連載

21世紀にふさわしい経済学を求めて

第 31 回

桑垣豊

（NPO 法人市民学研究室・特任研究員）

【これまでの連載（掲載ページへのリンク）】

第17章　統計学と経済学

統計学は数値をあつかうときはもちろん、アンケートのような選択肢の回答を集計するときにも必要です。当然、経済学にも必要ですが、使い方があやしい。でも今回説明するのは、今の経済学での使い方を検証して、正しい方法を示すのは一部だけです。

経済学であつかう予測の値打ちはどれくらいか、みかけの周期と意味のある周期の見分け方、気象学で使っている確率的予報（アンサンブル予報）の応用などを紹介します。

今回の話は、むずかしいと言えばむずかしい話です。書くのは結論だけにしてしまえば、簡単かも知れません。そうすると、「なるほど、そういうことか。しかし、常識的なことばかりで、

どこが目新しいのかな」と思われてしまいそうです。常識的なことは、その理由もそれほどむずかしくないのではないか。統計学を使うときはやり方さえわかっていたらいい、と思うかも知れません。

しかし、それがそんなに簡単ではないのが統計学です。統計学を使うときにもっとも頭を悩ませるのは、分析対象に対して、どんな方法を使ったらいいかがわからないことです。それ以前に、使うべきときに、こんな単純なことに統計学を使わなくてもいいと思ってしまいがちです。

最近、世論調査で選択肢の回答率の％を小数点以下１桁であらわすようになりました。例えば12.3％のように。これは、有効数字が３桁であることを意識したものです。その選択肢を選んだ人が1000人未満だと、割り算の２つの数の有効数字が小さいほうの数字で有効数字を考えないといけないからです。小学校６年で習うはずですが、みんな忘れてしまいます。今まで細かい数字の違いを気にして、誤差の範囲の大小を、意味のある違いとして大きく取り上げていました。

今回はむずかしい話や、いろいろ数式が出てきたりしますが、それはそんなもんだと思って、読み飛ばしてもらってもかまいません。ただし、人によってどこまでわかるか、様々なので、一通り載せることにしました。わかるところまで、読んでください。

これでも、数学者や統計学者、物理学者には、初歩的に見える内容です。そのせいか、今回書くことは、あまり本に載っていなかったり、簡単にしか書いてなかったりします。かと言って、中学校で習う数学でもわかるようにするには、高校の数学教科書、丸写しにたくさんのページをさく必要があるので、そこまではできませんでした。

また、今回とりあげたような数少ないことを書いている本が、間違っていることもあります。そういうことで、あえて今回はむずかしいことも載せることにしました。経済の連載ですから、経済学者が間違った方法を避ける手助けになる、とひそかに考えています。

今回の構成

17－１　クロス集計
17－２　予測と情報量
17－３　格差論争と分散分析
17－４　周期性（規則性）の検証
17－５　アンサンブル予測

クロス集計は今回の導入編で、きちんと分析するには案外簡単そうに見えるときにも、ふさわしい統計学の方法が必要だという話です。

予測の話は、的中率は同じでも、珍しい現象のほうが予測には価値があることを説明します。裏返して言うと、的中率が高くても意味のない予測もあるということです。その予測の値打ちが数字（情報量）であらわせるというのがミソです。

格差論争は、21世紀に入った頃にあった日本の格差論争が、統計学を無視して進み、まちがった結論で決着したのではないかと問います。

周期性と言えば、経済では景気の好況・不況の周期に注目が集まっています。経済学では周期性と言わず、循環と言っています。循環では必ず元に近い状態に戻る印象があるので、もっと一般化して周期性と呼ぶことにします。今回は、分かりやすい例として、富士山の噴火と地震の周期性を取り上げます。地震学会の発生確率予測にも疑問を投げかけます。

アンサンブル予報と言えば難しそうですが、台風の進路を予報円であらわすときに使っています。本当は、円でなく楕円で表示するべきだという話をします。経済予測で説明するにはデータ集めが大変で、計算も複雑になりすぎるので、説明には台風の進路予想を使います。それも、かなり単純化したモデル計算をします。それでも、「確率分布の合成」という微分積分の計算が必要です。数学的な部分は付録にしますので、読み飛ばしてください。これは、この連載史上もっともむずかしく、数式の連続です。

でも、数学の応用に関心のある方には、宇宙物理など各種シミュレーション計算で大変役立つ方法です。ネットや本で調べても、まとめて書いた本はありません。数学の部分だけなら、よい本が1冊だけ見つかりました。付録のところで紹介します。

ところで、すでにこの連載で紹介した統計的手法もあります。第７章の「価格分布のある市場モデル」と、第 10 章の「制御理論」です。統計学に関心のある方は、そのような目で読み返していただけると幸いです。

17－１　クロス集計

統計学はむずかしいので、どうしても必要なときだけ使えばいい、と考える人が多いかも知れません。しかし、かなり単純に見える場合にも使わなければいけないのに、使っていないだけのことがたくさんあります。

例えば、アンケートの分析で、10才きざみの年齢層別で回答数を比べるとき、その年齢層の中での回答を％であらわします。これをクロス集計分析と言います。分析するとき、この％どうしで比べるのですが、少し（例えば0.1％）でも違えば、違うと言えるかどうかはサンプル数によって変わってきます。新聞の世論調査などでは、単純に％の数字の大小だけでコメントをつけるのが普通です。それでは、誤差の範囲の違いを、意味のある違い（有意）にしてしまうかも知れません。

この場合は、カイ二乗検定という分析方法を使うことになっていますが、これでは年齢全体と回答の関係はわかっても、個々の20代、30代、……ごとの特徴はわかりません。それには一工夫して、２×２の単純な分割表に置き換えて分析することになるのですが、ほとんどだれもやっていません。やるにしても、この２×２分割表をカイ二乗検定で分析してしまいますが、結構誤差があるので危険です。

では、ふさわしい方法はというと、カイ二乗検定のイエーツの補正（誤差が少ない）や、フィッシャーの直接法（誤差がない）が正確です。フィッシャーの直接法は、電卓・そろばん時代には計算が大変でした。そのかわりに仕方なく誤差は大きくてもカイ二乗検定を使っていたのです。イエーツの補正なら大して計算は増えないので、使うべきなのに知らない人が大半です。

それにフィッシャーの直接法は、40年前くらいのパソコンでもすぐ計算できます。サンプル数が増えて、例えば１万にもなると大変な計算になるなどと、ネットを調べると書いてありますが、対数を使えば一瞬です。実は、フィッシャーの直接法は超幾何分布という名前だけ難しそうな分布を使う分析法ですが、Excelには超幾何分布の関数があるので、少し工夫すれば一瞬で計算してくれます。

