『数学嫌いのための社会統計学』プラスあるふぁ

１.ランダムサンプリング（本書第１章2-3の補足）

ここでは、ランダムサンプリング（無作為抽出法：本書第１章参照）の方法を学ぼう。

まず、母集団（本書第１章参照）となる調査対象者（団体）全員の名簿を用意する。そして、この名簿に記載された各メンバーをあらかじめナンバリングしておく。たとえば、名簿に記載されたメンバー数が100人なら、１から100までの番号をふっておく。ここから５人（５団体）をランダムサンプリングしたい場合は、１から100までの番号の中から、くじ引きに似た方法で５個の番号を選べばよい。番号を選ぶやり方は色々あるが、ここでは、**資料１**のような、乱数表を利用するやり方を紹介しよう。

乱数表には、0，1，2，3，……，9の数字が完全に同じ確率（本書第２章参照）で出現するように配列されている。乱数表から乱数を取り出す作業はくじ引きと同じ原理である。乱数表の横の並びを**行（row）**、縦の並びを**列（column）**という。くれぐれも行列を間違えないようにしたい。

乱数を取り出すには、つぎの手順を踏む。

ステップ１：スタートする乱数、すなわち最初の行列を決める。たとえば、２月10日が誕生日であれば、２行10列から始めてみよう。この場合、「0」がスタート地点になる。

ステップ２：方向を決める。右でも左でも上でも下でも任意の方向へ向かう（斜めでもかまわない）。この場合、右へ向かうことにする。

ステップ３：乱数を何桁ずつ拾うか決める。たとえば、母集団が100名であれば、母集団にふられる通し番号は「01」から「（1）00」だから、乱数を２桁ずつ拾えばよい。この場合、最初に拾う乱数は「07」になる。

ステップ４：データの個数にしたがって、必要な数の乱数を拾っていく。仮に５つのデータを必要とする場合、「07」「21」「21」「75」「14」になる。

ステップ１から４までを示すと、下の**図表１**になる。

図表１：乱数表の使い方

また、乱数を拾っていくと、同じ数が何度でも出てくることがある。その際、**復元抽出法**（sampling with replacement）、あるいは**非復元抽出法**（sampling without replacement）のどちらかの方法で処理をする。復元抽出法は同じ数を何度でも採用する方法、非復元抽出法は同じ数が出てきたら、２回目以降、それを捨てる方法である。この場合、「21」が２度出てきており、復元抽出法であれば「07」「21」「21」「75」「14」の５つがサンプルになり、非復元抽出法であれば２度目の「21」は捨て、「14」のつぎの乱数「84」を採用するため、「07」「21」「75」「14」「84」が５つのサンプルになる。もっとも、一人の調査対象者に複数回、同様の社会調査（調査票調査であれ、インタヴュー調査であれ）を実施することは稀である。調査対象者にとっては負担になるし、調査の信憑性の点から考えてもあまり好ましいことではないからだ。特別な理由がない限り、復元抽出法を用いることはない。

乱数表から必要な数の乱数を取り出したら、後は、調査対象者（団体）名簿と照らし合わせて、乱数に一致するナンバーの対象者（団体）を順に選んでいく。この一連の作業がランダムサンプリングである。

