Universidad de los Andes
Maestría en Lingüística

Metodología de la Investigación Lingüística
Prof. Valmore Agelvis C.

> Eco define humorísticamente la estadística como la ciencia según la cual, si dados dos hombres (AyB) y cuatro pollos, y el hombre A se come los cuatro pollos, la estadística dirá que la distribución de los pollos es del 50% por hombre.

1. ESTADÍSTICA LINGÜÍSTICA

Ch. Müller (1973) Es necesario insistir aquí sobre una función fundamental, que encontramos frecuentemente. La labor del estadístico comienza en verdad después de la elección y del recuento de los datos numéricos.
El texto no es una serie indeferenciada de elementos, sino una sucesión de unos «tipos» en número limitado. Sabemos que la estadística opera sólo sobre datos numéricos. No puede sometérsele por lo tanto ninguna materia si no está previamente cuantificada. Lo que supone que el objeto estudiado lleva consigo ciertos caracteres cuantificables, y que se juzga oportuno aislar algunos de estos caracteres para someterlos a las operaciones estadísticas.

G. Miller (1981) : Si le preguntamos a una persona normal qué es el lenguaje, lo más seguro es que para contestarnos recurra a las utilidades sociales y personales que tiene esta capacidad... tenemos que distinguir tres tipos de respuestas posibles para la pregunta.. En primer lugar, podríamos dar una respuesta estructural: una lengua es un conjunto de secuencias de palabras. En segundo lugar, podríamos responder haciendo referencia a los procesos que operan en el lenguaje: una lengua sería un conjunto de habilidades que capacitan a una persona para emitir y comprender esas secuencias de palabras. Y, en tercer lugar, podríamos dar una respuesta funcional, basada en el uso del lenguaje: una lengua seria un conjunto de convenciones sociales que regulan el empleo de

las habilidades mencionadas para articular las secuencias de palabras con el fin de alcanzar objetivos determinados.

Vamos a ocuparnos de la estructura. La característica estructural más evidente del lenguaje humano es que consiste en secuencias -a veces bastante largas -de símbolos. La primera cuestión que tenemos que abordar es, portanto, lo de por qué motivo esto es así.

Para estudiar el problema de la secuenciación nos resultará útil examinar antes un sistema de señalización más sencillo que el lenguaje humano, pero que también posee esa característica. Pensemos en un mecanismo que sólo fuera capaz de emitir dos señales distintas. Vamos a llamar a estas señales 0 y 1. Si el vocabulario total del sistema se redujera a estas dos señales, sólo podrían nombrarse dos cosas. Pero su vocabulario puede aumentar de tamaño si el sistema es capaz de emitir estas señales atómicas en parejas: 00, 01, 10, y 11. De esta manera e dispositivo sería capaz de poner nombre a cuatro cosas distintas. Si pudiera emitir ambas señales en tríos, entonces el sistema podría disponer hasta de ocho nombres: 000, 010, 011, 100, 101 110, 111. Cuanto mayor sea la secuencia de señales, más amplio será el vocabulario. Con combinaciones de ***n*** elementos, el vocabulario dispondría de 2^n nombres distintos. La regla general sería la siguiente: **m** señales atómicas distintas combinadas en secuencias de longitud de longitud **n** dan lugar a un total de $\mathbf{m}^n$ etiquetas.

Como el tamaño potencial del vocabulario aumenta de forma exponencial cuando el tamaño de las secuencias aumenta de forma lineal, el método de secuencación es un procedimiento excelente para conseguir un vocabulario de gran tamaño a partir de un número limitado de señales atómicas. De hecho, todos los lenguajes humanos emplean la secuenciación con este fin. Por eso es tan importante que poseamos la capacidad de producir secuencias diferentes de señales y responder de forma diferencial a ellas para poder hablar y entender un lenguaje humano.

En realidad la estrategia de secuanciación es tan poderosa que las lenguas humanas hacen doble uso de ellas en sonidos y en palabras... El peligro de

cometer errores es un factor que limita las posibilidades de utilización del principio secuencial. .. Si queremos que exista la posibilidad de reconocer los errores, lo que tenemos que hacer es, no usar todas las secuencias posibles. Los ingenieros de comunicación usan el término «redundancia» para referirse a la repetición innecesaria de palabras que se produce cuando el principio de secuenciación no se aprovecha al máximo. La existencia de redundancia significa que las secuencias que utilizamos de hecho son más largas de lo que en teoría deberían ser.

