大規模サンプルに対する一般化χ²適合度検定
― JGSS データへの適用例 ―

保 田 時 男

（甲子園大学人間文化学部）

The Generalized χ^2 Goodness-of-fit Test for Large-scale Sample

— Application to the data of JGSS —

Tokio YASUDA

Large-scale sample surveys cause the practical problem that any statistical tests for the survey data must reject null hypotheses easily. If sociologists receive the results of the statistical tests with no consideration, they should make an error to regard trivial characters of the population as important features. This paper applies 'the generalized χ^2 goodness-of-fit test' to the data of JGSS-2002, and show one of the methods to solve the problem. The generalized χ^2 goodness-of-fit test was developed in medical data analyses, but it should be useful in sociological data analyses. Because the method can test the null hypothesis that non-zero amount of lack of fit is present, and check whether there is 'the large amount' of divergence from a model or not. The result of the application of the method to JGSS-2002 data was very good, because the method could clarify whether the large amount of divergence was present. If we can establish the appropriate amount of divergence for null hypotheses, the generalized χ^2 goodness-of-fit test will be a very useful instrument for the analyses of large-scale samples.

Key words: JGSS, χ^2-test, large-scale sample

大規模サンプルの社会調査は、どのような統計的検定を行っても簡単に帰無仮説を棄却してしまうという実際的な問題を引き起こす。社会学者がこの検定の結果を何の思慮もなしに受け入れてしまうと、母集団の些細な特徴を重要な特徴とみなしてしまう可能性がある。この論文は、一般化χ²適合度検定と呼ばれる方法を JGSS-2002 データに適用し、この問題に対する 1 つの解決策を提示する。一般化χ²適合度検定は、医療データ分析の中で開発された方法であるが、社会学的なデータ分析にとっても有用なものである。というのは、この方法は不適合の量を 0 以外とする帰無仮説を検定することができるので、モデルからの〈大きな〉逸脱があるかどうかを確認にすることができるからである。JGSS データへの適用結果は非常に良好であり、大きな逸脱があるかどうかを明確にすることができた。帰無仮説の適切な逸脱量を定めることができれば、一般化χ²適合度検定は、大規模サンプルの分析にとって、非常に有効な道具となるであろう。

キーワード：JGSS、χ²検定、大規模サンプル

１．大規模サンプルの検定問題

無作為抽出された標本の調査データを分析するとき、分析者は単にその集計結果を提示するだけではなく、ほとんどいつでも統計的検定の結果を合わせて提示する。分析の結果分かった事実が標本集団について当てはまるだけではなく、母集団についても同様の傾向を読み取ることができることを示すためである。分析者は標本集団ではなく母集団について何らかの知見を得るために標本データを分析しているのであるから、統計的検定の手続きを取るのは当然である。

しかし、昨今のようにそのサンプル・サイズが大きくなった調査のデータに対して統計的検定を行う際には、共通して 1 つの問題がつきまとう。それは、サンプル・サイズが大きければ大きいほど、どんな検定を行っても有意という結果が出やすくなるという問題である。例えば、等しいサイズの 2 つのグループ間で、ある意見への賛成率が 20%と 22%というように 2%しか違わないとしても、7,000 サンプルという大規模サンプルでは、その差が 5%水準で有意となる（df＝1、χ^2＝4.22）。回帰分析では多くの独立変数の効果が有意になりやすくなるし、いくつかのモデルの適合度を調べる場合には、複雑な相互作用が有意性を示しやすくなるので、単純なモデルは適合しにくくなる。

ただ、統計的検定の仕組みから考えれば、サンプル数が大きくなるほど検定の結果が有意になりやすいのは当然のことであり、検定の結果は正しい知見を示しているにすぎない。例えば、先ほどの比率の差の検定では、「母集団では 2 つのグループ間でまったく差がないこと」を帰無仮説とするので、標本の分析から〈少しでも〉差があることを一定の確からしさで言えるのであれば、それは当然、統計的に有意となる。サンプル・サイズが大きくなれば、標本グループ間のわずかな違いからでも、母集団についての帰無仮説を棄却するのに十分な情報を得ることができるからである。したがって、検定の結果はまったく正しい情報を示しているにすぎない。もちろん、非標本誤差のせいで起こりえない規模の誤差が発生し、検定の結果を惑わしている場合（Fitzmaurice, 1997 などを参照）は、話が別であるが、ここではその問題を考えているわけではない。完全な無作為抽出ができていると仮定するならば、分析者が統計的検定の仕組みを正しく理解している限り、そこから誤解が生じる危険はないのである。

