"QUANTOS FALANTES PRECISO?" ou
A Questão do Tamanho da Amostra

Maria Luiza Braga — Universidade Federal Fluminense
Anthony J. Naro — Universidade Federal do Rio de Janeiro

Antes mesmo de se iniciar qualquer pesquisa empírica em lingüística coloca-se um problema prévio bastante sério: "quantos falantes preciso estudar?", isto é, qual é o tamanho da amostra necessário. E o pesquisador potencial pensa imediatamente no número imenso de falantes que existem por aí no mundo real, isto é, no universo de falantes, caindo a seguir num estado de depressão profundo do qual dificilmente poderá se libertar.

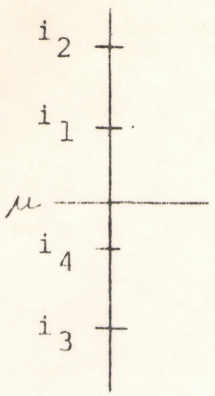
A mensagem que queremos transmitir é a de que este estado de depressão não se justifica porque o número total de falantes é, de fato, totalmente irrelevante. Do ponto de vista da precisão ou confiabilidade dos resultados obtidos não importa quantos falantes foram relegados ao esquecimento. Importa sim o número dos que foram efetivamente estudados bem como sua distribuição, isto é, se são representativos do grupo, não sendo apenas casos extremos ou pouco comuns. Repetindo, é o número de falantes estudados que determina a validade dos resultados, isto é, uma amostra de N informantes será tão boa ou tão ruim para uma população de 100 quanto para uma população de 1.000 ou de 1.000.000, as outras coisas sendo iguais. Mas, ainda assim, resta o problema principal: "quantos falantes?". A resposta depende do grau de variabilidade do fenômeno sob estudo: um fenômeno relativamente uniforme poderá ser estudado com menos falantes do que outro fenômeno que varia muito de falante para falante. Passamos agora a explicar o porquê destes fatos.

Seja uma variável lingüística qualquer que queremos estudar, digamos, por exemplo, o número de orações encaixadas por oração principal na fala natural. Queremos determinar o valor

médio deste índice de encaixe numa certa população, digamos, crianças de seis a oito anos de idade. Seria literalmente impossível estudar a fala de todas as crianças desta faixa etária. Portanto, teremos que nos contentar com uma amostra de crianças, tomadas todas as precauções necessárias para que esta amostra seja realmente representativa. Para cada criança verificamos o valor do índice em estudo, para depois calcular o valor médio, considerando todas as crianças da amostra. Poderemos afirmar que este valor médio encontrado é o valor médio do índice para as crianças na faixa de 6 a 8 anos de idade? É óbvio que não, pois se tivéssemos trocado uma ou mais das crianças da amostra por outras teríamos encontrado valores distintos, o que redundaria em outra média fatalmente diferente da primeira em maior ou menor grau.

A situação com que nos defrontamos é a seguinte: se pudéssemos determinar os valores do índice de cada criança da população poderíamos então determinar a média verdadeira, que convencionamos denotar com o caracter grego μ. Mas na realidade temos apenas um certo número N crianças, cada com um valor do índice. A partir destes N valores calculamos o índice médio da amostra, que convencionamos denotar com o caracter latino m. O problema que se nos coloca é então: qual é a relação entre μ, cujo valor queremos saber, e m, cujo valor encontramos experimentalmente?

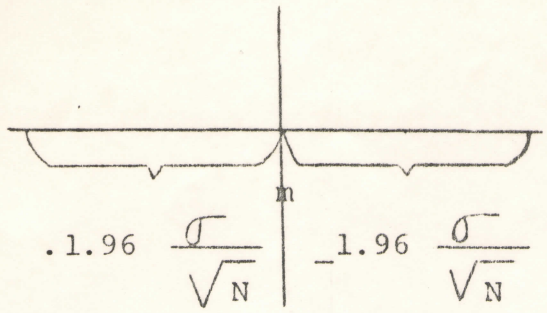
A resposta a esta pergunta vai depender em parte do grau de variabilidade do próprio índice de falante para falante. Se o índice for relativamente uniforme na população o seu valor será sensivelmente o mesmo em todas as crianças e, portanto, a troca de crianças da amostra, acima aventada, fará pouca diferença no valor calculado de m. O grau de variabilidade se mede através do desvio padrão σ.



$$\sigma^2 = \frac{(-i_1)^2 + (+i_2)^2 + \ldots + (-i_N)2}{N}$$

Para calcular o desvio padrão, tomamos a diferença entre μ e cada valor i, levamos ao quadrado para eliminar números negativos, somamos para todos os valores e dividimos pelo número total de valores. No entanto, já que levamos cada diferença ao quadrado, calculamos na verdade não σ mas seu quadrado σ^2, sendo portanto necessário ainda tirar a raiz quadrada. Como se vê, o essencial da fórmula do desvio padrão é a diferença entre a média populacional μ e cada valor individual desta média i. Se estes valores forem todos perto de μ (uniformes, como dissemos acima), as diferenças serão pequenas e portanto o desvio padrão será também pequeno. Se, ao contrário, forem grandes o desvio padrão será grande também.