Evidentemente, de acuerdo con esta definición, el lenguaje humano es redundante. No podemos pronunciar todas las secuencias posibles de sonidos lingüísticos. Para examinar esta cuestión vamos a utilizar el ejemplo del inglés escrito. En un alfabeto de 26 letras podría haber 26 palabras de una sola letra, 676 de dos, 17.576 de tres, y 456.254 de cuatro. Es decir, usando sólo como máximo cuatro letras por palabra, tendríamos un vocabulario de 475.254 palabras distintas; más o menos el número de vocablos que aparecen recogidos en el Webster's New International Dictionary. Pero las palabras más largas del diccionario tienen bastante más de cuatro letras.

Lo mismo sucede en el segundo nivel: no todas las secuencias de palabras constituyen una oración inteligible. Aunque somos capaces de pronunciar una secuencia de palabras como «vacas dientes de sable gas empapar incapacidad», al hacerlo le imprimimos una entonación de lectura de lista y no de oraciones, puesto que evidentemente estas palabras no forman una oración. Según algunos cálculos, si un idioma fuera capaz de aprovechar al máximo todas las secuencias posibles de letras para formar palabras, y todas las secuencias posibles de palabras para formar oraciones, sus libros tendrían un tamaño cuatro veces menor que el nuestro. Según este criterio, el 75% de lo que escribimos es redundante. Pero este tipo de redundancia no supone una total pérdida de tiempo, ya que gracias a ella disponemos de una especie de seguro contra el ruido. Nuestra capacidad para identificar mejor las palabras cuando van en oraciones que cuando las oímos aisladas se debe al hecho de que sabemos que algunas secuencias de palabras no pueden existir. (Cap. 7)

2.- Ley de Zipf

En numerosos escrutinios de idiomas y textos diversos, observa que si todas las palabras de un texto se ordenan según su frecuencia decreciente y si se les atribuye un numero de orden (rango), el producto de la frecuencia por el rango es sensiblemente constante (excepto para las primeras y las últimas frecuencias de la lista); tendría que aparecer que por ejemplo la 10° palabra es dos veces más frecuente que la 20°, 5 veces más frecuente que la 50°, etc. Zipf interpreta esto como un efecto de la ley del mínimo esfuerzo.

Malmberg (1973): Ya en 1897-98, F.W.Käding, en su *Häufigkeitswörterbuch der deutschen Sprache*, había demostrado (usando como material 11 millones de palabras) que las 15 palabras más comunes representan 25% del número total de palabras de un texto, que las 66 más comunes representan 50% del texto, y las 320 más comunes 72% del total. Así, con un vocabulario solamente de 320 palabras, una persona conseguiría entender tres cuartas partes de las palabras de cualquier texto.

Las cifras de Käding han sido confirmadas por posteriores investigaciones de otros lenguajes. En un examen general de los métodos y resultados de la estadística del vocabulario, Pierre Guiraud (*Les caractères statistiques du vocabulaire,* 1954) resume los resultados alcanzados hasta ahora en dos principios que son: 1) en cualquier texto dado se encontrará que un número muy pequeño de palabras constituye la mayor parte del texto; 2) un número muy reducido de palabras, bien elegidas, cubrirá la mayor parte de cualquier texto. Como ejemplos presenta las cifras siguientes: a) las 100 palabras más comunes cubrirán 60% de cualquier texto; b) las 1000 palabras más comunes cubrirán 85% de cualquier texto. Todo el resto del vocabulario, por tanto, no representará más que 2.5% del vocabulario de cualquier texto dado.

