METODOS ESTADISTICOS APLICADOS EN ARQUEOLOGIA

Trabajo Fin de Grado

Facultad de Matemáticas
Universidad de Sevilla

Foto de Portada: Altar del Dolmen de Matarrubilla
Valencina de la Concepción (Sevilla)
Fuente: Insituto Andaluz del Patrimonio Histórico

METODOS ESTADISTICOS APLICADOS EN ARQUEOLOGIA

Lucía Prada Domínguez

TRABAJO FIN DE GRADO
PRESENTADO PARA OPTAR
AL GRADO EN MATEMÁTICAS
UNIVERSIDAD DE SEVILLA
SEPTIEMBRE 2015

UNIVERSIDAD DE SEVILLA
DEPARTAMENTO DE
ESTADÍSTICA E INVESTIGACIÓN OPERATIVA

El abajo firmante ha leído este trabajo y recomienda a la Facultad de Matemáticas la aceptación del trabajo "**Métodos Estadísticos Aplicados en Arqueología** " realizado por **Lucía Prada Domínguez** como Trabajo Fin de Grado para obtener el Grado en Matemáticas conforme a lo dispuesto en la Ley.

Con fecha de: Septiembre 2015

Director:

Prof. Dr. José María Fernández Ponce

UNIVERSIDAD DE SEVILLA

Fecha: **Septiembre 2015**

Autora: **Lucía Prada Domínguez**

Título: **Métodos Estadísticos Aplicados en Arqueología**

Dpto: **Estadística e Investigación Operativa**

Firma de la autora

INDICE

Agradecimientos

A mis padres, por proporcionarme todos los medios necesarios para que pudiese cumplir mis objetivos.

A mi familia y amigos, por su apoyo incondicional. Porque para hacer un buen trabajo siempre es necesario rodearte de gente que confíe en ti.

A mi compañero, Luis, por hacer que en mi batalla personal contra el inglés yo saliese victoriosa.

A mi tutor, Jose María, por implicarse en este proyecto tanto como yo misma. Por estar siempre dispuesto a ayudarme en todo y por enseñarme a ver en la Arqueología una ciencia más que interesante.

Y en especial, a mi GUERRERA, porque le prometí dedicarle todos y cada uno de mis triunfos. Por enseñarme que cuando tienes un sueño hay que trabajar incansablemente hasta conseguirlo. Sin duda alguna, la constancia, la forma de afrontar los obstáculos y las ganas y alegría con la que hacía las cosas y que en su día me transmitió son las que me han llevado a finalizar este proyecto con éxito.

GRACIAS.

Resumen

La Arqueología no está tan lejos de ciencias como las Matemáticas, y en especial la Estadística, como podamos imaginar. Mostrar algunas aplicaciones de la Estadística en Arqueología es el objetivo de este proyecto.

En primer lugar, desarrollamos una técnica estadística como es el *Análisis de Datos Categóricos*. En segundo lugar, estudiaremos una técnica no paramétrica como la *Estimación Núcleo de la Densidad*. También presentaremos una herramienta multivariante como los *árboles de regresión y clasificación* (conocidos por sus siglas en inglés CART).

Y para finalizar estudiaremos un problema bastante común en Arqueología como es el caso de la datación por radiocarbono desde el punto de vista de la Estadística Bayesiana; enfoque que en los últimos años ha aumentado su contribución a las investigaciones estadísticas de manera espectacular.

Todas estas técnicas van acompañadas de ejemplos reales de estudios arqueológicos donde se emplean cada una de ellas, así como de los comandos del software `R` necesarios para su ejecución práctica.

Abstract

Archaeology is not so far from sciences like Mathematics, and specially Statistics, as we can believe. Analyzing some statistical applications in Archaeology is the purpose of this project.

First of all, a technique like the *Categorical Data Analysis* is explained. Secondly, we will study a non-parametric technique as the *Kernel Density Estimation* and a multivariate tool as the *Classification and Regression Trees*(known by its acronym CART).

Finally, we will study a very common problem in Archaeology like it is the radiocarbon dating from the point of view of the Bayesian Statistics, what has increased its contribution to Statistics research in a spectacular way lately.

All these techniques are illustrated with real examples from archaeological studies where each of them are used, as the R-software commands needed for its practical implementation.

Prólogo

SENATVSΣ*POPVLVSQVE*Σ*ROMANVS*
IMPΣ*CAESARI*Σ*DIVI*Σ*NERVAE*Σ*F*Σ*NERVAE*
TRAIANOΣ*AVG*Σ*GERM*Σ*DACICO*Σ*PONTIF*MAXIMOΣ*TRIB*Σ*POT*Σ*XVII*Σ*IMP*Σ*VI*Σ*COS*Σ*V I*Σ*P*Σ*P*Σ
AD*DECLARANDVM*Σ*QVANTAE*Σ*ALTITVDINIS*
MONS*ET*Σ*LOCVS*Σ*TANT* < ... > *IBVS*Σ*SIT*Σ*EGESTVS*

Es inevitable pensar en *Indiana Jones* cuando se habla de Arqueología y arqueólogos. Esa imagen romántica de esta Ciencia no sólo se debe al cine. En pleno Romanticismo, José Amador de los Ríos (Cordobés y Catedrático de Literatura de la Universidad Central de Madrid) denunció públicamente el expolio que se estaba cometiendo en las ruinas de Itálica (Santiponce, Sevilla). Su interés por estudiar y conservar los restos arqueológicos de la ciudad que vio nacer a dos emperadores de Roma hizo que en 1912 Itálica fuera declarada Monumento Nacional y que afortunadamente se llevaran acabo planes de rehabilitación, mantenimiento y reconstrucción de estos restos arqueológicos para encontrarse hoy en día en buen estado de conservación. Desde entonces la Arqueología ha cambiado mucho sus métodos de estudio y análisis debido al avance de la tecnología y de otras ciencias. De hecho en la actualidad no se concibe un arqueólogo sin conocimientos básicos de Química, Biología, Computación y Estadística. Por ejemplo, gracias a la computación y a la interpretación de los datos procedentes de las excavaciones, los arqueólogos han podido hacer reconstrucciones

3D de Itálica.

Uno de esos emperadores nacidos en Itálica fue Trajano (53 a.n.e.-117 a.n.e.) reconocido como uno de los mejores gobernadores por lo que contribuyó al progreso económico, social y cultural de la vieja Roma. Hizo construir la conocida *Columna de Trajano* como monumento conmemorativo de sus victorias frente a los dacios (actualmente Rumanía). Esta columna, aparte de su importancia histórica, tiene una característica que la hace original: su propia estructura en sí. Desde la base de la misma hasta la parte superior y en forma helicoidal se encuentra grabada la evolución de dichas batallas. Por ello puede ser considerada como la antesala del modelo de almacenamiento de datos masivos en discos duros que actualmente se lleva a cabo en computación. Hoy en día, los arqueólogos siguen tratando de descifrar e interpretar algunos de los mensajes que aparecen grabados en la columna.

Si viajamos en el tiempo unos 2500 años antes del nacimiento de Trajano nos encontraríamos con una zona del Aljarafe Sevillano y de la propia Sevilla muy diferente, no sólo a la actual sino incluso a la que conoció el afamado emperador Baetico. En aquella época, la Edad del Cobre, el lago Ligustinus llegaba hasta la actual Coria y en la cornisa del Aljarafe floreció una civilización dejándonos innumerables dólmenes, objetos de decoración y rituales de enterramientos únicos en Europa. Cabe resaltar el enterramiento encontrado en el Dolmen de Montelirio (Castilleja de Guzmán, Sevilla) en el año 2010, donde se encontraron los restos humanos de un varón de unos 40 años junto a los restos de 19 mujeres entre 20 y 30 años y los restos de tres varones (posiblemente *"los guardianes"*). Este tipo de enterramiento ritual es único en la cultura megalítica occidental y con cierto paralelismo con la tuma de Ur (Mesopotamia). Mediante técnicas de datación de radiocarbono se puede datar con cierto margen de error

la época del calcolítico en la que vivieron estas personas. También mediante técnicas geofísicas se puede detectar el emplazamiento de otros dólmenes y zonas domésticas que pudieran arrojar luz sobre esa época tan legendaria y mítica de nuestra tierra. No sin razón esta civilización podría ser considerada como la precursora de la mítica civilización tartésica (¿tal vez los Fenicios?), que surgió unos mil años después en las provincias de Huelva, Sevilla y Cádiz.

Por tanto, se pone de manifiesto la importancia que tienen otras ciencias en relación con la Arqueología. En particular, este Trabajo de Fin de Grado (TFG) se centrará en algunas aplicaciones de la Estadística en Arqueología, lo que viene a llamarse *Arqueoestadística*. La Arqueología como ciencia cuenta con una parte cuantitativa y a su vez de incertidumbre donde las tomas de decisiones son vitales tanto desde el punto de vista de inversión de dinero público en prospecciones arqueológicas con las mayores probabilidades de éxitos como en la interpretación y clasificación correcta de los restos encontrados. Por ello, la Arqueología y la Estadística deben ir de la mano en este campo de la investigación científica. Son numerosas estas aplicaciones, aquí se han seleccionado algunas para que puedan tratarse en este TFG por motivos lógicos de espacio.

Prof. Dr. José María Fernández Ponce

Dpto. Estadística e IO

Universidad de Sevilla

Capítulo 1

INTRODUCCION

"*ALEA IACTA EST*"

Julio César (100 a.n.e.-44 a.n.e.)

Resumen. En este Capítulo se describirán los conceptos básicos de partida para el posterior desarrollo del análisis estadístico. De esta forma se definirán los diferentes tipos de variables estadísticas y un breve resumen de lo que se pretende con este trabajo.

La Arqueoestadística es, como hacemos mención en el Prólogo, el estudio de fenómenos arqueológicos mediante herramientas estadísticas de diversa complejidad.

El objetivo de este Trabajo de Fin de Grado es hacer ver al lector la relación existente entre ambas ciencias, la Arqueología y la Estadística, apoyándonos en técnicas estadísticas que actualmente se emplean en los estudios arqueológicos.

Para contrastar nuestro estudio, emplearemos ejemplos arqueológicos reales en los cuales podremos observar la presencia e importancia de las técnicas estadísticas que iremos viendo a lo largo del proyecto.

En primer lugar, haremos una breve introducción de conceptos estadísticos necesarios para la correcta comprensión de los métodos estadísticos con los que trabajaremos. El concepto principal que debemos aclarar es el de *Variable estadística*. En Estadística, una variable aleatoria es aquella que nos permite representar y cuantificar los fenómenos aleatorios bajo estudio. Este estudio se lleva a cabo a través de la observación de datos.

Las variables aleatorias se pueden dividir en dos grandes grupos: Cuantitativas y Cualitativas.

- *Cuantitativas*: aquellas cuyos posibles valores son numéricos, por ejemplo, la estatura, el peso, la edad, la temperatura, etc. Estas se clasifican a su vez en variables continuas y discretas. Las variables continuas son aquellas que pueden tomar un conjunto infinito no numerable de valores y las discretas un conjunto finito o infinito numerable. Cabe destacar que la medición real de todas las variables se produce de una manera discreta, debido a las limitaciones de los instrumentos de medida. En la práctica, sin embargo, la distinción entre discreta y continua es una distinción entre variables que pueden tomar relativamente pocos valores y variables que pueden tomar muchos valores respectivamente.

- *Cualitativas* (o categóricas o atributos o factores): aquellas cuyos valores son un conjunto de cualidades no numéricas a las que se le suele llamar categorías o modalidades o niveles, por ejemplo, el sexo (mujer, hombre), el color de

pelo (moreno, rubio, castaño, pelirrojo), filosofía política (liberal, moderada, conservadora), etc.

Una propiedad deseable de las categorías es que sean exhaustivas (proporcionan suficientes valores para clasificar a toda la población) y mutuamente excluyentes (cada individuo se clasifica en una y solo una categoría).

A su vez, las variables cualitativas se pueden dividir en función de varios criterios. Dependiendo del criterio utilizado para su clasificación, las variables cualitativas se dividen en:

Variables dicotómicas y politómicas

Según el número de categorías las variables cualitativas se clasifican como

- *Dicotómicas*: tienen solo dos modalidades, por ejemplo, padecer una enfermedad (si, no), sexo (hombre, mujer), resultado de un examen (aprobar, suspender), en general los fenómenos de respuesta binaria, etc.
- *Politómicas*: tienen más de dos categorías, por ejemplo, los fenómenos de respuestas múltiples, lugar de nacimiento, clase social, etc.

Escalas nominal, ordinal y por intervalos

Según la escala de medida de sus categorías las variables cualitativas pueden ser clasificadas como

- *Nominales*: no se puede definir un orden natural entre sus categorías, por ejemplo, la raza (blanca, negra, otra), la religión (católica, judía, protestante, otra), etc.

- *Ordinales*: es posible establecer relaciones de orden entre las categorías lo que lleva a establecer relaciones de tipo mayor, menor, igual o preferencia entre los individuos. Por ejemplo, el rango militar (soldado, sargento, teniente, otro), la clase social (alta, media, baja), etc. Así, podemos decir que una persona de clase alta tiene mayor poder adquisitivo que una persona de clase media pero no podemos decir exactamente cuál es la diferencia en poder adquisitivo entre ambas.

- *Intervalo*: proceden de variables cuantitativas agrupadas en intervalos o que tienen un número pequeño de valores distintos. Estas variables pueden ser tratadas como ordinales pero para ellas se pueden calcular, además, distancias numéricas entre dos niveles de la escala ordinal, ejemplos de este tipo son el sueldo, la edad, los días del mes o el nivel de presión sanguínea.

Variables respuesta y explicativas

En muchos análisis es necesario distinguir entre variables que cambian en respuesta a condiciones fijadas (variables respuesta o dependientes) y variables que son tratadas como fijas y determinan la causa de la respuesta (variables explicativas o independientes).

Para este caso el análisis cualitativo dispone de técnicas similares a las de regresión, para describir como la distribución de una respuesta categórica cambia de acuerdo a

los niveles de variables explicativas, que pueden ser cuantitativas o cualitativas.

La razón por la que distinguimos entre datos cualitativos y cuantitativos es debido a que se usan métodos estadísticos diferentes para cada tipo de datos.

Esto es lo que podemos observar también a lo largo del proyecto, puesto que vamos a ver como en el capítulo 2 trataremos con datos cualitativos mientras que en el resto de capítulos trataremos con datos cuantitativos. También se podrá observar como se tratan técnicas tanto paramétricas como técnicas no paramétricas, conceptos que desarrollaremos en el capítulo 3.

Una vez hemos refrescado la memoria al lector sobre el concepto estadístico primordial en este estudio, veamos algunos de los aspectos más importantes que nos encontraremos a lo largo de cada capítulo.

En el capítulo 2, nos centraremos, como hemos dicho, en el estudio de datos categóricos. En primer lugar, vamos a comenzar con una breve introducción donde recordaremos los rasgos y aspectos más importantes que se presentan al tratar con este tipo de datos: los posibles problemas que podemos estudiar; las distribuciones muestrales más usadas; y el método de estimación de máxima verosimilitud.

Posteriormente, centraremos nuestro estudio únicamente en el problema del contraste de independencia. Comentaremos también algunas medidas de asociación para este tipo de problemas. Para llevar a cabo este análisis, nos apoyaremos en conceptos tales como el de *Tabla de Contingencia*; y discutiremos sobre las distintas técnicas existentes para la inferencia de dichas tablas, técnicas como el *Test Exacto de Fisher* y el *Test Chi Cuadrado*. Estas técnicas van acompañadas de las órdenes que debemos

emplear en el software R para aplicarlas a datos reales.

Por último, para completar el capítulo ilustraremos los conceptos vistos a lo largo del mismo mediante un ejemplo real aplicado a la revisión del registro arqueológico de Valencina de la Concepción (Sevilla).

En el capítulo 3, pasaremos a tratar con datos cuantitativos, particularmente, nos centraremos en el estudio de la estimación no paramétrica de la función de densidad de dichos datos.

Antes de entrar a comentar los conceptos teóricos necesarios para el estudio que se lleva a cabo en este capítulo, se hará una breve introducción donde recordaremos el concepto propio de función de densidad; discutiremos las ventajas y desventajas de las técnicas no paramétricas frente a las técnicas paramétricas, dejando claro en qué consisten cada una de ellas; y comentaremos los posibles problemas que se pueden tratar con las técnicas no paramétricas mediante una técnica específica que recibe el nombre de *Estimación Núcleo de la densidad*. También comentaremos alguna técnica similar a dicho método como es el *Histograma*. El objetivo de presentar técnicas similares a la dada es observar las ventajas que presenta el Método Núcleo frente a otras técnicas de uso similar. Estas ventajas son otra cuestión que trataremos en el capítulo.

Todos estos conceptos, al igual que en el capítulo anterior, van acompañados de los distintos paquetes del software R que podemos emplear para la ejecución de los mismos.

En segundo lugar, procederemos a presentar los conceptos teóricos en los que se basa el Método Núcleo. Para finalmente, ilustrar esta técnica mediante un ejemplo

real de un estudio arqueológico llevado a cabo en Pompeya.

Y por último, el proyecto finaliza con un cuarto capítulo donde se presentan varias técnicas estadísticas más, técnicas que al igual que la de los anteriores capítulos son usadas por los arqueólogos en sus estudios.

Así, en el capítulo 4 trataremos dos técnicas más. La primera de ellas tiene como objetivo abordar el problema de clasificación de objetos e individuos. Recibe el nombre de *CART*. Sobre dicha técnica veremos tanto los conceptos teóricos en los que se basa, de forma más breve que en los capítulos anteriores, así como una aplicación de la misma en un estudio arqueológico. En concreto, con dicha aplicación se pretende clasificar diferentes periodos de tiempo en base a las medidas de distintos cráneos Egipcios hallados en un yacimiento.

Anteriormente a la aplicación, y de forma similar a como se hizo en los capítulos anteriores, se comentan las distintas funciones del software `R` que permiten aplicar dicha técnica a un conjunto de datos.

La otra técnica abordada en este cuarto y último capítulo es la *Datación por Radiocarbono* desde el punto de vista de la Estadística Bayesiana.

El origen de la utilización de la estadística bayesiana en este tipo de estudios se remonta a principios de los años 90, donde gracias a la potencia de cálculo alcanzada por el empleo de los ordenadores, comenzaron a aparecer estudios que empleaban la estadística bayesiana en la interpretación de fechas radiocarbónicas. A partir de estos estudios se desarrollaron programas que permitían obtener cronologías empleando esta técnica, como OxCal o Bcal, programas que se comentaran en el desarrollo del

capítulo. Se llevará a cabo también una breve exposición sobre en qué consiste aplicar la estadística bayesiana al problema de datación por radiocarbono; para finalizar exponiendo la metodología en la que se basa dicha técnica ilustrándola con un ejemplo real en el que a través del empleo de muestras de ^{14}C se pretende estimar el periodo de tiempo en el que se desarrolló la cultura peruana pre-hispánica llamada Chancay.

Nuestra intención es que una vez vistas todas estas técnicas estadísticas y sus respectivas aplicaciones en los distintos estudios arqueológicos, el lector tenga una visión más cercana de la Estadística y la Arqueología así como un claro ejemplo de cómo la Estadística en los últimos años se ha convertido en una potente herramienta para muchas ciencias y actividades humanas.

Capítulo 2

ANALISIS DE DATOS CATEGORICOS

"No, no creo en la suerte, pero sí en asignar valor a las cosas."

John Nash, Jr.(1928-2015)

Resumen. En este capítulo se abordan las técnicas estadísticas cualitativas más importantes dentro de la Arqueología. En concreto, se describirán el test exacto de Fisher y los asintóticos Chi-cuadrado y de máxima verosimilud así como ventajas y desventajas de cada uno de ellos. Por último, se aplicará todo lo anterior a un ejemplo real de las excavaciones llevadas a cabo en la zona del Dolmen de La Pastora (Valencina de la Concepción, Sevilla).

2.1. Introducción

En esta sección introductoria se persiguen varios objetivos. El primero, justificar la necesidad del desarrollo de técnicas estadísticas específicas para el tratamiento de datos categóricos, que no son susceptibles de medida. El segundo, proporcionar una visión general del estado actual y de los antecedentes del análisis categórico, y en particular de las tablas de contingencia. El tercero, describir los elementos básicos de partida para el posterior desarrollo del análisis estadístico. Y el cuarto, hacer una revisión de algunos modelos de probabilidad que serán asumidos sobre los datos y que el lector interesado puede desconocer o haber olvidado.

2.1.1. Necesidad del análisis cualitativo

El *Análisis de Datos Cualitativos* se puede definir como un conjunto de técnicas estadísticas específicas para el estudio de las relaciones entre variables cualitativas, que son aquellas cuyos valores son cualidades no medibles de los individuos sujetos a estudio. Este tipo de variable aparece fundamentalmente en el campo de la medicina, las ciencias sociales, y más generalmente en las del comportamiento.

Mientras que en gran parte de las ciencias empíricas es posible medir con una escala el grado de presencia de las variables de interés, la dificultad inherente a la realidad social es que la mayoría de comportamientos sociales no son cuantificables, encontrándonos con un conjunto de cualidades para las que como mucho podremos ordenar sus distintas modalidades (variables cualitativas ordinales) o simplemente, formar grupos excluyentes y exhaustivos (variables cualitativas nominales). Al no ser susceptibles de medida, este tipo de variables no pueden ser analizadas con la metodología estadística convencional para datos cuantitativos. Esto ha dado origen

a una parte de la estadística que se conoce comúnmente con los nombres de *Análisis Cualitativo*, *Análisis de Datos Categóricos* o bien *Análisis de Datos Discretos* que da título al libro de [2] Bishop *et. al.* (1975).

El tratamiento matemático de estas variables se hace a partir de su único aspecto cuantificable: el número de veces que se presenta cada combinación de las modalidades de las variables estudiadas en una muestra, es decir, las frecuencias observadas. Estas frecuencias se presentan en tablas que reciben el nombre de *tablas de contingencia* o tablas cruzadas.

Los métodos de análisis estadísticos de datos categóricos, tanto nominales como ordinales, podrán ser aplicados a variables cuantitativas tomando como tablas de contingencia las tablas de correlación asociadas. Las variables cualitativas ordinales pueden ser tratadas con métodos específicos para el análisis de variables nominales pero el recíproco no es válido. A pesar de ello lo ideal es saber elegir en cada caso la técnica más apropiada en relación a la naturaleza de los datos.

2.1.2. Perspectiva histórica

Tradicionalmente, la inferencia estadística para tablas de contingencia se ha basado en gran medida en aproximaciones para muestras de tamaño grande. Muchas de estas aproximaciones son casos especiales de las que se aplican de manera más general a los datos categóricos (por ejemplo, aproximaciones chi-cuadrado para estadísticos de razón de verosimilitud y aproximaciones normales para estimadores de máxima verosimilitud de los parámetros del modelo). Con este énfasis en los métodos con muestras grandes, se podría decir que el desarrollo de los métodos inferenciales categóricos y el desarrollo inicial de los métodos continuos van de la mano.

De hecho, uno de los objetivos de la Estadística es inferir resultados observados en una muestra limitada del total de la población. Este nuevo objetivo y nuevas aproximaciones nacieron en torno a 1925, unos 20 años después de la publicación de la investigación de Gosset en la revista [31] Biometrika, basada en pequeñas muestras de "Guinnes beer", la compañía donde trabajaba debido a la escasez de trabajos académicos. A fin de no revelar secretos comerciales a compañías cerveceras rivales, el contrato de trabajo de Gosset le impedía publicar los resultados de su investigación. Para eludir este problema, publicó sus estudios usando el seudónimo "*A. Student*". Estos estudios fueron publicados entre 1907 y 1908, y los llevó a cabo usando la distribución t de Student, su logro más famoso.

Antes de la publicación de los estudios de Gosset, como hemos dicho anteriormente, los estadísticos estaban enfocados en la exploración de las distribuciones teóricas, llamada la distribución de la población completa, ya que trabajaba con muestras de gran tamaño. Uno de estos estadísticos fue Karl Pearson quien publicó un manuscrito sobre notas de estudiantes basado efectivamente en conjuntos de datos de gran tamaño. Fue este mismo quien al no llegar a entender la urgencia en el desarrollo de técnicas para muestras pequeñas criticó a Gosset diciendo "*Sólo los cerveceros traviesos negocian en muestras pequeñas*".

Más tarde, Fisher tomó la defensa de Gosset replicando "La maquinaria tradicional de los procesos estadísticos es totalmente inadecuada para las necesidades de la investigación práctica. No sólo se necesita un cañón para disparar a un gorrión, sino que encima falla". El elaborado mecanismo construido en la teoría de muestras infinitas no es adecuado para datos de laboratorio. Sólo abordando sistemáticamente problemas de muestras pequeñas parece posible aplicar tests adecuados a los datos prácticos. De

hecho, su libro [12] *R. A. Fisher's Statistical Methods for Research Workers* estaba en el primer plano de la defensa de procedimientos exactos para muestras pequeñas y fue en el prefacio de la primera edición de dicho libro en 1925 donde Fisher dijo las palabras anteriores.

La importancia de mejorar el ámbito de aplicación de métodos exactos para datos categóricos, así como el debate sobre el tamaño de las muestras ha ido incrementando con el paso de los años. Así, se han introducido nuevos métodos asintóticos que permiten que el número de casillas de una tabla de contingencia crezca a medida que crece el tamaño de la muestra (por ejemplo, [21] Morris, 1975), aunque a pesar de estos avances la información sobre la adecuación de estas aproximaciones asintóticas para los modelos estándar se encuentra en una etapa inicial.

Además, los estudios de simulación han demostrado que es imposible esperar pautas simples para indicar cuando las aproximaciones asintóticas con muestras grandes son adecuadas ([19]). Incluso cuando el tamaño muestral es bastante grande, [16] demostró que las aproximaciones para muestras grandes pueden ser muy pobres cuando la tabla de contingencia contiene a su vez valores pequeños y grandes de frecuencias esperadas.

Estos hechos, así como el tardío desarrollo de métodos para datos categóricos en comparación con los datos continuos son los que han provocado en parte el retraso en el desarrollo y uso de inferencias exactas para las tablas de contigencia. Otro motivo es la mayor complejidad computacional del mismo. Sin embargo, las mejoras concomitantes en la potencia de los ordenadores y los avances en la eficiencia de los algoritmos de cálculo han dado lugar a una mayor variedad de procedimientos exactos viables para el uso práctico, y a un aumento considerable en el tamaño de conjuntos

de datos y tablas a los que se pueden aplicar estos procedimientos.

2.1.3. Conceptos básicos

En este capítulo nos centraremos en el análisis de datos categóricos que se ocupa del estudio de las variables categóricas definidas anteriormente en el capítulo 1. La escala de medida de dichas variables será fundamental para la elección del procedimiento estadístico que usaremos para su estudio.

Cuestiones interesantes a estudiar cuando disponemos de variables categóricas podrían ser las siguientes:

- Podemos estudiar si los conteos observados de una variable categórica en cada una de sus categorías cumplen unas determinadas proporciones: **Contraste de Bondad de Ajuste**.
- Podemos estudiar si una variable categórica se comporta igual en varias subpoblaciones (o muestras): **Contraste de Homogeneidad**.
- Podemos estudiar la independencia o la posible relación entre varias variables categóricas: **Contraste de Independencia**.

En este capítulo abordaremos únicamente el estudio del contraste de independencia y veremos algunas medidas de asociación para el caso de variables asociadas. Para ello presentaremos los datos observados mediante tablas de contingencia, concepto que desarrollaremos más adelante, y discutiremos sobre las distintas técnicas existentes para la inferencia de dichas tablas. Entre las técnicas a discutir se encontraran principalmente el *Test Exacto de Fisher*, y el *Test Chi-cuadrado*, entre otras.

En los próximas secciones veremos un desarrollo teórico de dichas técnicas así como la aplicación de las mismas en la revisión de registros arqueológicos. Más específicamente, veremos cómo se aplican estas técnicas en la revisión del registro arqueológico disponible del sitio prehistórico de Valencina de la Concepción (Sevilla, España), uno de los asentamientos más importantes del Suroeste de la Península Ibérica durante los milenios III Y II ANE. A través de las técnicas que desarrollaremos a continuación examinaremos dos variables principales, la demografía y la metalurgia, con el objeto de valorar la más amplia cuestión de la complejidad social.

Con este ejemplo práctico pretendemos resaltar la importancia que ha cobrado hoy en día los métodos estadísticos en la Arqueología.

Por último, antes de entrar a desarrollar los conceptos teóricos de este capítulo es conveniente recordar las principales características de los modelos de probabilidad más usados en el análisis de datos categóricos. Así, mientras para los modelos de regresión con respuestas continuas, la distribución normal juega un papel central, las 4 distribuciones claves para respuestas categóricas son: Binomial, Multinomial, Poisson y Chi-cuadrado. También recordaremos en qué consiste el método de estimación de máxima verosimilitud.

Distribución binomial

Consideremos un experimento aleatorio con dos posibles resultados a los que llamaremos *éxito* y *fracaso* siendo la probabilidad de éxito igual a p $(0 < p < 1)$.

La variable aleatoria X que representa el número de éxitos en n realizaciones independientes de dicho experimento se dice que tiene una distribución de probabilidad

binomial de parámetros n y p, siendo su función de probabilidad la siguiente:

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

Su esperanza y su varianza son $\mathrm{E}[X] = np$ y $\mathrm{Var}[X] = np(1-p)$ y su notación abreviada es $B(n, p)$.

Distribución de Poisson

Una variable aleatoria X tiene distribución de probabilidad de Poisson de parámetro $\lambda > 0$ si su función de probabilidad es de la forma

$$P[X = x] = \frac{e^{-\lambda}\lambda^x}{x!}, \quad x = 0, 1, ..., \infty.$$

Su esperanza y su varianza son $\mathrm{E}[X] = \lambda$ y $\mathrm{Var}[X] = \lambda$, y suele abreviarse en la forma $P(\lambda)$.

La distribución de Poisson se obtiene como límite de una sucesión de distribuciones binomiales cuando $n \to \infty$, $p \to 0$ y np permanece fijo. En este sentido se conoce como distribución de los sucesos raros.

