A PERCENTILE REGRESSION MODEL FOR THE NUMBER OF ERRORS IN GROUP CONVERSATION TESTS

Erkki P. Liski (University of Tampere and the Academy of Finland)
Simo Puntanen (University of Tampere)

Introduction

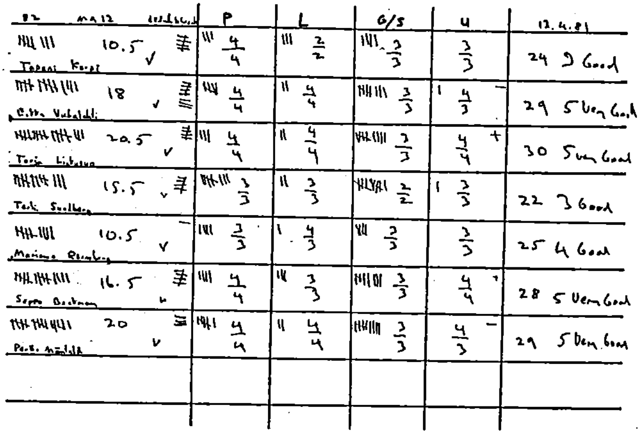
In this paper we put forward a statistical model for analysing the results of group conversation tests in spoken English. The variation of the number of errors was explained by the number of utterances spoken by a testee. Our data were obtained from the results of group conversation tests carried out at the University of Tampere from 1977 up to 1981. The number of students under consideration was 639. On the basis of these same data we analysed in another study (Liski and Puntanen 1978) the language performance of Finnish testees in relation to certain background variables like sex, school results etc. This earlier paper contains a thorough statistical description of our data.

In this study estimates of percentile curves for the number of errors are of greater interest than the mean regression line. In our situation the dependent variable turns out to be a compound Poisson variable. Explicit parametric assumptions are made using a linear model for scale and heterogeneous variance. We apply here the distribution theory described in more detail in two earlier papers (Liski 1984), Liski and Puntanen (1985). Also more far-reaching statistical discussion of the model can be found in the above papers.

1 Data

One day we happened to come upon Mr. John Clarkson, an English teacher in the University of Tampere, carrying some sheets full of figures and because statisticians are incorrigibly dedicated to finding facts from figures - whenever they meet figures - we asked what these sheets contained. He kindly told us; some consequences can be seen from the following pages.

John Clarkson's sheets looked like for example the followong:



82 [illegible]	P	L	G/S	U	12.4.81
[illegible] 10.5 ✓ Tapani Korpi	||| $\frac{4}{4}$	||| $\frac{2}{2}$	[illegible] $\frac{3}{3}$	$\frac{3}{3}$	24 3 Good
[illegible] 18 ✓ [illegible]	[illegible] $\frac{4}{4}$	|| $\frac{4}{4}$	[illegible] $\frac{3}{3}$		$\frac{4}{3}$ −	29 5 Very Good
[illegible] 20.5 ✓ [illegible]	||| $\frac{4}{4}$	|| $\frac{4}{4}$	[illegible] $\frac{3}{3}$	$\frac{4}{4}$ +	30 5 Very Good
[illegible] 15.5 ✓ [illegible]	[illegible] $\frac{3}{3}$	|| $\frac{3}{3}$	[illegible] $\frac{2}{2}$		$\frac{3}{3}$	22 3 Good
[illegible] 10.5 ✓ − [illegible]	[illegible] $\frac{3}{3}$		$\frac{4}{3}$	[illegible] $\frac{3}{3}$	$\frac{3}{3}$	25 4 Good
[illegible] 16.5 ✓ [illegible]	[illegible] $\frac{4}{4}$	[illegible] $\frac{3}{3}$	[illegible] $\frac{3}{3}$	$\frac{4}{4}$	28 5 Very Good
[illegible] 20 ✓ [illegible]	[illegible] $\frac{4}{4}$	|| $\frac{4}{4}$	[illegible] $\frac{3}{3}$	$\frac{4}{3}$ −	29 5 Very Good

The sheet above is a result of group conversation tests in spoken English. At first it struck us as somewhat surprising that a student's conversation ability could be measured by a quantitative method. Now that the problem field has become familiar, we see that there are indeed good possibilities for measuring such an ability, which is no doubt very important in life today.

As one may see, the examination sheet is full of figures implying that a kind quantitative assesment of the testee's ability to Speak English is applied. This testing system has been developed by Folland and Robertson at the University of Tampere, and it is described in Folland and Robertson (1976). In this section the reader is introduced to the data based on the examination sheets. To make the nature of data clear a brief description of the testing method will be next given.

