Statistics in the analysis of phonetic variation: application of the Goldvarb programme

by Josefina Carrera i Sabaté

Abstract

In this article the autor describes the functioning of the speadsheet programme Goldvarb taking into account its statistical applications which can be used for sociolinguistics analysis purposes, especially regarding phonetic variationism.

1. General aspects

Variationist sociolinguistics has taken major steps towards perfecting quantitative analysis techniques used to demonstrate the importance of social and linguistic contexts in variation. Hence, the "probability theory to the data allows us to extract higher-order regularities that govern variation in the community" (Labov, 1994:25). To this end, "the variationist method aims to calculate the probability that a given linguistic feature will appear in specific linguistic, sociological and contextual circumstances. On the basis of frequency data gathered from a group of speakers, a theoretical model is created from the probabilities of a certain phenomenon occurring when a number of circumstances converge. Statistics marks the extent to which the calculated probabilities are likely and the circumstances that, when occurring simultaneously, best explain a linguistic fact". (Moreno, 1994:95)

Sociolinguistics works with two types of statistics: a) descriptive statistics, which quantitatively counts and orders data extracted from reality; and b) inference statistics, which applies the results of descriptive statistics and adapts them to realities of a specific linguistic community that have not been studied.

The main object of study in inference statistics is the "dependent variable". This variable is determined by independent or explanatory variables which, in language, are linguistic and sociosituational elements. To establish the probability of a variable phenomena occurring in certain ways, first of all, we need to know how many times it has occurred in terms of all possible cases; this figure is obtained by counting the frequencies of appearance of a certain feature in each of the envisaged conditions and in the discourses gathered from a sample of speakers. Once the cases where a factor is present have been counted, we then turn to look for the frequency with which the phenomenon occurs when different explanatory factors coincide.

Probabilistic analysis allows us to find out: 1) the extent to which different groups of explanatory factors determine the variation of an element when all of these explanatory factors act together; and 2) the general behaviour of a community, even though data is only collated from a stratified sample of speakers. The probabilities are used to create a model of the sociolinguistic competence of speakers in order to predict future trends.

The first probabilistic model applied to linguistic analysis is based on an additive model; a logistic-multiplicative model[1] is then reached using this formula:

$$\frac{p}{1-p} = \frac{p_o}{1-p_o} \times \frac{p_i}{1-p_i} \times \frac{p_j}{1-p_j} \times \ldots$$

The mathematical advances in sociolinguistics that took place between 1969 – with Labov's work (1969) – and 1978, were in turn complemented by computer applications that perform statistical calculations. Two programmes that follow this line of research are: Varbrul for the PC and Vax, which performs multinomial analyses,[2] and Goldvarb, for the Macintosh, which calculates probability using a binomial dependent variable.

2. Aims

The aim of this article is to explain briefly the way in which the Goldvarb calculation programme works by studying a process of change in progress analysed in Alguaire (a town in the Segrià region), and to evaluate the use of the programme in variationist sociolinguistics.

The process of phonic variation analysed here forms part of the atonic vocalism of Lleidatà and involves initial e- in absolute initial pretonic position in words such as enciam, escala or eriçó. Diverse dialectological studies

[1] See Cedergren and Sankoff (1974 and 1988), López Morales (1989) and Moreno (1994) for a presentation of models. For a detailed explanation of the limitations of additive and multiplicative models, see Kay and McDaniels (1979).