A menudo la distribución de Poisson describe el número de ocurrencias aleatorias e independientes de un determinado suceso en un intervalo de tiempo.

Distribución multinomial

Consideremos I sucesos mutuamente excluyentes $A_1, ..., A_I$ que constituyen una partición del espacio muestral asociado a un experimento aleatorio del que se llevan a cabo n realizaciones independientes. Denotaremos por p_i $(i = 1, ..., I)$ a las probabilidades de ocurrencia de cada uno de estos sucesos que verifican $0 < p_i < 1$ y $\sum_{i=1}^{I} p_i = 1$.

Entonces el vector aleatorio X de dimensión I cuyas componentes X_i representan el número de veces que se repite cada uno de los sucesos A_i en las n realizaciones independientes del experimento, sigue una distribución de probabilidad multinomial de parámetros n y $p = (p_1, ..., p_I)$, cuya función de probabilidad es

$$P[X_1 = x_1, ..., X_I = x_I] = \frac{n!}{\prod_{i=1}^{I} x_i!} p_i^{x_i}, \quad x_i = 0, 1, ...n : \sum_{i=1}^{I} x_i = n.$$

La distribución multinomial se suele denotar por $M(n; p_1, ..., p_I)$. Las distribuciones marginales unidimensionales de la distribución multinomial son binomiales. En este sentido se demuestra que $X_i \to B(n, p_i)$. De ello se deduce que $\mathrm{E}[X_i] = np_i$ y $\mathrm{Var}[X_i] = np_i(1 - p_i)$.

La distribución multinomial puede ser obtenida también condicionando un conjunto de variables de Poisson independientes sobre su suma.

Distribución chi-cuadrado

Una variable aleatoria X tiene distribución chi-cuadrado con n grados de libertad ($n \in N$) si su función de densidad es

$$f(x) = \begin{cases} \dfrac{1}{2^{n/2}\Gamma(n/2)}\, x^{(n/2)-1} e^{-x/2} & \text{para } x \geq 0, \\ 0 & \text{para } x < 0 \end{cases}$$

Esta distribución será denotada por χ^2_n y verifica $\mathrm{E}[X] = n$ y $\mathrm{Var}[X] = 2n$

Funciones de verosimilitud y estimador de máxima verosimilitud

La noción de verosimilitud procede del término inglés “likelihood” que, desde sus orígenes estuvo vinculado al concepto de probabilidad, probability, aunque denotando

un vínculo de causalidad más débil. La comparación de hipótesis a través de la evaluación de verosimilitudes puede encontrarse en obras tan tempranas como *Aeropagitica* de John Milton.

Sin embargo, el uso más moderno del término apareció en las obras de Thiele, a quien se atribuye la invención, y Peirce. La fijación del término tal y como lo conocemos hoy en día es, sin embargo, obra de R.A. Fisher, que trata de él en su artículo *On the Mathematical Foundations of Theoretical Statistics*.

La función de verosimilitud (o, simplemente, verosimilitud) es una función de los parámetros de un modelo estadístico que permite realizar inferencias acerca de su valor a partir de un conjunto de observaciones.

No debe confundirse con el término probabilidad: ésta permite, a partir de una serie de parámetros conocidos, realizar predicciones acerca de los valores que toma una variable aleatoria.

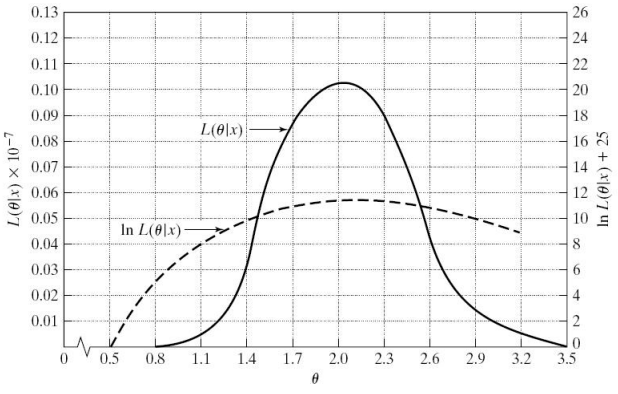

Figura 2.1: Función de verosimilitud y función de probabilidad para una distribución Poisson.

En cierto sentido, la verosimilitud es una versión inversa de la probabilidad condicional. Conocido un parámetro B, la probabilidad condicional de A es P(A|B), pero si se conoce A, pueden realizarse inferencias sobre el valor de B gracias al teorema de Bayes, según el cual

$$P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A)}$$

La función de verosimilitud, $L(b \mid A)$, definida como

$$L(b \mid A) = P(A \mid B = b)$$

desempeña el mismo papel bajo un enfoque no bayesiano. De hecho, lo relevante no es el valor en sí de $L(b \mid A)$ sino la razón de verosimilitudes,

$$\frac{L(b_2|A)}{L(b_1|A)},$$

que permite comparar cuanto más verosímil es el parámetro b_1 que el b_2 a la hora de explicar el evento A. De ahí que en ocasiones se entienda que la función de verosimilitud, más que una función en sí, sea la clase de funciones

$$L(b \mid A) = \alpha P(A \mid B = b),$$

donde α es una constante de proporcionalidad.

La función de verosimilitud, abundando en los razonamientos anteriores, abre la vía para dos técnicas muy habituales en inferencia estadística: las de la máxima verosimilitud y la del test de la razón de verosimilitudes.

La definición anterior es válida para distribuciones discretas. En el caso de distribuciones continuas, la función de verosimilitud se define de forma diferente. Supongamos que tenemos una variable aleatoria real de distribución desconocida X de la que se

extrae una muestra $x_1, ..., x_n$ de observaciones independientes. Supóngase también que se dispone de una familia parametrizada de funciones de densidad $f_\theta(x)$ (es decir, que existe una función de densidad $f_\theta(x)$ para cada valor del parámetro $\theta(x)$).

En este caso, $\theta(x)$ juega el papel de parámetro desconocido y es razonable definir la función de verosimilitud $L(\theta)$ de la siguiente manera:

$$L(\theta) = L(\theta \mid x_1, ..., x_n) = \prod_i f_\theta(x_i).$$

La función de verosimilitud se usa para la estimación de parámetros. De hecho, a partir de ella se definen los estimadores de máxima verosimilitud, que denotaremos por estimadores MV. El estimador MV es el valor del parámetro que maximiza la función de verosimilitud, esto es, el valor del parámetro bajo el cual los datos observados tienen la mayor probabilidad de ocurrencia.

Denotamos por β a un parámetro para un problema genérico y por $\hat{\beta}$ a su estimador. Denotamos también la función de verosimilitud por $l(\beta)$. Cabe destacar que el valor de β que maximiza $l(\beta)$ también maximiza $L(\beta) = log[l(\beta)]$, por ello en muchas ocasiones utilizamos la expresión $L(\beta)$ en vez de $l(\beta)$ pues es más simple maximizar una suma de términos que un producto. Para muchos modelos, $L(\beta)$ tiene forma cóncava y $\hat{\beta}$ es el punto en el cual la derivada es igual a 0. El estimador MV es entonces la solución de la ecuación de probabilidad

$$\frac{\partial L(\beta)}{\partial \beta} = 0.$$

A menudo, β es multidimensional, denotado por $\boldsymbol{\beta}$, y $\hat{\boldsymbol{\beta}}$ es la solución de un conjunto de ecuaciones de probabilidad. Sea cov($\boldsymbol{\beta}$) que denota la matriz de convarianzas asintótica de $\hat{\boldsymbol{\beta}}$. Bajo condiciones de regularidad, cov($\boldsymbol{\beta}$) es la inversa de la matriz

información. El elemento (j, k) de la matriz información es

$$-E\left(\frac{\partial^2 L(\beta)}{\partial\beta_j\partial\beta_k}\right). \tag{2.1.1}$$

Los errores estándar son las raíces cuadradas de los elementos de la diagonal de la inversa de la matriz información. Cuanto mayor es la curvatura de la función de verosimilitud, más pequeños son los errores estándar. Esto es razonable, ya que una gran curvatura implica que la log-verosimilitud cae rápidamente a medida que $\boldsymbol{\beta}$ se aleja de $\hat{\boldsymbol{\beta}}$; por lo tanto, los datos son mucho más probables de ocurrir si $\boldsymbol{\beta}$ toma un valor cercano a $\hat{\boldsymbol{\beta}}$ en vez de un valor lejano a $\hat{\boldsymbol{\beta}}$.

Los estimadores MV cobran tanta importancia en la inferencia de parámetros debido a que poseen propiedades deseables: tienen normalidad asintótica; son asintóticamente consistentes; convergen al parámetro estimado cuando n aumenta; y son asintóticamente eficientes, produciendo errores estándar no mayores que los cometidos por otros métodos de estimación.

2.2. Tablas de contingencia

El tratamiento estadístico de variables cualitativas se realiza a partir de su único aspecto cuantificable dado por las frecuencias observadas que se definen como el número de veces que se prsenta en una muestra cada combinación de niveles de las variables. Las frecuencias observadas se recogen en *tablas de contingencia* cuyo nombre es debido a [25] Pearson en 1904.

Consideremos un conjunto de n individuos clasificados según dos factores cualitativos A y B con I y J niveles, respectivamente. Si representamos por n_{ij} $(i =$

$1, ..., I$; $j = 1, ..., J$) el número de individuos de la muestra que se clasifican simultáneamente en el nivel A_i de A y B_j de B, la tabla bidimensional que contiene en cada una de sus $I \times J$ casillas las frecuencias observadas n_{ij} se llama *tabla de contingencia bidimensional* o *tabla cruzada*. A los niveles de A se le suelen llamar filas y a los de B columnas. La notación general de una tabla de contingencia es la que aparece en el Tabla 2.1.

Factor B / Factor A	B_1	B_2	...	B_J	Totales
A_1	n_{11}	n_{12}	...	n_{1J}	$n_{1\cdot}$
A_2	n_{21}	n_{22}	...	n_{2J}	$n_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_I	n_{I1}	n_{I2}	...	n_{IJ}	$n_{I\cdot}$
Totales	$n_{\cdot 1}$	$n_{\cdot 2}$	...	$n_{\cdot J}$	n

Tabla 2.1: *Notación para una tabla de contingencia $I \times J$.*

A partir de la tabla de contingencia se obtienen en la siguiente forma las distribuciones de frecuencias marginales.

$$Factor\ A: \quad n_{i\cdot} = \sum_{j=1}^{J} n_{ij}, \qquad (i = 1, .., I),$$

$$Factor\ B: \quad n_{\cdot j} = \sum_{i=1}^{I} n_{ij}, \qquad (j = 1, .., J),$$

$$\text{verificando } n = \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij} = \sum_{i=1}^{I} n_{i.} = \sum_{j=1}^{J} n_{.j}.$$

Modelos muestrales para las frecuencias observadas

Como dijimos previamente en el Capítulo 1, nuestro objetivo es el estudio del contraste de independencia entre variables categóricas. La forma de realizar dicho contraste entre dos factores dependerá de la distribución de frecuencias observadas que depende del procedimiento de muestreo considerado. Por lo tanto, en el análisis estadístico de tablas de contingencia las frecuencias observadas se consideran realizaciones de variables aleatorias, con valores enteros no negativos, cuyas esperanzas reciben el nombre de *frecuencias esperadas*. En el caso de una tabla bidimensional, las frecuencias observadas n_{ij} son realizaciones de variables aleatorias que denotaremos por N_{ij}, siendo sus frecuencias esperadas m_{ij}=E$[N_{ij}]$. Con objeto de simplificar notación, a partir de ahora daremos a las variables aleatorias que generan a las frecuencias observadas la misma notaación que a sus valores observados, de modo que el lector deberá diferenciarlas en función del contexto en que aparezcan.

A continuación estudiaremos los modelos de probabilidad más usuales que pueden considerarse en el diseño muestral como generadores de las frecuencias observadas.

Muestreo Poisson

El modelo muestral más simple supone que el tamaño muestral es desconocido a priori y que las frecuencias n_{ij} en cada una de las $I \times J$ posibles casillas de la tabla son variables aleatorias independientes con distribución de Poisson de parámetros las frecuencias esperadas m_{ij}.

Por lo tanto, la distribución de probabilidad conjunta de la tabla de frecuencias es el producto de las $I \times J$ distribuciones de Poisson independientes dado por

$$\prod_{i=1}^{I}\prod_{j=1}^{J}\frac{e^{-m_{ij}}m_{ij}^{n_{ij}}}{n_{ij}!}. \tag{2.2.1}$$

En la práctica, el muestreo de Poisson consiste en fijar un intervalo de tiempo y clasificar los individuos de forma independiente en las $I \times J$ categorías de la variables bidimensional de interés. De este modo el tamaño muestral queda determinado cuando pasa el intervalo de tiempo considerado.

Muestreo multinomial completo

La tabla de contingencia se genera tomando una muestra aleatoria simple de la población de tamaño muestral n fijado y clasificándola en las $I \times J$ posibles combinaciones de categorías de los dos factores considerados.

Entonces la distribución a priori de la variable de dimensión $I \times J$ que representa a las frecuencias observadas es una multinomial de parámetros $(n, \{p_{ij} : i = 1, ..., I;\ j = 1, ..., J\})$, siendo p_{ij} las probabilidades poblacionales de ocurrencia de cada una de las combinaciones de niveles de las variables que verifican $\sum_{i=1}^{I}\sum_{j=1}^{J} p_{ij} = 1$.

Por lo tanto, la probabilidad del conjunto de frecuencias observadas (n_{ij} viene dada por

$$\frac{n!\prod_{i=1}^{I}\prod_{j=1}^{J}p_{ij}^{n_{ij}}}{\prod_{i=1}^{I}\prod_{j=1}^{J}n_{ij}!}. \tag{2.2.2}$$

Como consecuencia las frecuencias esperadas en tablas $I \times J$ generadas por muestreo multinomial son de la forma $m_{ij} = np_{ij}$.

Lo que hace inusual el muestreo de Poisson es que el tamaño muestral n no es fijo sino aleatorio. Veamos a continuación que la distribución muestral de Poisson condicionada al tamaño muestral $n = \sum_i \sum_j n_{ij}$ da lugar a la distribución muestral multinomial para la tabla de contingencia.

Proposición 2.2.1. *Si la distribución muestral de una tabla de contingencia $I \times J$ es Poisson independiente del tipo (2.2.1), entonces esta distribución condicionada al tamaño muestral $n = \sum_i \sum_j n_{ij}$ es multinomial completa del tipo (2.2.2) con*

$$p_{ij} = \frac{m_{ij}}{\sum_i \sum_j m_{ij}}.$$

Demostración. Supongamos que el vector $I \times J$ de frecuencias observadas (n_{ij}) tiene una distribución muestral de Poisson dada por la ecuación (2.2.1). Entonces se tiene

$$\begin{aligned} P[(n_{ij})/\sum_i \sum_j n_{ij} = n] &= \frac{P[(n_{ij})]}{P[\sum_i \sum_j n_{ij} = n]} \\ &= \frac{\prod_{i=1}^{I} \prod_{j=1}^{J} (exp(-m_{ij}))m_{ij}^{n_{ij}}/n_{ij}!}{(exp(-\sum_i \sum_j m_{ij}))(\sum_i \sum_j m_{ij})^n/n!} \\ &= \frac{n!}{\prod_{i=1}^{I} \prod_{j=1}^{J} n_{ij}!} \prod_{i=1}^{I} \prod_{j=1}^{J} p_{ij}^{n_{ij}}. \end{aligned}$$

Donde la segunda igualdad se obtiene de la propiedad de reproductividad de la distribución de Poisson y quedando así demostrada la propiedad. □

Muestreo multinomial independiente

La tabla de contingencia se genera tomando muestras aleatorias simples independientes de tamaños fijados (los totales marginales de una de las variables) en cada

nivel de una de las variables y clasificando los individuos de cada muestra según las categorías de la otra variable.

Se toman, por ejemplo, I muestras aleatorias simples independientes de I subpoblaciones representads por los niveles de la variable fila y los individuos de cada muestra se clasifican según las categorías de la variable columna. Si denotamos por $n_{i.}$ $(i = 1, ..., I)$ al tamaño fijo de cada una de las I muestras, la variable de dimensión J que representa a las frecuencias obesrvadas en la i-ésima fila tiene distribución multinomial de parámetros $(n_{i.}, \{p_{j|i} : i = 1, ..., I;\ j = 1, ..., J\})$, donde $p_{j|i}$ representa la probabilidad poblacional de clasificación en la columns j para los individuos de la fila i verificando $\sum_{j=1}^{J} p_{j|i} = 1$.

Por lo tanto, la probabilidad de las frecuencias observadas en la i-ésima fila es

$$\frac{n_i!}{\prod_{j=1}^{J} n_{ij}!} \prod_{j=1}^{J} p_{j|i}^{n_{ij}} \tag{2.2.3}$$

Finalmente, como las I muestras son independientes, la probabilidad conjunta de la tabla de frecuencias completa es el producto de las I funciones de probabilidad multinomiales

$$\prod_{i=1}^{I} \frac{n_i!}{\prod_{j=1}^{J} n_{ij}!} \prod_{j=1}^{J} p_{j|i}^{n_{ij}}. \tag{2.2.4}$$

Por lo tanto, las frecuencias esperadas de tablas $I \times J$ generadas por muestreo multinomial independiente por filas son $m_{ij} = n_{i.} p_{j|i}$.

Independencia poblacional y muestral

Consideremos una tabla de contingencias $I \times J$ generada por muestreo multinomial completo. Denotando por p_{ij} a la probabilidad de que un individuo elegido aleatoriamente en la población se clasifique en el nivel A_i de A y en el nivel B_j de B, se obtiene la *distribución de probabilidad poblacional* de los factores A y B verificando $\sum_{i=1}^{I}\sum_{j=1}^{J} p_{ij} = 1$. Para representarla se usa el Tabla 2.2 que tiene la misma estructura que la tabla de contingencia

Factor A \ Factor B	B_1	B_2	...	B_J	Totales
A_1	p_{11}	p_{12}	...	p_{1J}	$p_{1\cdot}$
A_2	p_{21}	p_{22}	...	p_{2J}	$p_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_I	p_{I1}	p_{I2}	...	p_{IJ}	$p_{I\cdot}$
Totales	$p_{\cdot 1}$	$p_{\cdot 2}$	...	$p_{\cdot J}$	1

Tabla 2.2: *Tabla $I \times J$ de probabilidades poblacionales.*

De igual forma que para las frecuencias observadas, se definen las distribuciones de probabilidad marginales asociadas.

$$Factor\ A: \quad p_{i\cdot} = \sum_{j=1}^{J} p_{ij}, \qquad (i = 1, .., I),$$

$$Factor\ B: \quad p_{\cdot j} = \sum_{i=1}^{I} p_{ij}, \qquad (j = 1, .., J),$$

$$\text{verificando } \sum_{i=1}^{I} p_{i.} = \sum_{j=1}^{J} p_{.j} = 1.$$

Las distribuciones marginales dan información unidimensional sobre cada variable y no dicen nada sobre la asociación entre las dos variables.

Bajo muestreo multinomial completo, la hipótesis de independencia poblacional entre las dos variables cualitativas A y B es de la forma

$$p_{ij} = p_{i.}p_{.j} \quad \forall i = 1, ..., I;\ j = 1, ..., J, \tag{2.2.5}$$

que puede expresarse equivalentemente en términos de frecuencias esperadas como

$$m_{ij} = \frac{m_{i.}m_{.j}}{n}, \quad i = 1, ..., I;\ j = 1, ..., J, \tag{2.2.6}$$

donde $m_{i.}$ y $m_{.j}$ representan a las frecuencias marginales esperadas.

Esto significa que las distribuciones condicionales de B son iguales que su distribución marginal. Es decir, dos variables son independientes cuando la probabilidad de clasificarse en la columna j es igual en todas las filas.

De forma similar se puede definir el concepto de *independencia muestral*. Para ello, asociada a una tabla de contingencia del tipo correspondiente al Cuadro 2.1 se define también la *distribución muestral de proporciones* o probabilidades muestrales como $\hat{p}_{ij} = n_{ij}/n$ que es la proporción de individuos muestrales clasificados en la casilla (i, j). Entonces, la proporción de veces que un individuo de la fila i se clasifica en la columna j es $\hat{p}_{j|i} = \hat{p}_{ij}/\hat{p}_{i.}$, siendo $\hat{p}_{i.} = n_{i.}/n$ la proporción muestral de individuos en la fila i.

Las variables A y B son independientes en la muestra si se verifica

$$\hat{p}_{ij} = \hat{p}_{i.}\hat{p}_{.j}$$

que equivalentemente, en términos de frecuencias observadas es de la forma

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

2.3. Inferencia para tablas de contingencia bidimensionales

En la práctica, la distribución de probabilidad asumida por las variables respuestas tienen valores de parámetros desconocidos. En esta sección, revisaremos métodos de uso de datos muestrales para hacer inferencia sobre dichos parámetros. También abordaremos el estudio del contraste estadístico de la independencia entre dos variables cualitativas, tratando en primer lugar los métodos asintóticos y terminando con el estudio de los métodos basados en distribuciones exactas.

Estimación por MV de las frecuencias esperadas

Proposición 2.3.1. *Los estimadores MV de las probabilidades poblacionales p_{ij} bajo el muestreo multinomial son simplemente las proporciones muestrales*

$$\hat{p}_{ij} = \frac{n_{ij}}{n}. \tag{2.3.1}$$

Demostración. Denotemos por (p_{ij}) al vector $(p_{11}, p_{12}, \cdots, p_{IJ})$. Recordemos que en el caso de muestreo multinomial completo la función de verosimilitud es de la forma

$$L(p_{ij}) = \frac{n!}{\prod_{i=1}^{I}\prod_{j=1}^{J} n_{ij}!} \prod_{i=1}^{I}\prod_{j=1}^{J} p_{ij}^{n_{ij}}.$$

La parte de esta función de verosimilitud que involucra los parámetros es el núcleo de la verosimilitud. Está claro que maximizar L respecto de p_{ij} es lo mismo que maximizar su núcleo K. Además, como mencionamos anteriormente, es equivalente también maximizar K que maximizar su logaritmo. Por lo tanto, maximizaremos, el logaritmo del núcleo de la verosimilitud dado por

$$\log K\{p_{ij}\} = \sum_i \sum_j n_{ij} \log p_{ij} \tag{2.3.2}$$

sujeto a las restricciones $p_{ij} > 0$ y $\sum_i \sum_j p_{ij} = 1$.

A continuación, vamos a calcular el máximo directamente mediante el método de multiplicadores de Lagrange. Utilizando dicho método, se define la función

$$\phi = \sum_i \sum_j n_{ij} \log p_{ij} + \lambda \left(\sum_i \sum_j p_{ij} - 1 \right).$$

Derivando respecto a p_{ij} y respecto a λ, e igualando a cero, se tiene:

$$\frac{\Delta \phi}{\Delta p_{ij}} = \frac{n_{ij}}{p_{ij}} + \lambda = 0,$$

$$\frac{\Delta \phi}{\Delta \lambda} = \sum_i \sum_j p_{ij} - 1 = 0,$$

de donde se obtiene $\hat{p}_{ij} = \dfrac{-n_{ij}}{\hat{\lambda}}$, lo que implica:

$$\sum_i \sum_j p_{ij} = 1 = \sum_i \sum_j \frac{-n_{ij}}{\hat{\lambda}},$$

y por lo tanto, $\hat{\lambda} = -n$, que proporciona la expresión de los EMV de las probabilidades poblacionales

$$\hat{p}_{ij} = \frac{n_{ij}}{n}.$$

Esto quiere decir que los estimadores MV de las probabilidades poblacionales son simplemente las proporciones muestrales. □

Como consecuencia del resultado anterior, los estimadores MV de las frecuencias esperadas son las frecuencias observadas n_{ij}, y los estimadores MV de las probabilidades marginales son las proporciones muestrales marginales $(\hat{p}_{i.})$ y $(\hat{p}_{.j})$, ya que bajo la hipótesis de independencia poblacional, los estimadores MV de p_{ij} son

$$\hat{p}_{ij} = \hat{p}_{i.}\hat{p}_{.j} = \frac{n_{i.}n_{.j}}{n^2}.$$

Por lo tanto, la estimación MV de las frecuencias esperadas en el caso de independencia es

$$\hat{m}_{ij} = \frac{n_{i.}n_{.j}}{n^2}.$$

que se suelen llamar frecuencias esperadas estimadas y tienen los mismos totales marginales que la tabla de frecuencia observada. Por ejemplo, $\hat{m}_{i.} = \sum_j \hat{m}_{ij} = n_{i.}$.

Recordemos que en el caso de muestreo multinomial, cada frecuencia n_{ij} tiene distribución marginal $\mathrm{B}(n, p_{ij})$. Por lo tanto, aplicando la propiedad de invarianza de los estimadores MV, para las frecuencias esperadas m_{ij}=E$[n_{ij}]$=np_{ij} los estimadores MV serían $\hat{m}_{ij} = n\hat{p}_{ij} = n_{ij}$.

De la definición de las probabilidades marginales, $(p_{i.})$ y $(p_{.j})$, se deduce también que sus estimadores MV son las proporciones muestrales marginales $(\hat{p}_{i.})$ y $(\hat{p}_{.j})$.

A continuación se obtendrán los estimadores MV de las frecuencias esperadas bajo la hipótesis de independencia $p_{ij} = p_{i.}p_{.j}$.

Sustituyendo en $logK$ por esta última expresión, el problema se reduce a maximizar

$$\log K = \sum_i n_{i.} \log p_{i.} + \sum_j n_{.j} \log p_{.j},$$

bajo las restricciones $\sum_i p_{i.} = \sum_j p_{.j} = 1$.

Utilizando el método de los multiplicadores de Lagrange se define la función

$$\phi = \sum_i n_{i.} \log p_{i.} + \sum_j n_{.j} \log p_{.j} + \lambda_1 \left(\sum_i p_{i.} - 1 \right) + \lambda_2 \left(\sum_j p_{.j} - 1 \right).$$

Derivando convenientemente se tiene

$$\begin{aligned}
\frac{\Delta \phi}{\Delta p_{i.}} &= \frac{n_{i.}}{p_{i.}} + \lambda_1 = 0, \\
\frac{\Delta \phi}{\Delta p_{.j}} &= \frac{n_{.j}}{p_{.j}} + \lambda_2 = 0, \\
\frac{\Delta \phi}{\Delta \lambda_1} &= \sum_i p_{i.} - 1 = 0, \\
\frac{\Delta \phi}{\Delta \lambda_2} &= \sum_j p_{.j} - 1 = 0.
\end{aligned}$$

Despejando $\hat{p}_{i.} = -n_{i.}/\hat{\lambda}_1$ en la primera ecuación y sustituyendo en la tercera se tiene $\hat{\lambda}_1 = -n$, de donde se deduce $\hat{p}_{i.} = n_{i.}/n$.

Análogamente, despejando $\hat{p}_{.j} = -n_{.j}/\hat{\lambda}_2$ en la segunda ecuación y sustituyendo en la cuarta se tiene $\hat{\lambda}_2 = -n$, de donde se deduce $\hat{p}_{.j} = n_{.j}/n$.

Finalmente, aplicando la propiedad de invarianza de los EMV, se demuestra que los estimadores MV de p_{ij} bajo independencia son

$$\hat{p}_{ij} = \hat{p}_{i.}\hat{p}_{.j} = \frac{n_{i.}n_{.j}}{n^2}.$$

Por lo tanto, la estimación MV de las frecuencias esperadas en el caso de independencia es

$$\hat{m}_{ij} = n\hat{p}_{i.}\hat{p}_{.j} = \frac{n_{i.}n_{.j}}{n^2}.$$

2.3.1. Contrastes de independencia asintóticos

Contraste chi-cuadrado de independencia

Consideremos una tabla de contingencia $I \times J$ generada por muestreo multinomial completo, de modo que las frecuencias observadas n_{ij} tienen distribución multinomial, $\mathrm{M}(n; (p_{ij}))$, verificando $\sum_i \sum_j p_{ij} = 1$. Supongamos que queremos contrastar la hipótesis nula de independencia

$$\mathrm{H}_0 \,:\, p_{ij} = p_{i.}p_{.j} \quad \forall i = 1, ...I;\, j = 1, ..., J.$$

Para llevar a cabo este contraste se usa el estadístico chi-cuadrado de Pearson para una multinomial de parámetros p_{ij} que genera la tabla. Dado que bajo H_0 las probabilidades poblacionales dependen de un total de $(I-1)+(J-1)$ parámetros desconocidos, dados por las probabilidades marginales $p_{i.}$ y $p_{j.}$, el estadístico chi-cuadrado de Pearson para contrastar la independencia se define en la forma

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}}, \tag{2.3.3}$$

siendo $\hat{m}_{ij} = n\hat{p}_{i.}\hat{p}_{.j} = n_{i.}n_{.j}/n$ los estimadores MV de las frecuencias esperadas bajo la hipótesis de independencia, definidas por $m_{ij} = np_{ij} = np_{i.}p_{.j} = m_{i.}m_{.j}/n$.

El principal inconveniente del estadístico χ^2 es que su cálculo es muy laborioso

porque conlleva obtener en primer lugar las frecuencias esperadas. Mediante cálculos sencillos se obtiene la siguiente expresión operativa del estadístico chi-cuadrado para una tabla 2×2:

$$\chi^2 = \frac{n(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1.}n_{2.}n_{.1}n_{.2}}. \tag{2.3.4}$$

Se comprueba que el estadístico χ^2 así definido tiene una distribución asintótica $\chi^2_{(I-1)(J-1)}$. Por lo tanto, *se rechaza la hipótesis de independencia* al nivel de significación α cuando se verifica

$$\chi^2_{Obs} \geq \chi^2_{(I-1)(J-1);\alpha}$$

siendo el $p-$valor del test $\mathrm{P}[\chi^2 \geq \chi^2_{Obs}]$.

