A ciência da pesquisa bibliográfica: uso de catálogos de bibliotecas para levantamentos sistemáticos e análise estatística de resultados

Roberto de Andrade Martins

Doutor em Lógica e Filosofia da Ciência (UNICAMP). Professor do Instituto de Física “Gleb Wataghin”, UNICAMP.

Endereço:

Grupo de História e Teoria da Ciência

Universidade Estadual de Campinas

Caixa Postal 6059

13081-970 Campinas, SP

Telefone:

0xx19-788-5516

Fax:

0xx19-788-5512

E-mail:

rmartins@ifi.unicamp.br

A ciência da pesquisa bibliográfica: uso de catálogos de bibliotecas para levantamentos sistemáticos e análise estatística de resultados

Resumo

Este artigo apresenta, através de modelos estatísticos, uma análise de levantamentos bibliográficos sistemáticos baseados no estudo de acervos de bibliotecas. A partir de um primeiro modelo simples em que se supõe que os acervos das bibliotecas foram constituídos de forma independente e aleatória, a partir de um universo homogêneo de livros, é deduzida uma forma de estimar o número total de livros existentes, a partir dos números encontrados em duas bibliotecas e do número de coincidências de obras nessas mesmas bibliotecas. Esse modelo é aplicado a um caso particular (número de obras publicadas em Portugal no século XVI) e são discutidas suas limitações. O modelo é em seguida aperfeiçoado introduzindo-se o conceito de probabilidades intrínsecas de aquisição variáveis para diferentes tipos de livros, provando-se que o primeiro modelo sempre proporciona uma estimativa inferior ao número de livros. A análise apresentada indica que é razoável supor-se que o número total real pode ser aproximadamente 30% a 40% superior ao valor indicado pelo modelo de coleção aleatória homogênea.

Palavras-chave

Pesquisa bibliográfica; Avaliação de levantamentos; Estatística; Bibliotecas.

The science of bibliographic research: the use of library catalogues in systematic surveys and the statistical analysis of results

Abstract

This paper applies statistical models to the analysis of systematic bibliographic surveys based upon the study of library collections. The first simplest model presented in the article supposes that library collections are constituted in an independent and random way, choosing from a homogeneous universe of books. This model leads to a formula to estimate the total number of existing books, from the numbers found in two libraries and the number of agreeing works found in those libraries. This model is applied to a particular case (the number of 16th century Portuguese publications) and its limitations are discussed. The model is then improved by the introduction of the concept of variable intrinsic acquisition probabilities for different kinds of books, and a proof is presented that the first model will always lead to an estimation that is lower than the real number of books. The analysis shows that it is reasonable to estimate that the real number is about 30% to 40% larger than the value computed according to the model of random and homogeneous collecting.

Keywords

Bibliographic research; Survey evaluation; Statistics; Libraries.

INTRODUÇÃO

Quando se executa uma pesquisa bibliográfica sistemática, algumas vezes há o objetivo de realizar um levantamento o mais completo possível daquilo que se publicou sobre certo assunto, ou em certo período, ou em certa região geográfica, etc. Em alguns casos muito especiais, esse ideal pode ser atingido – por exemplo, quando se está realizado um levantamento da produção bibliográfica de um único autor recente. No entanto, quase sempre é inatingível o sonho do bibliógrafo de incluir em seus levantamentos “tudo” o que existe sobre o tema que está estudando.

Talvez mais angustiante do que saber que está realizando um trabalho incompleto é a dúvida permanente que o bibliógrafo tem com relação ao grau de perfeição de seu levantamento. Depois de um longo tempo de procura de referências, terá ele atingido cerca de 90% das informações existentes? Ou apenas 50%, ou mesmo 10%? Existirá algum modo de avaliar aquilo que ainda não foi encontrado, quando se realiza um levantamento sistemático?

Em alguns casos, é possível fazer uma estimativa sobre aquilo que se desconhece. Este artigo apresentará um modelo simples para pesquisas bibliográficas realizadas a partir de comparação entre várias bibliotecas, mostrando sua aplicação em dois casos práticos, e discutindo suas limitações.

O MODELO DE COLEÇÃO ALEATÓRIA HOMOGÊNEA

Freqüentemente são lançadas no comércio certas coleções de figurinhas de vários tipos, destinadas ao público infanto-juvenil. Suponhamos que duas crianças estão colecionando figurinhas de um mesmo tipo, independentemente uma da outra, e que nenhuma delas sabe qual o número total de figurinhas daquele tipo. Supondo-se que não existem “figurinhas difíceis” (ou seja, supondo-se que a probabilidade de adquirir qualquer figurinha é igual) e supondo que o processo de aquisição é aleatório, é possível estimar o número de figuras que nenhuma das duas possui, comparando as coleções das duas crianças. De fato, suponhamos que a coleção da primeira criança contenha N_A figuras diferentes, e que a coleção da segunda criança contenha N_B figuras diferentes. Se, comparando-se as duas coleções, encontrarmos N_{AB} figuras comuns às duas coleções, poderemos fazer uma estimava N* de que o total de figuras daquele tipo é dado aproximadamente por

$$N^* = (N_A \times N_B)/N_{AB}$$

Se, por exemplo, a primeira criança possui 80 figurinhas diferentes e a segunda possui 120 figurinhas diferentes, e 60 das figurinhas são comuns às duas coleções, pode-se estimar que existem aproximadamente 80×120/60 = 160 figurinhas diferentes.