[2] Multinomial analysis is based on the study of a dependent variable that can have more than two different values; in binomial analysis, the dependent variable has only two values

carried out during the twentieth century reveal that this vowel was traditionally pronounced with the solution [a]; however, we are now observing a process of phonic variation leaning towards replacement of the solution [a] with another corresponding to written forms of language: [e]. The analysis focuses on informants aged 3 to 80 years from the town of Alguaire and basically reveals the influence of written language on the formal model of North-western Catalan, since the phonic changes observed depend on the speaker's knowledge of Catalan, the type of education they had, and – linked to this – their age.[3]

3. Study variables

To enable us to work with the results and obtain the averages and probabilities of appearance of each phonetic segment, values were assigned to the dependent variable (the pretonic vowel) and to the independent variables (those that define the framework of appearance of the dependent variable, i.e. linguistic and extralinguistic factors). The linguistic and extralinguistic variables initially considered to be the most important are as follows:[4]

1) The linguistic variables taken into consideration at the start of the analysis were: type of pretonic vowels, position of stress in each word, consonantal context adjacent to the pretonic vowel, point and mode of articulation of sounds adjacent to the pretonic vowel, vocal chord vibration of the sound following the vowel, quality of the pretonic syllable, etymology of each word, tonic syllable of each word and the sound corresponding to the pretonic position of the same word in Spanish.
2) The extralinguistic variables, directly related to speakers, were: sex, sociocultural status, knowledge of written Catalan, studies and age.
Each of these variables contains different factors (e.g. the variable sex includes the factors female and male) and each factor is represented by a digit or letter (in this case, female is represented by a d and male, by an h). The result of these variables, taken together, gives a series of codifications used to qualify each pronunciation analysed. An example of such a codification is:

aa2cr4a2stbOch-1n9fA08836
which is decoded as follows:

DEPENDENT VARIABLE
a: pretonic sound [a]

INDEPENDENT VARIABLES
a: pretonic type vowel: absolute initial
2: position of stress: vv'v
c: adjacent context before the vowel: existent
r: point of articulation before the pretonic vowel
4: mode of articulation before the pretonic vowel: liquid
a: point of articulation after the pretonic vowel: alveolar
2: mode of articulation after the vowel: fricative
s: vocal chord vibration of the sound following the vowel: mute
t: quality of the pretonic syllable: closed
b: etymology: derivative word with the particle es-
O: tonic syllable of the word
c: initial sound of the corresponding word in Spanish: consonant
h: sex: male
-: sociocultural context: mid-low
1: level of studies: none
n: knowledge of standard Catalan: none
9: age: from 3 to 5 years
f: style of speech: formal
A: linguistic community: Alguaire
088: word code: estisores
36: informant code

4. Application of the Goldvarb statistical programme

The Goldvarb statistical programme contains four essential stages, as does the Varbrul programme. These four steps, which enable us to obtain descriptive statistics, are as follows:

a) Entry of survey results into the token file.[5] This file is the basis of all analyses so the results must be entered correctly. All codifications, similar to that above, for each enunciation by survey subjects are entered here.

[3] For more information, see Carrera (1999 and 2001).

[4] This information is presented as such because, once the statistical analysis had been applied, some factors were not relevant to the study and were disregarded.

[5] These results are codified using the parameters explained above.

b) Once the token file is complete, the condition file must be set up; this will contain the parameters that are to combine the results obtained with the dependent and independent variables of the analysis. Linguistic variation can be explained one way or another, depending on the structure of this file. However, this file is also modified somewhat throughout the analysis, since the combination of factors that best explains the variable phenomenon is required. Thus, among other changes, we find unmodified variables and factors, and regrouped factors and variables.
c) Once the condition file set-up has been completed, the programme creates an iteration file of independent variables: the "cell file". This part of the programme combines all independent factors (tonic syllable, age of informants, studies, etc.) and relates these to the real data of the analysis. The user simply has to click on a series of instructions by which these cells are created. However, each time the conditions are changed, this cell file must be re-created.
d) Lastly, the results of the combination of data and conditions are displayed automatically by the programme in percentages and real amounts.

Basically, it is the user who introduces the tokens and conditions with all the necessary changes in this first phase. The cells and results, however, are calculated by the programme when the appropriate option is chosen. By the end of this first stage, we have the descriptive statistics and real incidences of use for each independent variable, according to the dependent factors.

