Malva, M. (2000). Quem foi que? - Um Desafio à Estatística: Questões de Autoria em "Novas Cartas Portuguesas". *Millenium, 19*

Quem foi que? - Um Desafio à Estatística: Questões de Autoria em "Novas Cartas Portuguesas"

MADALENA MALVA

Departamento de Matemática - Escola Superior de Tecnologia de Viseu

O livro "Novas Cartas Portuguesas" é constituído por uma série de poemas e cartas baseados na história da freira Mariana Alcoforado. Para o estudo efectuado foram escolhidos treze textos em prosa, tendo a escolha destes ficado a dever-se ao facto de ser necessário comparar textos de autoria "desconhecida" com textos de autoria conhecida, para deste modo se estabelecerem comparações e assim se chegar a alguma conclusão sobre os autores dos textos desconhecidos. Como das três autoras em causa apenas Maria Teresa Horta tem poesia publicada, tal facto levou-nos a supor que os poemas que figuram nas Novas Cartas são de sua autoria, eliminando-se assim do estudo os poemas. Na escolha dos textos a estudar também se teve em consideração a hipótese, por nós assumida, de que as três autoras se apresentaram e se despediram; como tal, foram incluídas no estudo as três "Primeiras Cartas" e as três "Cartas Últimas"; note-se que a "Primeira Carta Última" é constituída por três partes distintas que se supôs serem da mesma autora.

Dos restantes cinco textos estudados, quatro foram seleccionados aleatoriamente e a inclusão do quinto ficou a dever-se ao facto de ser um dos textos preferidos da autora do trabalho.

Depois de escolhidos os textos de autoria desconhecida a estudar, foi necessário construir os blocos de texto de autoria conhecida, com os quais se "definiu", estatisticamente, o estilo de cada autora. Os blocos de texto conhecido foram retirados dos livros "Os Outros Legítimos Superiores" de Maria Isabel Barreno, "Ambas as Mãos Sobre o Corpo" de Maria Teresa Horta e "Maina Mendes" de Maria Velho da Costa. A escolha destas obras ficou a dever-se ao facto de serem antecessoras das "Novas Cartas Portuguesas", mas a sua publicação (e escrita) não se ter dado muito antes das Novas Cartas, evitando deste modo que possíveis evoluções estilísticas tornassem os textos dificilmente cotejáveis. Assim, para a Maria Isabel Barreno e para a Maria Velho da Costa foram construídos aleatoriamente doze blocos de texto de mil e quinhentas palavras cada. Para a Maria Teresa Horta construíram-se, também aleatoriamente, onze blocos de texto de mil e quinhentas palavras cada; a construção de onze blocos ficou a dever-se ao facto de a obra utilizada para a Maria Teresa Horta ser "pequena", e não ter sido possível construir mais blocos.

Depois de definidos os textos conhecidos e os textos desconhecidos passou-se à sua análise.

Como seria de esperar começou-se por estudar a frequência das palavras nos textos, mas, nem todas as palavras que figuram num texto têm interesse. Palavras que dependam do contexto — palavras contextuais — não interessam; as palavras que devem ser utilizadas em estudos desta natureza são não contextuais. Depois de separadas as palavras contextuais das não contextuais para cada um dos textos conhecidos, teve que se seleccionar, de entre as palavras não contextuais encontradas, as que iriam ser objecto de um estudo mais aprofundado.

A cada palavra não contextual associou-se o terno (x, y, z) onde x, y e z indicam que determinada palavra ocorreu em x textos da Maria Isabel Barreno, y textos da Maria Teresa Horta e z textos da Maria Velho da Costa. Retiveram-se as palavras que apresentaram um score do tipo |x-y|>2, ou |x-z|>2 ou ainda |y-z|>2. Para cada uma das palavras seleccionadas calculou-se o índice

$$Z = \frac{(x-y)^2}{x+y} + \frac{(x-z)^2}{x+z} + \frac{(y-z)^2}{y+z}$$

como medida do poder discriminador dessa palavra e retiveram-se as palavras para as quais $z \geq 8$. Este processo conduziu-nos a uma lista trinta e uma palavras não contextuais. Para cada uma das palavras não contextuais seleccionadas, calculou-se a frequência absoluta em cada um dos textos conhecidos das autoras e a partir das frequências encontradas calculou-se o número de palavras esperadas em blocos de mil palavras, i.e., a sua permilagem. Para as amostras de permilagens calculou-se a sua média, mediana e desvio-padrão.

De seguida, foi-se investigar quais das palavras seleccionadas ocorriam nos textos desconhecidos e qual a sua frequência de utilização em cada texto. Na posse das frequências absolutas calculou-se a permilagem e comparou-se o valor das permilagens com as médias da palavra para cada autora; se o valor encontrado no texto desconhecido estivesse mais perto da média da Maria Isabel, por exemplo, então o texto em causa tinha fortes possibilidades de ser da Maria Isabel. Fazendo um estudo análogo ao descrito, para todas as palavras que ocorriam num determinado texto desconhecido, a autoria "final" foi atribuída à autora que mais vezes surgiu como a mais provável. Textos houve para os quais foi "fácil" fazer uma

atribuição de autoria, pois a maioria das palavras indicou uma determinada autora; no entanto, para alguns dos textos foi impossível chegar a uma conclusão, pois verificou-se um empate entre algumas autoras.

