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$; con nuestros datos se obtiene que $\tilde{S} = \sqrt{\frac{1}{20} + \frac{4}{20}} = 0.5$.
- El estadístico de contraste es

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\tilde{S}}.$$

- En nuestro caso vale $Z = \frac{9.789 - 12.385}{0.5} = -5.192$

Ejemplo

- Si contrastamos contra $H_1 : \mu_1 \neq \mu_2$ entonces la región crítica del contraste para $\alpha = 0.05$ es

$$Z < z_{\frac{\alpha}{2}} \text{ o } Z > z_{1-\frac{\alpha}{2}}$$

- Para este nivel de significación $z_{1-\frac{\alpha}{2}} = z_{1-0.025} = z_{0.975} = 1.96$ y por lo tanto $z_{\frac{\alpha}{2}} = z_{0.025} = -1.96$
- No podemos aceptar que los tiempos de ejecución tiene medias iguales, contra que las tiene distintas, al nivel de significación $\alpha = 0.05$

Ejemplo

- Ahora podemos calcular un intervalo de confianza del 95 % para la diferencia de medias $\mu_1 - \mu_2$.
- Consultando las tablas de los resúmenes de los contrastes, tenemos que el intervalo pedido es

$$\left(\bar{X}_1 - \bar{X}_2 + z_{\frac{\alpha}{2}} \tilde{S}, \bar{X}_1 - \bar{X}_2 + z_{1-\frac{\alpha}{2}} \tilde{S} \right)$$

- Que en nuestro caso, para $\alpha = 0.05$ es
$$(9.789 - 12.385 + z_{0.025} \cdot 0.5, 9.789 - 12.385 + z_{0.975} \cdot 0.5)$$
$$= (-2.596 - 1.96 \cdot 0.5, -2.596 + 1.96 \cdot 0.5)$$
- Por lo tanto un intervalo de confianza al nivel del 95 % para $\mu_1 - \mu_2$ es

$$(-3.576, -1.616)$$

Ejemplo

- Notemos que en este caso el cero no se encuentra en el intervalo de confianza.
- Por último calculemos el p -valor para el contraste bilateral será el valor de α tal que

$$z_{\frac{\alpha}{2}} = -5.192$$

de donde (utilizando R `pnorm(0.025)` o en su caso las tablas) se tiene que

$\frac{\alpha}{2} = 1.040235e - 07$ y por lo tanto $\alpha = 2.08047e - 07$. Utilizando las tablas hubiéramos concluido que el p -valor es prácticamente cero.

- Se deja como ejercicio el cálculo de los dos contrastes bilaterales, sus p -valores y los intervalos de confianza unilaterales.

Ejemplo.

- Con los mismos datos que en el ejemplo anterior pero suponiendo ahora que las varianzas no son conocidas el test cambia.
- En primer lugar tendremos que estimar la varianza de otra forma.
- Lo podemos hacer de dos maneras: suponiendo que las varianzas poblacionales son iguales o que son distintas.
- En el primero de los casos se obtiene un estadístico que sigue la distribución t de Student y en el segundo otra vez un estadístico Z con distribución normal.

Ejemplo

- Necesitamos \tilde{S}_1 y \tilde{S}_2 las cuasi desviaciones típicas muestrales que valen 1.201323 y 1.579462 respectivamente.
- Supongamos las varianzas son iguales entonces el estadístico de contraste para la hipótesis nula bilateral es

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\tilde{S}_{1,2}}.$$

- Donde la desviación típica muestral se estima por

$$\tilde{S}_{1,2} = \sqrt{\frac{(n_1 - 1)\tilde{S}_1^2 + (n_2 - 1)\tilde{S}_2^2}{n_1 + n_2 - 2} \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}.$$

Ejemplo

- El estadístico t sigue una distribución $t_{n_1+n_2-2}$, es decir t de Student con $n_1 + n_2 - 2$ grados de libertad.

- Luego en nuestro caso

$$\tilde{S}_{1,2} = \sqrt{\frac{(20-1)(1.201323)^2 + (20-1)(1.579462)^2}{20+20-2}} \cdot \left(\frac{1}{20} + \frac{1}{20}\right) = 0.4437272$$

- Entonces el valor del estadístico de contraste es

$$t = \frac{9.789 - 12.385}{0.4437272} = -5.850441$$

- La región crítica es, rechazar H_0 si

$$t \leq t_{n_1+n_2-2, \frac{\alpha}{2}} \text{ o } t \geq t_{n_1+n_2-2, 1-\frac{\alpha}{2}}$$

Ejemplo

- Para el nivel de significación $\alpha = 0.05$ tenemos que $t_{n_1+n_2-2, 1-\frac{\alpha}{2}} = t_{38, 0.975}$ mirando las tablas de las t de Student aproximamos por $t_{40, 0.975} = 2.021075$ que es el valor más cercano (con R y la instrucción `qt(0.975, 38)` se obtiene 2.024394).
- Ahora tenemos que $t_{38, 0.025} = -t_{38, 0.975}$ y lo aproximamos por $-t_{40, 0.975} = -2.021075$
- Así rechazamos H_0 ya que $t = -5.850441 < -2.021075 \approx t_{38, 0.025}$.
- Para el cálculo del p -valor igualamos $t_{38, \frac{\alpha}{2}} = t = -5.850441$ con R hacemos `pt(-5.850441, 38)` y se obtiene que $\frac{\alpha}{2} = 4.565589e - 07$ luego el p -valor es $2 \cdot 4.565589e - 07 = 9.131178e - 07$ que es muy próximo a cero (ejercicio: resolverlo utilizando las tablas de la t de Student).

Ejemplo

- Por último el intervalo de confianza para la diferencia de las medias $\mu_1 - \mu_2$ es (donde $m = n_1 + n_2 - 2$)

$$(\bar{X}_1 - \bar{X}_2 + t_{m, \frac{\alpha}{2}} \tilde{S}_{1,2}, \bar{X}_1 - \bar{X}_2 + t_{m, 1 - \frac{\alpha}{2}} \tilde{S}_{1,2})$$

- Que en nuestro caso es
 $(9.789 - 12.385 - 2.021075 \cdot 0.4437272, 9.789 - 12.385 + 2.021075 \cdot 0.4437272) = (-3.492806, -1.699194).$
- Se deja como ejercicio el cálculo de los dos contrastes bilaterales, sus p -valores y los intervalos de confianza unilaterales.
- También se deja como ejercicio el caso en el que las varianzas son distintas.

Ejemplo

Como ejercicio resolver utilizando las tablas de contrastes el ejercicio anterior en el caso de varianzas desconocidas y distintas. !!Ahora!!

Contraste de dos proporciones muestras independientes

- El test de contraste de dos proporciones se enfrenta a la comparación del parámetro p de probabilidad de éxito en dos poblaciones Bernoulli de parámetros p_1 y p_2 , independientes de tamaños n_1 y n_2 .
- Este test es parecido al de dos medias y sólo se puede aplicar con tamaños muestrales grandes.
- Tendremos dos muestras y sus correspondientes proporciones muestrales \bar{p}_1 y \bar{p}_2 . El estadístico de contraste es

$$Z = \frac{\bar{p}_1 - \bar{p}_2}{\sqrt{\bar{p} \bar{q} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

Contraste de dos proporciones muestras independientes

- Donde $\bar{p} = \frac{n_1\bar{p}_1 + n_2\bar{p}_2}{n_1 + n_2}$ y $\bar{q} = 1 - \bar{p}$
- El estadístico Z sigue una ley normal estándar.
- La región de rechazo frente a la alternativa bilateral es, rechazar H_0 al nivel α si :

$$Z < Z_{\frac{\alpha}{2}} \text{ o } Z > Z_{1-\frac{\alpha}{2}}$$

- El intervalo de confianza para la diferencia de proporciones poblacionales $p_1 - p_2$ al nivel $(1 - \alpha) \cdot 100 \%$ es

$$\left(\bar{p}_1 - \bar{p}_2 + z_{\frac{\alpha}{2}} \sqrt{\bar{p} \cdot \bar{q} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}, \bar{p}_1 - \bar{p}_2 - z_{\frac{\alpha}{2}} \sqrt{\bar{p} \cdot \bar{q} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} \right)$$

Ejemplo.

- Se toman dos muestras de ADN de individuos con al menos tres generaciones familiares en la isla de Mallorca y otra, en las mismas condiciones, de Menorca.
- Se quiere saber si la proporción de la presencia de un determinado alelo en un gen es igual o distinta entre las dos muestras.
- La muestra de Mallorca tiene tamaño 100 y resultaron 20 individuos con el alelo, mientras que la de Menorca tiene tamaño 50 y resultaron 12 individuos con el alelo.
- Como ejercicio contrastar la hipótesis de igualdad de proporciones al nivel de significación 0.05, calcular el p -valor y los intervalos de confianza con el mismo valor de α (!!ahorajj).

Muestras emparejadas o dependientes

- Hasta ahora hemos considerado que las muestras de las dos poblaciones de las que teníamos que contrastar su media o su varianza eran elegidas de forma que fueran independientes.
- Otro caso distinto es cuando las dos muestras corresponden a los mismos individuos o a individuos emparejados por algún factor determinante.
- Ejemplos de este factor son pares de gemelos univitelinos, pares de proteínas a comparar u otros emparejamientos que se puedan considerar aceptables para el diseño del experimento, como coeficiente intelectual, por peso, por edad, por ideología etc....
- En estos casos se habla de un diseño de datos dependientes o emparejados. (En inglés: *paired* y por lo tanto se habla de *paired test*).

Muestras emparejadas o dependientes

- En este caso el contraste más común corresponde a calcular las diferencia de los valores de cada una de las muestras para cada individuo y realizar un contraste para averiguar si la media de las diferencias (o proporciones) es cero.
- Lo más importante de este caso es aprender que hay diferentes maneras de realizar un diseño experimental para contrastar una hipótesis.
- Este diseño debe haber sido fijado, justificadamente, antes de realizar la experiencia, es decir antes de la recogida de datos.
- La regiones de rechazo para estos contrastes están en la tabla de contrastes de hipótesis que hemos proporcionado.
- A continuación se presentan dos ejemplos. El primero es un contraste de medias para muestras emparejadas.

Ejemplo: Contraste de dos media muestras dependientes.

- Disponemos de dos algoritmos para alineamientos de proteínas. Ambos aportan resultados de la misma calidad.
- Estamos interesados en saber cuál de los dos tiene la media de tiempo de ejecución más pequeña.
- Para ello tomamos una muestra de pares de proteínas.
- Alineamos cada par de proteínas con cada uno de los algoritmos.
- Este es un ejemplo de diseño experimental de muestras emparejadas.
- Es evidente que las muestras no son independientes pues cada tiempo de ejecución corresponden a los mismos pares de proteínas.

Ejemplo: Comparación medias muestras dependientes

Los resultados obtenidos en los tiempos de ejecución de ambos algoritmos son:

ordenador i	1	2	3	4	5	6	7	8	9	10
antes	8.1	11.9	11.4	12.9	9.0	7.2	12.4	6.9	8.9	8.3
después	6.9	6.7	8.3	8.6	18.9	7.9	7.4	8.7	7.9	12.4
$d_i = \text{antes-después}$	1.2	5.2	3.1	4.3	-9.9	-0.7	5.0	-1.8	1.0	-4.1

Ejemplo: Comparación medias muestras dependientes

- Se tiene que $\bar{d} = 0.33$ y $\tilde{S}_d = 4.72$.
- Contrastar la igualdad de medias con el test que corresponda y que podéis encontrar en las tablas de contraste de hipótesis. Calcular el p -valor y un intervalo de confianza del 95 % para la diferencia de las dos medias.
- El estadístico de contraste es $t = \frac{\bar{d}}{\tilde{S}_d/\sqrt{n}} = \frac{0.33}{4.72/\sqrt{10}} = 0.22$
- Así para la región crítica bilateral el p -valor es el α tal que :

$$t_{9, 1-\frac{\alpha}{2}} = 0.22$$

Ejemplo: Comparación medias muestras dependientes

- Consultando las tablas de la distribución t de Student se obtiene que $t_{9,0.5871} \approx t_{9,0.9} = 1.383029$ (es una aproximación muy mala).
- Por lo tanto $1 - \frac{\alpha}{2} = 0.9$ y $\alpha = 0.2$.
- El código R es

```
> pt(9, 0.22)

[1] 0.7720373
```
- El p -valor exacto sería la solución de la ecuación $1 - \frac{\alpha}{2} = 0.7720373$.
- Como ejercicio calcular el intervalo de confianza para la diferencia de medias

Ejemplo: Contraste de dos proporciones muestras emparejadas.

- Este es un ejemplo de contraste de proporciones en la misma población antes y después (es lo mismo que el diseño emparejado) de un *evento* (Se deja como ejercicio)
- Se toma una muestra de 100 personas afectadas por migraña. Se les facilita un fármaco para que alivie los síntomas.
- Después de la administración se les pregunta si han notado alivio en el dolor.
- Después de un tiempo se les subministra a los mismos individuos un placebo y se les vuelve a preguntar si han notado o no mejoría.
- Los resultados son:

		Después	
		Sí	No
Antes	Sí	300	10
	No	100	590

Utilizar el contraste apropiado (mirar tablas de contrastes hipótesis).
 Calcular el p -valor y un intervalo de confianza del 95 % para la diferencia de las dos proporciones.

Comparación de dos varianzas muestras independientes

- Hemos visto la necesidad de comparar dos varianzas como paso previo a una comparación de medias de muestras independientes, aun que también puede tener sentido en si misma.
- El contraste corresponde con una hipótesis nula $H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1$ contra las alternativas unilaterales y bilateral habituales.
- El intervalo de confianza que se encuentra en las tablas de contrastes es para el cociente, no para la diferencia de varianzas. Así que si vamos a aceptar la igualdad de varianzas este intervalo debe contener a 1.
- El estadístico de contraste sigue una ley de distribución F de Fisher y tiene por parámetros dos grados de libertad n y m . Los valores de esta distribución se pueden obtener en <http://bioinfo.uib.es/recerca/mates2/tablasDistribuciones/Fisher.pdf> o en la carpeta de material adicional de Campus Extens.

Ejemplo.

En el ejemplo de los algoritmos para alineamiento de proteínas en el caso de muestras independientes. Se trata de contrastar la diferencia de las medias de los rendimientos en el caso que se determine según sean las varianzas iguales o distintas. Para ello se debe primero contrastar la igualdad de varianzas. Hacerlo ahora como ejercicio.

Parte VII

Análisis de la Varianza.

Introducción.

- El análisis de la varianza (ANOVA) es una técnica para comparar medias de más de dos poblaciones.
- En este capítulo introduciremos aspectos elementales del diseño experimental. El diseño experimental en estadística abarca los métodos para recoger y analizar datos, cuyos objetivos son aumentar al máximo la cantidad y mejorar la exactitud de la información proporcionada por un determinado experimento.
- La técnica ANOVA consiste básicamente en dividir la variación total del experimento en cuestión en componentes, algunas de las cuales son debidas a la diferencia entre los distintos grupos que componen las poblaciones y las demás a variaciones aleatorias.

Introducción

- Tenemos k poblaciones con una característica común que será la variable de estudio. Queremos ver si la media de dicha variable es la misma para todas las poblaciones.
- Sean μ_1, \dots, μ_k las medias de la variable a estudiar de cada una de las poblaciones.
- Cogemos una muestra de tamaño n_i con respecto de la población i , para $i = 1, \dots, k$. O sea, tenemos n_1 datos de la población 1, ..., n_k datos de la población k .
- Planteamos el contraste de hipótesis siguiente:

$$\left. \begin{array}{l} H_0 : \mu_1 = \mu_2 = \dots = \mu_k, \\ H_1 : \exists i, j \mid \mu_i \neq \mu_j. \end{array} \right\}$$

Ejemplo a estudiar. Enunciado.

El dióxido de carbono tiene un efecto crítico en el crecimiento microbiológico. Cantidades pequeñas de CO_2 estimulan el crecimiento de muchos microorganismos, mientras que altas concentraciones inhiben el crecimiento de la mayor parte de ellos. Este último efecto se utiliza comercialmente cuando se almacenan productos alimenticios perecederos. Se realizó un estudio para investigar el efecto del CO_2 sobre la tasa de crecimiento de *Pseudomonas fragi*, un corruptor de alimentos. Se administró dióxido de carbono a cinco presiones atmosféricas diferentes. La respuesta anotada fue el cambio porcentual en la masa celular después de un tiempo de crecimiento de una hora. Se utilizaron diez cultivos en cada nivel. Se obtuvieron los datos siguientes:

Ejemplo a estudiar. Tabla de datos.

Nivel del factor				
0.0	0.083	0.29	0.50	0.86
62.6	50.9	45.5	29.5	24.9
59.6	44.3	41.1	22.8	17.2
64.5	47.5	29.8	19.2	7.8
59.3	49.5	38.3	20.6	10.5
58.6	48.5	40.2	29.2	17.8
64.6	50.4	38.5	24.1	22.1
50.9	35.2	30.2	22.6	22.6
56.2	49.9	27.0	32.7	16.8
52.3	42.6	40.0	24.4	15.9
62.8	41.6	33.9	29.6	8.8

donde los niveles de CO_2 están en atmósferas. Queremos estudiar si el crecimiento del microorganismo está influido por la presión atmosférica.

Ejemplo a estudiar. Instrucciones de R

- Almacenamos los datos en la variable `ej_anova`. Dicha variable será una matriz con dos columnas; en la primera columna, almacenaremos los datos de la variable y en la segunda columna indicaremos el nivel donde pertenece dicho dato:

```
> (ej_anova <- matrix(c(62.6,50.9,45.5,29.5,24.9,  
+ 59.6,44.3, 41.1,22.8,17.2,64.5,47.5,29.8,19.2,  
+ 7.8,59.3, 49.5,38.3,20.6,10.5,58.6,48.5,40.2,  
+ 29.2,17.8, 64.6,50.4,38.5,24.1,22.1,50.9,35.2,  
+ 30.2,22.6, 22.6,56.2,49.9,27.0,32.7,16.8,52.3,  
+ 42.6,40.0, 24.4,15.9,62.8,41.6,33.9,29.6,8.8),  
+ 50,1))
```

Añadimos la columna que nos indica los niveles:

```
> (ej_anova <- cbind(ej_anova,rep(seq(1:5),10)))  
      [,1] [,2]  
[1,] 62.6   1  
[2,] 50.9   2  
...
```

Ejemplo a estudiar. Instrucciones de R

- Transformamos la variable ej_anova en data frame:

```
> ej_anova <- as.data.frame(ej_anova)
```

- Almacenamos las columnas en la memoria de R:

```
> attach(ej_anova)
```

- La variable ej_anova vale:

```
      V1 V2
1  62.6  1
2  50.9  2
3  45.5  3
4  29.5  4
5  24.9  5
6  59.6  1
7  44.3  2
8  41.1  3
9  22.8  4
10 17.2  5
11 64.5  1
12 47.5  2
13 29.8  3
14 19.2  4
15  7.8  5
16 59.3  1
17 49.5  2
18 38.3  3
...
```

Clasificación simple. Suposiciones.

- Vamos a estudiar la técnica ANOVA en los supuestos siguientes:
 - ▶ Clasificación simple o una vía. Se refiere a que solamente estudiamos un factor o una característica. Tenemos en total k niveles de dicho factor. En el ejemplo que estudiamos, $k = 5$ y $n_1 = n_2 = n_3 = n_4 = n_5 = 10$.
 - ▶ Diseño completamente aleatorio. La elección de las unidades experimentales o individuos de las muestras se ha realizado de forma totalmente independiente. Las k muestras son independientes unas de otras.
 - ▶ Efectos fijos. El experimentador selecciona específicamente los niveles del factor implicados. Más adelante hablaremos de efectos aleatorios donde los niveles se seleccionan aleatoriamente de entre un conjunto de niveles del factor. En el ejemplo, los niveles de CO_2 (0.0, 0.083, 0.29, 0.50 y 0.86) han sido elegidos por el experimentador.

Clasificación simple. Formato de los datos.

- Formato de los datos:

Nivel del factor			
1	2	...	k
X_{11}	X_{21}	...	X_{k1}
X_{12}	X_{22}	...	X_{k2}
...
X_{1n_1}	X_{2n_2}	...	X_{kn_k}

- donde n_i es el tamaño de la muestra del nivel i , para $i = 1, \dots, k$ y X_{ij} representa el valor de la característica o la variable que estudiamos del individuo j correspondiente al nivel i .

Clasificación simple. Estadísticos

Definimos los estadísticos siguientes:

- Suma total de los datos en el nivel i -ésimo:

$$T_{i\bullet} = \sum_{j=1}^{n_i} X_{ij}.$$

- Media muestral para el nivel i -ésimo:

$$\bar{X}_{i\bullet} = \frac{T_{i\bullet}}{n_i}.$$

- Suma total de los datos:

$$T_{\bullet\bullet} = \sum_{i=1}^k \sum_{j=1}^{n_i} X_{ij} = \sum_{i=1}^k T_{i\bullet}.$$

- Media muestral de todos los datos:

$$\bar{X}_{\bullet\bullet} = \frac{T_{\bullet\bullet}}{N},$$

donde $N = n_1 + \cdots + n_k$.

Cálculo de los estadísticos para el ejemplo estudiado.

- En primer lugar definimos una variable tipo factor que contendrá los niveles:
> niveles <- as.factor(ej_anova[,2])
- Seguidamente, definimos 5 variables que contendrán la información de la respuesta para cada uno de los cinco niveles:

```
> nivel1 <- ej_anova[niveles==1,]  
> nivel2 <- ej_anova[niveles==2,]  
> nivel3 <- ej_anova[niveles==3,]  
> nivel4 <- ej_anova[niveles==4,]  
> nivel5 <- ej_anova[niveles==5,]
```

Cálculo de los estadísticos para el ejemplo estudiado.

- Suma total de los datos para cada uno de los cinco niveles:

```
> (sumas_por_niveles <- c(sum(nivel1[,1]),  
sum(nivel2[,1]),sum(nivel3[,1]),sum(nivel4[,1]),  
sum(nivel5[,1])))  
[1] 591.4 460.4 364.5 254.7 164.4
```

- Media muestral por niveles:

```
> (medias_por_niveles<- c(mean(nivel1[,1]),  
mean(nivel2[,1]),mean(nivel3[,1]),mean(nivel4[,1]),  
mean(nivel5[,1])))  
[1] 59.14 46.04 36.45 25.47 16.44
```

Cálculo de los estadísticos para el ejemplo estudiado.

- Suma total de los datos:

```
> (suma_total <- sum(ej_anova[,1]))  
[1] 1835.4
```

- Media muestral de todos los datos:

```
> (media_total <- mean(ej_anova[,1]))  
[1] 36.708
```


Contraste a realizar

- Recordemos los parámetros que intervendrán en el contraste:
 - ▶ μ_i : promedio teórico o respuesta esperada al i -ésimo nivel, $i = 1, \dots, k$.
 - ▶ μ : promedio teórico o respuesta esperada, ignorando los niveles del factor.
- Expresión matemática del modelo a estudiar (modelo para la clasificación simple, diseño completamente aleatorizado con efectos fijos):

$$X_{ij} = \mu + (\mu_i - \mu) + (X_{ij} - \mu_i), \quad i = 1, \dots, k, \quad j = 1, \dots, n_i,$$

donde

- ▶ X_{ij} : respuesta del j -ésimo individuo dentro del nivel i -ésimo,
- ▶ μ : respuesta media global.
- ▶ $\mu_i - \mu$: desviación de la media global debida al hecho de que el nivel recibe el tratamiento i -ésimo.
- ▶ $X_{ij} - \mu_i$: desviación aleatoria de la media i -ésima debido a influencias aleatorias.

Suposiciones del modelo y estimadores de los parámetros

- Las suposiciones del modelo son:
 - ▶ Las k muestras representan muestras aleatorias independientes extraídas de k poblaciones específicas con medias μ_1, \dots, μ_k .
 - ▶ Cada una de las k poblaciones es normal.
 - ▶ Cada una de las k poblaciones tiene la misma varianza σ^2 .
- Los estimadores de los parámetros son los siguientes:
 - ▶ Parámetro μ : $\bar{X}_{\bullet\bullet}$.
 - ▶ Parámetro $\mu_i - \mu$: $\bar{X}_{i\bullet} - \bar{X}_{\bullet\bullet}$.

Identidad de la suma de cuadrados

Se cumple la identidad siguiente:

$$\sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{..})^2 = \sum_{i=1}^k n_i (\bar{X}_{i.} - \bar{X}_{..})^2 + \sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{i.})^2,$$

donde

- $\sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{..})^2$ es la suma de los cuadrados de las desviaciones de los datos respecto de la media global. A dicha cantidad le llamaremos suma total de cuadrados (SS_{Total}).
- $\sum_{i=1}^k n_i (\bar{X}_{i.} - \bar{X}_{..})^2$ es la suma ponderada de los cuadrados de las desviaciones del nivel o de las medias de los tratamientos respecto de la media global. Llamaremos a dicha cantidad suma de cuadrados de los tratamientos (SS_{Tr}).
- $\sum_{i=1}^k \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{i.})^2$ es la suma de los cuadrados de las desviaciones de los datos respecto de la media de los niveles asociados con dichos datos. Llamaremos a dicha cantidad suma de cuadrados de los residuos, o error (SS_E).

Cálculo de la suma de cuadrados para el ejemplo

- Suma total de cuadrados SS_{Total} :

```
> (SSTotal <- sum((ej_anova[,1]-  
media_total)^2))  
[1] 12522.36
```

- Suma de cuadrados de los tratamientos SS_{Tr} .

```
> (SSTr <- sum(table(ej_anova[,2])*  
(medias_por_niveles-media_total)^2))  
[1] 11274.32
```

- Suma de cuadrados de los residuos SS_E .

```
> (SSE <- sum((ej_anova[,1]-  
medias_por_niveles[ej_anova[,2]])^2))  
[1] 1248.038
```

- Podemos comprobar que se cumple la igualdad siguiente:

$$SS_{Total} = SS_{Tr} + SS_E.$$

Estadísticos para realizar el contraste

- Definimos los estadísticos siguientes:
 - ▶ Cuadrado medio de los tratamientos MS_{Tr} : $MS_{Tr} = \frac{SS_{Tr}}{k-1}$.
 - ▶ Cuadrado medio residual: $MS_E = \frac{SS_E}{N-k}$.
- Dichos estadísticos, al ser variables aleatorias, tienen una distribución y un valor medio y una varianza. Se pueden demostrar los resultados siguientes:
 - ▶ $E(MS_{Tr}) = \sigma^2 + \sum_{i=1}^k \frac{n_i(\mu_i - \mu)^2}{k-1}$.
 - ▶ $E(MS_E) = \sigma^2$.
- Notemos que el estadístico MS_E puede usarse para estimar la varianza común σ^2 .

Estadísticos para realizar el contraste

- ¿Cómo usar los estadísticos anteriores en nuestro contraste?
- Si H_0 es cierta, o $\mu_1 = \dots = \mu_k = \mu$, se cumple:

$$\sum_{i=1}^k \frac{n_i(\mu_i - \mu)^2}{k-1} = 0,$$

- y si H_0 no fuese cierta, la cantidad anterior sería positiva.
- Por tanto, si H_0 fuese cierta, los estadísticos MS_{Tr} y MS_E tendrían valores próximos y estimarían el mismo parámetro σ^2 .
- Por tanto, se considera el cociente $\frac{MS_{Tr}}{MS_E}$ como estadístico de contraste ya que si H_0 es cierta,
 - ▶ su valor sería próximo a 1 y
 - ▶ su distribución sigue la variable $F_{k-1, N-k}$ (F de Fisher-Snédecor con $k-1$ y $N-k$ grados de libertad).

Pasos a realizar en el contraste ANOVA

- En conclusión, el contraste ANOVA consta de los pasos siguientes:
 - ▶ En primer lugar, calculamos las sumas de cuadrados SS_{Total} , SS_{Tr} y SS_E .
 - ▶ Seguidamente, hallamos los cuadrados medios $MS_{Tr} = \frac{SS_{Tr}}{k-1}$ y $MS_E = \frac{SS_E}{N-k}$.
 - ▶ Luego, hallamos el estadístico de contraste $F = \frac{MS_{Tr}}{MS_E}$ y el valor crítico $F_{1-\alpha, k-1, N-k}$, donde dicho valor tiene área $1 - \alpha$ a la izquierda de la función de densidad de la variable $F_{k-1, N-k}$.
 - ▶ Si $F > F_{1-\alpha, k-1, N-k}$, rechazamos H_0 a un nivel de significación α y concluimos que los tratamientos son distintos para los k niveles. En caso contrario, aceptamos H_0 y concluimos que el tratamiento aplicado en todos los niveles tiene el mismo efecto.

Ejemplo anterior

- Los valores de las sumas de cuadrados eran: $SS_{Total} = 12522.36$, $SS_{Tr} = 11274.32$ y $SS_E = 1248.038$. Los cuadrados medios serán:

```
> (MSTr <- SStr/(5-1))  
[1] 2818.580  
  
> (MSE <- SSE/(dim(ej_anova)[1]-5))  
[1] 27.73418
```
- El estadístico de contraste F valdrá:

```
> (EstF <- MSTr/MSE)  
[1] 101.6284
```
- Para $\alpha = 0.05$, el valor crítico valdrá:

```
> (valor_critico <- qf(0.95,5,50-5))  
[1] 2.422085
```
- Como el estadístico de contraste es mayor que el valor crítico, rechazamos H_0 y concluimos que la presión atmosférica influye en el crecimiento del microorganismo *Pseudomonas fragi*.

Fórmulas para el cálculo de las sumas de los cuadrados

- En la práctica, las sumas de cuadrados deben realizarse usando las fórmulas siguientes: $SS_{Total} = \sum_{i=1}^k \sum_{j=1}^{n_i} X_{ij}^2 - \frac{T_{..}^2}{N}$,
 $SS_{Tr} = \sum_{i=1}^k \frac{T_{i.}^2}{n_i} - \frac{T_{..}^2}{N}$, $SSE = SS_{Total} - SS_{Tr}$.
- En el ejemplo tratado, tendríamos que usar las expresiones siguientes:

```
> (SSTotal <- sum(ej_anova[,1]^2)-suma_total^2/50)
[1] 12522.36
> (SSTr <- sum(sumas_por_niveles^2/
table(ej_anova[,2]))-suma_total^2/50)
[1] 11274.32
> (SSE <- SSTotal-SSTr)
[1] 1248.038
```

Tabla ANOVA

- El contraste ANOVA se resume en la tabla siguiente:

Origen de Variación	Grados libertad	Suma de cuadrados	Cuadrados medios	Estadístico de contraste
Nivel	$k - 1$	SS_{Tr}	$MS_{Tr} = \frac{SS_{Tr}}{k-1}$	$F = \frac{MS_{Tr}}{MS_E}$
Residuo	$N - k$	SS_E	$MS_E = \frac{SS_E}{N-k}$	
Total	$N - 1$	SS_{Total}		

- Ejemplo estudiado:

Origen de Variación	Grados libertad	Suma de cuadrados	Cuadrados medios	Estadístico de contraste
Nivel	4	11274.32	2818.58	$F = 101.63$
Residuo	45	1248.04	27.73	
Total	49	12522.36		

Tabla ANOVA

- En R el contraste ANOVA se puede realizar directamente:

```
> summary(aov(ej_anova[,1] ~ niveles))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
niveles	4	11274.3	2818.6	101.63	< 2.2e-16 ***
Residuals	45	1248.0	27.7		

```
---
```

- El valor de $\text{Pr}(> F)$ es el p-valor del contraste que, en nuestro caso, es despreciable. Por tanto, tenemos que rechazar la hipótesis nula como ya hemos comentado.

Comparaciones múltiples y por parejas

- En el caso de rechazar la hipótesis nula H_0 , hemos de plantearnos entre qué niveles existen diferencias. Veamos algunas técnicas para llevar a cabo dicha tarea.
- Contraste T de Bonferroni. Comparaciones por parejas. Si tenemos k niveles, pueden realizarse un total de $\binom{k}{2} = \frac{k(k-1)}{2}$ comparaciones posibles o contrastes de la forma:

$$\left. \begin{array}{l} H_0 : \mu_i = \mu_j, \\ H_1 : \mu_i \neq \mu_j. \end{array} \right\}$$

Comparaciones múltiples y por parejas

- El estadístico de contraste para realizar los contrastes anteriores es:

$$T = \frac{\bar{X}_{i\bullet} - \bar{X}_{j\bullet}}{\sqrt{MS_E(1/n_i + 1/n_j)}},$$

donde T sigue la distribución t de Student con $n_i + n_j - 2$ grados de libertad ($t_{n_i+n_j-2}$). Por tanto, diremos que existen diferencias entre los niveles i y j al nivel de significación α si $|T| > t_{1-\alpha/2, n_i+n_j-2}$. En caso contrario, concluiremos que no hay diferencia entre los dos niveles.

Comparaciones múltiples y por parejas

- La técnica anterior tiene que aplicarse teniendo en cuenta lo siguiente: Si se realizan c contrastes a un nivel de significación α , la probabilidad de realizar un contraste incorrecto, α' , es en general bastante mayor que α . De hecho, puede probarse que vale a lo sumo $1 - (1 - \alpha)^c$.
- Por ejemplo, en nuestro ejemplo estudiado, tenemos que $k = 5$. Si realizamos todos los posibles contrastes, tendremos que $c = \frac{5 \cdot 4}{2} = 10$ y, usando $\alpha = 0.05$, la probabilidad de que realicemos alguno incorrecto valdrá a lo sumo $1 - (1 - 0.05)^{10} \approx 0.40$. La probabilidad anterior no es aceptable, por tanto, sólo se tienen que realizar aquellos contrastes de interés real para el investigador ya que cuántos más contrastes hagamos, la probabilidad de que nos equivoquemos en alguno aumenta.

Comparaciones múltiples y por parejas

- Usando la aproximación $1 - (1 - \alpha)^c \approx c\alpha$, procederemos de la forma siguiente para hacer múltiples contrastes. Primeramente fijamos α' . Si queremos realizar c contrastes, el valor de significación α que tendremos que aplicar a cada uno de ellos será: $\alpha = \frac{\alpha'}{c}$.
- En el ejemplo tratado, si elegimos $\alpha' = 0.1$ y queremos realizar todos los posibles 10 contrastes entre todos los niveles, el nivel de significación α al que tenemos que realizar todos los contrastes será: $\alpha = \frac{0.1}{10} = 0.01$.

Ejemplo estudiado

- Hagamos una tabla con todos los posibles contrastes por parejas donde aparezcan los estadísticos de contraste y los valores críticos.
 - ▶ En primer lugar creamos una matriz donde aparezcan los niveles en los que realizamos las comparaciones múltiples:

```
> (parejas <- matrix(c(1,2,1,3,1,4,1,5,2,3,2,4,  
2,5,3,4,3,5,4,5),10,2,byrow=T))
```

	[,1]	[,2]
[1,]	1	2
[2,]	1	3
[3,]	1	4
[4,]	1	5
[5,]	2	3
[6,]	2	4
[7,]	2	5
[8,]	3	4
[9,]	3	5
[10,]	4	5

Ejemplo estudiado

- A continuación hallamos todos los estadísticos de contraste y añadimos los valores obtenidos como una columna más de la variable parejas.

```
> (parejas_est_contrastes <-  
abs(medias_por_niveles[parejas[,1]]-  
medias_por_niveles[parejas[,2]])/(sqrt(SSE*  
(1/table(niveles)[parejas[,1]]+  
1/table(niveles)[parejas[,2]])))  
niveles  
      1      1      1      1  
0.8291677 1.4361692 2.1311510 2.7027070  
      2      2      2      3  
0.6070014 1.3019832 1.8735393 0.6949818  
      3      4  
1.2665379 0.5715561
```

Ejemplo estudiado

```
• > (parejas <- cbind(parejas,  
  parejas_est_contrastes))
```

```
      parejas_est_contrastes
```

1	1	2	0.8291677
1	1	3	1.4361692
1	1	4	2.1311510
1	1	5	2.7027070
2	2	3	0.6070014
2	2	4	1.3019832
2	2	5	1.8735393
3	3	4	0.6949818
3	3	5	1.2665379
4	4	5	0.5715561

Ejemplo estudiado

- Hallamos los valores críticos:

```
> (valores_criticos <- qt(0.99,table(niveles)
[parejas[,1]]+table(niveles)[parejas[,2]]-2))
niveles
```

1	1	1	1	2
2.552380	2.552380	2.552380	2.552380	2.552380
2	2	3	3	4
2.552380	2.552380	2.552380	2.552380	2.552380

Ejemplo estudiado

- A continuación, añadimos los valores críticos a la variable parejas.

