VOLUMEN MONOGRÁFICO (1999), 263-290

INFORMÁTICA Y SOCIOLINGÜÍSTICA CUANTITATIVA

FERNANDO F. RAMALLO
Universidad de Vigo

RESUMEN. *La estadística es la herramienta más importante en la investigación de sociolingüística cuantitativa. El desarrollo de programas informáticos que permiten manipular enormes cantidades de datos, realizando una gran variedad de procedimientos estadísticos, ha contribuido de manera evidente a que las encuestas sociolingüísticas sean más fácilmente interpretables, menos onerosas y más fiables. Sin duda, la evolución de la sociolingüística cuantitativa ha estado determinada por el desarrollo y el perfeccionamiento de software estadístico. Hoy en día, disponemos de programas muy poderosos y de fácil uso que acercan al investigador a una amplia gama de herramientas de análisis exploratorio de datos, lo que hace que una persona con escasos conocimientos de estadística pueda resolver problemas impensables sin estas herramientas. En este trabajo presentamos una breve introducción al uso de la informática en la sociolingüística cuantitativa tomando como modelo el programa SPSS.*

PALABRAS CLAVE. *Cuantificación, sociolingüística, métodos estadísticos.*

ABSTRACT. *In this paper we present a brief introduction to the computer science and computer tools in quantitative sociolinguistics using SPSS. Statistics is one of the most important tools in quantitative sociolinguistics research. Development of computer programs allows to manipulate enormous quantities of data, carrying out a great variety of statistical procedures, and has contributed to surveys in sociolinguistics. Thus, they become more easily interpretable, less onerous and more reliable. There is no doubt that the evolution of quantitative sociolinguistics has been determined by development and improvement of statistical software. Today we have user friendly and powerful programs that bring the researcher near to a wide range of exploratory analysis tools. Therefore a person with scarce statistic knowledge may be able to solve problems with these tools.*

KEYWORDS. *Quantification, sociolinguistics, statistical methods*

1. INTRODUCCIÓN

Las nuevas tecnologías constituyen un capítulo imprescindible en el hacer de las ciencias sociales. El desarrollo de programas informáticos cada vez más sofisticados

ha proporcionado un marco excelente para llevar a cabo modelos de investigación impensables hace pocos años. En la actualidad, podemos manejar millones de datos en apenas unos minutos desde una máquina multimedia doméstica, lo que abarata de manera notable el proceso de investigación además de permitir un mayor control del mismo por parte del analista. Hasta hace pocos años, el elevado coste y la escasa presencia de recursos informáticos para ordenadores personales hacía casi inexcusable el trabajo en los centros de cálculo de las universidades o de alguna empresa privada.

En la medida en que los ordenadores se han constituido como instrumentos válidos para resolver problemas, la informática ha ido ocupando un lugar central dentro del conjunto de herramientas de trabajo necesarias para realizar una investigación. En este sentido, podemos afirmar que el desarrollo de la sociolingüística cuantitativa ha estado muy vinculado a la evolución de determinadas aplicaciones informáticas, en especial a los programas estadísticos, que permiten manipular ingentes cantidades de datos obtenidos en el trabajo empírico. En este relato ofrecemos una panorámica del aprovechamiento de la informática en el trabajo cotidiano de lo que se ha llamado sociolingüística cuantitativa.

Como es bien sabido, el par cuantitativo/cualitativo se expresa con cierta frecuencia en la metodología de las ciencias sociales en clara oposición. Lo que subyace a esta oposición es la distinción ontológica entre la cantidad (números) y la calidad (palabras). Los partidarios de alentar este debate de manera extrema suelen recurrir a argumentos que justifiquen su elección como el único referente válido en la investigación, ridiculizando, a veces, las técnicas opuestas, bien por frías, ideológicas, distantes, etc. (las cuantitativas) bien por carecer de mecanismos internos que garanticen la validez y la fiabilidad, por ser atomísticas, etc. (las cualitativas). En realidad, ambos procedimientos se complementan en la práctica, por lo que parece aconsejable renunciar a la creencia en la pureza de los conceptos: puede que los números nada *sean* sin palabras, pero estas no son inconmensurables. En palabras de Miguel Beltrán:

> Las ciencias sociales, por su parte, pueden y deben utilizar el método cuantitativo, pero sólo para aquellos aspectos de su objeto que lo exijan o permitan. Desde dos puntos de vista se ha vulnerado esta adecuación del método con el objeto: por una parte, un cierto humanismo delirante ha rechazado con frecuencia cualquier intento de considerar cuantitativamente fenómenos humanos o sociales apelando a una pretendida dignidad de la criatura humana que la constituiría en inconmensurable; de otro lado, una actitud compulsiva de constituir a las ciencias sociales como miembros de pleno derecho de la familia científica físico-natural ha llevado a despreciar toda consideración de fenómenos que no sea rigurosamente cuantitativa y formalizable matemáticamente. (Beltrán 1986: 33)

Por nuestra parte, creemos que lo que determina un paradigma frente a otro son los objetivos más que los propios objetos, ya que estos no son por naturaleza

cualitativos o cuantitativos; esta asignación procede más bien de la herramienta analítica utilizada. En consecuencia, el debate actual en las ciencias sociales propone la superación de esta vieja dicotomía, fomentando su integración y articulación para poder hacer frente de una manera más global a los nuevos desarrollos sociales, tal como se ha hecho en las ciencias naturales (Conde 1994: 68). Para comprender a fondo un objeto de estudio, tan necesario es profundizar en su dimensión cuantitativa como en la cualitativa. Así, en sociolingüística variacionista tan importante es conocer las variantes de una variable, por ejemplo, como la frecuencia y distribución de cada una de ellas. En este caso, es lógico pensar en una graduación dialéctica del estudio, es decir, para efectuar el análisis cuantitativo de la variación es necesario previamente conocer qué es lo que vamos a analizar (análisis cualitativo).

La cuantificación en lingüística tiene una historia claramente vinculada al desarrollo y crecimiento de la informática. El diseño de software relacionado con modelos lógicos y matemáticos de análisis ha permitido el trabajo con grandes cantidades de datos posibilitando su categorización y simplificando su tratamiento. Los trabajos pioneros en la cuantificación lingüística proceden, sin embargo, de una etapa anterior al uso de los ordenadores. Ya en las primera décadas del siglo XX encontramos intentos de cuantificar elementos lingüísticos, especialmente entre los componentes de la Escuela de Praga. Mathesius en 1911 inició una investigación que tenía como objetivo el estudio del sistema fonológico checo desde un punto de vista cuantitativo. Mostró, por ejemplo, que de todas las posibles combinaciones de fonemas, la lengua checa hacía uso de un 3.1%, mientras que en alemán este porcentaje alcanzaba el 5.4% (Tešitelová 1992). De esta manera abrió el camino a una fructífera metodología que llegará a nuestros días y que sirvió de modelo para muchos otros investigadores con intereses diversos, como el análisis de la extensión media de las palabras de textos diversos, la determinación de la frecuencia de un fenómeno lingüístico determinado, el análisis estadístico basado en el contenido o en la forma de un texto (estilometría), etc.

Este tipo de trabajos, con todo su interés, no debe llevarnos a pensar que la lingüística cuantitativa significa sólo *contar* fenómenos lingüísticos. Podríamos decir que dar cuenta de la cantidad es condición necesaria para hacer lingüística cuantitativa pero no suficiente. La cuantificación es una estrategia que podemos aprovechar, pero nunca debería ser el objeto de estudio de la lingüística cuantitativa, que, más allá de ofrecer recuentos, tratará de aportar interpretaciones sobre el objeto cuantificado. Como hemos dicho antes, partimos de que la investigación cuantitativa y la cualitativa se complementan e integran en la práctica investigadora.

Una de las disciplinas lingüísticas que más ha aprovechado la metodología cuantitativa ha sido la sociolingüística. Desde una perspectiva amplia, la sociolingüística se configura como un campo del saber cuyo objetivo principal es profundizar en las relaciones entre lo lingüístico y lo social. Es en el marco de estas relaciones donde se establecen los parámetros metodológicos y epistemológicos que

delimitarán el hacer sociológico-lingüístico frente a otras formas de conocimiento. Desde sus principios, la sociolingüística no permaneció ajena al debate cantidad vs. cualidad, tal como lo demuestra el reconocimiento general al menos de dos grandes corrientes denominadas sociolingüística cuantitativa y sociolingüística cualitativa[1].

