

みやこ わせ
都忘れ

人の名にも一度聞けば忘れられない名前がある。しかし、植物の名前にも、その優雅さに身も心もウツリするようなものがある。

静御前の舞い姿を形容した「一人静」や「二人静」、激流にのまれる時、岸边に立ちつくす彼女に Forget-me-not (私を忘れてくれるな) と叫んだ言葉がそのまま川辺の名もない野草の名前となった「忘れな草」などなど……。

「都忘れ」もそんな野草の一つだ。忘れるといいながら、実際は忘れられないのが人情というものらしい。深山嫁菜と本名で呼ぶよりもズーッといい。

都のような華麗さはないが、清そで気品があり、それでいて傲りがなく、紫の花には、日本女性の清潔さが凝縮されている。この花もきっとそんな女性にひかれて都をあきらめた男が、都を偲んで名づけた花なのだろう。

今月の主な行事

- 1 日 学校基本調査日
- 3 日 憲法記念日
- 5 日 子供の日
- 14日～16日 統計グラフ指導者講習会(水戸市・結城市・土浦市)
- 18日 市町村統計担当者会議(議会大会議室)
- 21日～22日 商業動態統計ブロック会議(大洗町)
- 24日～25日 関東ブロック統計主管課長会議(栃木県)
- 31日 法人企業投資動向調査基準日

標 本 設 計 (1)

— 確率抽出と標本誤差 —

今月から6回にわたって、標本設計の具体的な作業手順などについて書く予定ですが、多くの場合確率についての正しい理解が必要です。確率については、何となくわかるという人は多いようですが、正しく理解している人は少ないようです。追及していくと難しいことになりますが、標本調査の仕組みを理解するには、ごく基本的な知識だけでよいのです。そこで今回は、まず確率に関係した話を通じて感触を得ていただき、ついで標本誤差の意味、標本の規模との関係などにふれていきたいと思います。

§ 1 不思議な確率現象

1.1 身近かな例

ちょっと古い話で恐縮ですが、今年の正月はわが家に300枚ほどの年賀はがきがとどきました。大部分が抽籤による「お年玉つき」です。例年1月15日頃抽籤が行われたものですが、今年は郵便ストの影響で抽籤日がおくれ、月末になってしまいました。おくれたのはともかく、当たり番号が発表されると、はがきの番号を照らし合わせて、当たりのはがきを選び出すわけで、これは毎年娘が進んでやってくれます。今年は4等(下2桁、3通り)が3枚識別されました。3等以上は無し。「300枚でたった3枚か、少ないなあ」と娘は不満な顔。私は「抽籤だから仕方ないよ。当たりがよその人のところへ行ってしまったのさ。来年を楽しみにしよう。」となぐさめてやりました。

年賀はがきが「当たる」というのは確率的なことと考えられます。4等が当たる確率は——厳密なことはあとにして——下2桁3通りということから、直観的に3/100であると浮かびます。すると、常識的に、100枚のはがきに対しては3枚、300枚のはがきに対しては9枚ぐらい当たることを期待してもよいでしょう。こういう基準で期待しますと、8枚以下なら少ない、10枚以上なら多いと感じることとなるでしょう。娘が「3枚では少ない」といったのは、確率のことを考えた上かどうかわかりませんが、もっと多くを期待していたに違いありません。

1.2 確率の意味

さて、4等が当たる確率とは、統計理論の上ではおおむね次のように考えます。年賀はがきを順序不同の状態で1枚ずつ、当たりかはずれかを調べていき、調べた枚数に対する当たりの枚数の比を考えます。(もっと厳密には、無作為に、独立的に1枚ずつ調べるのがたてまえです。)調べる枚数を多くしていくと、その比は一定の値に近づくことが知られています。そこで、調べた枚数を非常に多くし

たときの、比の極限値を想定し、それを当たる確率を考えるのです。

年賀はがきの4等の場合、総枚数と当たりの枚数が初めからわかっていますから、比の極限値はすぐわかります。一般的にいきますと、一定の条件下で、ある事柄が起きる、あるいは起きないということが毎回独立的、偶発的にきまるとき、起きる確率というのは、同じ条件下での偶発の総度数に対する起きた度数の比の、総度数を非常に大きくしたときの極限値と考えます。この場合、その比の値について、予備的な知識がなければ極限値を知ることはできません。わかる限りの総度数を用いて近似するほかはありません。

