未貸出図書が図書館の蔵書に占める割合を推定するための方法

岸田和明（慶應義塾大学文学部）
kz_kishida@z8.keio.jp

1.研究の目的

蔵書回転率は長年に渡って蔵書評価のための指標として利用されてきた。蔵書回転率は業務統計から容易かつ客観的に計算可能であり，この点で，さらなる有効活用が望まれる。

統計学的には，蔵書回転率は，貸出回数による図書の度数分布における平均値に相当し，この分布が正規分布とならずに歪む際には，貸出記録には表れない未貸出図書の冊数を正確に見積もった上で蔵書回転率を計算する必要がある。つまり，未貸出図書の存在を十分に把握できず，計算から漏れてしまえば蔵書回転率は過大評価となり，逆に，書庫管理の不備などの理由で貸し出される見込みのない図書が未貸出図書として計数される際には過小評価となってしまう。

本研究の目的は，この未貸出図書が蔵書に占める割合を推計する方法の検討にある。具体的には，貸出回数による図書の度数分布が負の二項分布で記述できると仮定した上で，この分布を利用して簡便に未貸出図書の割合を推定する方法を提案する。

2.「不活性図書」とその把握

上で述べた「書庫管理の不備などの理由で貸し出される見込みのない図書」などの総称として，本稿では「不活性図書」という表現を便宜的に用いる。不活性図書の具体的な例としては，書庫が離れていてアクセスしにくいもの，書庫の乱れにより所定の位置に存在しなかったもの，極端に出版年が古いものなどが挙げられる。この概念を使うと，ある一定期間の貸出に基づいて，貸出可能な図書全体（以下，単に「蔵書」と呼ぶ）を図 1 のように区分できる。

ある図書館において，もし不活性図書がかなりの数存在し，それが蔵書回転率の計算に含められたとすれば，その値は歪んだ事実を伝える可能性がある。例えば，2 つの分野 a と b の蔵書回転率が，その図書館では，

a:1.0　b:0.8

であったとする。ここで，分野 b の蔵書の 30%が実は不活性であり，一方，分野 a の蔵書には不活性図書が含まれないと仮定する。ということは，分野 b についての実際の貸出は蔵書の 70%に対してなされていることになるので，分野 b の「本当の」蔵書回転率は

0.8÷(1-0.3)=1.1428…

と計算され，その大小関係は分野 a と分野 b とで逆転してしまう。

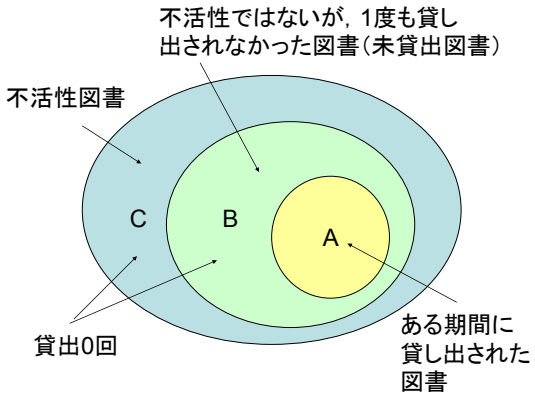

図 1　不活性図書の存在

この簡単な例から，もし不活性図書が存在する場合には，それらを取り除いて計算する必要のあることがわかる。具体的には，蔵書回転率は「貸出回数の合計÷蔵書冊数」で計算されるから，分母の蔵書冊数から不活性図書の冊数を差し引かなければならない。つまり，分母は「1 回以上貸出された図書の冊数+不活性ではない“本当の”未貸出図書の冊数」とすべきである。本稿では便宜的に，後者を単に「未貸出図書」と呼び（図 1 参照），不活性図書と区別する。

しかしながら，実際に不活性図書を識別するのは難しい。貸出回数 0 回の図書を実際に書架上であたってみればいくつかの不活性図書が判明するかもしれないが，すべての不活性図書を把握するにはたいへんな労力を要するため，図書館経営の観点からは，この確認作業は現実的ではない。そこで，以下では，業務統計として得られる単純な貸出統計だけで，不活性図書の冊数を近似的に推

定する方法を検討する。

3.未貸出図書の冊数の推定法

3.1 貸し出されない図書の集計法の問題点

ある一定期間の貸出回数を集計すると，図 2 に示すような度数分布を描くことができる。この種の分布は貸出頻度分布と呼ばれる。蔵書規模の大きな大学図書館では，この図のように，貸出回数 0 回に度数が集中した「歪んだ」分布になることが多い。

