講演５．北川源四郎（統計数理研究所所長）

［北　川］ 私は日ごろ学生や若い研究者達に、話の前に言い訳をするなと言い聞かせています。つまらない話かどうかは、言われなくても少し聞けばすぐに判ることです。それに言い訳は努力の放棄につながります。しかし、今日はそうそうたる先輩方を前に何か言い訳をしたい気持になっております。(笑)

お手元に資料をお配りしましたが、昼休みにご出席の方を考えて多少入れ替えしましたので順番は変わっております。

まず自分自身の研究所についてお話ししておいたほうがいいと思います。統計数理研究所は遺伝研、情報研、極地研とともに四つの大学共同利用機関で情報システム研究機構を構成しています。この機構の特徴は普通の自然科学の研究所とは違う立場で生命、地球、環境、社会などの複雑な現象を情報とシステムという観点から捉えようという新しいタイプの研究を目ざしていることにあります。

統計数理研究所は昔から文科省からも何をやっているかよくわからないとしばしば言われてきました。特定の対象分野がないので非常に説明が難しいのです。

もう一つ問題を複雑にしている要素ですが、統計学は17世紀、18世紀ごろから、スタティスティクスすなわち国の状態を計測する官庁統計としてスタートしたことにあります。ところが19世紀の終盤から20世紀にかけて、方法の学問としての統計学が出現し、現在までその二つが並立しています。世の中一般の方の多くは統計学というと前者をイメージされますので、我々の研究をお話しすると非常に面食らい、何が何だかわからないという反応をいただきます。

それでは、今日は、まず情報化に関連して社会、科学研究のスタイル、データ環境が変化しているということをお話しいたします。また、それに関連して知識とはどういうものかというイメージも変わってきているというお話をしたいと思います。また少し大げさになりますが、そういった状況を考えて新しい科学的な方法の確立を目指したお話をしたいと思います。そして、最後に少し具体的な話題についてもお話したいと思います。

まず情報化の影響についてです。ＩＴあるいはＩＣＴと略されるインフォメーション・コミュニケーション・テクノロジーあるいは情報通信技術の発展です。ご承知のように、センサー、コンピュータ、インターネット、データベースが20世紀後半に急激に発達しました。その結果、社会あるいは学術研究のいろいろな分野で大量、大規模なデータが蓄積し、なおかつ現在も得られつつあります。単に量が多いだけではなく、大規模で、しかもヘテロな、すなわちいろい

ろな種類が混在したデータが得られます。

ここに幾つか例を示しています。ライフサイエンスにおけるマイクロアレイデータや経済データです。私どもが研究所に入ったころ、経済データといえば年に１回、せいぜい月に１回のデータしかありませんでしたが、その後マーケティングの分野ではセブン・イレブンがＰＯＳシステムを作りました。よく知られているように皆様がお店で買うたびにすべてのアイテムが記録され、さらにその人の年齢、性別なども入力して、どの商品を買った人がどの商品を買うということがすべて把握できるようになっています。

ファイナンスの分野においても、現在では取引があるたびに分単位で記録されています。環境関係では環境に関するNO_xやCO_2などのデータが時々刻々観測されています。地球科学、天文、気象の分野も同様です。防災関係では1979年に東海、関東地方が重点観測地域に指定されて、いろいろな物理学の観測が始まっています。歪み、気圧、水位などを２分間隔で三十数年間観測していますので、現在までに膨大な、しかも多変量のデータが得られるようになっています。

天文の分野でも同様です。晴れた日には毎晩、全天をＣＣＤカメラで撮影していますので、原理的には移動する彗星や光度が変化する超新星は自動的に発見できるのではないかと思います。もちろん、これにはアマチュアの方の楽しみを奪うという問題があります。いずれにせよそういう形で大量データの利用環境は急速に発達してきています。

次に、このような変化に関連する社会の変化を考えてみたいと思います。左側が技術的な側面からの見方を示したものです。いわゆる工業化社会が情報化によって情報化社会あるいは情報社会に変化してきました。情報化社会というのは通信情報の技術が発達した社会ということです。これに対し、情報社会というのは物質・エネルギーと同等以上に情報が重要になってきた社会という意味だと思います。そこでは情報をたくさん持つ者が経済の競争に勝つという状況になってきました。

ところが21世紀に入ってさらに状況が変わってきました。私はこれをポストＩＴ社会と呼んでいます。ユビキタス社会という言葉がよく使われていますが、ユビキタスとは遍在するという意味です。どこでも、いつでもということです。それにさらに誰でもというのが加わったいわゆるユビキタス社会が実現しますと、その影響は甚大です。なぜなら情報社会においては多くの情報を持ったほうが勝ちだということですが、もしその情報が共有されるような状況、ユビキタスな社会が実現すると、原理的には全員が同じ情報を持ち得るわけですから、情報それ自体には相対的な価値はなくなります。したがって、ポストＩＴの社会では大量のデータある

いは情報からいかに知識を獲得するか、あるいは自分の知識をいかに発展させていくか、そういう知識発見あるいは知識創造の技術が非常に重要になってくると考えています。

右側の流れは社会制度の側面から見たものです。産業革命で資本主義が発達してきました。ドラッカーという人は本を非常にたくさん書いていますが、その人に言わせれば第二次大戦直後から資本主義は単純な資本主義社会ではなくて、ポスト資本主義に変質しました。彼によれば、ポスト資本主義社会における資源は資本でも土地でも労働でもなく、それは知識です。本当の資本主義社会であれば資本、労働、土地で生産量が決まってくるわけですが、現在はそうではないわけです。生産性の革命によって同じ資本、同じ労働を投入していても企業によって生産量は全く異なります。さらに、現代においてはそれだけではなくて知識が極めて重要な要素になってきています。

ドラッカー氏はポスト資本主義社会はいずれ知識社会に変わっていくに違いないということを1990年代に言っております。その意味で私の見たところ、左の図の流れから考えても右の流れから考えても、結局、21世紀社会は知識社会になっていくのだろうと考えております。

ここまでは社会の変化について考えてきましたが、次に科学的な研究の対象と方法の変化についてお話しさせていただきます。非常に大胆に単純化して書いてありますので、自然科学系の方から怒られるかもしれません。