しかしながら、それでもやはり、大規模サンプルに対する統計的検定が有意になりやすいことに、私はしばしば実際的な問題を感じる。それは 2 つの理由による。第 1 に、現実的に意味がない程度の傾向しか読み取れない分析結果（例えば、出身地によって年間平均収入が¥500 だけ異なるとか、職業によって生みたい子どもの平均数が 0.01 人だけ異なるというような場合）についても、検定の結果が有意になり、しかも全体的な論旨の都合上、どうしてもその検定結果を示すことを省けないことがある。このとき、「検定の結果は有意であるが、実際的には意味のない程度の差しかなく……」といった断り書きを書く必要が生じ、論旨の明確さが損なわれてしまう。第 2 に、検定の結果は別にして、示されてい

る傾向が実際的に意味のある程度のものかどうかを、分析者が確認しないといけないとするならば、何のために検定の手続きを行っているのかが分からないことになる。統計的検定が、客観的、形式的に母集団についての推測を行うための道具として十分に働いていないことに不満が感じられる。

この論文は、以上のような問題に対する 1 つの解決の試みである。この論文では、特に χ^2 統計量を利用した適合度検定（単純な独立性の検定や比率の差の検定を含む）に焦点を絞り、医療統計の分野から提案されている一般化 χ^2 適合度検定（the generalized χ^2 goodness-of-fit test）と呼ばれる方法が、大規模サンプルの社会調査データの検定に対しても、極めて有効であることを示す。全体の構成は以下のとおりである。次の節では、特に社会学的な仮説を検定する場合に、大規模サンプルが持つ問題性について整理する。第 3 節では、一般化 χ^2 適合度検定の考え方および手続きを概説する。そして第 4 節で、JGSS-2002 のデータを利用して、その方法の実際的な適用例を示す。最後の節では、一般化 χ^2 適合度検定の有効性と問題点を整理する。

２．社会学的仮説の検定に特有の問題性

大規模サンプルの検定が有意になりやすいことは、どのような標本データでも同じことであるが、社会学的な仮説を検証するために集められた社会調査データに対して検定を行う場合には、そこに特有の問題が発生する。この特有の問題性は、一般化 χ^2 適合度検定を導入する理由と密接に関係しているので、まずその問題性について説明する。

社会学者が検証しようとしている仮説の内容と、統計的検定によって検証される仮説の内容には、根本的に大きな齟齬がある。先にも記したように、統計的検定の帰無仮説は、一般に、平均の差が 0 であるとか、2 変数間が独立（関連性が 0）であるという具合に、母集団について何らかの傾向が〈まったくない〉という仮説である。逆に言えば、検定の結果が有意であるということは、平均の差が 0 ではないとか、2 変数間の関連性が 0 ではないという具合に、母集団について何らかの傾向が〈少しはある〉ということを実証したことになる。

しかし、社会学者が検証したいことは、平均の差が〈少しはある〉といったことではなく、その差がある程度〈大きくある〉ということのはずである。程度の問題という面はあるが、一般に自然科学に比べて、社会科学の統計分析は変数間の関連性の有無を厳密に特定することを目指してはいない。社会科学の中でも特に社会学にはその傾向が強いと言える。大きく社会を左右している仕組みがどこにあるのかを発見しようとしているからである。もし仮に、社会的な変数間の関連性の有無を厳密に調べるならば、ほとんどどの変数の間にも関連性が発見できてしまうに違いない。