2. SOCIOLINGÜÍSTICA CUANTITATIVA

La sociolingüística es una disciplina reciente que emerge en la intersección de dos ciencias que se desarrollaron independientemente en un período que abarca los dos últimos siglos, la sociología y la lingüística, y que se ha visto también influida por otros campos del conocimiento como la psicología social o la antropología. Sin embargo, la preocupación por la dimensión social en el lenguaje se inicia de manera sistemática a mediados del siglo XX, una vez que la enorme influencia de las diversas escuelas estructuralistas comienza a declinar. Con todo, sabemos que ha habido algunos precursores décadas antes. En este escenario debemos destacar la repercusión que tuvieron los estudios dialectológicos de finales del siglo pasado en la configuración de la actual sociolingüística. Tal como señala Mauro Fernández (1996: 126) "la conciencia de la existencia de diferencias lingüísticas entre los diversos estratos sociales fue uno de los más poderosos estímulos de las grandes investigaciones dialectales de fines del siglo XIX", que posiblemente dejaron una huella importante en los autores que en las primeras décadas de la segunda mitad del siglo XX diseñaron los primeros modelos teóricos y analíticos que sirvieron como referente de este nuevo campo de conocimiento.

El campo de estudio de la sociolingüística es muy amplio, tal como lo atestiguan diversos manuales que se han publicado en los últimos años (Fasold 1984, 1990; Holmes 1992; Lastra 1992; Wardhaugh 1992; Romaine 1994; Coulmas 1996; Hudson 1996; Moreno Fernández 1998; Spolsky 1998). El abanico de temas incluido en algunos de ellos es tan diverso que no es fácil dar una definición que satisfaga a todo el mundo. Incluso si partimos de una distinción entre macro y micro, encontramos tal cantidad de métodos, modelos e intereses en cada uno de los polos que obliga a hacer una matización del alcance de esta distinción, que sólo debe ser entendida epistemológicamente, es decir, como niveles de análisis. En el nivel micro se analizan los procesos interpersonales mientras que en el nivel macro son los procesos colectivos los pertinentes y el principal interés pasa a ser analizar la distribución de los miembros de una población a lo largo de una serie de parámetros lingüísticos y sociológicos que permiten agrupar o diferenciar a los individuos. Lo más importante en este último tipo de análisis es proporcionar modelos que den cuenta del grado en que estos parámetros correlacionan entre sí. Desde la perspectiva macrosociolingüística se da cuenta de aspectos tan diversos como las actitudes lingüísticas, el contacto de lenguas, el mantenimiento y cambio lingüístico, la demografía lingüística, la

planificación lingüística, las lenguas *pidgin* y las lenguas criollas, el multilingüismo, etc. Es decir, la macrosociolingüística toma las sociedades como punto de partida y analiza el lenguaje, como institución social que es, en su papel configurador de tales sociedades. Estos temas, en su mayoría, son analizados desde una perspectiva cuantitativa, aunque algunos de ellos, como las actitudes lingüísticas, por ejemplo, deben complementarse mediante técnicas cualitativas que proporcionen otra perspectiva de análisis (Lorenzo 1994: 101).

La aportación de la informática a la sociolingüística cuantitativa es muy notable. Prácticamente en todo el proceso de investigación nos hemos acostumbrado a recurrir al uso de herramientas informáticas, siendo imprescindibles en la parte analítica de explotación de los datos. En esta fase, el recurso más habitual es la estadística, técnica mediante la cual pasamos de los números aislados a su distribución probabilística en la realidad social que estamos analizando (inferencia). En este sentido, la estadística habría que entenderla como una iniciativa cualitativa. La importancia de la estadística es obvia en sociolingüística cuantitativa ("de la sociedad", en palabras de Fasold):

> En la investigación lingüística que presta atención sólo a la estructura del lenguaje, las pruebas de estadística inferencial generalmente no se consideran necesarias [...]. Sin embargo, cuando el objeto de estudio es el uso de la lengua en un contexto, como ocurre en la sociolingüística (especialmente en la sociolingüística de la sociedad), a menudo necesitamos un método para distinguir lo que es real de lo que es falso. La estadística nos proporciona este método (Fasold 1996: 141).

Con la finalidad de analizar la importancia de la informática en el estudio de la sociolingüística cuantitativa, desde las primeras fases del proyecto hasta su informe final, en este trabajo utilizaremos como referencia varias investigaciones demolingüísticas en las que participamos en los últimos años: por un lado el Mapa Sociolingüístico de Galicia (MSG, Seminario de Sociolingüística 1994, 1995, 1996) y por otro el estudio de las actitudes de los consumidores ante el uso del gallego en la publicidad, trabajo del que extraemos los ejemplos que aparecerán en las tablas y figuras (GALPU, Ramallo y Rei Doval 1995, 1997). En estos trabajos hemos recurrido a varias aplicaciones informáticas: bases de datos, gestores de gráficos y de mapas, procesadores de texto, etc. Pero sin duda el programa de mayor importancia en este tipo de investigaciones es el analizador estadístico, que en nuestro caso fue SPSS (acrónimo de *Statistical Package for the Social Sciences*). Previamente, ofrecemos una breve presentación del variacionismo, que es junto con la demolingüística la perspectiva más destacada dentro de la sociolingüística cuantitativa.

El variacionismo es la corriente de la sociolingüística que estudia la variación intralingüística con el objetivo de encontrar pautas que permitan explicar y predecir el cambio lingüístico. Los estudios pioneros dentro del variacionismo surgen en el seno de la gramática generativa como oposición a la preponderancia que el estudio del sistema había tenido entre las escuelas estructuralistas. Recordemos que para el

estructuralismo saussureano la variación quedaba fuera del alcance de la ciencia lingüística, del estudio de la *langue*. Los estudios variacionistas permitieron superar los conceptos de variación libre y de variación opcional aplicados a casi todo en la lingüística de buena parte del siglo XX. A partir de trabajos empíricos llevados a cabo con muestras reales de habla, el variacionismo permite descubrir el orden que hay en la variación. Así, con esta forma de proceder se pudo detectar que las distintas variantes de un mismo fenómeno están íntimamente relacionadas con el estrato sociocultural de los hablantes, con su edad, sexo, con la formalidad o informalidad de la interacción, etc. además de tener en cuenta características del contexto lingüístico y la función que cumple el elemento que varía dentro de la secuencia. En la actualidad se puede determinar con rigor qué hay detrás de una variación en una comunidad lingüística determinada, bien fenómenos lingüísticos (contextuales y situacionales), bien fenómenos extralingüísticos (sociales y situacionales). Los datos recogidos reciben un tratamiento estadístico, que ha sido posible gracias a la introducción del concepto de regla variable, definido inicialmente por Labov y sistematizado y formalizado después por autores como Cedergren, Sankoff o Rousseau, pertenecientes al *Centro de Investigaciones Matemáticas* de Canadá. Dicho concepto amplía el alcance generativista de la competencia lingüística incluyendo dentro de ésta la posibilidad de la variación, al menos la variación sistemática, que se ha demostrado que no es ni caprichosa ni errática. El análisis de regla variable tiene su fundamento en la elección que hace el hablante entre dos o más formas de decir lo mismo. En palabras de Sankoff:

> Whenever a choice among two (or more) discrete alternatives can be perceived as having been made in the course of linguistic performance, and where this choice may have been influenced by factors such as features in the phonological environment, the syntactic context, discursive function of the utterance, topic, style, interactional situation or personal or sociodemographic characteristics of the speaker or other participants, then it is appropriate to invoke the statistical notions and methods known to students of linguistic variation as *variable rules* (Sankoff 1988: 984).

Para establecer el análisis de regla variable se opera con modelos logísticos que son tratados mediante programas informáticos, destacando de manera especial el denominado de forma genérica VARBRUL, especialmente diseñado por los pioneros del variacionismo y disponible en versiones para IBM-PC y para Macintosh. Este programa recurre a un análisis logístico de regresión mediante el cual podemos determinar la secuencia de variación de una variable determinada, controlando tanto factores sociales como lingüísticos. El programa, que no tiene una distribución comercial, ha sido diseñado específicamente para este tipo de estudios y para su manejo no es necesario conocer los procedimientos matemáticos utilizados. Los resultados obtenidos son fácilmente interpretables por una persona con cierta destreza

en la lectura de tablas y con conocimientos básicos de estadística[2]. Como Sankoff y Labov indican, el análisis de regla variable realizado desde VARBRUL se convirtió en una técnica estándar para analizar datos lingüísticos

> For the present, however, the uniform use of the logistic model [...] seems the best strategy for systematic variation analysis suitable for comparison across different data sets and in different speech communities. The statistical and computational developments to date on variable rule models have brought us steadily closer to standard statistical approaches to the analysis of variation, still perserving the ability to deal with the special characteristics of linguistic data (Sankoff y Labov 1979: 201-202).

Durante los años 70 y 80, el análisis estadístico llegó a ser de rigor en los estudios sociolingüísticos y muchos autores recurrieron al programa VARBRUL para llevarlo a cabo, lo que repercutió en una mejora considerable de las aplicaciones de dicho programa. A pesar de la preponderancia de VARBRUL, algunos variacionistas han hecho uso de otras alternativas estadísticas, como el análisis de componentes principales (Horvath y Sankoff 1987), que es una técnica relativamente simple de reducción de datos que permite visualizar las principales relaciones en un conjunto de datos multivariantes.