```
> (parejas <- cbind(parejas, valores_criticos))
```

	parejas_est	contrastes	valores_criticos
1 1 2		0.8291677	2.552380
1 1 3		1.4361692	2.552380
1 1 4		2.1311510	2.552380
1 1 5		2.7027070	2.552380
2 2 3		0.6070014	2.552380
2 2 4		1.3019832	2.552380
2 2 5		1.8735393	2.552380
3 3 4		0.6949818	2.552380
3 3 5		1.2665379	2.552380
4 4 5		0.5715561	2.552380

Ejemplo estudiado

Vemos que los únicos niveles en los que el crecimiento del microorganismo es distinto serían entre los niveles 1 y 5 que corresponden a una presión atmosférica de 0.0 y 0.86 atmósferas, respectivamente.

Recordemos que hemos elegido un nivel de significación global de $\alpha = 0.10$ y hemos decidido realizar todas las posibles comparaciones entre todos los niveles.

Contraste de Duncan de rango múltiple

- El contraste de Duncan es otro método para ver en qué niveles hay diferencias.
- El contraste de Duncan, a diferencia del método de Bonferroni descrito anteriormente, tiene en cuenta las medias de los niveles ordenadas. O sea, se ordenan las medias de cada nivel $\bar{X}_{i\bullet}$, $i = 1, \dots, k$ de menor a mayor.

Contraste de Duncan de rango múltiple

- Los pasos a realizar en el contraste de Duncan son los siguientes:
 - ▶ Se ordenan en forma ascendente las k medias muestrales.
 - ▶ Se considera cualquier subconjunto de p medias muestrales, $2 \leq p \leq k$. Diremos que existe diferencia entre el nivel con media más grande del subgrupo anterior y la media más pequeña del subgrupo anterior si el valor absoluto de la diferencia entre las dos medias anteriores es mayor que $SSR_p = r_p \sqrt{\frac{MS_E(n_i + n_j)}{2n_i n_j}}$, donde n_i y n_j son los tamaños de los niveles de media más grande y más pequeña respectivamente y r_p se denomina el menor rango significativo y puede hallarse en la tabla:

[http://costaricalinda.com/
Estadistica/duncan1.htm](http://costaricalinda.com/Estadistica/duncan1.htm)

Ejemplo anterior

- Las medias de los cinco niveles eran las siguientes:

$$\begin{aligned}\bar{X}_{1\bullet} &= 59.14, \bar{X}_{2\bullet} = 46.04, \bar{X}_{3\bullet} = 36.45, \\ \bar{X}_{4\bullet} &= 25.47, \bar{X}_{5\bullet} = 16.44.\end{aligned}$$

- Si las ordenamos de menor a mayor, obtenemos el resultado siguiente:
 $\bar{X}_{5\bullet} < \bar{X}_{4\bullet} < \bar{X}_{3\bullet} < \bar{X}_{2\bullet} < \bar{X}_{1\bullet}$.
- Podemos considerar las diferencias siguientes:
 - ▶ $\bar{X}_{1\bullet} - \bar{X}_{5\bullet}$ ($p = 5$).
 - ▶ $\bar{X}_{1\bullet} - \bar{X}_{4\bullet}$ ($p = 4$).
 - ▶ $\bar{X}_{1\bullet} - \bar{X}_{3\bullet}$ ($p = 3$).
 - ▶ $\bar{X}_{1\bullet} - \bar{X}_{2\bullet}$ ($p = 2$).
 - ▶ $\bar{X}_{2\bullet} - \bar{X}_{5\bullet}$ ($p = 4$).
 - ▶ $\bar{X}_{2\bullet} - \bar{X}_{4\bullet}$ ($p = 3$).
 - ▶ $\bar{X}_{2\bullet} - \bar{X}_{3\bullet}$ ($p = 2$).
 - ▶ $\bar{X}_{3\bullet} - \bar{X}_{5\bullet}$ ($p = 3$).
 - ▶ $\bar{X}_{3\bullet} - \bar{X}_{4\bullet}$ ($p = 2$).
 - ▶ $\bar{X}_{4\bullet} - \bar{X}_{5\bullet}$ ($p = 2$).

Ejemplo anterior

- En nuestro ejemplo, tenemos que $n_i = 10$ para cualquier i .
- Por tanto, $SSR_p = r_p \sqrt{\frac{MS_E}{10}}$.
- Hagamos una tabla con los distintos valores de SSR_p según el valor de p según el nivel de significación $\alpha = 0.05$: (en la tabla de r_p usamos 40 como grados de libertad del error ya que es el valor más cercano a 45)

p	2	3	4	5
r_p	2.858	3	3.102	3.171
SSR_p	4.760	4.996	5.166	5.281

- En R, para obtener la tabla anterior, tendríamos que hacer lo siguiente:

```
> rp <- c(2.858,3,3.102,3.171)
> (SSRp <- rp*sqrt(MSE/10))
[1] 4.759594 4.996074 5.165941 5.280851
```

Ejemplo anterior

- Veamos en qué niveles hay diferencias:

Diferencias	d	p	SSR_p	$d > SSR_p?$	Conclusión
$\bar{X}_{1\bullet} - \bar{X}_{5\bullet}$	42.7	5	5.281	Sí	$\mu_1 \neq \mu_5$
$\bar{X}_{1\bullet} - \bar{X}_{4\bullet}$	33.67	4	5.166	Sí	$\mu_1 \neq \mu_4$
$\bar{X}_{1\bullet} - \bar{X}_{3\bullet}$	22.69	3	4.996	Sí	$\mu_1 \neq \mu_3$
$\bar{X}_{1\bullet} - \bar{X}_{2\bullet}$	13.1	2	4.760	Sí	$\mu_1 \neq \mu_2$
$\bar{X}_{2\bullet} - \bar{X}_{5\bullet}$	29.6	4	5.166	Sí	$\mu_2 \neq \mu_5$
$\bar{X}_{2\bullet} - \bar{X}_{4\bullet}$	20.57	3	4.996	Sí	$\mu_2 \neq \mu_4$
$\bar{X}_{2\bullet} - \bar{X}_{3\bullet}$	9.59	2	4.760	Sí	$\mu_2 \neq \mu_3$
$\bar{X}_{3\bullet} - \bar{X}_{5\bullet}$	20.01	3	4.996	Sí	$\mu_3 \neq \mu_5$
$\bar{X}_{3\bullet} - \bar{X}_{4\bullet}$	10.98	2	4.760	Sí	$\mu_3 \neq \mu_4$
$\bar{X}_{4\bullet} - \bar{X}_{5\bullet}$	9.03	2	4.760	Sí	$\mu_4 \neq \mu_5$

- Según el contraste de Duncan y a un nivel de significación de $\alpha = 0.05$, todos los niveles tienen medias distintas.

Introducción al modelo de efectos aleatorios

- En el modelo de efectos fijos, recordemos que el experimentador elegía los niveles o “tratamientos”.
- Para generalizar este hecho, consideremos ahora que el experimentador elige k muestras de un conjunto mucho más amplio de poblaciones o niveles. En este caso, el modelo se denomina de efectos aleatorios.
- El modelo se expresa de la forma siguiente:

$$X_{ij} = \mu + T_i + E_{ij}, \quad i = 1, \dots, k, \quad j = 1, \dots, n_i,$$

donde

- ▶ μ representa el efecto global medio,
- ▶ $T_i = \mu_i - \mu$, donde μ_i representa la media del i -ésimo nivel elegido aleatoriamente,
- ▶ $E_{ij} = X_{ij} - \mu_i$ es el error residual o residuo.

Supuestos del modelo

- Los k niveles representan muestras aleatorias independientes extraídas de k poblaciones seleccionadas aleatoriamente de un conjunto mayor de poblaciones.
- Todas las poblaciones del conjunto más amplio son normales. Por tanto, cada una de las k poblaciones muestreadas también será normal.
- Todas las poblaciones del conjunto más amplio tienen la misma varianza σ^2 y, por tanto, las k poblaciones muestreadas también tienen la misma varianza σ^2 .
- Las variables T_1, \dots, T_k son variables aleatorias independientes, cada una con media 0 y varianza común σ_{Tr}^2 .

Bloques completamente aleatorizados. Introducción

- Vamos a generalizar el contraste para datos emparejados.
- Más concretamente, supongamos que queremos comparar la media de un determinado tratamiento para k poblaciones en presencia de una variable extraña.
- Para neutralizar el efecto de dicha variable extraña, elegimos un individuo en cada una de las k poblaciones y les aplicamos el mismo tratamiento a todos estos individuos. Al conjunto de dichos individuos se le denomina bloque:

Bloque	Tratamiento			
	Población 1	Población 2	...	Población k
1	X_{11}	X_{21}	...	X_{k1}
2	X_{12}	X_{22}	...	X_{k2}
...
b	X_{1b}	X_{2b}	...	X_{kb}

Bloques completamente aleatorizados. Introducción

- El contraste a realizar es el siguiente:

$$\left. \begin{array}{l} H_0 : \mu_{1\bullet} = \mu_{2\bullet} = \cdots = \mu_{k\bullet}, \\ H_1 : \exists i, j \mid \mu_{i\bullet} \neq \mu_{j\bullet}, \end{array} \right\}$$

donde $\mu_{i\bullet}$ representa la media del i -ésimo tratamiento.

Modelo

- X_{ij} , $i = 1, \dots, k$, $j = 1, \dots, b$ representa la variable aleatoria que designa la respuesta del i -ésimo tratamiento en el j -ésimo bloque.
- El modelo es el siguiente:

$$X_{ij} = \mu + \tau_i + \beta_j + E_{ij}, \quad i = 1, \dots, k, \quad j = 1, \dots, b,$$

donde

- ▶ μ representa el efecto medio global,
- ▶ $\mu_{i\bullet}$ representa la media del tratamiento i -ésimo,
- ▶ $\mu_{\bullet j}$ representa la media del bloque j -ésimo,
- ▶ μ_{ij} representa la media para el i -ésimo tratamiento y el j -ésimo bloque,
- ▶ $\tau_i = \mu_{i\bullet} - \mu$ representa el efecto debido al tratamiento i -ésimo,
- ▶ $\beta_j = \mu_{\bullet j} - \mu$ representa el efecto debido al bloque j -ésimo,
- ▶ $E_{ij} = X_{ij} - \mu_{ij}$ representa el error residual o aleatorio.

Supuestos del modelo

- Las $k \cdot b$ observaciones constituyen muestras aleatorias independientes, cada una de tamaño 1, de $k \cdot b$ poblaciones con medias μ_{ij} , $i = 1, \dots, k$, $j = 1, \dots, b$.
- Cada una de las $k \cdot b$ poblaciones es normal.
- Cada una de las $k \cdot b$ poblaciones tiene la misma varianza σ^2 .
- Los efectos bloque y tratamiento son aditivos; es decir, no hay interacción entre los bloques y los tratamientos. Esto significa que la diferencia de los valores medios para dos tratamientos cualesquiera es la misma para todo un bloque y que la diferencia de los valores medios de dos bloques cualesquiera es la misma para cada tratamiento.

Supuestos del modelo

- Expliquemos el concepto de interacción con un ejemplo.
- Se han desarrollado tres programas para ayudar a que los pacientes que han sufrido un ataque cardíaco por primera vez se adapten física y psicológicamente a su situación. La variable de interés es el tiempo, en meses, necesario para que el paciente puede reiniciar una vida activa. Se sospecha que los varones pueden reaccionar a la enfermedad de forma distinta a las mujeres. Por tanto, se controla la variable de interés anterior mediante bloques: varones y mujeres. De esta forma, tendremos en total $k \cdot b = 3 \cdot 2 = 6$ poblaciones normales, todas ellas con la misma varianza:

Bloque	Tratamiento		
	A	B	C
Varones	$\mu_{11} = 4$	$\mu_{21} = 5$	$\mu_{31} = 7$
Mujeres	$\mu_{12} = 3$	$\mu_{22} = 4$	$\mu_{32} = 6$

Supuestos del modelo

- Podemos observar que las medias de los tratamientos tienen un comportamiento consistente en varones y en mujeres. O sea, μ_{11} es una unidad menor que μ_{21} al igual que μ_{12} también es una unidad menor que μ_{22} y lo mismo ocurre con μ_{11} y μ_{31} al igual que μ_{12} y μ_{32} . Para una interpretación gráfica, véase la figura siguiente:

file=NoInteraccio.pdf,height=4cm,clip=

- Diremos que cuando dicho efecto ocurre, no hay interacción entre los bloques y los tratamientos.

Estadísticos muestrales

- Se introducen los siguientes estadísticos muestrales:

- ▶ $T_{i\bullet} = \sum_{j=1}^b X_{ij}$: suma total de las respuestas al i -ésimo tratamiento, $i = 1, 2, \dots, k$.
- ▶ $\bar{X}_{i\bullet} = \frac{T_{i\bullet}}{b}$: media muestral del i -ésimo tratamiento, $i = 1, 2, \dots, k$.
- ▶ $T_{\bullet j} = \sum_{i=1}^k X_{ij}$: suma total de las respuestas en el j -ésimo bloque, $j = 1, 2, \dots, b$.
- ▶ $\bar{X}_{\bullet j} = \frac{T_{\bullet j}}{k}$: media muestral para el j -ésimo bloque, $j = 1, 2, \dots, b$.
- ▶ $T_{\bullet\bullet} = \sum_{i=1}^k \sum_{j=1}^b X_{ij} = \sum_{i=1}^k T_{i\bullet} = \sum_{j=1}^b T_{\bullet j}$: suma total de las respuestas.
- ▶ $\bar{X}_{\bullet\bullet} = \frac{T_{\bullet\bullet}}{k \cdot b}$: media muestral de todas las respuestas.
- ▶ $\sum_{i=1}^k \sum_{j=1}^b X_{ij}^2$: suma total de cuadrados de cada respuesta.

Ejemplo

Se realiza un experimento para comparar la energía que se requiere para llevar a cabo tres actividades físicas: correr, pasear y montar en bicicleta. La variable de interés es X , número de kilocalorías consumidas por kilómetro recorrido. Se piensa que las diferencias metabólicas entre los individuos pueden afectar al número de kilocalorías requeridas para llevar a cabo una determinada actividad, y se pretende controlar esta variable extraña. Para hacerlo, se seleccionan ocho individuos. Se le pide a cada uno que corra, camine y recorra en bicicleta una distancia medida, y se determina para cada individuo el número de kilocalorías consumidas por kilómetro durante cada actividad.

Ejemplo

Las actividades se realizan en orden aleatorio, con tiempo de recuperación entre una y otra. Cada individuo es utilizado como un bloque. Cada actividad se monitoriza exactamente una vez para cada individuo y de este modo se completa el diseño. Cualquier diferencia en el número medio de kilocalorías consumidas se atribuirá a diferencias entre las actividades mismas, puesto que se ha neutralizado el efecto de las diferencias individuales por medio de la construcción de bloques.

La hipótesis nula es:

$$H_0 : \mu_{1\bullet} = \mu_{2\bullet} = \mu_{3\bullet},$$

donde $\mu_{i\bullet}$, $i = 1, 2, 3$ representa el número medio de kilocalorías consumidas por kilómetro mientras se corre, se pasea o se monta en bicicleta, respectivamente.

Ejemplo

En la tabla siguiente se muestran los resultados obtenidos por los ocho individuos:

Bloque	Tratamiento		
	1 (corriendo)	2 (caminando)	3 (pedaleando)
1	1.4	1.1	0.7
2	1.5	1.2	0.8
3	1.8	1.3	0.7
4	1.7	1.3	0.8
5	1.6	0.7	0.1
6	1.5	1.2	0.7
7	1.7	1.1	0.4
8	2.0	1.3	0.6

Ejemplo

- Almacenamos los datos en la variable `kilocal` en R:

```
> (kilocal <- matrix(c(1.4,1.1,0.7,1.5,1.2,0.8,  
+ 1.8,1.3,0.7,1.7,1.3,0.8,1.6,0.7,0.1,1.5,1.2,  
+ 0.7,1.7,1.1,0.4,2.0,1.3,0.6),8,3,byrow=T))
```

	[,1]	[,2]	[,3]
[1,]	1.4	1.1	0.7
[2,]	1.5	1.2	0.8
[3,]	1.8	1.3	0.7
[4,]	1.7	1.3	0.8
[5,]	1.6	0.7	0.1
[6,]	1.5	1.2	0.7
[7,]	1.7	1.1	0.4
[8,]	2.0	1.3	0.6

Ejemplo

- De cara a trabajar cómodamente vamos a “apilar” la variable anterior de la forma siguiente:

```
kilocal2 <- matrix(kilocal,24,1)
```

```
  [,1]
```

```
[1,] 1.4
```

```
[2,] 1.5
```

```
[3,] 1.8
```

```
[4,] 1.7
```

```
[5,] 1.6
```

```
[6,] 1.5
```

```
[7,] 1.7
```

```
[8,] 2.0
```

```
[9,] 1.1
```

```
[10,] 1.2
```

```
[11,] 1.3
```

```
[12,] 1.3
```

```
[13,] 0.7
```

```
[14,] 1.2
```

```
...
```


Ejemplo

- Añadimos dos variables más que nos dirán el bloque y el tratamiento al que pertenece cada valor de la variable anterior `kilocal2`:

```
> (bloques <- rep(seq(1:8),3))  
[1] 1 2 3 4 5 6 7 8 1 2 3 4 5 6 7 8  
1 2 3 4 5 6 7 8  
> (tratam <- c(rep(1,8),rep(2,8),rep(3,8)))  
[1] 1 1 1 1 1 1 1 1 2 2 2 2 2  
[14] 2 2 2 3 3 3 3 3 3 3 3 3
```

Ejemplo

- A continuación añadimos como columnas las dos variables anteriores a la variable kilocal2:

```
> bloques <- as.factor(bloques)
> tratam <- as.factor(tratam)
> (kilocal2 <- cbind(kilocal2, tratam, bloques))
```

		tratam	bloques
[1,]	1.4	1	1
[2,]	1.5	1	2
[3,]	1.8	1	3
[4,]	1.7	1	4
[5,]	1.6	1	5
[6,]	1.5	1	6
[7,]	1.7	1	7
[8,]	2.0	1	8
[9,]	1.1	2	1
[10,]	1.2	2	2
[11,]	1.3	2	3
[12,]	1.3	2	4
[13,]	0.7	2	5
[14,]	1.2	2	6
[15,]	1.1	2	7

Ejemplo

- Calculemos los estadísticos para nuestro ejemplo:

- ▶ Sumas por bloques:

```
> num_bloques <- 8
> num_tratam <- 3
> sumas_por_bloques <- c()
> for (i in 1:num_bloques){
  sumas_por_bloques <- c(sumas_por_bloques,
    sum(kilocal2[bloques==i,1]))}
> sumas_por_bloques
[1] 3.2 3.5 3.8 3.8 2.4 3.4 3.2 3.9
```

- ▶ Sumas por tratamientos:

```
> sumas_por_tratam <- c()
> for (i in 1:num_tratam){
  sumas_por_tratam <- c(sumas_por_tratam,
    sum(kilocal2[tratam==i,1]))}
> sumas_por_tratam
[1] 13.2 9.2 4.8
```

Ejemplo

- Medias por bloques:

```
> medias_por_bloques <- c()
> for (i in 1:num_bloques){
  medias_por_bloques <- c(medias_por_bloques,
    mean(kilocal2[bloques==i,1]))}
> medias_por_bloques
[1] 1.066667 1.166667 1.266667 1.266667 0.800000
[6] 1.133333 1.066667 1.300000
```

- Medias por tratamientos:

```
> medias_por_tratam <- c()
> for (i in 1:num_tratam){
  medias_por_tratam <- c(medias_por_tratam,
    mean(kilocal2[tratam==i,1]))}
> medias_por_tratam
[1] 1.65 1.15 0.60
```

- Suma total:

```
> (suma_total <- sum(kilocal2[,1]))
[1] 27.2
```

- Media total:

```
> (media_total <- mean(kilocal2[,1]))
[1] 1.133333
```

Identidad de la suma de cuadrados

- Se cumple la igualdad siguiente:

$$\begin{aligned} \sum_{i=1}^k \sum_{j=1}^b (X_{ij} - \bar{X}_{..})^2 &= \sum_{i=1}^k \sum_{j=1}^b (X_{i.} - \bar{X}_{..})^2 \\ &+ \sum_{i=1}^k \sum_{j=1}^b (X_{.j} - \bar{X}_{..})^2 \\ &+ \sum_{i=1}^k \sum_{j=1}^b (X_{ij} - X_{i.} - X_{.j} + \bar{X}_{..})^2, \end{aligned}$$

donde

Identidad de la suma de cuadrados

- $SS_{Total} = \sum_{i=1}^k \sum_{j=1}^b (X_{ij} - \bar{X}_{..})^2$: variabilidad total de los datos.
- $SS_{Tr} = \sum_{i=1}^k \sum_{j=1}^b (X_{i.} - \bar{X}_{..})^2 = b \sum_{i=1}^k (X_{i.} - \bar{X}_{..})^2$: variabilidad de los datos atribuible a la utilización de distintos tratamientos.
- $SS_{Bloques} = \sum_{i=1}^k \sum_{j=1}^b (X_{.j} - \bar{X}_{..})^2 = k \sum_{j=1}^b (X_{.j} - \bar{X}_{..})^2$: variabilidad de los datos atribuible a la utilización de bloques diferentes.
- $SS_E = \sum_{i=1}^k \sum_{j=1}^b (X_{ij} - X_{i.} - X_{.j} + \bar{X}_{..})^2$: variabilidad de los datos debida a factores aleatorios.

Ejemplo

- Calculemos las sumas de cuadrados para nuestro ejemplo:

- ▶ Variabilidad total:

```
> (SS_total <- sum((kilocal2[,1]  
-media_total)^2))  
[1] 5.353333
```

- ▶ Variabilidad debida a tratamientos:

```
> (SS_Tr <- num_bloques*  
sum((medias_por_tratam-media_total)^2))  
[1] 4.413333
```

- ▶ Variabilidad debida a bloques:

```
> (SS_Bl <- num_tratam*  
sum((medias_por_bloques-media_total)^2))  
[1] 0.5533333
```

Ejemplo

- Variabilidad debido al error aleatorio. Creamos dos variables auxiliares aux y aux2 para poder realizar la suma usando las medias por tratamientos y por bloques:

```
> aux <- c()
> for (i in 1:num_tratam){
  aux <- c(aux,rep(medias_por_tratam[i],
    num_bloques))}
> aux
[1] 1.65 1.65 1.65 1.65 1.65 1.65 1.65 1.65 1.15
[10] 1.15 1.15 1.15 1.15 1.15 1.15 1.15 0.60 0.60
[19] 0.60 0.60 0.60 0.60 0.60 0.60
> (aux2 <- rep(medias_por_bloques,num_tratam))
[1] 1.066667 1.166667 1.266667 1.266667 0.800000
[6] 1.133333 1.066667 1.300000 1.066667 1.166667
[11] 1.266667 1.266667 0.800000 1.133333 1.066667
[16] 1.300000 1.066667 1.166667 1.266667 1.266667
[21] 0.800000 1.133333 1.066667 1.300000
```


Ejemplo

- A continuación calculamos la variabilidad del error:

```
> (SS_E <- sum((kilocal2[,1]-aux-aux2  
+media_total)^2))  
[1] 0.3866667
```

- Podemos comprobar que se verifica:

$$SS_{Total} = SS_{Tr} + SS_{Bloques} + SS_E.$$

Contraste a realizar y estadísticos de contraste

- Para contrastar si existen diferencias entre los tratamientos, tenemos que realizar el contraste siguiente:

$$\left. \begin{array}{l} H_0 : \mu_{1\bullet} = \dots = \mu_{k\bullet}, \\ H_1 : \exists i, j = 1, \dots, k \mid \mu_{i\bullet} \neq \mu_{j\bullet}. \end{array} \right\}$$

- Para realizar el contraste anterior, introducimos los estadísticos siguientes:
 - ▶ Cuadrado medio de los tratamientos: $MS_{Tr} = \frac{SS_{Tr}}{k-1}$.
 - ▶ Cuadrado medio de los bloques: $MS_{Bloques} = \frac{SS_{Bloques}}{b-1}$.
 - ▶ Cuadrado medio del error: $MS_E = \frac{SS_E}{(b-1)(k-1)}$.

Estadísticos de contraste

- Los valores esperados de los cuadrados medios son:

$$E(MS_{Tr}) = \sigma^2 + \frac{b}{k-1} \sum_{i=1}^k (\mu_{i\bullet} - \mu)^2, \quad E(MS_E) = \sigma^2.$$

- Para el contraste anterior, usamos como estadístico de contraste el cociente siguiente:

$$F = \frac{MS_{Tr}}{MS_E},$$

que, si H_0 es cierta, sigue la distribución $F_{k-1, (k-1)(b-1)}$ (distribución F de Fisher-Snédecor con $k-1$ y $(k-1)(b-1)$ grados de libertad) ya que, en este caso, $\sum_{i=1}^k (\mu_{i\bullet} - \mu)^2 = 0$ y tanto MS_{Tr} como MS_E estiman σ^2 . Si H_0 no fuese cierta, el valor del estadístico anterior sería mayor que 1.

Tabla del contraste

- Para realizar el contraste se construye la tabla siguiente donde se ha indicado el método de cálculo para calcular los estadísticos:

Fuente de variación	Grados de libertad	Suma de cuadrados	Cuadrados medios	Estadístico
Tratam.	$k - 1$	$\sum_{i=1}^k \frac{T_{i\bullet}}{b} - \frac{T_{\bullet\bullet}^2}{k \cdot b}$	$\frac{SS_{Tr}}{k-1}$	$\frac{MS_{Tr}}{MS_E}$
Bloque	$b - 1$	$\sum_{j=1}^b \frac{T_{\bullet j}}{k} - \frac{T_{\bullet\bullet}^2}{k \cdot b}$	$\frac{SS_{Bloques}}{b-1}$	
Error	$(k - 1)(b - 1)$	$SS_{Total} - SS_{Tr} - SS_{Bloques}$	$\frac{SS_E}{(k-1)(b-1)}$	
Total	$kb - 1$	$\sum_{i=1}^k \sum_{j=1}^b X_{ij}^2 - \frac{T_{\bullet\bullet}^2}{k \cdot b}$		

Ejemplo

- Calculemos la tabla ANOVA para nuestro ejemplo.
- En primer lugar, volvemos a calcular las sumas de cuadrados usando las reglas de cálculo vistas anteriormente:

- ▶ Sumas de cuadrados por tratamientos:

```
> (SS_Tr <- (1/num_bloques)
*sum(sumas_por_tratam^2)-
  suma_total^2/(num_bloques*num_tratam))
[1] 4.413333
```

- ▶ Suma de cuadrados por bloques:

```
> (SS_B1 <- (1/num_tratam)*
sum(sumas_por_bloques^2)-suma_total^2
/(num_bloques*num_tratam))
[1] 0.5533333
```

- ▶ Suma de cuadrados de error:

```
> (SS_E <- sum(kilocal2[,1]^2)-suma_total^2/
  (num_bloques*num_tratam)-SS_Tr-SS_B1)
[1] 0.3866667
```

Ejemplo

- A continuación hacemos la tabla ANOVA:

Fuente de variación	Grados de libertad	Suma de cuadrados	Cuadrados medios	Estadístico
Tratamiento	2	4.413	$\frac{4.413}{2} = 2.207$	$\frac{2.207}{0.028} = 79.897$
Bloque	7	0.553	$\frac{0.553}{7} = 0.079$	
Error	14	0.387	$\frac{0.387}{14} = 0.028$	
Total	23	5.353		

- A un nivel de significación $\alpha = 0.05$ buscamos el valor crítico $F_{0.95,2,14}$:
> qf(0.95,2,14)
[1] 3.738892

Ejemplo

- Como el valor del estadístico de contraste, 79.897, es mayor que el valor crítico, 3.739, rechazamos la hipótesis nula y concluimos que hay diferencias en la energía media requerida dependiendo de las actividades físicas.
- Si hallamos el p-valor del contraste, éste vale:

```
> 1-pf(79.896,2,14)
```

```
[1] 2.201343e-08
```

valor muy pequeño que nos reafirma en la decisión de rechazar la hipótesis nula.

Ejemplo

- En R se puede realizar el contraste ANOVA directamente:

```
> summary(aov(kilocal2[,1]~tratam+bloques))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
tratam	2	4.4133	2.2067	79.8966	2.201e-08
bloques	7	0.5533	0.0790	2.8621	0.04462
Residuals	14	0.3867	0.0276		


```
tratam      ***
bloques     *
Residuals
---
```


Efectividad en la construcción de los bloques

- Nos podemos preguntar si la elección de los bloques ha sido efectiva para nuestro experimento.
- Si la respuesta es afirmativa, la variabilidad debida a los bloques, $SS_{Bloques}$, explicaría una parte importante de la suma total de cuadrados. Este hecho, hace que el valor de la variabilidad del error, SS_E , disminuya, aumentando el valor del estadístico de contraste F y haciendo más difícil aceptar la hipótesis nula H_0 . En este caso, se mejoraría la potencia del contraste.
- Para averiguar la efectividad de la construcción de los bloques, se estima lo que se denomina la eficacia relativa (RE) del diseño de bloque completo aleatorizado comparada con la del diseño completo aleatorizado visto en secciones anteriores.
- La eficacia relativa RE se interpreta com el número de observaciones necesario para que los dos diseños sean equivalentes.

Efectividad en la construcción de los bloques

- Por ejemplo, si $RE = 3$, significa que el diseño completo aleatorizado requiere tres veces tantas observaciones como el diseño de bloque completo aleatorizado para producir un contraste con las mismas características. En este caso, sería deseable la construcción de bloques. En cambio, si $RE = 0.5$, entonces la construcción de bloques no es deseable.
- Si $RE = 1$, los dos métodos son equivalentes cuando los tamaños muestrales son idénticos.
- La manera de estimar RE es la siguiente:

$$\hat{RE} = c + (1 - c) \frac{MS_{bloques}}{MS_E},$$

donde $c = b(k - 1)/(bk - 1)$.

- Si \hat{RE} nos ha dado un valor mayor que 1, significa que la construcción de los bloques ha sido provechosa.

Ejemplo

- En nuestro caso el valor de \hat{RE} vale:

```
> (valor_c <- num_bloques*(num_tratam-1)/  
(num_bloques*num_tratam-1))  
[1] 0.6956522  
> (MS_B1 <- SS_B1/(num_bloques-1))  
[1] 0.07904762  
> (MS_E <- SS_E/((num_bloques-1)*(num_tratam-1)))  
[1] 0.02761905  
> (RE <- valor_c + (1-valor_c)*  
+ MS_B1/MS_E)  
[1] 1.566717
```

- Al darnos un valor mayor que 1, concluimos que la construcción de bloques ha sido útil en nuestro caso.

Comparaciones por parejas y múltiples

- En el caso en que hayamos rechazamos la hipótesis nula, podemos estar interesados en averiguar qué tratamientos son los que difieren.
- Al igual que hacíamos en la clasificación de una vía, en este caso también podemos realizar $\binom{k}{2}$ contrastes t de Bonferroni con el nivel de significación α elegido de tal forma que α' , la probabilidad de realizar un contraste incorrecto, se mantenga bajo control.
- También podemos realizar un contraste de Duncan de rango múltiple con $SSR_p = r_p \sqrt{\frac{MS_E}{b}}$, donde p es el tamaño del subconjunto de medias muestrales elegido.

Ejemplo

- Vamos a realizar el contraste de Duncan de rango múltiple para nuestro ejemplo.
- Recordemos que las medias por tratamientos eran:

$$\bar{X}_{1\bullet} = 1.65, \bar{X}_{2\bullet} = 1.15, \bar{X}_{3\bullet} = 0.60.$$

- A continuación, ordenamos las medias anteriores: $\bar{X}_{3\bullet} < \bar{X}_{2\bullet} < \bar{X}_{1\bullet}$.
- Podemos considerar las medias siguientes:
 - ▶ $\bar{X}_{1\bullet} - \bar{X}_{3\bullet}$ ($p = 3$).
 - ▶ $\bar{X}_{1\bullet} - \bar{X}_{2\bullet}$ ($p = 2$).
 - ▶ $\bar{X}_{2\bullet} - \bar{X}_{3\bullet}$ ($p = 2$).
- La tabla de los valores de r_p y SSR_p para $\alpha = 0.05$ es la siguiente (cogemos 14 como grados de libertad del error):

p	2	3
r_p	3.033	3.178
SSR_p	0.178	0.187

Ejemplo

- Para calcular la tabla anterior, hemos usado el siguiente código R:

```
> rp <- c(3.033,3.178)
> (SSR_p <- rp*sqrt(MS_E/num_bloques))
[1] 0.1782099 0.1867296
```

- Veamos en qué tratamientos hay diferencias:

Diferencias	d	p	SSR_p	$\hat{d} > SSR_p?$	Conclusión
$\bar{X}_{1\bullet} - \bar{X}_{3\bullet}$	1.05	3	0.187	Sí	$\mu_{1\bullet} \neq \mu_{3\bullet}$
$\bar{X}_{1\bullet} - \bar{X}_{2\bullet}$	0.50	2	0.178	Sí	$\mu_{1\bullet} \neq \mu_{2\bullet}$
$\bar{X}_{2\bullet} - \bar{X}_{3\bullet}$	0.55	2	0.178	Sí	$\mu_{2\bullet} \neq \mu_{3\bullet}$

- Concluimos que las medias de los tres tratamientos son diferentes a un nivel de significación de 0.05.

Introducción

- Vamos a generalizar el estudio que hemos realizado hasta ahora en el sentido de que consideraremos que los valores experimentales de los individuos pueden depender de dos o más variables.
- Si éste es el caso, el experimento se denomina **experimento factorial**.
- En este curso, estudiaremos la clasificación de dos vías, diseño completamente aleatorio con efectos fijos.
- Por tanto, en nuestro modelo tendremos dos factores, A y B donde el experimentador ha seleccionado los niveles de cada factor. (efectos fijos)

Formato de los datos y notación

- Como hemos comentado anteriormente, consideraremos que el experimento depende de dos factores, A y B .
- Supondremos que el factor A tiene a niveles y el factor B , b niveles.
- El número total de combinaciones de tratamientos o niveles será: $a \cdot b$.
- Supondremos que tenemos n observaciones para cada combinación de tratamientos. Por tanto, el número total de observaciones será:
$$N = n \cdot a \cdot b.$$
- La variable X_{ijk} , $i = 1, \dots, a$, $j = 1, \dots, b$, $k = 1, \dots, n$ es una variables aleatoria que nos da la respuesta de la k -ésima unidad experimental al i -ésimo nivel del factor A y del j -ésimo nivel del factor B .

Formato de los datos y notación

Todo queda reflejado en la tabla siguiente:

Factor B	Nivel del factor A			
	1	2	\dots	a
1	X_{111}	X_{211}	\dots	X_{a11}
	X_{112}	X_{212}	\dots	X_{a12}
	\dots	\dots	\dots	\dots
	X_{11n}	X_{21n}	\dots	X_{a1n}
2	X_{121}	X_{221}	\dots	X_{a21}
	X_{122}	X_{222}	\dots	X_{a22}
	\dots	\dots	\dots	\dots
	X_{12n}	X_{22n}	\dots	X_{a2n}
\vdots	\vdots	\vdots	\vdots	\vdots
b	X_{1b1}	X_{2b1}	\dots	X_{ab1}
	X_{1b2}	X_{2b2}	\dots	X_{ab2}
	\dots	\dots	\dots	\dots
	X_{1bn}	X_{2bn}	\dots	X_{abn}

Modelo

- El modelo es el siguiente:

$$X_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + E_{ijk}, \\ i = 1, \dots, a, \quad j = 1, \dots, b, \quad k = 1, \dots, n,$$

- donde

- ▶ μ : efecto medio global,
- ▶ $\mu_{i\bullet\bullet}$: media para el i -ésimo nivel del factor A .
- ▶ $\mu_{\bullet j\bullet}$: media para el j -ésimo nivel del factor B .
- ▶ $\mu_{ij\bullet}$: media para la $i - j$ -ésima combinación de tratamientos.
- ▶ $\alpha_i = \mu_{i\bullet\bullet} - \mu$: efecto debido al hecho de que la unidad experimental está en el i -ésimo nivel del factor A .
- ▶ $\beta_j = \mu_{\bullet j\bullet} - \mu$: efecto debido al hecho de que la unidad experimental está en el j -ésimo nivel del factor B .
- ▶ $(\alpha\beta)_{ij} = \mu_{ij\bullet} - \mu_{i\bullet\bullet} - \mu_{\bullet j\bullet} + \mu$: efecto de interacción entre el i -ésimo nivel del factor A y el j -ésimo nivel del factor B .
- ▶ $E_{ijk} = X_{ijk} - \mu_{ij\bullet}$: error residual o aleatorio.

Supuestos del modelo

- Nuestro modelo tiene las suposiciones siguientes:
 - ▶ Las observaciones para cada combinación de tratamientos constituyen muestras aleatorias independientes, cada una de tamaño n , de $a \cdot b$ poblaciones con media $\mu_{ij\bullet}$.
 - ▶ Cada una de las $a \cdot b$ poblaciones es normal.
 - ▶ Cada una de las $a \cdot b$ poblaciones tiene la misma varianza, σ^2 .