【クロス集計の例】

理屈だけではわけがわからない、という人も多いでしょう。具体的なアンケートの例で考えます。ある高校で、就職するなら学校関係と金融機関では、どちらに就職したいかを聞いたとします。表の上と下では、回答の割合は性別でまったく同じですが、調査した人数（サンプル数）が上のほうがちょうど２倍になっています。

表17－１　クロス集計の例

	学校	金融機関	小計	学校の率
男性	26	60	86	30.2%
女性	40	200	240	16.7%
小計	66	260	326	20.2%

	学校	金融機関	小計	学校の率
男性	13	30	43	30.2%
女性	20	100	120	16.7%
小計	33	130	163	20.2%

そこで、統計分析（検定）が必要になります。表17－２は、「カイ二乗（χ^2）検定」と「フィッシャーの直接法」で同じ集計結果に対して、検定を行なった結果です。仮説は「性別で回答に違いはない」というもので、％はその仮説が正しい率をあらわしています。％が少ないということは、「違いがない確率が低い」＝「違いがある確率が高い」ということです。こういう証明したいことと反対の仮説を立てて、それを否定することで証明する方法が「帰無仮説」による分析です。正面から証明するほうがむずかしいので、このような手のこんだ方法を使います。

表では、分析方法とサンプル数によって％が違います。サンプル数が326人のときは、いずれも％が低いので、性別によって回答に違いがあるという結論です。しかし、サンプル数を半分にして分析してみると、５％をはさんだ割合になっています。この％のことを有意水準といいます。通常５％未満で「違いがある」とします。検定方法によって差が出てしまいました。もっと結論を出すのを厳しくするときは、１％水準や0.1％水準のこともあります。

フィッシャー法のほうが正確な検定方法であることがわかっているので、サンプル数が少ないと、カイ二乗検定の誤差が大きくなることがわかります。カイ二乗検定では、サンプル数が50未満とか100未満で誤差が大きくなると言いますが、この例ではそれより多くても無視できない誤差があります。

表17－２　分析方法を比べる　％は有意水準

サンプル数	326人	163人
χ^2検定	0.72%	5.75%
フィッシャーの直接法	0.68%	4.95%

統計学の入門書には、この場合の計算方法がまだカイ二乗検定で書いてあることが多く、フィッシャーの直接法が載っていることもありますが、おまけで書いてあります。新しい入門書でも、統計学の初歩は内容が確定していることになっているので、昔の本と内容を変えていません。計算方法が進歩すれば、それを入門書にも反影させせないといけません。それをサボっているのです。そして、それに従って統計ソフトも昔の方法のままで、使う人も分析方法を意識しないので、それがずっと続くのです。

ちなみに、直接法のフィッシャーは統計学者で、経済学者のフィッシャーとは別人です。お間違えなく。

【参考文献】

『分割表の統計解析』宮川雅巳、青木敏　朝倉書店　2018年　2900円

クロス集計表（分割表）の分析方法は「カイ二乗分析で決まり」と思っている人が多い中、統計学的にどう考えるべきかを書いた珍しい本です。この本によると、統計学者の間では常識的な内容だが、それを解説した分割表の本がないので、この本を出版することにしたとあります。今回の連載でむずかしいこともあえて取り上げたことと、共通しています。

17－２　予測と情報量

リーマンショックは、１００年に一度の危機であったといいます。来年は危機ではないと毎年予想して、１００年に１回、予想がはずれて危機がおとずれても、的中率は99%だと言ったりします。しかし、このような予測に意味があるのでしょうか。その予測が持っている情報量を計算することで、その価値を求めます。

まず、発生確率だけがわかっている場合の情報量を算出します。次に、予測の的中率を加味して、発生確率だけがわかっているとき以上の情報が得られる予測であるか、どうかを調べます。

情報量は、確率の逆数を２の何乗になるか（２を底とする対数）で表わします。例えば、12.5%の確率（的中率）だと、８分の１なので逆数は８。８は２の３乗（２×２×２）なので、情報量は３になります。でも、的中率まで考えて、予測の情報量をどうやって求めるのでしょうか。具体的な数字を使って、計算してみます。先に情報量の公式を示しますが、数式が苦手な人は読み飛ばしてください。

１行目の式が、情報量ｈの計算方法です。雨の降る確率が12.5%だとすると、先ほどの計算のように情報量は３になります。エントロピーＨは、その情報量に同じ確率を掛けて、全部の場合を足しあわせて求めます。全体の情報量と思ってください。的中率を考えるときは、はずれる確率の分の情報量を引き算して計算します。

情報量ｈとエントロピーＨ

個々の情報量　$h = \log_2 1/p_i$

発生確率　$H = \Sigma\ (p_i \log_2 1/p_i)$

的中率考慮　$H = \Sigma\ (q_i (\log_2 1/p_i - \log_2 1/\text{的中率}))$

p：発生確率　q：その予測を出す割合

式(1)

●天気予報の例

数式だけで考えてもむずかしいので、次の計算例を見てください。１週間のうち、５日が晴で、２日が雨の単純なモデルを考えます。予報ではなく、晴れと雨の割合がわかっているとして、その情報量です。情報量は0.863になりますが、右の表のように半々の確率のときがもっとも情報量が多くて、片寄るほど情報量が減ります。

この割合は現実にそうなっているということなので、半々から離れるほど情報の値打ちがない、ということではありません。また、エントロピーを乱雑さととらえて、マイナスの値だとして引き算し、(－1.000)－(－0.836)＝＋0.137を情報量とすることがありますが、まちがいです。物理学（熱力学）の話と混同しています。

表17－３　天気割合の情報量

予報前

	晴	雨	合計
確率	５／７	２／７	
	0.714	0.286	1
情報量ｈ	0.485	1.807	
エントロピーＨ	0.347	0.516	**0.863**

予報前

	晴	雨	合計
確率	１／２	１／２	
	0.500	0.500	1
情報量ｈ	1.000	1.000	
エントロピーＨ	0.500	0.500	**1.000**

今度は、的中率も考慮に入れて計算してみましょう。

的中率については、的中率の負の情報量を求めて、元の情報量から引き算します。例えば、的中率 1/2 なら、情報量１をマイナスと見なします。式(1)の１行目、確率のところに的中率を代入します。的中率が低いほど、情報量ｈは大きくなります。不確実性の部分を情報量から取り除いたことになります。