El variacionismo fue el marco en el que se desarrolló la mayor parte de la dialectología urbana desde los años 60. Aunque ya en los años 30 se habían aplicado métodos cuantitativos en la dialectología, los estudios de dialectología urbana se inician a raíz de la publicación de los trabajos de W. Labov, en especial a partir de su trabajo en la ciudad de Nueva York, que supuso un modelo de análisis para los dialectólogos urbanos. Este tipo de trabajos se inicia una vez superado el mito del lenguaje puro, invariable y homogéneo que había originado los estudios de dialectología rural. La búsqueda de hablantes 'perfectos' en espacios menos contaminados por la industrialización y la urbanización fue infructuosa porque pronto se tomo conciencia de la importancia de la variación lingüística en todas sus dimensiones. Esto tuvo como resultado que los intereses de la investigación se trasladaron del campo a las ciudades. Las comunidades urbanas se convirtieron en un campo prioritario para la sociolingüística cuantitativa debido a su gran polimorfismo y al reciente interés por las variedades diastráticas y diafásicas en detrimento de las diatópicas, objeto de estudio de la dialectología tradicional, que por otro lado, como hemos sugerido, ha incorporado desde siempre cierta dimensión social. Desde los años 70, con los trabajos de Fasold y Trudgill, el uso de la estadística en la dialectología urbana alcanza una gran difusión.

La demolingüística o demografía de las lenguas estudia la distribución de las variedades lingüísticas entre una población determinada a partir de censos y mapas sociolingüísticos. La principal característica de este tipo de estudios es el hecho de trabajar con grandes cantidades de datos, lo que obliga a recurrir al uso de la

informática casi desde los inicios del diseño de la investigación. Este ha sido el caso del Mapa Sociolingüístico de Galicia (MSG) elaborado por el Seminario de Sociolingüística de la Real Academia Galega entre 1990 y 1997 y en el cual hemos participado desde sus inicios. Se trata de un dilatado estudio de la situación sociolingüística de Galicia en el que se analizaron exhaustivamente la lengua inicial, la competencia lingüística, el uso y las actitudes lingüísticas de la población gallega de 16 y más años. Para ello se seleccionó una muestra de casi 39.000 personas residentes en cada uno de los 313 ayuntamientos en los que Galicia estaba dividida administrativamente a principios de la presente década (actualmente son 315), lo que permitió que obtuviésemos datos significativos para cada uno de los 34 sectores en los que dividimos el territorio gallego. Dada la extraordinaria cantidad de datos manejados fue necesario desde el principio del trabajo recurrir a un tipo específico de software que cubriese nuestras necesidades. Junto con los procesadores de textos y las bases de datos, en esta investigación utilizamos como recurso principal el SPSS que es uno de los más celebrados programas de tratamiento estadístico que se pueden encontrar en el mercado y que, como su nombre indica, está específicamente diseñado para su uso en las ciencias sociales. Aunque en el momento de iniciar esa investigación no disponíamos de una versión que pudiese trabajar en Windows, en la actualidad ya son varias las generaciones que aprovechan este sistema operativo, por lo que los comentarios que aquí efectuaremos serán de la versión 7.5.2 para Windows 95 a falta de conocer con detalle la versión 8.0 ya disponible[3].

Además de unos conocimientos básicos de estadística descriptiva e inferencial, la persona que desee adentrarse en este tipo de investigaciones necesita cierta soltura en materias como las siguientes:

a) los rudimentos de la teoría y de la aplicación muestral, ya que raramente se trabaja con el censo completo de la población, por ser muy costoso y además, en muchos casos, innecesario.
b) la elaboración de cuestionarios, que es la técnica de recogida de información más adecuada a este tipo de estudios. Para elaborar un cuestionario, tan importante es elegir bien la formulación de la pregunta (procurando reducir al máximo los sesgos no intencionados) como diseñar el tipo de respuesta (normalmente se trabaja con preguntas cerradas, esto es, determinamos de antemano la batería de respuestas ofrecida). Para satisfacer esta estrategia debemos conocer las escalas de medición que se van a utilizar en la codificación de las variables. La elección de una escala en detrimento de otra será determinante para la posterior interpretación de los datos.
c) el trabajo de campo
d) la tabulación de variables y valores
e) construcción y lectura de tablas, gráficos, esquemas, mapas, etc.

Una vez realizado el trabajo de campo comienza a ser imprescindible el uso de la estadística, que permitirá explotar a fondo la información obtenida. En este punto, nos gustaría insistir que el propósito final del recurso a la estadística no es generar tests de significación sino más bien estudiar los datos en sí.

La mayoría de los programas estadísticos permiten, además de hacer un análisis muy completo de los datos, presentarlos de múltiples maneras, mediante tablas, esquemas o gráficos de alta calidad y de una manera sencilla. En el caso de SPSS, otros productos de la misma familia nos facilitarán el trabajo en distintas fases de la investigación. A continuación destacamos algunos de ellos:

Nombre del producto	*Función*
SamplePower	Determina el tamaño de la muestra que necesitemos
SPSS Data Entry	Diseño de cuestionarios y recogida de datos
Text Smart	Codificación de preguntas abiertas
SPSS ANSWER TREE	Creación de modelos generados en forma de árbol
MapInfo	Administra gráficamente los datos en forma de mapa

Tabla 1. *Productos informáticos de la familia SPSS*

3. EL FUNCIONAMIENTO DE SPSS

SPSS permite que personas sin mucho conocimiento en estadística puedan trabajar con él. Sin embargo, esto no significa que saber trabajar con el programa sea suficiente para analizar el conjunto de datos de que disponemos. El programa es una herramienta que simplifica y agiliza enormemente el trabajo pero debemos saber interpretar los resultados además de aplicar nuestros conocimientos básicos para determinar qué tipo de test estadístico debemos seleccionar según el tipo de variables con las que trabajemos.

SPSS es un programa modular que permite adaptar el sistema a las necesidades particulares de cada investigación. Cada uno de los módulos se puede adquirir independientemente, aunque el módulo base es esencial para poder utilizar cualquiera de los otros. La siguiente tabla resume la configuración de cada uno de los módulos del programa.

MÓDULO	ESTADÍSTICOS DISPONIBLES/PRESENTACIÓN
Base	Estadística descriptiva (frecuencias, porcentajes, medias, desviación típica, etc.) Tablas de contingencia Pruebas *t* Modelos ANOVA (Análisis de la Varianza): factorial simple Correlación Regresión lineal Pruebas no paramétricas (Chi-cuadrado, Kolmogorov-Smirnov para una muestra, etc.) Análisis factorial Análisis de conglomerados Gráficos
Estadísticas Profesionales	Regresión logística para variables dicotómicas Regresión no lineal Escalado multidimensional Análisis de conglomerados Test de fiabilidad
Estadísticas Avanzadas	Análisis multivariante de la varianza Análisis de supervivencia Técnicas Kaplan-Meier Regresión de Cox Modelos loglineales Modelo lineal general Modelos MANOVA
Tablas	Tablas de hasta tres dimensiones Anidación y concatenación de variables en todas las dimensiones
Tendencias	Modelos ARIMA, SEASON, SPECTRA, AREG
Categorías	Análisis de conjunto Análisis de correspondencias Análisis de homogeneidad Análisis de componentes principales Análisis de correlación canónica

Tabla 2. *Procesos estadísticos disponibles en los módulos de SPSS para Windows 7.5.2*

El programa permite visualizar los resultados por medio de diversas representaciones gráficas, muy mejoradas en las últimas versiones. Además para las personas menos familiarizadas con SPSS o con los procedimientos estadísticos, la versión 7.5 incluye un asesor en esta materia que puede ser muy útil. Mediante una serie de preguntas en lenguaje no técnico y a través de ejemplos visuales, el asesor ayuda a decidir los pasos que debemos seguir para lograr una mejor explotación de los datos en un nivel básico de análisis. Aunque si lo que se quiere es profundizar en los procedimientos estadísticos más complejos, lo más aconsejable es consultar un manual de estadística pensado para las ciencias sociales o bien consultar la ayuda de que dispone el programa que puede ser utilizada como un excelente compendio de la materia. A continuación, presentamos un breve recorrido de los procedimientos básicos para un análisis estadístico a partir de las bases de datos del MSG y de GALPU, relativamente modificados para esta ocasión.

3.1. *El editor de datos*

Las versiones para Windows del programa disponen de un editor de datos que tiene el aspecto de una hoja de cálculo aunque no permite introducir fórmulas. En este editor se pueden grabar directamente las respuestas recogidas mediante el cuestionario u otras técnicas de obtención de datos, utilizando una columna para cada variable (que habrá que definir) y una fila para cada caso, que en nuestra investigación representaba a una persona (figura 1).