Estadísticos muestrales

Se introducen los siguientes estadísticos muestrales:

- $T_{ij\bullet} = \sum_{k=1}^n$: suma total de las respuestas al i -ésimo nivel del factor A y al j -ésimo nivel del factor B .
- $\bar{X}_{ij\bullet} = \frac{T_{ij\bullet}}{n}$: media muestral para la $i - j$ -ésima combinación de tratamientos.
- $T_{i\bullet\bullet} = \sum_{j=1}^b T_{ij\bullet}$: suma total de las respuestas al i -ésimo nivel del factor A .
- $\bar{X}_{i\bullet\bullet} = \frac{T_{i\bullet\bullet}}{bn}$: media muestral para el i -ésimo nivel del factor A .

Estadísticos muestrales

- $T_{\bullet j \bullet} = \sum_{i=1}^a T_{ij \bullet}$: suma total de las respuestas al j -ésimo nivel del factor B .
- $\bar{X}_{\bullet j \bullet} = \frac{T_{\bullet j \bullet}}{an}$: media muestral para el j -ésimo nivel del factor B .
- $T_{\bullet \bullet \bullet} = \sum_{i=1}^a T_{i \bullet \bullet} = \sum_{j=1}^b T_{\bullet j \bullet} = \sum_{i=1}^a \sum_{j=1}^b T_{ij \bullet}$: suma total de las respuestas.
- $\bar{X}_{\bullet \bullet \bullet} = \frac{T_{\bullet \bullet \bullet}}{abn}$: media muestral para todas las respuestas.

Ejemplo

Estudio del efecto de la luz y la temperatura sobre el crecimiento del ovario del pez *Mirogrex terrae-sanctae*

El *Mirogrex terrae-sanctae* es un pez comercializado semejante a la sardina que se encontró en el Mar de Galilea. Se realizó un estudio para determinar el efecto de la luz y la temperatura sobre el índice gonadosomático (GSI), que es una medida de crecimiento del ovario. Se utilizaron dos fotoperíodos: catorce horas de luz, diez horas de oscuridad y nueve horas de luz, quince horas de oscuridad; y dos niveles de temperatura, 16 y 27 °C. De este modo, el experimentador puede simular situaciones de verano e invierno en la región. Se trata de un experimento factorial con dos factores, luz y temperatura, que son investigados cada uno a dos niveles.

Ejemplo

Datos obtenidos

El experimento se realizó sobre 20 hembras en junio. Se dividieron aleatoriamente las 20 hembras en 4 subgrupos de tamaño 5 cada uno. Después de tres meses se determinó el *GSI* para cada pez. Los resultados se muestran en la tabla siguiente:

Factor <i>B</i> (temperatura)	Factor <i>A</i> (fotoperíodo)	
	9 horas	14 horas
27°C	0.90	0.83
	1.06	0.67
	0.98	0.57
	1.29	0.47
	1.12	0.66
16°C	1.30	1.01
	2.88	1.52
	2.42	1.02
	2.66	1.32
	2.94	1.63

Ejemplo. Introducción de los datos en R

- En primer lugar, introducimos tres variables:
 - ① peces donde introduciremos los valores de la variable *GSI* de cada pez,
 - ② periodo donde introduciremos el valor del nivel para el factor *A* (fotoperíodo) y
 - ③ temperatura donde introduciremos el valor del nivel para el factor *B* (temperatura).

Ejemplo. Introducción de los datos en R

```
> (peces <- c(0.90,0.83,1.06,0.67,0.98,  
0.57,1.29,0.47,1.12,0.66,1.30,1.01,2.88,  
1.52,2.42,1.02,2.66,1.32,2.94,1.63))  
[1] 0.90 0.83 1.06 0.67 0.98 0.57 1.29  
[8] 0.47 1.12 0.66 1.30 1.01 2.88 1.52  
[15] 2.42 1.02 2.66 1.32 2.94 1.63  
> (periodo <- factor(rep(c(9,14),10)))  
[1] 9 14 9 14 9 14 9 14 9 14  
[12] 14 9 14 9 14 9 14 9 14  
Levels: 9 14  
> (temperatura <- factor(c(rep(27,10),  
rep(16,10))))  
[1] 27 27 27 27 27 27 27 27 27 27 16  
[12] 16 16 16 16 16 16 16 16 16 16  
Levels: 16 27
```

Ejemplo. Introducción de los datos en R

A continuación, juntamos las tres variables anteriores en un sólo data frame llamado peces2:

```
> (peces2 <- data.frame(peces,periodo,temperatura))
```

	peces	periodo	temperatura
1	0.90	9	27
2	0.83	14	27
3	1.06	9	27
4	0.67	14	27
5	0.98	9	27
6	0.57	14	27
7	1.29	9	27
8	0.47	14	27
9	1.12	9	27
10	0.66	14	27
11	1.30	9	16
12	1.01	14	16
13	2.88	9	16
14	1.52	14	16
15	2.42	9	16
16	1.02	14	16
...			

Ejemplo. Cálculo de los estadísticos muestrales

- Suma total de las respuestas al nivel i -ésimo para el factor A y al nivel j -ésimo para el factor B ($T_{ij\bullet}$):

```
> sumas_A_B <- c()
> for (i in 1:2){aux <- c(); for (j in 1:2) {
aux <- c(aux, sum(peces2[periodo==
unique(periodo)[i]& temperatura==
unique(temperatura)[j],1]))};
sumas_A_B <- cbind(sumas_A_B,aux)}
> sumas_A_B
      aux aux
[1,]  5.35 3.2
[2,] 12.20 6.5
```

- Media muestral para la $i - j$ -ésima combinación de tratamientos ($\bar{X}_{ij\bullet}$):

```
> (medias_A_B <- sumas_A_B/5)
      aux aux
[1,]  1.07 0.64
[2,]  2.44 1.30
```

Ejemplo. Cálculo de los estadísticos muestrales

- Suma total de las respuestas al i -ésimo nivel del factor A ($T_{i\bullet\bullet}$):

```
> sumas_A <- c()  
> for (i in 1:2){ sumas_A <- c(sumas_A,  
sum(peces2[periodo==unique(periodo)[i],1]))}  
> sumas_A  
[1] 17.55  9.70
```

- Media muestral para el i -ésimo nivel del factor A ($\bar{X}_{i\bullet\bullet}$):

```
> (medias_A <- sumas_A/(5*2))  
[1] 1.755 0.970
```

Ejemplo. Cálculo de los estadísticos muestrales

- Suma total de las respuestas al j -ésimo nivel del factor B ($T_{\bullet j \bullet}$):

```
> sumas_B <- c()  
> for (j in 1:2){ sumas_B <- c(sumas_B,  
sum(peces2[temperatura==  
unique(temperatura)[j],1]))}  
> sumas_B  
[1] 8.55 18.70
```

- Media muestral para el j -ésimo nivel del factor B ($\bar{X}_{\bullet j \bullet}$):

```
> (medias_B <- sumas_B/(5*2))  
[1] 0.855 1.870
```

Ejemplo. Cálculo de los estadísticos muestrales

- Suma total de las respuestas ($T_{\bullet\bullet\bullet}$):

```
> (suma_total <- sum(peces2[,1]))  
[1] 27.25
```

- Media total de las respuestas ($\bar{X}_{\bullet\bullet\bullet}$):

```
> (media_total <- suma_total/(5*2*2))  
[1] 1.3625
```

Identidad de la suma de cuadrados

Definimos las sumas de cuadrados siguientes:

- Variabilidad total: $SS_{Total} = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^n (X_{ijk} - \bar{X}_{\bullet\bullet\bullet})^2$.
- Variabilidad debida a la utilización de diferentes niveles del factor A :
 $SS_A = bn \sum_{i=1}^a (\bar{X}_{i\bullet\bullet} - \bar{X}_{\bullet\bullet\bullet})^2$.
- Variabilidad debida a la utilización de diferentes niveles del factor B :
 $SS_B = an \sum_{j=1}^b (\bar{X}_{\bullet j\bullet} - \bar{X}_{\bullet\bullet\bullet})^2$.
- Variabilidad debida a la interacción entre niveles de los factores A y B : $SS_{AB} = n \sum_{i=1}^a \sum_{j=1}^b (X_{ij\bullet} - \bar{X}_{i\bullet\bullet} - \bar{X}_{\bullet j\bullet} + \bar{X}_{\bullet\bullet\bullet})^2$.
- Variabilidad debida al error aleatorio: $SS_E = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^n (X_{ijk} - \bar{X}_{ij\bullet})^2$.

Cálculo de las sumas de cuadrados para el ejemplo

- Variabilidad total, SS_{Total} :

```
> (SS_total <- sum((peces2[,1]-media_total)^2))  
[1] 11.13377
```

- Variabilidad debida al factor A, SS_A :

```
> (SS_A <- 2*5*sum((medias_A-media_total)^2))  
[1] 3.081125
```

- Variabilidad debida al factor B, SS_B :

```
> (SS_B <- 2*5*sum((medias_B-media_total)^2))  
[1] 5.151125
```


Cálculo de las sumas de cuadrados para el ejemplo

- Variabilidad debida a la interacción entre A y B , SS_{AB} :

```
> SS_AB <- 0
> for (i in 1:2){for (j in 1:2){
SS_AB <- SS_AB+(medias_A_B[j,i]
-medias_A[i]-medias_B[j]+
media_total)^2}}
> (SS_AB <- 5*SS_AB)
      aux
0.630125
```

Cálculo de las sumas de cuadrados para el ejemplo

- Variabilidad debida al error aleatorio, SS_E :

```
> SS_E <- 0
> for (i in 1:2){for (j in 1:2) {
SS_E <- SS_E +sum((peces2[
periodo==unique(periodo)[i]&
temperatura==unique(temperatura)[j],1]
-medias_A_B[j,i])^2)}}
> SS_E
[1] 2.2714
```

- Podemos comprobar que se verifica:

$$SS_{Total} = SS_A + SS_B + SS_{AB} + SS_E.$$

Cálculo de las sumas de cuadrados

De cara a simplificar los cálculos, se usan las fórmulas siguientes para calcular las sumas de cuadrados:

- $SS_A = \frac{1}{bn} \sum_{i=1}^a T_{i\bullet\bullet}^2 - \frac{T_{\bullet\bullet\bullet}^2}{abn}.$
- $SS_B = \frac{1}{an} \sum_{j=1}^b T_{\bullet j\bullet}^2 - \frac{T_{\bullet\bullet\bullet}^2}{abn}.$
- $SS_{Total} = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^n X_{ijk}^2 - \frac{T_{\bullet\bullet\bullet}^2}{abn}.$
- $SS_{AB} = SS_{Tr} - SS_A - SS_B$, donde $SS_{Tr} = \frac{1}{n} \sum_{i=1}^a \sum_{j=1}^b T_{ij\bullet}^2 - \frac{T_{\bullet\bullet\bullet}^2}{abn}.$
- $SS_E = SS_{Total} - SS_{Tr}.$

Ejemplo

Vamos a usar las fórmulas anteriores para recalcular las sumas de cuadrados:

- SS_A :

```
> (SS_A <- (1/(2*5))*sum(sumas_A^2)
-(1/(2*2*5))*suma_total^2)
[1] 3.081125
```

- SS_B :

```
> (SS_B <- (1/(2*5))*sum(sumas_B^2)
-(1/(2*2*5))*suma_total^2)
[1] 5.151125
```

- SS_{Total} :

```
> (SS_total <- sum(peces2[,1]^2)
-(1/(2*2*5))*suma_total^2)
[1] 11.13377
```

Ejemplo

- SS_{AB} :

```
> (SS_Tr <- (1/5)*sum(sumas_A_B^2)
-(1/(2*2*5))*suma_total^2)
[1] 8.862375
> (SS_AB <- SS_Tr-SS_A-SS_B)
[1] 0.630125
```

- SS_E :

```
> (SS_E <- SS_total - SS_Tr)
[1] 2.2714
```

Contrastes a realizar

- Cuando realizamos un experimento factorial de dos vías, hemos de tener en cuenta los contrastes siguientes:

- 1 Contraste de no interacción. Contrastamos que no hay interacción entre los factores A y B :

$$\left. \begin{array}{l} H_0 : (\alpha\beta)_{ij} = 0, \\ H_1 : \exists i, j \mid (\alpha\beta)_{ij} \neq 0. \end{array} \right\}$$

- 2 Contraste de que no hay diferencias entre los niveles del factor A :

$$\left. \begin{array}{l} H_0 : \mu_{1\bullet\bullet} = \mu_{2\bullet\bullet} = \cdots = \mu_{a\bullet\bullet}, \\ H_1 : \exists i, i' \mid \mu_{i\bullet\bullet} \neq \mu_{i'\bullet\bullet}. \end{array} \right\}$$

- 3 Contraste de que no hay diferencias entre los niveles del factor B :

$$\left. \begin{array}{l} H_0 : \mu_{\bullet 1\bullet} = \mu_{\bullet 2\bullet} = \cdots = \mu_{\bullet b\bullet}, \\ H_1 : \exists j, j' \mid \mu_{\bullet j\bullet} \neq \mu_{\bullet j'\bullet}. \end{array} \right\}$$

Estadísticos de contraste

Los estadísticos de contraste que vamos a usar en los contrastes anteriores son los siguientes:

- Contraste de no interacción.

$$F = \frac{MS_{AB}}{MS_E},$$

donde $MS_{AB} = \frac{SS_{AB}}{(a-1)(b-1)}$, $MS_E = \frac{SS_E}{ab(n-1)}$ y la distribución del estadístico F , si la hipótesis nula es cierta, es la distribución F de Fisher-Snédecor con $(a-1)(b-1)$ y $ab(n-1)$ grados de libertad ($F_{(a-1)(b-1), ab(n-1)}$).

Estadísticos de contraste

- Contraste de que no hay diferencias entre los niveles del factor A :

$$F = \frac{MS_A}{MS_E},$$

donde $MS_A = \frac{SS_A}{a-1}$ y el estadístico F , si H_0 es cierta, sigue la distribución F de Fisher-Snédecor con $a - 1$ y $ab(n - 1)$ grados de libertad ($F_{a-1, ab(n-1)}$).

- Contraste de que no hay diferencias entre los niveles del factor B :

$$F = \frac{MS_B}{MS_E},$$

donde $MS_B = \frac{SS_B}{b-1}$ y el estadístico F , si H_0 es cierta, sigue la distribución F de Fisher-Snédecor con $b - 1$ y $ab(n - 1)$ grados de libertad ($F_{b-1, ab(n-1)}$).

Tabla ANOVA

Los tres contrastes anteriores se resumen en la siguiente tabla:

Variación	Grados de libertad	Suma de cuadrados	Cuadrados medios	Estadísticos F
Tratam.	$ab - 1$	SS_{Tr}	$SS_{Tr}/(ab - 1)$	MS_{Tr}/MS_E
A	$a - 1$	SS_A	$SS_A/(a - 1)$	MS_A/MS_E
B	$b - 1$	SS_B	$SS_B/(b - 1)$	MS_B/MS_E
AB	$(a - 1)(b - 1)$	SS_{AB}	$SS_{AB}/((a - 1)(b - 1))$	MS_{AB}/MS_E
Error	$ab(n - 1)$	SS_E	$SS_E/(ab(n - 1))$	
Total	$abn - 1$	SS_{Total}		

Tabla ANOVA para el ejemplo

Variación	Grados de libertad	Suma de cuadrados	Cuadrados medios	Estadísticos F
Tratam.	3	8.862	2.954	20.809
<i>A</i>	1	3.081	3.081	21.704
<i>B</i>	1	5.151	5.151	36.285
<i>AB</i>	1	0.630	0.630	4.439
Error	16	2.271	0.142	
Total	19	11.134		

p valores para los contrastes del ejemplo

Los p -valores para los tres contrastes valen en nuestro ejemplo:

- Contraste de no interacción entre los factores A y B :

$p = p(F_{1,16} > 4.439) \approx 0.051$. Es un p -valor pequeño. Por tanto, rechazaríamos la hipótesis nula y concluiríamos que sí hay interacción entre el fotoperíodo y la temperatura.

```
> 1-pf(4.439,1,16)
[1] 0.05126085
```

- Contraste de que no hay diferencias entre los niveles del factor A (fotoperíodo):

$p = p(F_{1,16} > 21.704) \approx 0.00026$. Es un valor muy pequeño; por tanto, rechazamos H_0 y concluimos que hay diferencias entre los niveles del fotoperíodo.

```
> 1-pf(21.704,1,16)
[1] 0.0002621106
```

- Contraste de que no hay diferencias entre los niveles del factor B (temperatura):

$p = p(F_{1,16} > 36.285) \approx 0.000018$. Es un valor muy pequeño; por tanto, rechazamos H_0 y concluimos que hay diferencias entre los niveles de la temperatura.

```
> 1-pf(36.285,1,16)
[1] 1.771135e-05
```

Resultado directo del ejemplo con R

Con R podemos obtener la tabla ANOVA anterior directamente:

```
> summary(aov(peces2[,1]~periodo*temperatura))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
periodo	1	3.0811	3.0811	21.7038	0.0002621	***
temperatura	1	5.1511	5.1511	36.2851	1.771e-05	***
periodo:temperatura	1	0.6301	0.6301	4.4387	0.0512685	.
Residuals	16	2.2714	0.1420			

Comparaciones usando el contraste de Duncan

- En el caso en que hayamos rechazado una (o las dos hipótesis nulas) respecto a las diferencias entre los niveles de los factores A o B , hemos de ver en qué niveles hay diferencias.
- Para hacerlo, podemos usar el contraste de Duncan comentado en secciones anteriores.
- En el caso de ver entre qué niveles del factor A hay diferencias, hay que usar el siguiente contraste de Duncan:

$$SSR_p = r_p \sqrt{\frac{MS_E}{bn}}.$$

- En el caso de ver entre qué niveles del factor B hay diferencias, hay que usar el siguiente contraste de Duncan:

$$SSR_p = r_p \sqrt{\frac{MS_E}{an}}.$$

Apuntes de Bioestadística.

R. Alberich y A. Mir

Departamento de Matemáticas e
Informàtica
Universitat Illes Balears

14 de julio de 2010

38 Pruebas de bondad de ajuste

- Un contraste de bondad de ajuste: distribución totalmente conocida

Introducción

- A lo largo de este tema se han tratado contrastes estadísticos de hipótesis sencillos para distintos parámetros y en muchos casos se ha supuesto que la población era normal. Consideraremos ahora una prueba para determinar si una población tiene una determinada distribución teórica.
- La prueba se basará en la diferencia entre las frecuencias observadas en la muestra y las frecuencias que se obtendrían con la distribución hipotética.

Ejemplo

Veamos el ejemplo más sencillo.

- Queremos saber si un dado está bien balanceado, es decir si la distribución teórica del dado es $P_X(x) = \frac{1}{6}$ para $x = 1, 2, 3, 4, 5, 6$.
- Supongamos que lanzamos el dado 120 veces y anotamos cada uno de los resultados. En teoría si el dado no está cargado esperaríamos obtener 20 veces cada resultado.
- Los resultados de la muestra se dan en la siguiente tabla:

Ejemplo

Frecuencia	Valor obtenido en el lanz.					
	1	2	3	4	5	6
Observada (o_i)	20	22	17	18	19	24
Esperada (e_i) si H_0 es cierta	20	20	20	20	20	20

Tenemos que “medir” de alguna manera la “distancia” entre los resultados observados y los teóricos.

Como vemos en la tabla tenemos que comparar $k = 6$ valores.

Un contraste de bondad de ajuste: distribución totalmente conocida

- Supongamos que tenemos n ($n \geq 25$ o 30) observaciones de las que se calculan sus frecuencias observadas en k clases (que debe ser $k \geq 5$).
- Queremos contrastar si los datos siguen una distribución totalmente conocida, es decir conocemos la forma de la distribución de contraste y todos sus parámetros.
- Denotemos por O_i las frecuencias absolutas observadas y por e_i las frecuencias esperadas condicionadas a que H_0 .
- Las frecuencias esperadas son $e_i = n \cdot p_i$ donde $p_i = P(\text{Clase } i / H_0)$. Los datos tienen esta distribución. Las hipótesis del contraste son:

$$\begin{cases} H_0 : \text{La población tiene esta distribución} \\ H_1 : \text{La población tiene otra distribución} \end{cases}$$

- Entonces el estadístico de contraste es:

$$\chi^2_{k-1} = \sum_{i=1}^k \frac{(o_i - e_i)^2}{e_i}$$

- Este estadístico tiene aproximadamente una distribución χ^2 con $k - 1$ g.l si todas las **frecuencias absolutas esperadas superan 5**.
- En contrario se pueden reagrupar los intervalos que no cumplan la condición con sus adyacentes. Reduciéndose los grados de libertad
- La regla de rechazo al nivel de confianza α es:
Rechazar H_0 si:

$$\chi^2_{k-1} > \chi^2_{k-1, 1-\alpha}$$

Ejemplo.

¿Podemos afirmar al nivel de significación $\alpha = 0.05$ que el dado del ejemplo anterior está bien balanceado (y por lo tanto conocemos su distribución y sus parámetros) a la vista de la muestra?

Solución

: Bajo estas condiciones conocemos completamente la distribución teórica, $k = 6$ y las frecuencia absolutas teóricas son superiores a 5 entonces:

$$\chi^2_{k-1} = \chi^2_5 = \frac{(20-20)^2}{20} + \frac{(22-20)^2}{20} + \frac{(17-20)^2}{20} + \frac{(18-20)^2}{20} + \frac{(19-20)^2}{20} + \frac{(24-20)^2}{20} = 1.7 \not\geq \chi^2_{5,1-\alpha} = \chi^2_{5,1-0.05} = 11.071.$$

No podemos rechazar H_0 al nivel de significación $\alpha = 0.05$.

Ejemplo.

Un ecólogo quiere estudiar el aumento de temperatura del agua a dos kilómetros de los vertidos de agua autorizados que realiza una planta industrial. El ecólogo afirma que el aumento de temperatura después de los vertidos no sigue una distribución normal. Lo que podría indicar que la empresa vierte en ocasiones agua demasiado caliente. La empresa afirma que la media del aumento de la temperatura es de 3.5 décimas de grado centígrado y con una desviación típica de 0.7 y que sigue una ley normal. El ecólogo toma muestra aleatoria de aumento de las temperaturas obteniéndose los siguientes resultados:

Límites de la clase	o_i
1.45 - - 1.95	2
1.95 - - 2.45	1
2.45 - - 2.95	4
2.95 - - 3.45	15
3.45 - - 3.95	10
3.95 - - 4.45	5
4.45 - - 4.95	3

¿A la vista de estos datos podemos afirmar que la información sobre la distribución de los datos de aumento de la temperatura que aporta la planta industrial es cierta al nivel de significación del 5 %?

Solución

: El contraste es:

$$\left\{ \begin{array}{l} H_0 : \text{La distribución de duración es normal} \\ \quad \text{con } \mu = 3.5 \text{ y } \sigma = 0.7 \\ H_1 : \text{La distribución es cualquier otra} \end{array} \right.$$

Solución

- Vamos a realizar el contraste de bondad de ajuste, para ello tenemos que calcular las frecuencias esperadas. La muestra es de tamaño $n = 40$.
- Sea X = el incremento de temperatura en una observación pila escogida al azar.
- Entonces: $P(1.95 \leq X \leq 2.45 / H_0) = P(1.95 \leq X \leq 2.45 / X \text{ sigue una distribución normal con } \mu = 3.5 \text{ y } \sigma = 0.7) = P\left(\frac{1.95-3.5}{0.7} \leq Z < \frac{2.45-3.5}{0.7}\right) = F_Z(-1.5) - F_Z(-2.21) \approx (1 - 0.9332) - (1 - 0.9864) = 0.0532$.

Solución

- Entonces la frecuencia esperada entre 40 pilas para el intervalo 1.95 - 2.45 es $e_1 = (40) \cdot (0.0532) = 2.128 \approx 2.1$ (nota: en este último cálculo se suele aproximar al primer decimal).
- De forma análoga se calculan los demás e_i y se obtienen los siguientes resultados:

El resto de resultados se resumen en la tabla siguiente:

Límites de la clase	o_i		e_i	
menor que 1.95	2		0.5	
1.95 - - 2.45	1		2.1	
2.45 - - 2.95	4	7	5.9	8.5
2.95 - - 3.45	15	15	10.3	10.3
3.45 - - 3.95	10	10	10.7	10.7
3.95 - - 4.45	5		7.0	
mayor que 4.45	3	8	3.5	10.5

Solución

- Donde se observa que las frecuencias esperadas de los dos primeros intervalos y el del último no superan 5. Así que agrupamos los tres primeros intervalos en uno y los dos últimos también, de forma que las frecuencias esperadas y observadas quedan como en las segundas columnas.
- Entonces $k = 4$ y el estadístico de contraste es

$$\chi_{k-1}^2 = \chi_3^2 = \frac{(7 - 8.5)^2}{8.5} + \frac{(15 - 10.3)^2}{10.3} + \frac{(10 - 10.7)^2}{10.7} + \frac{(8 - 10.5)^2}{10.5} = 3.05$$

- Ahora tenemos que $\chi_{k-1}^2 = 3.05 \not> \chi_{k-1, 1-\alpha}^2 = \chi_{3, 1-0.05}^2 = 7.815$ no hay razón para rechazar la hipótesis nula al nivel de significación $\alpha = 0.05$.
- **Nota:** Como se observa en la tabla el primer y último intervalo se consideran con toda la cola de la probabilidad.

Un contraste de bondad de ajuste: algún parámetro poblacional desconocido

- Supongamos que queremos contrastar si una población tiene una distribución por ejemplo normal, Poisson....
- Pero que no conocemos, o podemos determinar, los parámetros de estas distribuciones.
- Por ejemplo en la normal no conocemos μ o σ o ambas.
- El contraste que tenemos que realizar es similar al anterior pero el número de grados de libertad del estadístico de contraste será $k - m - 1$ donde k es el número de categorías y m es el número de parámetros que se estiman.

Ejemplo.

Durante la segunda guerra mundial se dividió el mapa de Londres en cuadrículas de 0.25 Km^2 y se contó el número de bombas caídas en cada cuadrícula durante un bombardeo alemán. Los resultados fueron:

num. impactos en la cuad. (x_i)	0	1	2	3	4	5
frecuencia (o_i)	229	211	93	35	7	1

Si realmente los bombardeos no seguían un plan prefijado la distribución del número de bombas en cada cuadrícula tendría que ser una $Po(\lambda)$. Contrastar esta hipótesis al nivel de significación $\alpha = 0.05$.

Solución:

- Sabemos el tipo de distribución pero no conocemos el parámetro λ lo tendremos que estimar por $\lambda = \frac{\sum_{i=0}^5 x_i o_i}{\sum_{i=0}^5 o_i} = \frac{535}{576} = 0.929$
- Calculemos las frecuencias esperadas e_i cuando la distribución de X =número de bombas por cuadrícula es una $Po(0.929)$, como sabemos que

$$P_X(x_i) = P(X = x_i) = \frac{0.929^{x_i}}{x_i!} e^{-0.929} = p_i$$

por lo tanto

x_i	0	1	2	3	4	$5 \geq$
p_i	.395	.367	.17	.053	0.012	0.003
$e_i =$ $p_i \cdot 576$	227.5	211.4	97.9	30.5	6.9	1.7

Tendremos que agrupar las dos últimas columnas. En resumen:

x_i	0	1	2	3	$4 \geq$
o_i	229	211	93	35	8
e_i	227.5	211.4	97.9	30.5	8.6

Solución

- Entonces tenemos que el número de clases es $k = 5$ y el número de parámetros estimados es $m = 1$.
- Entonces $\chi^2_{k-m-1} = \chi^2_3 = 0.961692$ y como $\chi^2_{3,1-0.05} = 7.815$ por lo tanto $0.961692 \not> 7.815$.
- No podemos rechazar la hipótesis nula con este nivel de significación.
- Por lo tanto podemos afirmar que el bombardeo era aleatorio y que no estaba dirigido a objetivos militares.

Prueba de bondad de ajuste de Kolgomorov-Smirnov (K-S)

El contraste de Kolgomorov-Smirnov es conocido con el acrónimo K-S. Dada una ley de **distribución continua** F el test K-S contrata las siguientes hipótesis:

$$\begin{cases} H_0 : \text{La distribución de la muestra sigue la ley de distribución } F(x) \\ H_1 : \text{no sigue esa ley de distribución} \end{cases}$$

- En principio la ley de distribución F puede ser cualquier **distribución continua**: normal, exponencial, uniforme, etc.
- Aunque en casos particulares, por ejemplo normalidad existen mejora de este test (por ejemplo para normalidad para algunos tamaños muestrales se debe aplicar el test de Kolgomorov-Smirnov-Lilliefors⁴)

⁴Ver: Daniel Peña, Sánchez Rivera. "*Estadística Modelos y métodos. 1 Fundamentos*". Segunda Edición. Ed. Alianza Universidad Textos. 1991. Pág.369

Test K-S

- Este contraste parte de una muestra aleatoria de una cierta variable X : x_1, x_2, \dots, x_n . A la muestra ordenada la denotaremos por $x_{(1)} \leq x_{(2)}, \dots, \leq x_{(n)}$.
- Entonces podemos definir la función de distribución muestral (empírica) de la variable X para su muestra de tamaño n a la que denotaremos por F_n .
- Donde

$$F_n(X) = \begin{cases} 0 & \text{si } x < x_{(1)} \\ \frac{k}{n} & \text{si } x_{(k)} \leq x \leq x_{(k+1)} \\ 1 & \text{si } x \geq x_n \end{cases}$$

- Definimos la máxima discrepancia para cada observación como:

$$D_n(x_h) = \max\{|F_n(x_{h-1}) - F(x_h)|, |F_n(x_h) - F(x_h)|\}$$

- Ahora definimos el estadístico D_n como la mayor de las discrepancias:

$$D_n = \max_{h=1, \dots, n} D_n(x_h)$$

- La regla de decisión es rechazar H_0 al nivel α si

$$D_n \geq D_{n,\alpha}$$

- Donde $D_{n,\alpha}$ es el cuantil de la distribución del test de Kolgomorv-Smirnov que podréis encontrar en las tablas de campus extens.

Ejemplo

Consideremos una serie de tiempos de vida de un cierto componente electrónico:

16, 8, 9, 12, 6, 11, 20, 7, 2, 24

Vamos a contrastar si proviene de una distribución exponencial.

Estimaremos primero el parámetro λ de la exponencial por la media muestral $\bar{x} = 11.5$

$$\begin{cases} H_0 : \text{los datos provienen de una } \exp(\frac{1}{11.5}) \\ H_1 : \text{siguen otra distribución} \end{cases}$$

Ejemplo

La muestra ordenada y los cálculos necesarios se muestran en la siguiente tabla:

h	x_h	$F_n(x_h)$	$ F(x_h) = 1 - e^{-x/11.5} $	$ F_n(x_{h-1}) - F(x_h) $	$ F_n(x_h) - F(x_h) $	máx
1	2	0.1	0.16	0.16	0.06	0.16
2	6	0.2	0.41	0.31	0.21	0.31
3	7	0.3	0.46	0.26	0.16	0.26
4	8	0.4	0.50	0.2	0.1	0.2
5	9	0.5	0.54	0.14	0.04	0.14
6	11	0.6	0.62	0.12	0.02	0.12
7	12	0.7	0.65	0.05	0.15	0.15
8	16	0.8	0.75	0.05	0.05	0.05
9	20	0.9	0.82	0.02	0.08	0.08
10	24	1	0.88	0.02	0.12	0.12

$D_n = 0.31$

Ejemplo

- Consultando la tablas del test K-S se tiene que $D_{10,0.01} = 0.490$ y $D_{10} = 0.31 \not\geq 0.490$, el estadístico no está en la región de rechazo y por lo tanto no podemos rechazar que estos datos se ajusten a esa exponencial al nivel $\alpha = 0.01$.
- El código en R para este ejemplo es:

```
> ks.test(c(16, 8, 9, 12, 6, 11, 20, 7, 2, 24), "pexp", 1/
```

One-sample Kolmogorov-Smirnov test