Lembremos que estamos tentando determinar se nossa média empírica m, determinada a partir da amostra, pode ser aceita como pelo menos próxima ao valor desconhecido da verdadeira média μ. Existe um teorema que nos informa o seguinte: em amostras de N elementos a média verdadeira μ estará num certo intervalo de distância ao redor de m. Esta distância é medida em termos do desvio padrão dividido pela raiz quadrada de N. Portanto, o intervalo ao redor de m, dentro do qual deverá se encontrar μ, será menor se os valores de i forem mais uniformes, reduzindo assim o seu desvio padrão. Este intervalo também será menor se o número de elementos da amostra for maior. De fato, pode-se provar que em 95 de 100 amostras de N elementos μ estará contido num intervalo ao redor de m como se mostra no diagrama:



— estará neste intervalo para 95 de 100 amostras de tamanho N

$$m \pm 1.96 \frac{\sigma}{\sqrt{N}}$$

Se quisermos aumentar o grau de certeza de μ cair no intervalo para 99 de 100 amostras o número 1.96 aumenta para 2.80, isto é, o intervalo fica maior.

Temos agora elementos que nos possibilitam responder à pergunta "quantos falantes?". Podemos utilizar a fórmula do intervalo. Mas primeiro temos que tomar algumas decisões preliminares:

1 - Qual é o grau de certeza que queremos ter que μ, a média verdadeira, estará no intervalo ao redor de m, a média empírica?

2 - Qual é o tamanho do intervalo ao redor de m que consideramos satisfatório?

3 - Qual é o grau de variabilidade do fenômeno em estudo, isto é, qual é o valor de σ?

Para responder 1, podemos utilizar a título de exemplo 95 de 100, correspondendo ao número 1.96. Para responder 2, podemos determinar qualquer número que nos pareça razoável, digamos 0.1, isto é, μ deverá cair entre 0.1 a mais ou a menos do que o valor empírico. Mas 3 é realmente um problema prático. Como seria possível determinar o desvio padrão (V. fórmula)
quando nem sequer sabemos μ. Aliás, é justamente μ que estamos tentando determinar. Em outras palavras se insistíssemos em responder

à pergunta "quantos falantes?" antes de fazer a pesquisa propriamente dita teríamos que fazer outra pesquisa prévia para determinar σ. Mas para calcular σ temos que saber μ, que é o objetivo da pesquisa propriamente dita. O que nos leva à confusão total: temos que fazer uma pesquisa para determinar μ antes de fazer a pesquisa para determinar μ. E realmente não há saída, a não ser a de postergar a resposta à pergunta "quantos falantes?" até que se tenha alguma idéia razoável quanto aos valores de μ e σ. Suponhamos que depois de estudar alguns falantes achamos que σ deve ser aproximadamente 0.15. Temos então que:

$$0.1 = 1.96 \frac{0.15}{N}$$

ou

$$\sqrt{N} = \frac{1.96 \times 0.15}{0.1} = 2.94$$

$$N = 8.64$$

Este resultado nos informa o seguinte: pressupondo um desvio padrão de 0.15, para termos a certeza de encontrar um valor de m, tal que, em 95 de 100 amostras, μ está no intervalo de 0.1 ao redor de m, devemos utilizar uma amostra contendo 9 falantes.

Levando em conta o exposto acima, temos que concluir que não é sensato insistir na questão do número de falantes antes de se começar uma pesquisa lingüística. O jeito prático é por as mãos à obra e, em algum ponto intermediário, parar para fazer um cálculo provisório baseado numa estimativa razoável do desvio padrão e nas decisões quanto aos pontos 1 e 2 acima.

É importante notar que o cálculo do número de falantes da amostra não leva em conta o número total de falantes do universo, confirmando nossas afirmações iniciais quanto a isto. Portanto, a proporção de falantes incluídos na pesquisa é irrelevante. O que é relevante para o cálculo é o grau de variabilidade do fenômeno lingüístico. Repetindo, quanto menor o grau de varia-

bilidade tanto menor a amostra e quanto maior o grau de variabilidade do fenômeno a ser estudado, tanto maior deverá ser a amostra. Mas, mesmo nestes casos, os padrões regulares que condicionam a variação lingüística atenuam a necessidade de uma amostra demasiadamente grande.

A este respeito, Labov observa: "... somos afortunados porque a padronização dentro da variação é fácil de se descobrir: ela não requer a análise estática de gravações de centenas de indivíduos como os lingüistas tradicionalmente receavam. Pelo contrário, descobrimos que os padrões básicos de estratificação de classe, por exemplo, emergem de amostras tão pequenas como 25 falantes... Ordenações regulares de estratificação social e estilística emergem mesmo quando nossas células contêm apenas cinco falantes e quando temos apenas cinco ou dez exemplos de uma dada variável para cada falante" (1972:204).