では、社会学者が検証したい仮説と統計的検定によって検証される仮説の間に、このような齟齬があるにもかかわらず、それが大きな問題とならないのはなぜであろうか。サン

プル・サイズがその鍵となっている。

いま、2 つのグループ間で何らかの社会的属性を表す得点の平均値を比較しているとする。2 つのグループとも標本の数は同じで、それぞれ標準偏差が 15 であるとする。この条件の下で、標本グループ間に何点以上の平均の差があれば、検定の結果が 5%水準で有意になるかは、サンプル・サイズだけに依存する。例えば、総サンプル数が 250 ならば、標本グループ間の平均の差が $1.96\times\sqrt{(15^2+15^2)/125}$　3.72 あれば、検定の結果、有意と判定される。いま、標本グループ間の平均の差が、このぎりぎりの臨界値を取っていると仮定する。このとき、95%の信頼度で母集団の平均の差を区間推定すると、その信頼区間は（0, 7.44）である。さまざまなサンプル・サイズについて、同様に、平均の差が有意になる臨界値を求め、標本平均の差が臨界値を取るときの信頼区間を算出した結果をまとめたものが図 1 である。

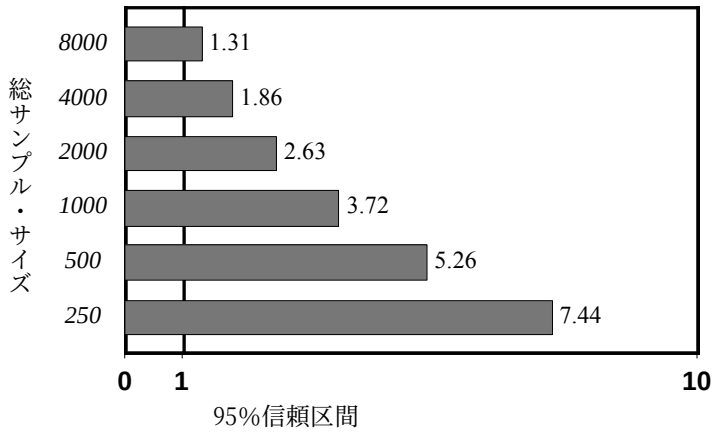

図１　サンプル・サイズと区間推定の関係

この図から、検定の結果と区間推定の間に存在する暗黙の関係が、そのサンプル・サイズに依存していることが分かる。総サンプル数が 250 と小さい場合、検定結果が有意であったならば、それはぎりぎりの水準で有意であったとしても、母集団について大きな平均の差が存在する可能性が高いことを表している。なぜならば、小さく見積もったとしても（0, 7.44）の間に真の平均の差があると推定できることが、暗に示されているからである。一方、総サンプル数が 8,000 と大きい場合、それがぎりぎりの水準で有意になったとするならば、真の平均の差は（0, 1.31）しかないことになる。

いま仮に、母集団の平均の差が 1 以上あれば、分析者にとって意味のある発見であるとしよう。つまり、分析を行っている社会学者は、この平均の差が 1 以上あれば、それを社会的に意味のある程度の〈大きな〉差とはみなし、1 未満であれば〈大きな〉差とはみなさないとする。もし、総サンプル数が少なく 250 や 500 しかない調査データにおいて、統

計的検定の結果が有意になったのであれば、その検定結果は高い確率で〈大きな〉差があることを暗に示している。つまり、統計的検定の手続きを取れば、分析者が本来持っている仮説との間に齟齬があるにもかかわらず、問題なく分析者の仮説を検証することができる。一方、総サンプル数が 4,000 も 8,000 もある大規模サンプルの調査データのおいて、統計的検定の結果が有意になったのであれば、それは〈大きな〉平均の差があることを必ずしも意味していない。つまり、統計的検定による検証は、分析者が本来持っている仮説を検証していることにはならない。

このように、統計的検定の手続きで検証している仮説と、社会学者が検証しようとしている仮説の間に根本的な齟齬があるにもかかわらず、それが大きな問題とならないのは、社会調査のサンプル・サイズが適切な小ささを持つという条件が満たされている場合に限られる。そして、幸か不幸か、この条件が犯される危険性は近年になるまでほとんどありえなかった。(日本の)社会学者が大規模サンプルの社会調査を企画し、そのデータを分析する機会が非常に少なかったからである。

