日本世論調査協会報

第67号

平成3年3月

日本世論調査協会

◇世論調査に関する文献紹介

(1) 統計的研究における パワー分析の勧め

How Many Subjects? Statistical Power Analysis in Research
Helena Chmura Kraemer and Sue Thiemann著 (Sage Publications 1987)

西澤由隆（明治学院大学）

「さて、サンプル数はいくらにしましょうか。」
「最近、調査環境がずいぶん厳しくなってきましたから・・・」
「回収率を考慮して、できれば有効数を1,000 にしたいところなんですが・・」
「いや、今回の研究助成金の額では、『サンプル総数 1,000』でお願いするのが精いっぱいではないですかね。」

＜サンプル数は大きい方がよい？＞

世論調査を調査会社に依頼する前に、依頼者として決定しておかなければならないことがいくつかある。その１つが調査の規模、つまりサンプル数である。調査を１度でも企画・実施した人なら、そのサンプル数の決定にあたっての上のような会話をだれしも経験したことだろうし、実際多くの場合このようにしてサンプル数は決まっていくようである。

驚くべきことは、――調査の経験がまだそれほどない私がこういうことを申し上げるのは多少気が引けるが――このような場面では「経験」や「相場」がずいぶん幅を利かせていることである。調査の目的、理論仮説についての議論や質問項目の選別・質問文のワーディングなどの質問票の設計にはかなりの時間を割いて準備が進められるのに、ことサンプル数の決定となると、理論的根拠より経済的な理由によって「簡単に」決められてしまう場合が実に多い。

分析する側の立場からいうと、「サンプル数は大きい方がよい」という心理が働く。サンプル数が十分でないために、仮説検証が思うようにできないという事態をできるかぎり避けたいからである。この事態が発生すると、たとえば、「理論的にはＸとＹには関係があると考えられるが、このデータからは必ずしもそれが確認できなかった」と、なんとも歯がゆい結論しか導けない。仮説に

自信があればあるほど、またその社会的・学問的意義が大きいときほど、サンプル不足によって検証が出来ないことによる「失意」は大きい。

ところが、調査のコストはサンプル数とおよそ比例して大きくなるので、無制限にサンプル数を大きくするわけにはいかない。したがって、予算の許すかぎりで最大限のサンプルを取る。つまり、予算の方が潤沢で、そちらのほうから上限が設定されることがない場合を除いて（そして、一般の研究者にとっては、そんなことは滅多にないことだが）、サンプル数についてはたいした議論の余地がないわけである。

＜「予算」が決めるサンプル数の２つの無駄＞

実は、ある特定な状況において、ある仮説を検証するために必要なサンプル数は理論的に決まる。にもかかわらず、その点の議論と検討を省略して、「予算の許す限り」という基準にしたがう場合、２つの「無駄」を発生させる危険性がある。１つは、研究の目的である仮説の検定に必要なサンプル数より大きなサンプルをとることの無駄である。そのデータをその他の目的で利用する時により多くのサンプルが必要になることも考えられるので、サンプル数の「余裕」は必ずしも無駄ではないかも知れないが、限られた資源の有効利用という観点からは無視できない問題である。第２の無駄はもっと深刻な問題である。予算が許すサンプル数が理論的に必要なサンプル数よりはるかに小さく、満足な結果の得られる可能性がほとんどないのに、調査に踏み込んでしまう危険である。あらかじめこの事態を予測できれば、取り返しのつかない無駄をするまえに、たとえば質問項目を減らしてでもサンプル数を増やす、あるいは世論調査ではなく実験的手法によるとか、仮説検定の方法について再考することもできるわけである。

＜「統計的研究におけるパワー分析」＞

では、合理的なサンプル数はどうして決まるのか。この問いに答えてくれるのが、ここで紹介するクレーマーとシーマンのHow Many Subjects? Statistic-al Power Analysis in Researchである。

その副題に「統計的研究におけるパワー分析」とあるように、本書の中心的概念はパワー(Power) である。ある特定な仮説が正しいときに、それが正しく

ないとする帰無仮説を棄却し、その対立仮説を採用することができる確率をパワーと呼ぶ。対立仮説を採用することが研究の具体的な課題であるから、当然のことながら、研究者としては研究デザインの「パワー＝アップ」を図りたい。前述したような調査研究の無駄を省き、効率よくパワー＝アップするには、研究デザインの設計の早い段階で、このパワーについて検討することがいかに大切であるかを、クレーマーとシーマンは具体例を交えて力説する。