On completion of the four phases of this initial stage, we move on to inference statistics or probabilistic analysis, where we can calculate: 1) the appearance of the dependent variable in relation to the factors of a range of explanatory or independent variables; in this case, the probability of appearance of the dependent variable (solution [a]) in an open or closed syllable, when the tonic vowel has a timbre or other quality, etc.; 2) the application of the theoretical model according to the data obtained; in this case, to observe the general probability of appearance of the vowel [a] in causes of independent variables, such as the quality of the pretonic syllable and the tonic vowel, age, studies, etc. As explained above, the final form of this part is directly related to the condition file.

The result of the iterations of all the factors, both individually and globally, is worked out by the input of the entry of the rule or application of the theoretical model, which gives the required signification by displaying whether or not the conditions chosen to explain the variation are relevant. If the input is greater than 0.5, the results generally facilitate the rule's application; if it is less than 0.5, they do not. There are two additional operations for observation of general input and the probability of maintenance or change in the variable under study in relation to the different independent factors:

-Binomial Up & Down (U&D)
-Binomial 1 level (1L)

U&D is a process of analysis that looks for significant independent variables to find out the probability of the variable rule application. It displays regression calculations by analysing all groups of independent factors, individually at first, and then in combination, until it arrives at the most likely composition of the variable rule. The result is not the definitive probability, but rather the weight of each independent factor in terms of the other factors of the analysis. We can find out the direction in which the dependent variable will change by the weight of each linguistic or extralinguistic factor: if it is greater than 0.5, it will be influential and, if not, this factor will not have a great deal of influence on the changes of the dependent variable. The U&D calculation also presents the chosen signification of the analysis as good in terms of reasonings calculated previously – before arriving at the optimum analysis, the programme performs a number of calculations and tests such as the logarithm of likelihood and the X-square test, which measures whether the variables analysed are independent. The outcome of this test is observed by the result of p, which must be less than 0.005 to reject what statistics terms the "null hypothesis", i.e. the hypothesis that variation is not caused by the independent factors chosen to explain it.

Once the best combination of factors has been revealed, using these data, we can then move on to observe the definitive probability of each independent factor in the incidence of the variable rule (in this case, maintenance of the pretonic vowel) using the other type of analysis: Binomial 1 level (1L). 1L is a relatively quick way of revealing whether or not the set of initial parameters - independent variables – is adequate to explain the initial rule. For example:

```
BINOMIAL VARBRUL, 1 step • 08/10/1•17:47 •••••••••••••••••••••••••••••••••••
Name of cell file:  01.Cel

Using more accurate method.
Averaging by weighting factors.
One-level binomial analysis…

Run # 1, 117 cells:
Iterations:  1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Convergence at Iteration 15
Input 0.636
```

Group Factor		Weight	App/Total	Input&Weight
1:	t	0.568	0.66	0.70
	o	0.253	0.36	0.37
2:	i	0.349	0.40	0.48
	e	0.489	0.61	0.63
	a	0.565	0.65	0.69
	o	0.413	0.54	0.55
	u	0.624	0.74	0.74
3:	1	0.877	0.92	0.93
	2	0.574	0.73	0.70
	3	0.594	0.72	0.72
	4	0.292	0.37	0.42
	5	0.343	0.46	0.48
	6	0.236	0.37	0.35
4:	9	0.395	0.90	0.53
	7	0.430	0.40	0.57
	5	0.554	0.59	0.68
	2	0.565	0.80	0.69

Cell	Total	App'ns	Expected	Error
tu67	8	4	3.759	0.029
tu65	17	9	10.095	0.292
tu57	21	11	12.594	0.504 (…up to 117)

The initial factors find their optimum point at the fifteenth of the programme's twenty possible iterations. The Input of application of the pretonic vowel [a] on these factors reveals that the rule "is applied" (0.636), i.e. generally speaking, the solution [a] is maintained as it is greater than 0.5, which figure can be attributed to the appearance of both dependent factors.