資料１：乱数表

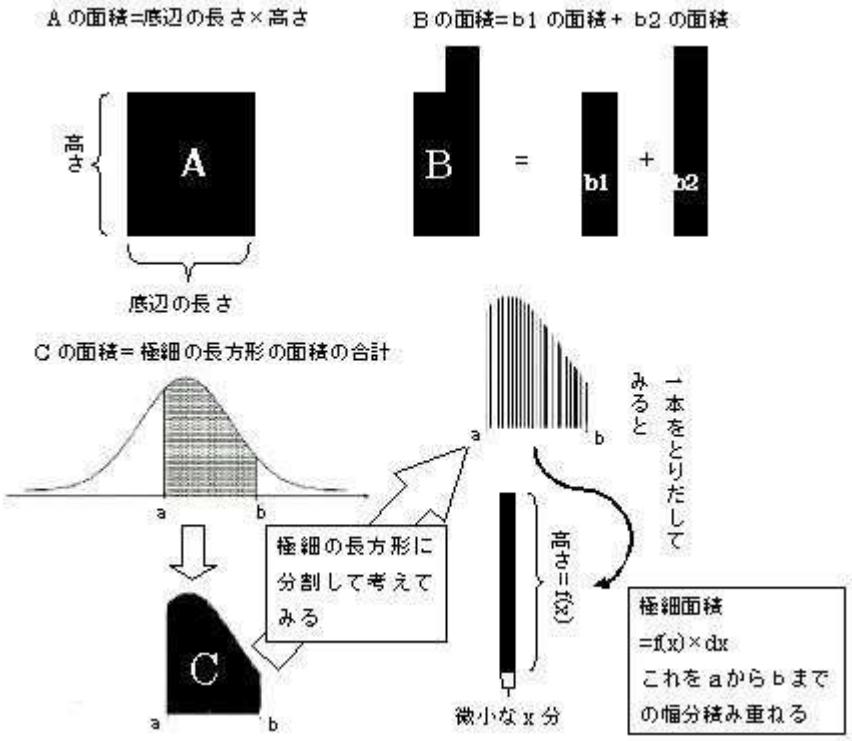
	5→	10→	15→	20→	25→	30→	35→	40→	45→	50→
1	14664	81013	28379	75318	22259	16319	30182	29997	44389	64528
2	85417	07210	72121	75148	45155	49377	90901	91589	32125	61696
3	99344	59450	76264	12225	20832	84709	57803	82669	32407	81846
4	54822	24431	05846	06100	57186	51081	07865	70579	69917	23861
5	98698	87213	93311	80589	25023	77942	26008	96572	54060	75769
6	16584	20859	07974	59979	17474	57221	94294	00062	05545	23582
7	38262	22355	76243	38112	16523	93583	93641	58354	47181	48291
8	54540	50920	43471	48980	81265	12743	97375	75093	71633	29883
9	89246	87636	50130	70181	29719	01322	61521	02478	95617	23587
10	08223	73757	44939	37434	72237	64171	75380	87173	27243	24444
11	20465	08490	32196	15891	12793	91085	88062	46555	49071	81649
12	22770	06995	56593	12156	17605	88471	44885	91447	51108	64590
13	90218	49606	26617	27417	11102	02260	26133	73465	41431	27325
14	06823	67873	69164	15525	78658	43431	43192	99007	94809	22308
15	48741	27907	70897	86653	47347	82283	82379	52023	91588	90925
16	84820	53447	70467	02696	74719	94532	58236	86287	13481	97414
17	96582	86642	79693	47774	94184	54633	11157	77496	25709	67205
18	36971	64118	61303	76138	48157	01771	80629	34388	01612	31984
19	61115	13565	41459	69063	64214	24044	90281	58990	01186	47047
20	57624	56475	18816	74471	77996	78791	75696	30035	04637	22934
21	46590	13334	50301	57996	11863	86284	43349	67681	46386	98306
22	88399	10949	03176	60598	86602	37811	13058	15717	87515	74950
23	29240	53051	07272	52855	19841	57999	19554	85474	78600	89273
24	03903	53344	63517	60018	23310	82969	31420	47678	66089	46035
25	39146	35181	23994	58273	17513	51325	90773	67520	85094	58192
26	15774	75209	95055	68234	78095	56508	29388	10275	89842	03173
27	65064	50788	82700	98676	20158	90378	06023	70340	04136	25008
28	18000	25979	84603	81491	43138	35434	96966	71814	53720	12713
29	31141	91534	72749	15605	72643	87847	48092	53395	20532	28368
30	65632	90950	54034	24748	30366	34394	21816	31321	31326	23939
31	60797	48144	94300	82984	57673	75080	99360	59345	64085	76822
32	73015	40556	05730	27608	30380	75767	06907	25162	07538	42488
33	37868	27051	35319	07228	03150	34607	01131	67281	36994	72850
34	85609	78445	61278	56005	69745	14798	30062	80561	42237	27453
35	54956	99557	66156	25604	25053	09067	48350	04657	14574	24865
36	70327	94249	12659	31541	79711	64856	60460	10375	09812	26998
37	76305	52386	56464	12157	32884	69350	32718	32445	90313	65876
38	61655	61906	44421	09282	76044	62675	71824	45918	39252	78625
39	74817	26939	48902	12334	98500	19043	27361	38689	02747	82065
40	07709	90337	07194	50109	72212	72935	80823	19555	56116	10669
41	84011	60228	68514	55458	01023	80627	87109	53678	32834	81002
42	92544	04800	75605	10856	55150	00459	31869	31990	91004	22812
43	58003	95135	00759	34198	98611	67950	56667	44538	89127	48929
44	45375	74306	60354	22213	06658	11415	38420	47490	39653	41818
45	73524	29812	02259	77064	38455	61521	75947	03947	94970	29484
46	29447	05792	19798	81336	63039	62827	80109	95055	92683	0.667
40	80146	54429	19684	40126	36921	37886	53971	79241	57114	72209
48	36534	37279	20616	30485	19483	87173	62172	01543	14220	21195
49	46716	89580	58255	32275	32602	57516	10942	33401	81418	42342
50	62074	75610	21798	33449	07578	50231	75626	99479	84080	25475