Archivo Edición Ver Datos Transformar Estadísticos Gráficos Utilidades. Ventana ?

1:cuest 52

	cuest	res	concello	sexo	clase1	idade2	nac
1	52	3	16	1	3	60	.
2	53	3	16	2	2	33	22
3	54	3	16	1	3	44	13
4	55	3	16	2	2	46	21
5	56	3	16	1	3	30	13
6	57	3	16	2	2	24	13
7	58	3	16	2	2	38	23
8	59	3	16	1	2	55	22

Figura 1. *Editor de datos de SPSS*

El editor permite mostrar las etiquetas que hay detrás de cada código numérico, en el caso de que éstas se hayan introducido, simplemente activando uno de los iconos de la barra de herramientas. Además de incorporar directamente los datos en el editor, con SPSS podemos importar una gran variedad de tipos de ficheros de datos: hojas de cálculo, bases de datos, ASCII, etc. En este último caso, debemos elaborar también un fichero de definición de datos, es decir, debemos proporcionarle al programa un conjunto de instrucciones que indiquen con toda precisión el lugar ocupado por cada variable, las etiquetas de cada una de las categorías y sus características (valores no computados, tipo de variable, etc.).

Para introducir los datos en el editor (bien sea directamente bien sea importándolos) es necesario que hayamos codificado las variables, es decir, que a cada respuesta le asignemos un código distinto y que para respuestas iguales debemos usar siempre el mismo código. Por regla general, cuando se trate de preguntas cerradas servirá la codificación que usamos en la confección del cuestionario. Con todo, el programa permite recodificar las variables según los objetivos de la investigación. En el caso de variables con preguntas abiertas debemos recurrir a programas de postcodificación, de manera especial si trabajamos con cientos o incluso miles de respuestas diferentes. En este caso podemos aprovechar el programa TextSmart, citado en la tabla anterior[4]. Este programa tiene un conjunto de herramientas que filtran las respuestas y crean una lista de palabras clave que obtiene del análisis de las contestaciones a las preguntas abiertas de un estudio concreto. Mediante procesos estadísticos, TextSmart analiza automáticamente esa lista de palabras clave y agrupa cada respuesta en categorías de similar significado en muy poco tiempo. Es decir, el programa no se limita a dar un índice de concordancias del corpus de preguntas abiertas sino que mediante la combinación de algoritmos y tecnología lingüística categoriza automáticamente las respuestas agrupando las palabras en familias léxicas, combinando sinónimos, excluyendo del análisis unidades con significado gramatical (si se desea), etc. Una ventaja de este programa es que una vez que tenemos las respuestas codificadas pueden ser tratadas con todos los recursos disponibles en SPSS. Hacer una postcodificación de preguntas abiertas manualmente para miles de respuestas puede llevar mucho tiempo, incrementando los costes de la investigación. Por lo tanto, si no se dispone de un programa de este tipo lo mejor es dejar las preguntas abiertas para las fases preparatorias del estudio, en las que se suele aplicar una versión previa del cuestionario a un número reducido de personas para ver sus posibilidades y cerrar en el cuestionario definitivo todas o la mayoría de las preguntas.

Para codificar una variable debemos tener claro el nivel de medición con el que queremos trabajar ya que la elección de las distintas escalas condicionará el tipo de test estadístico que podremos usar. De los cuatro tipos de escalas, nominal, ordinal, de intervalo y de razón, los dos primeros son los más usados en la investigación demolingüística:

a) nominal: se trata de meras clasificaciones sin implicación de orden de ningún tipo. La codificación asignada a cada etiqueta sólo sirve para identificar respuestas distintas

Ejemplo: *¿En que lengua aprendió a hablar?*

Gallego	1
Español	2
En ambas	3
En otra/s	4

b) ordinal: clasifica y ordena otorgando un significado a la secuencia ordenada de categorías

Ejemplo: ¿Qué le parece el uso del gallego en la publicidad?

Muy bien	1
Bien	2
Indiferente	3
Mal	4
Muy mal	5

Aunque para el tratamiento estadístico este tipo de escalas presenta muchas restricciones, en la mayoría de los trabajos sociolingüísticos (en general, en las ciencias sociales) es habitual tratar las escalas ordinales como escalas de intervalos, lo que amplia considerablemente los modelos de análisis (ANOVA, regresión, análisis de correspondencias, etc.). Es decir, pese a que teóricamente la medición ordinal no permite tomar en consideración las distancias entre los elementos ordenados, es muy frecuente en ciencias sociales dar un 'salto de nivel' y tomar la medición ordinal como si fuese interval, en la que sí ya podemos estar seguros de que las diferencias entre dos valores consecutivos en alguna sección de la escala es la misma que la que encontramos entre otros dos elementos correlativos en una parte diferente de la escala. Sirva como ejemplo de escala de intervalos la establecida en el cuadro que indica la velocidad de un automóvil. En ella, la diferencia que hay entre 20 y 30 kilómetros/hora es la misma que la que existe entre 110 y 120 kilómetros/hora. En definitiva, en una escala de intervalo diferencias iguales en los números representan diferencias iguales en los objetos medidos. Aunque este cambio de nivel no está exento de críticas por parte de muchos autores no deja de ser una práctica muy habitual, tal como acabamos de señalar.

En cualquier momento podemos crear variables nuevas a partir de transformaciones de otras variables así como recodificar los valores de una variable porque se ha comprobado, por ejemplo, que el plan de codificación inicial era poco operativo. Como ejemplo del primer caso, en el cuestionario del MSG se incluyeron

preguntas relacionadas con la lengua hablada por el entrevistado con sus hermanos, con sus hermanas, con sus hijos, con sus hijas, diferenciando en los cuatro casos el factor edad, es decir hermanos mayores y hermanos pequeños, hermanas mayores y hermanas pequeñas, etc. En total ocho preguntas, que, una vez analizadas las correlaciones y detectada la correspondiente ausencia de diferencias entre ellas, quedaron agrupadas en dos: lengua con los hermanos y lengua con los hijos. Los ejemplos de recodificación son muy habituales, en especial cuando se trata de variables sociales. En el MSG recodificamos las respuestas de adscripción a una clase social, ya que había muy pocos casos de clase alta a pesar del N tan alto de la muestra, lo que obligó a reducir la escala de cinco categorías a cuatro. La recodificación se puede hacer en la misma variable o en nuevas variables. Si optamos por esta última posibilidad dispondremos siempre de la variable original.

El editor de datos admite variables de distinto tipo. Aunque el programa establece por defecto todas las variables como numéricas, es muy sencillo modificar esta asignación en las variables que deban ser tratadas como alfanuméricas, fechas, notación científica, etc.

3.2. *Estadística descriptiva*

Una vez que tenemos el fichero de datos, podemos iniciar la explotación de los mismos. Debemos comenzar con la descripción univariable recurriendo a recuentos y porcentajes en el caso de variables nominales y a medidas de tendencia central (media, moda, mediana, etc.) y de dispersión (desviación típica, varianza, rango, etc.) en el caso de las variables codificadas ordinalmente.

La tabla 3 muestra un ejemplo de frecuencias de la variable nominal 'lengua inicial'. El programa hace una distinción entre valores válidos y valores perdidos o no computados. Se observará que en la tabla aparecen 7 valores perdidos, que en este caso se corresponden con la categoría 'otras'. En SPSS nos encontramos con dos tipos de valores perdidos. Por un lado, están los valores perdidos definidos por el propio sistema, que serían todas las casillas en blanco dentro de la matriz de datos del editor en el caso de las variables numéricas. Son casos de los que no tenemos información. Por otro lado, están los valores no computados asignados por el usuario a cualquier tipo de variable. Esto es lo que ocurre con la variable que aparece reflejada en la tabla anterior. En este caso, dado el reducido número de personas que tuvieron como lengua materna una distinta del gallego y del español, decidimos no tenerlos en cuenta en la investigación, por lo que fueron tratados como valores perdidos. En la columna que remite al porcentaje acumulado ya sólo aparecen los datos de las tres categorías válidas. SPSS permite asignar hasta un máximo de tres valores perdidos para cada variable. Si en algún momento decidimos que debemos trabajar también con los casos no computados asignados por el usuario, basta con modificar su estatus de perdidos a válidos.

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válidos	**Gallego**	724	59,8	60,2	60,2
	Español	315	26,1	26,2	86,4
	Ambas	163	13,5	13,6	100,0
	Total	1202	99,4	100,0	
Perdidos	**Otras**	7	,6		
	Total	7	,6		
Total		1209	100,0		

Tabla 3. *Lengua inicial (GALPU)*

La misma información que contiene la tabla anterior podemos presentarla de forma gráfica, aprovechando las posibilidades que en este sentido ofrece SPSS. Los gráficos permiten ver de manera rápida y precisa la concentración o dispersión de los datos, lo que evita la fatiga de enfrentarse a tablas de lectura oscura.