しかし、近年その状況は大きく変化し、この前提条件はもはやまったく保証されていない。多くの分析者の間での共同利用を目的とした、公開データの収集が本格化してきたからである(佐藤ほか編, 2000 などを参照)。個人研究者や少人数の研究者グループが自分たちの研究を進めるためだけに企画する調査は、当然、有効な分析結果を出すために必要な最小限のサンプル・サイズで収集される。これに対して、どのように利用されるかが特定されていない共有データのためには、比較的大きなサンプル・サイズの調査が企画される。日本版 General Social Surveys(JGSS)では、各年 3,000 弱のサンプルについてデータが回収されているし、全国家族調査(NFRJ)においては、7,000 に近いサンプルが回収されている。実際の分析では分析の課題により対象が限定されることがあるにしても、これらは十分に大きすぎるサンプル・サイズと言ってよい。

もし、小規模サンプルと同じ感覚で、これら大規模サンプルのデータを分析し、統計的検定の結果を社会学的な仮説の検証のために用いたとするならば、大きな誤解を生む恐れがある。そのため、この齟齬を埋めるための何らかの解決を模索することは、緊急の課題である。この問題は、推測統計法そのものが抱える問題ではなく、社会学的仮説を検証する必要がある社会学者が抱える問題であり、我々社会学者が積極的にその解決に当たらなければならない。

３．一般化 $\chi 2$ 適合度検定による解決

3.1　間に合わせの解決の問題性と根本的な解決の必要性

この問題に対処するための方法は、いろいろとありうる。1 つの有効な解決方法は、統計的検定の結果を偏重する分析姿勢を改め、区間推定の結果を積極的に示すことである。しかし、仮説の真偽を二者択一で表すという手続きの明確性が損なわれ、煩雑になる恐れ

がある。また、独立性の検定のように、母集団のある特定の統計量を推定することを直接的な目的としていないような場合には、そのような方法で対処することが難しい。別の方法として、あえて全サンプルの中から少数のサンプルを抽出し、小規模サンプルのデータに対して分析を行うやり方がありうる。しかし、この方法は、せっかく得られている情報を十分に活用していないので、無駄が大きい。

実際にもっとも頻繁に用いられている解決方法は、有意水準をより厳しくすることであろう。有意水準を 5%とするのではなく、1%、あるいは 0.1%として、検定を行えば、帰無仮説は棄却されにくくなる。したがって、大規模サンプルのデータに対する検定でも、検定の結果が有意であることが、ある程度〈大きな〉傾向が母集団に存在することを暗示することになる。しかし、この方法を認めるならば、大規模サンプルの調査データについては恣意的に分析の結果を左右できることになってしまう。有意性を示したくない検定については、有意水準を厳しくすればよく、有意性を示したい場合は、有意水準をゆるくすればよいからである。また、社会調査はデータ収集の過程で発生する非標本誤差を厳しく統制することができないので、1%水準や 0.1%水準のように分布の末端に棄却域を設定して検定が行えるほど、分布の近似性が確保されているとは信じがたい。以上の理由から、有意水準を厳しくするという頻繁に用いられている解決方法を、私は奨めることができない。

より根本的に、一般的な統計的検定で検証されている仮説と、分析者が本来検証しようとしている社会学的な仮説との間に存在する齟齬を埋めることができれば、望ましい。分析者が持っている仮説を変えることは本末転倒であるから、齟齬を埋めるためには、統計的検定で検証される仮説に修正を加えなければならない。つまり、母集団について何らかの傾向（グループ間の平均の差、2 変数間の関連など）が〈まったくない〉という仮説を帰無仮説とするのではなく、社会学者が意味を見出すほどの〈大きな〉傾向がないという仮説を帰無仮説として、検定を行うことができれば、根本的な解決となる。この論文で適用例を示す一般化 χ^2 適合度検定は、そのような根本的な解決を助けてくれる 1 つの有用な道具である。

3.2 一般化 χ^2 適合度検定の方法

この節では、一般化 χ^2 適合度検定の考え方と手続きを簡単に説明する。より詳しくは、McLaren et al.（1994）の論稿を参照していただきたい。ここまで、統計的検定全般について議論を進めてきたが、以降の解説と試論は、χ^2 統計量を用いたモデルの適合度検定に範囲を絞ることにする。適合度検定には、複雑な理論に基づいたモデルとの適合を検定するような場合だけではなく、より単純な独立性の検定や比率の差の検定が含まれるので、適合度検定についての議論は多くの分析者にとって有用である。また、考え方自体は、t 統計量や F 統計量を用いる検定であっても同じように適用できるので、限定的な議論で