[24] Pearson en el año 1900 estableció erróneamente que los grados de libertad del estadístico Chi-cuadrdado eran $(IJ-1)$ ya que hay en total IJ casillas en la tabla e $(IJ-1)$ parámetros libres de la distribución multinomial de las frecuencias observadas. Fue [11] Fisher en 1922 quién corrigió el error y estableció que los grados de libertad se obtienen como el número de parámetros libres menos el número de parámetros a estimar, en la forma $df = (IJ-1) - (I-1) - (J-1) = (I-1)(J-1)$.

Contraste de independencia de razón de verosimilitudes

Vamos a usar ahora el test de razón de verosimilitudes para contrastar la hipótesis de independencia

$$\mathrm{H}_0 \,:\, p_{ij} = p_{i.}p_{.j} \quad \forall i, j$$

en una tabla de contingencia $I \times J$ generada por muestreo multinomial. Para llevar a cabo este contraste se usa el estadístico de Wilks de razón de verosimilitudes que viene dado por

$$G^2 = -2\log\Lambda = 2\sum_i \sum_j n_{ij} \log \frac{n_{ij}}{\hat{m}_{ij}}, \tag{2.3.5}$$

siendo $\hat{m}_{ij} = n_{i.}n_{.j}/n$ la estimación MV de las frecuencias esperadas bajo la hipótesis de independencia. Se comprueba que el estadístico G^2 se distribuye asintóticamente como una variable aleatoria χ^2 con $(I-1)(J-1)$ grados de libertad.

Efectivamente, haciendo uso de los estimadores de MV para muestreo multinomial obtenidos anteriormente se tiene que el estadístico de razón de verosimilitudes es

$$\Lambda = \frac{\prod_{i=1}^{I}\prod_{j=1}^{J}(n_{i.}n_{.j})^{n_{ij}}}{n^n \prod_{i=1}^{I}\prod_{j=1}^{J} n_{ij}^{n_{ij}}},$$

de donde se deduce fácilmente la expresión (2.3.5).

Aplicando el test de razón de verosimilitudes se rechaza la hipótesis de independencia al nivel α si se verifica

$$G^2_{Obs} \geq \chi^2_{(I-1)(J-1);\alpha}.$$

Estudio comparativo de los estadísticos χ^2 y G^2

1. χ^2 y G^2 son asintóticamente equivalentes. De hecho, $\chi^2 - G^2$ converge en probabilidad a cero. Para ambos estadísticos valores grandes proporcionan más evidencia contra H_0.

2. Los resultados límites obtenidos para muestreo multinomial son también válidos para los otros tipos de muestreo. Es decir, los test chi-cuadrado y de razón de verosimilitudes de independencia son independientes del diseño muestral considerado.

3. El test χ^2 es más intuitivo porque mide distancias entre las frecuencias observadas y esperadas.

4. El test G^2 tiene la desventaja de usar logaritmos en su cálculo pero puede descomponerse para incrementar la potencia del test del contraste de independencia condicional en tablas múltiples. Esta descomposición del estadístico G^2 resulta especialmente útil en la selección de los modelos log-lineales.

5. El valor de los estadísticos χ^2 y G^2 depende de los totales marginales de filas y columnas y no del orden entre las filas y columnas. La invarianza frente a permutaciones de filas y columnas lleva a ignorar la información adicional en el caso de variables ordinales para las que se dispone de contrastes de independencia más potentes basados en alternativas más restringidas.

6. Estos dos contrastes son válidos para tamaños muestrales grandes. La eficiencia de la aproximación chi-cuadrado depende del tamaño muestral y de las frecuencias esperadas estimadas. Aunque no exista una regla simple para decidir el tamaño muestral adecuado para la aplicación de cada uno de estos contrastes usaremos el método propuesto por Cochran en 1954, que consiste en usar χ^2 cuando al menos el 80 % de las frecuencias esperadas sean mayores que 5 y todas ellas mayores que 1.

Se ha demostrado que para un número fijo de casillas χ^2 converge a la distribución chi-cuadrado más rápidamente que G^2. Por ello se puede utilizar para tamaños muestrales más pequeños y tablas más dispersas.

Para el caso de muestras pequeñas se considerará una solución alternativa que veremos a continuación. Dicha solución consiste en construir contrastes de independencia basados en distribuciones exactas en lugar de aproximadas para las frecuencias observadas.

2.3.2. Contrastes de independencia exactos

Hemos visto que los contrastes de independencia chi-cuadrado son válidos para tamaños muestrales grandes. Nos planteamos a continuación encontrar procedimientos alternativos para contrastar la independencia a partir de muestras pequeñas. Usaremos para ello la distribución exacta de la tabla de frecuencias observadas.

Supongamos que queremos contrastar la hipótesis nula de independencia entre dos factores cualitativos frente a la hipótesis alternativa de existencia de asociación entre ambos. El $p-$valor del test de independencia exacto es la probabilidad, bajo la hipótesis nula, de todas las tablas que obedecen al mismo diseño muestral que la observada y evidencian igual o mayor alejamiento de la hipótesis de independencia que la tabla de frecuencias observadas.

Por lo tanto un test exacto de independencia entre dos factores se puede resumir en los siguientes pasos:

1. Obtención del espacio muestral que está formado por todas aquellas tablas que obedecen el mismo diseño muestral que la tabla observada.

2. Selección de las tablas que se alejan de las hipótesis de independencia tanto o más que la tabla observada en la dirección marcada por la hipótesis alternativa.

3. Cálculo de las probabilidades exactas, bajo la hipótesis de independencia, de las tablas seleccionadas.

4. Cálculo del $p-$valor con el nivel de significación prefijado y decisión de no rechazo ($p-$valor mayor que el nivel de significación) o rechazo ($p-$valor menor o igual que el nivel de significación) de la hipótesis de independencia.

Se observa claramente que el test estará fuertemente ligado al tipo de diseño muestral considerado. Este tipo de contrastes exactos conlleva varios problemas. Por un lado, en la mayoría de los diseños muestrales considerados para generar la tabla, la distribución exacta de las frecuencias observadas bajo la hipótesis de independencia depende de parámetros desconocidos que se aproximan normalmente por sus estimadores MV. Por otro lado, al aumentar el tamaño muestral y el número de filas y columnas de la tablas, el espacio muestral de las tablas de frecuencias que obedecen al mismo diseño muestral que la observada aumenta considerablemente (especialmente para diseños muestrales que no fijan los totales marginales de algún factor) y el cálculo del $p-$valor es muy laborioso y casi impracticable a menos que se use un programa computacional.

Como consecuencia los contrastes exactos sólo son viables para muestreo hipergeométrico y multinomial independiente, que limitan el número de tablas del espacio muestral al fijar los totales marginales de filas y/o columnas.

Test exacto de Fisher para tablas 2×2

Se puede demostrar que la distribución de probabilidad exacta de una tabla de contingencia $I \times J$ con los totales marginales de ambas variables fijos es una hipergeométrica multivariante que bajo la hipótesis nula de independencia no depende de parámetros desconocidos, y se obtiene condicionando cualquiera de los diseños muestrales Poisson, multinomial o multinomial independiente a los totales marginales de filas y columnas.

Basándose en dicha distribución hipergeométrica [13] Fisher desarrolló en 1935 el test exacto que lleva su nombre para el contraste de independencia en tablas 2×2, que se presenta a continuación.

Consideremos una tabla 2×2 generada mediante muestreo hipergeométrico, lo que significa que los totales marginales $(n_{1\cdot}, n_{2\cdot}, n_{\cdot 1}, n_{\cdot 2})$ de ambas variables están fijos.

Supongamos que queremos contrastar la hipótesis de independencia poblacional que para una tabla 2×2, generada por muestreo multinomial independiente por filas, se puede expresar como

$$\mathrm{H}_0 : p_{1|1} = p_{1|2} = \pi.$$

Se comprueba que la distribución exacta bajo H_0 de cualquier distribución de frecuencias observadas cuyos valores marginales coincidan con los fijados de antemano se trata de una hipergeométrica y viene dada por la siguiente expresión

$$\begin{aligned} P[(n_{11}, n_{12}, n_{21}, n_{22})|(n_{,1}, n_{,2})] = \\ = \frac{P[(n_{11}, n_{12}, n_{21}, n_{22})]}{P[(n_{\cdot 1}, n_{\cdot 2})]} = \frac{n_{1\cdot}!n_{2\cdot}!n_{\cdot 1}!n_{\cdot 2}!}{n!n_{11}!n_{12}!n_{21}!n_{22}!}, \end{aligned} \tag{2.3.6}$$

que expresa la distribución de las cuatro casillas de la tabla en términos del elemento n_{11} debido a que, dados los totales marginales, si valor determina las otras tres casillas de la tabla. El rango de posibles valores para n_{11} es claramente

$$max\{0, n_{\cdot 1} - (n - n_{1\cdot})\} \leq n_{11} \leq min\{n_{\cdot 1}, n_{1\cdot}\}$$

.

Dado que aparecen en la fórmula de la distribución hipergeométrica un total de $IJ+I+J+1$ factoriales, el cálculo de las probabilidades de todas las posibles tablas es complicado y laborioso. Este proceso se simplifica aplicando la fórmula de Feldman y Kinger que calcula una de estas probabilidades, por ejemplo, la de la tabla observada, y las demás se obtienen a partir de ella. Si denotamos por $p_{n_{11}}$ a la probabilidad que asigna la distribución hipergeométrica a una tabla con frecuencia n_{11} en la casilla (1,1), la expresión para la probabilidad de las demás tablas es

$$p_{n_{11}+1} = \frac{n_{12}n_{21}}{(n_{11}+1)(n_{22}+1)} p_{n_{11}}, \tag{2.3.7}$$

$$p_{n_{11}-1} = \frac{n_{11}n_{22}}{(n_{12}+1)(n_{21}+1)} p_{n_{11}}. \tag{2.3.8}$$

El paso siguiente es fijar una hipótesis alternativa y seleccionar aquellas tablas que se alejan de H_0 tanto o más que la tabla observada en la dirección de la hipótesis alternativa considerada. La probabilidad anterior deberá calcularse para todas las tablas seleccionadas. Posteriormente, estas probabilidades se usan para calcular el $p-$valor asociado al test exacto de Fisher.

Sea α el nivel de significación prefijado de antemano,

$$\text{Si } p-\text{valor}<\alpha \Rightarrow \textit{Rechazamos Hipótesis nula}$$

y por tanto, debemos asumir que las dos variables no son independientes sino que están asociadas. En caso contrario, se dirá que no existe evidencia estadística de asociación entre ambas variables.

A continuación mostraremos varios métodos para el cálculo del $p-$valor:

1. Sumando las probabilidades de aquellas tablas con una probabilidad asociada menor o igual a la correspondiente a los datos observados.

2. Sumando las probabilidades asociadas a resultados al menos tan favorables a la hipótesis alternativa como los datos reales.

Este cálculo proporcionaría el p-valor correspondiente al test en el caso de un planteamiento unilateral. Duplicando este valor se obtendría el p-valor asociado a un test bilateral.

Extensión del test exacto de Fisher en una tabla $I \times J$

Consideramos el test de hipótesis de independencia para una tabla de contingencia $I \times J$ con muestreos multinomiales. Sea

$$\boldsymbol{X} = \begin{pmatrix} X_{11} & \cdots & X_{1J} \\ \vdots & \ddots & \vdots \\ X_{I1} & \cdots & X_{IJ} \end{pmatrix},$$

donde un valor particular se denotará como $\mathbf{x}$ y con una distribución multinomial

$$P[\mathbf{X} = \mathbf{x}] = n! \prod_{i=1}^{I} \prod_{j=1}^{J} \frac{p_{ij}^{x_{ij}}}{x_{ij}!},$$

donde $0 < p_{ij} < 1$, para $i = 1, ..., I$, $j = 1, ..., J$, $\sum_{i=1}^{I} \sum_{j=1}^{J} p_{ij} = 1$.

[15] Freeman and Halton (1951) propusieron una extensión del test exacto de Fisher en una tabla $I \times J$, en la cual el p-valor para el test de independencia se define como la probabilidad nula del conjunto de tablas que tienen probabilidad no mayor que la probabilidad de la tabla observada. Específicamente, baja la hipótesis nula de independencia, la probabilidad condicionada de observar una tabla muestral X dadas las filas y columnas marginales es

$$P_{H_0}(X) = \frac{\prod_{i=1}^{I} X_{i.}! \prod_{j=1}^{J} X_{.j}!}{n! \prod_{i=1}^{I} \prod_{j=1}^{J} X_{ij}!},$$

donde $X_{i.} = \sum_{j=1}^{J} X_{ij}$, $i = 1, ..., I$ y $X_{.j} = \sum_{i=1}^{I} X_{ij}$, $j = 1, ..., J$.

El $p-$valor $= \sum_{Y \in \mathcal{F}} P_{H_0}(Y)$, donde

$$\mathcal{F} = \left\{ Y : Y \text{ es una tabla } I \times J, P(Y) \leq P(X), \sum_{j=1}^{J} Y_{ij} = X_{i.}, \sum_{i=1}^{I} Y_{ij} = X_{.j} \right\}$$

Este test está condicionado a los valores observados de las filas y columnas marginales.

La extensión del test exacto de Fisher para una tabla de contingencia $I \times J$ parece que se usa a menudo, pero todavía no se ha demostrado ninguna muestra finita o propiedad óptima asintótica.

Dado que el test exacto de Fisher para tablas de contingencia $I \times J$ ordena únicamente los puntos muestrales sobre la base de la probabilidad de ocurrencia bajo la hipótesis nula, el test ha recibido fuertes críticas. Las críticas se deben a que la configuración de las frecuencias de las casillas puede ser menos probable que la tabla observada bajo la hipótesis nula, pero en algún sentido presentan menos discrepancia de la hipótesis nula que de la tabla observada.

2.4. Medidas de asociación en tablas bidimensionales

En caso de rechazar la independencia entre los dos factores de una tabla de contingencia, se plantea la necesidad de definir índices que describan no solo la intensidad de la asociación, sino también su dirección. El estudio de estos índices, que se conocen con el nombre genérico de *medidas de asociación*, es el objetivo principal de esta sección.

2.4.1. Funciones del cociente de ventajas

Consideremos una tabla de contingencia 2×2, como la reprentada en el Tabla 2.3, generada por muestreo multinomial completo, cuyo vector de probabilidades poblacionales, denotado por $\underline{p} = (p_{ij})$ $(i, j = 1, 2)$, verifica $\sum_{i=1}^{2} \sum_{j=1}^{2} p_{ij} = 1$.

	B_1	B_2	$n_{i\cdot}$
A_1	n_{11}	n_{12}	$n_{1\cdot}$
A_2	n_{21}	n_{22}	$n_{2\cdot}$
$n_{\cdot i}$	$n_{\cdot 1}$	$n_{\cdot 2}$	n

Tabla 2.3: *Tabla de contingencia* 2×2.

Cociente de ventajas

El *cociente de ventajas* o *razón de productos cruzados* se define como el cocientre entre la ventaja de la segunda columna para los individuos de la segunda fila y la misma ventaja para los individuos de la primera fila. Muchas veces se usa su denominación inglesa *odd ratio* cuya traducción literal es cociente de ventajas.

En general, para un suceso A de probabilidad p, su ventaja es la probabilidad de que ocurra A en lugar de que no ocurra A, dada por $p/(1-p)$.

Para la tabla 2×2 considerada, la ventaja de la segunda columna para los individuos de la primera fila viene dada por

$$\omega_1 = \frac{p_{2|1}}{p_{1|1}} = \frac{p_{12}}{p_{11}},$$

y representa la probabilidad de que un individuo elegido al azar en la primera fila se clasifique en la segunda columna en lugar de en la primera.

Análogamente, la ventaja de la segunda columna para los individuos de la segunda fila es

$$\omega_2 = \frac{p_{2|2}}{p_{1|2}} = \frac{p_{22}}{p_{21}}.$$

Observemos que $0 \leq \omega_i < \infty$ $(i = 1, 2)$, de modo que si $\omega_i > 1$ entonces, para los individuos de la fila i, la probabilidad de clasificarse en la segunda columna es mayor que en la primera.

El *cociente de ventajas poblacional* se define como

$$\theta = \frac{\omega_2}{\omega_1} = \frac{p_{11}p_{22}}{p_{12}p_{21}} = \frac{m_{11}m_{22}}{m_{12}m_{21}}, \tag{2.4.1}$$

expresión que justifica el nombre de razón de productos cruzados.

El rango de variación del cociente de ventajas θ es claramente el intervalo $[0, +\infty)$.

Interpretación y propiedades de θ

1. $\theta = 0 \Rightarrow p_{11} = 0$ o $p_{22} = 0$, que se interpreta como asociación perfecta de tipo II (cuando algún nivel del primer factor está relacionado con más de uno del segundo y al revés también, es decir, ningún factor puede ser determinado completamente a partir del otro), o bien asociación perfecta estricta negativa si ambas probabilidades son nulas (cuando cada nivel de un factor está asociado con uno y solo uno del otro factor).

2. $\theta = 1$ si y sólo si las variables A y B son independientes.

3. $\theta > 1 \Rightarrow \omega_2 > \omega_1$ lo que significa que la probabilidad de clasificarse en la segunda columna en lugar de en la primera es mayor para los individuos de la segunda fila que para los de la primera (asociación positiva).

4. $\theta < 1$ significa que la probabilidad de clasificarse en la segunda columna en lugar de en la primera es mayor para los individuos de la primera fila que para los de la segunda (asociación negativa).

5. Es invariante frente a cambios de escala en filas y columnas.

6. Al cambiar de orden las dos filas y las dos columnas, θ se convierte en el inverso del original representando el mismo grado de asociación pero en distinta dirección.

El estimador de MV del cociente de ventajas recibe el nombre de cociente de ventajas muestral y viene dado por

$$\hat{\theta} = \frac{n_{11}n_{22}}{n_{12}n_{21}}.$$

Se demuestra que bajo muestreo multinomial completo, $\hat{\theta}$ tiene la siguiente distribución asintótica normal multivariante:

$$(\hat{\theta} - \theta) \quad \xrightarrow[n\to\infty]{d} \quad N(0, \hat{\sigma}^2(\hat{\theta})),$$

donde $\hat{\sigma}^2(\hat{\theta})$ es de la forma

$$\hat{\sigma}^2(\hat{\theta}) = \hat{\theta}^2 \left[\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}\right].$$

Q de Yule

La medida de asociasión Q de Yule fue propuesta por [33] Yule en 1900 en honor al estadístico belga Quetelet. Su valor poblacional es

$$Q = \frac{p_{11}p_{22} - p_{12}p_{21}}{p_{11}p_{22} + p_{12}p_{21}} = \frac{\theta - 1}{\theta + 1}.$$

El rango de variación de Q es el intervalo [-1,1].

Interpretación y propiedades

1. $Q = 0$ si y sólo si las variables A y B son independientes.

2. $Q > 0 \Leftrightarrow$ asociación positiva.

3. $Q < 0 \Leftrightarrow$ asociación negativa.

4. Si hay asociación perfecta estricta Q vale 1 o -1.

La Q de Yule muestral es el EMV de Q dado por la siguiente expresión:

$$\hat{Q} = \frac{n_{11}n_{22} - n_{12}n_{21}}{n_{11}n_{22} + n_{12}n_{21}} = \frac{\hat{\theta} - 1}{\hat{\theta} + 1}.$$

2.4.2. Medidas para comparar proporciones

Consideremos ahora que se desea comparar dos grupos (representados, por ejemplo, por las dos filas de la tabla 2×2) sobre una variable respuesta cualitativa binaria (representada, por ejemplo, por las columnas de la tablas 2×2). Para estudiar la asociación en este caso se estudian a continuación dos medidas asimétricas para una tabla 2×2 generada mediante muestreo multinomial independiente por filas.

Diferencia de proporciones

Se puede tomar como medida de asociación la diferencia de probabilidades condicionadas de la primera respuesta en cada fila, definida por

$$p = p_{1|1} - p_{1|2},$$

que es equivalente a utilizar la diferencia entre las probabilidades condicionadas de la segunda respuesta en cada fila

$$p_{2|1} - p_{2|2} = -p.$$

El rango de variación de la diferencia de probabilidades condicionadas es claramente [-1,1].

Interpretación y propiedades

1. $p = 0 \Leftrightarrow$ las variables son independientes.

2. $p = 1 \Leftrightarrow$ hay asociación perfecta estricta positiva.

3. $p = -1 \Leftrightarrow$ hay asociación perfecta estricta negativa.

4. $-1 < p < 0 \Rightarrow$ asociación negativa.

5. $0 < p < 1 \Rightarrow$ asociación positiva.

La estimación muestral de p recibe el nombre de difrencia de proporciones y se obtiene como el estimador MV de p dado por

$$\hat{p} = \hat{p}_{1|1} - \hat{p}_{1|2} = \frac{n_{11}}{n_{1.}} - \frac{n_{21}}{n_{2.}}.$$

Riesgo relativo

El *riesgo relativo* para la respuesta representada por la primera columna se define como

$$R = \frac{p_{1|1}}{p_{1|2}},$$

siendo su rango de variación el intervalo $[0, +\infty)$.

Si se comparan las dos filas sobre la segunda respuesta el riesgo relativo es diferente

$$\frac{p_{2|1}}{p_{2|2}} = \frac{1 - p_{1|1}}{1 - p_{1|2}}.$$

Interpretación y propiedades

1. $R = 1 \Leftrightarrow$ A y B son independientes.

2. $R = 0 \Rightarrow$ asociación perfecta implícita de tipo II.

La estimación MV del riesgo relativo poblacional es

$$\hat{R} = \frac{\hat{p}_{1|1}}{\hat{p}_{1|2}} = \frac{n_{11}n_{2\cdot}}{n_{21}n_{1\cdot}}.$$

2.5. Aplicación

Como comentamos en el Capítulo 1, queremos ver como se aplican las técnicas mencionadas anteriormente en un estudio arqueológico concreto. Esta sección se desarrollará de la siguiente forma: en primer lugar vamos a dar una descripción exhaustiva de la zona a estudiar, en segundo lugar presentaremos los diferentes datos a estudiar y por último presentaremos el objetivo del análisis y comentaremos los resultados y conclusiones obtenidas en dicho estudio.

2.5.1. Descripción del yacimiento arqueológico

El yacimiento arqueológico de Valencina de la Concepción (Sevilla) está localizado al margen derecho del río Guadalquivir, a 6 km del centro de la ciudad de Sevilla,

ocupando una de las zonas más elevadas de la región del Aljarafe, principalmente dentro del municipio de Valencina de la Concepción, pero también en parte de Castilleja de Guzmán. La comunidad prehistórica vivía en un entorno físico muy diferente al actual. Las principales características del entorno prehistórico han comenzado recientemente a ser determinadas a partir de estudios geoarqueológicos de la parte baja del río Guadalquivir y de los depósitos fluviales en la ciudad de Sevilla y sus alrededores. Este entorno físico se caracterizó sobre todo por la riqueza y diversidad de los recursos naturales que ofrecía el gran golfo marino en el que fluía el Guadalquivir, por la compleja red de canales de ríos y pantanos que ocupó la desembocadura del río, y por las tierras fértiles del Aljarafe.

La investigación científica en el yacimiento prehistórico de Valencina se remonta a finales del siglo XIX. La orientación y el perfil de las numerosas excavaciones y los estudios han ido cambiando con el tiempo de acuerdo con el desarrollo general de la Arqueología como disciplina en España. La lista de los especialistas que han contribuido a esta investigación incluye algunos de los más famosos e influyentes investigadores a través de varias generaciones de prehistoriadores españoles.

2.5.2. El registro empírico

Osteología humana

Durante el largo periodo de estudio, el yacimiento arqueológico de Valencina ha revelado una importante colección de huesos humanos que, con algunas excepciones, nunca ha sido objeto de una investigación exhaustiva.

Los datos analizados aquí comprenden un total de 135 individuos. Estos datos provienen de un estudio reciente que incluye una recopilación y organización de los

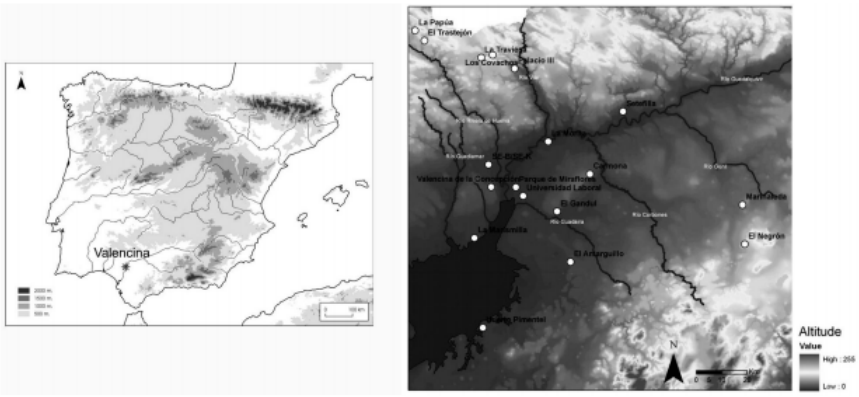

Figura 2.2: Izquierda: Localización del asentamiento de la Edad del Cobre de Valencina de la Concepción en la Península Ibérica. Derecha: Mapa de la localización de Valencina de la Concepción en relación con los asentamientos y sitios funerarios del 3rd y 2nd milenio excavado en el valle inferior del Guadalquivir. El mapa muestra el contorno de la costa estimado de la Prehistoria Reciente.

datos que ya están disponibles en informes publicados y no publicados, así como un estudio de algunos de los restos humanos del lugar que se conservan en el Museo Arqueológico de Sevilla, y que nunca antes se habían investigado.

El desglose general de la muestra antropológica considerada en este estudio requiere algunos comentarios y aclaraciones previas. Un punto a tener en cuenta es que el grado de conservación de los restos humanos en general no es muy bueno, y varía de una excavación a otra. En Corte A de La Perrera los individuos eran documentados casi completos en un buen estado de conservación. Por el contrario, en sitios como El Algarrobillo, La Cima, La Gallega y La Alcazaba, los esqueletos estaban muy fragmentados debido a los procesos tafonómicos: había algunos cráneos, pero no huesos largos y pelvis. Por esta razón la determinación del sexo se basó en las características sexuales difomórficas del cráneo descritas por [6] Buikstra y Ubelaker (1994), y la estimación de la edad se hizo de acuerdo con el grado de desgaste dental siguiendo

los métodos de [5], y por comparación con el grado de obliteración de las suturas craneales, de acuerdo con los datos de [26]. Para establecer un análisis comparativo de entre las diferentes metodologías de estimación de la edad, se utilizaron los siguientes rangos de edad: subadultos, 20-30 años, 30-45 años, más de 45 años, adultos e indeterminado.

Un segundo aspecto importante a considerar es el contexto funcional, espacial y arquitectónico para cada hallazgo, que se basan en las descripciones de los excavadores. Como hemos mencionado anteriormente, uno de los objetivos principales de este artículo es examinar la organización interna de Valencina estudiando en detalle la distribución espacial del registro osteológico humano. Para tal fin, los contextos en los que se han encontrado los restos humanos se han agrupado en dos categorías: Megalítico y no Megalítico.

El desglose total de los esqueletos de Valencina por edad y sexo se presenta en los Tablas 2.4 y 2.5. En general, esta población muestra una distribución equilibrada en cuanto a sexo.

SECTOR	FEMALE					MALE				S	A	?		TOTAL
	20-30	30-45	>45	A	YA	20-30	30-45	>45	A			?	30-45	
Matarrubilla	0	0	0	0	0	0	0	0	0	0	0	2	0	2
Los Cabezuelos	2	0	0	1	0	2	1	0	0	1	0	5	2	14
Cerro de la Cabeza	0	0	0	0	0	0	0	0	0	0	0	1	0	1
El Roquetito	0	0	0	2	0	0	0	0	0	0	0	46	0	48
Divina Pastora-Señorío de Guzmán	3	1	0	3	1	3	3	0	0	5	0	1	0	20
Norte Castilleja	0	0	0	0	0	0	1	0	0	0	0	0	0	1
Total	5	1	0	6	1	5	5	0	0	6	0	55	2	86

Tabla 2.4: Distribución del sexo y la edad en contextos funerarios megalíticos. Abreviaturas: (A): Adulto; (YA): Adulto Joven; (S): Subadulto; (?): Indeterminado.

En relación con el contexto de apariencia, el número de individuos encontrados en

SITE	FEMALE					MALE				S	A	?			TOTAL
	20-30	30-45	>45	A	YA	20-30	30-45	>45	A			?	20-30	30-45	
El Algarrobillo	2	0	0	0	1	4	0	0	2	0	4	0	3	0	16
La Alcazaba	0	0	0	0	0	0	0	0	1	0	3	1	2	0	7
La Cima	1	0	0	0	0	0	0	0	0	1	0	0	0	0	2
La Gallega	0	0	1	0	0	0	0	0	0	1	0	0	0	0	2
La Perrera	0	0	0	3	1	0	0	0	4	1	0	1	0	0	10
P.P. Matarrubilla	0	0	0	0	0	0	0	0	0	0	0	6	0	0	6
El Cuervo	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1
Mirador de Itálica	0	0	0	0	0	0	0	0	0	0	0	2	0	0	2
Mariana Pineda	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1
La Candelera-Emisora	0	0	0	0	0	0	0	0	0	0	0	2	0	0	2
Total	3	0	1	3	2	4	0	0	7	3	9	12	5	0	49

Tabla 2.5: Distribución del sexo y la edad en contextos funerarios no megalíticos. Abreviaturas: (A): Adulto; (YA): Adulto Joven; (S): Subadulto; (?): Indeterminado.

contextos megalíticos es 86 (que representa el 63.7 % de la muestra total), donde el 15.1 % son mujeres y el 11.6 % son hombres. En los contextos no megalíticos, el número mínimo de individuos considerados (MNI) es 49 (36.3 %), con 18.3 % de mujeres y 22.4 % de hombres. En ambos tipos de contextos el rango de edad predominante es el de 20-30 años (con 11.62 % en megalíticos y 24.4 % en no megalíticos). Los individuos subadultos representan un porcentaje muy pequeño, solo el 6.6 % del total. Esto representa una anomalía interesante que podría deberse a causas tafonómicas (deterioro de los huesos de los niños), causas culturales (los cadáveres de los individuos fueron sometidos a prácticas funerarias que dejaron no dejaron restos arqueológicos visibles) o causas epistemológicas (deficiencias en el estudios arqueológico y de observación del lugar). Debido a la falta de estudios de la antropología física en el yacimiento de Valencina, todavía hay una tasa muy alta de individuos indeterminados, un 66.2 % en contextos megalíticos y un 34.6 % en contextos no megalíticos. Estos datos pueden encontrarse en [7].