表17－４の左側は、晴と雨の割合は表17－３と同じですが、晴の予報は５回に１回、雨は２回に１回予報がはずれる場合の計算です。例えば、日本の天気の基本パターンは、５日晴れて２日雨が続きます。現実には、これからはずれることも多く、曇りや雪もあります。話を簡単にしました。

予報の方法は、例えば「明日の天気は今日と同じ」という単純なものを考えます。天気のパターンが、５日晴れて２日雨を正確に繰り返しているとします。そうすると、晴れの予報は５回に１回、雨は２回に１回はずれることになります。

表17－４　天気予報の情報量

前日と同じ天気　　　想定　○○○○○●●

予報＼実際	予報	晴 ５／７	雨 ２／７	予報後
晴	５／７	４／５	１／５	
確率　ｐ	0.714	0.8	0.2	
情報量ｈ	0.485	0.322		**0.163**
雨	２／７	１／２	１／２	
確率　ｐ	0.286	0.5	0.5	
情報量ｈ	1.807		1.000	**0.807**
エントロピーＨ	0.863			**0.347**
的中率			５／７＝	0.714

予報＼実際	予報	晴 ５／７	雨 ２／７	予報後
晴	５／７	５／５	０／５	
確率　ｐ	0.714	1	0	
情報量ｈ	0.485	0.000		**0.485**
雨	２／７	０／２	２／２	
確率　ｐ	0.286	0	1	
情報量ｈ	1.807		0.000	**1.807**
エントロピーＨ	0.863			**0.863**
的中率			５／７＝	0.714

右側は、的中率１００％の場合です。表17－３左と同じになることがわかります。この場合も、天気割合と天気予報の情報量を比べて、天気割合のほうが値打ちがあると解釈してはまちがいです。同じ予報の中で、的中率が高いほど予報の価値つまり情報量が多いと見るべきです。

ところで、左の情報量0.347は、何を意味しているのでしょうか。天気に連続性があるという情報を使っていて、その値打ちが情報量0.347だということです。これよりも的中率が下がれば、情報量がマイナスになることもあります。そうすると、予報をしないほうがいいことになります。

●経済予測の例

リーマンショックのような経済「危機がある」という予報は、天気予報の雨よりも確率が低く100分の１なので情報量が6.644と大きくなります。一方、「危機がない」という予想の情報量は、もともと実現確率が高いので、情報量はわずか0.0145にすぎません。これは、それぞれの予報を出して的中したときの情報量です。的中率を考慮して予報を評価すると、どうなるでしょうか。もし、危機と予想したときの的中率が、危機のおきる確率と同じ１％だすると、全体としての情報量はほぼゼロになります。表17－５を見てください。

表17－５　経済危機の情報量１

100年に一度の経済危機１　　想定 ほとんどはずれ

実際 ＼ 予報	予報	平常 99/100	危機 1/100	予報後
平常	99/100	98/99	1/99	
確率 p	0.990	0.9899	0.0101	
情報量 h	0.0145	0.0146		**-0.0001**
危機	1/100	99/100	1/100	
確率 p	0.010	0.99	0.01	
情報量 h	6.644		6.644	**0.000**
エントロピーH				**-0.0001**
的中率			98/100=	0.980

しかし、危機のおきる予想を１００年に２回して、50％の確率で当たるとします。表17－６を見ると、全体としての情報量は0.107ですが、危機を予想した場合に限ると極めて高い情報量であることがわかります。２回に１回でも危機を予想できれば、大きな価値があるからです。的中率98％は同じなのに。

表17－６　経済危機の情報量２

100年に一度の経済危機２

実際 ＼ 予報	予報	平常 99/100	危機 1/100	予報後
平常	98/100	97/98	1/98	
確率 p	0.980	0.9898	0.0102	
情報量 h	0.029	0.0148		**0.014**
危機	2/100	1／2	1／2	
確率 p	0.020	0.5	0.5	
情報量 h	5.644		1.000	**4.644**
エントロピーH				**0.107**
的中率			98/100=	0.980

この計算では、危機を予想できたときの利益（損失の縮小）や、予想がはずれたときの損失などの利得表がないので、実際の予想の価値を見積もることはできません。損得まで考えると、危機を予想する２回のうち１回あたるというのは大変なことです。単純な的中率の計算に惑わされないために、必要な知識です。

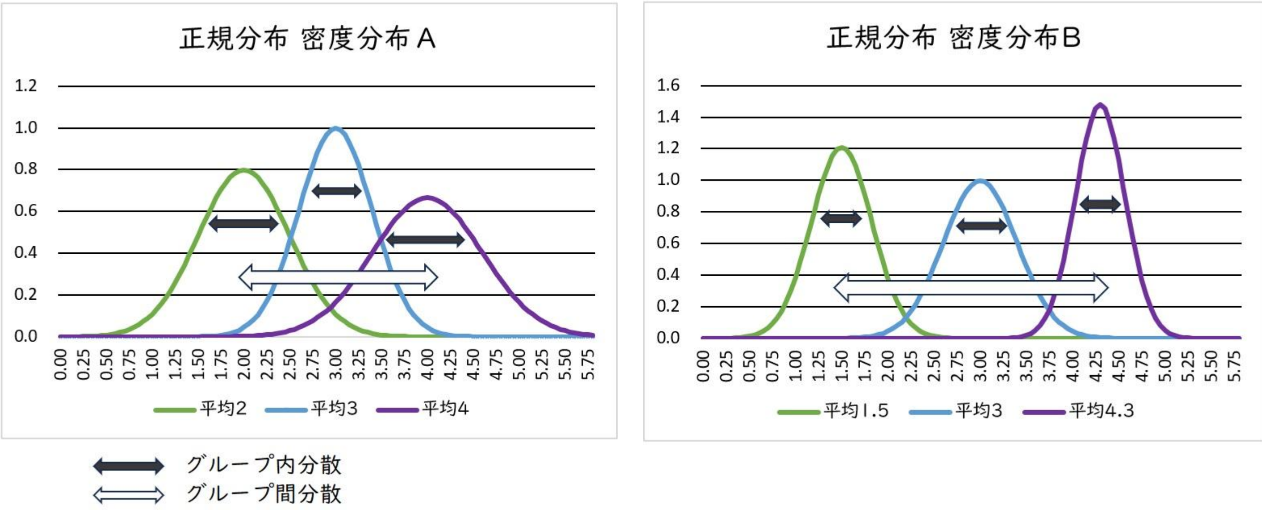
【参考文献】
『情報数学のはなし 改訂版』大村平　日科技連出版社　2018年　P38など参照
著者は、航空自衛隊のトップ、元航空幕僚長。ずっと技術畑を歩いて来られたようです。大村さんのおかげで、航空自衛隊の事故が少なくてすんでいるのかも知れません。情報数学だけで一冊の本は、この本だけのようです。大変参考になりますが、「予想割合の評価」と「的中率を考慮した計算法」がまちがっていました。今回書いた部分は全面的に関係するので、一から考え直しました。