Groups of students are tested in free discussion by a native English teacher for a minimum of five minutes per student. The typical group size is six, giving a 30-minute test for a group. At the beginning of the test a prerecorded tape is introduced for example as being a report to an international study group, after which the main points should be freely discussed in English. The examiners take no part in the discussion, the contents and development of which lie entirely with the students. During the previous terms the students have had considerable practice in this.

A testee's perfomance is assessed on a marking sheet under four catagories: *pronunciation, lexis, grammatical structures,* and *use.* The number of *major* and *minor* errors under the four categories are noted, together with *plus points* under use and lexis. Also *the number of utterances* each testee produces during the test is noted on the marking sheet, where one utterance (roughly estimated) is at least ten words and a part utterance less than ten words. At the end of the test the number of errors and pluses in each category per ten utterances for each testee is calculated, and these figures are then applied to an errors-points scale to arrive at the final mark.

Folland and Robertson transcribed a number of recorded tests and examined the errors which had occurred, devising definitions for them. An error occurs where *the speaker fails to follow the pattern or manner of speech of educated people in use in English-speaking countries today.* Taken statistically major errors were very seldom made.

On the basis of discussions with the examiners, we have combined the number of minor and major errors so that one major error corresponds to two minors (except in the pronunciation category, where one major corresponds to three minors) for the purpose of this study. The figure thus obtained is called simply the number of errors, thus excluding the major-minor division.

The main variables to be considered in this paper are

TFREQ = the *t*otal *freq*uency of utterances given by a testee,

EFR = the total *e*rror *fr*equency of a testee.

The examiners, however, consider errors separately in different categories which gives a more versatile picture of the testee's ability. In our data the variable EFR is simply the number of all errors made by the testee during the test.

It is obvious that the more the testee speaks the larger is the number of his errors. Hence the number or utterances should somehow be taken into account when assessing the testee's ability. As was earlier mentioned, the examiners simply consider the number of errors per ten utterances and then use their errors-points scale to arrive at the final mark. The errors-point scale is based on the examiners' *experience* of the "tolerable" number of errors. One simple *statistical* approach to the errors-points scale would be to consider the *conditional percentiles* of the number of errors (EFR) under different values of utterances (TFREQ).

The main interest of this paper lies between the dependence of the variables TFREQ and EFR, whose scatter diagram is shown in Figure 1 and frequency distributions in Figures 2 and 3. The objective of the study is to arrive at a statistical model of the data and particularly to estimate percentile curves for the variable EFR.

Our data were obtained from the results of group conversation tests carried out in the Universtiy of Tampere by John Clarkson from 1977 up to 1981. Also certain background variables such as sex and matriculation results were observed but they will not be considered in this paper. The number of students under consideration was 639. Originally the data consisted of 826 students but to reduce the heterogeneity of the students' background only students with the matriculation examination were included in the research. The examiners have found that there must be at least six utterances in order to assess the testee's ability in spoken English. On the other hand, ten utterances have always proved to be a sufficient sample size. Hence testees speaking less than six utterances are not included in the data here, thus the variable TFREQ varies from 6 to 40 in these data.

Figure 1

Scatter diagram between TFREQ and the number of errors

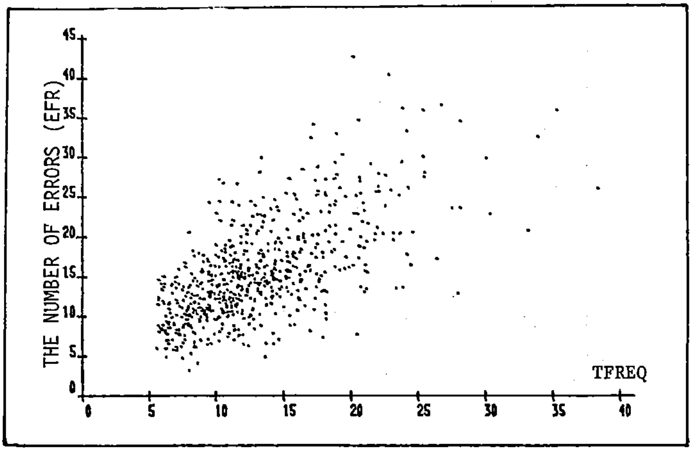


Figure 2

Distribution of the number of utterances (TFREQ)