Returning now to this programme's logistic model formula, we can see how the calculations in the analysis of 1L are reached:

$$\frac{p}{1-p} = \frac{p_o}{1-p_o} \times \frac{p_i}{1-p_i} \times \frac{p_j}{1-p_j} \times \ldots$$

p = **Input & Weight**
(i.e. probab. of
each independent factor)

p_0 **= Input**
the probab. of all
the variable rule.
Here, it is **0.636**

p_i **= Weight**
where factor
t of the 1 group is **0.568**

p_j **= Weight'**
where factor o
of the 1 group is **0.253**

This formula can be used to find out the definitive probability, in this case, of maintenance of the solution [a], for each independent factor if we relate the Input and Weight of each factor as follows:

$$\frac{p}{1-p} = \frac{p_o}{1-p_o} \times \frac{p_i}{1-p_i} \ ;\ where\ \ p_o = 0.636\ and\ p_i = 0.568$$

This equation gives p (Input & Weight) a value that is the probability of maintenance of [a] according to our variable rule: in the case of factor t of the 1 group, p= 0.696; if this figure is rounded up, it corresponds to the programme's third column of results – that is, 0.70.

Thus, 1L reveals the general input of maintenance of [a] (0.636), the weight of each factor, the percentage of use of [a] according to each independent variable, which appears under the name App/Total, and lastly, the probability of maintenance of [a] in relation to these independent factors – input & weight.

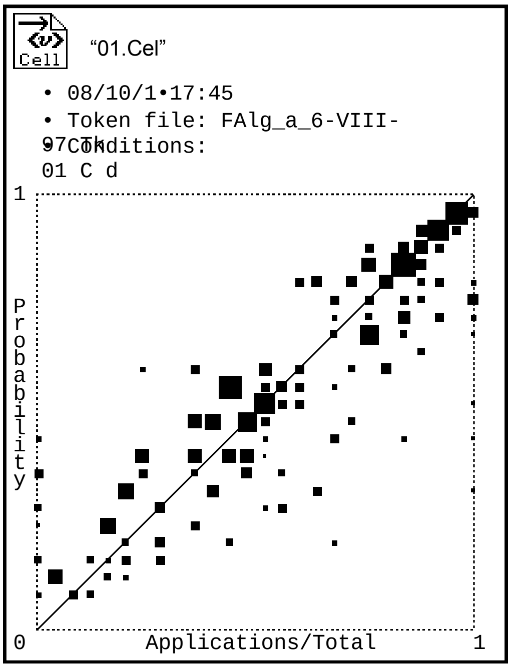
Once the results are obtained, this programme displays a description of the errors between the theoretical or expected probability and the sample used. Clearly, any probabilistic analysis will involve error; if it did not, the field of study would be functional mathematics. However, probability looks for the smallest margin of error between what was expected and the real data with which it works; thus, the closer the columns that reveal frequency (App/Total) and probability (Input & Weight), the more guarantees of success of the analysis.

This part contains three tests to determine whether the theoretical conditions adapt to the study data:

-Logarithm of likelihood (Log. likelihood)
-X-square test
-Scattergram

Using the example of the previous analysis, we obtain:

```
Total Chi-square = 150.8532
 Chi-square/cell = 1.2893
Log likelihood =  -1362.634
Maximum possible likelihood = -1277.040
Fit:  X-square(104) = 171.187, rejected, p = 0.0000
```



```
"01.Cel"
• 08/10/1•17:45
• Token file: FAlg_a_6-VIII-
97Conditions:
01 C d
```

The figure given by log. likelihood must be in relation to the maximum amount proposed by the programme under Maximum possible likelihood. If these two figures are similar, then this is an initial guarantee that the theory fits and that the rule is actually in place.