17－３　格差論争と分散分析

かつて、経済学界で格差論争がありました。日本の所得格差は、同志社大学橘木（たいばなき）教授が以前と比べて実質的に増えたと主張し、大阪大学大竹教授は増えたように見えても見かけにすぎないと言い、論争になりました。経済学界では、見かけにすぎないという主張が正しいとして決着しました。これを統計学の知識で考察すると、逆転するのではないかという話をします。
ジニ係数などの分散をあらわす指標には、気をつけないといけません。それは、層別に分散（標準偏差、ジニ係数）を求めて、それを重みづけ平均しても全体の分散を求めることはできない点です。層の間の分散も取り入れないと、全体の分散が計算できないからです。分散分析のひとつ一元配置分析の方法を知れば、理解できます。

●一元配置分散分析
一元配置分散分析というのは、群間変動と群内変動の分散を比べて検定し、群間変動が大きければ、「グループの平均値の間に差がある」と考える手法です。この群は、先ほどの層と同じ意味です。
次のページの図17－１で、分析方法を説明します。
１）「グループ間の分散」と「グループ内の分散」の和が、全体の分散と一致する。
２）従って、グループ内の分散をそれぞれのメンバー数で重みづけ平均しても、全体の分散とは一致しないので注意が必要。これを分散の合成は不可能と言う。
３）一方、平均値は、グループ内の平均値をあつめて、重みづけ平均すると全体の平均と一致する。

図17－１　一元配置分散分析



２）にもかかわらず、分散の合成が可能だという間違った理解で、経済学界は日本の所得格差を高齢化で説明できるという決着をつけました。つまり、年齢別所得格差を計算して、年齢別人口で重みづけして格差を平均して、全体の格差が決まるという認識です。実際に格差を計算するときは、分散でなくジニ係数を使いますが、同じ数学的性質があるので議論には影響しません。

これだと、高齢者のほうがそれ未満の年齢層よりも格差が大きいので、高齢化が進むとおのずと格差が開くので、経済政策や企業の経営の問題ではない、ということになります。

日本の所得格差論争は、高齢化で説明できるということで決着しましたが、その前から問題がありました。まず、2000年代前半にＮＨＫが格差社会の番組をつくったので注目を集めました。貧困層が増えたので、格差も広がっているに違いないという思い込みが前提になっていました。ＧＤＰは少しだけど増えているので、平均所得は増えているはずだ。貧困層が増えているということは、一方で高所得者の所得が増えて平均を押し上げているはずだ、と判断したのです。

実は、所得階層の全般で家計は所得を減らしました。ＧＤＰは増えているのになぜか。企業の内部留保増大を見逃したからです。経済学者も大部分が家計しか見ていませんから、見逃しました。その後、格差論争がおきましたが、その後から本当に格差が広がり始めました。残念ながら、高齢化だけのせいにしてしまいました。今はだれもが認める格差社会ですが、当時の論争の結論はそのままです。

【参考文献】

『日本の景気は賃金が決める』吉本佳生　講談社現代新書2205　2013年

分散やジニ係数を単純に合成してはいけないことに気が付いた統計学者は何人もいたようですが、経済学者で疑問を感じた人は少なく、吉本氏は珍しい例です。以前、この連載（第２章／第２回の【コラム】）でとりあげたＧＤＰギャップ計算方法にも疑問を投げかけておられて、これも数少ない指摘例です。

17－４　周期性（規則性）の検証

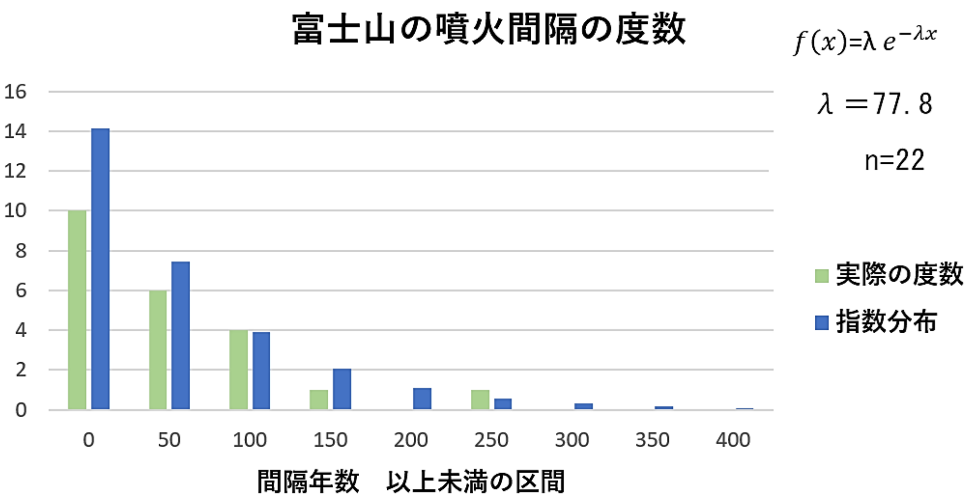
周期的な現象には、自己相関分析を使います。しかし、見かけの周期性を検出してしまう場合があります。景気循環（変動）の分析で注意すべき点です。一方、ある現象が集中しておこるように見えても、ランダムな発生（ポアソン過程）間隔が指数分布を示すことから、見かけの集中にすぎないこともあります。これらが、指数分布の分布形の検定で一定見分けられることを紹介します。

●富士山の噴火

災害の事例で集計をしてみました。富士山の噴火記録は23回あり平均間隔は80年程度です。噴火の起きる間の数は22回になるので、その間隔の度数（分布）を求めます。それが指数分布だと周期性がないことになります。実際の度数を緑の棒グラフにしました。例えば50年のところでは、50年以上100年未満の数を集計しています。指数分布は、平均値だけをパラメーターする変わった分布です。計算結果をグラフで青い棒にしました。

現実の度数は、指数分布より長い間隔で噴火することが多いことがわかります。このグラフから指数分布でないのは確かです。50年あたりに周期性があるかも知れません。周期性があるとすると時系列解析ができます。このグラフの範囲外に大変長い間隔の時期があり、それが指数分布のほうが総回数が多く見える理由です。