先ほどデカルトの話をしましたが、ごく単純に言えば17世紀以来の科学の研究はデカルトあるいはニュートンパラダイムで機械的世界観に基づいて物理的な世界を中心に発達してきたと思います。しかし19世紀の中ごろ、ダーウィンの進化論が一つの大きなきっかけになって、対象自体が進化し、変化する、確率的な要素を持った実世界も科学的研究の対象となりました。そしてこのような実世界の記述の方法としてゴルトンという人が相関や回帰という概念を提案しました。

この流れを受け継いだのが現代の統計学の始祖のカール・ピアソンです。この人はもともと数学者で、その後、法学やドイツ文学に転向した人でしたが、この様な動きに触発されて1991年、あらゆるものは科学研究の対象になりうることを主張し、その方法としての「科学の文法」を提案しました。

21世紀の状況を考えると科学的研究の対象はさらに広がっていると思います。現在の我々の生活を考えると、特に若い人はインターネットの情報なしに１日も生きることができないような状況になっています。このインターネットから得られる情報は必ずしも実世界を観察して得たもの、あるいは実験によって得られたデータではありません。ほとんどのものは人間が勝手

に作った情報で、物理法則など存在しないものです。そういう意味でサイバーワールドといわれていますが、サイバーワールドまでも研究の対象になりつつあります。別の言い方をすれば人工物です。人工物に関しては認識よりも設計が問題となりますが、人工物を対象とする科学も重要なものとして学術会議でも取り上げられていたと思います。

このスライドは、単純化しすぎと怒られるかもしれませんが科学的方法の変遷・発達をごく大雑把にまとめたものです。まず、デカルト、ニュートンの機械的世界観に基づく研究の手段たる科学の言語としての数学が生まれました。それを基礎に理論科学が発達して20世紀のいろいろな科学をドライブしてきたと思います。もちろん実験は昔からありましたが、先ほどお話ししたように進化論に触発され、K．ピアソンはあらゆるものが科学の対象たりうることを主張し、それを実現するために科学の文法を提案しました。さらにそれを実現する手段として記述統計学を提案します。その後20世紀の初頭にR．A．フィッシャーが推測統計学を発展させました。この流れが実験科学における方法論と言ってもいいと思います。この実験科学の方法の確立によって生物学だけではなく経済学あるいは心理学に関しても従来の物理的な現象の研究に近い形の研究が始まってきました。計量経済、計量心理学、計量生物学などのように、いろいろな学問分野で「計量」とついたものがあります。英語では、○○メトリックスと呼ばれるものです。

この図は科学的方法論の変化を表したものです。従来は理論科学と実験科学が車の両輪のように20世紀の科学を駆動してきました。しかし20世紀の後半になって、非線形ダイナミックス、複雑系、システムに関しては演繹的な範ちゅうではあっても解析解が得られないために、主にシミュレーションに基づく計算科学の方法が急速に発達してきました。

問題は現代ですが、さらにサイバー世界が加わってきました。また実世界に関しても大量大規模なヘテロデータが得られるようになりました。従来のいろいろな分野での実験、調査で得られてきた量と全く違う量のデータが出てきています。現在はそれをいかに使いこなして科学的発見につなげるかが重要な課題で、データマイニングや日本発のディスカバリー・サイエンス（発見科学）が提唱され、急速に使われるようになっています。私は個人的にはこれをベースに第4の科学というものが出てくるのではないかと期待しています。

これに関連して、我々の機構の機構長の堀田先生が面白いことを言われました。生物学というのは19世紀まで観察データによる博物学であった。20世紀は実験の手法を入れて実験科学になった。ところが21世紀になって大量観察データが出てきた。堀田先生は自分はデータを全部見ないで推論するのがサイエンスだと思ってきたけれど、今やゲノムが全部読める時代になっ

ている。下手をするとこれは新しい博物学になってしまうというものです。博物学が悪いというわけではありませんが、大量のデータから何か推論する、それが必要なことだと私は解釈しています。

面白いことに統計学にもやや似た状況があります。19世紀末に記述統計学が観察データを簡潔に記述するという形で提唱され、次に1920年ごろのフィッシャーあたりから、データを厳密にデザインして、少数実験データに基づいて推論を行うという推測統計学が発達しています。ところが、いまや大量・大規模なデータが得られるようになって、今までの手法と違う方法が必要になりつつあります。一つは超大量データの処理ですが、もう一つは必ずしもデザインされていないデータの利用です。そんなデータは役に立ちませんと言うわけにはいきません。貴重なデータです。それを使いこなすための方法が必要だろうと思います。

そういう流れはかなり前から統計の中ではありました。Tukeyというアメリカの人は、exploratory data analysis（探索的データ解析）を1960年ごろに提案していました。3代前の統数研の所長の林知己夫という方はデータ科学というものを提案し、比較的最近では慶應の柴田さんがデータサイエンスを提案しています。面白いことに昨年あたりから情報研の坂内所長がデータ・セントリック・サイエンス（データ中心科学）というものを言い出しまして、同じ流れではないかと考えています。新しい可能性がある一方、従来のを続けていくだけではうまくいかないだろうというのが我々の状況です。

20世紀の科学は、理論科学と実験科学すなわち理論主導型とデータ主導型の両輪によってドライブされてきましたが、20世紀後半に計算機の発達によって非線形系や複雑系を解明するための計算科学が確立しました。しかし、車の四輪をイメージすると、まだひとつ欠けています。我々は理論科学、実験科学、計算科学に続く第4の科学が必要と考えています。そのためには大量データに基づく科学的方法論が必要だろうと個人的には考えています。

サイバー情報に基づく生物学や統計学は現代の博物学に戻ってしまう危険性をもっています。それを超えていくためにはもう一段新しい方法論の進展が必要だろうと思います。そのためには何をやる必要があるか、まだ明確な戦略はできていませんが、大量・大規模データに基づく予測発見の方法の確立や原理主導型のアプローチとデータ主導型のアプローチの統合が必要です。このような統合の努力は地球環境シミュレーションではデータ同化と呼ばれ始まったところです。