```
data: c(16, 8, 9, 12, 6, 11, 20, 7, 2, 24)
```

```
D = 0.3065, p-value = 0.2486
```

```
alternative hypothesis: two-sided
```

- Notemos** que lo que la información que se nos da es el p -valor del test.

Parte IX

Fundamentos Estadísticos.

- En esta parte veremos las definiciones y fundamentos básicos de las variables aleatorias vectoriales.
- El objetivo es conocer estas construcciones y distinguirlas de las de las muestras de datos multivariantes.
- Recordar que son los modelos probabilísticos los que nos permiten contrastar hipótesis.

Variable aleatoria multidimensional.

- Una v.a. multidimensional o vector aleatorio es un vector compuesto por p v.a.

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix}$$

- Cada v.a. X_i puede ser discreta, continua o de otros tipos. Esto da lugar a una gran diversidad de vectores aleatorios.
- Diferenciamos entre un vector aleatorio multidimensional y una muestra de un vector aleatorio multidimensional.
- El vector aleatorio corresponde a un modelo teórico mientras que la muestra de ese vector corresponde a una recolección de datos del mismo medidos sobre distintos individuos.

Estas mediciones pueden corresponder a dos grandes grupos:

- Proviene de un experimento diseñado y reproducible para estudiar su comportamiento. Este experimento debe tener en cuenta la representatividad de la muestra, su tamaño y los métodos estadísticos que se utilizarán para inferir las conclusiones deseadas.
- Proviene de datos recopilados (si se quiere de forma estadística pero en su sentido etimológico): bases de datos de proteínas, datos estadísticos de Institutos Oficiales (INE, Eurostat, IBAE). O bien de diferentes procedencias.

La diferencia es clara.

- En un estudio inferencial, del que se desee extraer conclusiones estadísticas digamos *profesionales o científicas*, debemos contar con muestras aleatorias que respalden los resultados del experimento.
- Nos estamos refiriendo a resultados publicados en revistas científicas, o de la industria farmacéutica , encuestas profesionales de cualquier índole (políticas/opinión, sociológicas, sanitarias, ecológicas, de estudio de mercados etc.).
- Además el experimento debe ser reproducible.

Vector de valores esperados. Matriz de Covarianza y de Correlaciones.

- Al igual que en el caso unidimensional, los vectores aleatorios tienen función de probabilidad o de densidad, función de distribución, medias, varianzas y otros momentos asociados...
- En el caso de los vectores aleatorios, estas cantidades se convierten en vectores y en matrices que también medirán, por ejemplo, sus valores esperados y la variación conjunta de las variables.

Vector de medias.

De la misma forma que en el caso unidimensional un vector aleatorio \mathbf{X} posee un valor esperado que en este caso es el vector de valores esperados o medias formado por los valores esperados de cada una de las componentes. Se representa como

$$E(\mathbf{X}) = \begin{pmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_p) \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix} = \boldsymbol{\mu}.$$

Covarianzas.

- Dadas dos variables aleatorias, de las que se conoce su distribución conjunta, se define su covarianza como

$$\text{Cov}(X_1, X_2) = E((X_1 - \mu_1)(X_2 - \mu_2))$$

- Que también se puede calcular con la siguiente identidad

$$\text{Cov}(X_1, X_2) = E(X_1 X_2) - \mu_1 \mu_2.$$

- La covarianza puede tomar cualquier valor real y mide el grado de dependencia lineal entre las variables (es decir si existen α y β reales tales que $X_2 = \alpha X_1 + \beta$).
- La covarianza de una variable consigo misma es su varianza

$$\text{Cov}(X_1, X_1) = \text{var}(X_1) = E((X_1 - \mu_1)^2) = E(X_1^2) - \mu_1^2.$$

Así como para simplificar la notación se suele llamar μ a valor esperado de una variable, se suele utilizar σ para las covarianzas. Así pondremos

$$\text{Cov}(X_i, X_j) = \sigma_{ij}; \quad \text{Cov}(X_i, X_i) = \sigma_{ii} = \text{var}(X_i) = \sigma_i^2.$$

Notemos que las unidades de la varianza son unidades cuadradas. La raíz cuadrada de la varianza es σ_i , recibe el nombre de desviación típica o estándar de X_i y recupera las unidades de la variable.

Variable tipificada.

Propiedad Sean a y b dos números reales y X_i una variable aleatoria.

- Entonces podemos construir $X_i + b$ que es otra variable aleatoria. Se suele decir que esta variable es un cambio de origen de la variable X_i pues desplaza todos los valores de X_i una cantidad b .
- También podemos construir la v.a. aX_i . Se suele decir que esta variable es un cambio de escala de la variable X_i pues agranda si $a > 1$ o encoge si $0 < a < 1$; mientras que para valores negativos sucede un efecto parecido pero además cambia su signo.

Propiedades

- $E(aX_i + b) = aE(X_i) + b$. La esperanza es un operador lineal.
- $var(aX_i + b) = a^2 var(X_i)$. La varianza no varía con cambios de origen y queda multiplicada por el cuadrado del factor de cambio de escala.

Variable tipificada

- La transformación de una variable del tipo $Z_i = \frac{X_i - \mu_i}{\sigma_i}$ recibe en nombre de variable típica, tipificada o estándar de X_i .
- Se suele llamar tipificar la variable al proceso de calcular su variable típica.
- Utilizando las propiedades de los valores esperados y las varianzas se obtiene que (ejercicio) $E(Z_i) = 0$ y $var(Z_i) = 1$. O sea, cuando tipificamos una variable obtenemos otra variable que siempre tiene media cero y varianza uno. Este proceso es útil a la hora de comparar distribuciones de variables y en el caso multidimensional se utiliza para disminuir el efecto del tamaño y de las unidades en que estén medidas las variables.

Matriz de covarianzas.

De la misma forma que en el caso unidimensional, un vector aleatorio \mathbf{X} posee una medida de su dispersión respecto al valor medio; es la llamada matriz de covarianzas.

$$\text{Cov}(\mathbf{X}) = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2p} \\ \vdots & \vdots & & \vdots \\ \sigma_{p1} & \sigma_{p2} & \dots & \sigma_{pp} \end{pmatrix} = E((\mathbf{X} - \mu) \cdot (\mathbf{X} - \mu)^\top) = \circ.$$

O lo que es lo mismo

$$\text{Cov}(\mathbf{X}) = E(\mathbf{X}\mathbf{X}^\top) - \mu\mu^\top.$$

La matriz de covarianzas se suele representar por \circ .

Matriz de Correlaciones.

Como las covarianzas son difíciles de comparar, se utiliza el llamado coeficiente de correlación lineal de Pearson que es una medida adimensional de la variación lineal entre dos variables.

Definimos la correlación de las variables X_i y X_j como

$$\text{Cor}(X_i, X_j) = \rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j}.$$

Nota: Es evidente que la correlación de una variable con sí misma es uno; $\rho_{ii} = 1$ (ejercicio).

Y podemos construir la matriz de correlaciones

$$\text{Cor}(\mathbf{X}) = \begin{pmatrix} 1 & \rho_{12} & \dots & \rho_{1p} \\ \rho_{21} & 1 & \dots & \rho_{2p} \\ \vdots & \vdots & & \vdots \\ \rho_{p1} & \rho_{p2} & \dots & 1 \end{pmatrix}.$$

Propiedades

- $-1 \leq \rho_{ij} \leq 1$
- $\rho_{ij} = \rho_{ji}$; $\rho_{ii} = 1$.
- Salvo en el signo, ρ_{ij} es invariante a cambios de origen y escala.
- Si $\rho_{ij} = \pm 1$, las variables tienen una relación lineal perfecta. Es decir, existen α y β tal que $X_i = \alpha X_j + \beta$. La pendiente α tiene el mismo signo que la correlación.
- Si la correlación es cero se dice que las variables son incorreladas (notemos que la correlación es cero sii la covarianza es cero).

Propiedad La matriz de correlaciones de un vector aleatorio es igual a la matriz de covarianzas del vector aleatorio cuyas componentes son sus variables tipificadas. Es decir

$$\text{Sea } \mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix} \text{ y sea } \mathbf{Z} = \begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_p \end{pmatrix}$$

donde

$$Z_i = \frac{X_i - \mu_i}{\sigma_i} \text{ para } i = 1, \dots, p.$$

Entonces $\text{Cor}(\mathbf{X}) = \text{Cov}(\mathbf{Z})$

Nota: Evidentemente (ejercicio) $\text{Cov}(\mathbf{Z}) = \text{Cor}(\mathbf{Z})$.

Expresión matricial de la tipificación de un vector aleatorio.

La operación consistente en dado un vector aleatorio obtener el vector formado por sus componentes tipificadas admite una forma matricial que recuerda a la tipificación de una variable.

Sea $\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix}$, sea μ su vector de medias y

$$\mathbf{A} = \begin{pmatrix} \sigma_1^{-1} & 0 \dots & 0 \\ 0 & \sigma_2^{-1} \dots & 0 \\ \vdots & \vdots & \vdots \\ 0 & 0 & \dots & \sigma_p^{-1} \end{pmatrix}$$

la matriz que tiene en su diagonal el inverso de las desviaciones típicas y el resto de valores iguales a cero.

Entonces

$$\mathbf{Z} = \mathbf{A} \cdot (\mathbf{X} - \mu).$$

Comprobarlo como ejercicio el efecto de esta transformación para $p = 3$.

- La expresión anterior es un caso particular de transformación lineal multivariante.
- Una transformación lineal general se obtiene tomando cualquier otra matriz \mathbf{A} y sustituyendo por un vector constante cualquiera el vector de medias. Incluso podemos variar las dimensiones de estas matrices.

- Como hemos visto, la matriz de correlaciones es una matriz de covarianzas de un cierto vector aleatorio.
- De este hecho se sigue que las propiedades que verifican las matrices de covarianzas las deben de cumplir las de correlaciones. En particular, la siguiente propiedad.

Propiedad Las matrices de covarianzas son semidefinidas positivas. De esta propiedad se desprende que son simétricas (ejercicio) y que tienen todos sus valores propios no negativos.

¿Por qué hemos tenido que aprender las nociones de la sección anterior?

- Pues el motivo es que tenemos que distinguir entre muestras y variables. Este concepto es tan importante que es el fundamento de la estadística.
- Cualquier estudio científico pasa por crear un modelo, una teoría, un paradigma, que no sea contradictorio en sí mismo.

- El estudio estadístico se realiza de la forma siguiente:
 - ▶ Uno plantea una teoría y la modela estadísticamente mediante por ejemplo vectores aleatorios.
 - ▶ Luego se diseña un experimento en condiciones adecuadas y para que la teoría no sea falsa, los datos no deben contradecir la teoría.
- La estadística siempre nos dirá cuan probables son los resultados obtenidos suponiendo que la teoría es cierta.
- Así que la estadística se usa para cuantificar la veracidad de una teoría. Demostrar que algo es cierto mediante muestras y métodos estadísticos es algo más complicado.

Datos Multivariantes.

Supondremos que hemos observado, recopilado, obtenido etc... p variables (si no se dice lo contrario numéricas) en un conjunto de n individuos u objetos. Es decir, tenemos n observaciones de p variables.

Claramente se pueden expresar estas observaciones de forma matricial.

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} = \begin{pmatrix} \mathbf{x}^1{}^\top \\ \mathbf{x}^2{}^\top \\ \vdots \\ \mathbf{x}^n{}^\top \end{pmatrix}$$

Donde utilizamos las siguientes notaciones

- Denotamos por

$$\mathbf{x}^{i\top} = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \dots \\ x_{ip} \end{pmatrix}^{\top} = (x_{i1}, x_{i2}, \dots, x_{ip})$$

Es decir, $\mathbf{x}^{i\top}$ son los vectores filas compuestos por las observaciones de las p variables sobre el i -ésimo individuo.

- Denotamos por

$$\mathbf{x}_j = \begin{pmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{pmatrix}$$

- Es decir \mathbf{x}_j son los vectores columna compuestos por las n observaciones de la j -ésima variable.
- Así podemos expresar la matriz de datos como $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_p)$
- Por último denotamos por $\mathbf{X} = (X_1, X_2, \dots, X_p)$ el vector aleatorio cuyas componentes son las variables aleatorias observadas.

Vector de medias, varianzas.

- Con estas notaciones podemos recuperar algunas definiciones ya conocidas de los estadísticos más usuales de una muestra.
- La media aritmética que es un estimador del valor esperado de cada variable.
- La varianza y la desviación típica muestral estiman los parámetros poblacionales del mismo nombre.

Media de una variable

$\bar{x}_j = \frac{1}{n} \sum_{i=1}^n x_{ij}$ es la media aritmética de la variable j -ésima en esta muestra.

Así el vector de medias aritméticas al que denotaremos por

$$\bar{\mathbf{x}} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{pmatrix}$$

es un estimador de $E(\mathbf{X}) = \mu$.

Desviación de una observación respecto a la media

$$d_{ij} = x_{ij} - \bar{x}_j$$

Varianza muestral

La varianza muestral es el estadístico que estima la varianza σ_j^2 de la variable X_j .

Tiene dos versiones

- La varianza muestral MLE, acrónimo del inglés *maximum likelihood estimator* o estimador máximo verosímil:

$$s_j^2 = \frac{1}{n} \sum_{i=1}^n d_{ij}^2 = \frac{1}{n} \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2 = \frac{1}{n} \sum_{i=1}^n x_{ij}^2 - \bar{x}_j^2.$$

- El estimador insesgado de la varianza. Los estimadores insesgados son aquellos cuyo valor esperado es el verdadero valor del parámetro.

$$\begin{aligned}\tilde{s}_j^2 &= \frac{1}{n-1} \sum_{i=1}^n d_{ij}^2 = \frac{1}{n-1} \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2 \\ &= \frac{1}{n-1} \sum_{i=1}^n x_{ij}^2 - \frac{n}{n-1} \bar{x}_j^2.\end{aligned}$$

- La relación que existen entre ambas formulas de la varianza muestral es la siguiente:

$$\frac{n}{n-1} s_j^2 = \tilde{s}_j^2$$

Nota: Algunos autores utilizan el nombre de cuasivarianza para denominar a \tilde{s}_j^2 . Para valores muestrales grandes las diferencias entre ambos estimadores de la varianza se hacen pequeñas. Los paquetes estadísticos suelen calcular por defecto la cuasivarianza.

Desviación típica muestral

Es la raíz cuadrada positiva de la varianza (de la que estemos utilizando)

Coeficiente de variación

Es una medida de la variación muestral estandarizada (es mejor utilizar la solamente para variables positivas)

$$cv_j = \frac{s_j}{\bar{x}_j}$$

Coeficiente de asimetría, coeficiente de apuntamiento Otros coeficientes que veremos en prácticas.

Centralización de una matriz de datos.

- Al igual que el caso univariante, centrar una variable es transformarla en otra de media 0. Para ello bastará restar a sus observaciones la media.
- En el caso de observaciones multivariantes, tendremos que realizar esta operación en todas las columnas de datos:

$$\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}^i = \frac{1}{n} \sum_{i=1}^n \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix} = \frac{1}{n} \begin{pmatrix} \sum_{i=1}^n x_{i1} \\ \sum_{i=1}^n x_{i2} \\ \vdots \\ \sum_{i=1}^n x_{ip} \end{pmatrix} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{pmatrix}$$

Cálculo matricial del vector de medias.

- Esta igualdad admite cálculo matricial . Sea $\mathbf{1}_n$ un vector columna de n filas todas iguales a 1

Ejemplo El efecto que produce la multiplicación de una matriz 3×4 por el vector $\mathbf{1}_4$ queda muy claro con el siguiente ejemplo

$$\begin{pmatrix} 2 & 2 & 2 & 2 \\ 3 & 3 & 3 & 3 \\ 1 & 2 & 3 & 4 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 8 \\ 12 \\ 10 \end{pmatrix}.$$

Entonces

$$\begin{aligned}\frac{1}{n} \mathbf{X}^\top \mathbf{1}_n &= \frac{1}{n} \begin{pmatrix} x_{11} & x_{21} & \dots & x_{n1} \\ \vdots & \vdots & & \vdots \\ x_{1p} & x_{2p} & \dots & x_{np} \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \\ &= \frac{1}{n} \begin{pmatrix} x_{11} + x_{21} + \dots + x_{n1} \\ \vdots \\ x_{1p} + x_{2p} + \dots + x_{np} \end{pmatrix} \\ &= \begin{pmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_p \end{pmatrix} = \bar{\mathbf{x}}\end{aligned}$$

Propiedad

$$\bar{\mathbf{x}} = \frac{1}{n} \mathbf{X}^\top \mathbf{1}_n$$

Matriz de datos centrados

Dada una matriz de datos \mathbf{X} llamaremos matriz de datos centrados y la denotaremos por $\tilde{\mathbf{X}}$ a la matriz de datos resultante de restar a cada columna de \mathbf{X} su media aritmética.

$$\tilde{\mathbf{X}} = \begin{pmatrix} x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 & \dots & x_{1p} - \bar{x}_p \\ x_{21} - \bar{x}_1 & x_{22} - \bar{x}_2 & \dots & x_{2p} - \bar{x}_p \\ \vdots & \vdots & & \vdots \\ x_{n1} - \bar{x}_1 & x_{n2} - \bar{x}_2 & \dots & x_{np} - \bar{x}_p \end{pmatrix}$$

El resultado es una matriz de datos donde todas las variables tienen media aritmética cero.

Veamos que la operación de centrado tiene una expresión matricial:

Notación: Llamamos matriz centralizadora de orden n a:

$$\mathbf{H}_n = \mathbf{I}_n - \frac{1}{n} \mathbf{1}_n \cdot \mathbf{1}_n^\top$$

Es decir $\mathbf{H}_n = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix} - \frac{1}{n} \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \cdot (1, 1, \dots, 1)$

$$= \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix} - \frac{1}{n} \begin{pmatrix} 1 & 1 & \dots & 1 \\ 1 & 1 & \dots & 1 \\ \vdots & \vdots & & \vdots \\ 1 & 1 & \dots & 1 \end{pmatrix}$$

$$= \begin{pmatrix} 1 - \frac{1}{n} & -\frac{1}{n} & \dots & -\frac{1}{n} \\ -\frac{1}{n} & 1 - \frac{1}{n} & \dots & -\frac{1}{n} \\ \vdots & \vdots & & \vdots \\ -\frac{1}{n} & -\frac{1}{n} & \dots & 1 - \frac{1}{n} \end{pmatrix}$$

Propiedad

- $\tilde{\mathbf{X}} = \mathbf{H}_n \cdot \mathbf{X}$, que nos da una expresión matricial del centrado.
- La matriz \mathbf{H}_n cumple que $\mathbf{H}_n \cdot \mathbf{H}_n = \mathbf{H}_n$. Es lo que se llama una matriz idempotente (comprobadlo como ejercicio).
- Además \mathbf{H}_n es simétrica, tiene rango $n - 1$ y $\mathbf{H}_n \cdot \mathbf{1}_n = 0$ (ejercicio).

Tipificación Tabla de datos

- Dado un conjunto de datos, llamaremos datos tipificados a los datos resultantes de restar a cada valor la media de la variable o columna que corresponde y dividir el resultado por su desviación típica.
- De esta forma obtenemos datos tipificados que tienen media aritmética 0 y varianza 1.
- La tipificación se puede realizar de forma tradicional, o bien matricialmente.

Propiedad Sea \mathbf{Z} la matriz de datos resultante de tipificar la matriz de

datos \mathbf{X} . Sea $\mathbf{D}^{-\frac{1}{2}} = \begin{pmatrix} \frac{1}{s_1} & 0 & \dots & 0 \\ 0 & \frac{1}{s_2} & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & \frac{1}{s_p} \end{pmatrix}$ la matriz que contiene en la

diagonal la inversa de las desviaciones típicas. Entonces:

$$\mathbf{Z} = \mathbf{H}_n \cdot \mathbf{X} \cdot \mathbf{D}^{-1/2} = \tilde{\mathbf{X}} \cdot \mathbf{D}^{-1/2}.$$

Ejemplo Sea $\mathbf{X} = \begin{pmatrix} 1 & -1 & 3 \\ 1 & 0 & 3 \\ 2 & 3 & 0 \\ 3 & 0 & 1 \end{pmatrix}$ una matriz de datos con $p = 3$ variables

y $n = 4$ observaciones.

Calculemos la matriz de datos tipificado \mathbf{Z} .

La matriz de la inversa de las desviaciones típicas es (ejercicio):

$$\mathbf{D}^{-1/2} = \begin{pmatrix} \frac{1}{\sqrt{\frac{11}{16}}} & 0 & 0 \\ 0 & \frac{1}{\sqrt{\frac{9}{4}}} & 0 \\ 0 & 0 & \frac{1}{\sqrt{\frac{27}{16}}} \end{pmatrix}.$$

Entonces

$$\mathbf{Z} = \mathbf{H}_4 \cdot \mathbf{X} \cdot \mathbf{D}^{-1/2} = \begin{pmatrix} \frac{3}{4} & -\frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{4} & \frac{3}{4} & -\frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{4} & -\frac{1}{4} & \frac{3}{4} & -\frac{1}{4} \\ -\frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} & \frac{3}{4} \end{pmatrix} \cdot \begin{pmatrix} 1 & -1 & 3 \\ 1 & 0 & 3 \\ 2 & 3 & 0 \\ 3 & 0 & 1 \end{pmatrix}.$$

$$\begin{pmatrix} \frac{1}{\sqrt{\frac{11}{16}}} & 0 & 0 \\ 0 & \frac{1}{\sqrt{\frac{9}{4}}} & 0 \\ 0 & 0 & \frac{1}{\sqrt{\frac{27}{16}}} \end{pmatrix} = \begin{pmatrix} -3/\sqrt{11} & -1 & 5/(3\sqrt{3}) \\ -3/\sqrt{11} & -1/3 & 5/(3\sqrt{3}) \\ 1/\sqrt{11} & 5/3 & -7/(3\sqrt{3}) \\ 5/\sqrt{11} & -1/3 & -1/\sqrt{3} \end{pmatrix}$$

Covarianza.

- Se define la covarianza muestral de las variables \mathbf{x}_i y \mathbf{x}_j como

$$s_{ij} = \frac{1}{n} \sum_{k=1}^n (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j) = \frac{1}{n} \sum_{k=1}^n x_{ki}x_{kj} - \bar{x}_i\bar{x}_j.$$

- La expresión anterior es un estimador máximo verosímil de σ_{ij} . También se tiene un estimador insesgado que consiste en dividir por $n - 1$ en lugar de n : $\tilde{s}_{ij} = \frac{n}{n-1} s_{ij}$.
- La covarianza muestral estima la relación lineal entre las variables.
- Puede tomar cualquier valor. Si es cero se dice que las muestras son incorreladas.
- $s_{ij} = s_{ji}$.
- $s_{ii} = s_i^2$.

Matriz de covarianzas.

Dada una tabla de datos llamaremos matriz de covarianzas (o de varianzas-covarianzas) a la matriz

$$\mathbf{S} = (s_{ij})_{i,j=1,\dots,p} = \begin{pmatrix} s_{11} & s_{12} & \dots & s_{1p} \\ s_{21} & s_{22} & \dots & s_{2p} \\ \vdots & \vdots & & \vdots \\ s_{p1} & s_{p2} & \dots & s_{pp} \end{pmatrix}$$

- La matriz de covarianzas muestral representa la variabilidad conjunta de los datos multidimensionales.
- La matriz de covarianzas es simétrica.
- $\text{tr}(\mathbf{S}) = \sum_{i=1}^p s_i^2 \geq 0$.
- Las matrices de covarianzas son definidas positivas.
- Tienen todos los valores propios no negativos y por lo tanto $\det(\mathbf{S}) \geq 0$.

Expresión matricial de S .

$$\mathbf{S} = \frac{1}{n} \tilde{\mathbf{X}}^\top \cdot \tilde{\mathbf{X}} = \frac{1}{n} \mathbf{X}^\top \cdot \mathbf{H}_n \cdot \mathbf{X}.$$

$$\mathbf{S} = \frac{1}{n} \begin{pmatrix} x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 & \dots & x_{1p} - \bar{x}_p \\ x_{21} - \bar{x}_1 & x_{22} - \bar{x}_2 & \dots & x_{2p} - \bar{x}_p \\ \vdots & \vdots & & \vdots \\ x_{n1} - \bar{x}_1 & x_{n2} - \bar{x}_2 & \dots & x_{np} - \bar{x}_p \end{pmatrix}^\top.$$
$$\begin{pmatrix} x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 & \dots & x_{1p} - \bar{x}_p \\ x_{21} - \bar{x}_1 & x_{22} - \bar{x}_2 & \dots & x_{2p} - \bar{x}_p \\ \vdots & \vdots & & \vdots \\ x_{n1} - \bar{x}_1 & x_{n2} - \bar{x}_2 & \dots & x_{np} - \bar{x}_p \end{pmatrix}$$

$$\mathbf{S} = \frac{1}{n} \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}^{\top}$$

$$\begin{pmatrix} 1 - \frac{1}{n} & -\frac{1}{n} & \dots & -\frac{1}{n} \\ -\frac{1}{n} & 1 - \frac{1}{n} & \dots & -\frac{1}{n} \\ \vdots & \vdots & & \vdots \\ -\frac{1}{n} & -\frac{1}{n} & \dots & 1 - \frac{1}{n} \end{pmatrix}.$$

$$\begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}.$$

Ejemplo. Sea $\mathbf{X} = \begin{pmatrix} 1 & -1 & 3 \\ 1 & 0 & 3 \\ 2 & 3 & 0 \\ 3 & 0 & 1 \end{pmatrix}$ una matriz de datos con $p = 3$ variables y $n = 4$ observaciones. De la forma tradicional se ponen los datos en una tabla de la siguiente forma:

i	x_1	x_2	x_3	x_1^2	x_2^2	x_3^2	x_1x_2	x_1x_3	x_2x_3
1	1	-1	3	1	1	9	-1	3	-3
2	1	0	3	1	0	9	0	3	0
3	2	3	0	4	9	0	6	0	0
4	3	0	1	9	0	1	0	3	0
Σ	7	2	7	15	10	19	5	9	-3

Así tenemos que

- $\bar{x}_1 = \frac{7}{4}, \bar{x}_2 = \frac{2}{4} \text{ y } \bar{x}_3 = \frac{7}{4}$
- $s_1^2 = \frac{1}{4} \sum_{i=1}^4 x_{i1}^2 - \bar{x}_1^2 = \frac{15}{4} - \left(\frac{7}{4}\right)^2 = \frac{11}{16}$
- $s_2^2 = \frac{1}{4} \sum_{i=1}^4 x_{i2}^2 - \bar{x}_2^2 = \frac{10}{4} - \left(\frac{2}{4}\right)^2 = \frac{9}{4}$
- $s_3^2 = \frac{1}{4} \sum_{i=1}^4 x_{i3}^2 - \bar{x}_3^2 = \frac{19}{4} - \left(\frac{7}{4}\right)^2 = \frac{27}{16}$
- $s_{12} = \frac{1}{4} \sum_{i=1}^n x_{i1}x_{i2} - \bar{x}_1\bar{x}_2 = \frac{5}{4} - \frac{7}{4} \frac{2}{4} = \frac{3}{8}$
- $s_{13} = \frac{1}{4} \sum_{i=1}^n x_{i1}x_{i3} - \bar{x}_1\bar{x}_3 = \frac{9}{4} - \frac{7}{4} \frac{7}{4} = -\frac{13}{16}$
- $s_{23} = \frac{1}{4} \sum_{i=1}^n x_{i2}x_{i3} - \bar{x}_2\bar{x}_3 = \frac{-3}{4} - \frac{2}{4} \frac{7}{4} = -\frac{13}{8}$

Luego la matriz de covarianzas es

$$\mathbf{S} = \begin{pmatrix} 11/16 & 3/8 & -13/16 \\ 3/8 & 9/4 & -13/8 \\ -13/16 & -13/8 & 27/16 \end{pmatrix}$$

Hagamos los cálculos de forma matricial

En este caso la matriz $\mathbf{H}_4 = \begin{pmatrix} \frac{3}{4} & -\frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{4} & \frac{3}{4} & -\frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{4} & -\frac{1}{4} & \frac{3}{4} & -\frac{1}{4} \\ -\frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} & \frac{3}{4} \end{pmatrix}$

Entonces

$$\mathbf{S} = \frac{1}{4} \mathbf{X}^\top \cdot \mathbf{H}_4 \cdot \mathbf{X} = \frac{1}{4}$$

$$\begin{pmatrix} 1 & 1 & 2 & 3 \\ -1 & 0 & 3 & 0 \\ 3 & 3 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} \frac{3}{4} & -\frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{4} & \frac{3}{4} & -\frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{4} & -\frac{1}{4} & \frac{3}{4} & -\frac{1}{4} \\ -\frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} & \frac{3}{4} \end{pmatrix} \cdot \begin{pmatrix} 1 & -1 & 3 \\ 1 & 0 & 3 \\ 2 & 3 & 0 \\ 3 & 0 & 1 \end{pmatrix} =$$

$$\frac{1}{4} \begin{pmatrix} 1 & 1 & 2 & 3 \\ -1 & 0 & 3 & 0 \\ 3 & 3 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} -\frac{3}{4} & -\frac{3}{2} & \frac{5}{4} \\ -\frac{3}{4} & -\frac{1}{2} & \frac{5}{4} \\ \frac{1}{4} & \frac{5}{2} & -\frac{7}{4} \\ \frac{5}{4} & -\frac{1}{2} & -\frac{3}{4} \end{pmatrix} =$$

$$= \frac{1}{4} \begin{pmatrix} \frac{11}{4} & \frac{3}{2} & -\frac{13}{4} \\ \frac{3}{2} & 9 & -\frac{13}{2} \\ -\frac{13}{4} & -\frac{13}{2} & \frac{27}{4} \end{pmatrix} = \begin{pmatrix} \frac{11}{16} & \frac{3}{8} & -\frac{13}{16} \\ \frac{3}{8} & \frac{9}{4} & -\frac{13}{8} \\ -\frac{13}{16} & -\frac{13}{8} & \frac{27}{16} \end{pmatrix} = \mathbf{S}$$

También podemos calcular de forma matricial de $\bar{\mathbf{x}}$:

$$\bar{\mathbf{x}} = \frac{1}{4} \mathbf{X}^T \mathbf{1}_4 = \frac{1}{4} \begin{pmatrix} 1 & 1 & 2 & 3 \\ -1 & 0 & 3 & 0 \\ 3 & 3 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} = \frac{1}{4} \begin{pmatrix} 7 \\ 2 \\ 7 \end{pmatrix} = \begin{pmatrix} \frac{7}{4} \\ \frac{2}{4} \\ \frac{7}{4} \end{pmatrix}$$

Ejercicio Se deja como ejercicio el cálculo de la matriz centrada $\tilde{\mathbf{X}}$.

Variables redundantes

Decimos que en una tabla de datos hay variables redundantes cuando una o más variables aportan la misma información que otra.

La redundancia de variables se puede manifestar por ejemplo si una variable \mathbf{x}_i cumple que es combinación lineal de otras variables $\mathbf{x}_{i_1}, \dots, \mathbf{x}_{i_k}$:

$$x_i = a_1 \mathbf{x}_{i_1} + \dots + a_k \mathbf{x}_{i_k} + b.$$

La matriz de covarianzas es muy útil para descubrir las variables redundantes de tipo lineal. Lo veremos en la siguiente propiedad.

Propiedad

Sea \mathbf{S} un matriz de covarianzas de dimensión p .

- El número de variables redundantes es igual al número de valores propios de \mathbf{S} iguales a cero.
- Si $\det(\mathbf{S}) = 0$, entonces existe al menos una variable redundante.
- Si $rg(\mathbf{S}) = k$, entonces existen $p - k$ variables redundantes.

Ejemplo

i	x_1	x_2	x_3
1	1	0	-1
2	1	2	1
3	1	1	0
4	0	3	0

Entonces $\mathbf{X} = \begin{pmatrix} 1 & 0 & -1 \\ 1 & 2 & 1 \\ 1 & 1 & 0 \\ 0 & 3 & 0 \end{pmatrix}$

Calculemos $\bar{\mathbf{x}}$ en primer lugar.

$$\bar{\mathbf{x}} = \frac{1}{4} \cdot \mathbf{X}^T \cdot \mathbf{1}_4 = \frac{1}{4} \begin{pmatrix} 1 & 1 & 1 & 0 \\ 0 & 2 & 1 & 3 \\ -1 & 1 & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} \frac{3}{4} \\ \frac{3}{2} \\ 0 \end{pmatrix}$$

Ahora podemos restar manualmente las medias a \mathbf{X} para obtener $\tilde{\mathbf{X}}$.
(ejercicio)

O podemos calcular matricialmente:

$$\tilde{\mathbf{X}} = \mathbf{H}_4 \cdot \mathbf{X} = \begin{pmatrix} \frac{3}{4} & -\frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{4} & \frac{3}{4} & -\frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{4} & -\frac{1}{4} & \frac{3}{4} & -\frac{1}{4} \\ -\frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} & \frac{3}{4} \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 & -1 \\ 1 & 2 & 1 \\ 1 & 1 & 0 \\ 0 & 3 & 0 \end{pmatrix} =$$

$$\begin{pmatrix} \frac{1}{4} & -\frac{3}{2} & -1 \\ \frac{1}{4} & \frac{1}{2} & 1 \\ \frac{1}{4} & -\frac{1}{2} & 0 \\ -\frac{3}{4} & \frac{3}{2} & 0 \end{pmatrix}$$

$$\text{Ahora } \mathbf{S} = \frac{1}{4} \tilde{\mathbf{X}}^\top \cdot \tilde{\mathbf{X}} = \frac{1}{4} \begin{pmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & -\frac{3}{4} \\ -\frac{3}{2} & \frac{1}{2} & -\frac{1}{2} & \frac{3}{2} \\ -1 & 1 & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} \frac{1}{4} & -\frac{3}{2} & -1 \\ \frac{1}{4} & \frac{1}{2} & 1 \\ \frac{1}{4} & -\frac{1}{2} & 0 \\ -\frac{3}{4} & \frac{3}{2} & 0 \end{pmatrix} =$$

$$\begin{pmatrix} \frac{3}{16} & -\frac{3}{8} & 0 \\ -\frac{3}{8} & \frac{5}{4} & \frac{1}{2} \\ 0 & \frac{1}{2} & \frac{1}{2} \end{pmatrix}$$

Si calculamos $\det(\mathbf{S}) = 0$ (ejercicio). Luego existe al menos una variable redundante.

Veamos si hay más. Calculemos los valores propios.

El polinomio característico de **S** es

$$p_S = \begin{vmatrix} \frac{3}{16} - \lambda & -\frac{3}{8} & 0 \\ -\frac{3}{8} & \frac{5}{4} - \lambda & \frac{1}{2} \\ 0 & \frac{1}{2} & \frac{1}{2} - \lambda \end{vmatrix} = -\lambda^3 + \frac{31\lambda^2}{16} - \frac{9\lambda}{16}$$

Que sólo tiene una solución nula. Por lo tanto sólo hay una variable redundante.

¿A alguien se le ocurre una relación lineal entre las variables?

- La dificultad de la interpretación de la matriz de covarianzas como medida de variabilidad radica en que son muchas cantidades.