É fácil demonstrar essa fórmula. Se o número total (desconhecido) de figuras é N, e se a primeira coleção tem N_A figuras diferentes, então ela contém uma freqüência ou porcentagem $P_A = N_A/N$ das figuras existentes. A segunda

coleção, por sua vez, contém uma freqüência ou porcentagem $P_B = N_B/N$ das figuras existentes. Supondo-se que esses números são suficientemente grandes para aplicação das leis estatísticas, pode-se (de acordo com a "lei dos grandes números")[I] utilizar essas porcentagens P como valores representativos da probabilidade P' de que uma dada figura se encontre em cada uma das coleções, ou seja: $P'_A \cong P_A$ e $P'_B \cong P_B$, onde o símbolo $\cong$ (aproximadamente igual) será utilizado para relacionar as freqüências medidas (dados experimentais) com as probabilidades.

Admitindo-se a independência entre as duas coleções, então a probabilidade P'_{AB} de que uma mesma figura seja encontrada ao mesmo tempo nas duas coleções é

$$P'_{AB} = P'_A \times P'_B$$

O número efetivamente encontrado de figuras comuns às duas coleções é N_{AB} e portanto há uma freqüência ou porcentagem $P_{AB} = N_{AB}/N$ de figuras comuns às duas coleções. Admitindo-se que essa porcentagem é um valor representativo da probabilidade P'_{AB} de encontrar figuras comuns às duas coleções, temos:

$$N_{AB}/N = P_{AB} \cong P'_{AB} = P'_A \times P'_B \cong P_A \times P_B = (N_A/N) \times (N_B/N)$$

e, portanto,

$$N_{AB}/N \cong (N_A \times N_B)/N^2$$

de onde deduzimos o resultado final:

$$N \cong (N_A \times N_B)/N_{AB} = N^*$$

Se as freqüências fossem rigorosamente idênticas às probabilidades, essa fórmula proporcionaria o valor exato do número total de figurinhas. Na prática, como os números observados são finitos, as freqüências não são idênticas às probabilidades, e por isso, em vez de obtermos o número real de figurinhas, obtemos uma estimativa N* desse número.

Para que essa fórmula possa ser utilizada na prática, é importante que os números N_A, N_B e N_{AB} sejam relativamente elevados. A incerteza relativa dos resultados obtidos, devida a flutuações estatísticas para pequenos números, é medida pelo coeficiente de variação, que é dado por $(1/2n)^{1/2}$ (Green e Morgerison, 1978:84). Para números da ordem de 50, por exemplo, o coeficiente de variação é de 10%, e para números da ordem de 200 o coeficiente é de 5%.

Além disso, o modelo exige que esse tipo de figuras constitua um conjunto homogêneo, ou seja, todas os itens tenham igual probabilidade de aquisição (não podem existir "figurinhas difíceis"), e que as duas coleções tenham sido

---

[I] Ver, por exemplo, Moran (1968:56-57) e Gnedenko (1969:199-222).

formadas independentemente. Um contra exemplo simples seria o caso em que uma criança tivesse formado sua coleção com as figurinhas repetidas do seu irmão. Nesse caso, é violada a condição de independência das coleções.

APLICAÇÃO DO MODELO À PESQUISA BIBLIOGRÁFICA

Suponhamos que um bibliógrafo está tentando fazer um levantamento sistemático dos livros que foram publicados em Portugal durante o século XVI. Esse pesquisador poderia visitar várias bibliotecas portuguesas importantes, e verificar em cada uma delas quais as obras desse período e dessa delimitação geográfica que encontra. À medida que essa busca for se desenvolvendo, o pesquisador encontrará algumas obras repetidas, e evidentemente o número de repetições tenderá a crescer à medida que sua pesquisa vá se tornando cada vez mais completa.

Em 1926, o bibliotecário António Joaquim Anselmo, da Biblioteca Nacional de Lisboa, procurou efetivamente realizar esse tipo de levantamento, tendo publicado em 1926 o resultado de sua pesquisa (Anselmo, 1926). O catálogo de Anselmo apresenta não apenas as referências bibliográficas das obras descritas, mas também indica as bibliotecas em que foram encontrados exemplares das mesmas. Utilizando esse trabalho, podemos verificar a possibilidade de aplicação do modelo de coleção aleatória homogênea à análise da pesquisa bibliográfica sistemática realizada pela consulta a catálogos de bibliotecas.