統計学は従来サイエンスの対象になりにくかったところに焦点を当てて新しい方法論を作り上げてきました。19世紀末以来生物や経済を重要な対象としてきましたが、その後人間の行動、

診断、マネジメントへと広がり、今後はサービスなどが重要な課題になるのではないかと思います。

それを実現していくために我々はいくつか戦略的な課題を設定しています。今考えているのは予測と知識発見、リスクの評価、実世界のシミュレーションなどです。シミュレーションについては本島先生の核融合研も主要な研究機関ですが、ここで言っているのは確率的な要素、すなわち人間行動、ゲーム論的な要素、人間行動を含めてシミュレーションできるようなものにしていきたいと考えています。それから、日本が国家的な重要課題に挙げてきたサービスイノベーションも戦略的課題のひとつです。

それに必要な要素技術としては、いろいろな情報の統合技術、モデリングの方法、さらには原理主導とデータ主導を統合するデータ統合の方法があります。パーソナリゼーションについてはあとでお話しします。そういう課題に挑戦しながら要素技術を発展させて知識統合の方向を作っていく。それが第4の科学につながっていくのではないかと思います。

ここでまた一般的な話ですが、情報化の影響もあって知識のイメージが非常に変わりつつあることを感じています。別な意味にも使われるようになったと言ったほうがいいのかもしれませんが、その原因の一つは科学の対象が広がってきたということにあります。これまたドラッカーの『ポスト資本主義社会』に書いてありましたが、以前、知識は存在と認識に関するものだった。産業革命のときにはそれがものに適用された。道具に適用されて産業革命が起きた。さらに知識を仕事に使うことによって生産性の革命が起こって20世紀に産業が発展してきた。現在は知識は知識自身に適用されて知識を発展させる、マネジメント革命が起こりつつある。80年代か90年代のかなり古い時点ですが彼はそのように言っております。

こういうことを考えてくると、従来の知識は普遍の真理に関することで、これは現在も多くの人が固く信じていると思います。しかし、現在では必ずしもそれだけではなくて知識のイメージはかなり軽くなっています。人間の行動に有益な情報ととらえることが多くなってきていると思います。これは特に情報科学の方々に顕著で、多くの人がその様に考えています。

異分野の方と話をすると面白いことがあります。情報科学の方と話すと統計の人間と情報の人間では用語が全く違うことがわかります。我々がデータと言っているものが情報で、我々が情報と言っているものを知識と呼んで、その辺は気をつけて話さないと全く誤解したまま1日が過ぎてしまいます。(笑)

そういう知識の変化に対応して、統計科学も20世紀後半に大分変わってきています。このスライドの上側は従来の数理統計学です。フィッシャー以来のものですが、データは真の分布か

ら取られて、そのデータを用いて真の構造に関する推論を行いたい。これが推定論の基本的な枠組みです。

しかしながら最近、統計科学の考え方がだんだん変わってきて、もっと能動的に知識を獲得するという観点から、必ずしも普遍の真理を求めるということではなくて、何らかの意味で役に立つ知識を獲得しよう、あるいは知識を創造しようという観点になってきています。

そうすると、我々が考えるモデルというのは必ずしもtrueを表現したもの、あるいはtrueに非常に近い必要はなくて、推論を行うための根拠であればいいし、有益な推論を導くフレームワークであればいいということになってきます。

左の図が従来のイメージだと思います。真の分布があって、データはそこから出てきている。我々がデータを使ってモデルを推定するのは、ここを限りなく近づけたいからです。ところが予測の視点、主に赤池先生が言われたようなことですが、情報処理のためのモデリングにおいて我々が考えているモデルというのは必ずしも真の分布に少しでも近づけるためのものというわけではなくて、現在のデータに基づいてモデルを作って、そのモデルを使って予測などを行うわけです。例えば、同じ構造から出てくるデータを予測するときには、その予測がよければいいということです。この二つの考え方はあまり違わないようだけれども、実はテクニカルには非常に違います。予測のために作るモデルは、たとえ真のモデルがあったとしても、それに近い必要はない。なぜなら予測モデルの係数は、これが完全にわかればもちろんそれを使えばいい。しかし、実際にはデータから推定しなければいけないからです。

例えば非常に複雑なtrueがあって、データが100個しかなかったら、そのパラメータを全部推定したら目茶苦茶になります。一番典型的にはそういうことです。そういうときは真のモデルを再現することを考えずに予測をよくするという観点で初めからやったほうがいいということです。

ＡＩＣはそれを実現するものとして提案されました。ＡＩＣはそうでしたが、その考えをさらに進めていくと、また知識のイメージが変わってきたということも考慮すると対象に関するあらゆる知識、すなわち、理論やこれまでに得られた経験的な知識、観測データ、場合によっては何をしたいかという目的、それをすべてモデルに投入すればいいと考えています。そういう意味ではこの統計的なモデルは、情報抽出のいわばツールと考えられます。

いったんモデルができれば情報抽出、知識発見、予測、シミュレーション、制御や管理が比較的簡単というか、簡単ではなくても演繹的に実現できることになります。

それに対応して従来の20世紀前半の統計科学は厳格なデザインをして、実験、調査データか

らtruesに関する推論を実現しようとしていましたが、今はむしろ大量データを使って柔軟なモデリングをしていこうという方向になりつつあります。

ここで問題が生じるのは、trueの存在を前提にしないと何がtruesかということがないわけですから、これができればおしまいということがなくなります。モデルが非常に相対的になってしまいます。例えば、時系列解析では自己回帰モデルやＡＲモデルと呼ばれるモデルがよく使われますが、我々の立場ではこれらのモデルはtrueを表したものではなく、ある側面を捉えた非常に単純な近似であると考えます。そうすると、そのモデル族の中で一番いいものを探しても、それで満足すべきものということは決して言えないわけです。