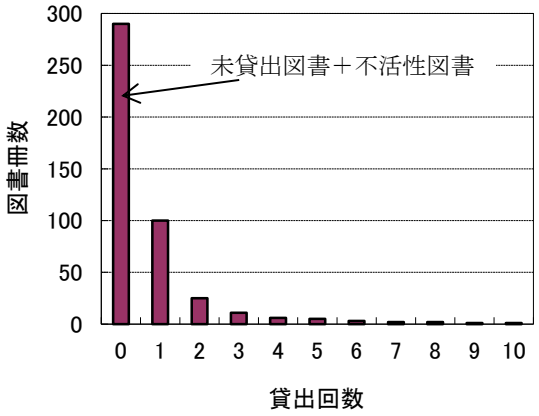

図 2　貸出頻度分布の例

この分布を「釣り鐘型」の正規分布で近似できる図書館の場合には，分布の左裾に位置する貸出回数 0 回の度数は相対的に多くはなく，蔵書回転率の計算には大きな影響を与えない。したがって，不活性図書あるいは未貸出図書の冊数の推定は実際的には必要なく，ここで考えるべき状況は図 2 のように貸出頻度分布が歪む場合である。

注意すべき点は，1 回以上貸し出された図書の冊数の集計とそうでない図書の冊数の集計とが，異なる種類の業務統計に基づくことである。この場合に，実際に発生した事象として把握しているのは「貸出」であり，現在の業務システムでは，この事象の生起回数は貸出処理記録から，誤差なしで正確に把握できる。それに対して，「貸し出されない」という事象は明示的に観察されないので，貸出業務に使用されている蔵書全体が記録されたファイル（蔵書ファイル）から，一定期間，貸し出されない図書を抽出することにより，貸出回数 0 回の図書の冊数を集計せざるを得ない。このため上で述べたような，いわば「目に見えない」誤差が生じる可能性がある。

この誤差は，例えば，蔵書ファイルによる集計の際に，貸し出されない図書の冊数を少なく見積もってしまうことで生じる可能性も考えられる。この場合には，不活性図書が含まれるときとは逆に，蔵書回転率が過大評価となる。

3.2 標準的な推定方法の応用

このように考えると，「業務統計は全数調査であるから標本誤差（sampling error）の考慮は不要」という捉え方に疑義が生じてくる。したがって，まず第 1 の方法として，

方法①:蔵書回転率を標本平均として捉え，95%信頼区間を計算して区間推定を行う

ことが考えられる。もちろん，本研究で考えている誤差は単純無作為抽出によるものではなく，方法①に十分な統計学的な根拠があるわけでない。当然，あくまで 1 つの（粗い）近似と考えるべきである。

もう少し厳密に考えると，貸出回数 1 回以上の部分には誤差が含まれないので，貸出回数 0 回の図書の割合（蔵書に占める比率）のみを推定したほうが状況には合っているかもしれない。したがって，第 2 の方法として，

方法②：貸出回数 0 回の図書の割合についての 95%信頼区間を求めて，その上限と下限に対応する平均を計算することにより蔵書回転率の区間推定を行う

ことが考えられる。ただし，方法①と同様に，単純無作為抽出による誤差ではないという限界に加えて，方法②の推定は「各図書が貸出回数 1 回以上の集合に属するのか，0 回の集合に属するのか」という意味での 2 項母集団を仮定しているという問題がある。この仮定での 2 値変数は，ここで問題としている「不活性図書であるか，（本当の）未貸出図書であるか」という 2 値変数とは，定義的にずれてしまっている。

別の大きな問題は，方法①と方法②では，単に区間推定の結果しか与えられず，不活性図書の存在により蔵書回転率が過小評価になっているのか，あるいは過大評価になっているのかに関する情報が得られない点である。何かの理由で過小評価・過大評価のどちらであるかがわかっていれば別であるが，これらの方法には，統計学的前提の問題に加えて，実用上の制約があるといえる。