Zipf había observado que el producto de la frecuencia de una palabra (o sea el número de veces que se presenta en un texto dado) y su posición ordinal o rango (en la lista de frecuencia: la palabra más común tiene rango 1, le sigue la de rango o posición ordinal 2, etc.) es constante. Esto se puede expresar mediante la fórmula **f x r = constante**, donde **f** es la frecuencia y **r** el rango de la palabra. Por ejemplo, en la novela Ulysses de james Joyce, la palabra que ocupa el décimo lugar en el orden de frecuencias es usada 2653 veces (fr = 26530), la de rango 100 aparece 265 veces (fr = 26500), orden 10000 sólo 2 veces (fr= 20000), y así sucesivamente. Fundándose en resultados de esta clase, Zipf forma una interesante hipótesis. Esta simetría es vista como un equilibrio entre dos fuerzas opuestas. El habla es gobernada por dos tendencias. El hablante tiende a repetir la misma palabra tanto como sea posible, es decir, a usar palabras como "cosa" y "bueno", o pronombres y otras palabras sustitutivas en lugar de la palabra exacta requerida por el contexto. El oyente, por su parte, necesita la máxima claridad, con descripciones precisas y la mayor variedad posible en palabras usadas. Entre los dos extremos de "la misma palabra para todos los conceptos", se establece un equilibrio expresado por la ecuación anterior, que representa en verdad el principio del mínimo esfuerzo. Zipf ha hallado confirmación de su aplicabilidad en numerosos textos, que van de la Biblia a T.S. Eliot, y en lenguajes tan diferentes como los de indios de Norteamérica y el chino.

El mismo principio opera también en el plano expresivo del lenguaje. La auténtica enunciación de la cadena de sonidos puede considerarse como el resultado de dos fuerzas opuestas: la tendencia a ejercer el mínimo esfuerzo y la necesidad de darse a entender.

3.-Diccionario de estadística

López Morales (1994): Aparte de conceptos como frecuencia, proporciones y porcentajes, la estadística descriptiva que interesa al lingüista trabaja con dos tipos de operaciones: 1) la determinación de parámetros de posición (media, mediana y moda) y 2) la determinación de parámetros de dispersión (como las diversas clases de desviación, los análisis de varianza y prueba t).

Magnitudes continuas y discontinuas (o «discretas»): Los caracteres cuantitativos son continuos o discontinuos. El número de niños (o de personas a cargo de un individuo es un caracter **discontinuo**: entre la clase de los individuos que tienen dos años y los que tienen tres no hay nada. Se pasa de un valor a otro sin que se pueda imaginar un valor intermedio. La talla de los individuos, por el contrario, es un caracter **continuo**: entre un individuo que mide 165 cms. Y otro que mide 166 existe teóricamente una infinidad de casos posibles. (Ch. M.)

Población y muestras: De sus aplicaciones demográficas, que son las más antiguas, la estadística ha conservado la costumbre de denominar **población** a todo el conjunto de objetos cualesquiera sometidos a análisis, e **individuo** a cada uno de estos objetos, a cada uno de los elementos del conjunto. Bajo este punto de vista, se puede considerar un texto como una «población» de frases, o de palabras, o de fonemas, etc. Todo el mundo conoce hoy el principio de los sondeos de opinión: para hacerse una idea de la actitud de una población (en el sentido restringido y humano del término, esta vez) sobre una cuestión cualquiera, no se interroga nunca a todos los individuos que la componen, sino a un cierto número de entre ellos, que constituyen entonces una **muestra**. (Ch. M.)

Caracteres cuantitativos y cualitativos: Los seres humanos que constituyen la población de un pueblo, de una ciudad, de un país tienen cada uno de ellos una edad, una talla, un peso, un cierto «número de personas a su cargo» (número que

puede ser cero), una superficie de habitabilidad, una renta media, etc. Cada uno de estos datos es un **caracter cuantitativo**, que, para cada individuo, se traduce por un número. Otros caracteres, unidos a los mismos individuos, son **cualitativos**, y no pueden traducirse por números: el sexo, el color de los ojos y el de los cabellos, la profesión, el origen geográfico y social, la religión o política que profesan, etc. Estos caracteres no responden a una cuestión «¿cuánto?», sino a una o varias cuestiones «¿qué?». Incluso si conviniésemos en representar algunos de entre ellos por números afectados a las categorías que determinan, no se trataría nada más que de un «código»; este procedimiento no justificaría los operaciones aritméticas, y sólo serviría de criterio de clasificación. Por ejemplo: personas nacidas en el municipio, 1; nacidas en la provincia, 2; en otras provincias de la misma región, 3; en otras regiones, 4; en un país extranjero europeo, 5; etc. (Ch. M.)