La tabla anterior es la más simple que ofrece el programa. La flexibilidad de este componente permite realizar tablas mucho más ricas en datos y de mayor complejidad de lectura. Puede ser necesario mostrar el cruce de una o más variables independientes (sociales en nuestro caso) con distintas variables dependientes, que podremos presentar anidadas. Esto es lo que se ofrece en el siguiente ejemplo:

					lengua habitual				
					Sólo español	Más español que gallego	Más gallego que español	Sólo gallego	Total de tabla
EDAD	18-25	sexo	Hombre	Recuento	7	28	48	32	116
				% col.	6,0%	24,4%	41,9%	27,8%	100,0%
			Mujer	Recuento	17	29	30	33	110
				% col.	15,7%	26,4%	27,6%	30,2%	100,0%
	26-35	sexo	Hombre	Recuento	11	29	41	59	140
				% col.	7,8%	21,0%	29,3%	41,9%	100,0%
			Mujer	Recuento	22	33	30	50	135
				% col.	16,1%	24,9%	22,1%	36,9%	100,0%
	36-45	sexo	Hombre	Recuento	11	15	38	50	115
				% col.	9,9%	13,4%	32,9%	43,8%	100,0%
			Mujer	Recuento	14	22	35	60	132
				% col.	10,6%	16,8%	26,8%	45,7%	100,0%
	46-55	sexo	Hombre	Recuento	9	16	30	68	123
				% col.	7,1%	13,2%	24,2%	55,5%	100,0%
			Mujer	Recuento	11	19	29	65	123
				% col.	8,8%	15,4%	23,2%	52,6%	100,0%
	56-65	sexo	Hombre	Recuento	2	13	29	58	101
				% col.	1,6%	12,4%	28,7%	57,2%	100,0%
			Mujer	Recuento	9	19	18	70	115
				% col.	8,1%	16,2%	15,2%	60,6%	100,0%
Total de tabla		Recuento			113	224	327	545	1209
		% col.			9,3%	18,5%	27,1%	45,1%	100,0%

Figura 2. *Lengua habitual según la edad y el sexo (GALPU)*

Si trabajamos con variables ordinales, podremos ofrecer una tabla descriptiva más completa, añadiendo a lo ya conocido otros cálculos estadísticos básicos como la media, moda, mediana, rango, desviación típica, varianza, etc. La media es el estadístico más utilizado con variables ordinales siempre y cuando la variable que se esté analizando no incluya valores muy dispares que afecten (sesguen) a esta medida. En estos casos, sería conveniente recurrir a la mediana, que es la puntuación que ocupa la posición central, o a la moda, que consiste en la puntuación más repetida de una serie. La media se calcula sumando todos los casos y dividiendo el resultado por el número total de casos. No siempre es fácil decantarse por una medida de tendencia central ya que cualquiera de ellas tiene ventajas e inconvenientes. Siempre que tengamos una distribución sin valores muy extremos que afecten de manera considerable a la media, esta debería ser la medida utilizada, ya que, como acabamos de decir, tiene en cuenta todos los datos. En caso contrario, la mediana será la mejor elección. Junto a las medidas de tendencia central siempre debemos ofrecer algún estadístico que dé cuenta de la mayor o menor variabilidad de los datos. En este caso, lo más habitual es recurrir a la desviación típica o estándar, que resulta de hallar el promedio de la desviación de los casos con respecto a la media. Datos muy homogéneos tendrán una desviación típica pequeña y datos muy heterogéneos lo contrario. De este modo, facilitaremos la interpretación de los datos de nuestra investigación.

Una vez realizada la descripción exploratoria de la muestra seleccionada, debemos pasar a los análisis que relacionan dos (análisis bivariantes) o más de dos variables (análisis multivariante). Cuando las variables dependientes son nominales lo correcto es elaborar tablas de contingencia del estilo de la ofrecida en la figura anterior[5]. Una tabla de contingencia no es más que una matriz de datos en la que las filas son los valores de una variable y las columnas los valores de otra. Cada casilla de la tabla es el cruce de un valor de fila con un valor de columna. Lo más aconsejable es indicar en la tabla tanto las frecuencias absolutas como las relativas (porcentajes), para facilitar la interpretación. Estas últimas, por lo general, deben estimarse con relación a la variable independiente, aunque los paquetes estadísticos, y en concreto SPSS, permiten dar los marginales tanto para la variable considerada independiente como para la dependiente.

A partir de las tablas de contingencia podemos pasar de la estadística descriptiva a la estadística inferencial, mediante la cual deducimos parámetros poblacionales a partir de los estadígrafos de la muestra. Esta es la finalidad de toda investigación que trabaje con muestras, que en sí mismas tienen poco interés si no podemos utilizarlas como representación de un conjunto más amplio. Pero no todo lo que aparentemente es significativo en una muestra hay que interpretarlo como igualmente relevante en la población de la que fue extraída. La solución a esta cuestión nos la proporcionan los tests estadísticos, que nos permiten pasar de los estimadores muestrales a los parámetros poblacionales, con un nivel de confianza predeterminado por el analista y que en la

demolingüística se suele situar en el 95%, lo que equivale a decir que de cada 20 veces que hagamos la inferencia 19 serán acertadas. Este nivel de confianza, aún siendo muy elevado, nunca nos garantiza que nuestras conclusiones sean las correctas, pero este es el arte de la estadística. Además, tenemos que contar siempre con el error muestral, que vendrá determinado por el tamaño de la muestra, el nivel de confianza adoptado, la varianza poblacional y el tipo de muestreo utilizado. En nuestro caso, el error máximo asumido para el conjunto de la muestra (N= 1209) fue del ± 2,9%, para un nivel de confianza del 95,5% y partiendo del supuesto de varianza poblacional más desfavorable (p=q=50%). Esto significa que el paso de la muestra a la población implica tener presente que el parámetro poblacional estará situado en un margen que va del –2,9 al +2,9 con relación al estadígrafo muestral en 95 de cada 100 veces.

3.3. *La reducción de los datos. El análisis factorial*

El cuestionario que utilizamos para analizar las actitudes de los consumidores residentes en Galicia ante el uso de la lengua gallega en la publicidad y en las relaciones comerciales (GALPU) constaba de 85 ítems organizados en 46 preguntas. Entre aquéllos había 18 que incidían sobre distintos objetos actitudinales, que sospechamos que podrían estar muy relacionados entre sí. Esta sospecha se justificaba porque en la mayoría de los trabajos de actitudes lingüísticas realizados en Galicia durante los últimos 15 años se había demostrado que las actitudes son bastante homogéneas, independientemente del objeto concreto a medir (Seminario de Sociolingüística 1996). Por tanto, para evitar el tratamiento de cada variable de forma individual, lo que haría el análisis muy tedioso y redundante, decidimos someter estas 18 variables a un análisis factorial. Denominamos análisis factorial a la técnica que se aplica para reducir un determinado número de ítems individuales en un número inferior de dimensiones latentes (factores). En esta prueba, los factores representan las variables originales con una pérdida mínima de información. Como condición previa para realizar un análisis factorial, las variables deben presentar una determinada correlación, que permita que sean consideradas como subconjuntos. Si esto no fuese así carecería de sentido realizar tal prueba. Una vez que hemos indicado qué variables serán sometidas al análisis, el programa calcula por defecto tantos factores como autovalores iniciales iguales o superiores a 1, aunque esta opción se puede modificar, tal como hemos hecho en el ejemplo que exponemos, si se observa que es necesario para poder interpretar alguna variable (comunalidad baja una vez realizada la extracción factorial). Lo normal es trabajar con autovalores superiores a 1, ya que cuanto más próximo está un autovalor a 1 más seguros estamos de que cada factor está asociado a una única variable (la suma de autovalores siempre será igual al número de variables, 18 en nuestro caso), por lo que no tendría sentido hacer un análisis factorial.

ANÁLISIS FACTORIAL

Varianza total explicada

Componente	Autovalores iniciales			Sumas de las saturaciones al cuadrado de la extracción			Suma de las saturaciones al cuadrado de la rotación		
	Total	% de la varianza	% acumulado	Total	% de la varianza	% acumulado	Total	% de la varianza	% acumulado
1	5,920	32,886	32,886	5,920	32,886	32,886	3,299	18,329	18,329
2	1,853	10,293	43,180	1,853	10,293	43,180	2,689	14,938	33,268
3	1,494	8,299	51,478	1,494	8,299	51,478	1,921	10,673	43,940
4	1,154	6,411	57,889	1,154	6,411	57,889	1,673	9,292	53,233
5	1,050	5,834	63,723	1,050	5,834	63,723	1,671	9,284	62,516
6	,908	5,044	68,767	,908	5,044	68,767	1,125	6,250	68,767
7	,797	4,427	73,194						
8	,701	3,893	77,087						
9	,628	3,489	80,576						
10	,603	3,349	83,925						
11	,497	2,760	86,685						
12	,461	2,559	89,244						
13	,440	2,443	91,687						
14	,396	2,202	93,890						
15	,327	1,818	95,707						
16	,295	1,637	97,345						
17	,262	1,453	98,798						
18	,216	1,202	100,000						

Método de extracción: Análisis de Componentes principales.