Consideremos algumas das bibliotecas portuguesas importantes incluídas no levantamento realizado por Anselmo: a Biblioteca Nacional de Lisboa, a Biblioteca da Ajuda, a Biblioteca Pública de Évora e a Biblioteca Pública do Porto. Foram escolhidas estas bibliotecas para a presente análise por serem aquelas onde foram localizados os maiores número de publicações portuguesas do século XVI. Abaixo estão indicados os números de obras encontradas por Anselmo em cada uma delas:

- BNL – Biblioteca Nacional de Lisboa: 670
- BA – Biblioteca da Ajuda: 262
- BE – Biblioteca Pública de Évora: 274
- BP – Biblioteca Pública do Porto: 184

Suponhamos que, após examinar os acervos dessas bibliotecas, o bibliógrafo quisesse avaliar o grau de perfeição do seu levantamento da bibliografia portuguesa do século XVI. Ele poderia aplicar o modelo de coleção aleatória homogênea, examinando primeiramente as obras comuns a cada par de bibliotecas. A partir da obra de Anselmo, verificamos os seguintes números de obras comuns às bibliotecas aqui estudadas:

BNL e BA: 205

BNL e BE: 168

BNL e BP: 135

BA e BE: 91

BA e BP: 59

BE e BP: 62

Aplicando-se a fórmula apresentada anteriormente a cada um desses casos, obtemos as estimativas para o número total (desconhecido) de obras publicadas em Portugal no século XVI mostradas na Tabela 1.

INSERIR AQUI A TABELA 1

Os resultados das estimativas (N*, na última coluna) variam bastante, como esperado, pois os números utilizados nos cálculos são pequenos, gerando grandes flutuações relativas (da ordem de 10%, como esperado).

A média simples dos vários resultados para N* é de 880, porém os resultados das primeiras linhas da tabela são mais confiáveis do que os das linhas inferiores, podendo ser estimados pesos dos diversos valores obtidos. Esses pesos W são calculados a partir da fórmula abaixo (Young, 1962:108).

$$W = 1 / (1/N_A + 1/N_B + 1/N_{AB})^{1/2}$$

A média dos valores de N*, utilizando-se esse tipo de ponderação, é de 998, com um desvio padrão igual a 68 (também calculado com o uso de ponderação dos dados). A partir desses valores, portanto, a aplicação do modelo de coleção aleatória homogênea nos levaria a esperar que existissem aproximadamente 1.000 obras publicadas em Portugal no século XVI.

DISCUSSÃO DO RESULTADO OBTIDO

Será este um bom resultado? Neste caso em particular, podemos comparar o valor obtido através do modelo de coleção aleatória homogênea com os resultados finais da pesquisa bibliográfica do próprio Anselmo. Em seu catálogo, ele indicou um total de 1.312 publicações portuguesas do século XVI – incluindo tanto as que ele de fato foi capaz de encontrar nas diversas bibliotecas, como também as que eram mencionadas por alguma obra de referência que utilizou. Algumas dessas 1.312 edições talvez não existam (podem ter ocorrido erros dos bibliógrafos anteriores) mas, por outro lado, após a publicação do trabalho de Anselmo foram descobertas outras obras do século XVI que ele não havia citado. Podemos afirmar com segurança, portanto, que o número de obras portuguesas do século XVI efetivamente existentes é bastante superior (30 a 40%) ao número que foi estimado empregando-se o modelo de coleção aleatória homogênea.

Essa diferença não pode ser explicada como resultado de uma mera flutuação estatística, pois é muito superior ao que esperaríamos, já que o desvio padrão da amostra utilizada corresponde a apenas 7% e portanto não esperaríamos

uma diferença superior a 20% entre o valor real e a média calculada[II]. Essa discrepância deve-se, provavelmente, ao fato de que a situação estudada não satisfaz o conjunto de hipóteses do modelo utilizado.

Com foi indicado acima, as duas hipóteses importantes do modelo de coleção aleatória homogênea são:

1. As coleções devem ter sido formadas independentemente.

2. Todos os itens devem ter igual probabilidade de aquisição (homogeneidade).

É plausível que a primeira condição tenha sido satisfeita na formação das bibliotecas estudadas. Não há motivo para suspeitar que a aquisição de uma determinada obra por uma biblioteca fosse de algum outro modo influenciada pela sua existência ou inexistência nas demais bibliotecas estudadas. É bem verdade que, se pensarmos que houvesse uma competição entre os responsáveis por essas bibliotecas, cada um querendo que sua coleção fosse melhor do que o das demais bibliotecas, poderia ter havido uma aquisição dirigida pela intenção de possuir, sempre que possível, as obras raras de que as demais bibliotecas também dispusessem. Nesse caso, o resultado seria um aumento do número de obras iguais nas várias bibliotecas, violando a hipótese de independência. Mas é pouco provável que tenha ocorrido um processo desse tipo.

Em outros casos, poderia ocorrer que a primeira condição não fosse válida por motivos geográficos. Consideremos, por exemplo, que esteja sendo feito um levantamento de obras publicadas em português no século XIX, e que sejam comparadas duas bibliotecas, uma de Portugal e outra do Brasil. Esperaríamos que a biblioteca portuguesa tivesse uma maior porcentagem de obras publicadas em Portugal, e que a biblioteca brasileira tivesse uma maior porcentagem de livros publicados no Brasil. No caso que estamos discutindo aqui, no entanto, não parece haver nenhum efeito geográfico importante.

A segunda condição (de igual probabilidade), no entanto, provavelmente não se aplica nesse tipo de situação. As diferentes obras portuguesas do século XVI foram publicadas em tiragens diferentes; algumas circularam de forma mais ampla pelo país, outras de forma mais restrita; algumas eram mais caras, outras mais baratas; algumas eram consideradas mais importantes e valiosas (sob o ponto de vista bibliográfico), outras eram consideradas de pequeno valor. Todos esses fatores devem ter influenciado a probabilidade de aquisição de exemplares pelas bibliotecas.