たとえば、あるパチンコの機械の当たり穴に入る確率というのは、

- (1) 同じ条件で多数回玉を打ち出す。(同じ条件とは、厳密には難しいけれども、ここでは、きめられた打ち方に従って打つという意味でかまいません。条件には幅があつてよいのです。)——例、4000回打った。
- (2) 当たり穴に入った回数を記録する。——例、200回入った。
- (3) (1)の回数に対する(2)の回数の比を出し、確率の近似値とする。(この確率は(1)できめた条件の中で定義されていると考えます。)——例、 $200/4000=0.05$ を確率の近似値と考える。

ということになります。

1.3 予期し難い小確率の事象

確率が3/100で起きるような事柄は、300回くり返すと何回起きるでしょうか?それは起きるまいかと解りません。心もとない答えですが、こう答えるよりほかはないのです。では、起きる回数について全く何もいえないのかというと、決してそうではないのです。300回くり返したとき、その事柄が起きる回数をすべて想定してみますと、0回、1回、2回、……、300回と、全部で301通りあります。そしてこの中のどれかに必ずあたります。どれにあたるかは、それぞれ確率として計算され、その合計は1になります。ちなみにこの場合の確率を計算してみますと表-1のようになります。特定の回数が起きる確率はいずれも小さく、しかも回ごとに相当の差異がみられます。表-1は、いかえれば、3/100の確率で起きる事柄が、300回の反復試行の中で何回起きるかについてのルールを示したものです。このルールは、300回くり返すことを1つの試みとして行うとき、その事柄が起きる回数を具体的に生み出す機能をもってお

表-1 確率3/100で起きる事柄が300回の試行中に起きる回数とその確率一覧表

起きる回数 r	起きる確率	確率の累計
0	0.0001	0.0001
1	0.0010	0.0011
2	0.0046	0.0057
3	0.0142	0.0199
4	0.0325	0.0524
5	0.0596	0.1120
6	0.0906	0.2026
7	0.1177	0.3203
8	0.1333	0.4536
9	0.1338	0.5874
10	0.1204	0.7078
11	0.0982	0.8060
12	0.0731	0.8791
13	0.0501	0.9292
14	0.0318	0.9610
15	0.0187	0.9797
16	0.0103	0.9900
17	0.0053	0.9954
⋮	⋮	⋮
⋮	⋮	⋮
120	1.70×10^{-99}	1
⋮	⋮	⋮
⋮	(以下更に)	⋮
⋮	(微小値)	⋮
300		1
計	1	—

〔表の説明〕

300回中 r 回現われる確率は一般に

$${}_{300}C_r \left(\frac{3}{100}\right)^r \left(\frac{97}{100}\right)^{300-r}$$

で表わされます。表の中央の欄が r = 0, 1, 2, …, 300 に対するこの式の値で、右の欄は r の小さい方からの累計です。

り、起きる回数の母集団といえます。そして、実際に起きた回数を標本といえます。

さきの年賀はがきの例では、標本の値は3となって現われたわけです。3となる確率はわずかに0.0142です。こういう小さな確率で起きる事柄は、偶発性が非常に高く、初めから期待して生じるわけではありません。ところが起きる回数に幅をつけると、確率は大きくなります。たとえば、4回～14回起きると、としますとその各回の確率を合計するので0.9411と大きくなります。また13回以下、としても0.9292と大きくなります。確率は、大きい場合ほどよく起きます。確率はちょうど、起きることへの期待に対する信頼度のような意味を持ちます。しかし、年賀はがきの例からもわかりますように、微小な確率でも起きるときは起きます。これは1に近い確率でも起きないことがあるのと同じです。いずれも期してかなえられず、期せずしてかなえられるという、意のままにならない代物です。確率とは不思議なものです。

標本調査の結果も、これと同様に考えることができます。たとえば、標本の平均が母集団の平均に一致する、あるいはそれに近い1つの値に一致するという確率を求めてみますと通常微小な値となりますから、初めから一致するのを期待するのは無理なことです。しかし、幅を作って、その幅の中に入る確率を計算しますと、幅の広さに従って確率は大きくなります。統計的推定とは、こうして、信頼に足る確率をもたらすような幅を作り、その確率を背景として標本と母集団を関係づけることをいうのです。