実際、情報量規準ＡＩＣは客観的にモデルの良さを評価できる基準ですが、絶対基準ではありません。ですから、ＡＩＣがゼロになったらそのモデルはベストで、それ以上の改善はあり得ないということではなくて、このように別の候補の族を持ってくれば、その中ではこれがいいということになります。ところが別のモデルクラスを持ってくれば、それまでのベストモデルよりもこの辺の適当に選んだモデルの方ががはるかに良いということはいくらでも起こり得るわけです。ですから、モデル選択は常に相対的な問題です。

実は進歩主義の後継ぎという今日のテーマにはちょっと困りました。そういう意味では近年の統計科学はむしろ進歩主義で、統計の研究者にとってはよいモデルを探す方法や理論を作るよりも、いい仮説や適切なモデルを提案することが必要になっています。そういう意味での進歩は常に要求されていて、データによくフィットするモデルができても、常により良いものがあるのではないか考えないといけない。絶対基準がない以上、統計的なモデリングはそういう宿命を負っています。

話は戻りますが、従来の統計はどちらかというと客観主義の立場をとっていました。データ以外のものを使うなという極端な立場もありましたが、今はむしろ利用できる理論、これまでの経験、あるいはほかのデータの持つ情報など現在のデータに限らず使えるものはすべて使うべしという方向になりつつあります。それを技術的に可能にするのは現時点ではベイズモデリング、ベイズの定理です。

ご承知かも知れませんが、ベイズの定理というのは18世紀に、数学者でお坊さんだったトーマス・ベイズという人が発見したものです。ベイズの理論の優れていることはわかっていたわけですが、ごく近年まで実際の問題にはほとんど適用されていなかったと言っていいと思います。パラメータに分布を考えるということはそもそもよいのかとか、確率をどう定義するかという哲学的な論争や事前分布をどうやって決めるのかという問題があり、厳格なベイズ主義者

はデータを見てから事前分布を変えていけないとか、いろいろな立場がありました。さらに現実的な問題としては計算困難性があって、高次元のシステム、特に非線形性、非正規性が入ってくると、ほとんどお手上げの状態がごく近年まで続いておりました。

しかしながら20世紀終盤にいろいろな発展がありました。例えば事前分布の決め方、ベイズの定理を使って求めたモデルの評価の仕方、それから計算アルゴリズムに関してもＭＣＭＣや逐次モンテカルロ法など乱数を使った方法が開発されて、近年急激に実用化されました。このような急速な転回は1980年以降、一般には90年以降と言っていいと思います。

ベイズモデルはグーグルでも使われていると言われています。ご承知と思いますが、ベイズの定理は逆確率とも言われていますが、事象Ｂを観測した下でのＡの確率を、逆の確率すなわち事象Ａが与えられたときのＢの確率とＡの事前確率から出せるという非常に簡単なものです。例えば最近、迷惑メールで大変迷惑している方もいると思います。それを排除するソフトですが、ある単語の組み合わせが幾つかあったとき、迷惑メールである確率が計算できれば、あるしきい値を超えたら排除することが可能です。

グーグルも驚くほど早く検索できるということを経験されている方もいると思いますが、地球上のあらゆる情報を探して0.2秒ぐらいで結果が出てきます。これは従来の情報科学のロジックベースの方法では絶対に不可能ですが､確率をうまく使ってやっているといわれています。

次は自分の研究所の宣伝になりますが、このような状況の中で大量データに基づく予測、知識発見、情報抽出が大事だということで、統数研では予測発見戦略研究センターを５年ほど前に作っております。このセンターには、地球環境シミュレーションに関してデータ同化、地震の予測解析、遺伝研にも関連しますが生物多様性、それからＤＮＡ等の情報を使った系統樹推定の４つの研究グループがあります。

我々は統計というのは、ものの見方を研究していると自負しておりますけれども、これがなかなか難しい。いわゆる帰納的方法、すなわち個々の具体的事実から一般的知識を得ようとする方法です。これは大変難しいことで、これを完璧にできるはずもなく、ある人は統計は帰納の原罪を負っていると言っております。これは論理的にはできないことだけれども、確率的には推論できるとか、あるいはやらざるを得ないという状況だと思います。それを実現するに当たって統計は、ものを集団として捉えることによって個々の観測値にあまりとらわれずに物事の本質をとらえようとしてきたと考えております。

これはある意味20世紀の大量生産、大量消費の世の中に合っていたのかもしれませんが、近年世の中は非常に変わってきて個人に焦点を当てた科学技術が必要になってきました。これは

倫理的な問題はあるかもしれませんが、テーラーメイドの創薬だとか、オーダーメイドの医療、マイクロマーケティングなどがあります。ダイレクトメールについていえば、昔はすべての人にいっせいに送りつけていましたが、今はそれぞれの人の購買履歴を見て、見込があるところに限定して送るということで非常に省資源かつ効果的になっています。それからサービスについても、教育や物流、医療を個人に合わせてやっていくということができるようになっています。

そうすると、こういう問題に係わっている人間にとっては、平均的にものを見るという従来の立場だけでは立ち行かなくなって、今後は個性をとらえていく方向に統計も進出していく必要があります。ただし統計が帰納を実現するに当たって使ってきた方法とかなり衝突する部分があります。統計は物事を見るときに、ひとつひとつの点ではなくて全体として分布を見ることを主張してきました。簡単にいえば、平均と分散でものごとをとらえていこうというわけですが、個人に焦点をあてるためにはまた構成要素の個々の事情に合わせて何かをしたいという事になります。もちろん、完全に個々に分解してしまったら統計的方法は破綻するわけだけど、全くできないかというとそうではありません。統計の本質である条件付け、コンディショニングをうまくやっていけばこれに適したことはできるだろうと考えています。

ただし、やはり難しい問題があって、非常にたくさんの説明変数を使って分類することが必要です。従来の統計の常識ではデザインマトリックスについて言えば、説明変数10個に対してデータが1,000個あるいは１万個あるという世界だったわけですが、その数が逆転した状況がいろいろな分野で出現しています。

例えばゲノム解析のマイクロアレイでは、サンプル数は100程度に対して項目は数千もあります。マーケティングの分野でも同様な問題があります。これは新ＮＰ問題と呼ばれているものです。もともとのＮＰ問題というのは計算量に関する未解決問題ですが、新ＮＰ問題はデータ数Ｎが説明変数Ｐよりはるかに小さいという問題です。現在は、そういう問題を考えていく必要があります。従来、そんなことはできないと言っていたわけですが、そういう問題も実際に出てきております。