Consideremos, por exemplo, o folheto *Compremisso* [sic] *da Irmandade da Casa da Sancta Misericordia da Cidade de Lisboa*, descrito por Anselmo

---

[II] No caso de uma distribuição normal (gaussiana) de erros, pode-se ter um grau de confiança de que em 95% dos casos os valores medidos encontram-se dentro de uma faixa de ±2 desvios padrões (nesse caso, 14%) e que em 99,7% dos casos encontram-se dentro de uma faixa de ±3 desvios padrões (nesse caso, 21%). Ver Young (1962:75).

(1926:14) apenas através de informação indireta de uma única fonte. Por causa do tema deste folheto, é plausível que ele tenha sido impresso em pequeno número de exemplares e que não tenha circulado de forma ampla, não despertando também grande interesse por sua conservação. Outro exemplo é a *Oratio de scientiarum omnium magnatumquem artium laude*, um discurso acadêmico pronunciado na Universidade de Coimbra por Antonio Pinto em 1555, citado por Anselmo (1926:19). Esperaríamos encontrar um exemplar desse folheto na própria Universidade de Coimbra, como de fato se encontra, mas trata-se de outro exemplo de publicação que deve ter tido pequena tiragem e fraca circulação, dificultando sua aquisição por outras bibliotecas. Por outro lado, uma outra oração acadêmica também apresentada no ano anterior em Coimbra, *Oratio de scientiarum, disciplinarumque omnium laudibus habita Conimbricae*, por Henrique de Brito, não foi conservada na biblioteca daquela Universidade nem em qualquer outra biblioteca consultada por Anselmo (1926:314), sendo conhecida apenas indiretamente.

Portanto, o modelo de coleção aleatória homogênea é apenas uma primeira aproximação da realidade, quando aplicado à aquisição de obras pelas bibliotecas.

SEGUNDA APROXIMAÇÃO: PROBABILIDADE INTRÍNSECA VARIÁVEL

Podemos formar um modelo um pouco melhor para descrever a formação de coleções por bibliotecas supondo a existência de dois ou mais grupos de obras que possuem probabilidades intrínsecas diferentes de serem adquiridas pelas bibliotecas. Essas probabilidades intrínsecas dependerão dos fatores indicados anteriormente (número de exemplares publicados, circulação, preço, etc.).

Suponhamos inicialmente a existência de dois grupos de obras: o grupo 1, com um total de N_1 obras, e o grupo 2, com um total de N_2 obras, formando um conjunto total de $N=N_1+N_2$ obras. Suponhamos que as obras do primeiro grupo têm uma probabilidade intrínseca P_1 de serem adquiridas por qualquer biblioteca, e que as do segundo grupo possuem a probabilidade intrínseca de aquisição P_2. Suponhamos que as coleções de livros de duas bibliotecas A e B são formadas aleatoriamente a partir desses dois grupos de obras, de tal modo que a probabilidade de aquisição de obras de cada grupo seria:

- αP_1 no caso de livros do grupo 1 adquiridos pela biblioteca A
- αP_2 no caso de livros do grupo 2 adquiridos pela biblioteca A
- βP_1 no caso de livros do grupo 1 adquiridos pela biblioteca B
- βP_2 no caso de livros do grupo 2 adquiridos pela biblioteca B

Considerando a lei dos grandes números, e considerando assim que as freqüências observadas são próximas às probabilidades em cada caso, os números de obras de cada tipo, adquiridas pelas duas bibliotecas, seriam:

- $N_{A1} \cong \alpha P_1.N_1$ (livros do grupo 1 adquiridos pela biblioteca A)

- $N_{A2} \cong \alpha P_2.N_2$ (livros do grupo 2 adquiridos pela biblioteca A)
- $N_{B1} \cong \beta P_1.N_1$ (livros do grupo 1 adquiridos pela biblioteca B)
- $N_{B2} \cong \beta P_2.N_2$ (livros do grupo 2 adquiridos pela biblioteca B)

e, portanto, o número de obras de cada biblioteca seria:

$$N_A = N_{A1} + N_{A2} \cong \alpha(P_1.N_1 + P_2.N_2)$$

$$N_B = N_{B1} + N_{B2} \cong \beta(P_1.N_1 + P_2.N_2)$$

O número de obras comuns às duas bibliotecas, para cada tipo de livro, seria:

- $N_{AB1} \cong \alpha P_1.\ \beta P_1.N_1$ (livros do grupo 1 comuns às duas bibliotecas)
- $N_{AB2} \cong \alpha P_2.\ \beta P_2.N_2$ (livros do grupo 2 comuns às duas bibliotecas)