§ 2 標本誤差の意味とその表わし方

2.1 標本誤差と非標本誤差

統計数字は誤差を伴います。誤差とは、本来得ようとしている値——これが唯一つあるとして、便宜上真の値と呼びます。——と統計数字とのくい違いのことです。誤差が生じる原因はさまざまですが、確率抽出に基づく標本調査では、標本を確率抽出することによって起きる誤差と、その他の原因によって起きる誤差に、大きく2つに分けることができます。標本を確率抽出することによって起きる誤差は、いいかえれば、くじの当たりはずれからくる誤差で、標本誤差といえます。標本誤差以外の誤差は、その原因は非常に多岐にわたりますが、一言でいえば、調査の仕組みからくるといってよいでしょう。非標本誤差といわれます。標本誤差は標本調査だけに生じますが、非標本誤差は標本調査だけでなく、全数調査にも生じます。

2.2 標本誤差の尺度としての標準誤差の意味

どんな原因による誤差も、統計数字を真の値とくい違わせる作用をもちます。したがって、どんな原因による誤差も、真の値との差異として考えたいわけですが、しかし、そういう誤差は考えるだけであって、数値として表わすことはできないのです。もしできるなら、それから真の値を逆算できることとなり、標本調査を論ずる意味がなくなってしまうでしょう。ですから、こういう意味での誤差というのは、存在するけれどもわからない、ちょうど真の値と同じ立場にあるわけです。しかし、こういうことでは役に立たない、何とかして数値として表わしたい、となると特別の約束がいります。話を標本調査に限りましょう。確率を用いた標本設計の場合、1つの統計数字は、一連のくじ引きによって現われたようなものですから、もし、同じ設計下で抽出し直したら、異なる標本が抽出され、異なる統計数字となるでしょう。また、これを何度もくり返したと仮定すれば、統計数字はその都度少しづつ、時には大きく異なるでしょう。このように、抽出を何度もくり返したときの統計数字の違いを考える意味は、さきに説明した、試行を何度もくり返して確率を求める意味と共通です。こうして得られる統計数値の間の違いは確率抽出によって生じたものですから、それらの違いが全体として小さければ、統計数字は抽出に対して安定しており、大きければ不安定であると解釈することができます。

このことに着目して、標本誤差を数値で表わす工夫をするわけです。いくつかの数値の全体としての違い（ばらつき）を測るには、分散、標準偏差、変動係数などによる測り方がふつうです。中でも標準偏差が最も基本的で、これで表わしておけば分散にも変動係数にも換算することができます。そこで、同じ設計のもとに、標本抽出を多数回くり返したとした場合に想定される多くの異なった統計数字についての標準偏差を考え、これを特に標準誤差と呼び、標本誤差の基本尺度とします。

なお、標本の規模を大きくしていった、全数調査(100%調査)をしたときは、確率にもとづく誤差はなくなりますから、標準誤差は0となります。このことは、統計数字から標本誤差を取り除いても、直ちに前述の真の値に一致するものでないことを意味します。すなわち、非標本誤差を含んだ値が残るわけです。以下、便宜上100%調査をした場合に得られるべき値を「100%調査値」といいます。標準誤差は、標本調査による統計数字が、100%調査値からどのくらい離れているかを平均的に表わした尺度であって、個

別の統計数字と100%調査値との差ではありません。

さて、こうして標準誤差を定義しても、すぐ数字で表わせるものにはなりません。なぜなら、同じ設計のもとに標本抽出を多数回くり返すということは、想定であって、モデル実験でもない限り実行できないからです。そこで仕方なく、1回の抽出で得られた標本から標準誤差を推定することとなります。これは理論的にできるのです。こうして、標本調査では100%調査値を標本から推定すると共に、その誤差もまた標本から推定します。よって、こうして表示した誤差にも標本誤差があるわけで、結局標本調査の結果から誤差を伴わずに断定的にいえることは1つもないのです。