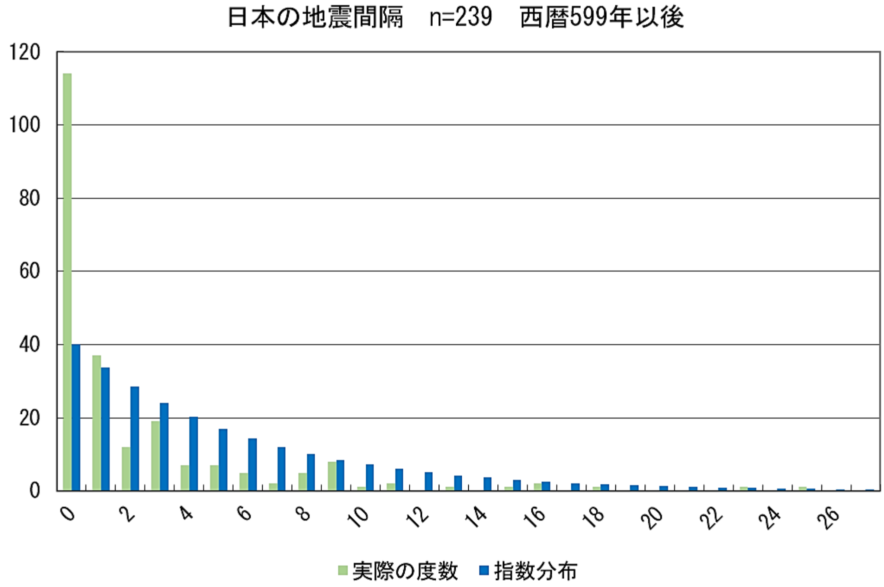
図17－２　富士山の噴火間隔の例



●日本全体の大きな地震

以下の図でわかるのは、地震間隔は１年未満が多いことです。全国なので連動に見えても、関連性のメカニズムをつかむのはむずかしい。一方、指数分布よりも多いところが周期性の候補ですが、９年のところにやや高く見える程度です。

図17－３　日本の地震間隔（年）の度数



日本の地震予知は、プレート型地震以外はすべて指数分布に従うということで、周期性もそのほかの規則性もない（ポアソン過程）ことになりました。つまり、頻度だけで予知は不可能。

プレート型地震は、ワルド分布というだんだんエネルギーがたまっていき、一定値になると地震がおきるというモデルを採用しています。時間に比例してたまる部分と、ランダムウォークのようにランダムにたまる部分の足し算するのが、ワルド分布です。しかし、東海地震の予知をあきらめたのは、エネルギーの蓄積がときどきからぶりになるからです。１周期の時間がすぎても、地震を１回パスすることもあるからです。そうすると、ワルド分布は当てはまらず、プレート型もポアソン過程の可能性が出てきます。つまり、予知もできず、発生確率もずっと同じ値が続くはずです。

ところが、地震がおきないと、毎年少しずつ発生確率が増えていくことになります。それで、南海地震のように、30年以内に80％の確率でマグニチュード８以上の地震がおきるという、過大な確率が出てきます。

ところで、指数分布に従った間隔分布を示せば、周期性だけでなくいろいろな規則性もないことになります。過剰証明です。周期性以外の短い間隔で起こりやすいことなどの規則性も、すべて否定することになります。ただし、未知の要素があることを否定するわけではなく、確かな規則性が今のところわからないという意味です。

経済学に目を向けてみると、景気循環に周期があるかどうかの検証にも使えそうです。でも、景気のピークや底がはっきりしないので、今のところむずかしそうです。今後の挑戦してみたいと思います。

【参考文献】

『デタラメにひそむ確率法則 地震発生確率 87％の意味するもの』
小林道正　岩波科学ライブラリー195　2012年

地震学者や学会が統計学に無理解で、できないはずの確率予報をできるとしていることを批判。ここまでの説明は、この本に頼りました。

付録１　規則性の検定　指数分布のコルモゴロフ・スミノルフ検定

　現実の分布データが、仮定した理論分布に当てはまるかどうかを検定する方法がある。その中の「コルモゴロフ・スミノルフ検定」を紹介する。例えば、シミュレーション結果の分布が、どのような分布か判断するのに使える。以下、地震や噴火のような現象がおきる時期に規則性があるかどうかを調べた。毎年、過去におきた時期に関係なく（無記憶性）、サイコロをふるように同じ確率で「できごと」がおきるとすると、その発生間隔は、指数分布に従う。指数分布は短い間隔ほど多いので、人間はそこに意味を見いだしたりするが、ランダムだからこそおこる特徴である。地震学では、活動期に入ったということが多いが、見かけだけかも知れない。

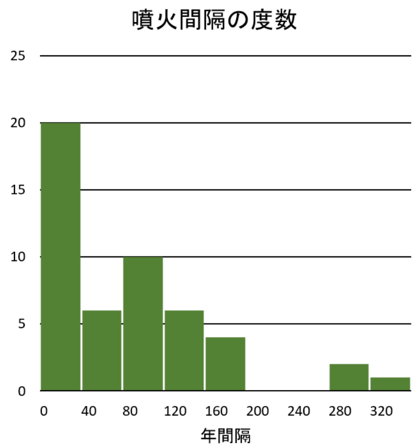
●富士山の噴火

　本文でとりあげた富士山の噴火間隔を検討してみる。24区間（図17－３よりも２回追加）の発生間隔を小さい方から並べて、コルモゴロフ・スミノルフ検定を行なった結果が以下のとおり。５％有意棄却限界よりも小さい数字がでたので、指数分布であることを棄却する（MAXと棄却限界を比べて）。つまり何らかの規則性があることになる。間隔の度数分布をみると80～120年が前後より多く、このあたりに周期成分があるのかも知れない。下の数直線グラフで、富士山の噴火時期をあらわした。