Figura 3. *Análisis factorial: varianza total explicada (GALPU)*

La figura 3 muestra los autovalores de cada uno de los factores y el porcentaje de varianza explicada por cada una de ellos, además de la varianza acumulada. Como se ve, el programa facilita los resultados de 18 factores, tantos como variables, por lo que la varianza acumulada total no puede ser otra que el 100%. Sin embargo, el interés de la prueba, como ya indicamos, es conseguir una reducción satisfactoria de las variables sin perder mucha información. En nuestro caso, con 6 factores tenemos un 68,7% de la varianza explicada, lo que supone una reducción importante. El porcentaje de varianza explicado por cada factor es el resultado de dividir su autovalor entre el número de variables y multiplicarlo por 100, que es la varianza total. En nuestro caso, el primer factor explica el 32,8%, mientras que el segundo ya sólo explica el 10,29% y así sucesivamente hasta llegar al sexto factor, que es el último con el que trabajamos.

Las puntuaciones factoriales de nuestro ejemplo presentan saturaciones altas en uno de los factores y bajas en el resto, tal como refleja la figura 4. El análisis detallado de las variables que fueron analizadas permite comprender las dimensiones subyacentes a cada uno de estos factores, que habrá que etiquetar de la mejor manera posible:

Factor 1: Adecuación del gallego para la publicidad en diferentes medios
Factor 2: Actitud ante usos concretos del gallego en la comunicación emitida en Galicia
Factor 3: Actitud ante el uso del español en la publicidad emitida en Galicia
Factor 4: Prestigio y confianza vehiculados por el gallego
Factor 5: La emisión de publicidad en gallego por los medios de comunicación masivos
Factor 6: La divergencia lingüística en las relaciones comerciales

Matriz de componentes rotados[a]

	Componente					
	1	2	3	4	5	6
El gallego es apropiado para la publicidad en TV/radio	,857					
El gallego es apropiado para el etiquetado de los productos	,845					
El gallego es apropiado para la publicidad en periódicos/revistas	,838					
El gallego es apropiado para el buzoneo	,820					
Aceptación de los empleados de comercio que hablan gallego		,720				
La rotulación de las tiendas debería estar en gallego		,681				
La publicidad de los establecimientos comerciales debe estar en gallego		,679				
El cien que se emite en las salas gallegas debe ser en gallego		,674				
Actitud ante el hecho de encontrar publicidad en gallego en el buzón		,506				
Los productos anunciados en gallego no merecen confianza			,883			
El gallego no es apropiado para anunciar productos de prestigio			,876			
Actitud ante la emisión de anuncios en español en la TVG				,811		
Los productos gallegos deberían anunciarse en español				,630		
La publicidad insititucional emitida en Galicia debe estar en español				,575		
Actitud ante la emisión de anuncios en gallego en las televisiones estatales					,804	
Actitud ante la publicidad en gallego en los periódicos hechos en Galicia					,700	
Actitud ante el hecho de hablar gallego en un comercio y ser constestado en español						,783
Actitud ante el hecho de hablar en español en un comercio y ser contestado en gallego						,679

Método de extracción: Análisis de componentes principales.
Método de rotación: Normalización Varinas con Kaiser.
a. La rotación ha convergido en 6 iteraciones

Figura 4. *Resultado factorial de la matriz de componentes rotados*

Una vez que contamos con los factores, podremos ya guardarlos como variables independientes y ubicarlos en el editor de datos.

3.4. *Pruebas paramétricas y no paramétricas*

Las pruebas estadísticas se dividen en *paramétricas* o dependientes de la distribución y en *no paramétricas* o de distribución libre. El uso de unas y otras no es arbitrario sino que viene determinado por una serie de supuestos. Aunque a veces estos términos son utilizados en diferentes sentidos, en general utilizar una prueba paramétrica implica asumir que los errores aleatorios que afectan a los resultados tienen una distribución normal, es decir, los contrastes paramétricos tienen que

suponer una normalidad aproximada de la distribución poblacional. Por su parte, las pruebas no paramétricas no requieren este supuesto (MacRae 1995: 154). Las pruebas paramétricas son más adecuadas cuando la medida se ha hecho con una escala de intervalo o de razón, en las que las diferencias entre los valores representan diferencias iguales en los objetos medidos. Las pruebas no paramétricas tienen normalmente menos potencia para rechazar la Hipótesis Nula cuando deberíamos hacerlo (error de tipo II, en términos estadísticos)[6].

A continuación presentaremos ejemplos de una prueba de cada tipo: el Chi-cuadrado (X^2), entre las no paramétricas y el Análisis de la Varianza (ANOVA) entre las paramétricas, por ser las más habituales en este tipo de trabajos y porque son las que hemos utilizado en nuestras investigaciones. Aunque SPSS ofrece muchas más posibilidades, en la práctica sociolingüística estas son quizás las más utilizadas.[7]

3.4.1. *La prueba Chi-cuadrado*

La prueba del Chi-cuadrado (X^2) se aplica para comprobar si la asociación de dos variables nominales es significativa o no. A partir de una tabla de contingencia, el estadístico compara las frecuencias observadas en la realidad analiza, también llamadas recuentos, con las 'esperadas' si no hubiese diferencias significativas en la correlación de dos variables (que sería la llamada Hipótesis Nula). Cuanto mayor sea la diferencia entre los recuentos y las frecuencias esperadas mayor será la probabilidad de que la muestra provenga de una población en la que las variables estén relacionadas (Sánchez Carrión 1995: 261). Podemos ver un ejemplo en la figura 5, que recoge el resultado de cruzar la lengua inicial con la edad en GALPU. Si observamos la columna correspondiente al gallego, apreciaremos que entre los entrevistados menores de 46 años las frecuencias observadas son inferiores a las esperadas y al contrario entre los de 46 o más años. El valor del X^2, 51.538, y la significación asociada, p=.000, nos muestra que estas diferencias no son debidas al azar (o no sólo) por lo que podremos concluir que hay que rechazar la hipótesis que establece que ambas variables son independientes. Si consultásemos una tabla de puntos críticos de X^2, para una nivel de significación de 0.05 y 8 grados de libertad (los de nuestra tabla) se vería que X^2 tiene que ser mayor o igual a 15.5 para que el resultado sea significativo.[8]

Como en este caso la variable independiente es la edad, los porcentajes que se ofrecen hacen 100 sus marginales, lo que permite comparar el porcentaje de sus categorías para cada una de las categorías de la variable dependiente (lengua inicial). El análisis de los datos nos permite confirmar que el gallego ha sido la lengua inicial de la población gallega de entre 18 y 65 años, independientemente del tramo concreto de edad. Así mismo, se detecta un aumento del bilingüismo inicial que en los últimos 40 años ha doblado su presencia, aunque su incidencia en el conjunto sigue siendo minoritaria.

Tabla de contingencia EDAD * lengua inicial

			lengua inicial			
			Gallego	Español	Ambas	Total
EDAD	18-25	Recuento	119	67	39	225
		Frecuencia esperada	135,4	58,9	30,7	225,0
		% de EDAD	52,9%	29,8%	17,3%	100,0%
	26-35	Recuento	134	102	36	272
		Frecuencia esperada	163,7	71,2	37,1	272,0
		% de EDAD	49,3%	37,5%	13,2%	100,0%
	36-45	Recuento	145	61	39	245
		Frecuencia esperada	147,4	64,2	33,4	245,0
		% de EDAD	59,2%	24,9%	15,9%	100,0%
	46-55	Recuento	162	52	31	245
		Frecuencia esperada	147,4	64,2	33,4	245,0
		% de EDAD	66,1%	21,2%	12,7%	100,0%
	56-65	Recuento	164	33	19	216
		Frecuencia esperada	130,0	56,6	29,4	216,0
		% de EDAD	75,9%	15,3%	8,8%	100,0%
Total		Recuento	724	315	164	1203
		Frecuencia esperada	724,0	315,0	164,0	1203,0
		% de EDAD	60,2%	26,2%	13,6%	100,0%

Pruebas de chi-cuadrado

	Valor	gl	Sig. asint. (bilateral)
Chi-cuadrado de Pearson	51,538[a]	8	,000
Razón de verosimilitud	51,839	8	,000
Asociación lineal por lineal	28,445	1	,000
N de casos válidos	1203		

a. 0 casillas (,0%) tienen una frecuencia esperada inferior a 5. La frecuencia mínima esperada es 29,45.