2.3 区間推定法の意味

ところで統計数字を中心として、その前後に標準誤差を単位とした幅を作ると、この幅の中に100%調査値が含まれる確率がきまります。確率の大きさは標本の設計の仕方などによって異なりますが最低値がきまっています。たとえば、統計数字の前後に標準誤差の2倍の幅(前後で計4倍)をとると、その幅の中に100%調査値が含まれる確率は最低75%となります(チェビシェフの不等式による下限値)。もし、統計数字の理論的分布型が正規分布になると見なせるような標本設計なら、確率は95%と高まります。こうして作った幅を信頼区間、その確率を信頼水準、信頼係数、信頼度などといいます。(信頼区間を作るにはこれ以外にも方法があります。)信頼水準は信頼区間に対応してきまりますが、前後に標準誤差の2倍づつの幅を作って直ちに95%の信頼水準とするのは、必ずしも当を得ません。根拠をしっかりとわきまえてからにすべきです。標本設計によっては意外に低いこともあれば、100%に近いこともあります。信頼水準を表わす根拠がはっきりしないときは、無理に95%などと言明せず、標準誤差を表示するにとどめる方がよいのです。下限がわかっているのですから。

さて、仮に1つの信頼区間を作り、その信頼水準が90%と示されたとします。これは、統計数字と100%調査値との確率的関係を表わすもので、1つの推定の形式です。こういう方法を区間推定法といいます。区間推定法では、統計数字と100%調査値との絶対的關係はわかりません。よって、その信頼区間の中に100%調査値が入っているかどうかということは議論に値しません。同じ条件下での標本抽出のし直しによってできる多くの信頼区間のうち、100%調査値を含むものが90%ぐらいいある、という意味に過ぎません。

2.4 標本の規模と標本誤差の関係

標本誤差はいろいろな要素によって異なります。主な要素の1つに標本の規模があります。大抵の場合、標本の規模を大きくすると標本誤差は小さくなります。具体的な関係は標本設計の内容によって異なりますが、その関係がわかっていると標本の規模を調節することによって標本誤差を管理することができます。

標本を確率的に抽出する方法のうち、最も基本的なものは単純任意抽出法といわれる抽出法ですが、この方法を用いて母集団の平均や総和などを推定しようとするときは、標本誤差は標本の規模の平方根に逆比例する、ということが知られています。(厳密には抽出率が非常に小さいという前提が必要です。)たとえば標本を2倍にすると標本誤差は $1/\sqrt{2}=0.7$ 倍に、3倍にすると $1/\sqrt{3}=0.6$ 倍に、4倍にすると $1/\sqrt{4}=0.5$ 倍と小さくなります。このように、標本誤差は標本の規模の変化の割には敏感に動きませんが、これらの増減関係は一般に次のようになります。すなわち、ある規模の標本とその標準誤差を基準としたとき、標準誤差を $\alpha\%$ 増や(減ら)してもよいなら、標本の大きさは $\beta=100\left(\frac{100}{100+\alpha}\right)^2-100\%$ 減ら(増や)してもよいのです。表-2は α のいろいろな値に対する β の値を表わしたもので、 α の少しの変化に対して β が大きく変化するのがわかるでしょう。

表-2 標準誤差の増減率(α)と標本の規模の追加削減率(β)との関係
+は増、-は減

α (%)	β (%)	α (%)	β (%)
-1	+2.0	+1	-2.0
-2	+4.1	+2	-3.9
-3	+6.3	+3	-5.7
-4	+8.5	+4	-7.5
-5	+10.8	+5	-9.2
-7	+15.6	+7	-12.7
-10	+23.5	+10 ⁽²⁾	-17.4
-15	+38.4	+15	-24.4
-20	+56.3	+20	-30.6
-30	+104.1	+30	-40.8
-50 ⁽¹⁾	+300.0	+50	-55.6
-90	+9900.0	+100	-75.0

〔表の見方の例〕

- (1) 標準誤差を現状より50%減らすためには、標本規模を現状より300%増やす必要がある。
- (2) 標準誤差が現状より10%増えてもよいなら、標本の規模は現状より17.4%減らすことができる。