鳥居泰彦『はじめての統計学』日本経済新聞社（25頁）より

２.連続分布における確率と積分（本書第２章2-2の補足）

本書第２章の2－2「確率分布」で説明したように、連続分布の場合、確率密度曲線と横軸の間の面積が確率を意味する。したがって、確率を知りたければこの面積を計算すればよいわけだ。面積の計算は、**図表２**（Cの面積の求め方に注目）のように、求める部分を細長い長方形に分割し、各長方形の面積を横に積み重ねていくといったアイディアである（積分）。

図表２



a≦X≦ｂの区間の確率はP（a≦X≦ｂ）だから、積分の計算式で表現すると確率は

$$P(a \leqq X \leqq b) = \int_a^b f(x)\,dx$$

となる。∫はインテグラル（integral）と読み、積分を表す記号である。dxのｄはdifferential「微小な」のｄで、dxは微小なx分ということである。

この式は高さf（x）×底辺dxで細い長方形の面積を出し、それをaからbの間において左から右へその面積を順に加えていく作業だといってよい。離散分布において、各棒の高さ

を順に加えていく作業と同様の発想である。さらに詳しく知りたい場合は各自、数学の参考書にあたってほしい。

図表２のａからｂの網掛け部分が全体に拡がれば、それは全体の面積だから100％、つまり確率は１である。この場合aが-∞、bが∞ということになり、確率はP（-∞≦X≦∞）で、P（-∞≦X≦∞）＝１である。これを積分の式で表せば、極細の長方形を-∞から∞までの幅で積み重ねることとなるから、

$$P\left(-\infty \leqq X \leqq \infty\right) = 1 = \int_{-\infty}^{\infty} f(x)dx$$

である。左から右へ順に加えて1になるのは、離散分布で棒を全部加えれば１になるのと同じである。

３.母平均の推定と信頼度（本書第12章2-1～2-2の補足）

ここでは、標本平均Xを使った母平均の推定の実例を紹介する。同一の母集団から標本を何度も繰り返して抽出し、それぞれから計算した標本平均Xを使って、母平均の区間を推定してみる。そこでは、推定の際、あらかじめ設定する信頼度（90%や95%）が、どのように機能するのか、そこに注視してもらいたい。

母集団として乱数表（資料１）を用いる。乱数表に列挙されている数全体を母集団、乱数の１ケタの数字を１つのデータとして捉える（事例としておもしろくないが、ガマンしよう）。母平均μは4.5、母標準偏差σは2.872である。（0，1，2，…，9の平均、標準偏差を算出すればよい。）

次に標本抽出について説明する。乱数表の１行には50のデータが存在する。それを一組の標本として見立て、標本の平均と標準偏差を算出する。最初の標本（１行にある50のデータ）は1，4，6，6，4，･･･，2，8で、平均X_1は4.60、標準偏差s_1は2.96となる。この値から母平均の信頼区間を推定する。以下のように、**平均μの推定（大標本の場合）の式**（本書第12章2-2参照）に値を入れればよい（母標準偏差σはわかっていないものとする）。信頼度は90%（α＝0.10）を用いる（Z＝1.64）。

$$\bar{X}-Z\frac{s}{\sqrt{N}} \leqq \mu \leqq \bar{X}+Z\frac{s}{\sqrt{N}}$$

$$4.60-1.64\frac{2.96}{\sqrt{50}} \leqq \mu_1 \leqq 4.60+1.64\frac{2.96}{\sqrt{50}}$$