The second test c2, seen earlier in the U&D, is used in many statistical analyses. The amount required to find out whether the results fit the theory is that following the letter p; the closer it is to 0, the more reliable the rule.

Finally, the scattergram of the data presents these results as a diagram. If the points of convergence between the theoretical model and the real data follow the line of the diagram, the confidence level of the analysis is very high.

It is clear from this brief introduction that, in order to obtain an analysis with a high confidence level using the Goldvarb programme, we need to present the Input of application of the initial rule and one of the tests guaranteeing the likelihood of the theoretical model constructed from the iteration of independent variables. Subsequently, we can move on to interpret the data, which is always related to the figure, 0.5. Thus, we can determine when a factor favours the presence of the initial rule (in this case, maintenance of the solution [a]) and when it does not facilitate application of the initial rule, as seen with factors o (of the variable 1) and 4, 5 and 6 (of the variable 3).

Finally, it is important to bear in mind that, although probabilistic analysis enables us to observe future tendencies, it does not provide an exhaustive analysis of all possible combinations of factors. It is therefore often interesting to complete the analysis by cross tabulating factors to detail the connections of the percentages obtained on each independent factor according to the different crossovers of variables. One of the applications of the above programme allows the creation of cross-tables of factors using the cross tabulation instruction, once the condition, cell and results files have been created. Thus, Goldvarb permits analysis of different features using the same initial classification of data.

5. Goldvarb in variationist sociolinguistics

Variationism has taken significant steps towards adapting a method capable of providing rigorous analyses and reliable linguistic interpretations, and the Goldvarb programme is one of the results.

As we have seen, this programme enables us to obtain descriptive statistics and to find out the inference statistics of variable phenomena very accurately. However, we need to be aware beforehand of the factors susceptible to variation (which can arise from phonetic/phonological, morphological, lexical, semantic or syntactic elements of grammar) and, above all, we need to know how to find the linguistic and extralinguistic factors that can explain, in one way or another, the different variable phenomena. To reach these intuitions, we need to carry out prior qualitative analyses on the variation and nature of the diverse phenomena to be analysed.

While finding the variable and explanatory factors of variation is important, knowing how to combine these adequately in order to obtain the most representative results of the variable reality is no less relevant. This means that we need to look and, eventually, know how to find the best combination of explanatory factors which is very often reached by re-doing the analysis time and time again.

In short, the Goldvarb programme allows us to explain any process of variation, so long as the initial theory is good and is accompanied by a series of factors capable of bearing it out.

6. Bibliography

CARRERA-SABATÉ, J. (1999): *L'Alternança a/e al Segrià.* Doctoral thesis. University of Barcelona. Barcelona.
CARRERA-SABATÉ, J. (2001): "La normativització del català modifica els hàbits fonètics dels parlants?" Llengua i literatura, 12: 175-199.
CEDERGREN, H. J.; SANKOFF, D. (1974): "Variable rules: Performance as a statistical reflection of competence". *Language.* 50: 333-355.
KAY, P.; McDANIEL, C. (1979): "On the logic of variable rules". *Language in Society.* 8: 151-187.
LABOV, W. (1969): *Contraction, Deletion, and Inherent Variability of the English Copula. Language.* 45: 715-762.
LABOV, W. (1994): *Principles of linguistic change. Internal factors.* Blackwell. Cambridge.
LÓPEZ MORALES, H. (1989): *Sociolingüística.* Gredos. Madrid.
MORENO, F. (1994): *Status quaestionis: sociolingüística, estadística e informática. Lingüística.* 6: 95-154.
SANKOFF, G. (1988): *Variable Rules*. U. Ammon; N. Dittmar; K. J. Mattheier (ed.) *Sociolinguistics. An international handbook of the science of language and society.* Walter de Gruyter. Berlin & New York: 984-997.

Josefina Carrera-Sabaté
Universitat de Lleida
Universitat de Barcelona
jcarrera@filcat.udl.es