＜パワー＝レベルを規定する３つの要因＞

パワー＝レベルは次の３つの要因によって決まる。

1) サンプル数(n)

2) 帰無仮説を棄却するレベル――つまり「有意水準」（α）

3) 検出が期待される効果の大きさ(d)

3)の「検出が期待される効果の大きさ」とは、おおむね次のようなことを指す。つまり、いま検証しようとしている仮説が正しい場合に、それが、どの程度の効果として表れるかという問題である。たとえば、ある工場の作業行程を改善すると作業の能率が上がると期待されるとするときに、いったい単位時間あたりの生産量を具体的にどれだけ上げることができるかということである。

そこで、パワーをＰとすると、

$$P = f(n, \alpha, d)$$

と、表すことができる。そして、サンプル数が大きくなれば、それだけパワーも増加する。一方、帰無仮説を棄却することのできるレベルを厳しくするとパワーは下がるので、パワーを一定のレベルに保つためには、サンプル数を大きくして補強しなければならないが、検出が期待される効果のレベルが大きい場合はサンプル数が少なくてすむ。これらの点については、以下でもう少し詳しく見てみることにしよう。

＜機械的に決まるサンプル数＞

さて、今われわれが関心のあるのは、P, α, dの３つの要因についてそれぞれ値を設定したときに、ｎをどのようにして決めるかという問題である。そこで、上記のＰについての関係式をｎについて整理すればよい。すると、

$$n = f(P, \alpha, d)$$

となる。さきに、サンプル数が理論的に決るといったのは、このことである。

ところでパワーと3要因とのこの関係についてのクレーマーとシーマンの指摘は、特に新しいものではない。たとえば、本書の基礎となっているJ.コーヘンの『行動科学における統計手法のパワー分析』は1969年に出版されている(Cohen 1969)。にもかかわらず、クレーマーとシーマンによる本書が画期的な点は、サンプル数算出の作業を単純化したことである。というのは、n についての上記の関係式の要素の1つであるdは、検定に用いる統計手法によって数学的な定義が異なる。つまり、nを求めるための換算表を統計手法ごとに用意しなければならないわけである。実際、400 ページにおよぶコーヘンの前掲書も、異なった形式の換算表があちこちに掲載されていて、ずいぶん使いづらいものになっている。*

そこで、クレーマーとシーマンは、統計手法ごとに使用する換算表が異なるという不都合を解消するために、共通の換算表(Master Table)を用意し、逆にdの方を換算表に対応させるように工夫したわけである。そしてそれが、Master Tableの直前に掲載されているSummary Table に、統計手法ごとに整理されている。おかげで、統計理論に不慣れな者にも、少しの練習でいろいろな状況に幅広くパワー分析ができるようになっている。

クレーマーとシーマンに従えば、サンプル数決定の手順は機械的である。有意水準を5%とするか1%にするか、さらに検定の条件が両側であるか片側であるかによって、Master Tableが4つに区分されているので、まず、そのいずれを使うかを選択する。そして、その換算表の縦・横の記載された「保障したいパワー＝レベル (P)」と「期待できる効果の差（具体的には、dではなく、それを Summary Tableの指示にしたがって修正した値; Δ）」の2つの値が交差する点の値（ν）を読む。最後に、もう一度Summary Table に戻り、指示されるとうりに自由度についての修正をした値が求めるサンプル数となる。これほど簡単な作業なら、調査を実施するまえに、パワーの視点から「必要サンプル数」を試算してみないほうはないと思う。

さて、読者のみなさんがどう判断されるだろうか。本書を手にとってみる価値があるかどうかの判断材料を提供する意味で、ここに具体的な利用例を紹介しようと思う。ただし、パワーの概念は、日本ではあまり知られていないうえに、それについての本書の説明も必ずしも十分でないので、上に挙げた3つの