2.5 標準誤差の大きさの目安

さて、1つの統計数字の標準誤差はどのくらいの大きさであれば満足できるのか、ということにはきまりがありません。許せる範囲で大きくするのがよいでしょう。なぜなら標本調査はもともと、ある程度の標本誤差を認めて行なうものですから。大き目の誤差で間に合うときは、標本の規模も小さくて済みます。標本設計の演習問題などでは、標準誤差の大きさは、推定値に対して5%ぐらいにしている場合が多いようですが、これは考えやすい基準の1つです。しかし、これはあくまで考えやすさからくる基準ですから、もっと精度を高めたければ、2%、1%などと小さくすればよいわけです。このように標準誤差を推定値との相対比に直した場合、それを相対標準誤差、標準誤差率または当該推定値の変動係数といいます。これに推定値を乗じると標準誤差になります。相対標準誤差が5%というのは、それを2倍すると10%で、これは標準誤差を2倍にすることに対応しますから、推定値の前後に10%づつの幅を考えると、これが $\pm 2 \times$ (標準誤差)の信頼区間にあたります。一方推定値から10%を増減すると上2桁目が動くことになり、有効数字はある意味で1桁と見なされます。ある意味とは、やや不明瞭なきらいはありますが、2倍の幅をとるとかなり高い確率が伴うという含みなのです。その意味で「相対標準誤差5%で有効数字1桁」は考えやすいということです。同じ発想で有効数字を2桁にするには、相対標準誤差を0.5%に下げなければなりません。すると、標準誤差も $1/2$ に、つまり現状より90%減らさなければなりませんから、表-2により標本は9900%、すなわち99倍増、すなわち現状の100倍にしなければなりません。となると驚くほど大規模な標本にしなければならないように思われるかもしれませんが、相対標準誤差5%という例は、標本の数が20~30でもしばしば見られることですから、これを100倍するといっても2000~3000ということでは驚くほど大きいものではありません。

行政資料室はこのように利用された ……………

行政資料室は、行政資料を集中管理し効率的に県職員の利用・活用に供し、情報化社会にふさわしい近代的・合理的な県行政に資するため昭和42年に設けられた施設である。室の運営管理は統計課(行政資料グループ)があたっている。場所は付属庁舎4階南側、総面積は123平方メートルである。

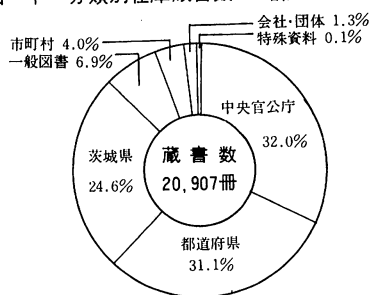
また、53年度からは新しい仕事として月刊誌である「統計いばらき」「新着資料情報」をはじめ、年1回発刊の「統計年鑑」「県勢要覧」「茨城県のすがた」「都道府県勢の展望」等を編集発行している。このことにより当室の資料を利用される方々に対する奉仕活動はより深められることになり、更には統計業務の相談の窓口としての機能が新しく生じはじめてきている。

当資料室の昭和53年度中における利用の状況は次のとおりである。

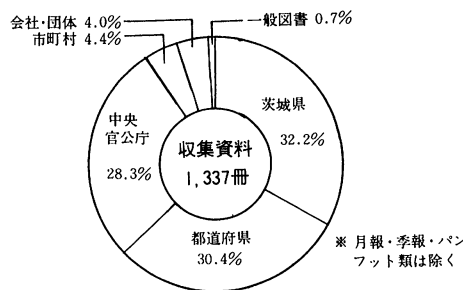
表一 昭和53年度収集資料及び在庫蔵書数

発行機関	中央官公庁 資 料	都道府県 資 料	茨 城 県 資 料	県内市町村 資 料	会社・団体 資 料	一般図書	特殊資料	計(冊)
昭53年度 収集図書	379	406	430	59	53	10	—	1,337
昭53年度 未蔵図書	6,695	6,496	5,135	846	263	1,449	23	20,907
同 上 構成比%	32.0	31.1	24.6	4.0	1.3	6.9	0.1	100

図一 分類別在庫蔵書数 昭和54.3.31現在



図二 昭和53年度収集資料数



利用者数(閲覧・貸出し)と利用冊数

閲覧利用者は、年間1,396人でその内訳は当室内閲覧は920人、貸出利用者476人、そのほかコピー複写利用者495人があり、1日平均7.0人である。利用者の多い月は2月、以下8月、9月、10月、7月、1月の順となっている。このほか電話、文書、統計調査方法等に関する照会や相談が293件寄せられた。

利用冊数は全体で5,955冊(コピー利用、レファレンス・サービスは除く)で1人当たり4.3冊、資料分類別にみると、本県関係資料の利用が3,421冊(57.4%)と最も多く、次いで中央官公庁資料1,903冊(32.0%)、一般図書194冊(3.3%)、会社・団体関係資料154冊(2.6%)、都道府県資料及び県内市町村資料となっている。また、利用された資料は、80.7%が本県や国の機関の公表した統計調査の結果表で、