Efectivos y clases: Pero cuando contamos los individuos que, para un caracter cualitativo dado, se ordenan en una misma **clase** (por ejemplo sexo masculino, o cabellos rubios, u originarios de un país extranjero), obtenemos los **efectivos**, o datos numéricos que podrán ser sometidos a operaciones estadísticas. (Ch. M.)

Frecuencia: La vuelta de un mismo carécter cualitativo, o del mismo valor de un carácter cuantitativo nos suministra los efectivos: efectivo de los adjetivos, efectivo de los versos de 8 palabras, de nueve palabras, etc. En Rodogune, hay 436 versos de 8 palabras (**frecuencia absoluta** - de una muestra de 100 sobre 1.844 versos-), o sea 23,64% ó 0,236 4 (**frecuencia relativa**). (Ch.M.)

Media, Mediana y Modo: Haremos esta experiencia sobre el texto de Rodogune de Corneille. Saco al azar un verso, el 675:

Est-il une constance à l'épreuve du foudre?

Donde la variable x toma el valor 9. Continúo de 10 en 10 versos (V. 685, 695..., 865), lo que me proporciona estos 20 valores x:

9 8 7 11 9 8 9 9 5 8 10 8 8 9 7 10 9 11 7 7 10

Ordenémosles en el orden creciente:

5 7 7 7 8 8 8 8 8 9 9 9 9 9 9 10 10 10 11 11.

Se puede comprobar que el valor que se encuentra en medio de esta serie, el que ocupa los rangos décimo y undécimo, es 9; es la **mediana**.
Se comprueba también que nueve es el valor cuya frecuencia es más elevada; se llama el **modo**.
Nos inclinamos más espontáneamente aún a buscar la **media** de estos 20 valores, es decir a sumarlos y a dividir su suma por su número, lo que nos da 172/20= 8,60.
Mediana, modo y media son los índices o parámetros de posición; muestran en qué parte del campo los valores tienden a acumularse.

La media (X) se saca aplicando la siguiente fórmula: $X = \sum x / N$ **,** donde x son los valores de la variable y N, el número total de casos, es decir, que dados los siguientes valores de una variable Z: 60, 64, 68, 73, 74, 79, 84, 89 y 92, la fórmula se llenaría así

X= 60 +64 + 68 + 73 + 74 + 79 + 84 + 89 + 92/ 9

X =683/ 9

X = 75.8

Las medias se utilizan principalmente para efectuar comparaciones. Por ejemplo: en una investigación sobre la disponibilidad léxica en niños de primero, tercero y quinto se trabajo con una muestra estratificada por niveles socioculturales: bajo, obrero y medio. A la hora de organizar los meteriales para interpretarlos se creyó oportuno conocer la media de disponibilidad de los grados analizados y costatar con ella los resultados de cada nivel sociocultural. (L.M. 94)

Primer Grado: B: 39,6; O: 54,5; M: 64,4. Localice la media:

Tercer Grado : B: 68,4; O 74,1; M: 84,2. Localice la media:

Quinto Grado : B:72,9; O:69,7; M: 100 Localice la media:

Sacar la **Media:**

Caso A	Caso B	
11	08	
14	20	
14	12	
13	15	X A=
12	18	
-	07	X B=

Ch. Müller (1973) define la **Media** (**ponderada)** así :

x_i	n_i	$n_i x_i$
5	1	5
6	0	0
7	3	21
8	5	40
9	6	54
10	3	30
11	2	22
	$\sum n_i = 20$	$\sum n_i X_i = 172$

La columna **xi** da los valores tomados por la variable, la columna **ni** los efectivos observados para cada valor, es decir el número de versos que tienen 5, 6, ..., 11 palabras en la muestra; el producto **ni xi** representa pues el número de palabras proporcionadas por cada uno de estos grupos de versos; nuestros 20 versos han proporcionado 172 palabras, lo que da la media de 8,60 palabras por verso. La **Media**, que será representada por x («x barra o « x sobrelineada») se formula pues:

$$X = \frac{\sum n_i x_i}{n}$$

y en nuestro ejemplo:

$$\bar{X} = \frac{(1x5) + (0x6) + (3x7) + (5x8) + (6x9) + (3x10) + (2x11)}{20} =$$