Efectuado o estudo anterior, ainda se estudaram mais algumas palavras. Por exemplo a palavra "certo", pois esta pode ser utilizada com dois significados — determinativo ou qualificativo; as palavras "pois" e "depois", dado que podem ser utilizadas no início ou no meio de frases; a frequência de utilização dos vocábulos "de um/dum" e "de uma/duma" que podem ser considerados sinónimos, e a frequência de utilização do vocábulo "não", que é dos vocábulos mais frequentes na língua portuguesa. O estudo dos vocábulos anteriores não foi muito proveitoso, uma vez que para a maioria deles deparamos com "falta de dados", pois ou não ocorriam nos textos desconhecidos ou a sua frequência de utilização era tão baixa que não nos pareceu legítimo tirar quaisquer conclusões. Nos casos para os quais nos atrevemos a tirar conclusões, estas, por vezes, vieram contradizer conclusões já obtidas anteriormente por estudo de outras variáveis.

De seguida, resolveu-se estudar o comprimento médio de frase. Para cada autora escolheram-se aleatoriamente cem frases dos textos conhecidos, e para as cem frases seleccionadas calculou-se a sua média, mediana e desvio-padrão. A primeira constatação efectuada foi a de que, para as três autoras, a média e a mediana eram substancialmente diferentes; as colecções recolhidas revelaram uma assimetria positiva forte. Face a estes resultados resolveu-se utilizar a mediana como medida de tendência central. Para cada um dos textos desconhecidos foi investigar-se o comprimento médio de frase, para, deste modo, se puder tirar mais algumas conclusões. Neste ponto resolveu-se recorrer às caixas-com-bigodes paralelas, para, à custa destas, efectuar algumas comparações, e assim chegar a algumas conclusões quanto à autoria dos textos. Se para alguns textos as conclusões a que se chegou foram de encontro às retidas à custa das palavras não contextuais, para a maioria tal não se verificou, e a possível autora do texto, segundo este critério, é outra. Também se utilizou o coeficiente de assimetria para estabelecer comparações, mas, mais uma vez, as conclusões a que se chegou foram de modo geral diferentes das já obtidas.

Outra das variáveis estudadas foi o comprimento de parágrafo. Para cada texto de autoria conhecida e desconhecida, verificou-se quantos parágrafos os constituíam e quantas palavras continha cada parágrafo. Para as amostras recolhidas efectuou-se um estudo, em tudo análogo ao efectuado para o comprimento de frase. Já sem grande surpresa verificou-se que esta variável, de modo geral, não reforçava a posição de nenhuma das escritoras como a possível autora de um ou mais textos, antes criava mais confusão.

Devido ao facto de os textos se encontrarem "contaminados" com sinais de pontuação, resolveu-se estudar quais os sinais de pontuação utilizados por cada autora e qual a sua frequência de utilização. Do

estudo efectuado eliminou-se o ponto final e a vírgula por serem, de algum modo, considerados sinais padrão da língua portuguesa. Assim, para cada sinal de pontuação utilizado nos textos conhecidos por cada autora, calculou-se a sua frequência absoluta e, a partir desta, a permilagem; para as amostras obtidas calcularam-se algumas medidas de tendência central e de dispersão. Constatou-se que a variabilidade de utilização dos sinais de pontuação é muito grande, pelo que se resolveu eliminar do estudo os sinais que apresentavam maior variabilidade (maior ou igual a dois). De seguida, foi investigar-se quais os sinais que ocorriam nos textos desconhecidos e qual a sua frequência; calculadas as permilagens para estes dados, mais uma vez se comparou o valor observado nos textos desconhecidos com as médias calculadas a partir dos textos conhecidos, estabelecendo-se que o valor observado provinha da população que tivesse valor médio mais próximo do valor observado. As conclusões foram "catastróficas"; com base nos dados obtidos a Maria Teresa Horta aparecia agora como a provável autora da maioria dos textos, o que nos pareceu absurdo. Saliente-se que, até este ponto, a referida autora apareceu sempre como a autora menos provável.

Numa derradeira tentativa de trazer alguma luz ao estudo, resolveu-se estudar ainda, o modo como as orações se encontravam ligadas nas frases, i.e., se as autoras utilizavam mais as conjunções coordenativas ou as conjunções subordinativas. Mais uma vez, para cada autora, recolheram-se aleatoriamente cem frases dos textos conhecidos, para as quais se estudou o modo como as orações se encontravam ligadas dentro de cada frase. Numa fase posterior do trabalho, efectuou-se o mesmo estudo para as frases dos textos desconhecidos, e, utilizando a técnica de comparar o valor médio, agora com a percentagem do valor observado, mais algumas conclusões foram retiradas. Como já se esperava, não foi esta variável que veio trazer a luz porque tanto ansiávamos.

Depois de estudadas as variáveis atrás apresentadas, algumas conclusões foram retiradas. Assim, analisando todas as conclusões parciais a que se chegou, e, ponderando em todas as hipóteses efectuadas, a "atribuição final de autoria" efectuada foi a seguinte:

Texto	**Possível autora**
Primeira Carta I	Maria Isabel
Primeira Carta II	Maria Teresa
Primeira Carta II	Maria Velho
Primeira Carta Última – Parte 1	Maria Velho
Primeira Carta Última – Parte 2	Maria Velho

Primeira Carta Última – Parte 3	Maria Velho
Segunda Carta Última	Maria Teresa
Terceira Carta Última	Maria Isabel
Paz	Maria Teresa
Lamento	Maria Velho
Monólogo	Maria Velho
Escriturário	Maria Isabel
Tarefas	Maria Isabel

Note-se que a Maria Isabel Barreno e a Maria Velho da Costa aparecem como as autoras mais prováveis enquanto, a Maria Teresa Horta aparece como a possível autora de apenas três textos.

Saliente-se que, o facto de uma variável apontar numa direcção e outra variável apontar noutra direcção, vem colocar a hipótese de que, provavelmente, uma das três autoras efectuou um trabalho de revisão dos textos; neste caso, a Maria Isabel Barreno a Maria Velho da Costa afiguram-se como as revisoras mais prováveis.