3.3　負の二項分布を応用した割合の推定

おおよそ 1970 年代から 90 年代にかけて，図 2

のような貸出頻度分布を計量書誌学的なアプローチを用いて研究する試みが多数行われた。そこでの１つの結論は, その理由は不明ではあるものの, ほとんどの場合に，貸出頻度分布は負の二項分布で記述できるということである [1]。この分布は, 貸出回数を確率変数xとして,

$$P(x) = \binom{x+k-1}{x} p^k (1-p)^x, x = 0,1,2, \dots$$

で表される [2]。ここでpとkはパラメータである。ただし，貸出頻度分布の分析では，kは整数ではなく実数となるので，式中の二項係数は,

$$\binom{\alpha}{x} = \frac{\Gamma(\alpha+1)}{\Gamma(x+1)\Gamma(\alpha-x+1)}$$

で定義される（$\Gamma(\cdot)$はガンマ関数)。

「貸出頻度分布は負の二項分布に従う」という仮定が妥当であれば, 負の二項分布から(本当の)未貸出図書の割合を$P(x=0)$として推定するのが自然である。すなわち,

方法③：負の二項分布を実際の貸出頻度分布にあてはめて，$P(x=0)$を未貸出図書の割合とする

ことが考えられる。この方法の最大の問題点は「貸出頻度分布が負の二項分布に従う」という前提が妥当かどうかであるが，もしこれが成立すれば，$P(x=0)$を未貸出図書の割合とすることは理に適っている。本稿では以下，この前提が経験的に成立していると仮定して議論を進める。

具体的には，以下のように$P(x=0)$を推定する。貸出回数１回以上の図書冊数には記録上の誤差が含まれないので固定し，これを$n(1), n(2), \dots$と表記する。例えば，$n(1)$は貸出回数１回の図書の冊数を示す。一方，貸出回数０回の図書冊数を変数yで表し，分布データ$y, n(1), n(2), \dots, n(x_{max})$と負の二項分布との乖離度が最小となる$y$の値を未貸出図書の冊数として採用することにする。ここで, x_{max}はデータにおける貸出回数の最大値である。すなわち，分布データと負の二項分布との間の乖離度を返す関数を$g(\cdot)$として,

$$y' = \arg\min_y \ g(y, n(1), n(2), \dots, n(x_{max}))$$

で求められたy'を未貸出図書の冊数の推定値とする。

データを負の二項分布にあてはめる際のパラメータの計算には積率推定を用いる。すなわち,

$$\hat{p} = \bar{x}/s^2, \qquad \hat{k} = \bar{x}^2/(s^2 - \bar{x})$$

であり [2]，ここで$\bar{x}$は平均，s^2は分散を意味する。例えば，平均は,

$$\bar{x} = \frac{1}{m} \sum_{x=1}^{x_{max}} x \times n(x)$$

として計算される。この式中のmは図書の全冊数を意味し,

$$m = y + n(1) + \cdots + n(x_{max})$$

である。つまり，単にyを$n(0)$として用い，普通に平均を求めればよい（分散も同様)。乖離度としては，本稿では，よく知られた,

$$\chi^2 = \frac{(y - mP(0))^2}{mP(0)} + \sum_{x=1}^{x_{max}} \frac{(n(x) - mP(x))^2}{mP(x)}$$

を用い, 関数$g(\cdot)$はこのχ^2の値を返すものとする。

なお，変数yは連続変数ではないため，上記の最小化問題をニュートン法等で解く必要はない。$y_1 \le y \le y_2$のように適当な範囲を決め, それぞれのyの値で平均, 分散, パラメータを順次計算し, それに基づいて乖離度χ^2を求めて記録していけば，簡単にその最小値を見つけることができる。

表１　貸出頻度分布のデータ

貸出回数	データ１		データ２	
0	8378	70.8%	5542	58.5%
1	1671	14.1%	1876	19.8%
2	781	6.6%	826	8.7%
3	398	3.4%	475	5.0%
4	241	2.0%	283	3.0%
5	146	1.2%	170	1.8%
6	80	0.7%	92	1.0%
7	59	0.5%	72	0.8%
8	40	0.3%	44	0.5%
9	26	0.2%	24	0.3%
10	12	0.1%	36	0.4%
11	3	0.0%	16	0.2%
12	0	0.0%	9	0.1%
13	1	0.0%	5	0.1%
14	2	0.0%	2	0.0%
15			2	0.0%
16			3	0.0%
17			0	0.0%
18			1	0.0%
19			0	0.0%
20			0	0.0%
21			2	0.0%
合計	11838	100%	9480	100%
平均	0.6557		0.9951	
分散	1.9728		3.2635	