話は変わりますが、情報化が世の中に大きなインパクトを与え、知識が重要になってきています。情報や、知識の特徴は、第一に瞬間的に移動できるということです。物質や資本、土地と違って知識はインターネットを使えば瞬時に移動できます。第二に共有可能性があります。ものは独り占めされるとほかの人は使えないし、資本はよくわかりませんが、知識は全員が共有することも原理的には可能です。

グローバル化あるいはアウトソーシングがよく話題になりますが、国際的なアウトソーシン

グが進行し、アメリカはＩＴ関連のかなりの部分、ソフト開発やコールセンターをインドに出し、特にマイクロソフトの人員の３分の１ぐらいはインドにいると言われています。社会や産業の情報化・知識化と情報、知識の特性がそれを可能にしています。

このグローバル化には二つの側面があります。経済を発展させる一方で、負の側面としては不確実性を増大させ、リスクを非常に大きくしたということがあります。そういう意味でこれからリスクの問題を考えていくことが我々統計の研究者にとっても、ほかの分野の方にとっても非常に大事だろうと考えています。

ところが、このリスクの概念というのが分裂して、現在、考えようによっては四つありますが、大きく分けて個別リスクと統合リスクがあります。本質はそれほど違わないのですが、マスコミなどである物質が危険だとなると一斉にたたかれます。ダイオキシンがホウレンソウに入っているとなると大騒ぎになります。もちろん原子力は安全なほうがよいわけですが、特定の現象のリスクだけを取り上げるわけです。ＢＳＥも同様です。もちろん一つのリスクだけを問題にしたら小さいほうがいいわけですが､それだけで世の中が考えられるかというわけです。

もう一方の考え方として統合リスクという考え方があります。代表的なものは医薬品です。医薬品は病気を治すためのもので、重大な場合には多少の副作用があっても使います。リスクとベネフィットの両方を考える必要があるのです。

リスクに対してもっと積極的なのはファイナンスです。ファイナンスの人は利益を得るためにリスクを取ります。ハイリターンはハイリスクに対応しています。そういう意味で、リスクとベネフィットを両方考えて最適化するという考え方が統合リスクです。残念ながらこの二つの考え方が分離した状態になっているというのが現在だろうと思います。

話は変わりますが、辻篤子さんという朝日新聞の論説委員が、あるところに今後の科学リテラシーとして、読み、書き、そろばん、リスクが必要だと書いていました。これまでは読み、書き、そろばんといわれていたが、今後は、リスクも加えるべきだという話をされています。ちょっと苦しいですが、reading、writing、arithmetic、riskで４Ｒだそうです。

中西準子先生という現在、産総研の化学物質リスク管理センターのセンター長をされている方がいます。この方は、ご自分に言わせると常に少数派だそうですが、私が個人的に賛同する部分は統合リスクが大事だという主張です。こういうふうに書いています。

我々の課題はリスクの比較とコストの比較です。リスク同士の比較が大事だし、リスクとコストの比較も大事です。なぜなら、あるリスクを削減するとそれに伴ってほとんど常に別のリスクが発生したりコストが発生するからです。

中西先生が示した１例を紹介します。米国の環境保護庁が発癌性物質を規制したことに追従して、ペルーでは水道の塩素消毒を1991年に中止しました。その結果として、コレラが蔓延して80万人がかかって7,000人が死亡したそうです。そういう意味で、発癌性物質のリスクと疫病大流行のリスクを両方考えるべきだということだと思います。

ほかの例として、魚の話も書いてありました。魚にダイオキシンが入っているという報道があるけれども、魚を食べることのベネフィットによって寿命が延びると思われる部分とダイオキシンによって減る部分を考えると、1,000倍ぐらいも違って食べたほうがよいことになると中西先生は言われていました。

問題は、いろいろな分野で算定しているリスクをどうすれば比較できるかです。いろいろな分野でいろいろな仮定を独自に入れて計算しているので、そのまま直接には比較できません。中西先生の仕事が面白いのは、リスクの問題を非常に単純な指標に落としていることです。損失余命です。これを食べると平均何日寿命が短くなるか。そういう指標にすると非常に簡単に比較することができます。

この方法はベストとは言えないにしても、統合リスクを考えるためには統合的な指標が必要だということは間違いないことだろうと思っています。ＢＳＥで発症する確率はアメリカでも日本でも１億人以上いて0.1人以下と書いてありました。一方、そのための対策に2,000億ぐらいはかかっているということだそうです。

統計数理研究所では３年前からリスク解析戦略研究センターを立ち上げてこの分野の研究を始めております。研究だけでなく私どもは、最近文科省もネットワークと言い出しましたが、network of excellence、ＮＯＥという名前をつけてリスク研究のネットワークを立ち上げました。リスクに関してはファイナンスから医薬品、環境、防災に至るまでいろいろ立派な研究機関がございますが、そのリスクの概念、リスクの問題をリスク評価するときの方法論に関する中心のところに統一的な方法がないように感じます。そういう意味で我々にも、この中心部分で貢献する可能性はあると考えて、リスク解析戦略研究センターが中心になってリスク研究ネットワークを立ち上げています。すでに39研究組織にご加盟いただいたところでございます。

今度は機構本部の話です。情報・システム研究機構の中に新領域融合センターというものを作りました。これは今まで四つの研究所がバラバラにやっていた研究を情報とシステムという観点から何か融合できないかということで始まった機構長主導のセンターです。その中のプロジェクトの一つとして、統数研が中心になって機能と帰納プロジェクトという語呂合わせをしたプロジェクトを作りました。これは従来のサイエンスが実体をモデル化しようとするのに対

して、機能に集目しモデル化するという観点でとらえようというものです。システムなど非常に複雑な対象をその機能に注目してモデル化しようというものです。

そのときに必要になってくるのはやはり帰納的な方法です。従来の統計的な方法とともに第４の科学や帰納的方法と演繹的な方法を統合したデータ同化の方法が必要になってくるだろうということで、いろいろな問題を解決しながら、そういう方法論を開発していくことを目指していきたいと考えております。