はない。あえて適合度検定に話を絞るのは、適用例が示しやすいことと、McLaren et al.（1994）によって議論の土台が整っているからにすぎない。

一般化χ^2適合度検定は、McLaren et al.（1994）が医療分野で提案した方法であり、端的に述べると以下のような問題に対処するために考え出されたものである。血液検査の機械は微量の血液サンプルからその人の血液中にどのような大きさの赤血球がどのように分布しているかを推し測ることができる。この赤血球の大きさの分布が、正常な人々に期待される分布から逸脱しているならば、その人は貧血の傾向があると判断される。しかし、近年の医療機器は、容易に大量のサンプルを処理できるため、ピアソンのχ^2統計量による適合度検定はわずかな適合の悪さであっても見逃さず、有意に適合が悪い（つまり、貧血である）と判断を下してしまう。ところが、大規模サンプルに見られるわずかな適合の悪さは、貧血によるものではなく、機器の測定誤差によるものであり、医療的には意味がない。このため、通常の適合度検定の手続きが適切な判断のために役に立たないという問題が引き起こされた。

そこで、McLaren らは正常な人々に期待される赤血球の分布からの逸脱が〈まったくない〉という仮説を帰無仮説として適合度検定をするのではなく、機器の測定誤差と考えられる程度の〈わずかな逸脱しかない〉という仮説を帰無仮説として適合度検定を行えばよいと考えた。通常の Pearson のχ^2適合度検定が逸脱量を 0 とする帰無仮説に限定されているのに対して、McLaren らが提案したものは、許容される逸脱の程度を自由に設定できる点で一般性が高い。そのため、彼らはこの方法を一般化χ^2適合度検定と呼んでいる。

一般化χ2 適合度検定の手続きは、以下のような考え方に基づいている。通常、Pearson の適合度検定は Pearson のχ^2統計量を利用して行われるが、モデルからの直接的な逸脱の量は、χ^2統計量を標本数 n で割った値によって測定することができる（Cressie and Read, 1984）。つまり、

$$\frac{X^2}{n}=\frac{1}{n}\sum_i\frac{(F_i-E_i)^2}{E_i}=\sum_i\frac{(F_i/n-P_i)^2}{P_i}$$

が逸脱量の尺度となる（n は総サンプル数、F_i は観察度数、E_i はモデルから期待される度数、P_i はモデルから期待される確率を表す）。母集団における真の逸脱量を d と表すならば、通常の適合度検定では $d=0$ が帰無仮説、$d>0$ が対立仮説である。これに対して、一般化χ^2適合度検定では、$d=d_0$（$d_0>0$）が帰無仮説、$d>d_0$が対立仮説となる。

このような検定は、Pearson のχ^2統計量を非心χ^2分布（non-central χ^2-distribution）に照らし合わせることで可能になる。非心χ^2分布とは、通常のχ^2分布（central χ^2-distribution）に対して中心が偏っているχ^2分布であり、通常のχ^2分布の平均値がその分布の自由度νと一致するのに対して、非心χ^2分布はその分布の非心度（non-centrality）δを加えた$\nu+\delta$を平均値に持つ。サンプル数 n が十分に大きいとき、非心度δと逸脱量

d との間には $d=\delta/n$ の関係が成り立つ。したがって、母集団の真の逸脱量 d が d_0 であるという帰無仮説が正しいとするならば、その母集団から選ばれた n 人のサンプルが示す Pearson の χ^2 統計量は、非心度 $\delta=n\cdot d_0$ の非心 χ^2 分布に近似する。これに照らし合わせて、帰無仮説を採択すべきか棄却すべきかを判断すればよい。非心 χ^2 分布への近似はサンプル数 n がおよそ 100 以上あれば成り立つ（Drost et al., 1989）ので、十分に通常の社会調査の範囲をカバーしている。

一般化 χ^2 適合度検定の具体的な手続きは、非常に簡単であり、何ら技術的な困難はない。まず、どの程度の逸脱量ならば意味のない逸脱とみなすのか、d_0 の値を定める。次に、通常使用している統計ソフトウェアなどを用いてふつうに χ^2 統計量による適合度検定を行い、χ^2 統計量の実現値を算出する。そして、算出された χ^2 値を非心度 $\delta=n\cdot d_0$ の非心 χ^2 分布に照らし合わせて、有意確率 p を算出する。非心 χ^2 分布における有意確率 p は、例えば統計ソフトウェアとして SPSS を利用している場合、次のシンタックス・コマンドを実行すれば、算出できる。

```
compute    p=1-NCDF.CHISQ(x2,df,nc).
execute.
```