- Desafortunadamente no hay una sola cantidad que mida la variabilidad multivariante de forma sobresaliente. Veamos dos
- **Varianza total**
 $T = \text{tr}(\mathbf{S}) = \sum_{i=1}^p s_i^2 = \sum_{i=1}^n \lambda_i$. La varianza media será $\frac{T}{p}$.
- **Varianza Generalizada**

$$\det(\mathbf{S}) = \lambda_1 \cdot \dots \cdot \lambda_p.$$

La desviación típica generalizada será $\sqrt{\det(\mathbf{S})}$ que cuando el conjunto de datos se representa en \mathbb{R}^p es el área, volumen o hipervolumen del conjunto de datos.

Correlación lineal de Pearson.

Se define la correlación lineal de Pearson de las muestras de las variables \mathbf{x}_i y \mathbf{x}_j como

$$r_{ij} = \frac{s_{ij}}{s_i s_j}$$

La correlación r_{ij} estima el parámetro poblacional $\rho_{ij} = \text{Cor}(X_i, X_j)$.

Propiedades

- $-1 \leq r_{ij} \leq 1$.
- $r_{ii} = 1$.
- La correlación tiene el mismo signo que la covarianza.
- $r_{ij} = \pm 1$ si y sólo si existe una relación lineal perfecta entre las variables \mathbf{x}_i i \mathbf{x}_j . O sea, existen valores a y b tal que $\mathbf{x}_j = a\mathbf{x}_i + b$. La pendiente de la recta a tiene el mismo signo que la correlación entre las variables.

Matriz de correlaciones

Llamaremos matriz de correlaciones de la tabla de datos \mathbf{X} a

$$\mathbf{R} = (r_{ij})_{i,j=1,\dots,p} = \begin{pmatrix} 1 & r_{12} & \dots & r_{1p} \\ r_{21} & 1 & \dots & r_{2p} \\ \vdots & \vdots & & \vdots \\ r_{p1} & r_{p2} & \dots & 1 \end{pmatrix}$$

Propiedades

- La matriz \mathbf{R} es semidefinida positiva.
- Si todas las variables son incorreladas entonces $\mathbf{R} = I_p$ y $\det(\mathbf{R}) = 1$.
- Respecto a las variables redundantes, \mathbf{R} cumple las mismas propiedades que la matriz de covarianzas. Por ejemplo si $\det(\mathbf{R}) = 0$, hay al menos una variable redundante.
- $\det(\mathbf{R}) \leq 1$.

Expresión matricial de la matriz de correlaciones.

$$\text{Sea } \mathbf{D}^{\frac{1}{2}} = \begin{pmatrix} s_1 & 0 & \dots & 0 \\ 0 & s_2 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & s_p \end{pmatrix}$$

$$\text{Entonces su inversa es } \mathbf{D}^{-\frac{1}{2}} = \begin{pmatrix} \frac{1}{s_1} & 0 & \dots & 0 \\ 0 & \frac{1}{s_2} & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & \frac{1}{s_p} \end{pmatrix}$$

Propiedades

- Expresión matricial de la matriz de correlaciones $\mathbf{R} = \mathbf{D}^{-\frac{1}{2}} \cdot \mathbf{S} \cdot \mathbf{D}^{-\frac{1}{2}}$.
- Usando la expresión anterior, podemos escribir la matriz de covarianzas como $\mathbf{S} = \mathbf{D}^{\frac{1}{2}} \cdot \mathbf{R} \cdot \mathbf{D}^{\frac{1}{2}}$
- La matriz de covarianzas de los datos tipificados es la matriz de correlaciones de \mathbf{X} . O sea, si \mathbf{Z} es la matriz de datos tipificados de \mathbf{X} entonces

$$\mathbf{S}_Z = \mathbf{R}_X$$

- Esta última propiedad confirma que la matriz de correlaciones es también de covarianzas y cumple sus propiedades.

Ejercicio

Consideremos la siguiente matriz de datos $\mathbf{X} = \begin{pmatrix} 1 & 0 & -1 \\ 1 & 2 & 1 \\ 1 & 1 & 0 \\ 0 & 3 & 0 \end{pmatrix}$

- a) Calcular la matriz de correlaciones.
- b) Obtener la tabla de datos tipificados \mathbf{Z} .
- c) Calcular la matriz de covarianzas de los datos tipificados y comprobar que es igual a la matriz de correlaciones de \mathbf{X} .
- d) Sin hacer cálculos, determinar la matriz de correlaciones de \mathbf{Z} .

Otros tipos de correlaciones.

- **Correlaciones parciales.** Hasta ahora hemos visto la relación lineal entre cada par de variables. Pero en un estudio conjunto nos podría interesar la relación lineal de dos variables eliminando el efecto de las demás, este concepto recibe el nombre de correlación parcial.
- **Correlaciones ordinales.** Otros tipos de correlaciones son las llamadas correlaciones ordinales. Lo que se busca es si existe correlación entre dos tipos de ordenaciones. Las más conocidas con la correlación ordinal de Spearman y la de Kendall.

Parte X

Regresión lineal simple

Introducción

Introducción

Consideremos los pares de observaciones de dos variables:

$$\{(x_i, y_i) | i = 1, 2, \dots, n\}$$

- La variable y es la variable dependiente o de respuesta.
- La variable x es la variable de control o independiente o de regresión.
- El problema que se intenta resolver es encontrar la mejor relación funcional que explique la variable y conocido el valor de la variable x : Y/x . En nuestro caso esta función será una recta.

Ejemplo

Ejemplo: Consideremos los datos siguientes donde x representa los meses e y representa el crecimiento de un determinado tipo de planta en mm.

Meses	Crecimiento
12	0
10	1
8	2
11	3
6	4
7	5
2	6
3	7
3	8

GrapRIS.

El modelo de Regresión lineal simple

En realidad, en un análisis más riguroso, el modelo de regresión lineal es el siguiente:

$$\mu_{Y/x} = \beta_0 + \beta_1 x,$$

donde $\mu_{Y/x}$ es el valor esperado que toma la variable y cuando la variable de control vale x , mientras que β_0 (término independiente) y β_1 (pendiente) son dos parámetros a determinar.

Dada una muestra calcularemos las estimaciones b_0 y b_1 de β_0 y de β_1 respectivamente. Notemos que para muestras diferentes, las estimaciones serán diferentes.

Una vez obtenidas las estimaciones podemos calcular la recta de regresión estimada, que es:

$$\hat{y} = b_0 + b_1 x.$$

Regresión lineal simple por Mínimos cuadrados

- Existen diversas maneras de calcular las estimaciones de los coeficientes de una regresión lineal: Regresión ortogonal, métodos robustos, regresión mínimo cuadrática o de mínimos cuadrados,... Nosotros optaremos por el método más habitual que es el de mínimos cuadrados (m.c.).

- Modelo

$$Y_i = \beta_0 + \beta_1 x_i + E_i,$$

Donde E_i es una nueva variable llamada error o residuo.

- Una vez planteado el modelo y dada una muestra el modelo se debe ajustar a los datos de ésta:

$$y_i = \beta_0 + \beta_1 x_i + E_i, \text{ para } i = 1, 2, \dots, n.$$

Regresión lineal simple por Mínimos cuadrados

- Cuando ajustamos por las estimaciones b_0 y b_1 obtenemos la recta de regresión ajustada

$$\hat{y} = b_0 + b_1x.$$

- Podemos calcular para cada par de observaciones:

$$y_i = b_0 + b_1x_i + e_i, \quad \hat{y}_i = b_0 + b_1x_i \text{ para } i = 1, 2, \dots, n.$$

- Entonces el error o residuo de la i -ésima observación, $i = 1, 2, \dots, n$, es

$$e_i = y_i - \hat{y}_i.$$

Regresión lineal simple por Mínimos cuadrados

Cálculo de los coeficientes

Cálculo de b_0 y b_1 por M.C.

- Los valores de b_0 y b_1 buscados son los que minimizan el error cuadrático:

$$SSE = \sum_{i=1}^n e_i^2.$$

- Estos valores serán los estimadores de β_0 y β_1 por el método de mínimos cuadrados.

Regresión lineal simple por Mínimos cuadrados

Cálculo de los coeficientes

- En primer lugar tenemos que:

$$SSE = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - b_0 - b_1 x_i)^2.$$

- Calculando las derivadas parciales respecto a b_0 y a b_1 , e igualando a cero:

$$\frac{\partial SSE}{\partial b_0} = -2 \sum_{i=1}^n (y_i - b_0 - b_1 x_i) = 0,$$

$$\frac{\partial SSE}{\partial b_1} = -2 \sum_{i=1}^n (y_i - b_0 - b_1 x_i) x_i = 0.$$

Regresión lineal simple por Mínimos cuadrados

Cálculo de los coeficientes. Ecuaciones normales

- Las ecuaciones anteriores reciben el nombre de ecuaciones normales:

$$\left. \begin{aligned} nb_0 + \sum_{i=1}^n x_i b_1 &= \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i b_0 + \sum_{i=1}^n x_i^2 b_1 &= \sum_{i=1}^n x_i y_i \end{aligned} \right\}$$

- Las soluciones de estas ecuaciones son:

$$b_1 = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2}, \quad b_0 = \frac{\sum_{i=1}^n y_i - b_1 \sum_{i=1}^n x_i}{n}.$$

Ejemplo

En el ejemplo anterior, tenemos:

$$b_1 = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2} = \frac{9 \cdot 175 - 62 \cdot 36}{9 \cdot 536 - 62^2} = -0.6704,$$

$$b_0 = \frac{\sum_{i=1}^n y_i - b_1 \sum_{i=1}^n x_i}{n} = \frac{36 - (-0.6704) \cdot 62}{9} = 8.6184.$$

Regresión lineal simple por Mínimos cuadrados

Cálculo de los coeficientes. Definición de los momentos de primer y segundo orden.

- Definimos las medias y varianzas de las variables x e y como:

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}, \quad \bar{y} = \frac{\sum_{i=1}^n y_i}{n},$$

$$s_x^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n} = \frac{\sum_{i=1}^n x_i^2}{n} - \bar{x}^2,$$

$$s_y^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n} = \frac{\sum_{i=1}^n y_i^2}{n} - \bar{y}^2,$$

Regresión lineal simple por Mínimos cuadrados

Cálculo de los coeficientes. Definición de los momentos de primer y segundo orden.

- Definimos la covarianza entre las variables x e y como:

$$s_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n} = \frac{\sum_{i=1}^n x_i y_i}{n} - \bar{x}\bar{y}.$$

Ejemplo anterior

- Los momentos de primer orden son para el ejemplo anterior:

$$\bar{x} = \frac{62}{9} = 6.89, \quad \bar{y} = \frac{36}{9} = 4.$$

- Los momentos de segundo orden son:

$$s_x^2 = \frac{536}{9} - 6.89^2 = 12.09877, \quad s_y^2 = \frac{204}{9} - 4^2 = 6.67,$$

$$s_{xy} = \frac{175}{9} - 6.89 \cdot 4 = -8.111111.$$

Regresión lineal simple por Mínimos cuadrados

Cálculo de los coeficientes en función de los momentos de primer y segundo orden

- Los coeficientes de la recta de regresión son en función de las medias, varianzas y covarianza entre las variables x e y :

$$b_1 = \frac{s_{xy}}{s_x^2}, \quad b_0 = \bar{y} - b_1 \bar{x}.$$

- Ejemplo anterior:

$$b_1 = \frac{-8.11}{12.09877} = -0.6704,$$

$$b_0 = 4 - (-0.6704) \cdot 6.89 = 8.6184.$$

Propiedades de los estimadores

- La recta de regresión pasa por el vector de medias (\bar{x}, \bar{y}) , es decir:

$$b_0 + b_1\bar{x} = \bar{y}$$

- La media de los valores estimados es igual a la media de los observados

$$\bar{\hat{y}} = \frac{\sum_{i=1}^n \hat{y}_i}{n} = \bar{y}$$

Ejemplo anterior

- Comprobemos que la recta de regresión pasa por el vector de medias que será (6.89, 4):

$$\hat{y} = b_0 + b_1x, \text{ si } x = 6.89, \text{ queda} \\ 8.6184 - 0.6704 \cdot 6.89 \approx 4.$$

- Veamos que la media de los valores estimados es la misma que los valores observados; o sea, 4. Los valores estimados son:

0.57347, 1.91429, 3.25510, 1.24388, 4.59591,
3.92551, 7.27755, 6.60714, 6.60714

Si hallamos la media, vale efectivamente 4.

Consideraciones sobre el modelo de regresión lineal

- Se supone que los errores del modelo E_i tienen una distribución normal de media 0 y desviación típica σ .
- Los errores de la estimación por mínimos cuadrados tienen media 0. Efectivamente,

$$\sum_{i=1}^n e_i = \sum_{i=1}^n (y_i - \hat{y}_i) = \sum_{i=1}^n (y_i - b_0 - b_1 x_i) = 0.$$

En conclusión

$$\bar{e} = \frac{\sum_{i=1}^n e_i}{n} = 0,$$

y

$$s_e^2 = \frac{\sum_{i=1}^n e_i^2}{n} = \frac{SSE}{n}.$$

Ejemplo anterior

- Para comprobar la normalidad se realiza un QQ test. Los errores son los siguientes:

−0.57347, −0.91429, −1.25510, 1.75612, −0.59591,
1.07449, −1.27755, 0.39286, 1.3928571,

El qq-test de los errores aparece en el gráfico siguiente:



Definición de las sumas de cuadrados

- Llamaremos suma de cuadrados de los residuales o del error a

$$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2.$$

- Llamaremos suma de cuadrados de totales a

$$SST = \sum_{i=1}^n (y_i - \bar{y})^2.$$

- Llamaremos suma de cuadrados de la regresión a

$$SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2.$$

Relación entre las sumas de cuadrados

- En una regresión lineal por el método de mínimos cuadrados se tiene que:

$$SST = SSR + SSE.$$

- La expresión anterior es equivalente a

$$S_y^2 = S_{\hat{y}}^2 + S_e^2.$$

Ejemplo anterior

- Las sumas de cuadrados son:

$$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = 11.06020,$$

$$SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = 48.9398,$$

$$SST = \sum_{i=1}^n (y_i - \bar{y})^2 = 60.$$

Se puede comprobar que se cumple la igualdad $SST = SSE + SSR$.

El coeficiente de determinación R^2 y la estimación de la varianza.

- Se define como

$$R^2 = \frac{SSR}{SST}.$$

- En el caso de regresión lineal m.c. se cumple que: $R^2 = \frac{S_y^2}{S_y^2}$,
 $R^2 = 1 - \frac{SSE}{SST}$, $R^2 = 1 - \frac{S_e^2}{S_y^2}$.
- $R^2 = r_{xy}^2$, donde $r_{xy} = \frac{s_{xy}}{s_x s_y}$.
- Por lo tanto R^2 es la proporción de varianza de la variable y que queda explicada por la regresión lineal.
- Una estimación insesgada de σ^2 (la varianza del error E) en m.c. es

$$S^2 = \frac{SSE}{n - 2}.$$

Ejemplo anterior

- El coeficiente R^2 valdrá: $R^2 = \frac{48.9398}{60} = 0.8157$. Por tanto, se explica el 81.57 % de la varianza del crecimiento de la planta.
- Estimación insesgada de la varianza:

$$S^2 = \frac{SSE}{n - 2} = \frac{11.06020}{7} = 1.5800.$$

Intervalos de confianza

- Suponemos de que los residuos siguen una ley normal.
- Intervalo de confianza al nivel $(1 - \alpha)100\%$ para el parámetro β_1 :
 $(\mu_{Y/x} = \beta_0 + \beta_1 x)$

$$b_1 - \frac{t_{n-2, \alpha/2} S}{\sqrt{n S_x^2}} < \beta_1 < b_1 + \frac{t_{n-2, \alpha/2} S}{\sqrt{n S_x^2}}$$

- Intervalo de confianza al nivel $(1 - \alpha)100\%$ para el parámetro β_0 :

$$b_0 - \frac{t_{n-2, \alpha/2} S \sqrt{\sum_{i=1}^n x_i^2}}{n S_x} < \beta_0 < b_0 + \frac{t_{n-2, \alpha/2} S \sqrt{\sum_{i=1}^n x_i^2}}{n S_x}$$

Intervalos de confianza

- Intervalo de confianza al nivel $(1 - \alpha)100\%$ para la respuesta media μ_{Y/x_0} :

$$\hat{y}_0 - t_{n-2, \alpha/2} S \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{n S_x^2}} < \mu_{Y/x_0} < \hat{y}_0 + t_{n-2, \alpha/2} S \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{n S_x^2}}$$

- Intervalo de confianza al nivel $(1 - \alpha)100\%$ para el valor de y_0 cuando $x = x_0$:

$$\hat{y}_0 - t_{n-2, \alpha/2} S \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{n S_x^2}} < y_0 < \hat{y}_0 + t_{n-2, \alpha/2} S \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{n S_x^2}}$$

Ejemplo anterior

- Intervalo de confianza al nivel 95 % para la respuesta media μ_{Y/x_0} :

$$\hat{y}_0 - t_{n-2, \alpha/2} S \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{n s_x^2}} < \mu_{Y/x_0} <$$

$$\hat{y}_0 + t_{n-2, \alpha/2} S \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{n s_x^2}}$$

$$\hat{y}_0 - t_{7, 0.025} \sqrt{\frac{1.58}{9} + \frac{1.58 \cdot (x_0 - 6.89)^2}{9 \cdot 12.09877}} < \mu_{Y/x_0} <$$

$$\hat{y}_0 + t_{7, 0.025} \sqrt{\frac{1.58}{9} + \frac{1.58 \cdot (x_0 - 6.89)^2}{9 \cdot 12.09877}}$$

$$\hat{y}_0 - 2.36 \cdot \sqrt{0.1756 + \frac{(x_0 - 6.89)^2}{68.917}} < \mu_{Y/x_0} <$$

$$\hat{y}_0 + 2.36 \cdot \sqrt{0.1756 + \frac{(x_0 - 6.89)^2}{68.917}}$$

Si cogemos $x_0 = 11$ meses, el intervalo anterior vale:

$$-0.287 < \mu_{Y/x_0} < 2.775.$$

ANOVA en la recta de regresión lineal

Muy brevemente el Análisis de la Varianza (ANAlisys Of VAriance) consiste en contrastar si la media de una variable en k poblaciones independientes, con distribución normal de igual varianza, son iguales contra que al menos dos son distintas.

$$\begin{cases} H_0 : \mu_1 = \mu_2 = \dots = \mu_k \\ H_1 : \text{no todas las medias son iguales} \end{cases}$$

En el caso de la regresión lineal, usaremos la técnica anterior para contrastar si las medias de los grupos que conforman las variables son iguales o no (el grupo k está formado por los valores cuya media vale μ_{Y/x_k}). En caso afirmativo, decir que las medias son iguales es equivalente a afirmar que $\beta_1 = 0$ y, por lo tanto, el modelo de regresión lineal no es bueno. Por tanto, para que el modelo sea bueno, hemos de rechazar la hipótesis nula en el contraste ANOVA

ANOVA en la recta de regresión lineal

- Test a realizar:

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}$$

- Tabla a calcular:

Fuente de variación	Suma de cuadrados	g. l.	Cuadrados medios	F
Regresión	SSR	1	SSR	SSR/S^2
Error	SSE	$n - 2$	$S^2 = \frac{SSE}{n-2}$	
Total	SST	$n - 1$		

ANOVA en la recta de regresión lineal

- Ahora rechazamos la hipótesis nula al nivel de significación α si $f > f_{\alpha,1,n-2}$ donde $f_{\alpha,1,n-2}$ es el valor de una distribución F de con grados de libertad 1 y $n - 2$.
- Esta prueba, en el caso de regresión lineal simple tiene un efecto a otra parecida en la que se contrasta con una t de student.

Ejemplo anterior

- Tabla ANOVA:

Fuente de variación	Suma de cuadrados	g. l.	Cuadrados medios	F
Regresión	48.9398	1	48.9398	30.974
Error	11.0602	7	$S^2 = 1.5800$	
Total	60	8		

- Cogemos $\alpha = 0.05$. El valor $f_{0.05,1,7}$ vale 5.59. Como $f = 30.974 > 5.59$, rechazamos la hipótesis nula y concluimos que $\beta_1 \neq 0$. Por tanto, nuestro modelo es adecuado según este análisis.

Parte XI

Regresión lineal múltiple

Introducción

- Tenemos k variables independientes x_1, \dots, x_k y una variable dependiente y .
- Postulamos el modelo de regresión lineal como:

$$\mu_{Y, x_1, \dots, x_k} = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k.$$

Los parámetros β_i son desconocidos y se pueden estimar a partir de una muestra:

$$\{(x_{i1}, x_{i2}, \dots, x_{ik}, y_i) | i = 1, 2, \dots, n\}$$

de la que se exige que $n > k$, es decir el número de observaciones sea mayor que el número de variables.

Introducción

- El modelo es el siguiente: Consideramos un conjunto de k variables aleatorias X_1, X_2, \dots, X_k . Suponemos que existen variables aleatorias respuestas Y_1, \dots, Y_k cuya relación con las anteriores es:

$$Y_i = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k + E_i,$$

donde E_i son variables aleatorias que representan el error aleatorio del modelo asociado a la respuesta Y_i .

- El problema es estimar los parámetros β_i a partir de una muestra de datos que representan una muestra aleatoria simple de las variables X_i y de la variable Y de tamaño n :

$$\{(x_{i1}, x_{i2}, \dots, x_{ik}, y_i) | i = 1, 2, \dots, n\}.$$

Introducción

- Llamaremos y_i al valor obtenido de la variable Y_i usando las estimaciones b_i de los parámetros β_i :

$$y_i = b_0 + b_1x_{i1} + \cdots + b_kx_{ik} + e_i \text{ para } i = 1, 2, \dots, n$$

donde e_i será la estimación de la variable error residual E_i asociado a la respuesta Y_i .

- Llamaremos

$$\hat{y}_i = b_0 + b_1x_{i1} + \cdots + b_kx_{ik}.$$

Entonces $e_i = y_i - \hat{y}_i$.

Introducción

- Definimos los vectores siguientes:

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \mathbf{b} = \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{pmatrix}, \hat{\mathbf{y}} = \begin{pmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{pmatrix}, \mathbf{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}.$$

- Definimos la matriz siguiente:

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix}$$

Introducción

- Podemos escribir el modelo de regresión múltiple matricialmente como:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e},$$

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{b},$$

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}.$$

Cálculo de los coeficientes b_i usando el método de mínimos cuadrados

- Definimos el error cuadrático SSE como:

$$\begin{aligned} SSE &= \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^n (y_i - b_0 - b_1 x_{i1} - \cdots - b_k x_{ik})^2. \end{aligned}$$

- Los estimadores por el método de mínimos cuadrados serán los valores b_0, b_1, \dots, b_k que minimicen SSE .
- Para resolver este problema calculamos las derivadas parciales de SSE respecto a cada b_i para $i = 1, 2, \dots, n$ y se obtiene el un sistema de ecuaciones que recibe el nombre de ecuaciones normales.

Cálculo de los coeficientes b_i usando el método de mínimos cuadrados

$$\left. \begin{aligned}
 nb_0 + b_1 \sum_{i=1}^n x_{i1} + \dots + b_k \sum_{i=1}^n x_{ik} &= \sum_{i=1}^n y_i \\
 b_0 \sum_{i=1}^n x_{i1} + b_1 \sum_{i=1}^n x_{i1}^2 + b_2 \sum_{i=1}^n x_{i1}x_{i2} + \dots + \\
 &\quad b_k \sum_{i=1}^n x_{i1}x_{ik} = \sum_{i=1}^n x_{i1}y_i \\
 &\quad \dots \quad \dots \\
 b_0 \sum_{i=1}^n x_{ik} + b_1 \sum_{i=1}^n x_{ik}x_{i1} + b_2 \sum_{i=1}^n x_{ik}x_{i2} + \dots + \\
 &\quad b_k \sum_{i=1}^n x_{ik}^2 = \sum_{i=1}^n x_{ik}y_i
 \end{aligned} \right\}$$

Cálculo de los coeficientes b_i usando el método de mínimos cuadrados

- El sistema anterior se puede expresar en forma matricial de la forma siguiente:

$$\left(\mathbf{X}^T \mathbf{X}\right) \cdot \mathbf{b} = \mathbf{X}^T \cdot \mathbf{y}.$$

- La solución buscada del sistema anterior será:

$$\mathbf{b} = \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \cdot \left(\mathbf{X}^T \mathbf{y}\right).$$

Ejemplo

- Se postula que la estatura de un niño recién nacido (y) tiene una relación con su edad en días x_1 , su estatura al nacer en cm. (x_2), su peso en Kg. al nacer (x_3) y el aumento en tanto por ciento de su peso actual con respecto a su peso al nacer (x_4). Se pudo obtener una pequeña muestra con $n = 9$ niños cuyos resultados fueron:

y	x_1	x_2	x_3	x_4
57.5	78	48.2	2.75	29.5
52.8	69	45.5	2.15	26.3
61.3	77	46.3	4.41	32.2
67	88	49	5.52	36.5
53.5	67	43	3.21	27.2
62.7	80	48	4.32	27.7
56.2	74	48	2.31	28.3
68.5	94	53	4.3	30.3
69.2	102	58	3.71	28.7

Ejemplo

- La matriz **X** es:

$$\mathbf{X} = \begin{pmatrix} 1 & 78 & 48.2 & 2.75 & 29.5 \\ 1 & 69 & 45.5 & 2.15 & 26.3 \\ 1 & 77 & 46.3 & 4.41 & 32.2 \\ 1 & 88 & 49 & 5.52 & 36.5 \\ 1 & 67 & 43 & 3.21 & 27.2 \\ 1 & 80 & 48 & 4.32 & 27.7 \\ 1 & 74 & 48 & 2.31 & 28.3 \\ 1 & 94 & 53 & 4.3 & 30.3 \\ 1 & 102 & 58 & 3.71 & 28.7 \end{pmatrix}$$

Ejemplo

- El vector \mathbf{y} es:

$$\mathbf{y} = \begin{pmatrix} 57.5 \\ 52.8 \\ 61.3 \\ 67 \\ 53.5 \\ 62.7 \\ 56.2 \\ 68.5 \\ 69.2 \end{pmatrix}$$

Ejemplo

- El producto $\mathbf{X}^T \mathbf{X}$ es el siguiente:

$$\begin{pmatrix} 9 & 729 & 439 & 32.68 & 266.7 \\ 729 & 60123 & 35947.2 & 2702.41 & 21715.3 \\ 439 & 35947.2 & 21568.18 & 1604.388 & 13026.01 \\ 66.07 & 6108.19 & 3541.008 & 128.66 & 1948.561 \\ 266.7 & 21715.3 & 13026.01 & 990.27 & 7980.83 \end{pmatrix}$$

- El producto $\mathbf{X}^T \mathbf{y}$ es el siguiente:

$$\mathbf{X}^T \mathbf{y} = \begin{pmatrix} 548.7 \\ 45001 \\ 26946.89 \\ 2035.52 \\ 16348.29 \end{pmatrix}$$

Ejemplo

- Resolviendo el sistema $(\mathbf{X}^\top \mathbf{X}) \cdot \mathbf{b} = \mathbf{X}^\top \mathbf{y}$ obtenemos como solución:

$$\mathbf{b} = \begin{pmatrix} 7.1475 \\ 0.1001 \\ 0.7264 \\ 3.0758 \\ -0.03 \end{pmatrix}$$

- La recta de regresión estimada es :

$$\hat{y} = 7.1475 + 0.1001x_1 + 0.7264x_2 + 3.0758x_3 - 0.03x_4.$$

Propiedades de la recta de regresión

- La recta de regresión ajustada pasa por el vector de medias. O sea, si llamamos $\bar{x}_j = \frac{\sum_{i=1}^n x_{ij}}{n}$, para $i = 1, \dots, k$, $\bar{y} = \frac{\sum_{i=1}^n y_i}{n}$, se verifica que:

$$\bar{y} = b_0 + b_1\bar{x}_1 + \dots + b_k\bar{x}_k.$$

- La suma de los errores e_i es 0: $\sum_{i=1}^n e_i = 0$ y por lo tanto su media también: $\bar{e} = 0$.
- La media de los valores estimados coincide con la media de los valores de la muestra: $\bar{\hat{y}} = \bar{y}$.

Ejemplo

- Veamos que la recta de regresión pasa por el vector de medias. Éste vale:

$$\bar{x}_0 = 1, \bar{x}_1 = 81, \bar{x}_2 = 48.778, \bar{x}_3 = 3.631, \bar{x}_4 = 29.633.$$

La media del vector **y** vale: $\bar{y} = 60.967$.

Se cumple:

$$\begin{aligned} 60.967 \approx & 7.1475 + 0.1001 \cdot 81 + 0.7264 \cdot 48.778 \\ & + 3.0758 \cdot 3.6311 - 0.03 \cdot 29.633. \end{aligned}$$

Ejemplo

- Veamos que la media de los errores es nula. Los valores de los vectores $\hat{\mathbf{y}}$ y \mathbf{e} son:

$$\hat{\mathbf{y}} = \mathbf{X} \cdot \mathbf{b} = \begin{pmatrix} 57.541 \\ 52.929 \\ 61.085 \\ 67.432 \\ 54.146 \\ 62.479 \\ 55.678 \\ 67.372 \\ 70.039 \end{pmatrix}, \quad \mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \begin{pmatrix} -0.0405 \\ -0.1290 \\ 0.2150 \\ -0.4324 \\ -0.6461 \\ 0.2214 \\ 0.5225 \\ 1.1277 \\ -0.8385 \end{pmatrix}$$

- Puede comprobarse que la media del vector \mathbf{e} es 0 y que $\overline{\hat{y}} = \bar{y} = 60.967$.

Sumas de cuadrados en la regresión

- Llamaremos suma de cuadrados de los residuales o del error a

$$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2.$$

- Llamaremos suma de cuadrados de totales a

$$SST = \sum_{i=1}^n (y_i - \bar{y})^2.$$

- Llamaremos suma de cuadrados de la regresión a

$$SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2.$$

- Se verifica:

$$SST = SSR + SSE,$$

o equivalentemente:

$$S_y^2 = S_{\hat{y}}^2 + S_e^2.$$

Ejemplo

- La suma de cuadrados del error vale en el ejemplo anterior:

$$SSE = (57.5 - 57.5405)^2 + \cdots + (69.2 - 70.0385)^2 = 2.9656.$$

- La suma de cuadrados totales vale:

$$SST = (57.5 - 60.967)^2 + \cdots + (69.2 - 60.967)^2 = 321.24.$$

- La suma de cuadrados de la regresión es:

$$\begin{aligned} SSR &= (57.5405 - 60.967)^2 + \cdots + (70.0385 - 60.967)^2 \\ &= 318.274. \end{aligned}$$

- Puede observarse que se cumple:

$$SST = SSE + SSR, \quad 321.24 = 2.9656 + 318.274.$$

Definición del coeficiente de determinación

- Definimos el coeficiente de determinación como

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST},$$

o también

$$R^2 = \frac{S_{\hat{y}}^2}{S_y^2} = 1 - \frac{S_e^2}{S_y^2}.$$

R^2 se interpreta como la proporción de varianza de la variable y que es explicada por el modelo de regresión múltiple.

- Definimos el coeficiente de determinación ajustado como:

$$R^2_a = R^2 - \frac{k(1 - R^2)}{n - k - 1}.$$

- Definimos el coeficiente de correlación múltiple r de la variable y respecto de las variables x_1, \dots, x_k como $r = \sqrt{R^2}$.

Ejemplo

- El coeficiente de determinación vale en nuestro ejemplo:

$$R^2 = \frac{318.274}{321.24} \approx 0.9908.$$

- El coeficiente de determinación ajustado vale:

$$R^2_a = 0.9908 - \frac{4 \cdot (1 - 0.9908)}{9 - 4 - 1} = 0.9815.$$

- El coeficiente de correlación múltiple r de la variable y respecto de las variables x_1, x_2, x_3, x_4 vale: $r = \sqrt{0.9908} = 0.9954$.

Consideraciones sobre el modelos de regresión múltiple

- Suponemos que las variables aleatorias error E_i son independientes e idénticamente distribuidas según una normal de media 0 y varianza σ^2 .
- Bajo el supuesto anterior, los estimadores b_0, \dots, b_k de β_0, \dots, β_k son insesgados. O sea, $E(b_i) = \beta_i$, $i = 0, \dots, k$.
- La matriz $(X^\top X)^{-1}\sigma^2$ es la matriz de covarianzas de β_0, \dots, β_k .
- Un estimador insesgado de σ^2 es

$$s^2 = \frac{SSE}{n - k - 1}.$$

Ejemplo

- Una estimación de la varianza σ^2 será:

$$S^2 = \frac{2.9656}{9 - 4 - 1} = 0.7414.$$

- Una estimación de la matriz de covarianzas de β_0, \dots, β_4 :
 $(X^\top X)^{-1} S^2 =$

$$\begin{pmatrix} 270.919 & 5.325 & -12.521 & -13.743 & -1.4 \\ 5.325 & 0.115 & -0.266 & -0.326 & -0.0176 \\ -12.521 & -0.266 & 0.618 & 0.742 & 0.0416 \\ -13.743 & -0.326 & 0.742 & 1.122 & -0.00598 \\ -1.4 & -0.0176 & 0.0416 & -0.00598 & 0.0277 \end{pmatrix}$$

ANOVA en regresión lineal múltiple

- El contraste ANOVA en la regresión lineal múltiple nos permite contrastar la adecuación del modelo. Se trata de contrastar:

$$\begin{cases} H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0 \\ H_1 : \text{hay alguna } \beta_i \neq 0 \end{cases}$$

- Si aceptamos H_0 estamos diciendo que la estimación dada por la regresión es constante. Por tanto el modelo no sería adecuado.
- En la tabla siguiente aparece los pasos necesarios para realizar el contraste. Como puede observarse, se usa como estadístico de contraste el cociente $\frac{MSR}{MSE}$ que, suponiendo normalidad, sigue una distribución F de Snédecor de $k, n - k - 1$ grados de libertad:

ANOVA en regresión lineal múltiple

F.V.	S.C.	g.l.	C.M.	f
Regresión	SSR	k	$MSR = \frac{SSR}{k}$	$f = \frac{MSR}{MSE}$
Error	SSE	$n - k - 1$	$MSE = \frac{SSE}{n - k - 1}$	
Total	SST	$n - 1$		

F.V.: Fuente de variación.

S.C.: Suma de cuadrados.

g.l.: grados de libertad.

C.M.: Cuadrados medios.

Rechazaremos la hipótesis nula $H_0 : \beta_1 = \dots = \beta_k = 0$ al nivel de significación α si $f > f_{\alpha, k, n-k-1}$ donde $f_{\alpha, k, n-k-1}$ es el valor crítico de una distribución F de con grados de libertad k y $n - k - 1$.

Ejemplo

- La tabla ANOVA es en nuestro ejemplo:

F.V.	S.C.	g.l.	C.M.	f
Regresión	318.274	4	$MSR = 79.569$	$f = 107.323$
Error	2.9656	4	$MSE = 0.7414$	
Total	321.24	8		

- Cogiendo $\alpha = 0.05$, el valor crítico $f_{0.05,4,4}$ vale 6.388. Como $f = 107.323 > f_{0.05,4,4} = 6.388$, rechazamos la hipótesis nula y concluimos que el modelo es adecuado según este análisis.

ANOVA en regresión lineal múltiple

- El modelo de regresión lineal con este conjunto de x puede no ser el único que se puede utilizar. Es posible que con algunas transformaciones de las x mejore el valor de f .
- El modelo podría ser más eficaz si se incluyen otras variables o podría continuar siendo casi igual de eficaz si se eliminan algunas (principio de parsimonia).

Intervalos de confianza

- Un intervalo de confianza al nivel $(1 - \alpha)100\%$ para la respuesta media $\mu_{Y/x_{10}, x_{20}, \dots, x_{k0}}$ es

$$\hat{y}_0 - t_{\alpha/2, n-k-1} S \sqrt{x_0^\top (X^\top X)^{-1} x_0} <$$

$$\mu_{Y/x_{10}, x_{20}, \dots, x_{k0}} < \hat{y}_0 + t_{\alpha/2, n-k-1} S \sqrt{x_0^\top (X^\top X)^{-1} x_0}$$

donde $t_{\alpha/2, n-k-1}$ es el valor crítico de una t de student con $n - k - 1$ grados de libertad, $x_0 = (1, x_{10}, x_{20}, \dots, x_{k0})^\top$ y $\hat{y}_0 = b_0 + b_1 x_{10} + \dots + b_k x_{k0}$.