Objetos metálicos y metalurgia

Los datos de los objetos metálicos considerados en este estudio provienen de una síntesis llevada a cabo recientemente ([8]) que incorpora diversas fuentes, incluidos los informes publicados en el *Anuario Arqueológico de Andalucía*, dos monografías que proporcionan los resultados de los estudios arqueométricos y que se derivan de dos tesis doctorales ([17];[27]), así como otros artículos de revistas científicas especializadas.

Actualmente hay 105 artefactos de metal registrados del yacimiento de Valencina. Esta cifra incluye 29 puntas de jabalinas encontradas fuera de los *tholos* de La Pastora, que diversos estudios coinciden en que pertenecen a la Edad del Bronce, es decir, en la fase posterior de la ocupación del asentamiento. Además sus morfologías, la cantidad y los contextos aparentes de sus deposiciones hacen de estos objetos metálicos unos objetos excepcionales.

Aunque ningún estudio aún ha analizado y evaluado toda la información disponible desde una perspectiva global y multidisciplinar, la colección de objetos metálicos de Valencina es una de las más importantes de la Edad del Cobre de la Península Ibérica.

Para su análisis, los objetos de metal se han dividido en varias categorías: herramientas, armas-herramientas, armas, adornos y objetos indeterminados.

2.5.3. Análisis de los datos

Objetivos

Sobre la base de los datos descritos en el apartado anterior, se ha llevado a cabo un análisis con el fin de examinar la organización del espacio en el asentamiento

de Valencina. Este objetivo se puede dividir en dos más específicos. El primero es determinar la validez de la división espacial convencional entre un sector doméstico-productivo en el norte y un sector funerario en el sur.

El segundo objetivo es investigar el grado de especialización funcional del espacio dentro de esta comunidad prehistórica, ambos en términos económicos y sociales, específicamente en relación con la producción metalúrgica y el uso y deposición de objetos metálicos. En un nivel metodológico, la especialización productiva y/o funcional del espacio, la existencia de diferencias significativas en el tamaño y la naturaleza de las estructuras domésticas y la presencia de zonas sociales, son parámetros generalmente considerados de relevancia en el análisis arqueológico del surgimiento de las sociedades altamente jerarquizadas, estratificadas y de bienestar social.

Los métodos de análisis de datos que se usan para examinar estos dos problemas incluyen tanto tests estadísticos convencionales como espaciales, los cuales se usan comúnmente en las investigaciones arqueológicas actuales, tales como los análisis de densidad, el test χ^2, el análisis del vecino más cercano y los tests índice de Moran de autocorrelación parcial.

Resultados

Análisis Osteológico

En relación con el primero de los objetivos de este estudio, se llevan a cabo varios procedimientos estadísticos para determinar la validez de la división espacial convencional que divide el terreno en un sector doméstico-productivo en el norte y un sector

funerario en el sur. Aunque hagamos referencia y nos ayudemos de los resultados obtenidos por varias técnicas estadísticas (el análisis de densidad, el análisis del vecino más cercano o el test índice de Moran) para apoyar las conclusiones a las que se llegan, no entraremos a analizarlas en detalle debido a que el objetivo de esta parte práctica es reflejar el uso en Arqueología únicamente de las técnicas vistas en la parte teórica del capítulo, técnicas como el test χ^2 o el test exacto de Fisher, las cuales si desarrollaremos de una forma más completa.

En primer lugar, se lleva a cabo un análisis de la densidad del MNI/m^2, el cual sugiere que el material osteológico aparece distribuido por todo el área del asentamiento. Encontrándose un mayor número de individuos (MNI=101) en el sector sur, convencionalmente conocido como "necrópolis"[1], mientras que en el sector norte (el área "doméstica" o "productiva") se encontraron una cantidad de individuos no del todo despreciable (MNI=34). En el sector norte, los contextos en los que se hallaron los restos humanos fueron descritos por los excavadores como silos[2], zanjas, pozos y estructuras subterráneas. En este estudio, todas estos contextos se han agrupado bajo el nombre de contexto "no megalítico". En la parte sur del asentamiento, sin embargo, los restos humanos aparecieron predominantemente en construcciones megalíticas[3].

[1]Una necrópolis es una especie de cementerio o lugar destinado a enterramientos. Etimológicamente significa ciudad de los muertos/cadáveres, pues proviene del idioma griego: necro, muerto o cadáver, y polis, ciudad. El término se emplea normalmente para designar cementerios pertenecientes a grandes urbes, así como para las zonas de enterramiento que se han encontrado cerca de ciudades de antiguas civilizaciones.

[2]Se tratan de pozos excavados en el terreno y que han sido interpretados como reflejo de antiguas prácticas de almacenaje y manifestación de procesos agrícolas intensivos. En definitiva, son estructuras que se utilizaban para almacenar alimentos como el grano y también como un hogar de arcilla en el suelo.

[3]Un megalito es un monumento prehistórico realizado con uno o varios bloques de piedra, de gran tamaño y sin labrar.El término procede de las palabras griegas mega, grande y lithos, piedra. El adjetivo megalítico describe tales estructuras, cuya construcción se realizó con un sistema de enclavamiento que no utiliza mortero ni cemento.

También se lleva a cabo un examen de la distribución de los depósitos osteológicos pero en este caso en relación con el tipo de contexto: megalítico y no megalítico. Anteriormente ya se ha indicado que la población funeraria es mayor en contextos megalíticos (MNI=86) que en los no megalíticos (MNI=49). Ahora se analiza la distribución de los individuos en función del sexo y la edad a través de los dos tipos de contextos. Para ello se utiliza el mencionado test χ^2.

Los datos muestrales se adjuntan en las siguientes tablas de contingencia bidimensionales:

- Caso I: Sexo vs Tipo de Enterramiento (Megalítico/No Megalítico)

Enterramiento / Sexo	Megalítico	No Megalítico	
Femenino	13	9	22
Masculino	10	11	21
	23	20	43

Se aplica como hemos dicho un test chi-cuadrado para contrastar la independencia de ambas variables. Para ello, se utiliza el software R mediante la orden **chisq.test**, obteniéndose un p-valor=0.45 para un nivel de significación del 0.05 que nos indica que no se rechaza la hipótesis de independencia; y por lo tanto, parece ser que el tipo de enterramiento de aquella época y zona no dependía del sexo del individuo.

- Caso II: Edad vs Tipo de Enterramiento (Megalítico/No Megalítico)

Enterramiento / Edad	Megalítico	No Megalítico	
Adultos	23	29	52
Subadultos	6	3	9
	29	32	61

Se procede de forma análoga que en el caso anterior. En este caso, se obtiene un p-valor=0.2134, lo cual nos indica que tampoco se rechaza la hipótesis nula de independencia. Y por tanto, parece ser que la edad tampoco influía en el tipo de enterramiento.

Es conveniente destacar que como hemos indicado en la sección 2.3. el test chi-cuadrado de independencia es válido únicamente para tamaños muestrales grandes. En estos ejemplos, no podemos considerar que el tamaño muestral sea suficientemente grande para que la realización de un test chi-cuadrado nos de unos resultados aceptables. Para evitar este problema, aplicaremos un test exacto, en concreto, el test exacto de Fisher.

Nuevamente se usa el software R para su ejecución, en este caso utilizando la orden **fisher.test**. Para el caso Sexo vs Enterramiento se obtiene un p-valor=0.54, lo cual indica que el razonamiento anterior es correcto, es decir, ambas variables son independientes. Para el caso Edad vs Enterramiento se obtiene un p-valor=0.287, lo cual coincide con el razonamiento anteriormente dado cuando aplicamos el test chi-cuadrado. Es decir, la edad tampoco influye en el tipo de enterramiento.

Para ilustrar que en este caso no es adecuado el uso del test chi-cuadrado debido

a que se trata de un test asintótico y el tamaño muestral no es suficientemente grande, hemos buscado un contrajemplo en el cual ambos test den resultados diferentes. Para ello hemos tomado todas las tablas 2×2 posibles cambiando las frecuencias observadas pero sin modificar las frecuencias marginales totales, y posteriormente le hemos aplicado tanto el test chi-cuadrado como el test exacto de Fisher. Para la tabla con frecuencias por columnas 18,5,9,11 sale un p-valor de Fisher de 0.03162 y del de Chi-cuadrado de 0.053. En este caso el test de Fisher nos indica que rechazamos la hipótesis de independencia, mientras que el test chi-cuadrado nos indica que se acepta aunque sea por muy poco. Este resultado contradictorio, sugiere que a pesar de que para la tabla estudiada ambos test llegan a las mismas conclusiones, esto no siempre tiene por qué ser así, y deja entrever que en este caso el uso del test chi-cuadrado no es adecuado debido a que el tamaño muestral no es suficientemente grande, uno de los requisitos principales para la aplicación del mismo.

En consecuencia, en este estudio nos quedamos con los resultados del test exacto de Fisher el cual como hemos dicho parece indicar que el tipo de enterramiento no dependía ni del sexo ni de la edad del individuo.

En conjunto, estos análisis empíricos muestran que en el asentamiento prehistórico de Valencina de la Concepción, no parece haber un patrón de concentración de los restos osteológicos humanos que justifique retener la noción de un sector “funerario” opuesto a un sector “doméstico”. Las prácticas de enterramiento documentadas se extienden a lo largo de toda las zonas conocidas del asentamiento, sin agrupaciones espaciales estadísticamente discernibles en cuanto al número de individuos enterrados

(el tamaño de los depósitos osteológicos), o en cuanto a la distribución de la población según el sexo o la edad, independientemente de la morfología del depósito y del contexto funerario.

Análisis de los objetos metálicos

En relación con el segundo objetivo del estudio, que es, investigar el grado de especialización funcional del espacio, se lleva a cabo el siguiente análisis de los objetos metálicos hallados.

Con respecto al contexto de deposición de dichos objetos, el número de objetos encontrados en contextos funerarios es 68 (65 %), mientras en aquellos contextos considerados como "domésticos" se encontraron 37 (35 %). Estos últimos porcentajes están sin embargo influenciados fuertemente por las 29 puntas de jabalinas encontradas en La Pastora, de forma que si eliminamos estos objetos del recuento, entonces el número de objetos asociados a contextos funerarios es 39, y los dos recuentos serían casi iguales. Por lo tanto, debido al fuerte efecto cuantitativo que las puntas de jabalinas tienen en el total de la muestra en este estudio, y con el fin de mejorar el valor comparativo de los resultados, los tests estadísticos se llevaran a cabo dos veces, en un caso incluyéndolas y en otro excluyéndolas.

A continuación se estudia la distribución espacial de los objetos metálicos en función de sus categorías funcionales básicas. Los datos muestrales se recogen en la siguiente tabla de contingencia, la cual no contempla las 29 puntas de jabalinas.

En este caso, se aplica el test χ^2 para investigar si hay diferencias estadísticas significativas en la distribución de las clases de artefactos metálicos entre los distintos

Contexto / Categorías Funcionales	Doméstico	Funerario	
Herramientas	30	7	37
Armas-Herramientas	4	20	24
Adornos de oro	0	10	10
Indeterminados	3	2	5
	37	39	76

contextos. Aplicando en R la función **chisq.test** de la misma forma que anteriormente, se obtiene un p-valor=0.00104, lo cual indica que se rechaza la hipótesis de independencia, lo que quiere decir que ambas variables están asociadas, y por lo tanto, que la categoría funcional de los objetos depende del contexto en el que han sido encontrados. Esto sugiere que hay un patrón significativo por el cual los objetos clasificados como herramientas tienden a aparecer en contextos domésticos, mientras los objetos clasificados como armas-herramientas y adornos de oro tienden a aparecer en contextos funerarios.

Ahora incluimos en el estudio las 29 puntas de jabalinas de La Pastora, de forma que la siguiente tabla de contingencia recoge dichos datos muestrales.

En este caso, al aplicar el test χ^2 se obtiene un p-valor=0.0000157 lo cual también indica que se rechaza la hipótesis de independencia, mostrando nuevamente que hay una diferencia en la distribución de ambas categorías de artefactos.

En ambos casos, debemos comentar que aunque se ha aplicado un test chi-cuadrado, los datos no cumplían las condiciones necesarias para su aplicación que vimos en la parte teórica del capítulo, es decir no todas las frecuencias son mayores que 1 (vemos

Contexto / Categorías Funcionales	Doméstico	Funerario	
Herramientas	30	7	37
Armas-Herramientas	4	20	24
Adornos de oro	0	10	10
Armas	0	29	29
Indeterminados	3	2	5
	37	39	76

la presencia de 0 en los datos), y además no cumplen la condición de que el 80 % de las frecuencias sean mayores que 5. Como hemos procedido anteriormente, para contrastar la validez de nuestras conclusiones aplicaremos el test de Fisher a cada uno de los caso. Para el primer caso, obtenemos un p-valor=0.000686 lo cual confirma la asociación de las variables como comentábamos. En el segundo caso, se obtiene un p-valor=0.00000273 lo cual confirma el resultado obtenido mediante el test chi-cuadrado.

Por último, se estudia la distribución de los objetos de metal dentro de los contextos de enterramiento (megalítico vs. no megalítico). Los datos muestrales se muestran a continuación.

Al aplicar el test χ^2 a estos datos se obtiene un p-valor=0.59 lo cual indica que no se rechaza la hipótesis de independencia, es decir, que la distribución de las armas vs. las armas-herramientas no parecer ser estadísticamente diferente. Para estudiar la relación entre el tipo de objeto y el contexto en el que fueron hallados, se utiliza la

Objetos \ Contexto	Funerario megalítico	Funerario no megalítico
Herramientas	2	5
Armas-Herramientas	8	12

medida de asociación **Q de Yule** mencionada en la parte teórica del capítulo. Como comentamos su valor muestral viene dada por la expresión siguiente

$$\hat{Q} = \frac{n_{11}n_{22} - n_{12}n_{21}}{n_{11}n_{22} + n_{12}n_{21}}$$

En nuestro caso,

$$\hat{Q} = \frac{2 \cdot 12 - 5 \cdot 8}{2 \cdot 12 + 5 \cdot 8} = -0{,}25$$

Lo cual indica una asociación negativa, es decir, los objetos clasificados como herramientas tienen una mayor probabilidad de encontrarse en contextos no megalíticos que los objetos clasificados como armas-herramientas.

Por otro lado, sin embargo, si añadimos las categorías de adornos y armas, entonces si hay una diferencia estadística significativa, independientemente de si se incluyen las puntas de jabalinas (donde se obtiene un p-valor=0.0047) o no (obteniéndose un p-valor= 0.003392). Los datos muestrales usados quedan recogidos en la siguiente tabla.

En definitiva, el test χ^2 sugiere por una parte, una tendencia en la cual los objetos clasificados como herramientas aparecen más frecuentemente en contextos domésticos, mientras que los objetos clasificados como armas-herramientas y adornos tienden a aparecer en contextos funerarios. Esto en principio puede ser interpretado como una

Objetos \ Contexto	Funerario megalítico	Funerario no megalítico
Herramientas	2	5
Armas-Herramientas	8	12
Adornos de oro	10	0
Armas	0/29	0
Indeterminados	2	0

sugerencia de que algunos tipos de artefactos específicos se seleccionan como ajuares funerarios sobre los demás. Las indicaciones de un posible patrón subyacente en el sentido de que algunos tipos de artefactos fueron usados y/o depositados en contextos específicos sugiere la posibilidad de que algunos tipos de artefactos transmiten un mayor significado ideológico y sociológico que otros. Esto es admisible dentro del contexto de las sociedades ibéricas del 3rd y 2nd milenio donde ciertos objetos metálicos se valoran cada vez más como marcadores de estatus.

Por otro lado, el test aplicado a los objectos metálicos dentro del contexto funerario muestra que la distribución de las herramientas vs. las armas-herramientas no presenta diferencias estadísticamente significativas entre los contextos megalítico vs. no megalítico, lo cual sugiere que ser enterrado en una zanja o en un megalito no es diferente en términos de la probabilidad de que una persona utilize cierto tipo de herramienta como un ajuar funerario.

Conclusión

En conjunto, las evidencias demográficas y arqueo-metalúrgicas consideradas en este estudio nos invita a reorganizar las ideas sobre la organización espacial de la comunidad prehistórica de Valencina de la Concepción. Nos sugiere considerarlo mas que como un asentamiento con un espacio marcadamente dual, en el que un sector fue ocupado para la vida (sector "doméstico-productivo") y otro para la muerte (sector "funerario"), como un gran espacio de ocupación y uso en el cual varias funciones y actividades (productivas, domésticas y funerarias) se solapan, tanto en tiempo como en espacio.

Capítulo 3

ESTIMACION NO PARAMETRICA DE LA FUNCION DE DENSIDAD

"Si la gente no piensa que las matemáticas son simples,
es sólo porque no se dan cuenta de lo complicada que es la vida."
John von Neumann (1903-1957)

Resumen. En este capítulo se aborda una de las técnicas estadísticas cuantitativas más importantes dentro de la Arqueología. En concreto, se describirá el método de estimación núcleo de las funciones de densidad que generaliza a los histogramas como primera aproximación de las mismas. Por último, se aplicará todo lo anterior a un ejemplo real de las excavaciones llevadas a cabo en Pompeya.

3.1. Introducción

Es difícil concebir la estadística actual sin el concepto de distribución de probabilidad de una variable aleatoria, entendiéndolo como un modelo matemático que describe el comportamiento probabilístico de la misma. La representación matemática más tangible de la distribución de una variable aleatoria se corresponde con las denominadas funciones de distribución y de densidad de probabilidad de la variable aleatoria, íntimamente relacionadas entre sí. Conocer la función de densidad de una variable aleatoria implica tener una completa descripción de la misma. Es por tanto un problema fundamental de la estadística la estimación de la función de densidad de una variable o vector aleatorio a partir de la información proporcionada por una muestra.

Un posible enfoque consiste en considerar que la función de densidad que deseamos estimar pertenece a una determinada clase de funciones paramétricas, por ejemplo a algunas de las clásicas distribuciones: normal, exponencial, Poisson, etc. Dicha suposición usualmente se basa en informaciones sobre la variable que son externas a la muestra, pero cuya validez puede ser comprobada con posterioridad mediante pruebas de bondad de ajuste. Bajo esta suposición la estimación se reduce a determinar el valor de los parámetros del modelo a partir de la muestra. Esta estimación es la que denominaremos *estimación paramétrica* de la densidad. La posibilidad alternativa es no predeterminar a priori ningún modelo para la distribución de probabilidad de la variable y dejar que la función de densidad pueda adoptar cualquier forma, sin más límites que los impuestos por las propiedades que se exigen a las funciones de densidad para ser consideradas como tales. Este enfoque, en el que se centra el presente

capítulo, es el que denominaremos *estimación no paramétrica* de la densidad, y tiene uno de sus orígenes más comúnmente aceptado en los trabajos de [14] que buscaban una alternativa a las técnicas clásicas de análisis discriminante que permitiera liberarse de las rígidas restricciones sobre la distribución de las variables implicadas. En cierta manera el enfoque no paramétrico permite que los datos determinen de forma totalmente libre, sin restricciones, la forma de la densidad que los ha de representar.

La controversia sobre la utilización de una estimación paramétrica o no paramétrica no ha cesado a lo largo de los años. A la eficiencia en la estimación que proporciona la estimación paramétrica se contrapone el riesgo que suponen desviaciones de las suposiciones que determinan el modelo y que pueden conducir a errores de interpretación que supongan mayor pérdida que la ganancia proporcionada por la eficiencia estimadora.

Entre las principales situaciones en las cuales la estimación no paramétrica de la densidad ha resultado ser de especial interés podemos destacar:

- *Análisis Exploratorio*: Diversas características descriptivas de la densidad, tales como multimodalidad, asimetrías, comportamiento en las colas, etc., enfocadas desde un punto de vista no paramétrico, y por tanto más flexible, pueden ser más reveladoras y no quedar enmascaradas por suposiciones más rígidas.

- *Técnicas Multivariantes*: Estimaciones no paramétricas de la densidad son utilizadas en problemas de discriminación, clasificación, contrastes sobre modas, etc.

- *Regresión*: Estimaciones no paramétricas de la densidad permiten estimar la *Curva de Regresión de la Media*, que sabemos que es la que minimiza la esperanza del error cuadrático.

- *Representación de Datos*: La representación gráfica de los resultados obtenidos en una estimación no paramétrica de la densidad es fácilmente comprensible e intuitivo para aquellas personas no especialistas en estadística que muy a menudo son los clientes de los servicios de estadística.

En esta última aplicación es en la cual centraremos el desarrollo del capítulo, haciendo hincapié en los métodos no paramétricos que suelen incluir algún tipo de aproximación o método de suavización (del inglés,*smoothing*), en particular, en los llamados métodos Núcleos. Estos métodos están normalmente indexados por un parámetro llamado *bandwidth, ancho de ventana o parámetro de suavización* que controla el grado de complejidad de los mismos. La elección de dicho parámetro es a menudo crucial para implementación del método. Los métodos no paramétricos que requieren de estos parámetros pero no tienen una regla de dependencia de datos explícita para su selección, son incompletos. Desafortunadamente, esto es bastante común, debido a la dificultad en el desarrollo de reglas rigurosas para la selección del ancho de ventana. A menudo en estos casos, el ancho de ventana es seleccionado basándose en un problema estadístico relacionado. Esto es factible pero un compromiso preocupante.

Una idea intuitiva de como estimar la función de densidad de una función a partir de una muestra es la siguiente:

Sea X una variable aleatoria con distribución continua $F(x)$ y densidad $f(x) = \frac{d}{dx}F(x)$. El objetivo es estimar $f(x)$ a partir de una muestra aleatoria $x_1, ..., x_n$.

La función de distribución $F(x)$ se estima naturalmente a través de la Función de Distribución Empírica (FDE) $\hat{F}(x) = n^{-1}I_A(x)$ donde $A = \{x : x_i \leq x\}$. Podría parecer natural estimar la densidad $f(x)$ como la derivada de $\hat{F}(x)$ pero este estimador sería un conjunto de puntos de probabilidad, no una función de densidad, y como tal no es un estimador útil de $f(x)$.

En su lugar, consideramos una derivada discreta. Para algún $h > 0$ pequeño,

$$\hat{f}(x) = \frac{\hat{F}(x+h) - \hat{F}(x-h)}{2h}.$$

Podemos escribir esto como

$$\frac{1}{2nh}\sum_{i=1}^{n} I_B(x) = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x_i - x}{h}\right),$$

donde $B = \{x : x + h \leq x_i \leq x + h\} = \{x : \frac{|x_i - x|}{h} \leq 1\}$ y

$$k(u) = \begin{cases} \frac{1}{2} & si \quad |u| \leq 1; \\ \\ 0 & si \quad |u| > 1, \end{cases}$$

es la función de densidad uniforme en [-1,1].

El estimador $\hat{f}(x)$ cuenta el porcentaje de observaciones que están cerca del punto x. Si muchas observaciones están cerca de x, entonces $\hat{f}(x)$ es grande. Por el contrario, si solo unas cuantas x_i están cerca de x, entonces $\hat{f}(x)$ es pequeño. El ancho de ventana h controla el grado de suavidad de la estimación.

$\hat{f}(x)$ es un caso especial de lo que llamamos estimador núcleo.

A continuación veremos en detalle la teoría en la que se basa el método Núcleo tanto para estimaciones de densidad univariante como para el caso multivariante. También abordaremos el problema de la elección del parámetro ancho de ventana proponiendo varios métodos para su cálculo, y analizaremos cómo el uso de un tipo de ancho de ventana u otro repercute en la estimación de la densidad. Antes de abordar de lleno la teoría del método Núcleo, comentaremos previamente una técnica más clásica en la estimación de la densidad, dicha técnica no es otra que el Histograma. Una vez analizadas ambas técnicas realizaremos una breve comparación entre ambas para que quede reflejado el por qué actualmente el uso del los métodos Núcleos frente al del histograma se prefiere cuando tratamos con grandes cantidades de datos y queremos obtener unos resultados más sofisticados.

Por último, ilustraremos como estas técnicas se pueden aplicar en estudios arqueológicos. En este capítulo, el estudio se llevará a cabo sobre artefactos, más concretamente piezas de cerámicas de telares, que fueron hallados en Ínsula VI,1, Pompeya. El objetivo de esta última parte es mostrar como los métodos estadísticos vistos a lo largo del capítulo nos permiten confirmar la validez de las observaciones que los arqueólogos encontraron en dicho yacimiento.

3.2. Estimación no paramétrica de la densidad

3.2.1. Del humilde histograma y sus virtudes

Es el más sencillo y mejor conocido de los estimadores no paramétricos de la densidad. Muchos autores distinguen la utilización del histograma como técnica de representación de datos o como estimador de la densidad, la diferencia básica es que

en este último caso debe estar normalizado.

Supongamos que f tiene soporte en $[a, b]$ generalmente deducido de los datos, efectuamos una partición en k intervalos no solapados $B_i = [t_i, t_{i+1})$ $i = 1, ..., k$ donde $a = t_1 < t_2 < ... < t_{k+1} = b$, el histograma viene definido por

$$\hat{f}(x) = \sum_{i=1}^{k} \frac{N_i/n}{t_{i+1} - t_i} I_{B_i}(x),$$

donde N_i es el número de datos dentro de B_i. Si la longitud de los intervalos es siempre la misma $h_n = t_{i+1} - t_i$, valor que denominaremos anchura del intervalo o *ancho de ventana*, la expresión resulta

$$\hat{f}(x) = \frac{1}{nh_n} \sum_{i=1}^{k} N_i I_{B_i}(x),$$

o en forma equivalente

$$\hat{f}(x) = \frac{1}{nh_n} \sum_{i=1}^{n} I_{B_i}(x) = \frac{N_i}{nh_n} \quad x \in B_i.$$

Si la longitud h_n del intervalo $[t_i, t_{i+1})$ tiende a cero cuando el número de datos tiende a infinito, cabe esperar que $\hat{f}(x)$ tienda hacia la "densidad instantánea" en el punto x que es precisamente la función de densidad. Hay que añadir solamente que h_n no debe tender a cero demasiado deprisa, para evitar quedarnos sin datos en muchos intervalos. De hecho, la condición que se requiere para que se produzca la convergencia, $\hat{f}(x) \xrightarrow[n\to\infty]{} f(x)$, es $nh_n \to \infty$, además de $h_n \to 0$.

Obsérvese que la amplitud h_n de los intervalos es elegida por el usuario y, en cierto modo, es arbitraria (aunque hay algunos criterios razonables para elegirla que no

discutiremos aquí). El aspecto del histograma podría cambiar considerablemente si este valor se cambia.

3.2.2. Los estimadores núcleos: una versión más sofisticada de los histogramas

Los histogramas pueden resultar útiles e ilustrativos para muchos propósitos pero son decididamente inadecuados bajo otros puntos de vista. En concreto:

- Los histogramas son siempre, por naturaleza, funciones discontinuas, sin embargo, en muchos casos es razonable suponer que la función de densidad de la variable que se está estimando es continua. En este sentido, los histogramas son estimadores insatisfactorios.

- Como los histogramas son funciones constantes a trozos, su primera derivada es cero en casi todo punto. Esto los hace completamente inadecuados para estimar la derivada de la función de densidad.

- Parcialmente relacionado con el punto anterior está el hecho de que los histogramas no son tampoco adecuados para estimar las modas (si se define **moda** como un máximo relativo de la función de densidad). A lo sumo, pueden proporcionar "intervalos modales", pero esto puede resultar demasiado burdo en casos en que se requiere mayor precisión.

Los estimadores de tipo **núcleo** (del inglés *kernel*) fueron diseñados para superar estas dificultades. La idea original es bastante antigua y se remonta a los trabajos de [28] y [22] en los años 50 y primeros 60. Los estimadores núcleo son, sin duda,

los más utilizados y mejor estudiados en la teoría no paramétrica. Antes de entrar a describir en qé consiste dicho método de estimación vamos a definir en primer lugar el concepto de función núcleo.

Definición 3.2.1. *Una función núcleo* $k(u) : \mathbb{R} \to \mathbb{R}$ *es una función que satisface*

$$\int_{-\infty}^{+\infty} k(u)\, du = 1.$$

Un núcleo se dice no negativo si verifica que $k(u) \geq 0\ \forall u \in [-1, 1]$. En este caso es una función de densidad. Los momentos de un núcleo se definen como

$$\kappa_j(k) = \int_{-\infty}^{+\infty} u^j k(u)\, du.$$

Una función núcleo simétrica satisface $k(u) = k(-u)\ \forall$u. En este caso, todos los momentos impares son cero. La mayoría de estimaciones no paramétricas usan núcleos simétricos, y aquí nos centraremos precisamente en esos casos.

El orden de un núcleo, ν, está definido como el orden del primer momento no nulo. Por ejemplo, si $\kappa_1(k) = 0$ y $\kappa_2(k) > 0$ entonces k es un núcleo de segundo orden y $\nu = 2$. Si $\kappa_1(k) = \kappa_2(k) = \kappa_3(k) = 0$ pero $\kappa_4(k) > 0$ entonces k es un núcleo de cuarto orden y $\nu = 4$. El orden de un núcleo simétrico es siempre par.

Los núcleos simétricos no negativos son núcleos de segundo grado. Un núcleo se dice que es de orden superior si $\nu > 2$. Esos núcleos tendrían partes negativas y no son densidades de probabilidad. Son también denominados como núcleos de sesgo reducido.