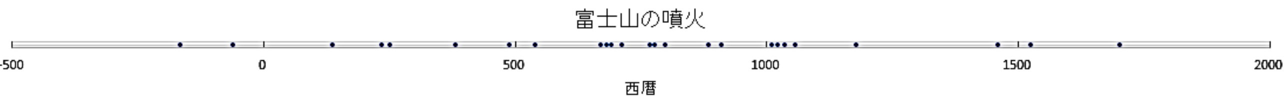
表17－７　富士山の噴火時期に規則性はあるか

富士山の噴火間隔　　ポアソン過程とは言えない（MAXが小さい）

					棄却限界1%	0.3260
n	平均1/λ		指数分布累積		棄却限界5%	0.2720
24	77.8		1-exp(-λx)		MAX=	**0.1364**
j	発生間隔年	j/n	p_j	(j-1)/n	$u_j=j/n-p_j$	$v_j=p_j-(j-1)/n$
1	9	0.042	0.1047	0.000	0.0630	0.1047
2	9	0.083	0.1047	0.042	0.0214	0.0630
3	13	0.125	0.1529	0.083	0.0279	0.0695
4	13	0.167	0.1529	0.125	0.0138	0.0279
5	13	0.208	0.1529	0.167	0.0555	0.0138
6	17	0.250	0.1984	0.208	0.0516	0.0099
7	22	0.292	0.2416	0.250	0.0501	0.0084
8	22	0.333	0.2416	0.292	0.0918	0.0501
9	22	0.375	0.2416	0.333	0.1334	0.0918
10	26	0.417	0.2824	0.375	0.1343	0.0926
11	52	0.458	0.4850	0.417	0.0267	0.0683
12	56	0.500	0.5127	0.458	0.0127	0.0544
13	65	0.542	0.5637	0.500	0.0221	0.0637
14	86	0.583	0.6691	0.542	0.0858	0.1275
15	99	0.625	0.7197	0.583	0.0947	0.1364
16	99	0.667	0.7197	0.625	0.0530	0.0947
17	103	0.708	0.7348	0.667	0.0264	0.0681
18	108	0.750	0.7490	0.708	0.0010	0.0407
19	120	0.792	0.7874	0.750	0.0043	0.0374
20	129	0.833	0.8097	0.792	0.0237	0.0180
21	129	0.875	0.8097	0.833	0.0653	0.0237
22	176	0.917	0.8964	0.875	0.0203	0.0214
23	198	0.958	0.9214	0.917	0.0369	0.0048
24	284	1.000	0.9740	0.958	0.0260	0.0157



320年以上の１回は宝永噴火以後現在



コルモゴロフ・スミノルフ検定の具体的な方法はむずかしいので、省略する。いろいろな本に載っているので調べられたい。ノンパラメトリックを書名に含む本に、たいてい載っている。ただし、計算経過の計算表を載せたので、自力で計算できるヒントになっている。

●東海地震

最後に、学会が予知のできないことを認めた東海地震の規則性を、コルモゴロフ・スミノルフ検定で検証した。その結果、指数分布は棄却された（0.5766＜0.6800）が、５％有意に近かったので、この４回の地震と３回の発生間隔から何らかの規則性を見つけるのはむずかしいのではないだろうか。たった４回の地震で、しかも間隔が100年程度が２回で１回は150年程度。はじめに述べた情報量の観点からも、予知できるとした40年前からもっと疑うべきあった。日本地震学会は、東海地震はあきらめたが、東南海地震等はまだ確率を発表している。前述のように、統計学者などから批判が出ている。

表17－８　東海地震の分析

東海地震の間隔

					棄却限界1%	0.8150
n	平均1/λ	指数分布累積			棄却限界5%	0.6800
3	118.7	1-exp(-λx)			MAX=	**0.5766**
j	発生間隔年	j/n	p_j	(j-1)/n	$u_j=j/n-p_j$	$v_j=p_j-(j-1)/n$
1	102	0.333	0.5766	0.000	0.2433	0.5766
2	107	0.667	0.5941	0.333	0.0726	0.2608
3	147	1.000	0.7103	0.667	0.2897	0.0436

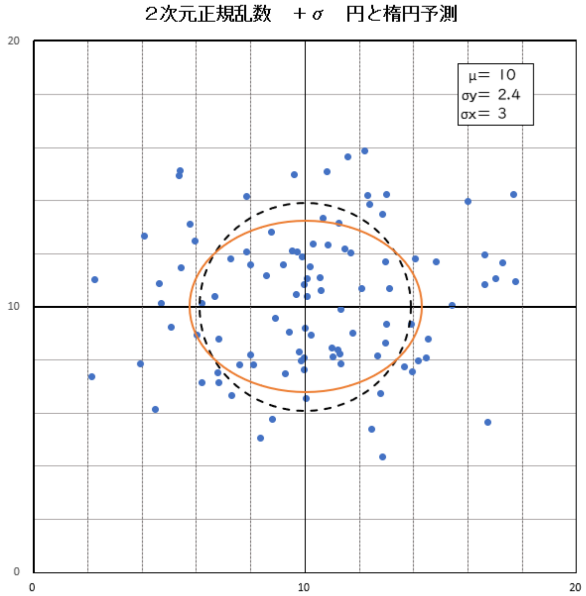
17－５　アンサンブル予測

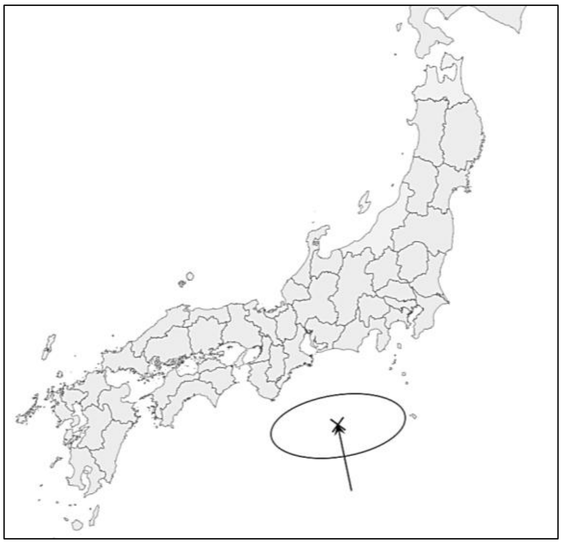
気象予報では 30 年前から導入が進んでいるアンサンブル予報を紹介します。初期条件を少しずつ変えて、結果の分布を得る方法です。カオス現象への対処法です。経済予測でも、初期条件を変えて多数回試行することで利用できます。経済の事例を探しましたが、今のところ見つかりませんでした。

台風の進路予測では、多数回、気象シミュレーションを行ない分布の中心から円を描き、一定割合の予想位置が入るように予報円を描いています。そこで、イメージ図をつくりました。真下から、台風がやってくる想定です。一つの青い点が、一つの計算結果（ここでは正規分布の乱数）です。分布は、２次元正規分布です。

なぜ円にするのか疑問がわきます。というのは、円は半径だけで決まるので１次元予測です。楕円にすれば、進行方向と進行速度の２つを表わすことができます。これを赤い円で書き入れました。特に停滞する台風は、予報円の中に現在の台風の位置を含んでしまう場合があり、円では不十分だと思うのです。停滞する台風は、進行方向につぶれたもっと細長い楕円になります。