要素とのパワーの関連を簡単に補足しておこう（なお、「仮説検定におけるタイプⅡのエラーを1から引いた値（$1-\beta$）がパワーである。」という定義だけで、その概念が理解しただける読者は、つづく3つの節を読みとばして進んでいただければと思う）。

＜仮説検定の手順＞

いま仮に、「教育程度は政治的関心に影響をおよぼす」という命題について検討したいという状況を想定してみよう。伝統的な仮説検定の手順は次のとうりである。

この仮説を実証するために、全国サンプルの世論調査を実施する。すると、選挙に関心を持ったと答えた回答者の平均就学年数が11.7年で、関心を持たなかったと答えた人の平均が11.2年というデータが得られた。グループ間の平均値の差は 0.5年、つまり半年ということになる。

そこで、「教育程度と政治的関心には関係がない」という帰無仮説(H_0)をたて、それが正しいときに、この半年の差が偶然発生する確率を求める。それが十分に小さければ（たとえば、1％以下）、帰無仮説が間違っていたと解するほうが妥当なので、それを棄却し、「教育と政治的関心には関連がある」との対立仮説を採用する。半年の差が偶然発生する確率が大きければ、帰無仮説は棄却できず、「教育と政治的関心には関係が認められなかった」と結論する。

＜仮説検定において犯す可能性のある2つのエラーと「パワー」＞

上のような手順で進められる仮説検定には、犯す可能性のあるエラーが2つある。1つは、帰無仮説が正しいにもかかわらず、それを誤って棄却してしまう誤りで、これをタイプⅠのエラー（α）と呼ぶ。たとえば、無罪である容疑者を誤って有罪としたり、重大な副作用のある新薬を誤って認可するような場合で、これはできるだけ避けたいエラーである。したがって、帰無仮説を棄却することを

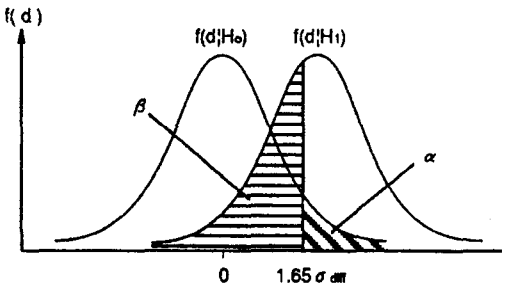

図1　仮説検定における2つのエラー

許すレベルは、一般にかなり厳しく設定されている。社会科学では、このαを5％や1％にするのが、慣例となっている。

これを確率分布として図解すると図1のようになる。先の教育程度と政治的関心の例を用いるならば、Hoについての左の山が、「関連がない」との仮定が正しい場合の両サンプル間の平均就学年数の差の分布をあらわす。とうぜん、ゼロを中心にした分布となっている。そしてその山の斜線の部分が有意水準を5％に設定したときのタイプⅠのエラー（α）である。標本平均の差が、σdiffの1.65倍（境界値）より偶然大きくなったとしても、帰無仮定は棄却され、本来は平均値に差が無いのにもかかわらず、誤って「ある」と結論してしまう確率である。

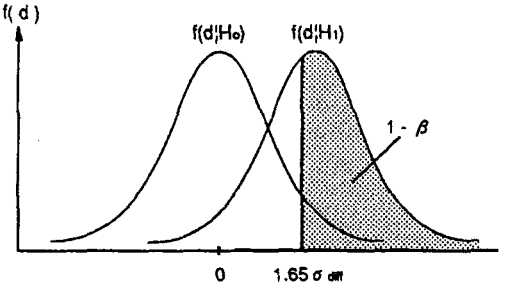

図2-1 確率分布にみるパワーの定義

一方、ほんとうは対立仮説が正しく、帰無仮説を棄却することができるのにもかかわらず、棄却しない誤りがある。タイプⅡのエラー（β）と呼ばれるこの誤りは、たとえそれを犯したとしても現状肯定ということになり、社会的影響という観点からはタイプⅠのエラーに比べて害が少ない。もっとも、研究者の立場からすると、せっかくの理論的「発見」を裏付ける機会を逸するわけで、また、多くの人命を救う可能性のある新薬を使用できずにいるということであれば、大きな社会的損失ということにもなる。