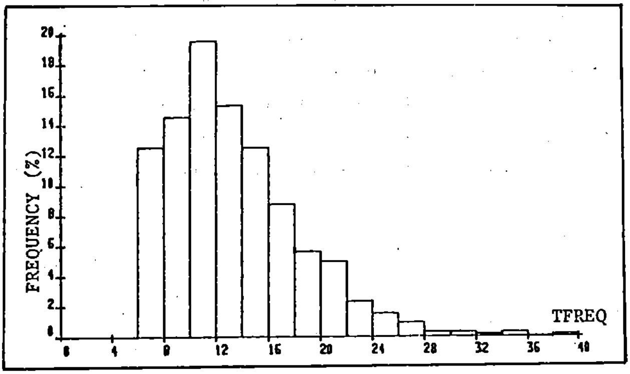
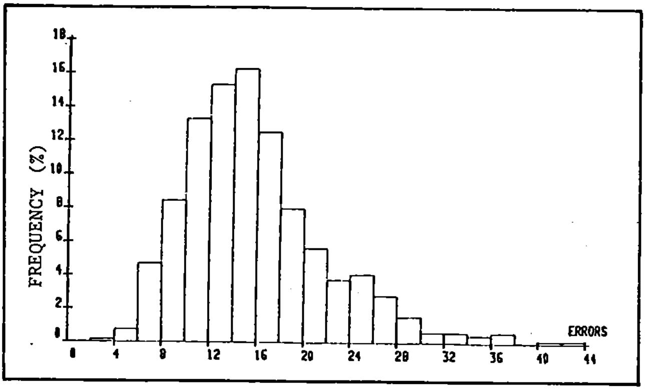


Figure 3

Distribution of the number of errors (EFR)



2 Modelling

As we explained in the previous section, testees spoke from 6 to 40 utterances during a test. The total number or errors made by a student varies from 3 to 43. In the whole sample the average number of errors per one utterance is 1.23. Let us consider first a particular student, who has spoken x utterances (in our sample $6 \leq x \leq 40$). Denote by y_j the number of errors made by a testee in the *jth* utterance he has spoken. Therefore, the set of utterances produced by a student can be characterized as follows:

utterance	1 2 ... x	the total number of errors
the number of errors in one utterance	y_1 y_2 ... y_x	$y = \sum_{j=1}^{x} y_j$

We assume that the number of errors in one utterance follows a Poisson distribution with the mean λ_x, where x is the number of utterances spoken by a student. Therefore we write

$$y_j \sim \text{Poisson}(\lambda_x), \text{ for all } j = 1,2,\ldots, x. \qquad (2.1)$$

Further, we assume that $y_1, y_2, \ldots, y_x$ are stochastically independent. As a sum of Poisson variables the total number of errors $y = y_1 + y_2 + \ldots + y_x$ made by a particular testee is also a Poisson variable:

$$y \sim \text{Poisson}(\lambda_x x). \qquad (2.2)$$

The value of the intensity parameter λ_x describes a student's proneness to errors. A high value of λ_x tells that a student's language ability is bad.

However, it is quite evident that proneness to errors, and hence the value of the intensity parameter λ_x varies from one student to another. Every particular student has a value of λ_x characteristic to him. Therefore, intensity parameter λ_x is a random variable. We have found (Liski and Puntanen 1983), on the basis of an empirical investigation that students who spoke seldom were more prone to errors than more talkative testees. Therefore we assume that the mean of λ_x depends on TFREQ(=x), and we denote $E\lambda_x = \mu_x$. Further, suppose that the variance of variable λ_x is σ^2, and $var(\lambda_x) = \sigma^2$ does not depend on x.

As was stated before, the mean μ_x decreases as x (the number of utterances) increases. In other, words talkative students tend to be better than students who speak seldom.

Taking the intensity parameter as a random variable yields the compound Poisson distribution (see e.g. Johnson & Kotz 1969). In this case the expectation and variance of (2.2) are

$$Ey = \underset{\lambda_x}{E}[E(y|\lambda_x)] = \mu_x x \qquad (2.3)$$

and

$$\mathit{var}\ y = \underset{\lambda_x}{E}[\mathit{var}(y|\lambda_x)] + \underset{\lambda_x}{\mathit{var}}[E(y|\lambda_x)]$$

$$= \mu_x x + \sigma^2 x^2 \ , \qquad (2.4)$$

where x is the number of utterances (see e.g. Rao 1973, p. 97).