O número total de livros comuns às duas bibliotecas seria, portanto:

$$N_{AB} = N_{AB1} + N_{AB2} \cong \alpha\beta(P_1^2.N_1 + P_2^2.N_2)$$

Se tentássemos calcular o número total de obras existentes (grupo 1 + grupo 2) a partir das informações sobre os livros encontrados nas duas bibliotecas, sem levar em conta a existência de diferentes probabilidades de aquisições intrínsecas, obteríamos o valor N*, de acordo com a fórmula apresentada anteriormente:

$$N^* = (N_A \times N_B)/N_{AB}$$

Substituindo os valores indicados acima e fazendo algumas simplificações, teremos:

$$N^* \cong (P_1.N_1 + P_2.N_2)^2/(P_1^2.N_1 + P_2^2.N_2)$$

O número total real de obras, no entanto, é $N=N_1+N_2$, e podemos provar que esse número real é sempre superior ao valor estimado N*. De fato, calculemos o seguinte fator:

$$\varepsilon = (N–N^*) / N^*$$

Substituindo os valores de N e de N* e fazendo algumas simplificações, obtemos:

$$\varepsilon = (N - N^*) / N^* \cong (P_1–P_2)^2.N_1.N_2 / (P_1.N_1+P_2.N_2)^2$$

É evidente que esse fator é sempre positivo. Portanto, como (N–N*) / N* é positivo, e como N* é positivo, a diferença N–N* é positiva, ou seja, N é sempre maior do que N*. Assim, a estimativa N* obtida pela fórmula do modelo de coleção aleatória homogênea irá proporcionar um resultado inferior ao número real de obras existentes.

O fator $\varepsilon = (N–N^*) / N^*$ indica o erro percentual do modelo de coleção aleatória homogênea, que não leva em conta a existência de diferentes probabilidades intrínsecas de aquisição das obras. Se representarmos P_1/P_2 por ρ e N_1/N_2 por ν, esse fator poderá também ser calculado por:

$$\varepsilon = (N – N^*) / N^* \cong \nu.[(\rho–1)/(\nu\rho+1)]^2$$

Suponhamos, por exemplo, os seguintes valores: $P_1 = 4.P_2$ e $N_1 = N_2$. Utilizando a fórmula acima, estima-se nesse caso que $\varepsilon = 0{,}36 = 36\%$.

A Tabela 2 indica os valores do erro percentual ε para diversos valores de $\rho=P_1/P_2$ e de $\nu=N_1/N_2$.

INSERIR AQUI A TABELA 2

Portanto, vê-se que o erro relativo no cálculo do número total de livros aumenta muito quando uma pequena fração dos livros (ν pequeno) tem uma grande probabilidade intrínseca de aquisição (ρ grande). O menor erro relativo ocorre quando as probabilidades intrínsecas não são muito diferentes (ρ próximo de 1) e quando os livros com maior probabilidade intrínseca de compra são os mais numerosos (ν grande).

DISTRIBUIÇÕES DE PROBABILIDADES INTRÍNSECAS

Podemos generalizar a análise anterior, que foi feita para apenas dois tipos de obras com diferentes probabilidades intrínsecas. Basta considerar a existência de n grupos de obras, cada grupo com uma probabilidade intrínseca diferente de ser adquirida. Nesse caso, as fórmulas anteriores se tornam:

- $N_{Ai} \cong \alpha P_i.N_i$ (livros do grupo i adquiridos pela biblioteca A)
- $N_{Bi} \cong \beta P_i.N_i$ (livros do grupo i adquiridos pela biblioteca B)

$$N_A = \sum N_{Ai} \cong \alpha.\sum P_i.N_i$$

$$N_B = \sum N_{Bi} \cong \beta.\sum P_i.N_i$$

O número de obras comuns às duas bibliotecas, para cada tipo, será:

$$N_{ABi} \cong \alpha\beta.P_i^2.N_i$$

e, portanto, o número total de obras comuns às duas bibliotecas será:

$$N_{AB} = \sum N_{ABi} \cong \alpha\beta.\sum P_i^2.N_i$$

Portanto, a estimativa N^* do número total de obras, utilizando o modelo de coleção aleatória, seria:

$$N^* = (N_A \times N_B)/N_{AB} \cong (\sum P_i.N_i)^2 / \sum P_i^2.N_i$$

O número real total de obras é simplesmente $N = \sum N_i$ e, portanto,

$$N^*/N \cong (\sum P_i.N_i)^2 / [(\sum N_i).(\sum P_i^2.N_i)]$$

Escolhendo-se uma distribuição de probabilidades intrínsecas, é possível assim calcular o erro percentual ε de aplicação da fórmula do modelo de coleção aleatória simples. Por exemplo: se considerarmos 5 grupos de livros, cada um com igual número W de obras, e com probabilidades intrínsecas de k, 2k, 3k, 4k e 5k, obteremos $N^*/N = 9/11$, e portanto $\varepsilon = (N - N^*) / N^* \cong 2/9 = 0,22 = 22\%$.