昭和三十五年行政資料室利用状況実績

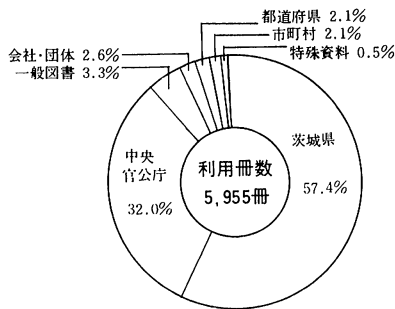
いずれも行政施策の見直しや企画立案の参考として使われている。利用の多い主な資料を例示してみると、農業関係の統計書が881冊(14.8%)で最も多く5年毎の農林業センサス、及び茨城県農業基本調査結果の利用が目立つ。次いで国勢調査、茨城の人口と世帯などの人口に関する統計が854冊(14.3%)、茨城県統計年鑑、要覧が844冊(14.2%)、商業及び工業統計調査結果が574冊(9.7%)、地域発展のた

めの県の方策や生活環境の整備の施策に関する現況、計画に関するもの438冊(7.4%)、統計学、郷土史等の一般図書が356冊(6.0%)、交通、運輸、観光等に関する統計が246冊(3.6%)、学校教育、社会教育関係の統計、219冊(3.7%)、市町村勢の現況の把握に関する資料、214冊(3.6%)、建築、道路等土木関係資料、195冊(3.3%)……(以下略)となっている。(表一・図一)

表一 昭和三十五年閲覧冊数及び利用者数

区分 内訳	閲 覧 冊 数								利 用 者 数 (人)					計 (人)
	中央官公庁資料	都道府県資料	茨城県資料	市町村資料	会社・団体資料	一般図書	特殊資料	計(冊)	室内利用	貸出利用	コピー利用		レファレンスサービス(件)	
室内閲覧資料	1,238	84	2,395	98	110	91	23	4,039	920	—	495	7,704	150	1,415
貸出資料	665	42	1,026	26	44	103	10	1,916	—	476	—	—	143	476
計	1,903	126	3,421	124	154	194	33	5,955	920	476	495	7,074	293	1,891
構成比%	32.0	2.1	57.4	2.1	2.6	3.3	0.5	100	48.7	25.1	26.2	—	—	100

図一 昭和三十五年資料別利用冊数



職業別資料利用状況

コピー利用者及びレファレンス・サービスがあったものを除く図書資料のみの利用者は、1,396人でこれを職業別にみると公務員が906人で利用者全体の64.9%を占め、行政事務執務のための利用が主である。そのほか公務員に対するレファレンス・サービスの分野が143件あった。

内容としては県職員及び市町村職員からの問い合わせ、照会をはじめ世論調査、実態調査上の標本設計の方法等の相談業務の傾向が多くなってきている。残りの150件については、一般の方々である。次いで学生が281人、20.1%あり、殆んどが県内の大学生で鹿島開発や筑波学園都市の現勢、県民福祉計画、生活環境整備に関するものの利用が大部分であった。会社員・団体等職員が81人、10.7%あった。特に会社員は東京からの来客が多く、その利用相談は本県の県政全般にわたる資料の要求である。次いで教員(主に

大学教授)、自由業、無職(主に市町村の郷土史編さんに関係の人の)の順となっている。(表一)

表一 昭和三十五年 職業別資料利用状況 (単位:人)

職業資料	公務員	教員	会社員	学生	自由業無職	計
図書利用	906	33	149	281	27	1,396
構成比%	64.9	2.4	10.7	20.1	1.9	100

電子コピーの利用状況

本室にある複写機は、いまでは旧式に属するが、表一のとおり495人で7,074枚の利用があった。前年度が150人、375枚の使用であったことに比べれば大幅な増加である。これは最近における行政情報の利用面での需要は、個別情報のものから関連情報を広く求める傾向が顕著で単一の資料ばかりでなく、研究、調査、郷土史、統計調査結果図書へとその関連分野への需要が目立ってきているためのものである。なお利用者の内訳は、公務員が243人、49.1%。学生が143人、28.9%。会社員、団体等職員が81人、16.4%。以下教員、自由業、無職の順となっている。(表一)

表一 昭和三十五年コピー複写利用状況 (単位:人)

職業資料	公務員	教員	会社員	学生	自由業無職	計
図書利用	243	16	81	143	12	495
構成比%	49.1	3.2	16.4	28.9	2.4	100

(行政資料室 環)