これは全然別な話かもしれませんが、日本はよくソフトが弱いと言われています。財政的にも次世代スパコンを1,150億円かけて作りますけれども、ソフト開発にはわずか20、30億ぐらいしか予算がついていないようです。そういう意味では日本にはソフト軽視の問題がありますが、我々はそれだけではなくて、それ以前のところがさらに弱いと思っています。なぜならばソフトウェアというのは特定の課題に対して突然開発できるものではなくて、それを支えるモデルやアルゴリズムという基礎的な研究が必要なはずです。そこをちゃんとやっていこうという国としての戦略が不可欠です。そういう意味で、造語ですがメタウェアと呼んで、我々としてはハードウェア、ソフトウェアだけではなくてメタウェアの研究・開発も必要だということを主張しています。

この帰納と機能のイメージを１枚のスライドにしています。機能のモデリングというのは対象そのものを実体的にモデル化するのではなくて､対象に関する入出力関係をうまく表現して､機能を模倣するようなモデルをつくるということです。

そんな方法はだめだと昔はよく言われていましたが、機能のモデリングによって最近ロボティクスが急速に発達しました。これは関節や腕の動作は微分方程式を書いて解いて、実現されたものではなくて、いろいろな入出力をうまくモデル化することによってアイボなどのロボットがこの10年ぐらい急速に発展してきたと考えています。そういう意味で機能をモデル化するということも今後ますます重要になってくるのではないかと思っております。

最後のまとめに代えましてひと言。確率的な考えは物理学においても量子力学など、いろいろなところで使われていますが、世の中一般では確率的な思考は欠けていると思います。しかしながら確率的思考は徐々に進展しています。我々の子供のころは天気予報は明日は雨ですと言っていましたけれども、最近は確率で表現します。地震予知も30年ぐらい前は何月何日どこでマグニチュードいくつの地震が発生するという予知が観測量を多くすれば実現できるといわれていたわけですが、今はそれはほとんど不可能ということで、あるエリアに例えばマグニチュード８以上の地震が30年間に起こる確率が0.6ぐらいというふうに確率予測を行うようにな

ってきています。それで一般の人にとってが何の役に立つかということがありますが、おそらくサイエンティフィックに言えるのはそこまでだろうと思います。

それから、景気予測等もサイエンティフィックになっていますし、このデータ同化は地球環境シミュレーションの方法です。従来は微分方程式を仮定して、それを地球シミュレータなどで解くという方法ですが、観測できるデータも取り入れながらやることによって従来よりもはるかにシミュレーションができるようになりつつあります。

意思決定やマネージメントなどの分野はもともと統計が関与してきたところですが、これからは特にサービスが重要なものになっています。リスクの問題もますます深刻になってきます。こういったところではやはり確率的な思考というのが必要になると思います。

進歩主義の後継ぎとは何かというテーマにふさわしい話となったかどうか心もとないですが、私の話は以上です。(拍手)

[廣　田] ありがとうございました。

北川源四郎氏の講演についての討議

[本　島] 第4の科学というのは私にとっても大変新鮮だったものですから質問させていただきたいと思います。実験と理論、順番は逆でしたけれども、私達もシミュレーションサイエンスを第3の科学分野とに発展させられないかということを理想として分野を立ち上げつつあります。つまりサイエンスとしてです。その場合、私が強く求めておりますのは、一つのサイエンスの分野を形づくるためには他の既存のサイエンスから独立している必要性があると言うことです。つまり理論について言えば、実験データがなくても優れた理論は作られてきましたし、現に存在しています。実験と理論を両方折り込んでということももちろんあります。実験もそうです。シミュレーションもある意味独立して、他の二つの既存の分野から遮断された状態でも結果を出して新しい法則、つまり真理を見つけ出し、作り上げて行くということが必要ですし理想とすべきだと思います。

さて、今の第4の分野の場合にデータベースの存在が不可欠と理解しております。その場合、分野としての独立性は十分保てていけるのか。またはどうやれば保てるのかというあたりを教えてください。

データを自ら生産するということは多分考えておられないのだろうなと思いましたので質問させていただきます。

科学としての独自性を保つにはどうすればいいのでしょうか。

［北　川］ 我々は自らデータを作ることはほとんどやっていません。それぞれの領域の方がデータを持ったときにいかに料理して理論あるいはモデルに使っていくか、その方法のところをやっているつもりです。

確かに自然科学の先生ですと、二つの方法論を統合するのはけしからん、そういう方もいらっしゃるかもしれない。しかし私は少なくとも予測やシミュレーションのように目標を定めた場合には両者の融合は必要と考えています。シミュレーションにおいてデータ同化は非常に有効ではないかと私は思っています。

［本　島］ 現在は確かにそうですね。

［北　川］ ええ。マルチフィジックスとマルチスケールのシミュレーションは重要になってきていますが、そこで必要となる違った階層の情報統合のためには理論だけではなくて、データ同化の考えも入れてうまく理論モデルとデータの両方を使いこなしていくことによって、それが実現できるのではないかと思います。

［本　島］ 理論と実験の独立性というのはかなり直感的にわかりますね。それをシミュレーションサイエンスにもやはり必要なこととして求めていくべきだ、と考えます。

［北　川］ 独立にできることが大事だというのはそうかもしれません。それをそのあとで統合してもいいのではないか。それは統計科学、システムサイエンスや制御工学のところで、その手続きを交互に繰り返していくことが自然に行われているからです。フィルタリングといっているのはほとんどそうです。したがって我々にとってはほとんど抵抗がないことですが、実験科学の方からは違和感があるというのは他の方からも指摘されています。

［本　島］ 考え方が違って当然ですし、それは我々にもものすごくプラスになるところが多いと思いながら聞いていました。例えば、その典型になるかどうかですが、配布された資料の16ページに統計のものの見方について説明があります。ここに端的に表れているだろうと思いますのは、平均を見るということがあります。これは、ナノサイエンスの考え方と全然違うところではないでしょうか。ナノサイエンスというのは物を小さく小さく分けていって、原子分子動力学で理解できるような、つまり原子分子を一つずつ勘定できるところまで分解して行く科学の手法をとります。そうしますと、逆にわずか100個粒子でもものすごい数の組み合わせの妙が出てくるわけです。ナノサイエンスの成功の秘訣はここにあって、そして非常に面白い結果を出していると言えます。