Los núcleos de segundo orden más comunes están anotados en la siguiente tabla

Núcleo	Ecuación	$R(k)$	$\kappa_2(k)$	$eff(k)$
Uniforme	$k_0(u) = \frac{1}{2} I_A(u)$	$1/2$	$1/3$	1,0758
Epanechnikov	$k_1(u) = \frac{3}{4}(1-u^2) I_A(u)$	$3/5$	$1/5$	1,0000
Biweight	$k_2(u) = \frac{15}{16}(1-u^2)^2 I_A(u)$	$5/7$	$1/7$	1,0061
Triweight	$k_3(u) = \frac{35}{32}(1-u^2)^3 I_A(u)$	$350/429$	$1/9$	1,0135
Gaussiano	$k_\phi(u) = \frac{1}{\sqrt{2\pi}} exp\left(\frac{-u^2}{2}\right)$	$1/2\sqrt{\pi}$	1	1,0513

Tabla 3.1: Núcleos de segundo orden más comunes.

Además de la fórmula del núcleo se ha anotado su aspereza $R(k)$ (del inglés *roughness*), el segundo momento $\kappa_2(k)$, y su eficiencia $eff(k)$, esta última propiedad será definida en secciones posteriores. La aspereza de una función se define como

$$R(g) = \int_{-\infty}^{+\infty} g(u)^2 \, du.$$

Los núcleos más usados habitualmente son el de Epanechnikov y el Gaussiano.

Para la propuesta de estimación no paramétrica, la escala del núcleo no está definida de forma única. Esto es, para cualquier núcleo $k(u)$ podríamos definir un núcleo alternativo

$$k^*(u) = \frac{1}{b} k\left(\frac{u}{b}\right), \text{ para algún } b > 0.$$

Estos dos núcleos son equivalentes en el sentido que producen el mismo estimador de densidad, es decir, si $\hat{f}(x)$ está calculado con el núcleo k y el ancho de ventana h, es numéricamente idéntico con uno calculado con el núcleo k^* y el ancho de ventana

$h^* = \dfrac{h}{b}$. Algunos autores usan diferentes definiciones para los mismos núcleos. El ancho de ventana de un núcleo es la semi-amplitud del núcleo en el intervalo de interés y por tanto el que controla el grado de suavidad.

El Método Núcleo

Un núcleo es una función de densidad. Si se coloca un núcleo en cada uno de los datos de la muestra, la suma ponderada de estas funciones también será una función de densidad. Esta suma es una función continua que suaviza el perfil de la distribución captando la influencia de los datos cercanos y constituye el estimador $\hat{f}(x)$ del modelo teórico del cual provienen los datos, permitiendo observar diferencias que los rectángulos del histograma no puede mostrar.

Así, sea $k(x) = \dfrac{1}{h}k\left(\dfrac{x - x_i}{h}\right),\ \ i = 1, 2, ..., n$. k es una función de densidad. Ahora si se multiplica cada núcleo por $1/n$, entonces la suma de los n núcleos también será una función de densidad. De esta forma, sea X una variable aleatoria con distribución de probabilidad continua, univariada y desconocida $f(x)$ de la cual se dispone de una muestra de n observaciones independientes $x_1, x_2, ..., x_n$, definimos el estimador por núcleos $\hat{f}(x)$ como

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} k(x)dx = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x - x_i}{h}\right), \tag{3.2.1}$$

siendo h el parámetro de ajuste o suavizado de $\hat{f}(x)$. La elección de dicho parámetro será crítica para el modelo y la analizaremos más adelante.
Mientras más pequeño es h, más concentrada está la contribución del núcleo en cada punto x_i. Mientras más grande es h, mayor es la influencia e interacción del núcleo

hacia los puntos vecinos. Si el ancho de ventana se elige demasiado pequeño, el estimador aparece "infrasuavizado", e incorpora demasiado "ruido", reflejado en la presencia de muchas modas (máximos relativos) que, de hecho no aparecen en la densidad que se quiere estimar. Por el contrario, si h se elige demasiado grande, se da el fenómeno contrario, de "sobresuavización" y el estimador es casi insensible a los datos.

Una de las principales aplicaciones prácticas de los estimadores núcleo es su utilidad para estimar las modas y el número de modas. Es curioso notar a este respecto que, en las primeras aproximaciones elementales a la Estadística, se suele hablar de **media**, **mediana**, **moda** como medidas de tendencia central, pero posteriormente en los cursos universitarios de Estadística y Probabilidad, la moda desaparece casi de escena. La razón de eso tiene que ver quizás con el hecho de que en los modelos paramétricos usuales, el número de modas aparece fijado de antemano desde el momento en que se elige el modelo (así, la distribución normal es unimodal) y, en muchos casos, la moda coincide necesariamente con media (de nuevo, la normal proporciona un ejemplo claro de esta situación). Por otra parte, la definición formal de moda de una variable aleatoria (y sobre todo su cálculo) resulta más "escurridiza" que la de la media. Si se define, como parece natural, la moda como un máximo local de la densidad, no resulta muy claro, si no se dispone de estimadores de densidad, como puede estimarse una moda a partir de una muestra. La utilización de estimadores de tipo núcleo proporciona una forma muy natural de estimar este parámetro: se define una **moda muestral** como un máximo local de un estimador núcleo $\hat{f}$ de la densidad poblacional f. En definitiva, los estimadores no paramétricos de la densidad proporcionan un marco natural para "rehabilitar" la noción de moda que resulta tan intuitiva y útil en un análisis estadístico. Los estimadores de densidad no fijan de antemano el número

de modas, como ocurre con los modelos paramétricos. Como ya hemos indicado antes, los enfoques no paramétricos tienen la ventaja de que "dejan hablar a los datos" y no prejuzgan de antemano algunas características importantes de los mismos, como ocurre frecuentemente con los modelos paramétricos.

3.3. Propiedades de los estimadores de densidad

En esta sección discutiremos algunas de las propiedades numéricas de los estimadores núcleos de la densidad

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}k\left(\frac{x - x_i}{h}\right),$$

vista como una función de x. Sea X una variable aleatoria con función de densidad $\hat{f}(x)$.

Primero, si $k(u)$ es no negativa, entonces es fácil ver que $\hat{f}(x) \geq 0$. Sin embargo, esto no se garantiza si k es un núcleo de orden superior. En este caso, es posible que $\hat{f}(x) < 0$ para algunos valores de x. Cuando esto ocurre es prudente quitar los valores no negativos y entonces reescalar:

$$\tilde{f}(x) = \frac{\hat{f}(x)I_C(x)}{\int_{-\infty}^{+\infty}\hat{f}(x)I_C(x)dx},$$

donde $C = \{x : \hat{f}(x) \geq 0\}$.

$\tilde{f}(x)$ es no negativa y tiene las mismas propiedades asintóticas que $\hat{f}(x)$. Dado que la integral del denominador no es realizable analíticamente, se debe calcular numéricamente.

Otras propiedades que cumplen los estimadores de densidad son las siguientes:

Proposición 3.3.1. *$\hat{f}(x)$ es una función de densidad cuando k es no negativa.*

Demostración. Para ver esto, primero notamos que por el cambio de variables $u = (x - x_i)/h)$, el cual tiene jacobiano h,

$$\int_{-\infty}^{+\infty} \frac{1}{h} k\left(\frac{x - x_i}{h}\right) dx = \int_{-\infty}^{+\infty} k(u)du = 1.$$

El cambio de variables anterior se usa frecuentemente, por lo que es útil estar familiarizado con esta transformación. Por lo tanto,

$$\int_{-\infty}^{+\infty} \hat{f}(x)dx = \int_{-\infty}^{+\infty} \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} k\left(\frac{x - x_i}{h}\right) dx = \frac{1}{n}\sum_{i=1}^{n} \int_{-\infty}^{+\infty} k(u)du = \frac{1}{n}\sum_{i=1}^{n} 1 = 1.$$

□

Proposición 3.3.2. *El valor esperado de la variable aleatoria X con densidad $\hat{f}(x)$ coincide con la media muestral, independientemente del núcleo k.*

Demostración. Para probar esto, volvemos a realizar el cambio de variables $u = (x - x_i)/h$, y tenemos en cuenta que $\int_{-\infty}^{+\infty} uk(u)du = 0$ por ser k simétrica. De esta

forma, el primer momento, es decir, la media o valor esperado de X es

$$\begin{aligned}
E[X] &= \int_{-\infty}^{+\infty} x\hat{f}(x)dx \\
&= \int_{-\infty}^{+\infty} \frac{1}{n}\sum_{i=1}^{n} x\frac{1}{h}k\left(\frac{x-x_i}{h}\right)dx \\
&= \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{+\infty}(x_i+uh)k(u)du \\
&= \frac{1}{n}\sum_{i=1}^{n} x_i\int_{-\infty}^{+\infty}k(u)du + \frac{1}{n}\sum_{i=1}^{n} h\int_{-\infty}^{+\infty}uk(u)du \\
&= \frac{1}{n}\sum_{i=1}^{n} x_i \\
&= \bar{x}.
\end{aligned}$$

□

Proposición 3.3.3. *La varianza de la densidad $\hat{f}(x)$ es $\hat{S}^2 + h^2\kappa_2(k)$ donde $\hat{S}^2$ es la varianza muestral.*

Demostración. El segundo momento de la densidad estimada es

$$\begin{aligned}
\int_{-\infty}^{+\infty} x^2\hat{f}(x)dx &= \int_{-\infty}^{+\infty} \frac{1}{n}\sum_{i=1}^{n} x^2\frac{1}{h}k\left(\frac{x-x_i}{h}\right)dx \\
&= \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{+\infty}(x_i+uh)^2k(u)du \\
&= \frac{1}{n}\sum_{i=1}^{n} x_i^2\int_{-\infty}^{+\infty}k(u)du + \frac{2}{n}\sum_{i=1}^{n} x_ih\int_{-\infty}^{+\infty}uk(u)du + \frac{1}{n}\sum_{i=1}^{n} h^2\int_{-\infty}^{+\infty}u^2k(u)du \\
&= \frac{1}{n}\sum_{i=1}^{n} x_i + h^2\kappa_2(k).
\end{aligned}$$

Se deduce que la varianza de la densidad $\hat{f}(x)$ es

$$\begin{aligned}\hat{\sigma}^2 &= E[X^2] - E^2[X] \\ &= \int_{-\infty}^{+\infty} x^2 \hat{f}(x)dx - (\int_{-\infty}^{+\infty} x\hat{f}(x)dx)^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} x_i + h^2\kappa_2(k) - (\frac{1}{n}\sum_{i=1}^{n} x_i)^2 \\ &= \hat{S}^2 + h^2\kappa_2(k).\end{aligned}$$

donde $\hat{S}^2$ es la varianza muestral. La varianza $\hat{\sigma}^2$ depende linealmente del factor $\kappa_2(k)$, pero cuadráticamente del ancho de banda h, por ello este es el factor crítico.

□

3.4. Eficiencia del estimador

Sean $f(x)$ la función de densidad teórica (desconocida) y $\hat{f}(x)$ el estimador de $f(x)$ basado en los datos y el núcleo elegido.

Es útil observar que las esperanzas de las transformaciones del núcleo pueden escribirse como integrales que toman la forma de una covolución del núcleo y de la función de densidad

$$E\left[\frac{1}{h}k\left(\frac{x-x_i}{h}\right)\right] = \int_{-\infty}^{+\infty} \frac{1}{h}k\left(\frac{x-z}{h}\right) f(z)dz,$$

usando el cambio de variables $u = \dfrac{x-z}{h}$, esto equivale a

$$\int_{-\infty}^{+\infty} k(u)f(x-hu)du.$$

Por la linealidad del estimador vemos que

$$E[\hat{f}(x)] = \frac{1}{n}\sum_{i=1}^{n} E\left[\frac{1}{h}k\left(\frac{x-x_i}{h}\right)\right] = \int_{-\infty}^{+\infty} k(u)f(x-hu)du.$$

La integral obtenida no es resoluble analíticamente, por lo tanto la aproximamos usando un desarrollo de Taylor de $f(x-hu)$ en el argumento hu, el cual es válido cuando $h \to 0$. Para un núcleo de orden ν tomamos el desarrollo hacia el término ν,

$$f(x-hu) = f(x) - f^{(1)}(x)hu + \frac{1}{2}f^{(2)}(x)h^2u^2 - \frac{1}{3!}f^{(3)}(x)h^3u^3 + ... + \frac{1}{\nu!}f^{(\nu)}(x)h^\nu u^\nu + O(h^\nu).$$

El resto es de orden menor que h^ν cuando $h \to \infty$, el cual se escribe como $O(h^\nu)$. (Este desarrollo asume la existencia de $f^{(\nu+1)}(x)$). Integrando término a término y usando $\int_{-\infty}^{+\infty} k(u)du = 1$ y la definición $\int_{-\infty}^{+\infty} k(u)u^j du = \kappa_j(k)$, obtenemos

$$\begin{aligned}\int_{-\infty}^{+\infty} k(u)f(x-hu)du &= f(x) - f^{(1)}(x)h\kappa_1(k) + \frac{1}{2}f^{(2)}(x)h^2\kappa_2(k) - ... \\ &\quad + \frac{1}{\nu!}f^{(\nu)}(x)h^\nu\kappa_\nu(k) + O(h^\nu) \\ &= f(x) + \frac{1}{\nu!}f^{(\nu)}(x)h^\nu\kappa_\nu(k) + O(h^\nu),\end{aligned}$$

donde la segunda igualdad usa la suposición que k es un núcleo de orden ν (por tanto $\kappa_j(k) = 0 \,\forall\, j < \nu$).

Esto significa que

$$E[\hat{f}(x)] = f(x) + \frac{1}{\nu!}f^{(\nu)}(x)h^\nu\kappa_\nu(k) + O(h^\nu).$$

3.4.1. Sesgo del estimador

Teniendo en cuenta el razonamiento anterior, concluimos que el sesgo del estimador $\hat{f}(x)$ es

$$\text{Sesgo}(\hat{f}(x)) = E[\hat{f}(x)] - f(x) = \frac{1}{\nu!} f^{(\nu)}(x) h^{\nu} \kappa_{\nu}(k) + O(h^{\nu}).$$

Para núcleos de segundo orden se simplifica a

$$\text{Sesgo}(\hat{f}(x)) = \frac{1}{2} f^{(2)}(x) h^{2} \kappa_{2}(k) + O(h^{2}).$$

En este caso, a medida que aumenta el cuadrado del ancho de banda aumenta el sesgo, por el contrario, valores pequeños del ancho de banda implicaran sesgos reducidos. El sesgo es también proporcional a la segunda derivada de la densidad $f^{(2)}(x)$. Intuitivamente, cuando $x_i = x$ el estimador $\hat{f}(x)$ suaviza los datos locales, por lo tanto es una estimación de la versión suavizada de $f(x)$. El sesgo resulta de este suavizado, y es más grande que la curvatura mayor en $f(x)$.

Cuando usamos núcleos de orden superior (y la densidad tiene suficientes derivadas), el sesgo es proporcional a h^{ν}, el cual es de un orden menor que h^2. Por lo tanto, el sesgo de los estimadores que usan núcleos de orden superior es de menor orden que los estimadores que usan núcleos de segundo orden, y es por esto que son llamados núcleos de reducción de sesgo. Esta es la ventaja de los núcleos de orden superior.

3.4.2. Varianza del estimador

Teniendo en cuenta las propiedades de la varianza siguientes:

- Sea X una variable aleatoria y a un número real cualesquiera, $Var[aX] = a^2 Var[X]$.

- Sean $X_1, ..., X_n$ variables aleatorias independientes e idéticamente distribuidas, $Var[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} Var[X_i]$.

se tiene que

$$\begin{aligned} Var[\hat{f}(x)] &= Var\left[\frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x-x_i}{h}\right)\right] \\ &= \frac{1}{n^2h^2}\sum_{i=1}^{n} Var\left[k\left(\frac{x-x_i}{h}\right)\right], \end{aligned}$$

donde

$$\begin{aligned} Var\left[k\left(\frac{x-x_i}{h}\right)\right] &= E\left[k^2\left(\frac{x-x_i}{h}\right)\right] - E\left[k\left(\frac{x-x_i}{h}\right)\right]^2 \\ &= \int_{-\infty}^{+\infty} k^2\left(\frac{x-z}{h}\right) f(z)dz - \left(\int_{-\infty}^{+\infty} k\left(\frac{x-z}{h}\right) f(z)dz\right)^2. \end{aligned}$$

Por lo tanto,

$$\begin{aligned} Var[\hat{f}(x)] &= \frac{1}{n^2h^2}\sum_{i=1}^{n}\int_{-\infty}^{+\infty} k^2\left(\frac{x-z}{h}\right) f(z)dz - \frac{1}{n^2h^2}\sum_{i=1}^{n}\left(\int_{-\infty}^{+\infty} k\left(\frac{x-z}{h}\right) f(z)dz\right)^2 \\ &= \frac{1}{n}\int_{-\infty}^{+\infty}\frac{1}{h^2}k^2\left(\frac{x-z}{h}\right) f(z)dz - \frac{1}{n}\left(\int_{-\infty}^{+\infty}\frac{1}{h}k\left(\frac{x-z}{h}\right) f(z)dz\right)^2 \\ &= \frac{1}{nh^2}E\left[k^2\left(\frac{x-x_i}{h}\right)\right] - \frac{1}{n}\left(\frac{1}{h}E\left[k\left(\frac{x-x_i}{h}\right)\right]\right)^2. \end{aligned}$$

De nuestro análisis del sesgo conocemos que $\frac{1}{h}E\left[k\left(\frac{x-x_i}{h}\right)\right] = f(x) + O(1)$, luego el segundo término es $O(n^{-1})$. Para el primer término procedemos de igual forma que al comienzo de esta sección, escribimos la esperanza como una integral, hacemos un cambio de variables y un desarrollo de Taylor de primer orden, y de esta forma obtenemos que

$$\begin{aligned}\frac{1}{h}E\left[k^2\left(\frac{x-x_i}{h}\right)\right] &= \frac{1}{h}\int_{-\infty}^{+\infty} k^2\left(\frac{x-z}{h}\right) f(z)dz \\ &= \int_{-\infty}^{+\infty} k(u)^2 f(x+hu)du \\ &= \int_{-\infty}^{+\infty} k(u)^2 (f(x)+O(h))du, \\ &= f(x)R(x) + O(h)\end{aligned}$$

donde $R(x) = \int_{-\infty}^{+\infty} k(u)^2 du$ es la aspereza del núcleo. Por lo tanto,

$$var[\hat{f}(x)] = \frac{f(x)R(x)}{nh} + O\left(\frac{1}{n}\right).$$

El resto $O(n^{-1})$ es de orden más pequeño que el resto $O(nh^{-1})$ del término principal, dado que $h^{-1} \to \infty$. Según este resultado, $Var[\hat{f}(x)]$ aumenta si h se reduce.

3.4.3. Error cuadrático medio (MSE)

Una común y conveniente medida de precisión de estimación es el error cuadrático medio (MSE). El MSE de un estimador, $\hat{f}(x)$, se define como

$$\text{MSE}(\hat{f}(x)) = E[\hat{f}(x) - f(x)]^2,$$

y para el caso particular del estimador núcleo, su cálculo sería el siguiente

$$\begin{aligned}
\text{MSE}(\hat{f}(x)) &= E[\hat{f}(x) - f(x)]^2 \\
&= \text{Sesgo}(\hat{f}(x))^2 + \text{Var}[\hat{f}(x)] \\
&\simeq \left(\frac{1}{\nu!} f^{(\nu)}(x) h^{\nu} \kappa_{\nu}(k)\right)^2 + \frac{f(x)R(x)}{nh} \\
&= \frac{\kappa_{\nu}^2(k)}{(\nu!)^2} f^{(\nu)}(x)^2 h^{2\nu} + \frac{f(x)R(x)}{nh} \\
&= \text{AMSE}(\hat{f}(x)).
\end{aligned}$$

Dado que esta aproximación se basa en desarrollos asintóticos recibe el nombre de error cuadrático medio asintótico (AMSE). Notar que el primer término (el sesgo al cuadrado) aumenta en h y el segundo término (la varianza) decrece en nh. Para que $\text{MSE}(\hat{f}(x))$ disminuya cuando $n \to \infty$ esos términos deben ser pequeños. Por lo tanto, cuando $n \to \infty$ debemos hacer $h \to 0$ y $nh \to \infty$. Esto es, el ancho de ventana debe disminuir, pero no en una proporción más rápida que el tamaño muestral. Esto es suficiente para establecer la consistencia puntual del estimador. Esto es, $\forall x, \hat{f}(x) \to_p f(x)$ cuando $n \to \infty$. Donde la convergencia puntual se ha denotado como $\to_p$.

Una medida de precisión global es el error cuadrático integrado por la media asintótica (AMISE). El AMISE de un estimador, $\hat{f}(x)$, se define como

$$AMISE = \int_{-\infty}^{+\infty} AMSE(\hat{f}(x))dx,$$

y para el caso particular del estimador núcleo, la expresión resultante sería la siguiente

$$AMISE = \int_{-\infty}^{+\infty} AMSE(\hat{f}(x))dx = \frac{\kappa_{\nu}^2(k)}{\nu!} R(f^{(\nu)}) h^{2\nu} + \frac{R(k)}{nh},$$

donde $R(f^{(\nu)}) = \int_{-\infty}^{+\infty}(f^{(\nu)}(x))^2 dx$ es la aspereza de $f^{(\nu)}$.

3.4.4. Consistencia del estimador

Teniendo en cuenta que la función núcleo es una función simétrica y acotada que verifica las siguientes propiedades

$$\int_{-\infty}^{+\infty} |k(x)|dx < \infty \tag{3.4.1}$$

$$\lim_{x\to\infty} |xk(x)| = 0 \tag{3.4.2}$$

$$\int_{-\infty}^{+\infty} k(x)dx = 1 \tag{3.4.3}$$

vamos a analizar la consistencia del estimador núcleo. Para ello es necesario presentar un par de resultados previos.

Teorema 3.4.1 ([3] Bochner, 1995)**.** *Sea k(z) una función Borel acotada que satisface las condiciones (3.4.1) y (3.4.2). Sea $g \in \mathcal{L}^1$. Sea*

$$g_n(x) = \frac{1}{h_n}\int_{-\infty}^{+\infty} k\left(\frac{z}{h}\right) g(x-z)dz, \tag{3.4.4}$$

donde h_n es una secuencia de constantes positivas que satisfacen $\lim_{n\to\infty} h_n = 0$. Entonces si x es un punto de continuidad de g,

$$\lim_{n\to\infty} g_n(x) = g(x)\int_{-\infty}^{+\infty} k(z)dz. \tag{3.4.5}$$

Demostración. Notemos en primer lugar que

$$g_n(x) - g(x)\int_{-\infty}^{+\infty} k(z)dz = \int_{-\infty}^{+\infty}(g(x-z)-g(x))\frac{1}{h_n}k\left(\frac{z}{h}\right)dz.$$

Sea ahora $\delta > 0$, y dividamos el dominio de integración en dos regiones, $|z| \leq \delta$ y $|y| > \delta$. Entonces,

$$\begin{aligned}
|g_n(x) - g(x)\int_{-\infty}^{+\infty} k(z)dz| &\leq \sup_{|z|\leq\delta} |g(x-z) - g(x)| \int_{|y|\leq\frac{\delta}{h_n}} |k(y)|dy \\
&\quad + \int_{|z|\geq\delta} \frac{|g(x-z)|}{z} \frac{z}{h_n} k\left(\frac{z}{h}\right) dz + |g(x)| \int_{|z|\geq\delta} \frac{1}{h_n} k\left(\frac{z}{h}\right) dz \\
&\leq \sup_{|z|\leq\delta} |g(x-z) - g(x)| \int_{-\infty}^{+\infty} |k(y)|dy \\
&\quad + \frac{1}{\delta} \sup_{|y|\geq\frac{\delta}{h_n}} |yk(y)| \int_{-\infty}^{+\infty} |g(z)|dz + |g(x)| \int_{|y|\geq\frac{\delta}{h_n}} |k(y)|dy.
\end{aligned}$$

Cuando $n \to \infty$, debido a que $h_n \to 0$, el segundo y tercer término tienden a cero, ya que $g \in \mathcal{L}^1$ y $\lim_{x\to\infty} |xk(x)| = 0$. Haciendo entonces $\delta \to 0$, el primer término tiende a cero debido a que $k \in \mathcal{L}^1$ y a que x es un punto de continuidad de g. □

Teniendo ahora en cuenta que

$$\begin{aligned}
E[\hat{f}(x)] &= \frac{1}{n}\sum_{i=1}^{n} E\left[\frac{1}{h_n} k\left(\frac{x-x_i}{h}\right)\right] \\
&= E\left[\frac{1}{h_n} k\left(\frac{x-z}{h}\right)\right] = \int_{-\infty}^{+\infty} \frac{1}{h_n} k\left(\frac{x-z}{h}\right) f(z)dz,
\end{aligned}$$

del teorema anterior se deduce el siguiente Corolario:

Corolario 3.4.1. *El estimador $\hat{f}(x)$ definido en (3.2.1) es asintóticamente insesgado en todos los puntos x en los cuales la función de densidad es continua si las constantes h_n satisfacen $\lim_{n\to\infty} h_n = 0$ y si la función k satisface las propiedades (3.4.1), (3.4.2) y (3.4.3).*

Con estos resultados podemos pasar a demostrar la consistencia del estimador tipo núcleo,

Teorema 3.4.2. *El estimador $\hat{f}_n(x)$ definido en (3.2.1) es consistente, es decir $MSE[\hat{f}_n(x)] \to 0\ \forall x \in \mathbb{R}$ cuando $n \to \infty$, si añadimos la condición adicional de que $\lim_{n\to\infty} nh_n = \infty$.*

Demostración. En efecto, tengamos en cuenta que

$$Var[\hat{f}(x)] = \frac{1}{n} Var\left[\frac{1}{h} k\left(\frac{x-z}{h}\right)\right].$$

Además

$$\begin{aligned}\frac{1}{n} Var\left[\frac{1}{h} k\left(\frac{x-z}{h}\right)\right] &\leq \frac{1}{n} E\left[\left(\frac{1}{h} k\left(\frac{x-z}{h}\right)\right)^2\right] \\ &= \frac{1}{nh}\left[\frac{1}{h}\int_{-\infty}^{+\infty}\left(k\left(\frac{x-z}{h}\right)\right)^2 f(z)dz,\right]\end{aligned}$$

y por el Teorema 3.4.1

$$\frac{1}{h}\int_{-\infty}^{+\infty}\left(k\left(\frac{x-z}{h}\right)\right)^2 f(z)dz \to f(x)\int_{-\infty}^{+\infty} k^2(z)dz,$$

ya que $\int_{-\infty}^{+\infty} k^2(z)dz < \infty$. Es por tanto evidente que

$$\lim_{n\to\infty} Var[\hat{f}(x)] \to 0 \quad \text{si} \quad \lim_{n\to\infty} nh_n = \infty.$$

Finalmente al ser

$$MSE[\hat{f}(x)] = Var[\hat{f}(x)] + Sesgo^2[\hat{f}(x)],$$

teniendo en cuenta el Corolario 3.4.1 el Teorema queda demostrado. □

Este resultado ilustra perfectamente el problema básico de la estimación no paramétrica. Una rápida convergencia al cero del parámetro h provoca una disminución

del sesgo, pero sin embargo la varianza aumentaría de forma considerable. El ancho de ventana ideal debe de converger a cero pero a un ritmo más lento que n^{-1}, y es lo que veremos en la siguiente sección.

3.5. Ancho de ventana óptimo asintótico

La fórmula del AMISE expresa el MSE como una función de h. El valor de h que minimiza esta expresión se llama ancho de ventana óptimo asintótico. La solución se obtiene tomando la derivada del AMISE con respecto a h y igualándola a cero

$$\begin{aligned}\frac{d}{dh}AMISE &= \frac{d}{dh}\left(\frac{\kappa_\nu^2(k)}{\nu!}R(f^{(\nu)})h^{2\nu} + \frac{R(k)}{nh}\right)\\ &= 2\nu h^{2\nu-1}\frac{\kappa_\nu^2(k)}{\nu!}R(f^{(\nu)}) - \frac{R(k)}{nh^2}\\ &= 0,\end{aligned}$$

con solución

$$h_0 = C_\nu(k,f)n^{-1/(2/nu+1)},$$

$$C_\nu(k,f) = R(f^{(\nu)})^{-1/(2/nu+1)}A_\nu(k),$$
$$A_\nu(k) = \left(\frac{(v!)^2R(k)}{2\nu\kappa_\nu^2(k)}\right)^{1/(2\nu+1)}.$$

El ancho de ventana óptimo es proporcional a $n^{-1/(2\nu+1)}$. Decimos que el ancho de ventana es de orden $O(n^{-1/(2\nu+1)})$. Para núcleos de segundo orden el orden óptimo es $O(n^{-1/5})$. Para núcleos de orden superior el orden es más lento, lo que sugiere que los anchos de ventana son generalmente más grandes que para los núcleos de segundo orden. La intuición es que dado que los núcleos de orden superior tienen sesgos más pequeños, pueden permitirse un ancho de ventana más grande.

La constante de proporcionalidad $C_\nu(k, f)$ depende del núcleo a través de la función $A_\nu(k)$ (que se puede calcular), y de la densidad a través de $R(f^{(\nu)})$ (que es desconocido).

Si el ancho de ventana se ajusta a h_0, entonces con algunas simplificaciones el AMISE es igual a

$$AMISE_0(k) = (1 + 2\nu) \left(\frac{R(f^{(\nu)}) k_\nu^2 R(k)^{2\nu}}{(\nu!)^2 (2\nu)^{2\nu}} \right)^{1/(2\nu+1)} n^{-2\nu/(2\nu+1)}.$$

Para núcleos de segundo orden, esto equivale a

$$AMISE_0(k) = \frac{5}{4} \left(\kappa_2^2(k) R(k)^4 R(f^{(2)}) \right)^{1/5} n^{-4/5}.$$

Cuando ν es grande, el orden de convergencia aproxima el orden paramétrico a n^{-1}. Por lo tanto, al menos asintóticamente, la lenta convergencia de la estimación paramétrica puede ser mitigada por el uso de núcleos de orden superior.