Este ponto de vista já havia sido expresso no Colóquio em Sociolingüística, organizado pelo Central Institute of Indian Languages em Mysore, India. Ao tratar do problema do tamanho da amostra, numa sociedade tão complexa quanto a indiana, Labov menciona: "Geralmente, tem-se descoberto que dados sociolingüísticos a respeito de uma variável são bastante confiáveis se há quatro ou cinco falantes em cada célula" (1972:38).

A veracidade desta afirmação tem sido confirmada por várias pesquisas sociolingüísticas. Assim, Guy (1977), ao estudar o cancelamento de t, d em inglês, utilizou uma amostra de 23 falantes. Os dados foram analisados, primeiro, individualmente para cada falante e, posteriormente, analisados para seus diversos agrupamentos. Guy prova que os padrões controlando a variação no comportamento do t d são os mesmos para os diversos indivíduos e grupos.

Resultados semelhantes foram obtidos por Shuy, Wolfram e Riley (1967). Estes três autores, analisando o inglês falado em Detroit, descobriram padrões extremamente regulares de estratificação social para diversas variáveis lingüísticas a partir da análise de 25 entrevistas.

Utilizando uma amostra de 23 falantes, Sankoff

(1980) analisou os condicionamentos que controlam o cancelamento do que no francês de Montreal, enquanto que Dittmar e associados estudaram a aquisição da língua alemã por imigrantes espanhóis e italianos em uma amostra de 48 falantes, distribuídos de acordo com sexo, origem e anos de permanência na Alemanha.

Naro (1981) mostrou que a estruturação do uso da concordância verbal no português do Brasil é válida para os diversos indivíduos que compõem o grupo. A única vantagem do grupo sobre o indivíduo é de ter maior número de dados, possibilitando assim maior precisão na análise. Este mesmo ponto de vista tem sido comprovado em diversas teses sobre o português brasileiro.

BIBLIOGRAFIA

GUY, Gregory. 1977. A New Look at t, d Deletion. In: FASOLD, R. W. & SHUY, R. W. ed. Studies in Language Variation. Washington, D.C., Georgetown University Press.

HEIDELBERGGER Forschungsprojekt "Pidgin-Deutsch". 1978. The Acquisition of German Syntax by Foreign Imigrant Workers. New York Academic Press.

LABOV, William. 1972. Sociolinguistic Patterns. Philadelphia, University of Pennsylvania Press.

_____. 1977. The Design of a Sociolinguistic Research Project. Report of the Sociolinguistic workshop, Chapter II, Central Institute of Indian Languages.

NARO, Anthony J. 1981. The Social and Structural Dimensions of a Syntactic Change. Language, 57: 63-98.

SANKOFF, G. 1980. Above and Beyond Phonology in Variable Rules. The Social Life of English. Philadelphia, University of Pennsylvania Press.

SHUY, R., WOLFRAM, W. & RILEY, W. K. 1967. A Study of Social Dialects in Detroit. Final Report, Project 6-1347. Washington D. C., Office of Education.

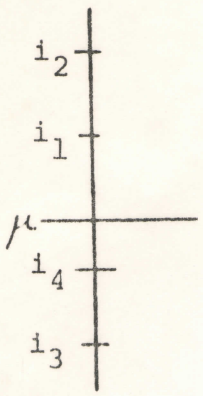
'QUANTOS FALANTES PRECISO?' ou
A Questão do Tamanho da Amostra

Maria Luísa Braga (UFF)
Anthony J. Naro (UFRJ)

I. μ -- média verdadeira a partir dos valores do índice de todos os elementos da amostra

m -- índice médio da amostra

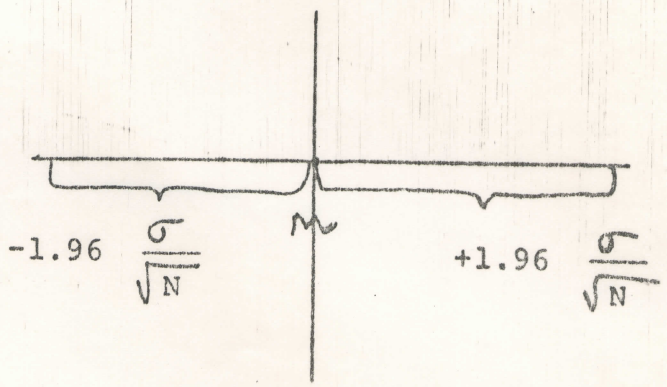
II. Desvio Padrão: σ



$$\sigma^2 = \frac{(\mu - i_1)^2 + (\mu - i_2)^2 + \ldots + (\mu - i_N)^2}{N}$$

III. Intervalo

μ está neste intervalo para 95 de 100 amostras de tamanho N



IV. Decisões

1. Qual é o grau de certeza que queremos ter que μ, a média verdadeira, estará no intervalo ao redor de m, a média empírica?
2. Qual é o tamanho do intervalo ao redor de m que consideramos satisfatória?
3. Qual é o grau de variabilidade do fenômeno em estudo, i.e., qual é o valor de σ?

V. Cálculo do tamanho da amostra

$$0.1 = 1.96 \frac{0.15}{\sqrt{N}} \qquad \sqrt{N} = \frac{1.96 \times 0.15}{0.1} = 2.94$$

$$N = 8.64$$