もちろんナノサイエンスでも平均するという考え方があることは承知しています。ただし、

全然違う考え方だと受け取ったほうがよろしいのでしょうね。

［**北　川**］ 我々は統計科学の中では平均、分散だけを議論するなんていうのは数10年前の古いものだと言ってきていますが、統計解析の基本的考えはそのデータだけに興味があるのではなくて、データの背後にあるものや、次に出てくるデータに関する推論をするためにデータを利用しているという立場です。そうすると、現在のデータを幾ら精緻に見ても将来には役に立たない。将来にも役に立つような本質的な情報を取る一つの手段として、平均や分散を見るという立場だったと思います。だけど、それだけではだんだん済まなくて、全体を一つのものと見て平均をとるなんていうことでは現在の要求にとても役に立たないわけです。けれども条件付き分布を推定していくにしても何らかの汎化（抽象化）の操作がないと、帰納によって一般的な知識を取り出すのは難しいと考えています。

情報科学の人はわりと大胆に考えますが、私達とは少し溝があります。私達は帰納の原罪というのでしょうか、それがロジカルには非常に難しいことだということを自覚しながらやってきました。

［**本　島**］ 大変面白いことと思います。

［**小　林**］ 現在の統計はtrueが何であるかは気にしない、有効な予測をおこなうことだというお話がありましたが、第４の科学とおっしゃっているときは真理なり何かを明らかにするという。

［**北　川**］ 大量データの問題自体は、どちらでもあり得るわけです。したがって、それはモデルに対する立場とは別の次元のことと言えると思います。けれどもやはりデータの環境といいますか、大量のデータが利用できるようになったという、そういう違いがあって、従来の少数のデータを用いて精密な推論を行うというイメージの方法が必要になっていると思います。

［**小　林**］ ただ、有効な予測というのを超える必要があるのではないかと。

［**北　川**］ 従来からの実験科学の立場ではデータ量が多くなっても対象は本質的に変わらないので精度がよくなると考えられます。けれど、先ほどお話ししたように必ずしもtrueを考えない、情報抽出が目的のモデリングでは事情が異なります。そのような知識創造といわれる場面で、新しい方法の成果が出やすいのではないかと思います。

［**高　畑**］ 簡単な質問を３つほど。先ほど本島先生がおっしゃった平均からといいますか、私の拙い理解で言えば今の統計でも一種のアウトライアーを見つけるというレベルのことなのか。先生は絵を描かれましたが、色が変わって個性がそれぞれ持つようなイメージだったので、それとは違う、もっと違った意味合いで個性まで考慮するとおっしゃっているのか。

［**北　川**］ 深い意味ではないです。むしろ量的に大変になったということだと思います。従来

の統計がそういうものを考えていなかったかというと、全くうそで、実は回帰分析やその他の方法でも別の情報を持ってきて、それで説明するというのは常にやってきています。だから、常に何らかの意味で分類したり、ほかの情報も利用して推論することは実質的には行われていた。ただ、将来は本質的な問題も起こるかもしれないし、現在でもテクニカルには全く違います。行列が縦長と横長の場合では数学的に解ける解けないの違いがあって、横長のマトリックスの問題を解くためには別の情報を入れなければ絶対に意味のある解は得られません。そういう意味では質的に違います。従来の状況では最小二乗とか最尤法で簡単に解けたわけです。

［高　畑］ もう１つはベーズの定理に関係したことです。私たちの分野でも系統樹を作るときにベーズ的方法をよく用いるようになっていますが、問題があるのではないかと思っているのは事前確率をどうやって決めるのかということです。その辺は適当に、例えばユニホームディストリビューションでも何でもいいのでやってしまっていいのかどうかという、正当性についての議論はどこまで浸透しているのでしょうか。

［北　川］ それにはっきりお答えするのは難しいと思いますが、我々がやるべきことは、モデルや事前分布の中に持っている知識をなるべくたくさん投入することです。一様分布を使うにしても、その分布の取り方になるべく依存しないところまでぎりぎりモデルを改良していくということが不可欠で、いきなりベイズの定理を使えばいいということはほとんどないと思います。決して自動的推定できることにはならないと思います。

［高　畑］ 東大のデータスモールダッチなんていうのはいろいろな分野の情報も放り込んだうえでのトライアルという。

［北　川］ そうですね。ただ技術的には個々のモデルを選ぶことによるリスクというのは小さくできる。そのための一般的な方法はあります。その意味では改良はできます。ただ、その改良はそれほど大きくなくて、本質的にはベイズモデルで構築していくためにはモデリングのところで頑張って良いモデルを作っていく必要があります。

もう一つは、例えば系統樹など離散的な構造が入っているようなモデルでは組合せの数が爆発して最尤法などで推定できないようなときにむしろベイズでやったほうが現実的になるわけです。

［高　畑］ 最後、ディスカバリーサイエンスの基盤についてです。それにしても何でもかんでも情報だといってため込んでいます。ごみだらけと言っては何ですが、不必要なデータもいっぱいあるのではないかと思うのですが。

現在、蓄積する事自体に問題はないかもしれませんが、ユーザーから見たらごみではない方

がよいわけですので、選択した情報だけをアーカイブ化していく段階もあり得るのか。あるいは、将来を見越して何が起こるかわからないので、とにかくありとあらゆる情報をため込んでいくことしかできないのか。その辺はどうでしょうか。

［北　川］ 今の流れはその問題は考えずに、とりあえず全部ため込んでいこうというもので、従来の統計の立場とは違います。例えば東大の喜連川先生は世界のインターネットの情報を全部ディスクに入れているそうです。それをやっても実はそれほど多くはなくて、１ペタバイトは要らないのではないでしょうか。実験で大量に出てくるのは別として、人が書いたようなインターネットの情報というのはそれほど多くないといわれています。