Esto parece un poco extraño. øCuál es el motivo? Por un lado, la mejora en el orden de convergencia requiere que la densidad sea suficientemente suave para que la derivadas existan hasta el orden $\nu + 1$. A medida que la densidad se vuelve cada vez más suave, es más fácil aproximarla por una curva de baja dimensión, y se acerca a un problema de tipo paramétrico. Esto es explotar la suavidad de f, la cual es inherentemente desconocida. El otro motivo es que hay una cierta evidencia de que los beneficios de los núcleos de orden superior se desarrollan sólo cuando el tamaño muestral es bastante grande. La sensación es que en muestras pequeñas, un núcleo de segundo orden sería la mejor elección, en muestras medianas un núcleo de cuarto orden, y en muestras grandes se podría usar un núcleo de sexto orden.

3.6. Núcleo Óptimo Asintótico

Dado que hemos escogido el orden del núcleo, ¿qué núcleo deberíamos usar? Examinando la expresión $AMISE_0$ podemos ver que par un valor fijado ν la elección del núcleo afecta a la precisión asintótica a través de la cantidad $\kappa_\nu(k)R(k)^\nu$. En iguales condiciones, el AMISE será minimizado seleccionando el núcleo que minimice esa cantidad. Como discutimos anteriormente, solo la forma del núcleo es importante, no su escala, por lo tanto podemos establecer que $\kappa_\nu = 1$. Entonces el problema se reduce a la minimización de $R(k) = \int_{-\infty}^{+\infty} k(u)^2 du$ bajo las restricciones $\int_{-\infty}^{+\infty} k(u)du = 1$ y $\int_{-\infty}^{+\infty} u^\nu k(u)du = 1$. Este problema es un problema en el cálculo de variaciones. Como la escala es irrelevante, esto significa que para la estimación de la función de densidad, el núcleo Epanechnikov de orden superior $\kappa_{\nu,1}$ con ancho de ventana óptimo produce el AMISE más bajo posible. Por esta razón, el núcleo Epanechnikov se suele llamar "núcleo óptimo".

Para comparar los núcleos, se define su eficiencia relativa como

$$\begin{aligned} eff(k) &= \left(\frac{AMISE_0(k)}{AMISE_0(k_{\nu,1}} \right)^{(1+2\nu)/2\nu} \\ &= \frac{(\kappa_\nu^2(k))^{1/2\nu} R(k)}{(\kappa_\nu^2(k_{\nu,1}))^{1/2\nu} R(k_{\nu,1})}. \end{aligned}$$

Como para n grande la razón del AMISE está elevada a la potencia $(1 + 2\nu)/2\nu$, el AMISE será el mismo si usamos n observaciones con el núcleo $\kappa_{\nu,1}$ o $n \cdot eff(k)$ observaciones con el núcleo k. Por lo tanto la penalización $eff(k)$ se expresa como un porcentaje de observaciones.

Las eficiencias de varios núcleos están dadas en la Tabla 3.1. Examinando los núcleos de segundo orden, podemos ver que en relación con el núcleo Epanechnikov, el

núcleo uniforme paga una penalización alrededor del 7 %, el núcleo Gaussianiano una penalización alrededor del 5 %, el núcleo Triweight sobre 1.4 %, y el núcleo Biweight menos del 1 %.

Las diferencias no son muy grandes. Sin embargo, el cálculo sugiere que los núcleos Epanechnikov y Biweight son una buena elección para la estimación de la densidad.

3.7. Selección del ancho de ventana

Siguiendo a [18] Jones, Marron y Sheather (1996a) podemos clasificar las técnicas de selección del ancho de ventana basadas en una muestra en *métodos de primera generación* y *métodos de segunda generación*. La clasificación tiene su origen principal en la superioridad que han mostrado las técnicas desarrolladas recientemente, a partir de 1990 frente a las técnicas de primera generación desarrolladas en su mayoría con anterioridad a 1990.

Entre los métodos de primera generación incluimos:

- Reglas basadas en las distribuciones paramétricas. "Rules of Thumb".
- Sobresuavización.
- Reglas de Validación cruzada.

y entre los de segunda:

- Métodos Plug-in.
- Boostsrap suavizado.

El método más utilizado para la elección del ancho de ventana en estimación de densidad univariante es el "Rules-of-Thumb", por ello a continuación explicaremos con más detalle en qué consiste dicho método.

3.7.1. Rules-of-Thumb

El ancho de ventana óptimo depende de la cantidad desconocida

$$R(f^{(\nu)}) = \int_{-\infty}^{+\infty} f^{(\nu)}(u)^2 du.$$

[30] propuso que podíamos intentar calcular el ancho de ventana remplazando $R(f^{(\nu)})$ en la fórmula óptima por $R\left(g_\sigma^{(\nu)}\right)$ donde g_σ es una densidad de referencia (un posible candidato para f), y $\hat{\sigma}^2$ es la desviación típica muestral. La elección estándar es tomar $g_\sigma = \phi_{\hat{\sigma}}$, la distribución normal $N(0, \hat{\sigma}^2)$ de media cero y varianza $\hat{\sigma}^2$. La idea es que si la densidad real es normal, entonces el ancho de ventana calculado será óptimo. Si la densidad real está razonablemente cerca de la normalidad, entonces el ancho de ventana estará cerca del óptimo. Aunque no es una solución perfecta, es una buena alternativa.

Para cualquier densidad g, si tomamos $g_\sigma(x) = \sigma^{-1} g(x/\sigma)$, entonces $g_\sigma^{(\nu)}(x) = \sigma^{-1} g^{(\nu)}(x/\sigma)$. Por lo tanto,

$$\begin{aligned} R\left(g_\sigma^{(\nu)}\right)^{-1/(2\nu+1)} &= \left(\int g_\sigma^{(\nu)}(x)^2 dx\right)^{-1/(2\nu+1)} \\ &= \left(\sigma^{-2-2\nu} \int g^{(\nu)}(x/\sigma)^2 dx\right)^{-1/(2\nu+1)} \\ &= \left(\sigma^{-1-2\nu} \int g^{(\nu)}(x)^2 dx\right)^{-1/(2\nu+1)} \\ &= \sigma R\left(g^{(\nu)}\right)^{-1/(2\nu+1)}. \end{aligned}$$

Además,

$$R\left(\phi^{(\nu)}\right)^{-1/(2\nu+1)} = 2\left(\frac{\pi^{1/2}\nu!}{(2\nu)!}\right)^{1/(2\nu+1)}.$$

Por lo tanto,

$$R\left(\phi_{\hat{\sigma}}^{(\nu)}\right)^{-1/(2\nu+1)} = 2\hat{\sigma}\left(\frac{\pi^{1/2}\nu!}{(2\nu)!}\right)^{1/(2\nu+1)}.$$

El ancho de ventana obtenido a través de este método es entonces

$$h = \hat{\sigma}C_\nu(k)n^{-1/(2\nu+1)},$$

donde

$$\begin{aligned} C_\nu(k) &= R\left(\phi^{(\nu)}\right)^{-1/(2\nu+1)} A_\nu(k) \\ &= 2\left(\frac{\pi^{1/2}(\nu!)^3 R(k)}{2\nu(2\nu)!\kappa_\nu^2(k)}\right)^{1/(2\nu+1)}. \end{aligned}$$

A continuación ilustraremos una tabla con los valores de las constantes $C_\nu(k)$ para varios tipos de núcleos y para varios órdenes.

Núcleo	$\nu = 2$	$\nu = 4$	$\nu = 6$
Epanechnikov	2.34	3.03	3.53
Biweight	2.78	3.39	3.84
Triweight	3.15	3.72	4.13
Gaussiano	1.06	1.08	1.08

Tabla 3.2: Constantes Rule-of-Thumb.

Teniendo en cuenta dicha tabla, podemos decir por ejemplo que el ancho de ventana óptimo para un núcleo de segundo orden Gaussiano viene dado por

$$h = 1{,}06\hat{\sigma}n^{-1/5}.$$

3.8. Estimación de Densidades Multivariantes

3.8.1. Definición y propiedades básicas

Dada la muestra aleatoria $\mathbf{x}_1, ..., \mathbf{x}_n$ de elementos $\mathbf{x}_i \in \mathbb{R}^d$, definimos la estimación de la densidad por núcleos multivariantes, con función núcleo $k : \mathbb{R}^d \to \mathbb{R}$ de la forma

$$\hat{f}(\mathbf{x}) = \frac{1}{n \cdot \det(\mathbf{H})} \sum_{i=1}^{n} k\left(\mathbf{H}^{-1}(\mathbf{x} - \mathbf{x}_i)\right), \tag{3.8.1}$$

donde $\mathbf{H}$ es una matriz simétrica y definida positiva de orden $d \times d$ que será la denominada matriz de anchos de ventana y donde la función núcleo es generalmente una función de densidad multivariante.

$$\int_{\mathbb{R}^d} k(\mathbf{x}) d\mathbf{x} = 1. \tag{3.8.2}$$

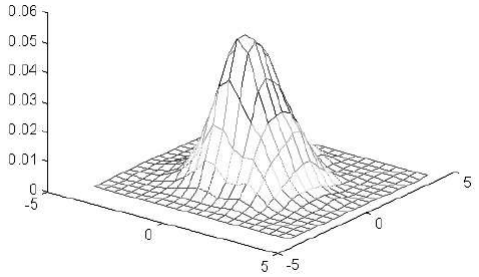

Figura 3.1: Estimación núcleo multivariante.

Las más usuales en $\mathbb{R}^d$ son:

- Función núcleo multivariante de Gauss

$$k_N(\mathbf{x}) = (2\pi)^{-d/2} e^{-1/2\mathbf{x}^T\mathbf{x}}.$$

- Función núcleo multivariante de Barlett-Epanechnikov

$$k_e(\mathbf{x}) = \begin{cases} \frac{1}{2}c_d^{-1}(d+2)(1-\mathbf{x}^T\mathbf{x}), & \text{si } \mathbf{x}^T\mathbf{x} < 1, \\ 0 & \text{en caso contrario,} \end{cases}$$

donde c_d es el volumen de la esfera unidad de dimensión d dado por

$$c_d = \pi^{d/2}/\Gamma((d/2)+1),$$

por ejemplo: $c_1 = 2$, $c_2 = \pi$, $c_3 = 4\pi/3$, etc.

- Otras funciones útiles para el caso $d = 2$ son:

$$k_2(\mathbf{x}) = \begin{cases} 3\pi^{-1}(1-\mathbf{x}^T\mathbf{x})^2, & \text{si } \mathbf{x}^T\mathbf{x} < 1, \\ 0 & \text{en caso contrario.} \end{cases}$$

$$k_3(\mathbf{x}) = \begin{cases} 4\pi^{-1}(1-\mathbf{x}^T\mathbf{x})^3, & \text{si } \mathbf{x}^T\mathbf{x} < 1, \\ 0 & \text{en caso contrario.} \end{cases}$$

En la práctica una de las opciones más recomendada es la utilización del producto de funciones núcleo univariante que se define como sigue.

- Producto de funciones núcleo univariantes

$$k(\mathbf{x}) = \prod_{i=1}^{d} k(x_i).$$

Algunas de las condiciones generalmente exigidas a la función núcleo $k(\mathbf{x})$ vienen dadas por las siguientes ecuaciones matriciales

$$\int_{\mathbb{R}^d} k(\mathbf{x})d\mathbf{x} = 1, \quad \int_{\mathbb{R}^d} \mathbf{x}k(\mathbf{x})d\mathbf{x} = 0 \quad \text{y} \quad \int_{\mathbb{R}^d} \mathbf{x}\mathbf{x}^T k(\mathbf{x})d\mathbf{x} = I_d.$$

Si k es una densidad de probabilidad multivariante, las dos últimas ecuaciones anteriores resumen muchas propiedades de las funciones núcleo marginales. La segunda ecuación dice que las medias de las marginales son iguales a cero y la tercera que los núcleos marginales son incorrelacionados dos a dos y con varianza unidad.

Volviendo a la matriz $\mathbf{H}$ podemos considerar algunas clases de valores posibles para dicha matriz

$$\mathcal{H}_1 = \{h_1\mathbf{I} : h_1 > 0\}, \quad \mathcal{H}_2 = \{diag(h_1, ..., h_d > 0\},$$

o en el caso bivariante ($d = 2$)

$$\mathcal{H}_3 = \left\{ \begin{pmatrix} h_1 & h_{12} \\ h_{12} & h_2 \end{pmatrix} : h_1, h_2 > 0, h_{12}^2 < h_1 h_2 \right\}.$$

Notemos que $\mathcal{H}_1 \subseteq \mathcal{H}_2 \subseteq \mathcal{H}_3$ y que en el caso bivariante cada clase representa estimadores con uno, dos o tres parámetros de suavización independientes.

Es fácil observar que utilizando el núcleo Gaussiano

$$k_{\mathbb{H}}(\mathbf{x}) = |\mathbb{H}|^{-1} k(\mathbb{H}^{-1}\mathbf{x}) = (2\pi)^{-1} |\mathbb{H}|^{-1} e^{\left(\frac{-1}{2}\mathbf{x}^T \mathbb{H}^{-2}\mathbf{x}\right)},$$

que es la densidad de una distribución normal multivariante con vector de medias $\mathbf{0}$ y matriz de covarianzas $\mathbf{H}^2$. La pertenencia a $\mathcal{H}_1$ significa que la masa del núcleo será esférica, a $\mathcal{H}_2$ significa que será elíptica con los ejes ortogonales y en el caso $\mathcal{H}_3$ elíptica con los ejes en cualquier orientación.

Bajo la axiomática anterior y la parametrización

$$\mathbf{H} = h \cdot A,$$

donde A es una matriz $d \times d$ con $|A| = 1$ y $h > 0$, en [29] Scott (1922) a través de la forma multidimensional del desarrollo de Taylor pero siguiendo el mismo esquema que en el caso univariante, se muestra que para una estimación como la definida en (3.8.1) el error cuadrático medio asintótico toma la forma

$$AMISE = \frac{R(k)}{nh^d} + \frac{1}{4}h^4 \int_{\mathbb{R}^d} [tr\{AA^T \nabla^2 f(\mathbf{x})\}]^2 d\mathbf{x},$$

donde $R(k) = \int_{\mathbb{R}^d} k(\mathbf{x})^2 d\mathbf{x}$ y $\nabla^2 f(\mathbf{x}) = \partial^2 f/(\partial_{x_i}\partial_{x_j})$.

Bajo la parametrización anterior se tiene que si por ejemplo $\mathbf{H} \in \mathcal{H}_2$

$$H = \begin{pmatrix} h_1 & & 0 \\ & \ddots & \\ 0 & & h_d \end{pmatrix}; \text{ entonces } H = h \cdot \begin{pmatrix} h_1/h & & 0 \\ & \ddots & \\ 0 & & h_d/h \end{pmatrix},$$

donde $h = \left(\prod_{i=1}^{d} h_i\right)^{1/d}$.

3.8.2. Selección del ancho de ventana

La elección óptima de la matriz de anchos de ventana será aquella que minimiza el AMISE. [30] presenta algunos resultados para el parámetro de suavización en el caso $\mathbf{H} \in \mathcal{H}_1$, es decir $\mathbf{H} = h\mathbf{I}$, se obtiene

$$AMISE = \frac{1}{nh^d} R(k)^d + \frac{1}{(\nu!)^2} h^{2\nu} \kappa_\nu^2 \int [\nabla^\nu f(\mathbf{x})]^2 d\mathbf{x},$$

y se obtiene un parámetro óptimo

$$h^* = \left(\frac{(\nu!)^2 dR(k)^d}{2\nu\kappa_\nu^2 \int [\nabla^\nu f(\mathbf{x})]^2 d\mathbf{x} n} \right)^{1/(2\nu+d)},$$

versión multivariante de la forma obtenida en el caso univariante. Una posibilidad es considerar los datos procedentes de una distribución normal multivariante de varianza unidad, obteniéndose una valor óptimo para el ancho de ventana que minimiza el AMISE

$$h^* = C_\nu(k,d) n^{-1/(2\nu+d)},$$

donde la constante $C_\nu(k,d)$ depende del núcleo utilizado según se muestra en la tabla siguiente.

Función núcleo	Dimensión	$C_\nu(k,d)$
Mult. Gauss	2	1
Mult. Gauss	d	$(4/(d+2))^{1/(d+4)}$
Mult. Epanechnikov	2	2.40
Mult. Epanechnikov	d	$(Sc_d^{-1}(d+4)(2\sqrt{\pi})^d)^{1/(d+4)}$
k_2	2	2.78
k_3	2	3.12

Tabla 3.3: Valor de la constante $C_\nu(k,d)$ para diversos núcleos multivariantes.

3.8.3. Normalidad asintótica

Como hemos definido anteriormente, el estimar kernel multivariante es el promedio muestral

$$\hat{f}(\mathbf{x}) = \frac{1}{n|\mathbf{H}|}\sum_{i=1}^{n} k\left(\mathbf{H}^{-1}(\mathbf{x}-\mathbf{x}_i)\right).$$

Por lo tanto, podemos aplicar el Teorema Central del Límite (CLT), pero la tasa de convergencia no es $\sqrt{n}$. Sabemos que

$$Var[\hat{f}(x)] = \frac{f(x)R(k)^d}{nh_1h_2\cdots h_d} + O\left(\frac{1}{n}\right).$$

por lo tanto la tasa de convergencia es $\sqrt{nh}$. Cuando aplicamos el CLT escalamos por esta tasa, en lugar del convencional $\sqrt{n}$.

Como es un estimador insesgado, también nos centramos en su esperanza.

Así

$$\begin{aligned}\sqrt{nh_1h_2\cdots h_d}\left(\hat{f}(x) - E[\hat{f}(x)]\right) &= \frac{\sqrt{nh_1h_2\cdots h_d}}{n}\sum_{i=1}^{n}\frac{1}{|\mathbf{H}|}k(\mathbf{H}^{-1}(\mathbf{x}-\mathbf{x}_i)) - \\ &\quad E\left(\frac{1}{|\mathbf{H}|}k(\mathbf{H}^{-1}(\mathbf{x}-\mathbf{x}_i))\right) \\ &= \frac{\sqrt{h_1h_2\cdots h_d}}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{1}{|\mathbf{H}|}k(\mathbf{H}^{-1}(\mathbf{x}-\mathbf{x}_i)) - E\left(\frac{1}{|\mathbf{H}|}k(\mathbf{H}^{-1}(\mathbf{x}-\mathbf{x}_i))\right)\right) \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_{ni}\end{aligned}$$

donde

$$Z_{ni} = \sqrt{h_1h_2\cdots h_d}\left(\frac{1}{|\mathbf{H}|}k(\mathbf{H}^{-1}(\mathbf{x}-\mathbf{x}_i)) - E\left(\frac{1}{|\mathbf{H}|}k(\mathbf{H}^{-1}(\mathbf{x}-\mathbf{x}_i))\right)\right).$$

Podemos ver que

$$Var[Z_{ni}] \simeq f(x)R(k)^d.$$

Por lo tanto por el CLT,

$$\sqrt{nh_1h_2\cdots h_d}\left(\hat{f}(x) - E[\hat{f}(x)]\right) \to_d N\left(0, f(x)R(k)^d\right).$$

3.9. Implementación de los conceptos estudiados en R

Como hemos visto a lo largo de este capítulo, la estimación de la densidad es una herramienta estadística importante, y dentro de R hay más de 20 paquetes que lo implementan: tantos que a menudo es difícil saber cuál usar.

En esta sección, veremos algunos de esos paquetes presentando una visión general del código a utilizar. De manera que el objetivo de esta sección es en definitiva proporcionar un código R para los conceptos más importantes vistos durante el capítulo.

3.9.1. Código para histogramas

Una de las funciones que podemos utilizar para la representación de un histogramas es la función **hist**. Dicha función permite a los usuarios generar un histograma de los datos **x**. El argumento **breaks** especifica el número deseado de barras, o la frontera de cada barra o bien una función que calcula el número de barras automáticamente. Más allá de la representación gráfica que la función devuelve por defecto, la función también proporciona los siguientes datos:

- Una lista con los límites de las barras.
- Una lista con el recuento de las frecuencias observadas.

- Los valores de la densidad estimada (estandarizados por el ancho de ventana).
- Los puntos medios de cada barra.

3.9.2. Código para estimaciones núcleo

Para el caso de la estimación núcleo vamos a presentar tres paquetes de R diferentes.

El paquete sm

El paquete **sm** puede realizar una estimación de la densidad tipo núcleo desde 1 dimensión a 3 dimensiones. La función que utiliza para ello es **sm.density**. Dicha función toma de entrada los datos **x**, el ancho de banda **h** o la matriz de anchos de banda **H**, y un vector de enteros que representa las frecuencias de las observaciones individuales **h.weights**. Si no se especifica el ancho de banda, **h.select** utiliza un estimador suavizado óptimo normal. Esta función devuelve por defecto una representación gráfica de la estimación, y además

- Una lista que contiene los valores de la estimación de la densidad en los puntos de evaluación.
- El ancho de ventana.
- Los pesos de los anchos de ventana.
- Los pesos del núcleo.
- Para los datos de uno y de dos dimensiones, también se suministran el error estándar de la estimación y los extremos superior e inferior de una banda de

variabilidad.

Para el caso de 2 y 3 dimensiones se necesitan además instalar los siguientes paquetes: "**misc3d**"; "**rpanel**"; "**rgl**".

El paquete stats

Utiliza la función **density** que realiza una estimación de la densidad univariante con los núcleos Gaussiano, rectangular, triangular y coseno. El ancho de ventana se especifica con el parámetro **bw**, en caso de no especificarse, la función toma por defecto el dado por la regla "Rule of Thumb"de Silverman, **bw.nrd0**. En este caso la función density no proporciona por defecto una representación gráfica de la estimación, sin embargo para conseguirla, podemos usar simplemente la orden **plot**. Esta función nos devuelve los siguientes datos:

- Las n coordenadas de los puntos donde se estima la densidad (x).
- Los valores de la densidad estimada (y).
- El ancho de ventana usado.
- El tamaño muestral.
- Un resumen de las componentes x e y.

De esta función queremos destacar que es útil para el cálculo del ancho de ventana mediante la regla "Rule of Thumb"de Silverman vista en el capítulo.

El paquete kedd

Este paquete se utiliza principalmente para calcular las derivadas de la densidad núcleo, aunque nosotros la utilizaremos simplemente para obtener las funciones núcleos y la densidad núcleo en vez de sus derivadas usando el comando *deriv.order=0* en la entrada de las funciones que utilizaremos para su cálculo.

Para calcular las funciones núcleos utilizaremos la función **kernel.fun** que recibe como argumentos de entrada, los puntos en los cuales se quiere evaluar la función núcleo,**x**; el orden de la derivada que como hemos dicho usaremos siempre 0, **deriv.order**; y el núcleo que usaremos, **kernel**. Nuevamente esta función no produce una representación gráfica y si queremos obtenerla debemos utilizar la función plot. Los datos que esta función proporciona son

- Las n coordenadas de los puntos donde se evalúa la función núcleo.
- Los valores de la función núcleo.

Si en lugar de calcular la función núcleo queremos calcular la estimación de la densidad tipo núcleo utilizamos la función **dkde** que recibe como argumentos de entrada: los datos muestrales, **x**; el orden de la derivada que como hemos dicho usaremos siempre 0, **deriv.order**; el ancho de banda, **h**, y el núcleo que se usa, **kernel**, que por defecto el software usa el gaussiano si no indicamos lo contrario. Esta función puede hacer estimaciones con los núcleo de Epanechnikov, Uniforme, Triangular, Triweight, Biweight o Cuártico, y Coseno, además del ya mencionado Gaussiano, y nos devuelve los siguientes datos:

- Las coordenadas de los puntos donde se ha estimado la densidad (eval.points).

- Los valores de la densidad estimada (est.fx).
- El ancho de ventana utilizado.
- Un resumen de las componentes eval.points y est.fx.

Y por último comentaremos la función **h.amise**. Esta función proporciona el ancho de ventana óptimo bajo el AMISE, recibe como argumentos de entrada: los datos muestrales, **x**; el orden de la derivada que como hemos dicho usaremos siempre 0, **deriv.order**; y nos devuelve

- El valor del ancho de ventana óptimo.
- El valor del AMISE.

De este paquete queremos destacar la función **h.amise** que es útil para el cálculo del ancho de ventana óptimo y del error cuadrático medio asintótico vistos en el capítulo.

3.10. Aplicación

Al igual que se hizo en el capítulo segundo, vamos a dedicar una sección a ver cómo se pueden aplicar las técnicas estadísticas vistas a lo largo del capítulo en un estudio arqueológico. Esta sección se desarrollará de la siguiente forma: en primer lugar, vamos a presentar una breve motivación del estudio que se llevará a cabo donde se pondrá de manifiesto el objetivo del mismo; en segundo lugar se pasará a describir los datos con los cuales se trabaja y por último se aplicará varias técnicas estadísticas a dichos datos, haciendo especial hincapié a las mencionadas anteriormente, y se proporcionará los resultados y conclusiones a las que se llegan en el estudio estadístico.

3.10.1. Motivación

Dentro del registro arqueológico hay muchos tipos de artefactos que atraen poca atención incluso en la literatura especialista. Normalmente esto es porque son utensilios funcionales cuyas formas básicas no han cambiado desde hace siglos. Como tales, no son ni útiles para fines de datación, ni suficientemente atractivos en sí mismos para generar interés desde el punto de vista de la historia del arte. Sin embargo, incluso estos objetos aparentemente mundanos pueden proporcionar información útil sobre las personas que los fabricaron y usaron si se analizan apropiadamente. El objetivo de este estudio es precisamente tomar objetos de este tipo y mostrar cómo el análisis estadístico puede proporcionar información útil. El tipo de objeto elegido para ilustrar esto son piezas de cerámicas que se usaban en las maquinarias para hacer telas, objetos de los que a continuación daremos una descripción más exhaustiva.

El análisis estadístico inicial de los pesos de las piezas de cerámicas bajo estudio se llevó a cabo en el campo. Para sorpresa de quiénes llevaron a cabo dicho estudio, la distribución de los pesos parecía ser claramente bimodal. El análisis posterior, que veremos aquí, parece que confirma esto y, adicionalmente sugiere patrones en los datos asociados con las formas de la parte superior y la base de las piezas de cerámica. La interpretación completa de los resultados en términos culturales y arqueológicos tiene que esperar al análisis completo de la estratigrafía de la zona, pero es posible avanzar algunas conclusiones previas que ofreceremos más adelante.

3.10.2. Registro empírico

Las piezas de cerámicas de las que hablamos fueron usadas en los urdimbres[1] de los telares para mantener los hilos de los urdimbres bajo tensión. Funcionalmente, su característica más importante es su peso, ya que los hilos delanteros y los traseros deben mantenerse bajo la misma tensión.

Los datos de este estudio se derivan de ejemplares recuperados durante la excavación de la Ínsula VI.1 por el Proyecto Anglo-Americano en Pompeya. Esta ínsula se encuentra junto a la Puerta de Herculano y fue una de las primeras zonas despojadas de escombros volcánicos a finales del siglo XVIII. Incluye la famosa Casa del Cirujano así como la Casa de los Vestales. Algunas de las decoraciones de la pared y el suelo se retiraron cuando fue excavada por primera vez. La erosión que ha sufrido en los dos siglos desde entonces, incluyendo los daños causados por la Segunda Guerra Mundial, donde fue alcanzada por una bomba de los Aliados, da lugar a que muy pocas de las superficies del piso originales presentes en el momento de la erupción en el año 79 d.C ahora sobreviven. Esto ha hecho posible el proyecto de excavar la ínsula dentro y alrededor de las paredes aún de pie para descubrir su historia desde el siglo IV a.C (cuando comenzó la ocupación) hasta la erupción.

Las excavaciones se concluyeron en 2006. La mayoría de las piezas de los telares tenían la forma típica de pirámide truncada con una perforación que atraviesa la parte superior (Fig. 3.2, n. 142). Un número pequeño de ellas tenían un contorno cuadrado pronunciado con una sección transversal rectangular (Fig. 3.2, n. 116). Una minoría habían sido decoradas con una o más abolladuras circulares grabadas en la parte superior.

[1]Conjunto de hilos que se colocan en el telar longitudinal para formar un tejido.

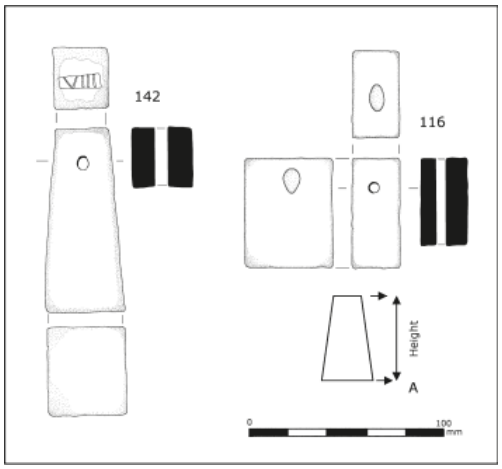

Figura 3.2: Ejemplos de las piezas de cerámicas bajo estudio.

Aunque dichas piezas se encontraron en toda la ínsula, mostraron una marcada concentración en el área ocupada por la Casa del Cirujano. Cuando se llevo a cabo este estudio, la información estratigráfica no estaba aún disponible para todas las partes de la ínsula, pero era casi completo para la zona de la Casa del Cirujano, de ahí que solo se puedan ofrecer conclusiones arqueológicas de dicha zona, conclusiones que comentaremos al final de esta sección.

El protocolo seguido en la recogida de estos objetos fue que se recogieron aquellos que pesaran en torno a los 2 gramos. Posteriormente se obtenían medidas de la altura, y de la parte superior y la base y los datos obtenidos se redondeaban al milímetro. En lo que sigue solamente se discutirán como hemos dicho las 95 piezas completas halladas, considerando ser completa el conservar todos sus bordes.

3.10.3. Análisis estadístico

Análisis unidimensional

En primer lugar se lleva a cabo un análisis para conocer la distribución que sigue el peso en las piezas de cerámica a estudiar. Como hemos mencionado en la parte teórica del capítulo, el histograma es el estimador no paramétrico de la densidad más sencillo y mejor conocido, y por ello, es la elección común para una exploración inicial de datos continuos.