図1では、H_1の山が、標本平均の差の標準誤差（σdiff）の2倍に当たるだけの差が両グループ間にはあるとの仮定のもとでの差の分布の状態をあらわしている。そしてその横線の部分がタイプⅡのエラー（β）にあたる。これは、本来は平均値に差があるのにもかかわらず、たまたまデータとしして手に入れたサンプルの平均値が境界値より小さな値を示したために、誤って帰無仮説を採用してしまう確率である。

さて、問題のパワーは、すでに紹介したとうり、対立仮説を誤ることなく採用する確率だが、図1の右の山から横線の部分を除いたところ、あるいは、図2-1の陰のかかっている部分になる。f(d : H1)で囲まれた部分の総面積は1なので、陰の部分は1-βとなる。

＜パワーを規定する3要素との関係 ** ＞

さて、研究デザインのパワーを上げるためには、サンプル数を増やすこと・有意水準下げること・検出が予測される効果をより大きく設定することであるとすでに指摘した。その点を図2-1 との比較で図示するとその関連がわかりやすい。

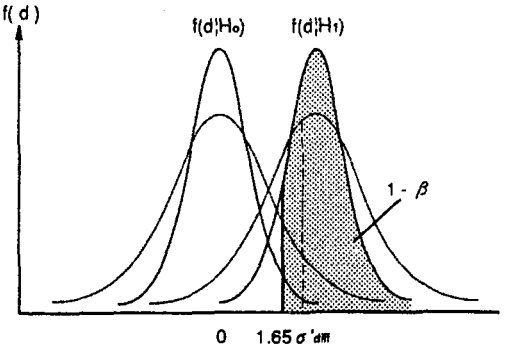

図2-2

たとえば、図2-2 はサンプル数を増やした場合である。サンプル数を増やすと平均値の差についての標準誤差が小さくなる。その分だけ境界値が左に移動し、1-βが大きくなる。図2-3 は、有意水準を5％から10％に下げることにより、境界値をσdiffの1.65倍から1.28倍の位置に移動させた場合である。もっとも簡単な操作だが、先にも触れたように、αを大きくすることについては慎重でなければならない。図2-4 は、政治的関心の高いグループとそうでないグループの間には、「より大きい標本平均の差（σdiffの2倍ではなく3倍の差）が期待できる」と効果のレベルを高くした場合である。境界値はそのままでも、H_1の山だけがσdiffひとつ分だけ右に移動するため、その分1-βが大きくなる。いずれも、図2-1 の陰の部分の面積を大きくする、あるいは、対立仮説の中央値に対する境界値の相対的位置を左に移動させるような措置だということができる。

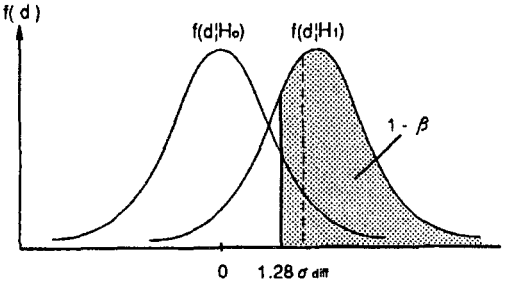

図2-3

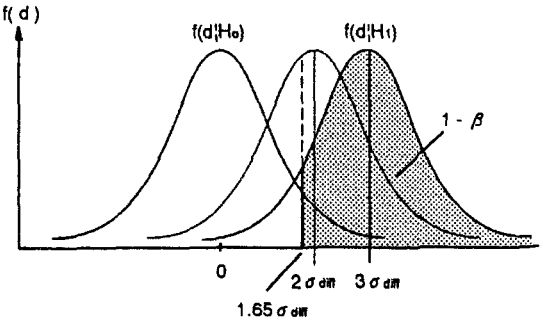

図2-4

ここでは、手元にある全国サンプルの世論調査の結果を参考に、

$$\sigma^2 = 5,\ p/q = 0.2/0.8$$

と設定することにする。

これらの値を、図3にあるΔについての公式に代入すると、Δ＝0.089 を得ることができる。

そこで、次に、Master Tableの片側検定／1％有意水準に対応する部分（P.107)を開ける。P=90、Δ=0.09 に対応するｖ値は、1600である（図4参照）。最後に、もう1度Summary Table のｖの項にしたがって、自由度について修正する。

$$n = v + 2 = 1,602$$

これが、上記のような条件で調査を実施しようとするときの、理論的に必要とされるサンプル数である。かなり、大きなサンプル数が必要だということになる。ちなみに、他の条件は変えずに効果の差だけを「半年」から「2年」に変えるとどうなるか。実は、Δが0.34になり、必要サンプル数は 103と、飛躍的に小さくなる。