（x2 には χ2 統計量の実現値、df には自由度、nc には非心度を代入する）

４．JGSS データに対する適用例

では、JGSS-2002 のデータを用いて、一般化 χ^2 適合度検定の適用例を示し、その有効性を確認しよう。JGSS データは、十分に大きなサンプル・サイズを持ち、公開データのため再確認が容易なので、適用例としては格好の材料である。

用いる設問項目は、各種組織・団体への信頼の程度（留置票 Q16）である。「次にあげる A〜O について、あなたはどれくらい信頼していますか」という質問文で、大企業、宗教団体など 15 個の組織・団体への信頼の程度が 3 段階で尋ねられている。ここでは、単純化して、「とても信頼している」「少しは信頼している」をまとめて「信頼している」とし、「ほとんど信頼していない」を「信頼していない」とすることで、信頼の有無を表す 2 値の変数と考えることにする。そして、検定を行うのは、ある組織を信頼している者の割合が、性別によって（ある程度大きく）異なるのかどうかである。つまり、性別と信頼が独立であるというモデルに対して適合度検定を行う。

例えば、「学者・研究者」に対する信頼を男女別に集計すると、表 1 のようになる。Pearson の χ^2 統計量で独立モデルへの適合度検定を行うと、自由度 1 で $\chi^2=3.93$ なので、5%水準で有意な逸脱があると明確に判断できる。しかし、実際にこの 2.5%程度の信頼率の男女差が社会学者にとって意味があるかどうかは、微妙な問題である。もちろん、分析の文脈によるが、多くの分析者は重要な意味があるとは読み取らないと予想される。社会

表１　学者・研究者への信頼（男女別）

	信頼 している	信頼 していない	合計
男性	961 （89.9%）	108 （10.1%）	1069 （100%）
女性	966 （92.4%）	80 （7.6%）	1046 （100%）

df=1, $\chi^2=3.93$ *

学者が実際に検証したいのは、性別と信頼が〈少しでも〉関連しているかどうかではなく、〈大きく〉関連しているかどうかなのである。そこで、このデータに対して、一般化χ^2適合度検定を適用し、通常の適合度検定と対比してみよう。

まず、独立モデルからのどの程度の逸脱であれば意味がないとみなすのかを判断し、帰無仮説の逸脱量 d_0 の値を定めなければならない。ここでは、次のように d_0 を定めることにした。男女別の信頼率が、60%と 50%の場合と、15%と 5%の場合では、男女差は等しく 10%であるが、前者の方がその差が持つ意味は弱い。後者が男女で信頼率が 3 倍異なるのに対して、前者は 1.2 倍にしかならないからである。一般に、同じ比率の差であっても、両グループの比率が 50%付近にある方がその意味は弱くなる。いま、男女間でちょうど 50%をはさんで 3%の信頼率の差があると仮定する。つまり、51.5%と 48.5%の信頼率を持つと仮定する。3%という微量の差を持ち、その意味がもっとも弱くなる 50%前後の場合において、この 3%の差に社会学者が重要な意味を読み取ることは、おそらくほとんどないであろう。このとき、男女比が 1 対 1 であるとすると、その逸脱量 d は以下のように.0009 と算出される。

$$d=\frac{(.2575-.2500)^2}{.2500}+\frac{(.2425-.2500)^2}{.2500}+\frac{(.2575-.2500)^2}{.2500}+\frac{(.2425-.2500)^2}{.2500}=.0009$$

そこで、今回の試論では、帰無仮説の逸脱量 d_0 を.0009 と定め、一般化χ^2適合度検定を行うことにした（帰無仮説：$d=.0009$、対立仮説：$d>.0009$）。3%の差という基準は、まったく恣意的なものであるが、計量社会学者一般の感覚からかけ離れて強すぎるような仮定ではないであろう。