A continuación en la Fig. 3.3 podemos ver dos histogramas con distintos anchos de banda. El primero de ellos usa el ancho de banda predeterminado por el software utilizado para su construcción (en este caso de 100g). Aparentemente no hay nada inusual en los datos, aparte de algunos pesos pequeños y grandes atípicos, sin embargo, como se tiene la sospecha que los anchos de banda predeterminados por los software tienden a sobresuavizar los datos, en el segundo de ellos se usa un ancho de banda diferente (en este caso de 25g). Esta segunda gráfica, a diferencia de la primera, parece sugerir que los pesos observados están divididos en dos "subgrupos" correspondientes a las dos "modas" que se observan en el gráfico. Para obtener una visión mejorada de esto, se superpone en la gráfica un estimador de densidad tipo núcleo, el cual nos sugiere más claramente que los datos son bimodales, y que pueden aproximarse de forma adecuada por una mixtura de dos distribuciones normales. Con este ejemplo queda reflejado el por qué decimos que los estimadores núcleos son una versión mejorada y más sofisticada que el histograma a la hora de estimar densidades.

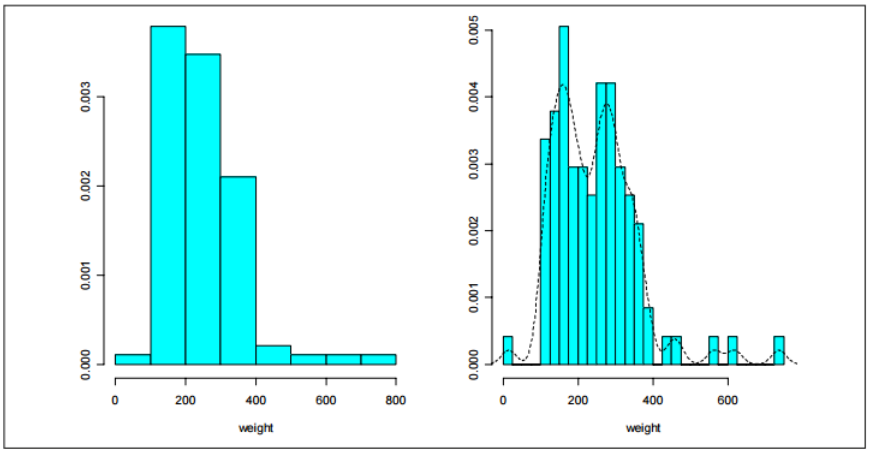

Figura 3.3: Dos histogramas que usan diferentes anchos de banda que muestran la distribución de los pesos de las piezas de cerámica encontradas en Pompeya.

Como hemos dicho anteriormente, hay algunos datos inusuales, en particular, un peso muy pequeño y 5 muy grandes, mayores de 400g. Para algunos análisis posteriores estos datos se eliminarán, y nos referiremos entonces a los datos como el conjunto de datos modificados.

Para confirmar la hipótesis de bimodalidad, se usó un software que proporciona no solo la representación gráfica del estimador de la densidad núcleo de los datos, si no que también permite hacer un test para conocer el número de componentes normales de la mixtura así como ofrece las medias y desviaciones típicas de las componentes. Dicha representación podemos verla en la Fig 3.4, y se basa en los datos modificados. Efectivamente, el test proporcionado por el software confirma que la mixtura de dos componentes normales es óptima, teniendo iguales desviaciones típicas estimadas $\sigma_1 = \sigma_2 = 41{,}5$ y medias $\mu_1 = 166$ y $\mu_2 = 300{,}7$, con 45 y 44 casos clasificados en los dos grupos.

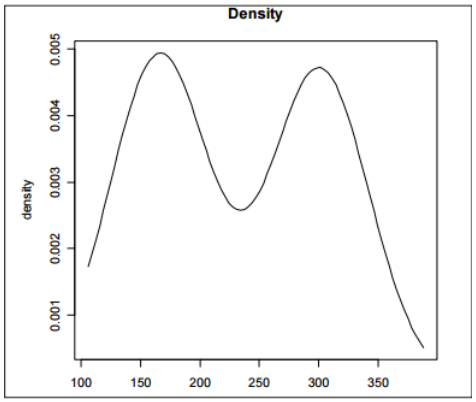

Figura 3.4: Una estimación de la mixtura de dos componentes normales para los pesos de las piezas de cerámica.

Se propone la siguiente regla para clasificar las piezas de cerámica:

"Las piezas con 230g o menos se asignan al primer grupo y las piezas con 239g o más se asignan al segundo grupo"

Esta regla da grupos de tamaño 46 y 49 respectivamente si se usa para clasificar todos los pesos.

A continuación se lleva a cabo un análisis similar pero en este caso para estudiar la distribución de la altura de las piezas. Así, en la Fig. 3.5 podemos ver dichos datos representados a través de un histograma con ancho de banda 5 con un estimador núcleo superpuesto de ancho de banda 18. Observamos que también se sugiere la bimodalidad en la distribución de la altura.

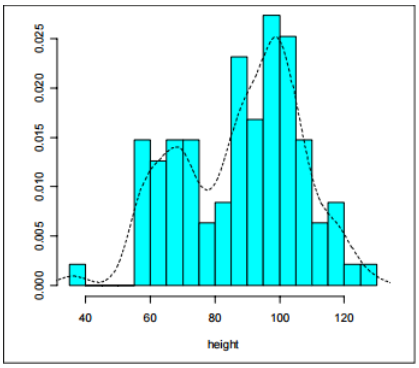

Figura 3.5: Histograma que representa la altura de las piezas de cerámica con un estimador de densidad núcleo superpuesto.

Análisis bidimensional: peso y altura

En el apartado anterior hemos analizado las piezas exclusivamente en función de su peso ya que como dijimos en la descripición de las mismas esta es su característica principal. Ahora, sin embargo, vamos a analizarlas en función de algunas carácteristicas más.

Una forma rápida de observar los datos (en este caso vamos a usar todas las piezas disponibles) es ver todas las posibles gráficas bivariantes, como observamos en la Fig. 3.6.

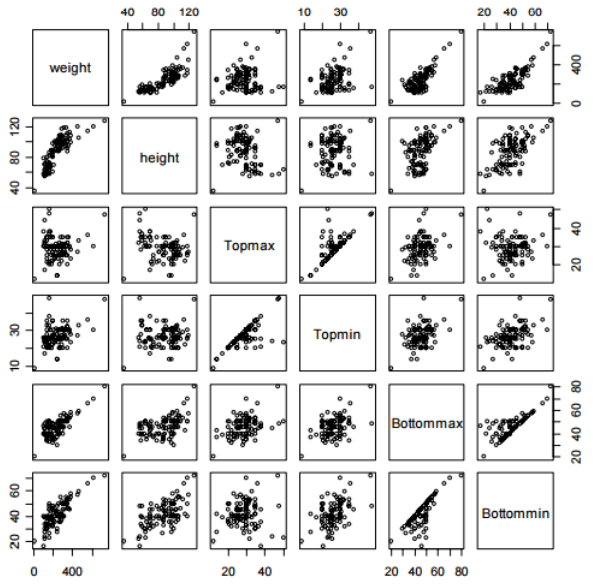

Figura 3.6: Pares de gráficas para seis variables características de las piezas de cerámica, que muestra todas las gráficas bivariantes posibles. El triángulo superior del dibujo es el mismo que el inferior, a excepción que los ejes están cambiados. Tomando la base y el vértice de las piezas como rectangulares, Topmax es la longitud de los lados mayores del rectángulo en el vértice y Topmin se refiere al lado más pequeño. Bottommax y Bottommin se refieren a dimensiones similares para la base.

En general las variables están correlacionadas positivamente, como era de esperar, aunque dicha correlación en algunos casos no es tan fuerte como pensábamos. La mayor correlación, de $r = 0{,}82$, en las gráficas mostradas se produce entre el peso y la altura, y es por eso que prestaremos especial atención a estas características.

A continuación mostramos un gráfico de dispersión de la altura frente al peso usando los datos modificados, etiquetados según la clasificación sugerida por el análisis de la mixtura para el peso. Algunos pesos podrían reclasificarse en el grupo 2 según la evidencia visual. Analizaremos esto más tarde.

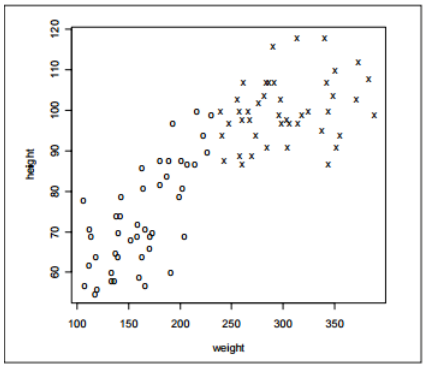

Figura 3.7: Gráfica de la altura frente al peso, con los casos etiquetados por la clasificación sugerida por el análisis de la mixtura.

La cuestión nuevamente aquí se trata de saber como se distribuyen ahora de forma simultánea el peso y la altura de las piezas. Para su análisis, utilizaremos estimadores de densidad núcleo bidimensionales, que se aplicarán una vez más a los datos modificados. Los resultados de este análisis se pueden mostrar de varias formas, como vemos en la Fig. 3.8. Normalmente sólo una de las gráficas es necesaria, pero se muestran todas de forma ilustrativa.

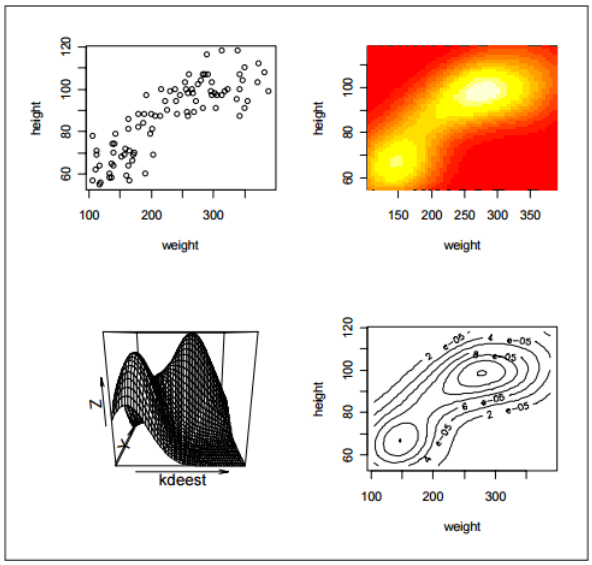

Figura 3.8: Diferentes formas de mostrar las relación entre la altura y el peso. Los datos en bruto se muestran en la parte superior izquierda; en la parte superior derecha podemos observar la imagen gráfica; un dibujo en perspectiva se muestra abajo a la izquierda; y a la derecha tenemos el dibujo de contornos.

A la vista de las gráficas, vemos que todas sugieren también dos concentraciones principales en los datos.

Otras dimensiones

Volviendo a la Fig. 3.6, vemos que también existen patrones evidentes en las dimensiones máximas y mínimas de la parte superior y la base de las piezas, de ahí que también vamos a analizar su relación con el peso de las mismas.

En dicha gráfica se puede apreciar que dichas variables muestran características lineales distintivas. Estas corresponden a piezas donde o bien la parte superior o bien la base eran cuadradas. Para las piezas con bases que no eran cuadradas, la diferencia

mínima entre los dos lados era de 2mm, pero normalmente superaban los 5mm. Para las piezas con partes superiores no cuadradas, las cuales eran más pequeñas que las bases, hubo más casos de diferencias pequeñas en las dimensiones incluyendo la diferencia de 1mm.

Estas diferencias pueden verse en la siguiente gráfica:

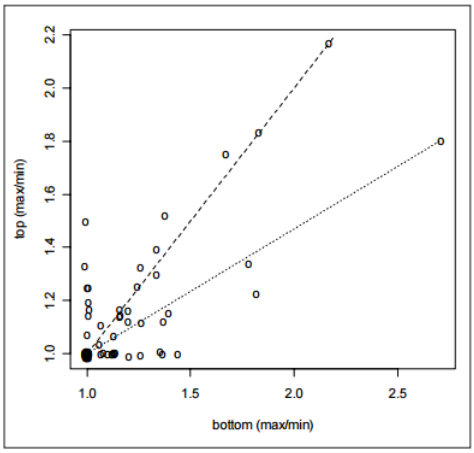

Figura 3.9: Gráfica de las proporciones máximas/mínimas de las partes superiores de las piezas frente a unas proporciones similares para las bases de las mismas.

A raíz de este análisis, se propone dividir las piezas en función de su forma (parte superior y base) según la siguiente clasificación:

- Tipo 1: base y parte superior cuadradas (por ejemplo, Fig. 3.2, n. 142);
- Tipo 2: base rectangular (no cuadrada) y parte superior de dimensiones relativas similares;
- Tipo 3: base cuadrada y parte superior rectangular;
- Tipo 4: base rectangular y parte superior cuadrada (por ejemplo, (Fig. 3.2, n. 166);

- Tipo 0: otra.

En los apartados anteriores, hemos visto que los análisis sugieren que, en base al peso y la altura, es posible dividir las piezas de los telares en dos clases o grupos de tamaños. El objetivo de este apartado, es relacionar, si es posible, dicha clasificación con esta otra nueva sugerida según la forma, en definitiva comprobar si existe relación entre la forma y el peso o la altura de las piezas que estamos estudiando.

En primer lugar vamos a representar los datos mediante un gráfico de dispersión. Dicho gráfico enfrenta nuevamente la altura frente al peso pero en este caso los objetos estarán etiquetados según la clasificación dada anteriormente basada en la forma de dichos objetos.

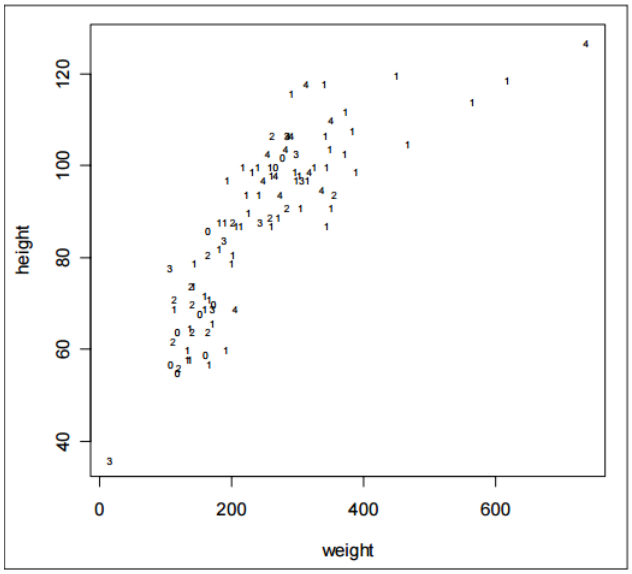

Figura 3.10: Gráfica de la altura frente al peso, etiquetado de acuerdo a la clasificación basada en la forma.

A la vista del gráfico, podemos ver ciertos indicios de que las piezas de mayor peso (más de 375g) tienden a ser de Tipo 1.

Una forma alternativa de ver los datos es a través de una tabla de clasificación cruzada según el tamaño. Para esta propuesta se ha modificado ligeramente la clasificación sugerida por el modelo mixtura, para tener en cuenta la evidencia visual y no separar los valores inusuales. Llamaremos a estas nuevas clases, "Pequeña" y "Grande"; la clasificación modificada se muestra en la Fig. 3.11, mientras que la clasificación cruzada se da en la Tabla 3.4.

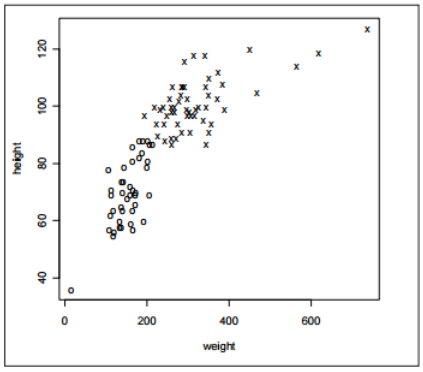

Figura 3.11: Similar a la Fig. 3.7 pero usando todos los datosy con una clasificación del tamaño modificada.

Tamaño	Tipo					Total
	0	1	2	3	4	
Grande	2	32	4	4	12	54
Pequeño	7	20	9	4	1	41

Tabla 3.4: Clasificación cruzada del tamaño por el tipo de forma, basada en las clasificaciones descritas en el texto.

Por último para confirmar si existe o no relación entre la clasificación basada en la forma y la clasificación basada en el tamaño se aplica un test χ^2 convencional (técnica

comentada en el capítulo segundo) dando como resultado un valor del estadístico de $\chi^2 = 15{,}48$ con 4 grados de libertad y un p-valor=0.0041. Antes de aceptar este resultado como válido, debemos destacar que el subconjunto de objetos usados en el análisis es muy pequeño, por lo que al igual que ocurría en el estudio que se llevó a cabo en Valencina, el test χ^2 podría no dar un resultado válido puesto que se trata de un test asintótico. También nos encontramos ante el inconveniente de que la tabla no cumple las condiciones exigidas para su aplicación, pues más del 20 % de las casillas no superan el 5. Por lo tanto, y procediendo de igual modo que en el capítulo anterior, se aplica el test de fisher para corroborar el resultado obtenido. Dicho test proporciona un p-valor=0.0027. A raíz de estos resultados, comprobamos que existe por lo tanto una clara asociación entre la forma y el tamaño de las piezas estudiadas.

Conclusiones arqueológicas

Como mencionamos en la motivación, el análisis de la estratigrafía no es lo suficientemente completo para que sea posible datar la mayoría de las piezas por su contexto. Sin embargo, en el caso de las piezas halladas en la Casa del Cirujano es posible aislar grupos pequeños de contextos de fechas diferentes. Nos encontramos con un grupo de 5 objetos de características que son anteriores a la construcción de la casa en el 200 a.C. Otro grupo, de 5 piezas también, se encontró en el hoyo excavado para extraer material de construcción para extender el *triclinio*[2]. Este estaba relleno de basura doméstica datada sobre el 100 a.C. Finalmente, 9 pueden datarse a mediados del primer siglo d.C ya que se recuperaron de compensar y nivelar las capas de las últimas plantas en la Casa del Cirujano. Esta fase de reconstrucción se

[2]Un triclinio es una estancia destinada a comedor formal en un edificio romano o grecorromano.

cree que se llevó a cabo entre el terremoto, convencionalmente datado en el 62 d.C, y la erupción en el 79 d.C. Estos datos se encuentran resumidos en la Tabla 3.5 en función del peso y las formas definidas anteriormente.

Grupo	Fecha	Pequeño				Grande			Total
		0	2	3	1	1	3	4	
1	Pre c. 200 a.C	2	2	-	1	-	-	-	5
2	c. 100 a.C	-	-	-	3	2	-	-	5
3	c. 62-79 d.C	-	-	1	3	3	1	1	9
Total		2	2	1	7	5	1	1	19

Tabla 3.5: Piezas de cerámica datadas independientemente de la Casa del Cirujano.

A continuación representaremos los ejemplos que caen en el conjunto de los datos modificados etiquetando los puntos según el Grupo en el que se clasifican. Esta representación puede verse en la Fig 3.12, donde el peso se ha medido de acuerdo a las medidas de gramo modernas mientras los Grupos se han etiquetado según las medidas de *unciae* Romanas.

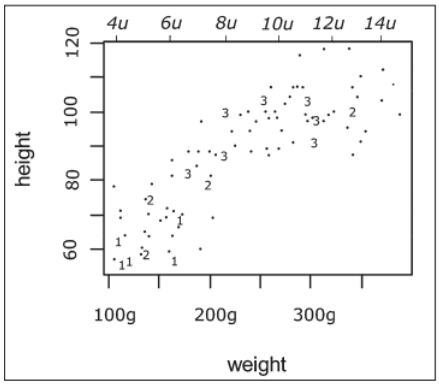

Figura 3.12: Gráfica del conjunto de datos modificados etiquetados con los grupos de la Casa del Cirujano resumidos en la Tabla 3.5.

A la vista del gráfico, se observa como las piezas del tercer siglo a.C (Grupo 1) se agrupan entre 4 y 6 *unciae* y las de mediado del primer siglo a.C del Grupo 3 oscilan entre 6 y 12 *unciae*. Esto posiblemente puede sugerir que los fabricantes de estas piezas estaban trabajando hacia la producción de piezas de pesos específicos y que éstos cambiaron con el tiempo. Es cierto que se necesitan más evidencias y más piezas a datar de las que aquí se han analizado, pero existe la posibilidad de que el peso de las piezas de cerámicas de los telares pueda tener un significado cronológico en Pompeya. Si se acepta que hubo cambios en el tamaño de las piezas, se pueden plantear otras preguntas como por ejemplo si hubo cambios en la naturaleza de los textiles que se producían. La creciente estandarización de la forma con el tiempo también podría apuntar a un aumento del nivel de centralización en la producción de estos artefactos.

Si el patrón se reproduce en otros lugares de la ínsula, las piezas de cerámica de los telares pueden pasar a la categoría de hallazgos que son cronológicamente sensibles y, como hemos señalado en la motivación, siempre se presta más atención a este tipo

de hallazgos.

Capítulo 4

OTROS METODOS ESTADISTICOS APLICADOS EN ARQUEOLOGIA

"Si tu experimento necesita de la Estadística,
entonces hubiese sido necesario hacer un experimento mejor"
Ernest Rutherford, (1871-1937)

Resumen. En este capítulo veremos dos métodos estadísticos que se usan en la investigación arqueológica con el fin de reflejar la relación existente entre ambas ciencias. Se hará una introducción de cada uno de los métodos, para posteriormente ilustrarlas con ejemplos prácticos donde veremos para qué se usan dichas técnicas en Arqueología.

Hay muchas etapas en un proceso de investigación arqueológica donde los problemas estadísticos están presentes. En primer lugar nos encontramos con el problema de la recolección de datos, pues es de remarcar otra vez más que una apropiada selección de los mismos es crucial para validaciones posteriores de los procedimientos de inferencia. Las técnicas de muestreo y el diseño experimental ofrecen buenas soluciones a esto.

Una vez tenemos los datos recogidos, el objetivo de los arqueólogos es representarlos gráficamente para así tener una representación visual de los mismos. En el capítulo anterior hemos visto varios métodos muy útiles para ello, como son los *Histogramas* y las *Estimaciones núcleo de la densidad* .

La siguiente cuestión a resolver por parte de los arqueólogos es realizar diferentes estudios inferenciales a los datos con el fin de obtener conclusiones de los mismos y así ya con los datos obtenidos de los diferentes estudios estadísticos poder responder a las diferentes cuestiones arqueológicas que se presenten. Los métodos de datación, el análisis de conglomerados y el análisis discriminante son quizás los procedimientos más populares para llevar a cabo esta etapa, y serán en algunos de ellos en los que nos centremos en este capítulo. Entre las técnicas que comentaremos destacan los *Árboles de Regresión y Clasificación (CART)* para el análisis discriminante y los *Métodos bayesianos*, punto de vista a partir del cual trataremos el problema de la datación por radiocarbono.

4.1. Árboles de regresión y clasificación (CART)

Como podemos imaginar la clasificación de individuos u objetos hallados en grupos o poblaciones conocidas, como por ejemplo separar objetos hallados según la época

a la que pertenecen, los materiales con los que se fabricaron o las carácterísticas que presentan, entre otros; puede ser de gran ayuda a la hora de llevar a cabo un estudio arqueológico.

Desde el punto de vista estadístico, el problema de la clasificación es también de gran interés y por ello se han desarrollado técnicas para cumplir este objetivo, técnicas de las que se ayudan los arqueólogos en sus estudios. La técnica más utilizada es el análisis discriminante, pero requiere de unas condiciones previas, normalidad y homocedasticidad, que no se cumplen con frecuencia. Por este motivo se han desarrollado otras técnicas basadas en árboles de decisión, una de ellas y la cual comentaremos aquí, es los Árboles de Regresión y Clasificación, en adelante CART (de sus siglas en inglés, Classification And Regression Trees), propuesta por [4] Breiman et al. (1984).

El objetivo de esta técnica es, conocidos los grupos o categorías en los que se quiere clasificar los individuos u objetos, ubicar dichos individuos dentro de estas categorías a partir de los valores de ciertos parámetros.

Breiman, desarrolló el algoritmo CART cuyo resultado es en general, un árbol de decisión, las ramas representan conjuntos de decisiones y cada decisión genera reglas sucesivas para continuar la clasificación (partición) formando así grupos homogéneos respecto a la variable que se desea discriminar. Las particiones se hacen en forma recursiva hasta que se alcanza un criterio de parada, el método utiliza datos históricos para construir el árbol de decisión, y este árbol se usa para clasificar nuevos datos.

CART es un método no-paramétrico de segmentación binaria donde el árbol es

construído dividiendo repetidamente los datos. En cada división los datos son partidos en dos grupos mutuamente excluyentes. El nodo inicial es llamado nodo raíz o grupo madre y se divide en dos grupos hijos o nodos, luego el procedimiento de partición es aplicado a cada grupo hijo por separado. Las divisiones se seleccionan de modo que "la impureza" de los hijos sea menor que la del grupo madre y éstas están definidas por un valor de una variable explicativa ([10] Deconinck et al., 2006). El objetivo es particionar la respuesta en grupos homogéneos y a la vez mantener el árbol razonablemente pequeño. Para dividir los datos se requiere un criterio de particionamiento el cual determinará la medida de impureza, esta última establecerá el grado de homogeneidad entre los grupos.

El análisis de árboles de clasificación y regresión (CART) generalmente consiste en tres pasos ([32] Timofeev, 2004):

1. Construcción del árbol máximo.
2. Poda del árbol.
3. Selección del árbol óptimo mediante un procedimiento de validación cruzada ("cross-validation").

Construcción del árbol máximo

El árbol máximo es construido utilizando un procedimiento de partición binario, comenzando en la raíz del árbol, este árbol es un modelo que describe el conjunto de entrenamiento (grupo de datos original) y generalmente es sobreajustado, es decir, contiene gran cantidad de niveles y nodos que no producen una mejor clasificación

y puede ser demasiado complejo. Cada grupo es caracterizado por la distribución (respuesta categórica), o por la media (respuesta numérica) de la variable respuesta, el tamaño del grupo y los valores de las variables explicativas que lo definen. Gráficamente, el árbol se representa con el nodo raíz (los datos sin ninguna división), al iniciar y las ramas y hojas debajo (cada hoja es el final de un grupo).

Calidad del Nodo: Función de Impureza

La función de impureza es una medida que permite determinar la calidad de un nodo, esta será denotada por $i(t)$. Existen varias medidas de impureza (criterios de particionamiento) que nos permiten analizar varios tipos de respuesta, las dos medidas más comunes presentadas por Breiman, para árboles de clasificación son:

- El índice de información o entropía el cual se definde como:

$$i(t) = \sum_{j} p\left(j|t\right) \ln p\left(j|t\right) \tag{4.1.1}$$

  El objetivo es encontrar la partición que maximice $\Delta i(t)$ en la ecuación 4.1.2

$$\Delta i(t) = -\sum_{j=1}^{k} p\left(j|t\right) \ln p\left(j|t\right), \tag{4.1.2}$$

  donde $j = 1, ..., k$ es el número de clases de la variable respuesta categórica y $p\left(j|t\right)$ la probabilidad de clasificación correcta para la clase j en el nodo t.

- El índice Gini tiene la forma

$$i(t) = \sum_{i \neq j} p\left(j|t\right) p\left(i|t\right) \tag{4.1.3}$$

  Encontrar la partición que maximice $\Delta i(t)$ en 4.1.4

$$\Delta i(t) = -\sum_{j=1}^{k} [p\left(j|t\right)]^2. \tag{4.1.4}$$

Este índice es el más utilizado. En cada división el índice Gini tiende a separar la categoría más grande en un grupo aparte, mientras que el índice de información tiende a formar grupos con más de una categoría en las primeras decisiones.

Poda del árbol

El árbol obtenido es generalmente sobreajustado por tanto es podado, cortando sucesivamente ramas o nodos terminales hasta encontrar el tamaño "adecuado" del árbol. Breiman et al. introducen algunas ideas básicas para resolver el problema de seleccionar el mejor árbol. Computacionalmente el procedimiento descrito es complejo. Una forma es buscar una serie de árboles anidados de tamaños decrecientes ([9] Deáth & Fabricius, 2000), cada uno de los cuales es el mejor de todos los árboles de su tamaño. Estos árboles pequeños son comparados para determinar el óptimo. Esta comparación está basada en una función de costo complejidad , $R_\alpha(T)$. Para cada árbol T, la función costo-complejidad se define como:

$$R_\alpha(T) = R(T) + \alpha|\tilde{T}|, \tag{4.1.5}$$

donde $R(T)$ es el promedio de la suma de cuadrados entre los nodos, puede ser la tasa de mala clasificación total o la suma de cuadrados de residuales total dependiendo del tipo de árbol, $|\tilde{T}|$ es la complejidad del árbol, definida como el número total de nodos del sub-árbol y α es el parámetro de complejidad.

El parámetro α es un número real mayor o igual a cero. Cuando $\alpha = 0$ se tiene el árbol más grande y a medida que α se incrementa, se reduce el tamaño del árbol. La función $R_\alpha(T)$ siempre será minimizado por el árbol más grande, por tanto se necesitan mejores estimaciones del error, para esto Breiman et al. proponen obtener

estimadores "honestos" del error por "validación cruzada". Computacionalmente el procedimiento es exigente pero viable, pues solo es necesario considerar un árbol de cada tamaño, es decir, los árboles de la secuencia anidada.

Selección del árbol óptimo

De la secuencia de árboles anidados es necesario seleccionar el árbol óptimo y para esto no es efectivo utilizar comparación o penalización de la complejidad, por tanto se requiere estimar con precisión el error de predicción y en general esta estimación se hace utilizando un procedimiento de validación cruzada. El objetivo es encontrar la proporción óptima entre la tasa de mala clasificación y la complejidad del árbol, siendo la tasa de mala clasificación el cociente entre las observaciones mal clasificadas y el número total de observaciones. El procedimiento de validación cruzada puede implementarse de dos formas:

- Si se cuenta con suficientes datos se parte la muestra, sacando la mitad o menos de los datos y se construye la secuencia de árboles utilizando los datos que permanecen, luego predecir, para cada árbol, la respuesta de los datos que se sacaron al iniciar el proceso; obtener el error de las predicciones; seleccionar el árbol con el menor error de predicción.