図17－５　アンサンブル予報　円（点線）と楕円（赤線）／台風の楕円形の進路予想図





【参考文献】
『メソ気象の監視と予測』斉藤和雄、鈴木修　朝倉書店　2016 年

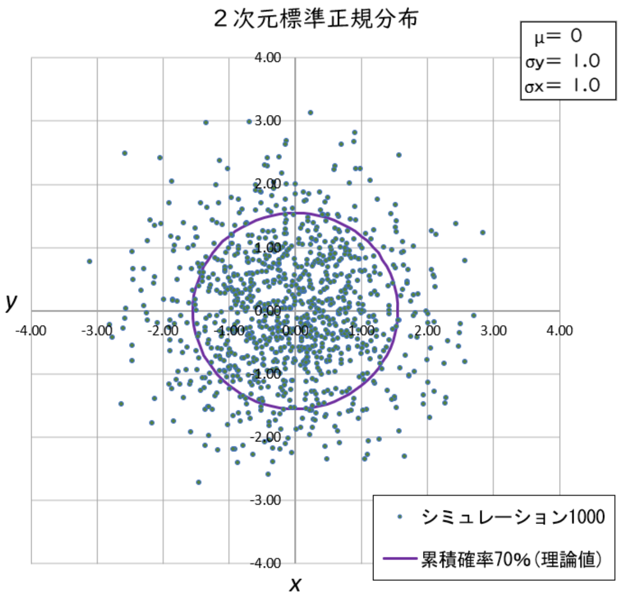
付録２　アンサンブル予測と分布の合成

１）シミュレーションと分布の合成

本文では、シミュレーションしないで乱数をそのまま図に載せた。ここでは、２次元標準正規分布の中心からの距離の分布をとりあげる。台風の予測の状況に似ている。「分布の合成」で正確に分布の形を計算（理論値）して、それとシミュレーションを比べる。シミュレーションで、正確に再現するのにどれくらい計算回数が必要かを検討した。まずシミュレーション結果の図を示す。円の中に70％を含むのは、半径ｒ＝1.55の内側である。分布の合成は実用性が高いが、さらにむずかしいので、付録３に載せた。多変数の変数変換で積分を行なう方法だが、なかなか本に載っていないので、あえて載せる。

計算がしやすい極座標への変換を行なった。これに具体的に２つの分布標準正規分布を当てはめる。極座標で新たに導入する変数ｒ、θの一方のｒを変数にしたいので、θを消す。ちなみにθの分布は一様分布になる。図でも角度（方向）に分布のムラはない。区間は［－π，π］で、値はｆ(θ)＝１／２πで一定。

図17－６　試行1000回のシミュレーション結果



２つの標準正規分布を平面展開したときの中心からの距離分布

●一般的な２つの分布

距離なので

$$r^2 = {X_1}^2 + {X_2}^2$$

$Y_2 = X_2$ とはしないでθを導入　直交座標から極座標への変換

$$X_1 = r\cos\theta$$

$$X_2 = r\sin\theta$$

ヤコビアンは

$$J = \begin{vmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{vmatrix} = r\cos^2\theta + r\sin^2\theta = r \geq 0$$

求めたい合成分布

$$h(r) = \int f_{X_1}(r,\theta) f_{X_2}(r,\theta) r d\theta$$

●標準正規分布どうしの場合

分布関数１　$f_{X_1}(x_1) = \dfrac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}{x_1}^2}$

分布関数２　$f_{X_2}(x_2) = \dfrac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}{x_2}^2}$

求めたい合成分布

$$h(y_1) = \int_{-\pi}^{\pi} f_{X_1}(r,\theta) f_{X_2}(r,\theta) r d\theta$$

$$= \int_{-\pi}^{\pi} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(r\cos\theta)^2} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(r\sin\theta)^2} r d\theta$$

$$= \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-\frac{1}{2}r^2} r d\theta$$　θが消える

$$= \frac{1}{2\pi} r e^{-\frac{1}{2}r^2} \int_{-\pi}^{\pi} d\theta$$

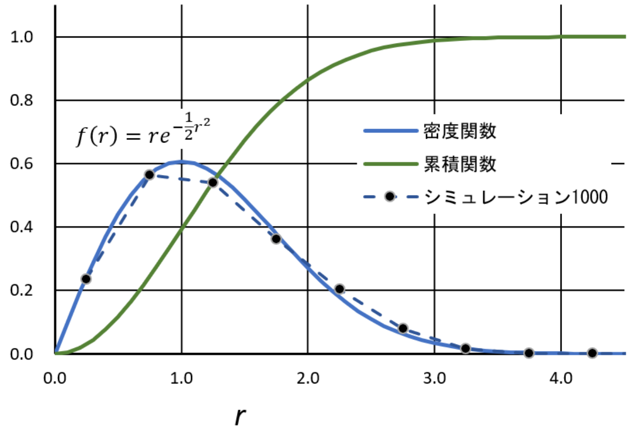
$$= \frac{1}{2\pi} r e^{-\frac{1}{2}r^2} [\theta]_{-\pi}^{\pi}$$　定積分は２π

$$= r e^{-\frac{1}{2}r^2}$$

試行1000回のシミュレーション結果を、集計区間0.5で計算した結果で書き入れた。このような粗い度数分布では、シミュレーションは1000回でも正確に見える。青い曲線が理論値。