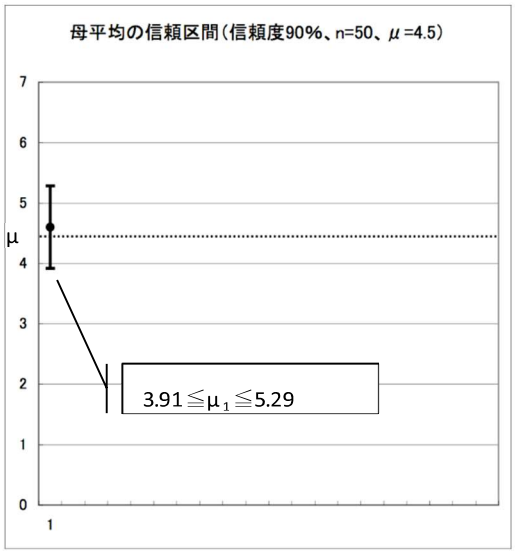
$$3.91 \leqq \mu_1 \leqq 5.29$$

標本１から推定される母平均の信頼区間は、$3.91 \leqq \mu_1 \leqq 5.29$となる。信頼区間を図にしたものが、**図表３**である。（本書の**図表12-1**では区間を横にとっていたが、ここでは便宜的に縦にとっている。）図中の点線は母平均μ（＝4.5）である。図を見ると、標本１から推定される母平均の信頼区間（$3.91 \leqq \mu_1 \leqq 5.29$）が母平均μ（4.5）のラインと交差している、つまり母平均を含んでいることがわかる。これは、母平均の推定に「成功」していること

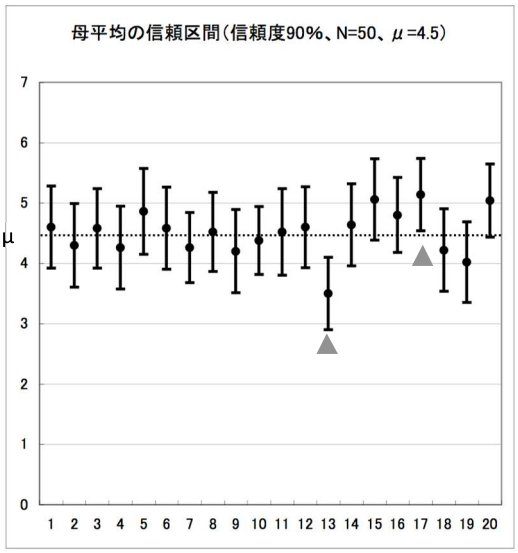
を意味する。

次に、この作業を標本２から標本20まで繰り返して、それら20の信頼区間を表したものが**図表４**である。

図表３



図表４



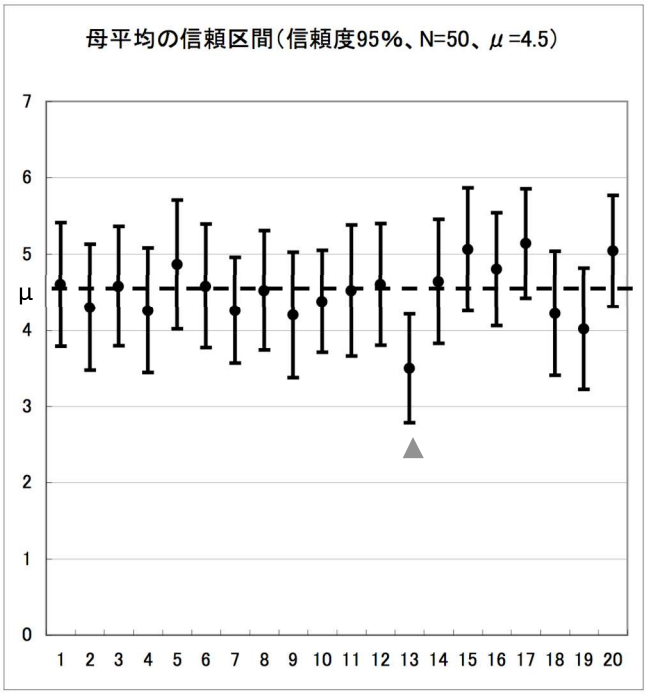
縦線で表されている推定された信頼区間の多くは、母平均μを含んでいることがわかる。例外が、標本13と標本17の２つの信頼区間である。これら２つの標本の推定区間は母平均μを含んでおらず、母平均の推定に「失敗」したわけである。

ポイントは、見誤った標本は全体20のうち２つであったという事実である。確率で言えば、推定の90%は成功して、10%は失敗したのである。この確率こそ、母平均の推定の際にあらかじめ設定した信頼度90%なのである[1]。すなわち、信頼度90%とは、同一の母集団から抽出された標本の「仲間内輪」で母平均を推定した場合、約９割の標本が母平均を言い当てる（１割は言い間違える）ということを意味する。