ただ問題はそこからいかに役に立つ情報を取るかです。最近、下手をするとインターネット検索がだんだん難しくなって、あまりにも情報が多すぎて、自分が欲しいものにヒットしないということがありますね。情報研なども連想計算でなるべく、その人が知りたい情報を提示するために、過去にどういうものに興味を持って検索したかを知りながら結果を出してくる方式などを研究しているようです。そういうことをやらない限り単純な従来の情報検索だとすぐに破綻するし、役に立たなくなります。

［本　島］ 確かに今、一般の方が検索に使う時間がパソコンに向かっている時間の大体３割ぐらいだと言われています。業績審査をする機会などに思うのですが、知らない方の業績を評価する場合、情報不足の状態になりがちです。論文等のリストがありますが、それだけで済まなくてインターネットで個人情報をチェックすることになります。私は真面目にやっているほうだと思いますが、学振の委員などをしているものですから。（笑）

ぜひ階層化をうまく進めて、情報理論ももっと進歩させていただきたい。高畑先生もおっしゃるように検索に使う時間はどんどん増えていくでしょう。今後何かを調べる際に検索に５割以上の時間を使うことになる可能性があります。その結果、考える時間が反比例に減ってしまいます。非常に難しい世の中というか時代がもうすぐ来るわけです。

私は先のミレニアム紀でインターネットの利用者が１億人突破したのはほんの少し前の1998年と申しました。今は何億人ぐらい使っているのでしょうか。おそらく全人口の３分の１ぐらい使っているでしょうからすごい数ですね。

情報流通の起源であるシルクロードのことで思うことがあります。時代は大分前の話になりますが、日本の位置づけと当時の朝鮮半島の位置づけ、中国の位置づけを地勢学的に見てみますと、シルクロード上で大きく違う所は日本は東の端に位置して情報が全部たまってしまう国だったということです。ヨーロッパから来た情報が中国で加工されて、韓半島へ渡り日本へ来

ます。韓国の人は左から右へと要らないものは流せばいいということができたと思います。我々の場合はこれができません。日本の東は海だったからです。これで日本人はずいぶん苦労したと思います。つまり消化しないと消化不良になってしまうからです。だから平仮名もできたのだと思います。そういう観点でため込まないようにするために行動したり考えたりすることのメリットには大きなものがあったのではないでしょうか。

現在、日本のインターネット回線は全部東に向いていますね。ヨーロッパとつなぐときでもアメリカ経由です。

［北　川］ 新しい情報研のスーパーサイネットはアジアにはつなげています。

［本　島］ それはごく最近なんだろうと思います。インターネットの権威である慶應大学の村井純先生は、ロシア経由でヨーロッパと繋ぐことの必要性をポリティカルにも強く主張されています。つまり、全方向的にいろいろなことをする必要があるかと思います。

［北　川］ 現在の世の中の流れは、すべての情報をため込むという方向でまだ進んでいます。個人の行動、履歴まで全部とってしまうというか、生まれてから死ぬまでの行動や書いたものすべて記録しようという、そういう流れにあると思います。

データが余りにも急激に増えたのでそれに対して情報処理の技術が追いついていないと思います。ある時点である量でうまく処理できても、それが１億倍になったらそのままうまく使えるということは期待できなくて、従来の手法がかなり破綻しているのが現状です。

私は個人的には、そこで一つメタな情報を使うことが不可欠だと考えています。そのときは統計の手法が大事になってくると思います。従来の情報処理はダイレクトに情報を操作していましたが、それはあるサイズの場合に適当なのではないかと思います。

［廣　田］ 一種の階層化みたいなものですね。

［北　川］ そうですね。知識もいったん上に上げて、メタなものにしていかないと量が増えたらどうしようもない。

［廣　田］ 私なんかはバイアスがかかることを非常に心配しています。例えば本島先生などはよく知っているのでやっていらっしゃるけれども、それで本当にバイアスがかかっていない情報が取れるかというと、非常に心配しています。

［北　川］ それは間違いなくバイアスがあって、学生なんか宿題があるとインターネットで調べます。そうすると自分のキーワードで入れたのに都合のいい情報だけが出てくるから、あたかもそれが正しいように思っているけれど、実は逆のキーワードを入れると反対の意見がいっぱい出てくるという問題があります。調査でもそうです。

［**廣　田**］ 便利になったような、非常に危険もあります。それがだんだんと増幅されてくるから。

［**北　川**］ 昔の日本では非常にシンプルな方法を持っていて、それを適用するという形に自然に汎化の操作をやっていますけれど、今の情報科学はどうしてもダイレクトにやっているので、この操作が必要だろうと思います。

［**塩　谷**］ 情報の議論は置いて、どういう情報を捨てたらいいか。捨てていいかどうかの判断とか、そっちの方向への統計処理といいますか、何ていいますか、そういうものがすごく必要な気がします。

先ほど本島先生も言われましたが、科学の対象として統計処理していいものとしていけないものというか、統計の処理方法も慎重にやらないといけないものもたくさんあると思います。

例えば稚拙な例ですが、人間の顔も全部平均値をとると美男、美女だった。例えば平均化することで非常に重要な情報が出てくる場合もあります。ナノテクノロジーで例えば固体表面の上で１つの分子が例えば回転することを見ているとき、みんなそれぞれ違う振る舞いをしている。それを平均化したらどういう意味があるのか。もしくは１つ見たことによって、１つのところから非常に重要な知見がひらめく場合もあります。その辺は科学の対象物によって統計処理の仕方というか、使っていい場合と使っていけない場合、その辺の判断というか、その辺はこれからすごく大事になるのではないか。

［**北　川**］ そうですね。非常にラフなイメージで言えば、統計のモデルで平均的なものは構造のところに出てきて、残差が個々の特異性を表しているものに対応しているわけです。その辺のうまい使い分けが必要です。残差というのは必ずしも要らないものでなくて、時系列で、そこをイノベーションといって、それでドライブされて時系列は変動していると考えます。むしろ積極的にそこを利用する立場です。その両方が必要になってきます。

［**廣　田**］ どうもありがとうございました。