- La cantidad $S \sqrt{x_0^\top (X^\top X)^{-1} x_0}$ recibe el nombre de error estándar de predicción.

Ejemplo

- Para $\alpha = 0.05$, hallemos un intervalo de confianza para la respuesta media $\mu_{Y/x_{10}, x_{20}, x_{30}, x_{40}}$, para $x_{10} = 69$, $x_{20} = 45.5$, $x_{30} = 2.15$, $x_{40} = 26.3$.
- El valor crítico $t_{0.025,4}$ vale 2.776. El valor \hat{y}_0 valdrá:

$$\begin{aligned}\hat{y}_0 &= b_0 + \sum_{i=1}^4 b_i x_{i0} \\ &= 7.1475 + 0.1001 \cdot 69 + 0.7264 \cdot 45.5 + 3.0758 \cdot 2.15 \\ &\quad - 0.03 \cdot 26.3 = 52.929.\end{aligned}$$

- El valor de $x_0^\top (X^\top X)^{-1} x_0$ es:

$$(1, 69, 45.5, 2.15, 26.3) \cdot (X^\top X)^{-1} \cdot \begin{pmatrix} 1 \\ 69 \\ 45.5 \\ 2.15 \\ 26.3 \end{pmatrix} = 0.3615.$$

Ejemplo

El intervalo de confianza será:

$$\begin{aligned} 52.929 - 2.776\sqrt{0.7414 \cdot 0.3615} &< \mu_{Y/X_{10}, X_{20}, X_{30}, X_{40}} \\ &< 52.929 + 2.776 \cdot \sqrt{0.7414 \cdot 0.3615} = \\ 51.492 &< \mu_{Y/X_{10}, X_{20}, X_{30}, X_{40}} < 54.366. \end{aligned}$$

Intervalos de confianza

- Un intervalo de confianza al nivel $(1 - \alpha)100\%$ para una predicción individual y_0 para los valores de la variables dependientes $x_{10}, x_{20}, \dots, x_{k0}$ es

$$\hat{y}_0 - t_{\alpha/2, n-k-1} S \sqrt{1 + x_0^\top (X^\top X)^{-1} x_0} < y_0 <$$

$$\hat{y}_0 + t_{\alpha/2, n-k-1} S \sqrt{1 + x_0^\top (X^\top X)^{-1} x_0},$$

donde $t_{\alpha/2, n-k-1}$ es el valor crítico de una t de student con $n - k - 1$ grados de libertad, y $x_0 = (1, x_{10}, x_{20}, \dots, x_{k0})^\top$.

Intervalos de confianza

- Un intervalo de confianza al nivel $(1 - \alpha)100\%$ para el parámetro β_k es:

$$b_k - t_{\alpha/2, n-k-1} s_{\beta_k} < \beta_k < b_k + t_{\alpha/2, n-k-1} s_{\beta_k},$$

donde s_{β_k} es la raíz cuadrada del elemento k -ésimo de la diagonal de la matriz $(X^T X)^{-1} S^2$: $s_{\beta_k} = \sqrt{((X^T X)^{-1} S^2)_{kk}}$.

Ejemplo

- Para $\alpha = 0.05$, hallemos un intervalo de confianza para el parámetro β_2 .
- El valor crítico $t_{0.025,4}$ vale 2.776. Los valores de la diagonal de la matriz $(X^T X)^{-1} S^2$ son:

$$270.919, \quad 0.1154, \quad 0.6176, \quad 1.1219, \quad 0.02775.$$

El valor que nos interesa es para el parámetro β_2 : 0.726.

- El intervalo será:

$$\begin{aligned} 0.726 - 2.776 \cdot \sqrt{0.6176} &< \beta_2 \\ &< 0.726 + 2.776 \cdot \sqrt{0.6176}, \\ -1.4556 &< \beta_2 < 2.908. \end{aligned}$$

El problema de la selección del modelo. Colinealidad

- Dado un problema regresión lineal múltiple podemos ajustar todos los submodelos lineales posibles

$$Y = \beta_0 + \beta_1 x_1,$$

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2,$$

...

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k,$$

ya que es posible que entre variables x_i exista una fuerte relación lineal y entonces *sobren del modelo*. Este problema recibe el nombre de colinealidad.

- Existen también otro tipo de problemas que se pueden resolver usando la técnica de la regresión lineal múltiple. Véase como ejemplo el problema de la regresión polinomial:

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k.$$

El problema de la selección del modelo. Colinealidad

Una solución para ver qué modelo lineal es el más simple y adecuado es recurrir a los llamados métodos secuenciales de selección del modelo lineal, como son los siguientes:

- Regresión paso paso (Stepwise).
- Selección hacia adelante (Forward).
- Selección hacia atrás (Backward).

Parte XII

Análisis de correspondencias

Introducción

- El análisis de correspondencias es una técnica descriptiva para representar y estudiar tablas de contingencia de variables cualitativas. Es decir, vamos a estudiar las frecuencias de aparición de dos o más variables cualitativas en un conjunto de elementos.
- Dicho análisis constituye la aplicación de técnicas de componentes principales y escalado multidimensional vistas anteriormente para variables cualitativas.
- Nuestra información de partida es una matriz \mathbf{X} de dimensiones $I \times J$ donde cada valor de la matriz representa la frecuencia absoluta observada de dos variables cualitativas en n elementos.
- Suponemos que la primera variable puede tomar I valores diferentes y que la segunda variable, J . Entonces, x_{ij} sería el número de elementos de entre los n que toma el valor i -ésimo para la primera variable y toma el valor j -ésimo para la segunda.

Ejemplo

- Consideremos el siguiente experimento: tenemos 61 ratas adultas y 61 crías de rata. Cada rata puede tener un genotipo diferente de entre cuatro: A, B, I y J. Ponemos cada cría aleatoriamente junto con una rata adulta para que crezca. Al cabo de 28 días, anotamos el porcentaje de peso adquirido por la cría de rata. Anotamos los resultados obtenidos en la tabla siguiente:

	1	2	3	4	5	6	7
A A	0	0	0	0	1	3	1
A B	0	1	0	1	1	0	0
A I	0	0	1	2	0	1	0
A J	1	1	1	1	1	0	0
B A	0	0	2	1	1	0	0
B B	0	0	1	0	0	4	0
B I	0	0	1	1	2	0	0
B J	1	0	0	1	0	0	0
I A	2	0	0	0	0	0	1
I B	0	0	0	0	1	0	2
I I	1	1	0	2	0	1	0
I J	0	1	1	1	0	0	0
J A	0	0	1	1	2	0	0
J B	0	0	0	2	1	0	0
J I	0	1	0	0	1	1	0
J J	0	2	0	3	0	0	0

Ejemplo

- La primera fila nos indica el porcentaje de peso adquirido por la cría de rata (1 indica que ha adquirido poco peso y 7 que ha adquirido mucho peso) y la primera columna nos indica el cruce que hemos hecho (por ejemplo B A indica que hemos puesto una cría de rata de genotipo B con una rata adulta de genotipo A).
- La tabla nos da el número de crías de rata que han aumentado un determinado porcentaje de peso en 28 días usando un cruce determinado.

Búsqueda de la mejor proyección

- Llamaremos $\mathbf{F} = (f_{ij})_{i=1,\dots,I,j=1,\dots,J}$ a la matriz de frecuencias relativas de la tabla de contingencia. Esto es, dividimos cada elemento de la matriz \mathbf{X} por el número total de elementos n : $f_{ij} = \frac{x_{ij}}{n}$.
- Los elementos de la matriz anterior verifican: $\sum_{i=1}^I \sum_{j=1}^J f_{ij} = 1$.
- Cualquier estudio aplicado a dicha matriz debe ser equivalente al estudio aplicado a su traspuesta ya que elegir la variable que va por filas o por columnas es una elección arbitraria y no debe influir en el análisis.

Ejemplo anterior

- En el ejemplo anterior la matriz **F** será:

	1	2	3	4	5	6	7
A A	0/61	0/61	0/61	0/61	1/61	3/61	1/61
A B	0/61	1/61	0/61	1/61	1/61	0/61	0/61
A I	0/61	0/61	1/61	2/61	0/61	1/61	0/61
A J	1/61	1/61	1/61	1/61	1/61	0/61	0/61
B A	0/61	0/61	2/61	1/61	1/61	0/61	0/61
B B	0/61	0/61	1/61	0/61	0/61	4/61	0/61
B I	0/61	0/61	1/61	1/61	2/61	0/61	0/61
B J	1/61	0/61	0/61	1/61	0/61	0/61	0/61
I A	2/61	0/61	0/61	0/61	0/61	0/61	1/61
I B	0/61	0/61	0/61	0/61	1/61	0/61	2/61
I I	1/61	1/61	0/61	2/61	0/61	1/61	0/61
I J	0/61	1/61	1/61	1/61	0/61	0/61	0/61
J A	0/61	0/61	1/61	1/61	2/61	0/61	0/61
J B	0/61	0/61	0/61	2/61	1/61	0/61	0/61
J I	0/61	1/61	0/61	0/61	1/61	1/61	0/61
J J	0/61	2/61	0/61	3/61	0/61	0/61	0/61

Ejemplo anterior

- O, si se quiere:

	1	2	3	4	5	6	7
A A	0.00	0.00	0.00	0.00	0.02	0.05	0.02
A B	0.00	0.02	0.00	0.02	0.02	0.00	0.00
A I	0.00	0.00	0.02	0.03	0.00	0.02	0.00
A J	0.02	0.02	0.02	0.02	0.02	0.00	0.00
B A	0.00	0.00	0.03	0.02	0.02	0.00	0.00
B B	0.00	0.00	0.02	0.00	0.00	0.07	0.00
B I	0.00	0.00	0.02	0.02	0.03	0.00	0.00
B J	0.02	0.00	0.00	0.02	0.00	0.00	0.00
I A	0.03	0.00	0.00	0.00	0.00	0.00	0.02
I B	0.00	0.00	0.00	0.00	0.02	0.00	0.03
I I	0.02	0.02	0.00	0.03	0.00	0.02	0.00
I J	0.00	0.02	0.02	0.02	0.00	0.00	0.00
J A	0.00	0.00	0.02	0.02	0.03	0.00	0.00
J B	0.00	0.00	0.00	0.03	0.02	0.00	0.00
J I	0.00	0.02	0.00	0.00	0.02	0.02	0.00
J J	0.00	0.03	0.00	0.05	0.00	0.00	0.00

Proyección de las filas

- Vamos a realizar un análisis de la matriz de frecuencias relativas \mathbf{F} por filas.
- Consideramos las I filas como I puntos en el espacio \mathbb{R}^J .
- El objetivo de nuestro análisis es buscar una representación de estos I puntos en un espacio de dimensión menor que nos permita apreciar sus distancias relativas.
- En nuestro análisis debemos tener en cuenta:
 - ▶ No todas las filas (puntos en \mathbb{R}^J) tienen el mismo peso ya que algunas filas contienen más datos que otras. Por tanto debemos dar más peso a aquellas filas que contengan más datos.
 - ▶ La distancia euclídea utilizada en el análisis multidimensional no es una buena medida en este caso para estudiar la proximidad entre las filas.

Proyección de las filas

- Definimos la frecuencia relativa de la fila i -ésima como: $f_{i\bullet} = \sum_{j=1}^J f_{ij}$.

Llamando \mathbf{f} al vector de frecuencias relativas de las filas $\mathbf{f} = (f_{i\bullet})_{i=1,\dots,I}$, podemos escribir matricialmente: $\mathbf{f} = \mathbf{F}\mathbf{1}$.

- Sea la matriz $\mathbf{D}_f = \text{diag}(f_{1\bullet}, \dots, f_{I\bullet})$.
- De la misma forma, definimos la frecuencia relativa de la columna j -ésima como: $f_{\bullet j} = \sum_{i=1}^I f_{ij}$. Llamando \mathbf{c} al vector de frecuencias relativas de las columnas $\mathbf{c} = (f_{\bullet j})_{j=1,\dots,J}$, podemos escribir matricialmente: $\mathbf{c} = \mathbf{F}^\top \mathbf{1}$.
- De la misma forma que antes, definimos la matriz $\mathbf{D}_c = \text{diag}(f_{\bullet 1}, \dots, f_{\bullet J})$.

Proyección de las filas

- Seguidamente definimos la matriz siguiente que nos permitirá realizar la proyección de las frecuencias por filas. Dicha matriz es

$$\mathbf{Z} = \mathbf{D}_f^{-1/2} \mathbf{F} \mathbf{D}_c^{-1/2} = \left(\frac{f_{ij}}{\sqrt{f_{i\bullet} \cdot f_{\bullet j}}} \right)_{i=1, \dots, I, j=1, \dots, J}.$$

- Para obtener la mejor representación bidimensional de las filas de la tabla de contingencia, hay que seguir los pasos siguientes:
 - ▶ Calcular la matriz $\mathbf{Z}^\top \mathbf{Z}$ y obtener sus vectores y valores propios.
 - ▶ Tomar los dos vectores propios, \mathbf{v}_1 y \mathbf{v}_2 ligados a los dos mayores valores propios menores que la unidad de esta matriz.
 - ▶ Calcular las proyecciones siguientes $\mathbf{D}_f^{-1} \mathbf{F} \mathbf{D}_c^{-1/2} \mathbf{v}_i$, $i = 1, 2$ y representarlas gráficamente en un espacio bidimensional.

Ejemplo anterior

- El valor del vector \mathbf{f} vale en nuestro ejemplo:

$$\mathbf{f} = \begin{pmatrix} AA & 0.08 \\ AB & 0.05 \\ AI & 0.07 \\ AJ & 0.08 \\ BA & 0.07 \\ BB & 0.08 \\ BI & 0.07 \\ BJ & 0.03 \\ IA & 0.05 \\ IB & 0.05 \\ II & 0.08 \\ IJ & 0.05 \\ JA & 0.07 \\ JB & 0.05 \\ JI & 0.05 \\ JJ & 0.08 \end{pmatrix}$$

Ejemplo anterior

- La matriz D_f será:

$$\begin{pmatrix} 0.08, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00 \\ 0.00, 0.05, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00 \\ 0.00, 0.00, 0.07, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00 \\ 0.00, 0.00, 0.00, 0.08, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00 \\ 0.00, 0.00, 0.00, 0.00, 0.07, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00 \\ 0.00, 0.00, 0.00, 0.00, 0.00, 0.08, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00 \\ 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.07, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00 \\ 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.03, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00 \\ 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.05, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00 \\ 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.05, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00 \\ 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.08, 0.00, 0.00, 0.00, 0.00, 0.00 \\ 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.05, 0.00, 0.00, 0.00, 0.00, 0.00 \\ 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.07, 0.00, 0.00, 0.00 \\ 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.05, 0.00, 0.00 \\ 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.05, 0.00 \\ 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.08 \end{pmatrix}$$

Ejemplo anterior

- El vector \mathbf{c} que nos da la suma de columnas vale en nuestro ejemplo:

$$\mathbf{c} = \begin{pmatrix} 0.08 \\ 0.11 \\ 0.13 \\ 0.26 \\ 0.18 \\ 0.16 \\ 0.07 \end{pmatrix}$$

- La matriz \mathbf{D}_c valdrá, por tanto:

$$\mathbf{D}_c = \begin{pmatrix} 0.08 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.11 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.13 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.26 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.18 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.16 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.07 \end{pmatrix}$$

Ejemplo anterior

- La matriz **Z** será:

$$\mathbf{Z} = \begin{pmatrix} 0.00 & 0.00 & 0.00 & 0.00 & 0.13 & 0.42 & 0.22 \\ 0.00 & 0.22 & 0.00 & 0.14 & 0.17 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.18 & 0.25 & 0.00 & 0.16 & 0.00 \\ 0.20 & 0.17 & 0.16 & 0.11 & 0.13 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.35 & 0.12 & 0.15 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.16 & 0.00 & 0.00 & 0.57 & 0.00 \\ 0.00 & 0.00 & 0.18 & 0.12 & 0.30 & 0.00 & 0.00 \\ 0.32 & 0.00 & 0.00 & 0.18 & 0.00 & 0.00 & 0.00 \\ 0.52 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.29 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.17 & 0.00 & 0.58 \\ 0.20 & 0.17 & 0.00 & 0.22 & 0.00 & 0.14 & 0.00 \\ 0.00 & 0.22 & 0.20 & 0.14 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.18 & 0.12 & 0.30 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.29 & 0.17 & 0.00 & 0.00 \\ 0.00 & 0.22 & 0.00 & 0.00 & 0.17 & 0.18 & 0.00 \\ 0.00 & 0.34 & 0.00 & 0.34 & 0.00 & 0.00 & 0.00 \end{pmatrix}$$

Ejemplo anterior

- La matriz $\mathbf{Z}^\top \mathbf{Z}$ vale:

$$\mathbf{Z}^\top \mathbf{Z} = \begin{pmatrix} 0.45 & 0.07 & 0.03 & 0.12 & 0.03 & 0.03 & 0.15 \\ 0.07 & 0.31 & 0.07 & 0.23 & 0.10 & 0.06 & 0.00 \\ 0.03 & 0.07 & 0.31 & 0.18 & 0.18 & 0.12 & 0.00 \\ 0.12 & 0.23 & 0.18 & 0.44 & 0.18 & 0.07 & 0.00 \\ 0.03 & 0.10 & 0.18 & 0.18 & 0.36 & 0.09 & 0.13 \\ 0.03 & 0.06 & 0.12 & 0.07 & 0.09 & 0.58 & 0.09 \\ 0.15 & 0.00 & 0.00 & 0.00 & 0.13 & 0.09 & 0.47 \end{pmatrix}$$

- Los valores propios de la matriz anterior son:

1.00, 0.57, 0.52, 0.37, 0.24, 0.12, 0.11.

Tenemos que considerar, por tanto los vectores propios asociados a los valores propios 0.57 y 0.52.

Ejemplo anterior

- Los vectores propios asociados a los valores propios anteriores son:
(por columnas)

$$\begin{pmatrix} -0.28 & 0.57 \\ 0.29 & 0.10 \\ 0.20 & -0.15 \\ 0.42 & 0.17 \\ 0.06 & -0.00 \\ -0.37 & -0.74 \\ -0.69 & 0.27 \end{pmatrix}$$

Ejemplo anterior

- Antes de hallar las proyecciones, calculamos la matriz $\mathbf{D}_f^{-1}\mathbf{F}\mathbf{D}_c^{-1/2}$:

$$\mathbf{D}_f^{-1}\mathbf{F}\mathbf{D}_c^{-1/2} = \begin{pmatrix} 0.00 & 0.00 & 0.00 & 0.00 & 0.47 & 1.48 & 0.78 \\ 0.00 & 0.98 & 0.00 & 0.65 & 0.78 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.69 & 0.98 & 0.00 & 0.62 & 0.00 \\ 0.70 & 0.59 & 0.55 & 0.39 & 0.47 & 0.00 & 0.00 \\ 0.00 & 0.00 & 1.38 & 0.49 & 0.59 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.55 & 0.00 & 0.00 & 1.98 & 0.00 \\ 0.00 & 0.00 & 0.69 & 0.49 & 1.18 & 0.00 & 0.00 \\ 1.75 & 0.00 & 0.00 & 0.98 & 0.00 & 0.00 & 0.00 \\ 2.33 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 1.30 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.78 & 0.00 & 2.60 \\ 0.70 & 0.59 & 0.00 & 0.78 & 0.00 & 0.49 & 0.00 \\ 0.00 & 0.98 & 0.92 & 0.65 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.69 & 0.49 & 1.18 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 1.30 & 0.78 & 0.00 & 0.00 \\ 0.00 & 0.98 & 0.00 & 0.00 & 0.78 & 0.82 & 0.00 \\ 0.00 & 1.18 & 0.00 & 1.17 & 0.00 & 0.00 & 0.00 \end{pmatrix}$$

Ejemplo anterior

- Para hallar las proyecciones, basta multiplicar la matriz anterior por la matriz de los dos vectores propios considerados anteriormente:

$$\begin{pmatrix} -1.07 & -0.89 \\ 0.60 & 0.21 \\ 0.32 & -0.39 \\ 0.28 & 0.44 \\ 0.52 & -0.12 \\ -0.63 & -1.54 \\ 0.41 & -0.02 \\ -0.09 & 1.16 \\ -1.56 & 1.67 \\ -1.75 & 0.69 \\ 0.11 & 0.23 \\ 0.75 & 0.07 \\ 0.41 & -0.02 \\ 0.59 & 0.22 \\ 0.02 & -0.51 \\ 0.83 & 0.32 \end{pmatrix}$$

Ejemplo anterior

- El gráfico bidimensional de las proyecciones es el siguiente donde puede observarse por ejemplo que las crías de rata con genotipo B criadas por ratas con genotipo B tienen un porcentaje de aumento de peso muy distinto al cabo de 28 días que crías de rata con genotipo I criadas por ratas con genotipo B:



Proyección de las columnas

- Vamos a realizar el mismo análisis de la matriz de frecuencias relativas \mathbf{F} que hemos hecho anteriormente pero ahora por columnas.
- Consideramos las J columnas como J puntos en el espacio \mathbb{R}^I .
- El objetivo de nuestro análisis es buscar una representación de estos J puntos en un espacio de dimensión menor que nos permita apreciar sus distancias relativas.
- Debemos tener en cuenta las mismas consideraciones que teníamos por filas.

Proyección de las columnas

- Para obtener la mejor representación bidimensional de las filas de la tabla de contingencia, hay que seguir los pasos siguientes:
 - ▶ Calcular la matriz $\mathbf{Z}\mathbf{Z}^\top$ y obtener sus vectores y valores propios. Los valores propios de la matriz anterior son los mismos que los valores propios de la matriz $\mathbf{Z}^\top\mathbf{Z}$ calculada anteriormente.
 - ▶ Tomar los dos vectores propios, \mathbf{w}_1 y \mathbf{w}_2 ligados a los dos mayores valores propios menores que la unidad de esta matriz.
 - ▶ Calcular las proyecciones siguientes $\mathbf{D}_c^{-1}\mathbf{F}^\top\mathbf{D}_f^{-1/2}\mathbf{w}_i$, $i = 1, 2$ y representarlas gráficamente en un espacio bidimensional.

Ejemplo anterior

- La matriz \mathbf{ZZ}^T valdrá:

$$\begin{pmatrix} 0.25, 0.02, 0.07, 0.02, 0.02, 0.24, 0.04, 0.00, 0.06, 0.15, 0.06, 0.00, 0.04, 0.02, 0.10, 0.00 \\ 0.02, 0.10, 0.04, 0.08, 0.04, 0.00, 0.07, 0.03, 0.00, 0.03, 0.07, 0.07, 0.07, 0.07, 0.08, 0.12 \\ 0.07, 0.04, 0.12, 0.06, 0.09, 0.12, 0.06, 0.04, 0.00, 0.00, 0.08, 0.07, 0.06, 0.07, 0.03, 0.08 \\ 0.02, 0.08, 0.06, 0.12, 0.09, 0.03, 0.08, 0.08, 0.10, 0.02, 0.09, 0.09, 0.08, 0.06, 0.06, 0.09 \\ 0.02, 0.04, 0.09, 0.09, 0.16, 0.06, 0.12, 0.02, 0.00, 0.03, 0.03, 0.09, 0.12, 0.06, 0.03, 0.04 \\ 0.24, 0.00, 0.12, 0.03, 0.06, 0.35, 0.03, 0.00, 0.00, 0.00, 0.08, 0.03, 0.03, 0.00, 0.10, 0.00 \\ 0.04, 0.07, 0.06, 0.08, 0.12, 0.03, 0.14, 0.02, 0.00, 0.05, 0.03, 0.05, 0.14, 0.09, 0.05, 0.04 \\ 0.00, 0.03, 0.04, 0.08, 0.02, 0.00, 0.02, 0.13, 0.16, 0.00, 0.10, 0.03, 0.02, 0.05, 0.00, 0.06 \\ 0.06, 0.00, 0.00, 0.10, 0.00, 0.00, 0.00, 0.16, 0.35, 0.17, 0.10, 0.00, 0.00, 0.00, 0.00, 0.00 \\ 0.15, 0.03, 0.00, 0.02, 0.03, 0.00, 0.05, 0.00, 0.17, 0.36, 0.00, 0.00, 0.05, 0.03, 0.03, 0.00 \\ 0.06, 0.07, 0.08, 0.09, 0.03, 0.08, 0.03, 0.10, 0.10, 0.00, 0.14, 0.07, 0.03, 0.06, 0.06, 0.13 \\ 0.00, 0.07, 0.07, 0.09, 0.09, 0.03, 0.05, 0.03, 0.00, 0.00, 0.07, 0.11, 0.05, 0.04, 0.05, 0.12 \\ 0.04, 0.07, 0.06, 0.08, 0.12, 0.03, 0.14, 0.02, 0.00, 0.05, 0.03, 0.05, 0.14, 0.09, 0.05, 0.04 \\ 0.02, 0.07, 0.06, 0.06, 0.00, 0.09, 0.05, 0.00, 0.03, 0.06, 0.04, 0.09, 0.11, 0.03, 0.10 \\ 0.10, 0.08, 0.03, 0.06, 0.03, 0.10, 0.05, 0.00, 0.00, 0.03, 0.06, 0.05, 0.05, 0.03, 0.11, 0.07 \\ 0.00, 0.12, 0.08, 0.09, 0.04, 0.00, 0.04, 0.06, 0.00, 0.00, 0.13, 0.12, 0.04, 0.10, 0.07, 0.23 \end{pmatrix}$$

- Los valores propios de la matriz anterior son:

$$1.00, 0.57, 0.52, 0.37, 0.24, 0.12, 0.11, 0.00, \\ 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00.$$

Obsérvese que los valores propios no nulos ya estaban calculados al hacer el análisis por filas.

Ejemplo anterior

- Los vectores propios correspondientes a los valores propios 0.57 y 0.52 son:

$$\mathbf{w} = \begin{pmatrix} 0.41 & 0.35 \\ -0.18 & -0.06 \\ -0.11 & 0.14 \\ -0.10 & -0.17 \\ -0.18 & 0.04 \\ 0.24 & 0.61 \\ -0.14 & 0.01 \\ 0.02 & -0.29 \\ 0.46 & -0.51 \\ 0.52 & -0.21 \\ -0.04 & -0.09 \\ -0.22 & -0.02 \\ -0.14 & 0.01 \\ -0.17 & -0.07 \\ -0.01 & 0.16 \\ -0.32 & -0.13 \end{pmatrix}$$

Ejemplo anterior

- La matriz $\mathbf{D}_c^{-1} \mathbf{F}^\top \mathbf{D}_f^{-1/2}$ vale en nuestro caso:

$$\begin{pmatrix} 0.00, 0.00, 0.00, 0.70, 0.00, 0.00, 0.00, 1.10, 1.80, 0.00, 0.70, 0.00, 0.00, 0.00, 0.00, 0.00 \\ 0.00, 0.64, 0.00, 0.50, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.50, 0.64, 0.00, 0.00, 0.64, 1.00 \\ 0.00, 0.00, 0.49, 0.44, 0.98, 0.44, 0.49, 0.00, 0.00, 0.00, 0.00, 0.00, 0.56, 0.49, 0.00, 0.00, 0.00 \\ 0.00, 0.28, 0.49, 0.22, 0.24, 0.00, 0.24, 0.35, 0.00, 0.00, 0.44, 0.28, 0.24, 0.56, 0.00, 0.65 \\ 0.32, 0.41, 0.00, 0.32, 0.36, 0.00, 0.71, 0.00, 0.00, 0.41, 0.00, 0.00, 0.71, 0.41, 0.41, 0.00 \\ 1.05, 0.00, 0.39, 0.00, 0.00, 1.40, 0.00, 0.00, 0.00, 0.00, 0.35, 0.00, 0.00, 0.00, 0.45, 0.00 \\ 0.87, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 1.13, 2.25, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00 \end{pmatrix}$$

- Las proyecciones $\mathbf{D}_c^{-1} \mathbf{F}^\top \mathbf{D}_f^{-1/2} \mathbf{w}$ calculadas por columnas valen:

$$\begin{pmatrix} 0.75 & -1.43 \\ -0.65 & -0.21 \\ -0.43 & 0.29 \\ -0.61 & -0.24 \\ -0.10 & 0.01 \\ 0.70 & 1.31 \\ 2.03 & -0.75 \end{pmatrix}$$

Ejemplo anterior

- La representación gráfica de las proyecciones anteriores es:



AnCoi

Ejemplo anterior

- En el gráfico anterior se puede observar las similitudes o las diferencias del genotipo entre las crías de rata y las ratas adultas que las crían con diferente porcentaje de aumento de peso al cabo de 28 días.
- Por ejemplo, vemos que las crías de rata y ratas adultas con porcentaje de aumento de peso en segundo y en cuarto lugar tienen un genotipo muy parecido.
- En cambio, la mayor diferencia entre el genotipo de las crías y las ratas adultas se encuentra en las crías de rata con porcentaje de aumento de peso en primer y sexto lugar.

Análisis conjunto

- Debido a la simetría del problema, conviene representar conjuntamente las proyecciones de las filas y las columnas en el mismo gráfico.
- Antes de dar los pasos para hacer tal representación conviene tener en cuenta que
 - ▶ si \mathbf{v} es vector propio de la matriz $\mathbf{Z}^T \mathbf{Z}$ de valor propio λ , entonces \mathbf{Zv} es vector propio de la matriz \mathbf{ZZ}^T del mismo valor propio y
 - ▶ viceversa: si \mathbf{w} es vector propio de la matriz \mathbf{ZZ}^T de valor propio λ , entonces $\mathbf{Z}^T \mathbf{w}$ es vector propio de la matriz $\mathbf{Z}^T \mathbf{Z}$ del mismo valor propio.

Análisis conjunto

- Para hacer la representación conjunta de las proyecciones de las filas y las columnas hay que realizar los pasos siguientes:
 - ▶ Se calcula la matriz de frecuencias relativa \mathbf{F} .
 - ▶ Se calcula la matriz estandarizada \mathbf{Z} .
 - ▶ Se busca de las dos matrices siguientes, $\mathbf{Z}^\top \mathbf{Z}$ o $\mathbf{Z} \mathbf{Z}^\top$ la que tenga menor dimensión. Supongamos para fijar ideas que es la matriz $\mathbf{Z}^\top \mathbf{Z}$. Se calculan los dos valores propios menores que 1 más grandes de la matriz anterior. Sean \mathbf{v}_1 y \mathbf{v}_2 los dos vectores propios asociados a los dos valores propios anteriores. La proyección de las filas vendrá dada por $\mathbf{D}_f^{-1/2} \mathbf{Z} \mathbf{v}_i$, $i = 1, 2$.
 - ▶ Sean $\mathbf{w}_i = \mathbf{Z} \mathbf{v}_i$, $i = 1, 2$ los vectores propios de la matriz $\mathbf{Z} \mathbf{Z}^\top$ asociados a los valores propios anteriores. La proyección de las columnas vendrá dada por: $\mathbf{D}_c^{-1/2} \mathbf{Z}^\top \mathbf{w}_i$, $i = 1, 2$.

Ejemplo anterior

- En el ejemplo anterior la matriz de menor dimensión era $\mathbf{Z}^T \mathbf{Z}$ (7×7).
- Los valores propios a considerar eran: 0.57 y 0.52.
- Los vectores propios eran:

$$\begin{pmatrix} -0.28 & 0.57 \\ 0.29 & 0.10 \\ 0.20 & -0.15 \\ 0.42 & 0.17 \\ 0.06 & -0.00 \\ -0.37 & -0.74 \\ -0.69 & 0.27 \end{pmatrix}$$

Ejemplo anterior

- La proyección de las filas era:

$$\mathbf{v} = \begin{pmatrix} -1.07 & -0.89 \\ 0.60 & 0.21 \\ 0.32 & -0.39 \\ 0.28 & 0.44 \\ 0.52 & -0.12 \\ -0.63 & -1.54 \\ 0.41 & -0.02 \\ -0.09 & 1.16 \\ -1.56 & 1.67 \\ -1.75 & 0.69 \\ 0.11 & 0.23 \\ 0.75 & 0.07 \\ 0.41 & -0.02 \\ 0.59 & 0.22 \\ 0.02 & -0.51 \\ 0.83 & 0.32 \end{pmatrix}$$

Ejemplo anterior

- Busquemos ahora los vectores propios de la matriz \mathbf{ZZ}^\top : $\mathbf{w}_i = \mathbf{Z}\mathbf{v}_i$, donde \mathbf{v}_i son los vectores propios hallados anteriormente por columnas:

$$\mathbf{w} = \begin{pmatrix} -0.31 & -0.25 \\ 0.13 & 0.05 \\ 0.08 & -0.10 \\ 0.08 & 0.13 \\ 0.13 & -0.03 \\ -0.18 & -0.44 \\ 0.10 & -0.01 \\ -0.02 & 0.21 \\ -0.35 & 0.37 \\ -0.39 & 0.15 \\ 0.03 & 0.06 \\ 0.17 & 0.02 \\ 0.10 & -0.01 \\ 0.13 & 0.05 \\ 0.00 & -0.11 \\ 0.24 & 0.09 \end{pmatrix}$$

Ejemplo anterior

- Las proyecciones de las columnas será la matriz $\mathbf{D}_c^{-1/2} \mathbf{Z}^\top \mathbf{w}$:

$$\begin{pmatrix} -0.56 & 1.03 \\ 0.49 & 0.15 \\ 0.32 & -0.21 \\ 0.46 & 0.17 \\ 0.07 & -0.00 \\ -0.52 & -0.95 \\ -1.53 & 0.54 \end{pmatrix}$$

Ejemplo anterior

- En el gráfico siguiente se puede ver la proyección conjunta donde en rojo están las proyecciones por filas y en negro, las proyecciones por columnas:



AnCorrC

Parte XIII

Escalado multidimensional.

Introducción

- Las técnicas del escalado multidimensional son una generalización de las técnicas de componentes principales cuando se tiene una matriz \mathbf{D} $n \times n$ de distancias o similaridades en lugar de una matriz de observaciones.
- El objetivo de nuestro análisis es representar dicha matriz mediante un conjunto de variables ortogonales $\mathbf{y}_1, \dots, \mathbf{y}_p$ que llamaremos coordenadas principales donde suponemos que $p < n$ de manera que las distancias euclídeas al cuadrado entre las coordenadas de los elementos respecto a estas variables sean iguales o lo más próximas posible.

Introducción

- Dicho de forma más explícita, dada la matriz de distancias o similitudes \mathbf{D} , queremos obtener una matriz $\mathbf{Y} = (y_{ij})_{i=1,\dots,n,j=1,\dots,p}$ $n \times p$ que puede interpretarse como la matriz de datos de p variables en los n individuos y donde la distancia euclídea al cuadrado entre los n individuos venga dada aproximadamente por la matriz $\mathbf{D} = (d_{ij}^2)_{i,j=1,\dots,n}$.
- O sea, para todo $i, j = 1, \dots, n$, $d_{ij}^2 = \sum_{k=1}^p (y_{ik} - y_{jk})^2$.
- El escalado multidimensional comparte con el análisis de componentes principales el objetivo de describir e interpretar los datos.
- Dicho análisis nos permitirá encontrar la estructura de los datos estudiando la similitud de las observaciones, estudiando grupos entre las mismas, viendo si hay observaciones atípicas, etc.

Pasos a realizar para la obtención de la matriz \mathbf{Y}

- En el primer paso obtenemos una matriz auxiliar $\mathbf{Q} = (q_{ij})_{i,j=1,\dots,n}$ $n \times n$ llamada matriz de similitud a partir de la matriz $\mathbf{D} = (d_{ij}^2)_{i,j=1,\dots,n}$ usando la expresión:

$$q_{ij} = -\frac{1}{2} (d_{ij}^2 - d_{i\bullet}^2 - d_{\bullet j}^2 + d_{\bullet\bullet}^2),$$

donde:

$$d_{i\bullet} = \frac{1}{n} \sum_{j=1}^n d_{ij}^2, \quad d_{\bullet j} = \frac{1}{n} \sum_{i=1}^n d_{ij}^2, \quad d_{\bullet\bullet} = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n d_{ij}^2.$$

Matricialmente: $\mathbf{Q} = -\frac{1}{2}\mathbf{PDP}$, donde la matriz \mathbf{P} es $\mathbf{P} = \text{Id} - \frac{1}{n}\mathbf{11}^\top$.

Pasos a realizar para la obtención de la matriz \mathbf{Y}

- En el segundo paso obtendremos la matriz \mathbf{Y} a partir de la matriz de similitud \mathbf{Q} . Para ello, suponiendo que dicha matriz es definida positiva, podemos descomponerla como:

$$\mathbf{Q} = \mathbf{V} \cdot \tilde{\cdot} \cdot \mathbf{V}^T,$$

donde \mathbf{V} es $n \times p$ y contiene los vectores propios correspondientes a valores propios no nulos de la matriz \mathbf{Q} y $\tilde{\cdot}$ es diagonal $p \times p$ y contiene los valores propios no nulos de \mathbf{Q} .

Pasos a realizar para la obtención de la matriz \mathbf{Y}

- Usando que los valores propios de la matriz \mathbf{Q} son positivos, podemos escribir:

$$\mathbf{Q} = \left(\mathbf{V} \cdot \sim^{1/2} \right) \left(\sim^{1/2} \cdot \mathbf{V}^T \right).$$

- Sea $\mathbf{Y} = \mathbf{V} \cdot \sim^{1/2}$. Dicha matriz es $n \times p$ y reproduce la métrica inicial \mathbf{D} . Las columnas de dicha matriz son las variables $\mathbf{y}_1, \dots, \mathbf{y}_p$ ortogonales que queríamos obtener.

Propiedades de la matriz obtenida \mathbf{Y}

- La matriz obtenida por el método descrito anteriormente no es única.
- De hecho, si cogemos una matriz de datos $\mathbf{X} = (x_{ij})_{i=1,\dots,n,j=1,\dots,p}$ $n \times p$, calculamos una matriz de distancias entre sus elementos (filas) usando la expresión:

$$d_{ij}^2 = \sum_{s=1}^p (x_{is} - x_{js})^2,$$

y aplicamos el método descrito anteriormente, lo más probable es que no obtengamos la matriz original \mathbf{X} . O sea, en general $\mathbf{Y} \neq \mathbf{X}$.

Propiedades de la matriz obtenida \mathbf{Y}

- La razón de que no se obtenga la matriz original es que no existe una biyección entre una matriz de datos y una matriz de distancias calculada como se ha indicado anteriormente.
- Dada una matriz de datos \mathbf{X} , la matriz de distancias obtenida a partir de dicha matriz no varía si:
 - ▶ modificamos las medias de las variables,
 - ▶ rotamos los puntos. O sea, multiplicamos la matriz \mathbf{X} por una matriz ortogonal.

Ejemplo

- Consideremos la siguiente matriz de distancias entre 10 proteínas:

$$\begin{pmatrix} 0.00, & 0.09, & 0.17, & 0.19, & 0.15, & 0.08, & 0.05, & 0.20, & 0.07, & 0.18 \\ 0.09, & 0.00, & 0.17, & 0.17, & 0.06, & 0.09, & 0.06, & 0.14, & 0.04, & 0.11 \\ 0.17, & 0.17, & 0.00, & 0.05, & 0.20, & 0.13, & 0.17, & 0.11, & 0.16, & 0.16 \\ 0.19, & 0.17, & 0.05, & 0.00, & 0.20, & 0.13, & 0.18, & 0.13, & 0.16, & 0.17 \\ 0.15, & 0.06, & 0.20, & 0.20, & 0.00, & 0.15, & 0.12, & 0.14, & 0.10, & 0.08 \\ 0.08, & 0.09, & 0.13, & 0.13, & 0.15, & 0.00, & 0.06, & 0.17, & 0.06, & 0.17 \\ 0.05, & 0.06, & 0.17, & 0.18, & 0.12, & 0.06, & 0.00, & 0.18, & 0.03, & 0.15 \\ 0.20, & 0.14, & 0.11, & 0.13, & 0.14, & 0.17, & 0.18, & 0.00, & 0.16, & 0.07 \\ 0.07, & 0.04, & 0.16, & 0.16, & 0.10, & 0.06, & 0.03, & 0.16, & 0.00, & 0.14 \\ 0.18, & 0.11, & 0.16, & 0.17, & 0.08, & 0.17, & 0.15, & 0.07, & 0.14, & 0.00 \end{pmatrix}$$

Ejemplo

- Calculamos la matriz de similitud **Q**. Para ello, primero tenemos que hallar la matriz **P**:

$$\begin{pmatrix} 0.90, -0.10, -0.10, -0.10, -0.10, -0.10, -0.10, -0.10, -0.10, -0.10 \\ -0.10, 0.90, -0.10, -0.10, -0.10, -0.10, -0.10, -0.10, -0.10, -0.10 \\ -0.10, -0.10, 0.90, -0.10, -0.10, -0.10, -0.10, -0.10, -0.10, -0.10 \\ -0.10, -0.10, -0.10, 0.90, -0.10, -0.10, -0.10, -0.10, -0.10, -0.10 \\ -0.10, -0.10, -0.10, -0.10, 0.90, -0.10, -0.10, -0.10, -0.10, -0.10 \\ -0.10, -0.10, -0.10, -0.10, -0.10, 0.90, -0.10, -0.10, -0.10, -0.10 \\ -0.10, -0.10, -0.10, -0.10, -0.10, -0.10, 0.90, -0.10, -0.10, -0.10 \\ -0.10, -0.10, -0.10, -0.10, -0.10, -0.10, -0.10, 0.90, -0.10, -0.10 \\ -0.10, -0.10, -0.10, -0.10, -0.10, -0.10, -0.10, -0.10, 0.90, -0.10 \\ -0.10, -0.10, -0.10, -0.10, -0.10, -0.10, -0.10, -0.10, -0.10, 0.90 \end{pmatrix}$$

Q será entonces: $\mathbf{Q} = -\frac{1}{2}\mathbf{PDP}$.

Ejemplo

- La matriz **Q** será en este caso:

$$\begin{pmatrix} 0.06, & 0.00, & -0.02, & -0.02, & -0.01, & 0.01, & 0.03, & -0.03, & 0.01, & -0.03 \\ 0.00, & 0.04, & -0.03, & -0.03, & 0.02, & -0.00, & 0.01, & -0.02, & 0.01, & -0.00 \\ -0.02, & -0.03, & 0.07, & 0.05, & -0.03, & -0.00, & -0.03, & 0.02, & -0.03, & -0.01 \\ -0.02, & -0.03, & 0.05, & 0.08, & -0.03, & -0.00, & -0.03, & 0.01, & -0.02, & -0.01 \\ -0.01, & 0.02, & -0.03, & -0.03, & 0.06, & -0.02, & -0.01, & -0.00, & -0.00, & 0.02 \\ 0.01, & -0.00, & -0.00, & -0.00, & -0.02, & 0.05, & 0.01, & -0.03, & 0.01, & -0.03 \\ 0.03, & 0.01, & -0.03, & -0.03, & -0.01, & 0.01, & 0.04, & -0.03, & 0.02, & -0.02 \\ -0.03, & -0.02, & 0.02, & 0.01, & -0.00, & -0.03, & -0.03, & 0.07, & -0.03, & 0.03 \\ 0.01, & 0.02, & -0.03, & -0.02, & -0.00, & 0.01, & 0.02, & -0.03, & 0.03, & -0.02 \\ -0.03, & -0.00, & -0.01, & -0.01, & 0.02, & -0.03, & -0.02, & 0.03, & -0.02, & 0.07 \end{pmatrix}$$

- Los valores propios de la matriz **Q** son aproximadamente:

0.22, 0.16, 0.05, 0.04, 0.03, 0.03, 0.02, 0.02, 0.01, 0.00

Ejemplo

- Sólo vamos a tener en cuenta los 9 primeros valores propios ya que el último es nulo. Los vectores propios correspondientes a estos 9 valores propios son:

$$\begin{pmatrix} -0.33, -0.22, 0.41, 0.64, 0.00, -0.15, -0.33, 0.07, 0.16 \\ -0.23, 0.17, -0.29, -0.11, -0.42, -0.10, -0.12, 0.65, -0.30 \\ 0.44, -0.34, -0.03, 0.26, -0.16, -0.15, 0.67, 0.14, 0.03 \\ 0.44, -0.36, -0.40, 0.04, 0.06, 0.33, -0.53, -0.11, -0.07 \\ -0.10, 0.48, -0.51, 0.28, 0.15, -0.35, 0.07, -0.42, 0.02 \\ -0.17, -0.31, 0.02, -0.47, 0.60, -0.41, 0.00, 0.13, -0.01 \\ -0.35, -0.12, 0.17, -0.08, -0.05, 0.39, 0.24, -0.39, -0.60 \\ 0.41, 0.28, 0.49, -0.31, -0.34, -0.31, -0.23, -0.25, -0.05 \\ -0.29, -0.06, -0.09, -0.32, -0.30, 0.28, 0.11, -0.13, 0.71 \\ 0.20, 0.50, 0.23, 0.07, 0.45, 0.47, 0.10, 0.33, 0.10 \end{pmatrix}$$

- La matriz anterior será la matriz **V** del algoritmo.

Ejemplo

- Puede comprobarse que, en este caso: $\mathbf{Q} = \mathbf{V}\tilde{\mathbf{V}}^T$, donde

$$\tilde{\mathbf{V}} = \text{diag}(0.22, 0.16, 0.05, 0.04, 0.03, 0.03, 0.02, 0.02, 0.01).$$

- Las coordenadas principales serán:

$$\mathbf{Y} = \mathbf{V}^{-1/2} = \begin{pmatrix} -0.15, -0.09, 0.10, 0.13, 0.00, -0.02, -0.05, 0.01, 0.02 \\ -0.11, 0.07, -0.07, -0.02, -0.07, -0.02, -0.02, 0.08, -0.03 \\ 0.20, -0.14, -0.01, 0.05, -0.03, -0.02, 0.10, 0.02, 0.00 \\ 0.20, -0.14, -0.09, 0.01, 0.01, 0.05, -0.08, -0.01, -0.01 \\ -0.05, 0.19, -0.12, 0.06, 0.03, -0.06, 0.01, -0.05, 0.00 \\ -0.08, -0.13, 0.01, -0.09, 0.11, -0.07, 0.00, 0.02, -0.00 \\ -0.16, -0.05, 0.04, -0.02, -0.01, 0.06, 0.04, -0.05, -0.06 \\ 0.19, 0.11, 0.11, -0.06, -0.06, -0.05, -0.03, -0.03, -0.01 \\ -0.14, -0.02, -0.02, -0.06, -0.05, 0.04, 0.02, -0.02, 0.08 \\ 0.09, 0.20, 0.05, 0.01, 0.08, 0.08, 0.01, 0.04, 0.01 \end{pmatrix}$$

- Podemos comprobar que si hallamos la matriz de distancias al cuadrado de la matriz \mathbf{Y} obtenemos la matriz inicial \mathbf{D} .

Matrices compatibles con métricas euclídeas.

- Para poder usar el algoritmo indicado anteriormente, necesitamos que la matriz \mathbf{Q} sea semidefinida positiva o que no tenga ningún valor propio no negativo. Dicha condición no se cumple siempre.
- Diremos que una matriz de distancias \mathbf{D} es compatible con una métrica euclídea si la matriz de similitud $\mathbf{Q} = -\frac{1}{2}\mathbf{PDP}$ obtenida a partir de ella es semidefinida positiva.
- Se puede demostrar que la condición anterior es necesaria y suficiente. O sea, si \mathbf{D} se ha construido a partir de una métrica euclídea, \mathbf{Q} es semidefinida positiva y si \mathbf{Q} es semidefinida positiva, es posible encontrar una métrica euclídea que reproduzca \mathbf{D} .

Cálculo de las coordenadas principales en general.

- En el caso en que la matriz de distancias **D** no sea compatible con métricas euclídeas, es posible en algunos casos hallar una aproximación de las coordenadas principales.
- Para hallar dicha aproximación, hay que realizar los pasos siguientes:
 - ▶ Hallar la matriz $\mathbf{Q} = -\frac{1}{2}\mathbf{PDP}$ de similitud.
 - ▶ Obtener los valores propios de **Q**. Aunque pueden aparecer valores propios negativos, es posible que existan r valores propios positivos que en módulo sobresalgan sobre los anteriores. Si éste es el caso, tomamos estos r valores propios.
 - ▶ Sea \mathbf{V}_r la matriz de los vectores propios correspondientes a los r valores propios anteriores.
 - ▶ Definimos la aproximación de las coordenadas principales como:
 $\mathbf{Y}_r \approx \mathbf{V}_r \tilde{\Lambda}_r^{1/2}$, donde $\tilde{\Lambda}_r$ es una matriz diagonal con los r valores propios en la diagonal.

Ejemplo anterior

- Aunque en el ejemplo anterior, la matriz de distancias **D** es compatible con métricas euclídeas, hallemos una aproximación de las coordenadas principales.
- En el cálculo de los valores propios de la matriz **Q** vimos que había dos que sobresalían sobre los demás. Éstos eran 0.22 y 0.16.
- La matriz **V_r** de vectores propios correspondiente a dichos valores propios es:

$$\mathbf{V}_r = \begin{pmatrix} -0.33 & -0.22 \\ -0.23 & 0.17 \\ 0.44 & -0.34 \\ 0.44 & -0.36 \\ -0.10 & 0.48 \\ -0.17 & -0.31 \\ -0.35 & -0.12 \\ 0.41 & 0.28 \\ -0.29 & -0.06 \\ 0.20 & 0.50 \end{pmatrix}$$

Ejemplo anterior

- La aproximación de las coordenadas principales será:

$$\mathbf{Y}_r \approx \mathbf{V}_r \mathbf{\Lambda}_r^{1/2} = \begin{pmatrix} -0.33 & -0.22 \\ -0.23 & 0.17 \\ 0.44 & -0.34 \\ 0.44 & -0.36 \\ -0.10 & 0.48 \\ -0.17 & -0.31 \\ -0.35 & -0.12 \\ 0.41 & 0.28 \\ -0.29 & -0.06 \\ 0.20 & 0.50 \end{pmatrix} \begin{pmatrix} 0.22 & 0.00 \\ 0.00 & 0.16 \end{pmatrix}^{1/2}$$

Ejemplo anterior



$$\mathbf{Y}_r \approx \begin{pmatrix} -0.15 & -0.09 \\ -0.11 & 0.07 \\ 0.20 & -0.14 \\ 0.20 & -0.14 \\ -0.05 & 0.19 \\ -0.08 & -0.13 \\ -0.16 & -0.05 \\ 0.19 & 0.11 \\ -0.14 & -0.02 \\ 0.09 & 0.20 \end{pmatrix}$$

- Definimos en general grado de bondad de la aproximación al valor:

$$m = 100 \cdot \frac{\sum_{i=1}^r \lambda_i}{\sum_{i=1}^p |\lambda_i|} \%,$$

donde los λ_i representan los valores propios de la matriz de similitud **Q**

Ejemplo anterior

- El grado de bondad valdrá en nuestro caso:

$$\begin{aligned} m &= 100 \cdot \frac{0.22 + 0.16}{(0.22 + 0.16 + 0.05 + 0.04 + \dots + 0.00)} \% \\ &= 65.25 \% \end{aligned}$$

- Podemos concluir que las dos variables tenidas en cuenta explican el 65.25 % de la variabilidad entre las proteínas.

Ejemplo anterior

- En el gráfico siguiente podemos ver la representación de la coordenadas principales halladas anteriormente:



EscalM

Relación entre las coordenadas principales y las componentes principales

- Sea $\mathbf{X} = (x_{ij})_{i=1,\dots,n,j=1,\dots,p}$ una matriz de datos de n datos y p variables. Sea $\tilde{\mathbf{X}}$ la matriz centrada construida a partir de la matriz \mathbf{X} .
- Sea $\mathbf{D} = (d_{ij}^2)_{i,j=1,\dots,n}$ la matriz de distancias euclídeas al cuadrado construida a partir de la matriz \mathbf{X} :

$$d_{ij}^2 = \sum_{k=1}^p (x_{ik} - x_{jk})^2, \quad \forall i, j = 1 \dots, n.$$

- Entonces las coordenadas principales obtenidas de la matriz \mathbf{D} son equivalentes a las componentes principales de la matriz \mathbf{X} usando la matriz de covarianzas.

Relación entre las coordenadas principales y las componentes principales

- Recordemos que las componentes principales se calculaban como: $\mathbf{CP} = \tilde{\mathbf{X}}\mathbf{u}$, donde \mathbf{u} es la matriz de los vectores propios de la matriz de covarianzas (donde cada vector tiene módulo unidad) $\tilde{\mathbf{X}}^T \tilde{\mathbf{X}}$.
- Sea \mathbf{V} la matriz de vectores propios de la matriz $\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T$ correspondientes a valores propios no nulos (donde cada vector tiene módulo unidad). Sean $\lambda_1, \dots, \lambda_p$ los valores propios no nulos. Entonces las coordenadas principales \mathbf{Y} pueden calcularse como $\mathbf{Y} = \mathbf{V}\text{diag}(\sqrt{\lambda_1}, \dots, \sqrt{\lambda_p})$.

Relación entre las coordenadas principales y las componentes principales

- Puede demostrarse que si \mathbf{u} es un vector propio de valor propio λ de la matriz $\tilde{\mathbf{X}}^T \tilde{\mathbf{X}}$, entonces $\tilde{\mathbf{X}}\mathbf{u}$ es un vector propio de la matriz $\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T$ con el mismo valor propio.
- Concluimos que tanto las componentes principales como las coordenadas principales son vectores propios de la matriz $\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T$. Por tanto, aparte de un factor de escala, las componentes principales y las coordenadas principales son iguales.
- Tanto en la técnica de componentes principales como en la técnica de coordenadas principales, tratamos de reducir la dimensionalidad de los datos.
- Como hemos visto, si la matriz de similaridades \mathbf{Q} proviene de una métrica euclídea, ambos métodos son equivalentes. Sin embargo, la técnica de coordenadas principales puede aplicarse a un conjunto más general de problemas.

Parte XIV

Análisis de correspondencias

Introducción

- El análisis de correspondencias es una técnica descriptiva para representar y estudiar tablas de contingencia de variables cualitativas. Es decir, vamos a estudiar las frecuencias de aparición de dos o más variables cualitativas en un conjunto de elementos.
- Dicho análisis constituye la aplicación de técnicas de componentes principales y escalado multidimensional vistas anteriormente para variables cualitativas.
- Nuestra información de partida es una matriz \mathbf{X} de dimensiones $I \times J$ donde cada valor de la matriz representa la frecuencia absoluta observada de dos variables cualitativas en n elementos.
- Suponemos que la primera variable puede tomar I valores diferentes y que la segunda variable, J . Entonces, x_{ij} sería el número de elementos de entre los n que toma el valor i -ésimo para la primera variable y toma el valor j -ésimo para la segunda.

Ejemplo

- Consideremos el siguiente experimento: tenemos 61 ratas adultas y 61 crías de rata. Cada rata puede tener un genotipo diferente de entre cuatro: A, B, I y J. Ponemos cada cría aleatoriamente junto con una rata adulta para que crezca. Al cabo de 28 días, anotamos el porcentaje de peso adquirido por la cría de rata. Anotamos los resultados obtenidos en la tabla siguiente:

	1	2	3	4	5	6	7
A A	0	0	0	0	1	3	1
A B	0	1	0	1	1	0	0
A I	0	0	1	2	0	1	0
A J	1	1	1	1	1	0	0
B A	0	0	2	1	1	0	0
B B	0	0	1	0	0	4	0
B I	0	0	1	1	2	0	0
B J	1	0	0	1	0	0	0
I A	2	0	0	0	0	0	1
I B	0	0	0	0	1	0	2
I I	1	1	0	2	0	1	0
I J	0	1	1	1	0	0	0
J A	0	0	1	1	2	0	0
J B	0	0	0	2	1	0	0
J I	0	1	0	0	1	1	0
J J	0	2	0	3	0	0	0

Ejemplo

- La primera fila nos indica el porcentaje de peso adquirido por la cría de rata (1 indica que ha adquirido poco peso y 7 que ha adquirido mucho peso) y la primera columna nos indica el cruce que hemos hecho (por ejemplo B A indica que hemos puesto una cría de rata de genotipo B con una rata adulta de genotipo A).
- La tabla nos da el número de crías de rata que han aumentado un determinado porcentaje de peso en 28 días usando un cruce determinado.

Búsqueda de la mejor proyección

- Llamaremos $\mathbf{F} = (f_{ij})_{i=1,\dots,I,j=1,\dots,J}$ a la matriz de frecuencias relativas de la tabla de contingencia. Esto es, dividimos cada elemento de la matriz \mathbf{X} por el número total de elementos n : $f_{ij} = \frac{x_{ij}}{n}$.
- Los elementos de la matriz anterior verifican: $\sum_{i=1}^I \sum_{j=1}^J f_{ij} = 1$.
- Cualquier estudio aplicado a dicha matriz debe ser equivalente al estudio aplicado a su traspuesta ya que elegir la variable que va por filas o por columnas es una elección arbitraria y no debe influir en el análisis.

Ejemplo anterior

- En el ejemplo anterior la matriz **F** será:

	1	2	3	4	5	6	7
A A	0/61	0/61	0/61	0/61	1/61	3/61	1/61
A B	0/61	1/61	0/61	1/61	1/61	0/61	0/61
A I	0/61	0/61	1/61	2/61	0/61	1/61	0/61
A J	1/61	1/61	1/61	1/61	1/61	0/61	0/61
B A	0/61	0/61	2/61	1/61	1/61	0/61	0/61
B B	0/61	0/61	1/61	0/61	0/61	4/61	0/61
B I	0/61	0/61	1/61	1/61	2/61	0/61	0/61
B J	1/61	0/61	0/61	1/61	0/61	0/61	0/61
I A	2/61	0/61	0/61	0/61	0/61	0/61	1/61
I B	0/61	0/61	0/61	0/61	1/61	0/61	2/61
I I	1/61	1/61	0/61	2/61	0/61	1/61	0/61
I J	0/61	1/61	1/61	1/61	0/61	0/61	0/61
J A	0/61	0/61	1/61	1/61	2/61	0/61	0/61
J B	0/61	0/61	0/61	2/61	1/61	0/61	0/61
J I	0/61	1/61	0/61	0/61	1/61	1/61	0/61
J J	0/61	2/61	0/61	3/61	0/61	0/61	0/61

Ejemplo anterior

- O, si se quiere:

	1	2	3	4	5	6	7
A A	0.00	0.00	0.00	0.00	0.02	0.05	0.02
A B	0.00	0.02	0.00	0.02	0.02	0.00	0.00
A I	0.00	0.00	0.02	0.03	0.00	0.02	0.00
A J	0.02	0.02	0.02	0.02	0.02	0.00	0.00
B A	0.00	0.00	0.03	0.02	0.02	0.00	0.00
B B	0.00	0.00	0.02	0.00	0.00	0.07	0.00
B I	0.00	0.00	0.02	0.02	0.03	0.00	0.00
B J	0.02	0.00	0.00	0.02	0.00	0.00	0.00
I A	0.03	0.00	0.00	0.00	0.00	0.00	0.02
I B	0.00	0.00	0.00	0.00	0.02	0.00	0.03
I I	0.02	0.02	0.00	0.03	0.00	0.02	0.00
I J	0.00	0.02	0.02	0.02	0.00	0.00	0.00
J A	0.00	0.00	0.02	0.02	0.03	0.00	0.00
J B	0.00	0.00	0.00	0.03	0.02	0.00	0.00
J I	0.00	0.02	0.00	0.00	0.02	0.02	0.00
J J	0.00	0.03	0.00	0.05	0.00	0.00	0.00

Proyección de las filas

- Vamos a realizar un análisis de la matriz de frecuencias relativas \mathbf{F} por filas.
- Consideramos las I filas como I puntos en el espacio \mathbb{R}^J .
- El objetivo de nuestro análisis es buscar una representación de estos I puntos en un espacio de dimensión menor que nos permita apreciar sus distancias relativas.
- En nuestro análisis debemos tener en cuenta:
 - ▶ No todas las filas (puntos en \mathbb{R}^J) tienen el mismo peso ya que algunas filas contienen más datos que otras. Por tanto debemos dar más peso a aquellas filas que contengan más datos.
 - ▶ La distancia euclídea utilizada en el análisis multidimensional no es una buena medida en este caso para estudiar la proximidad entre las filas.

Proyección de las filas

- Definimos la frecuencia relativa de la fila i -ésima como: $f_{i\bullet} = \sum_{j=1}^J f_{ij}$.

Llamando \mathbf{f} al vector de frecuencias relativas de las filas $\mathbf{f} = (f_{i\bullet})_{i=1,\dots,I}$, podemos escribir matricialmente: $\mathbf{f} = \mathbf{F}\mathbf{1}$.

- Sea la matriz $\mathbf{D}_f = \text{diag}(f_{1\bullet}, \dots, f_{I\bullet})$.
- De la misma forma, definimos la frecuencia relativa de la columna j -ésima como: $f_{\bullet j} = \sum_{i=1}^I f_{ij}$. Llamando \mathbf{c} al vector de frecuencias relativas de las columnas $\mathbf{c} = (f_{\bullet j})_{j=1,\dots,J}$, podemos escribir matricialmente: $\mathbf{c} = \mathbf{F}^\top \mathbf{1}$.
- De la misma forma que antes, definimos la matriz $\mathbf{D}_c = \text{diag}(f_{\bullet 1}, \dots, f_{\bullet J})$.

Proyección de las filas

- Seguidamente definimos la matriz siguiente que nos permitirá realizar la proyección de las frecuencias por filas. Dicha matriz es

$$\mathbf{Z} = \mathbf{D}_f^{-1/2} \mathbf{F} \mathbf{D}_c^{-1/2} = \left(\frac{f_{ij}}{\sqrt{f_{i\bullet} \cdot f_{\bullet j}}} \right)_{i=1, \dots, I, j=1, \dots, J}.$$

- Para obtener la mejor representación bidimensional de las filas de la tabla de contingencia, hay que seguir los pasos siguientes:
 - ▶ Calcular la matriz $\mathbf{Z}^\top \mathbf{Z}$ y obtener sus vectores y valores propios.
 - ▶ Tomar los dos vectores propios, \mathbf{v}_1 y \mathbf{v}_2 ligados a los dos mayores valores propios menores que la unidad de esta matriz.
 - ▶ Calcular las proyecciones siguientes $\mathbf{D}_f^{-1} \mathbf{F} \mathbf{D}_c^{-1/2} \mathbf{v}_i$, $i = 1, 2$ y representarlas gráficamente en un espacio bidimensional.

Ejemplo anterior

- El valor del vector \mathbf{f} vale en nuestro ejemplo:

$$\mathbf{f} = \begin{pmatrix} AA & 0.08 \\ AB & 0.05 \\ AI & 0.07 \\ AJ & 0.08 \\ BA & 0.07 \\ BB & 0.08 \\ BI & 0.07 \\ BJ & 0.03 \\ IA & 0.05 \\ IB & 0.05 \\ II & 0.08 \\ IJ & 0.05 \\ JA & 0.07 \\ JB & 0.05 \\ JI & 0.05 \\ JJ & 0.08 \end{pmatrix}$$

Ejemplo anterior

- La matriz D_f será:

$$\begin{pmatrix} 0.08, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00 \\ 0.00, 0.05, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00 \\ 0.00, 0.00, 0.07, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00 \\ 0.00, 0.00, 0.00, 0.08, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00 \\ 0.00, 0.00, 0.00, 0.00, 0.07, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00 \\ 0.00, 0.00, 0.00, 0.00, 0.00, 0.08, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00 \\ 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.07, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00 \\ 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.03, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00 \\ 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.05, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00 \\ 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.05, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00 \\ 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.08, 0.00, 0.00, 0.00, 0.00, 0.00 \\ 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.05, 0.00, 0.00, 0.00, 0.00, 0.00 \\ 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.07, 0.00, 0.00, 0.00 \\ 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.05, 0.00, 0.00 \\ 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.05, 0.00 \\ 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.08 \end{pmatrix}$$

Ejemplo anterior

- El vector \mathbf{c} que nos da la suma de columnas vale en nuestro ejemplo:

$$\mathbf{c} = \begin{pmatrix} 0.08 \\ 0.11 \\ 0.13 \\ 0.26 \\ 0.18 \\ 0.16 \\ 0.07 \end{pmatrix}$$

- La matriz \mathbf{D}_c valdrá, por tanto:

$$\mathbf{D}_c = \begin{pmatrix} 0.08 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.11 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.13 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.26 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.18 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.16 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.07 \end{pmatrix}$$

Ejemplo anterior

- La matriz **Z** será:

$$\mathbf{Z} = \begin{pmatrix} 0.00 & 0.00 & 0.00 & 0.00 & 0.13 & 0.42 & 0.22 \\ 0.00 & 0.22 & 0.00 & 0.14 & 0.17 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.18 & 0.25 & 0.00 & 0.16 & 0.00 \\ 0.20 & 0.17 & 0.16 & 0.11 & 0.13 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.35 & 0.12 & 0.15 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.16 & 0.00 & 0.00 & 0.57 & 0.00 \\ 0.00 & 0.00 & 0.18 & 0.12 & 0.30 & 0.00 & 0.00 \\ 0.32 & 0.00 & 0.00 & 0.18 & 0.00 & 0.00 & 0.00 \\ 0.52 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.29 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.17 & 0.00 & 0.58 \\ 0.20 & 0.17 & 0.00 & 0.22 & 0.00 & 0.14 & 0.00 \\ 0.00 & 0.22 & 0.20 & 0.14 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.18 & 0.12 & 0.30 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.29 & 0.17 & 0.00 & 0.00 \\ 0.00 & 0.22 & 0.00 & 0.00 & 0.17 & 0.18 & 0.00 \\ 0.00 & 0.34 & 0.00 & 0.34 & 0.00 & 0.00 & 0.00 \end{pmatrix}$$

Ejemplo anterior

- La matriz $\mathbf{Z}^T \mathbf{Z}$ vale:

$$\mathbf{Z}^T \mathbf{Z} = \begin{pmatrix} 0.45 & 0.07 & 0.03 & 0.12 & 0.03 & 0.03 & 0.15 \\ 0.07 & 0.31 & 0.07 & 0.23 & 0.10 & 0.06 & 0.00 \\ 0.03 & 0.07 & 0.31 & 0.18 & 0.18 & 0.12 & 0.00 \\ 0.12 & 0.23 & 0.18 & 0.44 & 0.18 & 0.07 & 0.00 \\ 0.03 & 0.10 & 0.18 & 0.18 & 0.36 & 0.09 & 0.13 \\ 0.03 & 0.06 & 0.12 & 0.07 & 0.09 & 0.58 & 0.09 \\ 0.15 & 0.00 & 0.00 & 0.00 & 0.13 & 0.09 & 0.47 \end{pmatrix}$$

- Los valores propios de la matriz anterior son:

1.00, 0.57, 0.52, 0.37, 0.24, 0.12, 0.11.

Tenemos que considerar, por tanto los vectores propios asociados a los valores propios 0.57 y 0.52.

Ejemplo anterior

- Los vectores propios asociados a los valores propios anteriores son:
(por columnas)

$$\begin{pmatrix} -0.28 & 0.57 \\ 0.29 & 0.10 \\ 0.20 & -0.15 \\ 0.42 & 0.17 \\ 0.06 & -0.00 \\ -0.37 & -0.74 \\ -0.69 & 0.27 \end{pmatrix}$$

Ejemplo anterior

- Antes de hallar las proyecciones, calculamos la matriz $\mathbf{D}_f^{-1}\mathbf{F}\mathbf{D}_c^{-1/2}$:

$$\mathbf{D}_f^{-1}\mathbf{F}\mathbf{D}_c^{-1/2} = \begin{pmatrix} 0.00 & 0.00 & 0.00 & 0.00 & 0.47 & 1.48 & 0.78 \\ 0.00 & 0.98 & 0.00 & 0.65 & 0.78 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.69 & 0.98 & 0.00 & 0.62 & 0.00 \\ 0.70 & 0.59 & 0.55 & 0.39 & 0.47 & 0.00 & 0.00 \\ 0.00 & 0.00 & 1.38 & 0.49 & 0.59 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.55 & 0.00 & 0.00 & 1.98 & 0.00 \\ 0.00 & 0.00 & 0.69 & 0.49 & 1.18 & 0.00 & 0.00 \\ 1.75 & 0.00 & 0.00 & 0.98 & 0.00 & 0.00 & 0.00 \\ 2.33 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 1.30 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.78 & 0.00 & 2.60 \\ 0.70 & 0.59 & 0.00 & 0.78 & 0.00 & 0.49 & 0.00 \\ 0.00 & 0.98 & 0.92 & 0.65 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.69 & 0.49 & 1.18 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 1.30 & 0.78 & 0.00 & 0.00 \\ 0.00 & 0.98 & 0.00 & 0.00 & 0.78 & 0.82 & 0.00 \\ 0.00 & 1.18 & 0.00 & 1.17 & 0.00 & 0.00 & 0.00 \end{pmatrix}$$

Ejemplo anterior

- Para hallar las proyecciones, basta multiplicar la matriz anterior por la matriz de los dos vectores propios considerados anteriormente:

$$\begin{pmatrix} -1.07 & -0.89 \\ 0.60 & 0.21 \\ 0.32 & -0.39 \\ 0.28 & 0.44 \\ 0.52 & -0.12 \\ -0.63 & -1.54 \\ 0.41 & -0.02 \\ -0.09 & 1.16 \\ -1.56 & 1.67 \\ -1.75 & 0.69 \\ 0.11 & 0.23 \\ 0.75 & 0.07 \\ 0.41 & -0.02 \\ 0.59 & 0.22 \\ 0.02 & -0.51 \\ 0.83 & 0.32 \end{pmatrix}$$

Ejemplo anterior

- El gráfico bidimensional de las proyecciones es el siguiente donde puede observarse por ejemplo que las crías de rata con genotipo B criadas por ratas con genotipo B tienen un porcentaje de aumento de peso muy distinto al cabo de 28 días que crías de rata con genotipo I criadas por ratas con genotipo B:



Proyección de las columnas

- Vamos a realizar el mismo análisis de la matriz de frecuencias relativas \mathbf{F} que hemos hecho anteriormente pero ahora por columnas.
- Consideramos las J columnas como J puntos en el espacio \mathbb{R}^I .
- El objetivo de nuestro análisis es buscar una representación de estos J puntos en un espacio de dimensión menor que nos permita apreciar sus distancias relativas.
- Debemos tener en cuenta las mismas consideraciones que teníamos por filas.

Proyección de las columnas

- Para obtener la mejor representación bidimensional de las filas de la tabla de contingencia, hay que seguir los pasos siguientes:
 - ▶ Calcular la matriz $\mathbf{Z}\mathbf{Z}^\top$ y obtener sus vectores y valores propios. Los valores propios de la matriz anterior son los mismos que los valores propios de la matriz $\mathbf{Z}^\top\mathbf{Z}$ calculada anteriormente.
 - ▶ Tomar los dos vectores propios, \mathbf{w}_1 y \mathbf{w}_2 ligados a los dos mayores valores propios menores que la unidad de esta matriz.
 - ▶ Calcular las proyecciones siguientes $\mathbf{D}_c^{-1}\mathbf{F}^\top\mathbf{D}_f^{-1/2}\mathbf{w}_i$, $i = 1, 2$ y representarlas gráficamente en un espacio bidimensional.

Ejemplo anterior

- La matriz \mathbf{ZZ}^T valdrá:

$$\begin{pmatrix} 0.25, 0.02, 0.07, 0.02, 0.02, 0.24, 0.04, 0.00, 0.06, 0.15, 0.06, 0.00, 0.04, 0.02, 0.10, 0.00 \\ 0.02, 0.10, 0.04, 0.08, 0.04, 0.00, 0.07, 0.03, 0.00, 0.03, 0.07, 0.07, 0.07, 0.07, 0.08, 0.12 \\ 0.07, 0.04, 0.12, 0.06, 0.09, 0.12, 0.06, 0.04, 0.00, 0.00, 0.08, 0.07, 0.06, 0.07, 0.03, 0.08 \\ 0.02, 0.08, 0.06, 0.12, 0.09, 0.03, 0.08, 0.08, 0.10, 0.02, 0.09, 0.09, 0.08, 0.06, 0.06, 0.09 \\ 0.02, 0.04, 0.09, 0.09, 0.16, 0.06, 0.12, 0.02, 0.00, 0.03, 0.03, 0.09, 0.12, 0.06, 0.03, 0.04 \\ 0.24, 0.00, 0.12, 0.03, 0.06, 0.35, 0.03, 0.00, 0.00, 0.00, 0.08, 0.03, 0.03, 0.00, 0.10, 0.00 \\ 0.04, 0.07, 0.06, 0.08, 0.12, 0.03, 0.14, 0.02, 0.00, 0.05, 0.03, 0.05, 0.14, 0.09, 0.05, 0.04 \\ 0.00, 0.03, 0.04, 0.08, 0.02, 0.00, 0.02, 0.13, 0.16, 0.00, 0.10, 0.03, 0.02, 0.05, 0.00, 0.06 \\ 0.06, 0.00, 0.00, 0.10, 0.00, 0.00, 0.00, 0.16, 0.35, 0.17, 0.10, 0.00, 0.00, 0.00, 0.00, 0.00 \\ 0.15, 0.03, 0.00, 0.02, 0.03, 0.00, 0.05, 0.00, 0.17, 0.36, 0.00, 0.00, 0.05, 0.03, 0.03, 0.00 \\ 0.06, 0.07, 0.08, 0.09, 0.03, 0.08, 0.03, 0.10, 0.10, 0.00, 0.14, 0.07, 0.03, 0.06, 0.06, 0.13 \\ 0.00, 0.07, 0.07, 0.09, 0.09, 0.03, 0.05, 0.03, 0.00, 0.00, 0.07, 0.11, 0.05, 0.04, 0.05, 0.12 \\ 0.04, 0.07, 0.06, 0.08, 0.12, 0.03, 0.14, 0.02, 0.00, 0.05, 0.03, 0.05, 0.14, 0.09, 0.05, 0.04 \\ 0.02, 0.07, 0.06, 0.06, 0.00, 0.09, 0.05, 0.00, 0.03, 0.06, 0.04, 0.09, 0.11, 0.03, 0.10 \\ 0.10, 0.08, 0.03, 0.06, 0.03, 0.10, 0.05, 0.00, 0.00, 0.03, 0.06, 0.05, 0.05, 0.03, 0.11, 0.07 \\ 0.00, 0.12, 0.08, 0.09, 0.04, 0.00, 0.04, 0.06, 0.00, 0.00, 0.13, 0.12, 0.04, 0.10, 0.07, 0.23 \end{pmatrix}$$

- Los valores propios de la matriz anterior son:

$$1.00, 0.57, 0.52, 0.37, 0.24, 0.12, 0.11, 0.00, \\ 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00.$$

Obsérvese que los valores propios no nulos ya estaban calculados al hacer el análisis por filas.

Ejemplo anterior

- Los vectores propios correspondientes a los valores propios 0.57 y 0.52 son:

$$\mathbf{w} = \begin{pmatrix} 0.41 & 0.35 \\ -0.18 & -0.06 \\ -0.11 & 0.14 \\ -0.10 & -0.17 \\ -0.18 & 0.04 \\ 0.24 & 0.61 \\ -0.14 & 0.01 \\ 0.02 & -0.29 \\ 0.46 & -0.51 \\ 0.52 & -0.21 \\ -0.04 & -0.09 \\ -0.22 & -0.02 \\ -0.14 & 0.01 \\ -0.17 & -0.07 \\ -0.01 & 0.16 \\ -0.32 & -0.13 \end{pmatrix}$$

Ejemplo anterior

- La matriz $\mathbf{D}_c^{-1} \mathbf{F}^\top \mathbf{D}_f^{-1/2}$ vale en nuestro caso:

$$\begin{pmatrix} 0.00, 0.00, 0.00, 0.70, 0.00, 0.00, 0.00, 1.10, 1.80, 0.00, 0.70, 0.00, 0.00, 0.00, 0.00, 0.00 \\ 0.00, 0.64, 0.00, 0.50, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.50, 0.64, 0.00, 0.00, 0.64, 1.00 \\ 0.00, 0.00, 0.49, 0.44, 0.98, 0.44, 0.49, 0.00, 0.00, 0.00, 0.00, 0.00, 0.56, 0.49, 0.00, 0.00, 0.00 \\ 0.00, 0.28, 0.49, 0.22, 0.24, 0.00, 0.24, 0.35, 0.00, 0.00, 0.44, 0.28, 0.24, 0.56, 0.00, 0.65 \\ 0.32, 0.41, 0.00, 0.32, 0.36, 0.00, 0.71, 0.00, 0.00, 0.41, 0.00, 0.00, 0.71, 0.41, 0.41, 0.00 \\ 1.05, 0.00, 0.39, 0.00, 0.00, 1.40, 0.00, 0.00, 0.00, 0.00, 0.35, 0.00, 0.00, 0.00, 0.45, 0.00 \\ 0.87, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00, 1.13, 2.25, 0.00, 0.00, 0.00, 0.00, 0.00, 0.00 \end{pmatrix}$$

- Las proyecciones $\mathbf{D}_c^{-1} \mathbf{F}^\top \mathbf{D}_f^{-1/2} \mathbf{w}$ calculadas por columnas valen:

$$\begin{pmatrix} 0.75 & -1.43 \\ -0.65 & -0.21 \\ -0.43 & 0.29 \\ -0.61 & -0.24 \\ -0.10 & 0.01 \\ 0.70 & 1.31 \\ 2.03 & -0.75 \end{pmatrix}$$

Ejemplo anterior

- La representación gráfica de las proyecciones anteriores es:



AnCoi

Ejemplo anterior

- En el gráfico anterior se puede observar las similitudes o las diferencias del genotipo entre las crías de rata y las ratas adultas que las crían con diferente porcentaje de aumento de peso al cabo de 28 días.
- Por ejemplo, vemos que las crías de rata y ratas adultas con porcentaje de aumento de peso en segundo y en cuarto lugar tienen un genotipo muy parecido.
- En cambio, la mayor diferencia entre el genotipo de las crías y las ratas adultas se encuentra en las crías de rata con porcentaje de aumento de peso en primer y sexto lugar.

Análisis conjunto

- Debido a la simetría del problema, conviene representar conjuntamente las proyecciones de las filas y las columnas en el mismo gráfico.
- Antes de dar los pasos para hacer tal representación conviene tener en cuenta que