Essa análise pode ser facilmente estendida a uma distribuição contínua. Suponhamos que existam obras com diferentes probabilidades intrínsecas de aquisição, dependendo de certa característica x da obra, com uma densidade de probabilidade[III] p dada por

$$p = dP/dx = g(x)$$

Por simplicidade, suponhamos que essa relação é biunívoca, e que sabendo-se a densidade de probabilidade pode-se determinar o característica da obra, ou seja, existe uma função inversa à função g(x), de modo que x é uma função de p:

$$x = h(p)$$

Suponhamos, além disso, que a densidade de obras existentes com característica x dependa do valor de x, ou seja:

$$dn/dx = q(x)$$

Como a característica x é uma função da densidade de probabilidade p, existirá também uma relação do tipo

$$dn/dp = f(p)$$

É fácil mostrar que, nesse caso, a fórmula anteriormente apresentada para distribuição descontínua de probabilidades intrínsecas se transforma em:

$$N^*/N \cong [\int p.f(p).dp]^2 / \{[\int f(p).dp].[\int p^2.f(p).dp]\}$$

Seria difícil justificar qualquer modelo mais detalhado de distribuição de probabilidades intrínsecas. Vamos, no entanto, considerar três modelos, para efeito de comparação, descritos abaixo.

O primeiro caso seria uma situação em que o número de obras existentes é proporcionalmente maior para as obras que possuem maior probabilidade intrínseca de aquisição, com uma distribuição do tipo $dn/dp = f(p) = k.p$ que pode ser representada pelo gráfico da Fig. 1a.

INSERIR AQUI A FIGURA 1

[III] Green & Margerison (1978:12-13).

Um segundo caso seria o inverso deste, ou seja, uma situação em que as obras mais numerosas são as que possuem menor probabilidade intrínseca de aquisição, enquanto as que possuem maior probabilidade intrínseca são poucas, com uma distribuição do tipo dn/dp = f(p) = k.(p”–p) que pode ser representado pelo gráfico da Fig. 1b.

Por fim, um terceiro caso, intermediário, seria aquele em que houvesse iguais porcentagens de obras com grande ou pequena probabilidade intrínseca de aquisição, ou seja, dn/dp = f(p) = k que pode ser representado pelo gráfico da fig. 1c.

Efetuando-se os cálculos, os resultados obtidos são os seguintes: no primeiro caso, $\varepsilon = 1/8 = 12{,}5\%$, ou seja, quando os livros mais numerosos são os que possuem maior probabilidade intrínseca de serem adquiridos, a estimativa N* do número total de obras, utilizando o modelo de coleção aleatória homogênea, não é muito inferior ao número real N de obras existentes.

No segundo caso, $\varepsilon = 1/2 = 50\%$, ou seja, quando os livros mais numerosos são os que possuem menor probabilidade intrínseca de serem adquiridos, a estimativa N* do número total de obras, utilizando o modelo de coleção aleatória homogênea, é muito inferior ao número real N de obras existentes.

Por fim, no terceiro caso, $\varepsilon = 1/3 = 33\%$, ou seja, quando há iguais porcentagens de obras com grande ou pequena probabilidade intrínseca de aquisição, a estimativa N* do número total de obras, utilizando o modelo de coleção aleatória homogênea, é bastante inferior ao número real N de obras existentes.

No caso dos três modelos acima discutidos, nota-se que o valor real de obras existentes é sempre superior ao número N* estimado pelo modelo de coleção aleatória homogênea. Poderíamos criar situações hipotéticas em que N fosse dez ou cem vezes maior do que N*, mas isso não ocorreu em nenhum dos casos estudados. Nos três exemplos, N variou entre 9/8 de N* (primeiro caso) e 2.N* (segundo caso). Como esses são casos bastante extremos, é plausível que as situações reais sejam normalmente intermediárias entre essas, correspondendo aproximadamente ao terceiro caso, ou seja, que o número total N seja normalmente entre 30 e 40% superior à estimativa N*.

No caso dos livros portugueses do século XVI, vimos que realmente o número total de obras era cerca de 30 a 40% superior à estimativa do modelo de coleção aleatória homogênea, concordando muito bem com o modelo de probabilidade intrínseca variável apresentado acima.

SEGUNDO EXEMPLO PRÁTICO

Vejamos um outro exemplo de pesquisa bibliográfica baseado em coleções de bibliotecas. Trata-se de um levantamento das diferentes edições, traduções e comentários sobre o famoso *Tratado da esfera* de Johannes de Sacrobosco, publicadas do século XV ao século XVII. Essa obra foi possivelmente o texto astronômico mais popular de toda a história, tendo sido utilizado em universidades de toda a Europa, principalmente em latim, mas também em

traduções para o francês, italiano, alemão, espanhol e português (Thorndike, 1949).

Procurando fazer um levantamento sistemático das edições da obra de Sacrobosco, foi realizada uma pesquisa utilizando bases de dados de algumas bibliotecas que estão disponíveis na Internet. As bibliotecas consultadas foram:

- LC = Library of Congress
- YU = Yale University
- NY = New York Public Library
- HU = Harvard University
- BL = British Library

Além dessas bibliotecas, foi também utilizada uma obra de referência bibliográfica: a *Bibliographie astronomique* de Lalande (1803), que será indicada pela sigla LL. Embora não se trate nesse caso de um catálogo de uma biblioteca, essa obra foi utilizada pelo grande número de edições que cita, e por se tratar de uma fonte de informação baseada principalmente no estudo realizado por Lalande em bibliotecas francesas e européias, podendo ser considerado independente das informações sobre as edições existentes nas bibliotecas acima referidas.