Figura 5. *Prueba del Chi-cuadrado para el cruce de lengua inicial y edad (GALPU)*

Con todo, que las diferencias entre dos variables sean significativas no quiere decir que lo sean todos y cada uno de los cruces entre ambas. A través del análisis de los residuos podremos precisar más la interpretación de los resultados. El análisis de residuos es una prueba que se detiene en cada pareja de categorías y ofrece datos sobre su significación. SPSS dispone de varios análisis de residuos siendo el más recomendable el que opera con los residuos tipificados corregidos (ajustados), esto es, con la diferencia entre el valor observado y el valor esperado dividido por una estimación de su error típico. De esta manera, el residuo tipificado viene expresado en unidades de desviación típica, por encima o por debajo de la media. Así, para un nivel

de confianza del 95,5%, todos los residuos ajustados que se encuentren entre la media y ±1,96 unidades de desviación típica, no serán estadísticamente significativos. Por el contrario, todos los residuos ajustados que queden por encima de esta distancia habrá que interpretarlos como un ejemplo de dependencia entre los valores del cruce en cuestión. En el caso que acabamos de mostrar, los residuos ajustados son los siguientes:

	Gallego	Español	Ambas
18-25 años -	2,5	1.4	1,8
26-35 años	-4,2	4.8	-0,2
36-45 años -	0,4	-0,5	1,2
46-55 años	2,1	-2,0	-0,5
56-65 años	5,2	-4,0	-2,3

Tabla 4. *Residuos ajustados del cruce lengua inicial y edad (GALPU)*

Cuanto mayor sea el valor del residuo mayor será la relación que recoge el cruce de las dos categorías. Los residuos de signo negativo indican una relación contraria. La tabla permite una lectura más precisa del cruce de la figura 5. A modo de ejemplo, destacaremos dos aspectos: por ejemplo, el grupo de edad que mantiene una relación más ajustada a lo esperado de no haber dependencia entre ambas variables es el de 36 a 45 años: en ningún caso los valores superan el valor de referencia: ±1,96. Por otro lado, en cuanto al bilingüismo inicial, debemos descartar la hipótesis recogida más arriba de que hay diferencias significativas con relación a la edad. Sólo entre los mayores de 55 años hay una relación estadísticamente significativa al 95,5%. Es decir, la edad es una variable que permite explicar el descenso del bilingüismo inicial sólo en el grupo de 56 a 65 años. El resto de las diferencias no se deben al factor edad.

3.4.2. *El análisis de la varianza*

El análisis de la varianza es una de las pruebas más relevantes de la estadística multivariante. Hay varias razones para preferir un análisis multivariante a un análisis univariante. Quizá la más importante sea que los tests que tienen en cuenta sólo una variable ignoran información importante como por ejemplo la correlación entre variables. Además, sabemos que aunque las variables puedan no ser significativas tomadas individualmente sí pueden serlo analizadas de manera conjunta. Con todo, cuando sólo es necesario estudiar la influencia de una única variable independiente sobre una variable dependiente cuantitativa podemos recurrir al análisis de la varianza

de un solo factor; en este caso, asumiremos que el resto de las 'causas' de variación debemos vincularlas al componente aleatorio. Así pues, podemos recurrir a alguna de las siguientes formulaciones:

ANOVA de un factor	1 variable dependiente/1 variable independiente
Análisis factorial de la varianza	1 variable dependiente/2 o más variables independientes
Análisis multivariante de la varianza (MANOVA)	2 o más variables dependientes/2 o más variables independientes

Tabla 5. *Análisis de la varianza en SPSS*

Aunque el verdadero análisis multivariante es aquél que analiza el comportamiento de varias variables dependientes a partir de dos o más variables independientes, en la práctica es usado casi exclusivamente en la investigación experimental. Lo más frecuente en sociolingüística es explorar el análisis factorial de la varianza, calculando las medias de una variable dependiente para cada uno de los grupos de las variables independientes y analizando las diferencias observadas entre ellas.

Como ya hemos dicho, una de las condiciones para poder realizar un análisis de la varianza es que la variable dependiente sea cuantitativa. En demolingüística, y en general en la investigación no experimental, es muy infrecuente utilizar este tipo de variables como dependientes por lo que, en rigor, el análisis de la varianza tiene un interés menor. Las variables utilizadas son ordinales en su mayoría y no cumplen con los requisitos que exige la utilización del análisis de la varianza, que son los siguientes:

- supuesto de normalidad: la distribución de la población de la que sacamos la muestra es normal.
- supuesto de homoscedasticidad: las varianzas de los grupos con los que se trabaja son iguales.

Sin embargo, cuando las muestras son grandes (N>30), este tipo de pruebas no se ven muy afectadas a pesar de que no se cumplan los supuestos necesarios (Sánchez Carrión 1995: 334 y siguientes). Además, como ya se ha dicho, la solución más habitual pasa por dar un salto de nivel y considerar las escalas ordinales como si fuesen de intervalo, asumiendo que las respuestas dadas por los individuos pueden ser interpretadas como puntuaciones que permitan calcular medias. Esto no quiere decir que un individuo que diga que habla sólo gallego (puntuación 4 en nuestro cuestionario) hable el doble de gallego que alguien cuya conducta lingüística es 'más español que gallego' (puntuación 2 en nuestro cuestionario).

Una vez que tenemos las medias de la variable dependiente para cada uno de los valores de las variables independientes, SPSS dispone de varias pruebas que deciden si las diferencias son significativas y por lo tanto podemos extrapolarlas de la muestra

con la que trabajamos al conjunto de nuestra población. Veamos un ejemplo tomado de nuestro trabajo:

Variable dependiente: lengua habitual
Variables independientes: edad, estudios, hábitat de residencia

En primer lugar, el programa presenta un análisis estadístico básico, donde podemos ver la distribución de las frecuencias y las medias de cada categoría y de las interacciones entre las variables independientes con relación a la variable dependiente. La parte central del análisis es la tabla de la varianza que podemos ver a continuación. Si observamos el nivel de significación de *F*, que es el contraste utilizado por ANOVA podemos concluir que los tres efectos principales que estamos analizando son estadísticamente significativos por ser menores que 0,05. Además, el modelo permite concluir que parte de la varianza viene explicada por la interacción de la edad con estudios (Sig=0,026), siendo irrelevantes los otros dos cruces de variables. Si revisamos la columna de *F* podemos ver que la variable más explicativa del modelo es el hábitat de residencia seguida de los estudios. En el cuadro de la bondad de ajuste del modelo disponemos de un coeficiente de correlación múltiple, *R*, y de su cuadrado, *R* cuadrado. Este último hay que interpretarlo como la proporción de varianza explicada por todos los efectos, que en nuestro caso es del 42%. El resto de la varianza no estaría controlada con estas variables. Aunque podríamos incluir más variables en el análisis para ver si de esta manera aumenta la proporción de varianza, en el caso que exponemos este aumento es poco relevante y no compensa por el alto grado de complicación que supone la inclusión de más efectos.

				Método jerárquico			
			Suma de cuadrados	gl	Media cuadrática	F	Sig
LENGUA HABITUAL	Efectos principales	(Combinadas)	500,643	9	55,627	96,650	,000
		EDAD	39,508	4	9,877	17,161	,000
		ESTUDIOS	195,846	3	65,282	113,426	,000
		HÁBITAT	265,290	2	132,645	230,467	,000
	Interacciones de orden 2	(Combinadas)	23,335	26	,898	1,559	,037
		EDAD * ESTUDIOS	13,448	12	1,121	1,947	,026
		EDAD * HÁBITAT	2,397	8	,300	,521	,842
		ESTUDIOS * HABITAT	6,258	6	1,043	1,812	,093
	Modelo		523,978	35	14,971	26,011	,000
	Residual		666,665	1158	,576		
	Total		1190,644	1193	,998		

	R	R cuadrado
LENGUA HABITUAL POR EDAD, ESTUDIOS Y HABITAT DE RESIDENCIA	,648	,420

Figura 6. *Resultado del análisis de la varianza de la lengua habitual (GALPU)*

Estos ejemplos que hemos dado no agotan las posibilidades de análisis del programa. Cualquier programa estadístico de los citados en este trabajo permite efectuar análisis similares y otros muchos de distinta complejidad. Para ampliar los conocimientos sobre estos programas, el lector puede consultar las referencias bibliográficas relacionadas con los mismos y que se incluyen a continuación.

4. CONCLUSIONES

En estas páginas hemos presentado algunas de los usos más frecuentes de la aplicación de la estadística en el campo de la sociolingüística cuantitativa. Como se ha visto a lo largo de estas páginas, esta disciplina no se puede entender sin el recurso a los modernos instrumentos aportados por la informática tanto por las mejoras técnicas de las máquinas como por el desarrollo y el perfeccionamiento de aplicaciones de enorme utilidad en las tareas de cuantificación y análisis estadístico. En concreto, el programa SPSS se ha convertido en un estándar de trabajo por su versatilidad y adaptación a los intereses particulares de cada usuario. Con todo, debemos seguir insistiendo en que estos programas, aunque facilitan nuestras tareas permitiendo un ahorro extraordinario de tiempo entre otras cosas, deben complementarse con unos conocimientos básicos de los rudimentos de la estadística.