However, the number of errors per utterance (EFR/TFREQ) is used instead of EFR when assessing the student's ability in pronunciation. These figures are then applied to an errors-points scale to arrive at the final mark in pronunciation. As a direct consequence of (2.3) and (2.4) we obtain

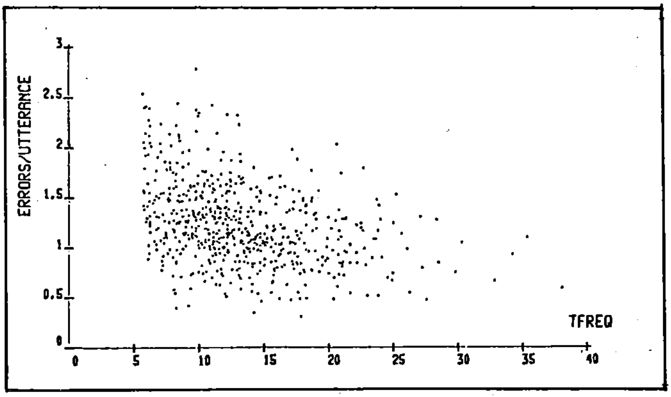
$$E(y/x) = \mu_x \ , \qquad (2.5)$$

$$\mathit{var}(y/x) = \mu_x/x + \sigma^2 \ . \qquad (2.6)$$

Thus we find that also the variance of y/x depends on x .

Figure 4

Scatter diagram between TFREQ and the number of errors per one utterance



3 Regression

We assume a linear relationship between y and x. That is

$$y = \alpha + \beta x + \varepsilon \ , \tag{3.1}$$

where

$$E\varepsilon = 0 \tag{3.2a}$$

and

$$var\ \varepsilon = \mu_x x + \sigma^2 x^2 \ .$$

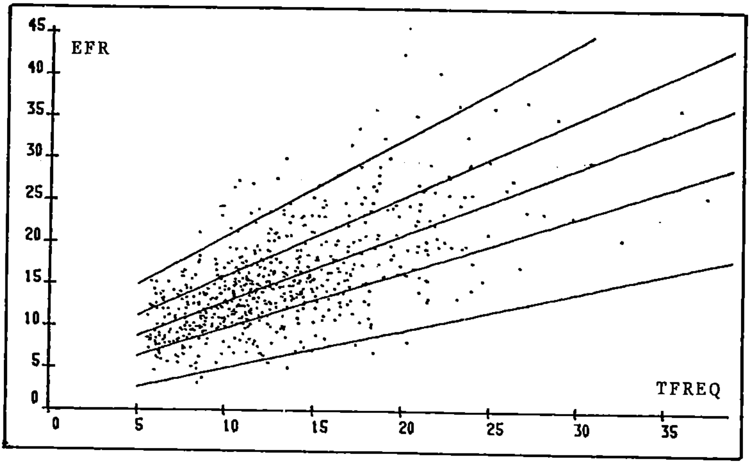
However, we are more interested in estimating different percentile curves as a function of x than in estimates of regression parameters α and β. Since final marks can be determined on the basis of the figures y, the examiners simply need various percentiles as a function of x. The percentile curves are useful both in constructing errors-point scales and generally in comparison of error frequencies with different values of x.

We estimate parameters α, β and σ^2 by the method of maximum likelihood, when the distribution of ε was assumed to be normal. The estimates were found to be

$$\begin{aligned}\hat{\alpha} &= 4.788 \text{ (sd. } 0.481),\\ \hat{\beta} &= 0.806 \text{ (sd. } 0.039),\\ \hat{\sigma}^2 &= 0.032 \text{ (sd. } 0.007).\end{aligned}$$

Figure 5

Estimates of the 5th, 25th, 50th, 75th and 90th percentile curves for the regression of EFR on TFREQ.



The estimated median line is

$$EFR_{0.50} = 4.778 + 0.806 \text{ TFREQ}$$

and the variance of EFR can be expressed as a function of TFREQ

$$\mathit{var}(\mathrm{EFR}) = 4.778 + 0.806\ \mathrm{TFREQ} + 0.032\ \mathrm{TFREQ}^2.$$

The estimate of the 100 *p*th percentile curves for the regression of EFR on TFREQ were found to be of the form

$$\mathrm{EFR}_p = 4.778 + 0.806\ \mathrm{TFREQ} + (4.778 + 0.806\ \mathrm{TFREQ} + 0.032\ \mathrm{TFREQ}^2)^{1/2} Z_p,$$

where Z_p is the 100 *p*th percentile of the standardised normal distribution. As can be seen from Figure 5 percentile curves are rather linear.

When assessing a student's language ability, we usually consider the number of errors per utterance (u = EFR/TFREQ). Percentile curves for this variable are of the form

$$u_p = 0.806 + (4.778/\mathrm{TFREQ}) + (0.032 + 0.806/\mathrm{TFREQ} + 4.778/\mathrm{TFREQ}^2)^{1/2} Z_p.$$

We see from Figure 6 that proneness to errors decreases when TFREQ increases. Thus more talkative testees are better than the testees who rarely speak. On the other hand, when the value of EFR/TFREQ is low (the best students) proneness to errors does not seem to depend on TFREQ (the 5th percentile curve). When TFREQ increases, the value of $u_{0.50}$ approaches 0.806 and the standard deviation of u approaches $0.032^{1/2} = 0.179$.