$d_0=.0009$ と定めたならば、自由度 1、非心度$(1069+1046)\times.0009=1.9035$ の非心χ^2分布に照らし合わせて、Pearson のχ^2値（3.93）の位置から、その適合の悪さの有意度を確認すればよい。実際に SPSS を用いて有意度を算出してみると、$p=.274$ となった。有意水準を 5%とするならば、この結果は有意ではないので、帰無仮説は棄却される。つまり、学者・研究者への信頼率は、男女の間で〈大きく〉異なっているとは言えない。この検定

結果は、通常の適合度検定の結果とは異なっており、有効に働いている。

同じ手続きを 15 個の組織・団体すべてについて適用した結果をまとめたものが表 2 である。帰無仮説の逸脱量 d_0 はすべて等しく.0009 とした。見やすさのために、信頼率の男女差が高かったものから順に結果を並べ直してある。一番右側の列が 5%水準での一般化 χ^2 検定の結果であり、*印のある箇所が、帰無仮説が棄却され、有意な男女差があると判定された項目である。母集団において男女間で〈大きな〉信頼率の差があると判断されるのは、労働組合、中央官庁、金融機関の 3 つのみであることが分かる。その左には、有意水準を 5%、1%、0.1%とした場合の、通常の χ^2 適合度検定の結果を併記している。通常の適合度検定を 5%水準や 1%水準で行った場合に比べて、わずかな男女差しか持たない項目は有意と判定されていないので、やはり一般化 χ^2 適合度検定は有効に働いている。

表２　通常のχ2 適合度検定と一般化χ2 適合度検定の対比

	各組織を信頼する割合			χ^2 値	通常の χ^2 検定			一般化 χ^2 検定
	男性	女性	男女差		5%	1%	0.1%	5%
労働組合	56.3%	65.4%	9.1%	15.37	*	*	*	*
中央官庁	57.6%	64.9%	7.3%	12.00	*	*	*	*
金融機関	61.4%	68.4%	7.0%	12.4	*	*	*	*
大企業	65.9%	69.8%	3.9%	3.75				
テレビ	84.6%	88.4%	3.8%	8.52	*	*		
国会議員	34.2%	37.6%	3.4%	3.01				
学校	87.6%	90.9%	3.3%	7.20	*	*		
学者・研究者	89.9%	92.4%	2.5%	3.93	*			
市区町村議会議員	45.3%	47.6%	2.3%	1.34				
新聞	95.1%	96.1%	1.0%	1.50				
裁判所	92.1%	93.0%	0.9%	.70				
宗教団体	17.1%	17.8%	0.7%	.27				
病院	93.3%	92.8%	-0.5%	.18				
警察	79.2%	79.7%	0.5%	.08				
自衛隊	74.8%	75.3%	0.5%	.07				

注：それぞれの項目について、無回答者および「わからない」と回答した者は、分析対象から除外している。分析対象のサンプル数は、労働組合から自衛隊まで順に、男性が 959, 1113, 1161, 1032, 1270, 1171, 1234, 1069, 1169, 1305, 1144, 1117, 1299, 1236, 1101 であり、女性が 824, 1023, 1140, 1049, 1423, 1143, 1337, 1046, 1148, 1445, 1151, 1153, 1457, 1284, 993 である。

ただ、一般化χ^2適合度検定の結果を通常のχ^2適合度検定の結果と見比べると、それが単に有意水準を 0.1%と厳しくしたときの通常の検定結果と同じものにすぎない、と思えるかもしれない。しかし、それは誤りである。なぜならば、仮に有意水準を 0.1%として、通常のχ^2適合度検定を行うならば、その検定結果はサンプル・サイズに大きく依存するからである。もし、実際のデータとまったく同じ配分のままサンプル数が 2 倍のデータが得られていると仮定するならば、0.1%水準で検定を行った際には、5 つの項目でその男女差が有意と判定される。さらに、元のサンプル数の 4 倍のデータが得られていると仮定してみると、8 個の項目で男女差が有意となる（いずれも分析結果の表は省略)。いくらでもサンプル・サイズを大きくできるならば、(男女の信頼率が完全に一致していない限り）必ずいつかはわずかな適合の悪さが有意と判定されるのである。これに対して、一般化χ^2適合度検定を行った場合、サンプル・サイズをいくら大きくしても、最初に定めた逸脱量 d_0 よりも小さな無視すべき適合の悪さが有意と判定されることはない。そのため、サンプル・サイズの大小にかかわらず、一定の有意水準で検定を行うことができる。