A Tabela 3, semelhante à utilizada para analisar as obras portuguesas do século XVI, indica o número de edições do *Tratado* de Sacrobosco encontrado em cada biblioteca, bem como o número de coincidências e as estimativas do número total de edições existentes, utilizando o modelo simplificado.

INSERIR AQUI A TABELA 3

A média simples dos valores estimados N* é 191. A média ponderada proporciona um valor quase igual: 192. Assim, de acordo com o modelo de coleção aleatória homogênea, a previsão seria de que deveriam existir cerca de 190 edições diferentes da obra de Sacrobosco. Portanto, de acordo com a análise acima apresentada, poderíamos esperar que o número total real de edições da obra de Sacrobosco seria cerca de 30 a 40% superior a esse número, ou seja, aproximadamente 250 a 270.

Utilizando informações não apenas dos acervos dessas bibliotecas, mas de várias outras, assim como diversas fontes de informação bibliográfica, a pesquisa realizada permitiu identificar, até o momento, 232 edições diferentes da obra de Sacrobosco, ou seja, um valor 20% superior a N*. Como esse levantamento não pode ser considerado completo, é razoável supor-se que existam ainda mais 20-40 edições dessa obra que não foram identificadas, de acordo com a previsão acima.

APLICAÇÕES

O tipo de análise aqui apresentado pode ser utilizado ao se fazer pesquisas bibliográficas de vários tipos, a partir de informações sobre os acervos encontrados em bibliotecas. Pode-se procurar determinar, por exemplo, o número de diferentes edições da Bíblia publicadas no século XV, a partir de uma análise desse tipo, analisando-se as edições encontradas nas maiores bibliotecas do mundo. Pode-se tentar verificar o número de obras sobre medicina publicadas em um país, em certo século, fazendo-se um levantamento dos livros encontrados nas principais bibliotecas daquele país.

É necessário sempre tomar o cuidado de procurar analisar tipos de obras que, por sua natureza, apresentem uma certa homogeneidade, e escolher bibliotecas tais que não haja motivos *a priori* para suspeitar que elas privilegiariam fortemente a aquisição de apenas algumas das obras desse tipo, e não outras. Não é conveniente, por exemplo, tentar fazer uma análise do número de obras publicadas na Inglaterra do século XV ao século XX empregando esse método, pois trata-se de um objeto de estudo altamente heterogêneo. É preferível dividir o estudo por séculos (ou, no caso de períodos recentes, em décadas) e fazer as estimativas para cada época separadamente.

Também não é conveniente estudar o número de obras portuguesas do século XVII comparando uma biblioteca de uma faculdade de medicina com uma biblioteca de uma faculdade de direito, pois cada uma delas tenderá a privilegiar a aquisição de livros relacionados à sua especialidade. Bibliotecas públicas (que possuem acervos genéricos) são mais adequadas, para esse tipo de estudo. Por outro lado, para fazer um levantamento das obras *médicas* portuguesas do século XVII, pode-se utilizar tanto bibliotecas especializadas como bibliotecas gerais.

Outra condição prática importante é aplicar esse modelo apenas quando se percebe haver um número suficiente de obras que permita uma análise estatística “razoável”. Quando se lida com números em torno de 50, como já foi indicado, pode-se esperar flutuações da ordem de 10%. Quando a quantidade de obras (e de coincidências encontradas em pares de bibliotecas) é superior a 50, e quando são utilizados vários pares de bibliotecas para comparação, os resultados serão bons. Quando se utilizam poucas bibliotecas e os números envolvidos nos cálculos são inferiores a 50, os resultados são inadequados.

CONCLUSÃO

Através de modelos simplificados de formação de coleções, pode-se fazer estimativas sobre o grau de perfeição de levantamentos bibliográficos baseados no estudo dos acervos de bibliotecas. O modelo mais simples aqui discutido é aquele em que se supõe que os acervos das bibliotecas foram formados de forma independente e aleatória, a partir de um universo homogêneo de livros, todos com a mesma probabilidade de serem adquiridos. Esse modelo conduz a uma forma de estimativa do número total de livros existentes, a partir dos números encontrados em duas bibliotecas e do número de coincidências de obras nessas mesmas bibliotecas.

No entanto, a hipótese de que todas as obras possuem a mesma probabilidade de serem adquiridas pelas bibliotecas geralmente não pode ser aceita em estudos bibliográficos. Introduzindo-se o conceito de probabilidades intrínsecas variáveis para diferentes grupos de livros, foi provado que a estimativa do primeiro modelo é sempre inferior ao número real de livros.

Portanto, em pesquisas bibliográficas sistemáticas que se baseiem no estudo de acervos de bibliotecas, pode-se utilizar como primeira aproximação o modelo de coleção aleatória homogênea, tomando o cuidado de utilizar várias bibliotecas para comparação, e não apenas duas, de modo a obter uma estatística mais confiável. Porém, mesmo tomando esse cuidado, deve-se ter em mente que essa estimativa será normalmente inferior ao valor real total de obras. Sem a escolha de um modelo detalhado de distribuição de probabilidades intrínsecas é impossível chegar-se a uma avaliação mais exata, porém a análise apresentada indica que é razoável supor-se que o número total real pode ser aproximadamente 30% a 40% superior ao valor indicado pelo modelo de coleção aleatória homogênea.