4 Discussion

We clearly perceive that the distributions of EFR and of the relative number of errors (EFR/TFREQ) depend on TFREQ. Although talkative testees make more errors than testees who rarely speak, proneness to errors decreases when TFREQ increases. Thus talkative students do better in this conversation test than testees who seldom speak. It also proves out that male students make more errors than female students, but there is almost no difference in the relative number of errors between boys and girls (Figure 6). In fact, talkative boys were less prone to errors than talkative girls.

Figure 6

Estimates of the 5th, 25th, 50th, 75th and 95th percentile curves for the regression of EFR/TFREQ on TFREQ.

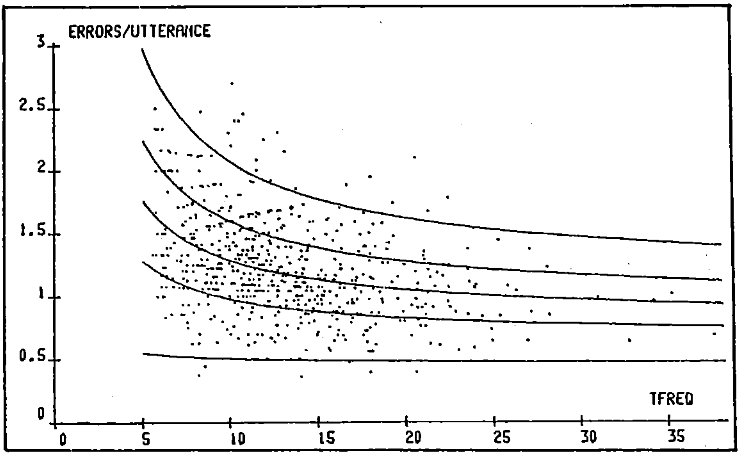
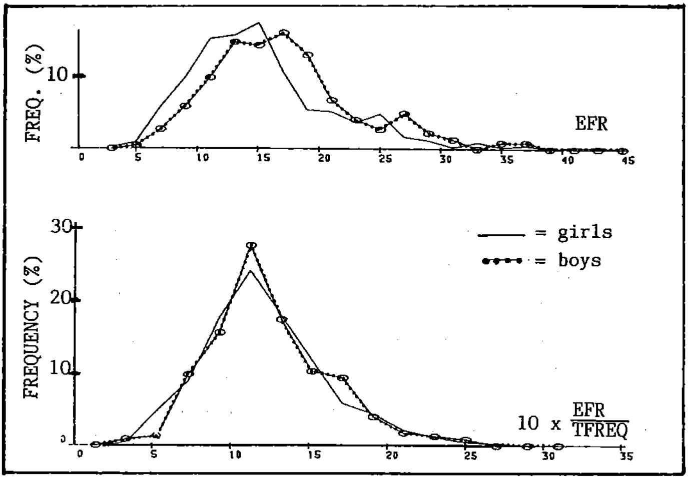


Figure 7

Distributions of the number of errors(EFR) and EFR/TFREQ.



Bibliography

Folland, D., and D. Robertson 1976. "Towards objectivity in group oral testing". English Language Teaching Journal 30: 156-167.

Johnson, N.L. and S. Kotz 1969. Discrete Distributions. New York: John Wiley & Sons.

- 1970. Continuous Univariate Distributions - 1. New York: John Wiley & Sons.

Liski, E.P. 1984. "Regression analysis with two mixed forms of heteroscedasticity". Communications in Statistics - Theory and Methods 13: 1015-1030.

Liski, E.P. and S. Puntanen 1978. "A statistical analysis of the results of group conversation tests in spoken English". Report A 25, Department of Mathematical Sciences, University of Tampere, Finland.

- 1985. "A percentile regression model with an application to error frequency in group conversation tests". The Canadian Journal of Statistics 13, No. 1, in press.

- 1983. "A study of the statistical foundations of group conversation tests in spoken English". Language Learning 33: 225-246.

- 1984. "Statistical modelling of the group conversation tests in spoken English". An abstract in the Proceedings of the 7th World Congress of Applied Linguistics, Brussels, Belgium, August 4 - 10, 1984.

Rao, C.R. 1973. Linear Statistical Inference and Its Applications. Second Edition. New York: John Wiley & Sons.