En general no se cuenta con suficientes datos como para utilizar el procedimiento anterior, de modo que otra forma sería:

- Validación cruzada con partición en V, (v-fold cross validation, se menciona más adelante).

La idea básica de la "Validación cruzada" es sacar de la muestra de aprendizaje una muestra de prueba, con los datos de la muestra de aprendizaje se calculan los estimadores y el subconjunto sacado es usado para verificar el desempeño de los estimadores obtenidos utilizandolos como "datos nuevos". El desempeño entendido como el error de predicción, es acumulado para obtener el error medio absoluto del conjunto de prueba.

Como se mencionó anteriormente, para la metodología CART generalmente se utiliza Validación Cruzada con partición en V (v-fold cross validation), tomando $V = 10$ y el procedimiento es el siguiente:

- Dividir la muestra en diez grupos mutuamente excluyentes y de aproximadamente igual tamaño.
- Sacar un conjunto por vez y construír el árbol con los datos de los grupos restantes. El árbol es usado para predecir la respuesta del conjunto eliminado.
- Calcular el error estimado para cada subconjunto. Repetir los items dos y tres para cada tamaño de árbol.
- Seleccionar el árbol con la menor tasa de mala clasificación.

Al llegar a este punto se procede a analizar el árbol obtenido.

4.1.1. Modelado CART en R

Los árboles de Clasificación y Regresión se pueden generar a través del paquete **rpart**. A continuación se proporcionan los pasos generales para su implementación.

1. Construcción del árbol

Para construir el árbol, usamos

rpart(*formula*, **data**=, **method**=,**control**=) donde

formula	está en el formato resultado ∼ predictor1+predictor2+predictor3+ect.
data=	especifica el marco de datos
method=	"class" para un árbol de clasificación "anova" para un árbol de regresión
control=	parámetros opcionales para el control de crecimiento de los árboles. Por ejemplo, control = rpart.control (minsplit = 30, cp = 0,001) requiere que el número mínimo de observaciones en un nodo sea 30 antes de intentar una división y que una división debe disminuir la falta general de ajuste por un factor de 0.001 (factor de complejidad coste) antes de ser tratado.

2. Examinar los resultados

Las siguientes funciones nos ayudan a examinar los resultados.

printcp(*fit*)	muestra la tabla cp
plotcp(*fit*)	dibuja los resultados de la cross-validation
rsq.rpart(*fit*)	dibuja los residuos cuadrados aproximados y el error relativo para diferentes divisiones (2 dibujos). Las etiquetas son solo apropiadas para el método "anova".
print(*fit*)	muestra los resultados
summary(*fit*)	resultados detallados incluyendo divisiones sustitutas
plot(*fit*)	dibuja el árbol de decisión
text(*fit*)	etiqueta el diagrama del árbol de decisión

En árboles creados por rpart (), pasar a la rama izquierda cuando la condición establecida es cierta.

3. Podar el árbol

Como hemos dicho se debe podar el árbol para evitar sobreajustes en los datos. Por lo general, queremos seleccionar un tamaño de árbol que minimice el error de validación cruzada, la columna de la **xerror** impreso por **printcp ()**.

Para podar el árbol hasta el tamaño usamos

prune(fit, cp=)

En concreto, usaremos **printcp ()** para examinar los resultados de error con validación cruzada, seleccionaremos el parámetro complejidad asociada con el

error mínimo, y lo colocaremos en la función **prune ()**. Alternativamente, se puede utilizar el fragmento del código siguiente
`fit$cptable[which.min(fit$cptable[,"xerror"]),ÇP"]` para seleccionar automáticamente el parámetro complejidad asociada con el error de validación cruzada más pequeño.

4.1.2. Aplicación

Los datos consisten en las medidas de 150 cráneos de varones Egipcios de 5 periodos de tiempo diferentes (-4000, -3300, -1850, -200, 150). Los datos y la fuente original pueden encontrarse en [20]. El objetivo es discriminar (diferenciar) los diferentes periodos de tiempo según las medidas de los cráneos. Se han medido 30 cráneos para cada periodo. Se han tomado cuatro medidas de cada cráneo:

$V1 \rightsquigarrow$ Amplitud máxima del cráneo.

$V2 \rightsquigarrow$ Altura máxima del cráneo (Altura Basibregmatic).

$V3 \rightsquigarrow$ Longitud basialveolar del cráneo (mínima distancia entre los puntos basion y alveolar).

$V4 \rightsquigarrow$ Altura nasal del cráneo.

La Figura 4.1 muestra el árbol de clasificación final. Cada nodo final contiene una etiqueta que indica en cual de los 5 periodos está clasificado un cráneo medido según el camino que va desde el nodo original al nodo final.

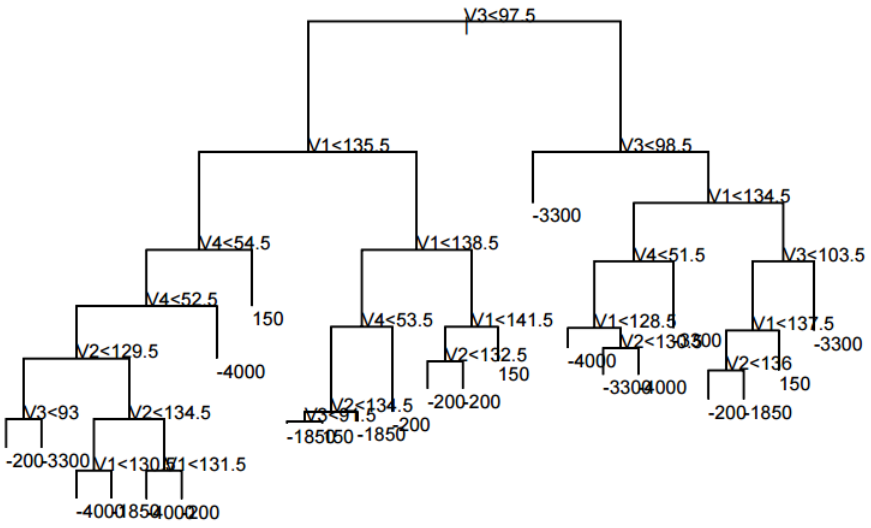

Figura 4.1: CART aplicado a la muestra de cráneos egipcios.

4.2. Métodos Bayesianos. Datación por radiocarbono

El enfoque Bayesiano en estadística no es en absoluto nuevo. Sin embargo consideramos apropiado incluirlo en este capítulo por varias razones. En primer lugar, en los últimos años la contribución Bayesiana a las investigaciones estadísticas ha aumentado de manera espectacular. Por otra parte, el uso de ordenadores potentes permite dar una respuesta Bayesiana a muchos problemas que hace unos cuantos años atrás eran inaccesibles para la estadística Bayesiana. Un resumen de la metodología Bayesiana podría ser el siguiente:

1. La información a priori se expresa como una distribución de probabilidad sobre el espacio paramétrico.

2. La función de verosimilitud es, de hecho, la distribución condicional de las observaciones dado los valores de los parámetros.

3. El Teorema de Bayes se usa para combinar la información a priori con la información experimental y las transforma en la información posterior: otra distribución de probabilidad sobre el espacio paramétrico.

Algunos aspectos positivos de la Estadística Bayesiana son los siguientes. El enfoque Bayesiano es conceptualmente atractivo (y simple). Por ejemplo, la probabilidad de que un parámetro pertenezca a un intervalo de confianza al 95 % Bayesiano es realmente 0.95. Además, es posible incluir antes la información cualitativa en el proceso de inferencia. También es posible actualizar progresivamente las creencias: las información "a posteriori" de hoy es la información "a priori" de mañana.

Por otro lado, esta metodología también presenta algunas dificultades. Siempre se necesita una distribución a priori, incluso si no tenemos tal información "a priori", y algunos resultados dependen fuertemente de ella. Además, en los problemas con tamaños de muestras grandes y medianos, el cálculo de la distribución a posteriori es extremadamente difícil. Muchas veces solo están disponibles soluciones aproximadas.

A continuación veremos como esta metodología está presente en algunos de los procesos arqueológicos más importantes y más útiles en la actualidad. Hablamos del proceso de **datación por radiocarbono.** En las próximas secciones describiremos en qué consiste dicho proceso para posteriormente pasar a ver un ejemplo práctico; pero antes de eso haremos una breve exposición sobre el significado que tiene aplicar el enfoque bayesiano a esta técnica.

La estadística bayesiana se basa en el teorema que enunció Thomas Bayes, sobre la probabilidad de un suceso condicionado por la ocurrencia de otro suceso. En esencia, los seguidores de la estadística tradicional sólo admiten probabilidades basadas en experimentos repetibles y que tengan una confirmación empírica mientras que los llamados estadísticos bayesianos permiten probabilidades subjetivas. El teorema puede servir entonces para indicar cómo debemos modificar nuestras probabilidades subjetivas cuando recibimos información adicional de un experimento. La estadística bayesiana demuestra su utilidad en ciertas estimaciones basadas en el conocimiento subjetivo a priori y en el hecho de permitir revisar esas estimaciones en función de la evidencia empírica. De este modo, como indica [1], la "cronología objetiva" de Renfrew se ve contaminada por las opiniones arqueológicas conocidas de forma previa a la realización de las dataciones.

Los estimadores "a priori" son aquellos conocimientos que se tienen del yacimiento (estratigrafía, tipología, monedas, crecimiento de anillos de árboles, etc.). Incluyendo estos datos en un modelo cronológico durante la calibración de las fechas radiocarbónicas, podemos obtener distribuciones de probabilidad de las fechas, ya calibradas, que incorporan estos condicionantes, que a partir de ahora no podrán ser consideradas como evidencias científicas independientes.

Dada la importancia que en el desarrollo de la cronología empleando estadística bayesiana tienen los estimadores "a priori" es crucial insistir en que esta información debe conocerse de forma previa a la realización de las dataciones. El caso más paradigmático es la estratigrafía, donde se establece una relación entre las distintas unidades estratigráficas, y, de este modo, entre las muestras que en ellas aparecen.

Aquellas muestras que aparecen en niveles superiores deben normalmente ser más modernas que las que se encuentran en los niveles inferiores. Así, resulta fundamental establecer de un modo indiscutible que la muerte y deposición del material que va a ser analizado ha sido de forma coetánea a la formación del contexto en el que se halla. Más que nunca, conviene recordar que la datación por carbono-14 se realiza sobre materiales y, por tanto, solo determina la edad de éstos, no de los contextos en los que se encuentran. No solo se requieren muestras de vida corta e identidad única, sino que se precisan datos incuestionables sobre la deposición de los mismos en el depósito arqueológico.

Además de este tipo de relaciones, se incluye en el modelo cronológico todos aquellos datos susceptibles de aportar información cronológica: agrupación de fechas por fases de actividad arqueológica, fechas que son el origen o la terminación de una secuencia, fechas que están ordenadas dentro de una secuencia, eventos puntuales incluidos entre dos fases de actividad, y un largo etcétera.

En estudios cronológicos existen dos tipos de información: fechas de calendario, que sitúan los sucesos en escalas de tiempo absolutas (reinado, documento fechado, fechas obtenidas por métodos de datación); y fechas relativas, que son aquellas procedentes de la estratigrafía, estudio de los materiales, agrupamiento de unidades estratigráficas en fases ordenadas o no, etc. La primera suele referirse normalmente a fechas de una muestra en concreto, mientras que las fechas relativas generan relaciones más complejas entre los momentos en los que se desarrollan los eventos del yacimiento en cuestión. Por esta razón, al aplicar la estadística bayesiana las fechas de calendario se emplean como los intervalos temporales de probabilidad, por ejemplo, una fecha carbono-14 calibrada, mientras que las fechas relativas se muestran como las

probabilidades a priori.

4.2.1. Introducción a la Datación por Radiocarbono

La datación por radiocarbono es una técnica científica usada rutinariamente por arqueólogos para datar la materia orgánica encontrada en los yacimientos arqueológicos. Por lo general, se toman varias muestras de material orgánico de cada yacimiento o grupo de yacimientos para datarlas, y posteriormente inferir respecto al intervalo o periodo de tiempo representado por estas muestras.

Como consecuencia de ello, es necesario resumir el conjunto de muestras de radiocarbono obternidas las cuales, como veremos, pueden formularse como un problema de inferencia estadística. En este trabajo nos centraremos en estudiar dichos problemas mediante un enfoque Bayesiano. La datación por radiocarbono es un método que involucra tanto procesos físicos como químicos, a través del cual midiendo la proporción de carbono-14, ^{14}C, y carbono-12, ^{12}C, en un objeto, estiman su edad usando la ley de la decadencia del radiocarbono. El resultado final de este proceso de datación, es una muestra de radiocarbono que consiste en una estimación de los "años de radiocarbono" denotados por y BP (antes del presente, concretamente es el número de años antes del 1950 d.C) y en una desviación típica σ que refleja la incertidumbre en el proceso. De forma que la muestra de radiocarbono se expresa como $y \pm \sigma$.

Las muestras de radiocarbono necesitan calibrarse para transformar los años de radiocarbono en años naturales; actualmente, esto se lleva a cabo mediante el uso de la curva de calibración lineal a trozos acordada internacionalmente, la cual se denota por $\mu(\theta)$. El modelo de probabilidad de estas muestras desde un punto de vista Bayesiano consite en:

Sea y un año de radiocarbono, asumimos que $y \,|\, \sigma, \theta \sim N\{\mu(\theta), \sigma^2\}$, donde θ es el año natural en el cual el material orgánico contenido en el objeto datado murió (el año natural asociado al objeto) y σ es la desviación típica reportada por el laboratorio. Por lo tanto el modelo expone que una muestra de radiocarbono se distribuye según una normal de media $\mu(\theta)$ (el año de radiocarbono correspondiente al año natural θ) y varianza σ^2. La forma de proceder es asumir σ como conocido. Así para facilitar la notación eludimos condicionar en σ y simplemente escribimos

$$y \,|\, \theta \sim N\{\mu(\theta), \sigma^2\}$$

Suponemos ahora que tenemos un conjunto de muestras de radicarbono $y_1 \pm \sigma_1, y_2 \pm \sigma_2, ..., y_m \pm \sigma_m$ asociadas con los años naturales desconocidos $\theta_1, \theta_2, ..., \theta_m$. Supongamos además que las muestras de radiocarbono pertenecen a objetos relacionados con una etapa arqueológica dada (e.g. un yacimiento arqueológico en particular, una cultura dada, etc), con la consecuencia que los años naturales pertenecen al periodo de tiempo de dicha etapa. Por ejemplo, a menudo las muestras que se usan para la datación por radiocarbono son tomadas deliberadamente de contextos que muestran una asociación clara y sin ambigüedades con tipos de cerámica o artefactos similares a los que usan para definir una etapa.

Comúnmente los arqueólogos tienen esa información, posiblemente con algún conocimiento previo más acerca de la duración del periodo de tiempo de la etapa. En general, sin embargo, sólo hay informaciones a priori imprecisas sobre la relación interna entre los θ_j. En tales circunstancias, los arqueólogos desean resumir las muestras de radiocarbono y inferir sobre el periodo de tiempo de la etapa, intentando combinar las evidencias aportadas por dichas muestras con la información arqueológica que

tienen previamente sobre la etapa.

Metodología

Supongamos que tenemos un conjunto de muestras de radiocarbono $y_1 \pm \sigma_1, y_2 \pm \sigma_2, ..., y_m \pm \sigma_m$ asociadas con los años naturales desconocidos $\theta_1, \theta_2, ..., \theta_m$ y con la información a priori de que esos años naturales pertenecen al periodo de tiempo de una etapa arqueológica en particular. El enfoque que proponemos para este problema es suponer que la distribución a priori de cada θ_j está en una forma paramétrica dada por un vector de parámetros ψ; entonces representamos la información arqueológica de que todos los θ_j pertenecen a una etapa individual haciendo ψ común para todos los θ_j. Podemos poner esto en términos probabilísticos diciendo que la distribución a priori para cada θ_j es $f(\theta_j \,|\, \psi)$ para $j = 1, 2, ..., m$. Una elección de $f(\theta_j \,|\, \psi)$ que puede ser adecuada para varias aplicaciones específicas es $\psi = (\alpha, \beta)$ y

$$\theta_j \,|\, \alpha, \beta \sim U(\alpha, \beta), j = 1, 2, ..., m.$$

es decir, los θ_j están uniformemente distribuidos en un intervalo de tiempo que comienza en β y acaba en α (años BP). Es muy útil exponer el modelo en esta forma jerárquica dado que los arqueólogos normalmente tienen algún tipo de información a priori sobre el periodo de tiempo y la posición absoluta en la escala natural de la etapa bajo estudio, y esto se podría trasladar en una distribución para ψ. Sea esta distribución a priori $f(\psi)$.

Nuestro cometido ahora es obtener la distribución a posteriori de ψ dadas las muestras de radiocarbono $y_1 \pm \sigma_1, y_2 \pm \sigma_2, ..., y_m \pm \sigma_m$. Si aceptamos los supuestos y la información a priori indicada anteriormente, debemos tomar esta distribución para

representar nuestro conocimiento actual del periodo de tiempo de la etapa arqueológica bajo consideración. Un ejemplo para las distribuciones a priori de α y β que puede utilizarse para la especificación de la información a priori sobre varios casos diferentes de etapas arqueológicas es

$$\alpha \sim U(a_1, b_1) \text{ y} \beta \sim U(a_2, b_2) \tag{4.2.1}$$

para unas constantes positivas $a_1 < b_1 < a_2 < b_2$, es decir, distribuciones uniformes con algunos márgenes que no se solapan. Como las distribuciones no se solapan podemos suponer que $f(\alpha, \beta) = f(\alpha)f(\beta)$.

Distribuciones a posteriori

Para obtener las distribuciones a posteriori de α y β usamos el "Muestreo de Gibbs". El algoritmo de muestreo de Gibbs se usa iterativamente y nos permite actualizar la información y volver a muestrear hasta obtener la convergencia para la distribución bajo investigación. Suponemos que tenemos un parámetro n-dimensional u, cuya distribución a posteriori se denota por $f(u) = f(u_1, u_2, ..., u_{i-1}, u_{i+1}, ..., u_n)$. El esquema de muestreo de Gibbs requiere la elección inicial de valores para $u_2, u_3, ..., u_n$ (es decir $u^{(0)} = (u_1^{(0)}, u_2^{(0)}, ..., u_n^{(0)})$. Entonces $u_1^{(1)}$ se genera de la distribución condicionada $f(u_1 \,|\, u_2^{(0)}, u_3^{(0)}, ..., u_n^{(0)})$. El siguiente, $u_2^{(1)}$, se genera de la distribución condicionada $f(u_2 \,|\, u_1^{(0)}, u_3^{(0)}, ..., u_n^{(0)})$. El proceso continúa hasta generar $u_n^{(1)}$. Este proceso de pasar de $u^{(k)}$ a $u^{(k+1)}$ forma un proceso de Markov y Geman y Geman (1984) demostraron que

$$u^{(k)} \to^d u \sim f(u) \quad \text{cuando} \quad k \to \infty.$$

Por lo tanto, para grandes k, $u^{(k)}$ puede verse como una realización de nuestro vector de parámetros u. Repitiendo el proceso s veces daremos una muestra $u_1^{(k)}, u_2^{(k)}, ..., u_s^{(k)}$ de u y a cada muestra podemos aproximar $f(u)$ o cualquiera de sus marginales a posteriori $f(u_i)$, o momentos a posteriori.

Software informáticos

Por supuesto, la verdadera potencia de trabajar con muchas fechas de Carbono 14 es que, como hemos dicho, se pueden calibrar y someter a distintos procedimientos estadísticos. Por ello, mencionaremos los software más destacados a la hora de trabajar con datos de ^{14}C:

- **Bcal**: Es una herramienta en línea para hacer calibraciones bayesianas de dataciones de Carbono 14. Tiene una interfaz compleja, porque no sólo sirve para calibrar fechas, sino que también permite otras opciones: Se puede introducir información a priori que el programa tiene en cuenta, y calibrar de forma conjunta grupos de fechas para hallar la antigüedad de un estrato, una estructura, etc.

- **CalPal**: Aunque hace tiempo que no se actualiza, Calpal sigue siendo una buena opción para calibraciones rápidas y usuarios "no avanzados". Tiene tanto un calibrador on-line (con las opciones más básicas) como un completo software de calibración para descargar e instalar en tu sistema. Las curvas de calibración de CalPal, además, incluyen la información paleoclimática de las últimas glaciaciones.

- **14 Chrono Center**: Esta página de la Queen's University de Belfast contiene dos programas de utilidad general y otro que es más bien una çuriosidad". Los programas más generales son el Calib, una herramienta de calibración bastante completa, y una herramienta online para la corrección del .[ef]ecto reservorio.[en] las muestras de origen marino. El programa tipo çuriosidad.[es] Calibomb, una herramienta para calibrar dataciones de la época post-atómica, es decir de los últimos setenta años.

- **OxCal**: es la herramienta de Calibración de la Universidad de Oxford. El programa OxCal está destinado a proporcionar calibraciones de radiocarbono y análisis de información cronológica ambiental y arqueológica.

4.2.2. Aplicación

Esta sección de aplicación se basará en los estudios llevados a cabo por [23] sobre la cultura peruana pre-hispánica llamada "Chancay". Allí encontramos 13 determinaciones de radiocarbono procedentes de muestras de carbón tomadas de las tumbas asociadas con esta cultura (Tabla 4.1).

Identificación de la muestra	Determinación ^{14}C (años BP)
Gd-2819	520 ± 60
Gd-3396	430 ± 30
Gd-5304	460 ± 50
Gd-5307	970 ± 50
Gd-5309	910 ± 35
Gd-5310	1000 ± 50
Gd-5312	390 ± 45
Gd-5672	830 ± 50
Gd-5823	670 ± 40
Gd-5824	1140 ± 50
Gd-6189	1070 ± 60
Gd-6196	810 ± 70
Gd-6197	900 ± 70

Tabla 4.1: Determinaciones de radiocarbono para la cultura Chancay, Perú.

Padzur y Krzanowski usaron una combinación de técnicas heurísticas y software computacionales para investigar sus conclusiones sobre el periodo de tiempo de existencia de la cultura Chancay en base a las muestras de radiocarbono que estaban disponibles.

Nuestra intención presentando este ejemplo no es dar un estudio arqueológico de la cultura Chancay sino más bien es ilustrar las técnicas presentadas en esta sección y demostrar cómo se pueden aplicar usando un conjunto específico de muestras de

radiocarbono. Usamos el modelo dado anteriormente con $\psi = (\alpha, \beta)$ y $m = 13$ y tomamos distribuciones a priori uniformes como el el modelo (4.2.1). La cultura Chancay es pre-hispánica, y esto significa que el final de su intervalo de tiempo (α) debe ser antes de la invasión de Perú por Pizarro en el siglo XVIII. Decidimos fijar $a_1 = 400$ BP (1500 d.C) como una cota final para la distribución a priori de α. Fijamos b_1, a_2, b_2 de una manera menos informativa, dándoles valores extremos. En conjunto, la información a priori dada es imprecisa (aparte del valor de a_1), lo que significa que las distribuciones a posteriori de α y β se basarán más en los datos y en las muestras de radiocarbono, y menos en las consideraciones arqueológicas.

Usando las distribuciones a posteriori condicionadas obtenidas anteriormente para el caso $\psi = (\alpha, \beta)$ y considerando las distribuciones a priori ya fijadas, obtenemos las distribuciones a posteriori condicionadas para este ejemplo. El muestreo de Gibbs se implementó en un ordenador y obtuvimos las distribuciones a posteriori marginales de α y β mostradas como histogramas en la Fig. El algoritmo de muestreo Gibbs se ha ejecutado con diferentes valores iniciales, con $k_1 = 5000$, $q = 12$ y $s = 10000$, y también con $q = 1$ y $s = 30000$. Los histogramas resultantes fueron casi idénticos para todas las ejecuciones.

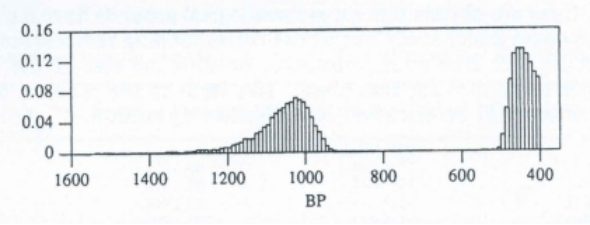

Figura 4.2: Histogramas de las distribuciones a posteriori de β (lado izquierdo) y α (lado derecho).

De las distribuciones marginales de α y β vemos que, dada la muestra actual, al 95 % la región de densidad a posteriori más alta (HPD) para α es aproximadamente (480,400) BP (1470-1550 d.C) y para β (1200,950) BP (850-1100 d.C), con modas en $\alpha = 440$ y $\beta = 1020$ BP (1510 y 930 d.C respectivamente). En [23] encontramos una estimación anterior ,que no se basa en muestras de radiocarbono, para el rango de la cultura Chancay que es 900-1479 d.C y esto es consistente con los resultados aquí comentados.

Bibliografía

[1] A. Bayliss. Rolling out revolution: using radiocarbon dating in archaeology. *Radiocarbon*, 51(1):123–147, 2009.

[2] S. Holland P. Bishop, Y.; Fienberg. *Discrete Multivariate Analysis: Theory and practice*. The MIT Press, 1975.

[3] S. Bochner. *Harmonic analysis and the Theory of Probability*. Univ. of California Press., 1955.

[4] Friedman J. H. Olshen R. A. Stone C. G. Breiman, L. *Classification and Regression Trees*. 1984.

[5] D. Brothwell. *Desenterrando Huesos. La Excavación, Tratamiento y Estudio de Restos del Esqueleto Humano*. Fondo de Cultura Económica, 1987.

[6] J.E. Buikstra and D.H. Ubelaker. Standards for data collection from human skeletal remains. *Arkansas Archaeological Survey Research* ., 44, 1994.

[7] M.; García Sanjuán L. y Wheatley D. W. Costa Caramé, M.E.; Díaz Zorita Bonilla. The copper age settlement of valencina de la concepción (seville, spain): Demography, metallurgy and spatial organization. 2010.

[8] M.E Costa Caramé. Las producciones metálicas del iii y ii milenio, cal ane en el suroeste de la península ibérica. 2010.

[9] K. E. Death, G. Fabricius. Classification and regression trees: A powerful yet simple technique for ecological data analysis. *Ecology*, 81(11):3178–3192, 2000.

[10] Zhang M. H. Coomans D. Heyden Y. V. Deconinck, E. Classification tree models for the prediction of blood-brain barrier passage of drugs. *Journal of Quemical Information and Modeling*, 46(3):1410–1419., 2006.

[11] R. A. Fisher. On the interpretation of chi^2 from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*, 85(1):87–94, 1922.

[12] R. A. Fisher. *Statistical Methods for Research Workers*. Oliver and Boyd, 1 edition, 1925.

[13] R. A. Fisher. The logic of inductive inference. *J.R. Statist. Soc.*, 98:39–54, 1935.

[14] E. Fix and J.L. Hodges. *Discriminatory analysis, nonparametric estimation: consistency properties*. Report No 4, Project no 21-49-004, USAF School of Aviation Medicine, 1951.

[15] Halton J.H. Freeman, G.h. Note on an exact treatment of contingency, goodness-of-fit and other problems of significance. *Biometrika*, 38:141–149, 1951.

[16] S. J. Haberman. A warning on the use of chi-squared statistics with frequency tables with small expected cell counts. *J. Amer. Statist. Assoc.*, 83:555–560, 1988.

[17] M.A. Hunt Ortiz. Prehistoric mining and metallurgy in southwest iberian peninsula. 2003.

[18] Marron J.S. Jones, M.C. and Sheather S.J. 1966.

[19] K. Koehler and K. Larntz. An empirical investigation of goodness-of-fit statistics for sparse multinomials. *J. Amer. Statist. Assoc.*, 75:336–344, 1980.

[20] B. F. J. Manly. *Multivariate Statiscal Methods: A Primer. 2nd Edition).* 1994.

[21] C. N. Morris. Central limit theorems for multinomial sums. *Ann. Statist.*, 3:165–188, 1975.

[22] E. Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076, 1962.

[23] M. F. Pazdur and A. Krzanowski. Fechados radiocarbónicos para los sitios de la cultura chancay. *Estudios Sobre la Cultura Chancay, Perú*, pages 155–132, 1991.

[24] K. Pearson. Mathematical contributions to the theory of evolution. vii. on the correlation of characters not quantitatively measurable. *Philosophical Transactions of the Royal Society of London. Series A, containing Papers of a Mathematical or Physical Character*, 195:1–47, 1900.

[25] K. Pearson. Report on certain enteric fever inoculation statistics. *British Medical Journal*, 3:1243–1246, 1904.

[26] W.R. Perizonius. Closing and nono.closing sutures in 256 crania of known age and sex from amsterdam a.d. 1883-1909. *Journal of Human Evolution*, 13(2):201–216, 1984.

[27] M. Rodríguez Bayona. La investigación de la actividad metalúrgica durante el iii milenio a.n.e en el suroeste de la península ibérica. la arqueometalurgia y la aplicación de análisis metalográficos y composicionales en el estudio de la producción de artefactos de metal. 2008.

[28] M. Rosenblatt. Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 27(3):832–837, 1956.

[29] D. W. Scott. *Multivariate density estimation: Theory, practice, and visualization*. John Wiley Sons., 1992.

[30] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman Hall, 1986.

[31] Student. The probable error of a mean. *Biometrika*, 6:1–25, 1908.

[32] R. Timofeev. *Classification and regression trees (cart). theory and applications. Master thesis*. 2004.

[33] G.U. Yule. On the association of attributes in statistics. *Philosophical Transactions of the Royal Society A*, 75:257–319, 1900.