SUMMARY TABLE

TEST	*SPECIFICATION*	H_0	ν	Δ	*SECTION*
Single-Sample Normal Test	$X_i \sim N(\mu, \sigma^2)$ $i = 1, 2, \ldots, n$ σ^2 *known*	$\mu = \mu_0$	$n + 1$	$\Delta = (e^{2\delta} - 1)/(e^{2\delta} + 1)$ $\delta = (\mu - \mu_0)/\sigma$	4.1
Single-Sample t Matched Pair t	$X_i \sim N(\mu, \sigma^2)$ $i = 1, 2, \ldots, n$	$\mu = \mu_0$	$n - 1$	$\Delta = \delta/(\delta^2 + 1)^{1/2}$ $\delta = (\mu - \mu_0)/\sigma$	4.2
Two-Sample t	$X_i \sim N(\mu_x, \sigma^2)$ $i = 1, 2, \ldots, np$ $Y_i \sim N(\mu_y, \sigma^2)$ $i = 1, 2, \ldots, nq$ $p + q = 1$	$\mu_x = \mu_y$	$n - 2$	$\Delta = \delta/(\delta^2 + 1/pq)^{1/2}$ $\delta = (\mu_x - \mu_y)/\sigma$	4.3
Intraclass ρ	$(x_i y_i)$ Bivariate Normal corr $(X_i, Y_i) = \rho$	$\rho = \rho_0$	$n - 1$	$\Delta = (\rho - \rho_0)/(1 - \rho\rho_0)$	5.1

(continued)

図3　Summary Table （Kraemer and Thiemann 1987 p.101より転写）

＜Summary Table とMaster Tableの利用例＞

ではここで、「教育程度と政治的関心」の例を用いて、クレーマーとシーマンのサンプル数算出の手順を紹介しよう。

検討したい仮説は、「教育程度が高いほど、政治的関心が高い」であるとしよう。具体的には、政治的関心の高いグループとそうでないグループの平均就学年数には差があるということを、統計的に裏付けようというのである。これは、サンプル間の平均値の差についての検定であるので、用いる統計手法はT-検定である。また、一方のグループの方が平均値が高いと、あらかじめ仮定しているので、これは片側検定になる。

そして、パワーを90%とする。多少高望みの傾向があるが、相当な費用をかけて調査を実施するのであるから、すこしは高目に設定しておきたい。さらに、有意水準は１%とする。そして最後に、期待できる平均就学年数の差をとりあえず半年とすることにした。***

ここまでを整理すると次のようになる。

検定方法：T-検定
検定の方向性：片側検定
パワー（P）：90%
有意水準（α）：１%
効果の差（$d = \mu x - \mu y$）：0.5(ただし、μx ・μy はそれぞれ関心の高いグループの就学年数の標本平均)

Master Tableを開ける前に、Δを求める必要がある。そこで、さっそくSummary Table のT-検定の部分(p.101) を見てみよう。Two-Sample tの項目がそれにあたる（図３参照）。

T-検定の場合は、上に整理した情報の他に、就学年数の平均値の分散（σ^2）と関心の高いグループと低いグループの２グループに、総サンプル数がどのように分配されているかを示す分配比（p/q:ただしp+q=1)を知る必要がある。それぞれ、調査を実施するまでわからない情報で、実際には、予備調査によるデータか過去の調査データなどを参考にするか、あるいは研究者の「知的な推測(educated guess)」によらざるを得ない。