NOTAS

1. Otras oposiciones que se han usado son macrosociolingüística vs. microsociolingüística o sociología del lenguaje vs. sociolingüística.
2. El lector podrá encontrar una clara exposición del funcionamiento de este conjunto de aplicaciones en Moreno Fernández (1994).
3. Además del sistema SPSS hay otros programas estadísticos que pueden ser utilizados en este tipo de investigaciones. Si lo que se necesita es un análisis estadístico elemental algunas de las hojas de cálculo más conocidas incluyen herramientas que satisfacen esta demanda. Otros programas específicos son MINITAB, SAS, BMDP (adquirido por SPSS Inc. en 1995), ESP, OSIRIS, etc. Aunque para su uso se requiere un conocimiento de fundamentos estadísticos no es necesario que dichos conocimientos sean exhaustivos. La mayoría de estos programas es modular por lo que el investigador podrá adquirir los componentes que realmente serán imprescindibles para su trabajo. Todos ellos tienen gran versatilidad y su manejo carece de complicaciones, entre otras cosas debido a que disponen de una interfaz muy intuitiva. Una breve comparación de MINITAB y SPSS, basada en las versiones de hace algunos años, puede encontrarse en Butler (1985: 154-167). Por su parte, Stevens (1996: 23-34) ofrece un resumen de las potencialidades de SAS y de SPSS. Una comparativa de estos dos últimos programas y su aplicación al estudio de datos similares a los del MSG puede encontrarse en Kretzschmar, Jr. W. A. y E. W. Schneider (1996: cap. 4). Por lo que respecta a los fundamentos de BMDP, el lector puede consultar Álvarez González (1996).
   Sobre SPSS existen diversos libros que presentan, por lo general de una forma gráfica y clara, los procedimientos de este programa en alguna de sus últimas versiones. Entre ellos, podemos destacar dentro del entorno hispánico, Ferrán Aranaz (1996) y Visauta Vinacua (1997), además del propio

manual del programa. A través de Internet se puede consultar la dirección de SPSS en el sitio Web *http://www.spss.com/SPAIN* que ofrece información sobre 30 productos de la familia SPSS, además de bibliografía, cursos on-line y la posibilidad de incorporar diversas demos de los distintos módulos.

4. Otros programas que incluyen la posibilidad de codificar preguntas abiertas procedentes de cuestionarios son ANACONDA (Mochmann, 1985) e INTEXT (Klein, 1991).
5. En la investigación social se habla de variables dependientes en relación a aquéllas que hay que explicar y que conforman el objeto de estudio. Las variables independientes (predictoras) son las que se utilizan para analizar su efecto sobre las dependientes. Por regla general, el carácter de independiente o dependiente no viene dado por la propia variable sino por los intereses del investigador.
6. La Hipótesis Nula es una manera precisa de afirmar que las diferencias que encontramos entre dos o más variables son debidas al azar. Lo que se busca en una investigación es comprobar si podemos determinar una Hipótesis Alternativa que corrobore que "algo" distinto del azar es imprescindible para explicar los resultados obtenidos. El error de tipo I se comete al descartar la Hipótesis Nula cuando no debería hacerse, mientras que el error de tipo II se produce cuando concluimos que el resultado no es significativo y en realidad lo es. Es imposible eliminar los dos tipos de errores. Si pretendemos reducir el error de tipo I aumentamos las posibilidades de cometer error de tipo II y viceversa.
7. Junto al chi-cuadrado y al análisis de la varianza, Fasold (1984: cap. 4) destaca la distribución *t* de Student y la correlación como las otras dos pruebas más utilizadas en los trabajos sociolingüísticos.
8. Los grados de libertad en la prueba del chi-cuadrado se corresponden con el número de frecuencias observadas que pueden variar libremente sin modificar ninguna frecuencia esperada.

REFERENCIAS BIBLIOGRÁFICAS

Álvarez González, F. 1996. *Investigación estadística con BMPD.* Cádiz: Universidad de Cádiz.

Beltrán, M. 1986. "Cinco vías de acceso a la realidad social". El *análisis de la realidad social.* Comp. M. García Ferrando, J. Ibáñez y F. Alvira. Madrid: Alianza. 17-47.

Butler, C. S. 1985. *Statistics in Linguistics.* Oxford: Blackwell.

Conde, F. 1994. "Las perspectivas metodológicas cualitativa y cuantitativa en el contextos de la historia de las ciencias". *Métodos y técnicas cualitativas de investigación en ciencias sociales.* Coord. J. M. Delgado y J. Gutiérrez. Madrid: Síntesis. 53-68

Coulmas, F. ed. 1996. *The Handbook of Sociolinguistics.* Oxford: Blakwell.

Fasold, R. 1984. *The Sociolinguistics of Society.* Oxford: Blackwell. [cito por la traducción española: *La Sociolingüística de la Sociedad.* Madrid: Visor, 1996].

Fasold, R. 1990. *Sociolinguistics of Language.* Oxford: Blackwell.

Fernández, M. 1997. "Los orígenes de la Sociolingüística". *II Jornadas de Lingüística.* Dir. M. Casas. Cádiz: Universidad de Cádiz. 105-132.

Ferrán Aranaz, M. 1997. *SPSS para Windows. Programación y análisis estadístico.* Madrid: McGraw-Hill.

Holmes, J. 1992. *An Introduction to Sociolinguistics.* Londres: Longman.

Horvath, B. y D. Sankoff. 1987. "Delimiting the Sydney speech community". *Languaje in Society* 16: 179-204.

Hudson, R. A. 1996 (1980). *Sociolinguistcs*. Cambridge: CUP.

Klein, H. 1991. "INTEXT/PC-A Program Package for the Analysis of Texts in the Humanities and Social Sciences". *Literary and Linguistics Computing* 6-2: 108-111.

Kretzschmar, Jr. W. A. y E. W. Schneider. 1996. *Introduction to Quantitative Analysis of Linguistic Survey Data*. California: Sage

Lastra, Y. 1992. *Sociolingüística para hispanoamericanos. Una introducción*. México, D. F: El Colegio de México.

Lorenzo, A. M. 1994. "O tratamento informático do material sociolingüístico: algunhas suxestións a partir dos protocolos de investigación". Ed. Gómez Guinovart, J. *Aplicaciones lingüísticas de la informática*. Santiago: Tórculo. 97-112.

MacRae, S. 1995. *Modelos y métodos para las ciencias del comportamiento*. Barcelona: Ariel.

Mochmann, E. 1985. "Análisis de contenido mediante ordenador aplicado a las ciencias sociales". *Revista Internacional de Sociología* 43-1: 11-43.

Moreno Fernández, F. 1994. "Status Quaestionis: Sociolingüística, estadística e informática". *Lingüística* 6: 95-154.

Moreno Fernández, F. 1998. *Principios de sociolingüística y sociología del lenguaje*. Barcelona: Ariel.

Ramallo, F. F. y G. Rei Doval. 1995. *Publicidade e Lingua*. Santiago: Consello da Cultura Galega.

Ramallo, F. F. y G. Rei Doval. 1997. *Vender en galego*. Santiago: Consello da Cultura Galega.

Romaine, S. 1994. *The language of Society*. Oxford: Blackwel.

Sánchez Carrión, J. J. 1995. *Manual de análisis de datos*. Madrid: Alianza.

Sankoff, D. 1988. "Variable Rules". Eds. Ammon, U., N. Dittmar y K. J. Mattheier. *Sociolinguistics. An International Handbook of the Science of Languaje and Society*. Berlin, Walter de Gruyter, vol.2. 984-997.

Sankoff, D. y W. Labov. 1979. "On the uses of variables rules". *Language in Society* 8: 189-222

Seminario de Sociolingüística. 1994. *Lingua Inicial e Competencia Lingüística en Galicia*. A Coruña: Real Academia Galega.

Seminario de Sociolingüística. 1995. *Usos Lingüísticos en Galicia*. A Coruña: Real Academia Galega.

Seminario de Sociolingüística. 1996. *Actitudes Lingüísticas en Galicia*. A Coruña: Real Academia Galega.

Spolsky, B. 1998. *Sociolinguistics*. Oxford: Oxford University Press.

Stevens, J. 1996. *Applied multivariate statistics for the social sciencies*. New Jersey: Lawrence Erlbaum Ass.

Tešitelová, M. 1992. *Quantitative Linguistics*. Amsterdam: John Benjamins.

Visauta Vinacua, B. (1997). *Análisis estadístico con SPSS para Windows*. Madrid: McGraw-Hill.

Wardhaugh, R. 1992 (1986). *An introduction to sociolinguistics*. Oxford: Blackwell.