Resuelva este ejercicio

X_i	n_i	$n_i x_i$	
4	1	4	
5	0	0	
6	0	0	
7	1	7	$X =$
8	5	40	
9	11	99	
10	6	60	
11	5	55	

$$X = \frac{12}{\frac{1}{12}}$$

La Mediana (Md) suele aplicarse muy poco en lingüística. Se trata de una medida que neutraliza el efecto desequilibrador de las series que cuentan con valores muy extremos, ampliamente apartados del patrón regular de los otros datos. La serie

8, 64 68, 73, 74, 79, 84, 89, 92

esta aparecería desequilibrada, entre 8 y 64 hay 56. Sacar una media sobre este conjunto de valores nos alejaría de la realidad: **631/9=70.1** Para esto se acude a la mediana, que busca el valor medio del conjunto, sin más operación que un conteo de miembros de la serie. En este ejemplo es **74**, ya que es el punto intermedio. Cuando el conjunto está integrado por un número par de miembros, la mediana es un decimal. (L.M.)

El **Modo** per se, es una auténtica rareza en nuestros estudios. Se calcula a través de la siguiente fórmula:

$$Mo = 3Md - 2x$$

donde Md= mediana y X= la media **Mo= 3x74-2x75.8**

Mo= 222-151,6

Mo= 70,4

Parámetro de Dispersión (Fluctuación) : Se concibe que dos variables tengan la misma media, pero varianzas muy desiguales: dos alumnos pueden tener, en sus notas escolares, la misma media, pero las notas de uno pueden ser más estables, más centradas que las del otro. La varianza será un **parámetro de dispersión.** (Ch.M.)

Se saca restando los valores extremos de la variable. Ej. 12-14= 2 caso A;
20-07= 13, caso B del ejercicio de la *Media*.

La Desviación estándar o típica. Se propone medir cuánto se apartan los datos de la investigación de la media, dándonos así información sobre cómo se distribiyen los elementos de análisis alrededor de este punto. La desviación (%) junto con la varianza (v) si son muy manejadas en lingüística.

Su fórmula es:

$$\sigma = \sqrt{\frac{\Sigma(Xi - X)_2}{N}}$$

donde **X**i= punto medio de cada intervalo, **X**= la media y N= número de puntuaciones.

La desviación tipo, que es utilizada como parámetro de dispersión, es la raíz cuadrada de la varianza **(S=sigma)**

xi	**ni**	**(xi--x)**	**(xi-x)2**	**ni (xi __x)2**
5	1	__3,6	12,96	12,96
6	0	__2,6	6,76	0,00
7	3	__1,6	2,56	7,68
8	5	__0,6	0,36	1,80
9	6	+0,4	0,16	0,96
10	3	+1,4	1,96	5,88
11	2	+2,4	5,76	11,52
	20			40,80

La varianza se obtiene realizando la media de los cuadrados de las desviaciones.

$$S = \sum n_i (X_i - \bar{X})^2 = 40,80 = 2,04$$

$$S = \sqrt{ni(Xi - X)^2} = \sqrt{2,04} = 1,43$$

Variable: los conceptos cuantificables y no relacionales de una proposición científica son conocidos con el nombre de **variables**. Si se desea investigar, por ejemplo, la relación entre el "nivel socioeconómico y el tipo de personalidad en adolescentes", las variables de investigación son: nivel socioeconómico y tipo de personalidad, mientras que adolescencia es una constante. Al planear una investigación conviene confeccionar una lista en la cual se clasifiquen las variables en **dependientes**, o efectos, e **independientes**, o causas, y estas últimas en controladas y no controladas. (V.M.)

Variables independientes, si una no influye en otra: nacionalidad e inteligencia, sexo y riqueza; edad y realizaciones oclusivas. Pero si una variable determina la otra, entonces es dependiente, como p. ej. : edad y menstruación.

(Ch.M.) Charles Müller (1973). *Estadística Lingüística*. Madrid,Gredos.
(L.M.) Humberto López Morales (1994). *Métodos de Investigación Lingüística.*. España, Colegio de España
(G.M.) George A. Miller.(1985) *Lenguaje y habla*. Madrid, Alianza Editorial
(V.M.) VictorMorles(1992). *Planeamiento y análisis de investigaciones*. Caracas, Ediciones El Dorado.