AGRADECIMENTO

O autor agradece à FAPESP (Fundação de Amparo à Pesquisa do Estado de São Paulo) e ao CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) o apoio recebido, que permitiu a realização da presente pesquisa.

REFERÊNCIAS BIBLIOGRÁFICAS

1. ANSELMO, António Joaquim. *Bibliografia das obras impressas em Portugal no século XVI*. Lisboa: Biblioteca Nacional, 1926.

2. GNEDENKO, B. V. *The theory of probability*. Moscow: Mir, 1969.

3. GREEN, J. R. & MARGERISON, D. *Statistical treatment of experimental data*. Amsterdam: Elsevier, 1978

4. LALANDE, Jérôme de. *Bibliographie astronomique*. Paris: Imprimérie de la République, 1803.

5. MORAN, P. A. P. *An introduction to probability theory*. Oxford: Clarendon Press, 1968.

6. THORNDIKE, Lynn. *The Sphere of Sacrobosco and its commentators.* Chicago: University of Chicago Press, 1949.

7. YOUNG, Hugh D. *Statistical treatment of experimental data*. New York: McGraw Hill, 1962.

TABELA 1

Número de publicações portuguesas do século XVI encontradas em algumas bibliotecas, com a indicação do número N_{AB} de obras comuns a cada par de bibliotecas e estimativa de número total de livros N* existentes desse tipo, utilizando-se o modelo mais simples

Biblioteca A	N_A	Biblioteca B	N_B	N_{AB}	$N^*=N_A \times N_B / N_{AB}$
BNL	670	BA	262	205	856
BNL	670	BE	274	168	1092
BNL	670	BP	184	135	913
BE	274	BA	262	91	788
BE	274	BP	184	62	813
BA	262	BP	184	59	817

BNL = Biblioteca Nacional de Lisboa; BA = Biblioteca da Ajuda; BE = Biblioteca Pública de Évora; BP = Biblioteca Pública do Porto

TABELA 2

Valores do erro relativo $e = (N-N^*)/N^* \approx n.[(r-1)/(nr+1)]^2$ da estimativa que utiliza o modelo de coleção aleatória homogênea, supondo-se a existência de dois tipos de obras com diferentes probabilidades aquisição, onde $r = P_1/P_2$ é a razão entre as probabilidades intrínsecas de aquisição e $n = N_1/N_2$ é a razão entre o número de obras de cada tipo.

	ρ=2	ρ=3	ρ=4	ρ=5
ν=1/3	12%	33%	55%	75%
ν=1/2	12%	32%	50%	65%
ν=1	11%	25%	36%	44%
ν=2	8%	16%	22%	26%
ν=3	6%	12%	16%	19%

TABELA 3

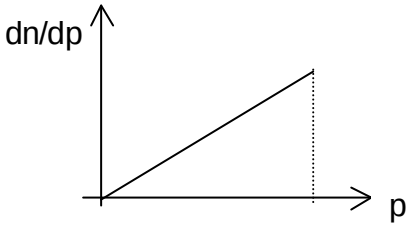
Número de edições e comentários do *Tratado da Esfera* de Sacrobosco encontrados em diversas bibliotecas, com a indicação do número N_{AB} de obras comuns a cada par de bibliotecas e estimativa de número total de livros N* existentes desse tipo, utilizando-se o modelo simplificado

Biblioteca A	N_A	Biblioteca B	N_B	N_{AB}	$N^*=N_A \times N_B / N_{AB}$
LC	55	YU	45	20	124
LC	55	NY	122	37	181
LC	55	HU	42	14	165
LC	55	BL	63	18	193
LC	55	LL	81	21	212
YU	45	NY	122	35	157
YU	45	HU	42	14	135
YU	45	BL	63	16	177
YU	45	LL	81	20	182
NY	122	HU	42	18	285
NY	122	BL	63	38	202
NY	122	LL	81	45	220
HU	42	BL	63	16	165
HU	42	LL	81	14	243
BL	63	LL	81	23	222

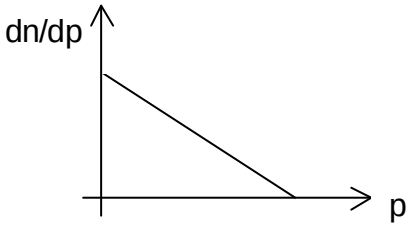
LC = Library of Congress; YU = Yale University; NY = New York Public Library; HU = Harvard University; BL = British Library; LL = *Bibliographie astronomique* (Lalande, 1803).

FIGURA 1

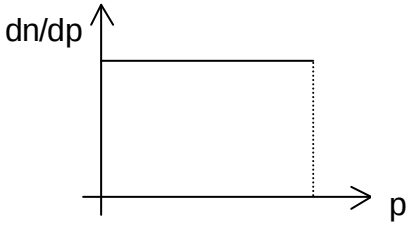
Três modelos lineares de distribuição de densidade de probabilidade intrínseca: (a) quando é maior o número de obras com maior probabilidade de aquisição; (b) quando as obras mais numerosas são as que possuem menor probabilidade intrínseca de aquisição; e (c) quando há iguais porcentagens de obras com grande ou pequena probabilidade intrínseca de aquisição



(a)



(b)



(c)