５．まとめと課題

この論文では、「大規模サンプルに対する検定は、わずかな傾向でも有意になりやすい」という問題に対する 1 つの解決策として、一般化χ^2適合度検定の可能性を示した。社会学的な仮説を検証するためには、ある程度〈大きな〉傾向が母集団にあるのかどうかを判定する必要があるので、この問題は重要である。大規模サンプルの公開データが普及した現在、計量社会学者がこの問題を避けて通ることはできない。一般化χ^2適合度検定は、母集団についてモデルからの逸脱が〈まったくない〉ことを帰無仮説とするのではなく、〈大きな逸脱がない〉ことを帰無仮説とする。その手続きは、非常に簡単であり、通常の Pearson のχ^2統計量をそのまま活用することができる。実際に JGSS-2002 データに対してこの方法の適用を試みたところ、非常に有効に機能することが確認された。

しかし、適合度検定の手続きをよく理解している分析者にとっては、もしかすると、一般化χ^2適合度検定はあまり魅力的なものに映らないかもしれない。単純にある程度の逸脱が存在する場合の期待度数を自分で算出し、その期待度数に対する適合度を通常の適合度検定の方法で検定すれば、ほぼ同じ結果が得られるはずだからである。しかし、複雑なモデルにおいて一定の逸脱量を持つような各セル度数を算出することは、非常に手間のかかることであるし、人々の納得を得られるようにその算出手続きを示すことは困難である。これに対して、一般化χ^2適合度検定は、統計ソフトウェアを利用しても算出される通常の Pearson のχ^2統計量をそのまま活用することができる。分析者にとっての新たな負担は、帰無仮説の逸脱量 d_0 を定めることと、非心度 $n \cdot d_0$ の非心χ^2分布における有意度 p を求めることだけである。非常に小さな手間で問題を根本的に解決することができる、実用的な方法と言える。

唯一の残される問題は、帰無仮説の逸脱量 d_0 をどのようにして定めるかということである。今回は、これを便宜的に.0009 と定めたが、その定め方について何らかの納得できる基準を確立することができなければ、それは検定の有意水準を任意に変更するのと同じ恣意性を逃れることができない。どのような基準に沿って d_0 を定めることが妥当なのかを慎重に検討することが、次の課題であろう。

この課題さえ克服することができれば、一般化 χ^2 適合度検定は、計量社会学者にとって非常に有効な道具となるはずである。また、非心 t 分布や非心 F 分布を利用することによって、一般化 t 検定や一般化 F 検定のような方法も可能と考えられ、その可能性は統計的検定一般に広がっている。

[Acknowledgement]

日本版 General Social Surveys（JGSS）は、大阪商業大学比較地域研究所が、文部科学省から学術フロンティア推進拠点としての指定を受けて（1999-2003 年度）、東京大学社会科学研究所と共同で実施している研究プロジェクトである（研究代表：谷岡一郎・仁田道夫、代表幹事：佐藤博樹・岩井紀子、事務局長：大澤美苗）。データの入手先は、東京大学社会科学研究所附属日本社会研究情報センターSSJ データ・アーカイブである。

[参考文献]

Cressie, N. and Read, T. R. C., 1984, “Multinomial Goodness-of-fit Tests,” *Journal of Royal Statistical Society. ser. B*, 46(3), 440-464.

Drost, F. C., Kallenberg, W. C. M., Moore, D. S. and Oosterhoff, J., 1989, “Power approximations to multinominal tests of fit,” *Journal of American Statistical Association*, 84(405), 130-141.

Fitzmaurice, Garrett M., 1997, “Model Selection with Overdispersed Data,” *The Statistician*, 46(1), 81-91.

McLaren, C. E., Legler, J. M., and Brittenham, G. M., 1994, “The Generalized $\chi 2$ Goodness-of-fit Test,” *The Statistician*, 43(2), 247-258.

佐藤博樹・石田浩・池田謙一編, 2000, 『社会調査の公開データ: 2 次分析への招待』, 東京大学出版会.