1 % Level, One-Tailed Test

	POWER										
Δ	99	95	90	80	70	60	50	40	30	20	10
0.01	216463	157695	130162	100355	81264	66545	54117	42972	32469	22044	10917
0.02	54106	39417	32535	25085	20313	16634	13528	10742	8117	5511	2730
0.03	24040	17514	14456	11146	9026	7391	6011	4773	3607	2449	1214
0.04	13517	9848	8128	6267	5075	4156	3380	2684	2029	1378	683
0.05	8646	6299	5200	4009	3247	2659	2163	1718	1298	882	437
0.06	6000	4372	3609	2783	2254	1846	1501	1192	901	612	304
0.07	4405	3209	2649	2043	1655	1355	1102	876	662	450	224
0.08	3369	2455	2027	1563	1266	1037	843	670	507	344	171
0.09	2660	1938	1600	1234	999	819	666	529	400	272	136
0.10	2152	1568	1295	998	809	663	539	428	324	220	110
0.11	1776	1294	1069	824	668	547	445	354	268	182	91
0.12	1490	1086	897	692	560	459	374	297	225	153	77
0.13	1268	924	763	589	477	391	318	253	191	130	65
0.14	1092	796	657	507	411	337	274	218	165	112	56
0.15	949	692	571	441	357	293	238	190	144	98	49
0.16	833	607	501	387	314	257	209	1661	126	86	43
0.17	736	537	443	342	277	277	185	147	112	76	39
0.18	655	478	395	305	247	202	165	131	100	68	3
0.19	587	428	353	273	221	181	148	118	89	61	
0.20	528	385	318	246	199	163	133	106	81	55	

(continued)

図4 Master Table (Kraemer and Thiemann 1987 p.107 より転写)

＜パワー分析の勧め＞

さて、120 ページ足らずの本書だが、もっとも重要な部分は巻末の12ページに納められた2つの表ということになる。1章から3章までの概念説明と換算表についての論理的根拠の説明の部分を除くと、その大部分が換算方法の具体的な説明に当てられている。だからといって、本書が単なるサンプル数算定のためのマニュアルにすぎないと、過少評価するわけにはいかない。具体的な説明の中で、研究デザインについての貴重な助言が随所に見られる。なかでも、重要なポイントはパワーについての検討が、研究そのものの意義を再考する機会を与えてくれることだろう。

ここで取り上げた、教育程度と政治的関心の2変数間の関連性の問題であるが、1,600 のサンプルをとれば、かなりの確率で「半年」の差を立証することができることがわかった。ところで、この「半年」の差とは実質的にはどのような意味を持っているのだろうか。「高校を1学期分よけいに行くことで、政治的関心が高くなることが、統計的に立証された」と主張したところで、耳を傾ける人はいないだろう。これが2年分の学校教育の効果とでもなれば、多少は説得力が増すかも知れない。ところが、上の例で見た通り、「2年の差」を立証するのであれば、100 あまりのサンプルですむ。

実は、ここで用いた「半年の差」という値は、私が手元に持っている全国サンプルの世論調査のデータの分析の結果でてきた平均値の差である。「今回の選挙に興味を持った」と答えた人の方が「持たなかった」と答えた人より半年分就学年数が長い。そして、この調査サンプルは、有効数が 2,000を超えているので、１％の有意水準でもなるほどその差が認められた。一方、もしこれが 100そこそこのサンプル数の調査であったとしたら、それが有意な差であることを確認できなかったことだろう。ただし「１学期分の差が認められた」と主張することと、「差が確認されなかった」と結論することと、実質的差はないと考えられる。

私がここで紹介した「効果の差」の定義は、クレーマーとシーマンのそれとは少し違う。「社会が『重要である』と評価するに足る最低限の効果」が、彼女達の定義である(p.24)。社会科学を「実践」する者への貴重なメッセージだと、わたしは思う。

＜注＞

* コーヘンのこの本は、1977年に改訂版がでている。残念ながら、脱稿の時点までにその改訂版を手にすることができなかった。ただし、クレーマーとシーマンのこの改訂版についてのコメントから類推すると、繁雑であることには変わりがないようである。

** ここでの議論はWinkler & Hays (1971) に負うところが大きい。

*** なぜ半年としたかは、後述する。

＜参考文献＞

Cohen, J. (1969). Statistical Power Analysis for the Behavioral Sciences. New York: Academic Press.

Winkler, R. and Hays, W. (1971). Statistics: Probability, Inference, and Decision (2nd. ed.). New York: Holt, Rinehart and Winston.