4. 未貸出図書冊数の推定の実際例

未貸出図書の冊数（割合）を推定するための上記の方法を実際のデータに適用してみる。使用するデータを表 1 に示す。これらの 2 つのデータは，異なる 2 つの私立大学図書館の貸出記録を集計したものであり，「データ 1」は 1985 年度に受け入れた図書の 1986 年度の貸出頻度分布，「データ 2」は 1982 年度に受け入れた図書の 1983 年度の貸出頻度分布である [3]。それぞれの蔵書回転率は，0.6557，0.9951 となっている。

表 2　区間推定の結果（方法①と②）

方法	データ	区間		
		下限	平均	上限
①	1	0.6304	0.6557	0.6810
	2	0.9588	0.9951	1.0315
②	1	0.6504	0.6557	0.6611
	2	0.9854	0.9951	1.0051

これらのデータに対して，方法①と方法②とを適用した結果を表 2 に示す。方法①での 95%信頼区間は，データに含まれる図書冊数が多く，十分大きな標本と見なせることから，母分散を標本分散で単純に置き換えた，

$$\bar{x} \pm 1.96\frac{s}{\sqrt{m}}$$

により算出した。ここでmは前と同様に図書の総数を示し，データ 1 では$m = 11838$，データ 2 では$m = 9480$である（表 1 参照）。一方，方法②では，標本サイズについての同じ理由から，2 項母集団における割合qの 95%信頼区間を，標本における割合q'をqと同一視して，

$$q' \pm 1.96\frac{\sqrt{q'(1-q')}}{\sqrt{m}}$$

によって求めてある。なお，データ 1 では実際にこの区間は 70.0%～71.6%，データ 2 では 57.5%～59.5%であった。表 2 からは，貸出回数 0 回の図書冊数のみに誤差が含まれると仮定した方法②での区間のほうが，当然，狭くなっていることがわかる。

一方，方法③を用いて，データ 1 での未貸出図書の冊数を推定した結果を図 3 に示す。図が示すとおり，未貸出図書を「5207 冊」とした場合の乖離度が最小であり（χ^2 =19.6），割合に直せば，約 44.0%となる。これは，3171 冊の図書が何らかの理由でこの年は不活性であったことを意味し，したがって，蔵書回転率は，0.6557 から 0.8956 へと増加する。

それに対して，データ 2 で未貸出図書冊数を推定すると，「7774 冊」が最小の乖離度（χ^2 =27.3）を与えるため，蔵書回転率は，0.9951 から 0.8055 に減少する。本研究の前提に立てば，データ 2 の場合には，何らかの理由で，記録上，未貸出図書が補足されていない可能性がある。

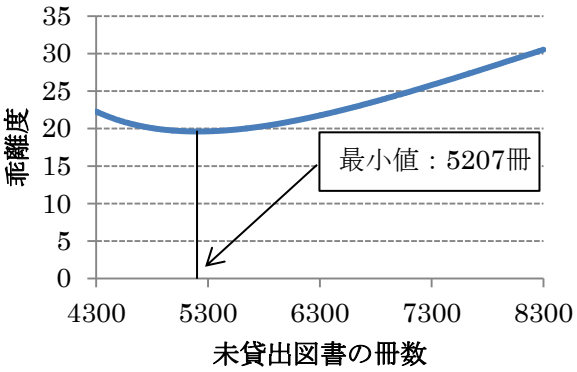

図 3　未貸出図冊数の推定（方法③，データ 1）

5. 問題点と課題

利用者は，新しい図書を利用する傾向があるとしても，受入年度を明示的に意識して貸出を行うわけでないため，本報告で用いたデータが，未貸出図書の冊数（割合）の推定を試すのに適切ではないことも考えられる。そして何よりも「貸出頻度分布が負の二項分布に従う」という経験的事実をあまりにも過信しすぎているという批判は現時点では避けられない。この点，今後の検証がさらに必要である。

引用文献

1)岸田和明．蔵書管理のための数量的アプローチ：文献レビュー．Library and Information Science. 1995, No.33, p.39-69.

2)竹内啓，藤野和建．2 項分布とポアソン分布．東京大学出版会．1981．262p.

3)岸田和明ほか．大学図書館における館外貸出データの分析手法：オブソレッセンスと貸出頻度分布の分析を中心として．図書館研究シリーズ. 1994, No.31, p.79-127.