[1] 常にこのように一致するわけではない。しかし、推定を何回も繰り返し、標本の数を多くすれば、信頼区間に母平均をふくむ確率は、かならず設定した信頼度に近くなる。

最後に、信頼度を95%に上げて、先と同じ母平均の区間推定の作業を、標本１から標本20まで繰り返してみる。結果を図にしたものが**図表５**である。

図表５



区間推定の幅は広がり、その結果、母平均を外す信頼区間が２つから１つに減っている（標本13は依然として母平均を外しているが、標本17が母平均を含むことになった）。このように、信頼度を90%から95%に上げた結果、20の標本のうち19の標本が、母平均の区間推定に成功したことになる。つまり成功率は95%、あらかじめ設定した信頼度に一致したことになる。

４. 仮説検定と反証主義について（本書第13章の補足）

本書第13章でみたように、仮説検定の論法は一見するとややこしい。わたしたちが本来、検証したいと考える仮説と反対の仮説を帰無仮説として立て、それを棄却するか、しないかという形をとるからである（本書第13章2-1参照）。なぜ、このようにややこしいことをするのであろうか。だが、実のところこうした考え方は、科学方法論で有力な立場である反証主義（falsificationism）と同じものである。

反証主義は、イギリスの哲学者カール・ポパー（Karl Popper）によって唱えられた。科学は、基本的に、現象の背後に存在する法則を発見しようという営みである。近代科学の科学方法論として支持されていたのは帰納法にもとづく経験主義（empiricism）ないし実証主義（positivism）であった。たとえば、あのカラスも黒い、このカラスも黒い。そのようにして、非常に多くのカラスを調べてみたら、とりあえず調べた限りのカラスは、みな黒かった。そこから、「カラスの色は黒い」ということができるかもしれない。これが帰納法である。なにげないことのように思えるかもしれない。けれども、これは部分（個々のデータ）から全体を推測するという作業をしているのであり、こうしたことによって科学は推進されていったのである。

しかし、である。たった1羽の白いカラスが見つかったとしよう。そうしたら、もはや「カラスの色は黒い」という命題は成り立たないのではないか。要するにつまり、世の中に存在するすべてのカラスを調べることができないことは明白であり、つねに調べた範囲で、仮説は正しいと主張しているにすぎない、ということである。これでは、得られた知識はきわめて不安定で、とても真理と呼べるようなものではない。このように、帰納法から仮説の正しさを検証するということは、実のところとても難しいのである。対して、反証するということは比較的容易である。したがって、ポパーは、仮説が正しくないということに焦点を当てたのである。かりに、それまで正しいと信じられていた説に反証例が見出されたとするなら、もはやその仮説は正しくない。逆にいえば、反証がなされない限りは、仮説が正しくないとは言えないという消極的な形で、科学が追及する客観的な真理が主張されるのである。

仮説検定もこうした考え方を採用していると言えるだろう。仮説検定の手続きは、帰無仮説を棄却することによって、調査仮説を採択するというものであった。つまり、「調査

仮説が正しくないとは言えない」という形で調査仮説の正しさを主張しているのであり、この論法は「カラスの色は黒くないとは言えない（今のところ黒以外のカラスは発見されていない）」という形で、「カラスの色は黒い」ことを主張する反証主義の論法と一致する。すなわち、データから発見することのできる事実に対して、きわめて慎重にその正しさを主張しているのである。

ところで、本書第13章2－3において、仮説検定には第１種の誤りと第２種の誤りがあるが、どちらかといえば第１種の誤りを極力少なくすることが優先されると述べた。この点についても、説明を加えよう。なぜこのような方針が取られるかといえば、第１種の誤りを犯すことの問題のほうが大きいと考えられるからである。第２種の誤りを犯したとするならば、それは変数間に関係があるのにそれを見失ってしまったということである。つまり、調査仮説が支持されるのに、誤ってそれを捨ててしまったということである。むろん、これも望ましいことでは決してない。だが、第１種の誤りを犯したとすればどうだろう。本来何の関連もないものに、なんらかの関係があると見誤ってしまうことになる。つまり、間違っている調査仮説を正しいと主張してしまうことになるのである。たいていの場合、調査仮説として設定することは重要なことであるはずなので、これを誤って支持してしまうことのほうが、問題はより大きくなる可能性が高いであろう。したがって、基本的には、第１種の誤りを犯さないことが優先されるのである。