 - ▶ si \mathbf{v} es vector propio de la matriz $\mathbf{Z}^T \mathbf{Z}$ de valor propio λ , entonces \mathbf{Zv} es vector propio de la matriz \mathbf{ZZ}^T del mismo valor propio y
 - ▶ viceversa: si \mathbf{w} es vector propio de la matriz \mathbf{ZZ}^T de valor propio λ , entonces $\mathbf{Z}^T \mathbf{w}$ es vector propio de la matriz $\mathbf{Z}^T \mathbf{Z}$ del mismo valor propio.

Análisis conjunto

- Para hacer la representación conjunta de las proyecciones de las filas y las columnas hay que realizar los pasos siguientes:
 - ▶ Se calcula la matriz de frecuencias relativa \mathbf{F} .
 - ▶ Se calcula la matriz estandarizada \mathbf{Z} .
 - ▶ Se busca de las dos matrices siguientes, $\mathbf{Z}^\top \mathbf{Z}$ o $\mathbf{Z}\mathbf{Z}^\top$ la que tenga menor dimensión. Supongamos para fijar ideas que es la matriz $\mathbf{Z}^\top \mathbf{Z}$. Se calculan los dos valores propios menores que 1 más grandes de la matriz anterior. Sean \mathbf{v}_1 y \mathbf{v}_2 los dos vectores propios asociados a los dos valores propios anteriores. La proyección de las filas vendrá dada por $\mathbf{D}_f^{-1/2} \mathbf{Z} \mathbf{v}_i$, $i = 1, 2$.
 - ▶ Sean $\mathbf{w}_i = \mathbf{Z} \mathbf{v}_i$, $i = 1, 2$ los vectores propios de la matriz $\mathbf{Z}\mathbf{Z}^\top$ asociados a los valores propios anteriores. La proyección de las columnas vendrá dada por: $\mathbf{D}_c^{-1/2} \mathbf{Z}^\top \mathbf{w}_i$, $i = 1, 2$.

Ejemplo anterior

- En el ejemplo anterior la matriz de menor dimensión era $\mathbf{Z}^T \mathbf{Z}$ (7×7).
- Los valores propios a considerar eran: 0.57 y 0.52.
- Los vectores propios eran:

$$\begin{pmatrix} -0.28 & 0.57 \\ 0.29 & 0.10 \\ 0.20 & -0.15 \\ 0.42 & 0.17 \\ 0.06 & -0.00 \\ -0.37 & -0.74 \\ -0.69 & 0.27 \end{pmatrix}$$

Ejemplo anterior

- La proyección de las filas era:

$$\mathbf{v} = \begin{pmatrix} -1.07 & -0.89 \\ 0.60 & 0.21 \\ 0.32 & -0.39 \\ 0.28 & 0.44 \\ 0.52 & -0.12 \\ -0.63 & -1.54 \\ 0.41 & -0.02 \\ -0.09 & 1.16 \\ -1.56 & 1.67 \\ -1.75 & 0.69 \\ 0.11 & 0.23 \\ 0.75 & 0.07 \\ 0.41 & -0.02 \\ 0.59 & 0.22 \\ 0.02 & -0.51 \\ 0.83 & 0.32 \end{pmatrix}$$

Ejemplo anterior

- Busquemos ahora los vectores propios de la matriz \mathbf{ZZ}^\top : $\mathbf{w}_i = \mathbf{Z}\mathbf{v}_i$, donde \mathbf{v}_i son los vectores propios hallados anteriormente por columnas:

$$\mathbf{w} = \begin{pmatrix} -0.31 & -0.25 \\ 0.13 & 0.05 \\ 0.08 & -0.10 \\ 0.08 & 0.13 \\ 0.13 & -0.03 \\ -0.18 & -0.44 \\ 0.10 & -0.01 \\ -0.02 & 0.21 \\ -0.35 & 0.37 \\ -0.39 & 0.15 \\ 0.03 & 0.06 \\ 0.17 & 0.02 \\ 0.10 & -0.01 \\ 0.13 & 0.05 \\ 0.00 & -0.11 \\ 0.24 & 0.09 \end{pmatrix}$$

Ejemplo anterior

- Las proyecciones de las columnas será la matriz $\mathbf{D}_c^{-1/2} \mathbf{Z}^\top \mathbf{w}$:

$$\begin{pmatrix} -0.56 & 1.03 \\ 0.49 & 0.15 \\ 0.32 & -0.21 \\ 0.46 & 0.17 \\ 0.07 & -0.00 \\ -0.52 & -0.95 \\ -1.53 & 0.54 \end{pmatrix}$$

Ejemplo anterior

- En el gráfico siguiente se puede ver la proyección conjunta donde en rojo están las proyecciones por filas y en negro, las proyecciones por columnas:



AnCorrC

Parte XV

Métodos de clasificación automática.

Descripción del problema

- Tenemos un conjunto de individuos con unas ciertas medidas multidimensionales.
- Queremos ver si existe una "forma natural" de clasificar a dichos individuos en grupos.
- Los grupos que formemos tienen que ser lo más homogéneos posible y las diferencias entre los grupos, lo más acentuadas posible.

Introducción al algoritmo de las k -medias

- En todo método de partición, partimos de n individuos de los que se han tomado p mediciones. O sea, tenemos una matriz $n \times p$ de n

individuos con p variables: $\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}$

- El objetivo del algoritmo de las k -medias es dividir los n individuos en un número de conjuntos o grupos prefijado G .

Etapas del algoritmo de las k -medias

- Se seleccionan G puntos como centros de los puntos iniciales. Hay distintas formas de realizar dicha selección:
 - ▶ asignando aleatoriamente los objetos a los grupos y tomando los centros de los grupos formados,
 - ▶ tomando como centros los G más alejados entre sí,
 - ▶ seleccionando los centros “a priori”.
- Calcular las distancias euclídeas de cada elemento a los centros de los G grupos, y asignar cada elemento al grupo de cuyo centro esté más próximo. La asignación se realiza secuencialmente y al introducir un nuevo elemento en un grupo se recalculan las coordenadas del nuevo centro del grupo.
- Definir un criterio de optimalidad y comprobar si reasignando alguno de los elementos mejora el criterio.
- Si no es posible mejorar el criterio de optimalidad, terminar el proceso.

Criterio de optimalidad.

- Sea x_{ijg} la medida j -ésima del individuo i -ésimo que pertenece al grupo g . Definimos la suma de cuadrados dentro de los grupos $SCDG$ como:

$$SCDG = \sum_{g=1}^G \sum_{j=1}^p \sum_{i=1}^{n_g} (x_{ijg} - \bar{x}_{jg})^2,$$

donde \bar{x}_{jg} es la media de esta variable en el grupo g y n_g representa el número de elementos de grupo g .

- Nuestro objetivo será encontrar aquella partición que minimice $SCDG$.
- La suma de cuadrados $SCDG$ puede escribirse como:

$$SCDG = \sum_{g=1}^G \sum_{j=1}^p n_g s_{jg}^2,$$

donde s_{jg}^2 es la varianza de la variable j en el grupo g .

Criterio de optimalidad.

- Escribamos matricialmente el criterio de optimalidad:

$$SCDG = \sum_{g=1}^G \sum_{i=1}^{n_g} (\mathbf{x}_{ig} - \bar{\mathbf{x}}_g)^\top (\mathbf{x}_{ig} - \bar{\mathbf{x}}_g),$$

donde \mathbf{x}_{ig} sería el vector de los j valores del individuo i -ésimo dentro del grupo g y $\bar{\mathbf{x}}_g$ sería el vector de medias de los g grupos.

- El valor anterior de dentro del sumatorio es la distancia euclídea entre el individuo i -ésimo dentro del grupo g y la media dentro del grupo g . Si nombramos $d^2(i, g)$ a dicha distancia, podemos escribir:

$$SCDG = \sum_{g=1}^G \sum_{i=1}^{n_g} d^2(i, g).$$

Criterio de optimalidad.

- Usando que $(\mathbf{x}_{ig} - \bar{\mathbf{x}}_g)^\top (\mathbf{x}_{ig} - \bar{\mathbf{x}}_g) = \text{tr}((\mathbf{x}_{ig} - \bar{\mathbf{x}}_g)(\mathbf{x}_{ig} - \bar{\mathbf{x}}_g))^\top$ (ejercicio), podemos escribir la suma de cuadrados como:

$$SCDG = \text{tr} \left(\sum_{g=1}^G \sum_{i=1}^{n_g} (\mathbf{x}_{ig} - \bar{\mathbf{x}}_g)(\mathbf{x}_{ig} - \bar{\mathbf{x}}_g)^\top \right),$$

donde tr significa traza.

- Nombrando \mathbf{W} a la matriz $\sum_{g=1}^G \sum_{i=1}^{n_g} (\mathbf{x}_{ig} - \bar{\mathbf{x}}_g)(\mathbf{x}_{ig} - \bar{\mathbf{x}}_g)^\top$, tenemos que el criterio de optimalidad equivale a minimizar la traza de la matriz \mathbf{W} .

Ejemplo

- La siguiente tabla muestra el gasto medio en varios tipos de comida para distintos tipos de familias en Francia: trabajadores manuales (TM), empleados (EM) y directivos (DIR) con distintos número de hijos: 2, 3, 4 o 5 hijos.

Significado de las siglas:

TF Tipo familia

P Pan

VE Vegetales

F Fruta

C Carne

A Aves

L Leche

VI Vino

Ejemplo

TF	P	VE	F	C	A	L	VI
TM2	332	428	354	1437	526	247	427
EM2	293	559	388	1527	567	239	258
DIR2	372	767	562	1948	927	235	433
TM3	406	563	341	1507	544	324	407
EM3	386	608	396	1501	558	319	363
• DIR3	438	843	689	2345	1148	243	341
TM4	534	660	367	1620	638	414	407
EM4	460	699	484	1856	762	400	416
DIR4	385	789	621	2366	1149	304	282
TM5	655	776	423	1848	759	495	486
EM5	584	995	548	2056	893	518	319
DIR5	515	1097	887	2630	1167	561	284

Ejemplo

- Elegiremos $G = 3$. Elegiremos como centros de los valores iniciales las medias de los 4 primeros individuos, luego las medias de los individuos 5 al 8 y por último las medias de los individuos 9 al 12. O sea, elegimos como grupos iniciales los siguientes: Grupo 1: individuos 1, 2, 3 y 4. Grupo 2: individuos 5, 6, 7 y 8 y Grupo 3: individuos 9, 10, 11 y 12. Las coordenadas de los 3 centros iniciales son las siguientes:

$$\begin{pmatrix} 350.75 & 579.25 & 411.25 & 1604.75 & 641.0 & 261.25 & 381.25 \\ 454.50 & 702.50 & 484.00 & 1830.50 & 776.5 & 344.00 & 381.75 \\ 534.75 & 914.25 & 619.75 & 2225.00 & 992.0 & 469.50 & 342.75 \end{pmatrix}$$

Ejemplo

- En el gráfico siguiente observamos la proyección de las dos primeras variables de los individuos junto con los centros de los grupos iniciales (en rojo):



kmeans

Ejemplo

- Vamos a aplicar el algoritmo de k-means a nuestro ejemplo. El valor inicial de *SCDG* es 1916875. Calculemos primero el grupo más cercano de cada individuo hallando la distancia entre éste y el centro correspondiente:

- ▶ Individuo 1: distancias de (332, 428, 354, 1437, 526, 247, 427) a cada uno de los centros:

$$\begin{aligned}d((332, 428, \dots, 427), (350.75, 579.25, \dots, 381.25)) &= 264.895, \\d((332, 428, \dots, 427), (454.50, 702.50, \dots, 381.75)) &= 579.945, \\d((332, 428, \dots, 427), (534.75, 914.25, \dots, 342.75)) &= 1114.883.\end{aligned}$$

Por tanto, el grupo correspondiente al primer individuo es el primero. No hay ningún cambio.

- ▶ Con el individuo 2, tampoco hay cambios. En cambio con el individuo 3, resulta que el grupo más cercano es el grupo 2. Por tanto, cambiamos los grupos por:

$$G1 = \{1, 2, 4\}, \quad G2 = \{3, 5, 6, 7, 8\}, \quad G3 = \{9, 10, 11, 12\}.$$

Ejemplo

- El nuevo valor de $SCDG$ es 1622739. Ha habido por tanto una mejora. Los nuevos centros son:

343.667	516.667	361.000	1490.333	545.667	270.000	364.000
438.000	715.400	499.600	1854.000	806.600	322.200	392.000
534.750	914.250	619.750	2225.000	992.000	469.500	342.750

- Volvemos a mirar el grupo más cercano de cada individuo: individuo 1: grupo 1 (no hay cambios); individuo 2: grupo 1 (no hay cambios); individuo 3: grupo 2 (no hay cambios); individuo 4: grupo 1 (no hay cambios); individuo 5: grupo 1. Por tanto, cambiamos los grupos por:

$$G1 = \{1, 2, 4, 5\}, \quad G2 = \{3, 6, 7, 8\}, \quad G3 = \{9, 10, 11, 12\}.$$

El nuevo valor de $SCDG$ es 1367967. Se sigue mejorando. Los nuevos centros son:

354.250	539.500	369.750	1493.000	548.750	282.250	363.750
451.000	742.250	525.500	1942.250	868.750	323.000	399.250
534.75	914.250	619.750	2225.000	992.000	469.500	342.750

Ejemplo

- La tabla siguiente va mostrando los pasos del algoritmo donde la primera columna nos indica el individuo que va cambiando de grupo:

	Grupo 1	Grupo 2	Grupo 3	<i>SCDG</i>
6	{1, 2, 4, 5}	{3, 7, 8}	{6, 9, 10, 11, 12}	1072675
10	{1, 2, 4, 5}	{3, 7, 8, 10}	{6, 9, 11, 12}	709037.8
11	{1, 2, 4, 5}	{3, 7, 8, 10, 11}	{6, 9, 12}	632336.1
7	{1, 2, 4, 5, 7}	{3, 8, 10, 11}	{6, 9, 12}	564800.9

Propiedades del algoritmo de k -means

- El algoritmo no es invariante ante cambios de escala. Por tanto, si las variables van en unidades distintas, conviene estandarizarlas antes de aplicar el algoritmo.
- El algoritmo produce grupos aproximadamente esféricos ya que tiende a minimizar las distancias euclídeas entre los puntos y sus medias de grupo.

Elección del número de grupos

- El procedimiento habitual de elegir cuántos grupos hay que usar en el algoritmo de k -means es el llamado test F de reducción de variabilidad.
- Dicho test consiste en calcular:

$$F = \frac{SCDG(G) - SCDG(G + 1)}{SCDG(G + 1)/(n - G - 1)},$$

donde se compara la variabilidad al aumentar un grupo con la varianza promedio. El valor obtenido se compara con el valor crítico de una distribución F de p y $p(n - G - 1)$ grados de libertad. Hay que decir que dicha regla no está muy justificada ya que los datos no tienen porqué verificar las hipótesis necesarias para aplicar la distribución F .

Ejemplo

- En la tabla siguiente observamos los valores de $SCDG$ para distintas elecciones de G :

G	$SCDG$	F
3	564800.9	
4	363066.1	$\frac{(564800.9 - 363066.1)}{363066.1 / (12 - 3 - 1)} = 4.44$
5	255134.8	$\frac{(363066.1 - 255134.8)}{255134.8 / (12 - 4 - 1)} = 2.96$
6	153977.9	$\frac{(255134.8 - 153977.9)}{153977.9 / (12 - 5 - 1)} = 3.94$
7	109217.7	$\frac{(153977.9 - 109217.7)}{109217.7 / (12 - 6 - 1)} = 2.05$

- La reducción de la variabilidad más acentuada se produce al pasar de 3 a 4 grupos. Si hallamos el valor crítico (para $\alpha = 0.05$) de la distribución F con $p = 7$ y $p(n - G - 1) = 7 \cdot (12 - 3 - 1) = 56$ grados de libertad, obtenemos $F_{0.95, 7, 56} = 2.18$. Como $4.44 > 2.18$, concluimos que la variabilidad se ha reducido y, según este criterio, escogeríamos $G = 4$ como el número de grupos óptimo.

Introducción

- Los métodos jerárquicos parten de una matriz de distancias o similaridades. O sea, suponemos que tenemos una matriz D $n \times n$ que nos dice la distancia entre los individuos o elementos de la muestra:
 d_{ij} : distancia entre el individuo i y el individuo j .
- A partir de la matriz de distancias, se define el algoritmo de construcción de grupos. O sea, usando la matriz de distancias definida anteriormente, indicar cómo se formarán los distintos grupos.

Cómo calcular la matriz de distancias

- Recordemos que tenemos n individuos de los que se han tomado p

mediciones: $\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}$

- A partir de la matriz \mathbf{X} hemos de construir una matriz de similaridad-disimilaridad \mathbf{D} $n \times n$ entre los objetos:

$$\mathbf{D} = \begin{pmatrix} d_{11} & d_{12} & \dots & d_{1n} \\ d_{21} & d_{22} & \dots & d_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ d_{n1} & d_{n2} & \dots & d_{nn} \end{pmatrix}$$

- En el caso en que \mathbf{D} sea una matriz de similaridad, cuánto menor sean los valores d_{ij} , más alejados estarán los individuos i y j y en el segundo caso, más cercanos estarán.

Introducción

- Supongamos que la matriz \mathbf{X} sólo coge los valores 0 o 1. En este caso, diremos que la estructura de los datos es binaria.
- De cara a definir la matriz de similaridad, consideremos los datos referentes al individuo i (columna i de la matriz \mathbf{X} : $(x_{i1}, \dots, x_{ip})^\top$) y al individuo j (columna j de la matriz \mathbf{X} : $(x_{j1}, \dots, x_{jp})^\top$)
- Definimos las cantidades siguientes:

$$\begin{aligned}a_1 &= \sum_{k=1}^p \mathcal{I}(x_{ik} = x_{jk} = 1), \\a_2 &= \sum_{k=1}^p \mathcal{I}(x_{ik} = 0, x_{jk} = 1), \\a_3 &= \sum_{k=1}^p \mathcal{I}(x_{ik} = 1, x_{jk} = 0), \\a_4 &= \sum_{k=1}^p \mathcal{I}(x_{ik} = x_{jk} = 0),\end{aligned}$$

donde $\mathcal{I}(\text{condición})$ cuenta cuantos elementos de la matriz \mathbf{X} cumplen la condición.

Definición de la matriz de distancias

- Definimos la similaridad entre los individuos i y j como:

$$d_{ij} = \frac{a_1 + \delta a_4}{a_1 + \delta a_4 + \lambda(a_2 + a_3)}.$$

- Los parámetros δ y λ son factores de peso. La tabla siguiente muestra ejemplos de similaridades más comunes:

Nombre	δ	λ	Definición
Jaccard	0	1	$\frac{a_1}{a_1 + a_2 + a_3}$
Tanimoto	1	2	$\frac{a_1 + a_4}{a_1 + 2(a_2 + a_3) + a_4}$
Simple	1	1	$\frac{a_1 + a_4}{p}$
Rusell y Rao	-	-	$\frac{a_1}{p}$
Dice	0	0.5	$\frac{2a_1}{2a_1 + a_2 + a_3}$
Kulczynski	-	-	$\frac{a_1}{a_2 + a_3}$

¿Por qué no se usa la distancia euclídea?

- La distancia euclídea no sería adecuada en este caso ya que trata los valores 0 y 1 de la misma forma y en la mayoría de aplicaciones donde se tratan datos binarios, deben ser tratados de formas distintas.
- Por ejemplo, si $x_{ik} = 1$ significa que el individuo i tiene un cierto conocimiento sobre el lenguaje k y si $x_{ik} = 0$ significa que no lo tiene, el valor $x_{ik} = 0$ tiene que tratarse de forma distinta del valor $x_{ik} = 1$.

Ejemplo

- Consideremos 3 tipos de marca de automóviles, Renault, Rover y Toyota. Se ha realizado una encuesta en 40 personas para que opinen de las variables siguientes: Economía (EC), servicio (SER), valor no depreciado (V), precio (P)(valor 1 para los coches más baratos), diseño (DI), deportividad (DE), seguridad (SEG) y facilidad de Manejo (FM). Los resultados van de 1 (muy bueno) a 6 (muy malo).
- Las medias de las valoraciones de los 40 encuestados son:

Modelo	EC	SER	V	P	DI	DE	SEG	FM
Renault	2.7	3.3	3.4	3.0	3.1	3.4	3.0	2.7
Rover	3.9	2.8	2.6	4.0	2.6	3.0	3.2	3.0
Toyota	2.5	2.9	3.4	3.0	3.2	3.1	3.2	2.8

Ejemplo

- Sea \mathbf{X} la matriz de datos anterior:

$$\mathbf{X} = \begin{pmatrix} 2.7 & 3.3 & 3.4 & 3.0 & 3.1 & 3.4 & 3.0 & 2.7 \\ 3.9 & 2.8 & 2.6 & 4.0 & 2.6 & 3.0 & 3.2 & 3.0 \\ 2.5 & 2.9 & 3.4 & 3.0 & 3.2 & 3.1 & 3.2 & 2.8 \end{pmatrix}$$

Se trata de una matriz 3×8 de 3 individuos con 8 variables.

- A partir de la matriz anterior, construimos la matriz binaria siguiente:

$$y_{ik} = \begin{cases} 1, & \text{si } x_{ik} > \bar{x}_k, \\ 0, & \text{en caso contrario.} \end{cases}$$

- La matriz \mathbf{Y} es:

$$\mathbf{Y} = \begin{pmatrix} 0 & 1 & 1 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \end{pmatrix}$$

Ejemplo

- Si aplicamos la medida de Jaccard, obtenemos la matriz siguiente de similitud:

$$\mathbf{D} = \begin{pmatrix} 1.000 & 0.000 & 0.400 \\ & 1.000 & 0.167 \\ & & 1.000 \end{pmatrix}$$

- La medida de Tanimoto da el siguiente resultado:

$$\mathbf{D} = \begin{pmatrix} 1.000 & 0.000 & 0.455 \\ & 1.000 & 0.231 \\ & & 1.000 \end{pmatrix}$$

- La medida simple da como matriz de similitud:

$$\mathbf{D} = \begin{pmatrix} 1.000 & 0.000 & 0.625 \\ & 1.000 & 0.375 \\ & & 1.000 \end{pmatrix}$$

Introducción

- En el caso general de datos continuos, se usan las distancias generadas por las normas L_r , $r \geq 1$. Esto es, la distancia entre el individuo i (fila i de la matriz \mathbf{X} : $(x_{i1}, \dots, x_{ip})^\top$) y al individuo j (fila j de la matriz \mathbf{X} : $(x_{j1}, \dots, x_{jp})^\top$) viene dada por:

$$d_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|_r = \left(\sum_{k=1}^p |x_{ik} - x_{jk}|^r \right)^{\frac{1}{r}}.$$

- Es claro que en este caso $d_{ii} = 0$. Por tanto, usamos una matriz de distancias o disimilaridad. Si $r = 2$, tenemos la distancia euclídea usual.

Escalado de los datos

- Una suposición usual es que los datos estén en la misma escala. Si no fuese el caso, la definición de la matriz de distancias se realiza mediante una métrica \mathbf{A} . En el caso de la distancia euclídea, la modificación sería:

$$d_{ij}^2 = \|\mathbf{x}_i - \mathbf{x}_j\|_{\mathbf{A}} = (\mathbf{x}_i - \mathbf{x}_j)^{\top} \mathbf{A} (\mathbf{x}_i - \mathbf{x}_j).$$

- La métrica \mathbf{A} más usada es la siguiente:

$$\mathbf{A} = \text{diag} \left(s_{\mathbf{x}_1\mathbf{x}_1}^{-1}, \dots, s_{\mathbf{x}_p\mathbf{x}_p}^{-1} \right),$$

donde $s_{\mathbf{x}_k\mathbf{x}_k}$ es la varianza de la variable k -ésima.

- La distancia euclídea quedará de la forma siguiente:

$$d_{ij}^2 = \sum_{k=1}^p \frac{(x_{ik} - x_{jk})^2}{s_{\mathbf{x}_k\mathbf{x}_k}}.$$

Ejemplo

- Recordemos el ejemplo del gasto medio en varios tipos de comida para distintos tipos de familias en Francia: trabajadores manuales (TM), empleados (EM) y directivos (DIR) con distintos número de hijos: 2, 3, 4 o 5 hijos.

Significado de las siglas:

TF Tipo familia

P Pan

VE Vegetales

F Fruta

C Carne

A Aves

L Leche

VI Vino

Ejemplo

TF	P	VE	F	C	A	L	VI
TM2	332	428	354	1437	526	247	427
EM2	293	559	388	1527	567	239	258
DIR2	372	767	562	1948	927	235	433
TM3	406	563	341	1507	544	324	407
EM3	386	608	396	1501	558	319	363
• DIR3	438	843	689	2345	1148	243	341
TM4	534	660	367	1620	638	414	407
EM4	460	699	484	1856	762	400	416
DIR4	385	789	621	2366	1149	304	282
TM5	655	776	423	1848	759	495	486
EM5	584	995	548	2056	893	518	319
DIR5	515	1097	887	2630	1167	561	284

Ejemplo

- La matriz de distancia euclídea para los datos anteriores es:

0.00												
241.34	0.00											
762.82	645.96	0.00										
188.21	212.95	664.36	0.00									
226.89	171.16	633.21	87.42	0.00								
1230.63	1098.12	491.16	1130.71	1100.22	0.00							
411.24	367.75	547.30	237.01	238.69	982.71	0.00						
601.26	503.84	285.75	465.88	445.55	694.00	303.37	0.00					
1216.50	1078.47	505.67	1117.97	1089.81	134.14	974.02	685.23	0.00				
719.99	660.90	456.21	558.64	555.18	777.94	332.66	248.34	781.53	0.00			
1012.71	876.39	450.59	854.25	820.89	537.55	648.97	433.11	544.21	397.83	0.00		
1648.73	1506.22	941.37	1521.75	1485.66	543.70	1341.05	1063.15	564.44	1077.54	733.29	0.00	

Ejemplo

- Vamos a escalar los datos. Calculemos primero

$$\mathbf{A} = \text{diag}(s_{x_1 x_1}^{-1}, \dots, s_{x_7 x_7}^{-1}):$$

$$\mathbf{A} = \text{diag}(9.50 \cdot 10^{-5}, 3.05 \cdot 10^{-5}, 4.00 \cdot 10^{-5}, \\ 6.97 \cdot 10^{-6}, 1.75 \cdot 10^{-5}, 7.95 \cdot 10^{-5}, \\ 2.12 \cdot 10^{-4})$$

- La distancia euclídea escalada será:

0.00												
2.62	0.00											
3.16	3.62	0.00										
1.30	2.57	2.83	0.00									
1.63	1.94	2.70	0.80	0.00								
4.99	4.49	2.23	4.48	4.12	0.00							
2.88	3.62	3.04	1.66	1.88	4.18	0.00						
2.93	3.52	1.97	1.95	1.95	3.13	1.34	0.00					
4.96	3.99	2.73	4.43	3.97	1.26	4.23	3.24	0.00				
4.64	5.61	3.86	3.58	3.87	4.62	2.10	2.39	4.99	0.00			
5.54	5.07	3.88	4.39	4.11	3.37	3.15	2.82	3.32	3.02	0.00		
7.58	6.83	5.19	6.71	6.31	3.66	5.80	4.94	3.62	5.46	3.06	0.00	

Caso en que la matriz \mathbf{X} es una tabla de contingencia

- \mathbf{X} es una tabla de contingencia cuando representa una matriz de frecuencias.
- Definimos $x_{i\bullet} = \sum_{j=1}^p x_{ij}$ y $\frac{x_{ij}}{x_{i\bullet}}$ la distribución condicional de la fila i -ésima. De la misma forma, definimos $x_{\bullet j} = \sum_{i=1}^n x_{ij}$ y $\frac{x_{ij}}{x_{\bullet j}}$ la distribución condicional de la columna j -ésima. Sea $x_{\bullet\bullet} = \sum_{i=1}^n x_{i\bullet} = \sum_{j=1}^p x_{\bullet j}$.
- En este caso la distancia entre las filas i_1 e i_2 es la siguiente:

$$d^2(i_1, i_2) = \sum_{j=1}^p \frac{1}{\left(\frac{x_{\bullet j}}{x_{\bullet\bullet}}\right)} \left(\frac{x_{i_1 j}}{x_{i_1 \bullet}} - \frac{x_{i_2 j}}{x_{i_2 \bullet}} \right)^2.$$

Introducción

- Existen dos métodos jerárquicos de “clustering”: algoritmos aglomerativos y algoritmos de división.
 - ▶ Los algoritmos aglomerativos empiezan con la partición más fina posible (cada individuo constituye un clúster) y los va agrupando.
 - ▶ Los algoritmos de división empiezan con la partición más burda posible (todos los individuos constituyen el clúster) y va dividiendo los clústers en clústers más pequeños.
 - ▶ Los algoritmos aglomerativos requieren menos tiempo de cálculo y son los más utilizados.

Pasos a realizar en los algoritmos aglomerativos

- Construir la partición más fina posible.
- Calcular la matriz de distancias D .
- Realizar los pasos siguientes hasta que todos los individuos estén en un mismo clúster:
 - ▶ Encontrar los dos clústers con la distancia más próxima.
 - ▶ Definir un nuevo clúster compuesto por los dos clústers anteriores encontrados.
 - ▶ Calcular la nueva matriz de distancias reducida D entre los nuevos clústers.

Pasos a realizar en los algoritmos aglomerativos

- Si dos grupos, P y Q tienen que agregarse en un nuevo clúster, el cálculo de la distancia entre el nuevo clúster $P + Q$ y otro clúster cualquiera R se realiza de la forma siguiente:

$$d(R, P + Q) = \delta_1 d(R, P) + \delta_2 d(R, Q) + \delta_3 d(P, Q) + \delta_4 |d(R, P) - d(R, Q)|,$$

donde los δ_j son parámetros a escoger. Cada elección de dichos parámetros da lugar a un algoritmo aglomerativo distinto.

Tabla de algoritmos aglomerativos

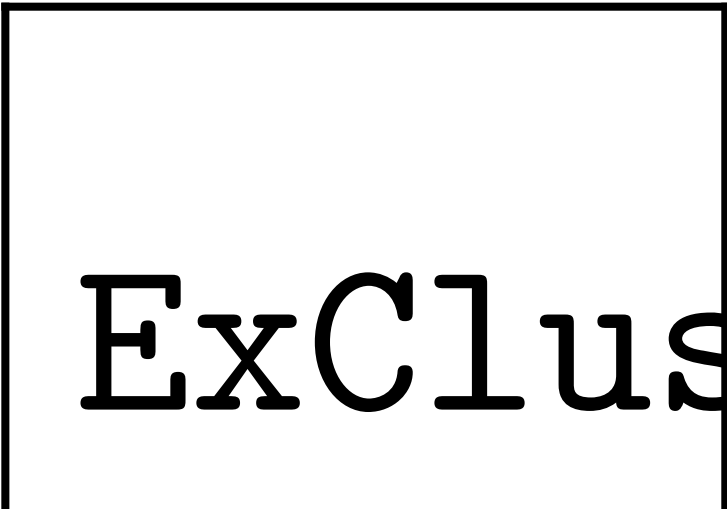
Definimos n_P el número de objetos en el clúster P .

Nombre	δ_1	δ_2	δ_3	δ_4
Enlace simple	$1/2$	$1/2$	0	$-1/2$
Enlace completo	$1/2$	$1/2$	0	$1/2$
Enlace promedio	$1/2$	$1/2$	0	0
Enlace promedio con pesos	$\frac{n_P}{n_P+n_Q}$	$\frac{n_Q}{n_P+n_Q}$	0	0
Centroide	$\frac{n_P}{n_P+n_Q}$	$\frac{n_Q}{n_P+n_Q}$	$-\frac{n_P n_Q}{(n_P+n_Q)^2}$	0
Mediana	$1/2$	$1/2$	$-1/4$	0
Ward	$\frac{n_R+n_P}{n_R+n_P+n_Q}$	$\frac{n_R+n_Q}{n_R+n_P+n_Q}$	$-\frac{n_R}{n_R+n_P+n_Q}$	0

Ejemplo

- Consideremos los puntos siguientes en

$$\mathbb{R}^2 : \quad (5, -3), (2, -4), (-2, -1), (-3, 0), \\ (-2, -2), (-2, 4), (1, 2), (1, 4).$$



ExClust

Ejemplo

- La matriz de distancias (distancia euclídea) entre los 8 puntos es:

$$D = \begin{pmatrix} 0.00 & 3.16 & 7.28 & 8.54 & 7.07 & 9.90 & 6.40 & 8.06 \\ & 0.00 & 5.00 & 6.40 & 4.47 & 8.94 & 6.08 & 8.06 \\ & & 0.00 & 1.41 & 1.00 & 5.00 & 4.24 & 5.83 \\ & & & 0.00 & 2.24 & 4.12 & 4.47 & 5.66 \\ & & & & 0.00 & 6.00 & 5.00 & 6.71 \\ & & & & & 0.00 & 3.61 & 3.00 \\ & & & & & & 0.00 & 2.00 \\ & & & & & & & 0.00 \end{pmatrix}$$

Ejemplo

- Aplicación del algoritmo. Como vemos, los puntos más cercanos son el 3 y el 5. Por tanto, los nuevos clústers serán:

$$\{1\}, \{2\}, \{3, 5\}, \{4\}, \{6\}, \{7\}, \{8\}.$$

- Calculemos la nueva matriz de distancias usando el enlace simple entre los 7 clústers:

$$D_2 = \begin{pmatrix} 0.00 & 3.16 & 7.07 & 8.54 & 9.90 & 6.40 & 8.06 \\ & 0.00 & 4.47 & 6.40 & 8.94 & 6.08 & 8.06 \\ & & 0.00 & 1.41 & 5.00 & 4.24 & 5.83 \\ & & & 0.00 & 4.12 & 4.47 & 5.66 \\ & & & & 0.00 & 3.61 & 3.00 \\ & & & & & 0.00 & 2.00 \\ & & & & & & 0.00 \end{pmatrix}$$

Ejemplo

- La distancia más cercana está ahora entre los clústers $\{3, 5\}$ y el clúster $\{4\}$. Los nuevos clúster son:

$$\{1\}, \{2\}, \{3, 4, 5\}, \{6\}, \{7\}, \{8\}.$$

La nueva matriz de distancias será:

$$D_3 = \begin{pmatrix} 0.00 & 3.16 & 7.07 & 9.90 & 6.40 & 8.06 \\ & 0.00 & 4.47 & 8.94 & 6.08 & 8.06 \\ & & 0.00 & 4.12 & 4.24 & 5.66 \\ & & & 0.00 & 3.61 & 3.00 \\ & & & & 0.00 & 2.00 \\ & & & & & 0.00 \end{pmatrix}$$

Ejemplo

- La distancia más cercana se encuentra entre los clústers $\{7\}$ y $\{8\}$. Los nuevos clústers serán:

$$\{1\}, \{2\}, \{3, 4, 5\}, \{6\}, \{7, 8\}.$$

La nueva matriz de distancias será:

$$D_4 = \begin{pmatrix} 0.00 & 3.16 & 7.07 & 9.90 & 6.40 \\ & 0.00 & 4.47 & 8.94 & 6.08 \\ & & 0.00 & 4.12 & 4.24 \\ & & & 0.00 & 3.00 \\ & & & & 0.00 \end{pmatrix}$$

Ejemplo

- La distancia más cercana se encuentra entre los clústers $\{6\}$ y $\{7, 8\}$. Los nuevos clústers serán:

$$\{1\}, \{2\}, \{3, 4, 5\}, \{6, 7, 8\}.$$

La nueva matriz de distancias será:

$$D_5 = \begin{pmatrix} 0.00 & 3.16 & 7.07 & 6.40 \\ & 0.00 & 4.47 & 6.08 \\ & & 0.00 & 4.12 \\ & & & 0.00 \end{pmatrix}$$

Ejemplo

- La distancia más cercana se encuentra entre los clústers $\{1\}$ y $\{2\}$. Los nuevos clústers serán:

$$\{1, 2\}, \{3, 4, 5\}, \{6, 7, 8\}.$$

La nueva matriz de distancias será:

$$D_6 = \begin{pmatrix} 0.00 & 4.47 & 6.08 \\ & 0.00 & 4.12 \\ & & 0.00 \end{pmatrix}$$

Ejemplo

- La distancia más cercana se encuentra entre los clústers $\{3, 4, 5\}$ y $\{6, 7, 8\}$. Los nuevos clústers serán:

$$\{1, 2\}, \{3, 4, 5, 6, 7, 8\}.$$

La nueva matriz de distancias será:

$$D_7 = \begin{pmatrix} 0.00 & 4.47 \\ & 0.00 \end{pmatrix}$$

- El último paso del algoritmo será juntar los dos últimos clusters obteniendo un clúster único compuesto por los 8 puntos.

Ejemplo

Todo el ejemplo queda resumido en lo que se denomina un dendograma:



Dendog

Propiedades de los algoritmos aglomerativos.

- En el procedimiento del enlace simple, la distancia entre el clúster R y el clúster unión de P y Q , $P + Q$ se puede hallar como:

$$d(R, P + Q) = \min(d(R, P), d(R, Q)).$$

Por dicho motivo, se denomina algoritmo del vecino más próximo. Dicho algoritmo tiende a construir clústers grandes. Clústers que difieren pero que no están bien separados pueden unirse en un sólo clúster si tienen dos individuos próximos.

Propiedades de los algoritmos aglomerativos.

- En el procedimiento del enlace completo, se trata de corregir este tipo de agrupamiento considerando distancias grandes. De hecho, la distancia entre el clúster R y el clúster unión de P y Q , $P + Q$ se puede hallar como:

$$d(R, P + Q) = \max(d(R, P), d(R, Q)).$$

Por dicho motivo, se denomina el algoritmo del vecino más alejado. Dicho algoritmo agrupa clústers donde todos los puntos están próximos.

- En el procedimiento del enlace promedio, se propone una solución intermedia entre los dos procedimientos anteriores. En este caso, se calcula la distancia promedio:

$$d(R, P + Q) = \frac{n_P}{n_P + n_Q} d(R, P) + \frac{n_Q}{n_P + n_Q} d(R, Q).$$

Propiedades de los algoritmos aglomerativos.

- El procedimiento del centroide es similar al procedimiento promedio pero usa la distancia geométrica entre el clúster R y el centro de gravedad promediado entre los clústers P y Q .
- El algoritmo de Ward se diferencia de los otros procedimientos respecto de la unificación. Dicho algoritmo no junta grupos con distancias pequeñas sino que junta grupos en los que no se incrementa la heterogeneidad de los mismos “demasiado”. El objetivo de dicho procedimiento es que la variación dentro de los clústers no aumente de forma drástica. El resultado es que los grupos son lo más homogéneos posible.

Ejemplo

El dendograma usando el método completo para el ejemplo anterior es:



Dendog

Ejemplo

El dendograma usando el método promedio para el ejemplo anterior es:



Dendog

Ejemplo

El dendograma usando el método de Ward para el ejemplo anterior es:



Dendog

Propiedades de los algoritmos aglomerativos.

- Se define la heterogeneidad de un clúster R como:

$$I_R = \frac{1}{n_R} \sum_{i=1}^{n_R} d^2(x_i, \bar{x}_R),$$

donde \bar{x}_R es el centro de gravedad o la media del clúster R . Si d es la distancia euclídea, I_R representa la varianza del clúster R .

- Cuando dos clústers P y Q se unen, el nuevo clúster $P + Q$ tiene una heterogeneidad I_{P+Q} . Puede demostrarse que el aumento de heterogeneidad viene dada por:

$$\Delta(P, Q) = \frac{n_P n_Q}{n_P + n_Q} d^2(P, Q).$$

- El algoritmo de Ward, por tanto, puede definirse como el algoritmo que minimiza $\Delta(P, Q)$.

Parte XVI

Bibliografía