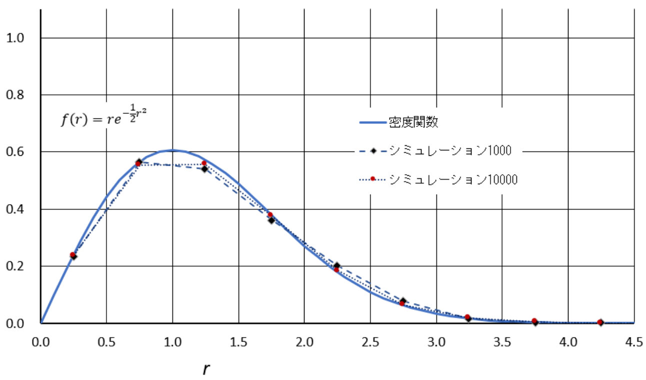
図17－７　距離の分布　シミュレーション1000回　区間0.5

２次元同時標準正規分布の合成



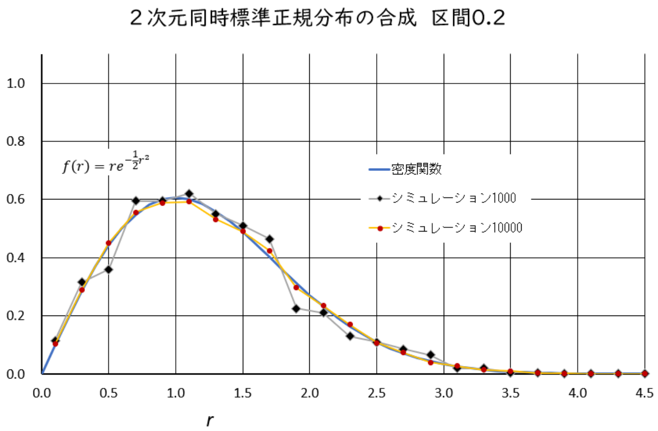
試行回数１万回のシミュレーション計算をした。より正確になっている。

図17－８　シミュレーション１万回　区間0.5



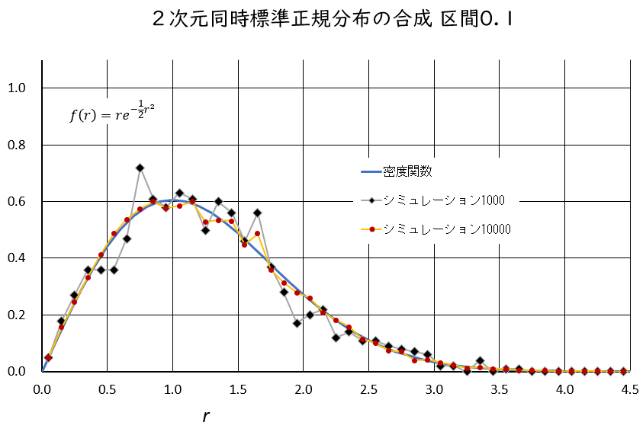
超過確率などを求めるにはこれで十分かも知れないが、分布形が粗すぎるので集計区間を0.2にする。試行回数1000回ではいびつになる。これがシミュレーションするたびに結果が変わる現象の正体である。超過確率などを求める精度は、円の面積をシミュレーションで求めるのと同じ計算である。

図17－9　シミュレーション１万回　区間0.2



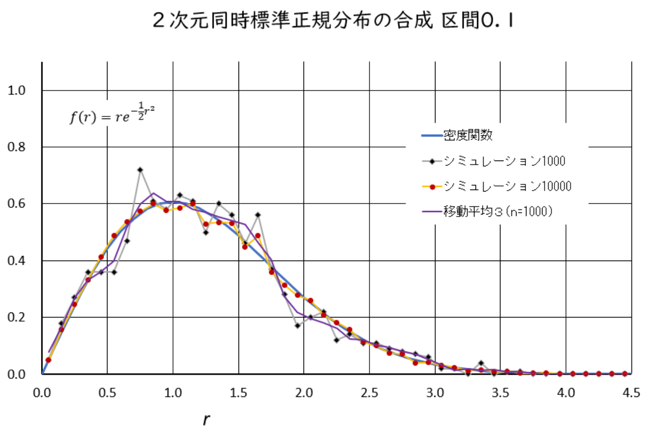
さらに細かい分布形を求めて、集計区間を0.1にすると、シミュレーション回数１万回でもいびつになる。ピークはｒ＝１なので、区間0.1は決して細かすぎることはない。（以下の図の区間の「すそ」の値が約ゼロまで35区間程度）

図17－10　シミュレーション１万回　区間0.1



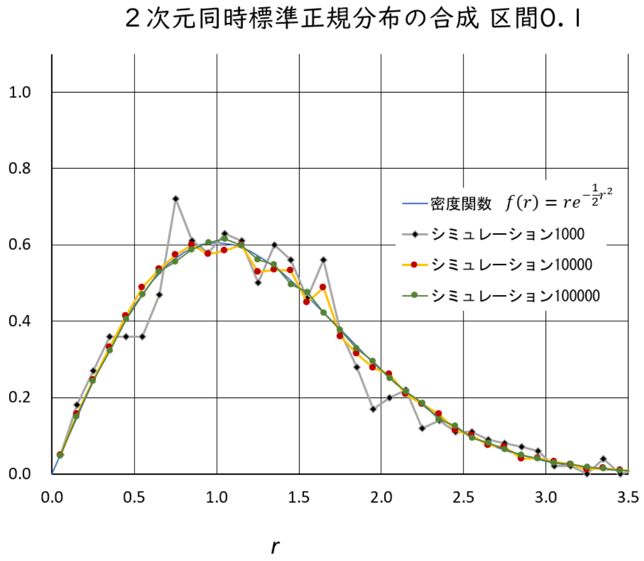
求めたい密度関数はなだらかなので、試行回数1000回のデータを移動平均（３つの平均）してみた。紫色をの線である。だいぶ、なだらかだが、１万回におよばない。

図17－11　移動平均



シミュレーション回数を10万回に増やした。回数による違いがわかりやすいように、グラフでは右すそを切って大きく表示した。10万回で、ようやく分布の合成（理論値）と同じ程度の精度になった。変数１つあたり1000回必要だとすると今回は２変数なので100万回になるが、それより１桁少なかった。この場合、一変数で300回程度必要だということになる。

図17－12　シミュレーション10万回



２）「分布のあてはめの検定」と「カーネル法」

シミュレーションを始めるとき、モンテカルロ法で乱数を発生するときの分布を、現実の分布を使って行なうこともできる。理論的分布に当てはめることができれば、その分布に基づいて乱数を発生させればいい。累積確率分布の逆関数を使えば、一様乱数から任意の分布の乱数がつくれる。

理論分布にあてはめることが許容できるか否かの検定には、「コルモゴロフ・スミノルフ検定」などがある。もし、却下されれば、現実データに近い分布をなぞる密度分布をカーネル法などでつくれる。累積分布ならそのまま得られる。「コルモゴロフ・スミノルフ検定」も累積分布を使う。

シミュレーション結果の分布をさらに別のシミュレーションに使うとき、上記の方法が使える。しかし、分布の合成が変数変換で解析的に求まれば、そのままシミュレーションに使える。ここにも分布の合成が有利な点がある。

付録３　統計分布の合成　変数変換

ヤコビ行列式を使った統計分布の合成方法を、シミュレーションに取り入れることができる。モンテカルロ法と統計学的分析を組み合わせることで、試行回数の大幅な節約が可能になる。シミュレーション結果の不安定性から逃れる手法の一つである。

ヤコビ行列式は、変数変換で使う方法だということだが、確率密度関数の間の演算を変数変換と見なして、２つの確率密度関数から新たな確率密度関数を求めることができる。モンテカルロ法で、確率変数が増えれば試行回数を爆発的に増やさないと、結果は安定しない。例えば、乱数

を発生させる変数が４つあるとすると、分布形を再現するにはおのおの1000回くらいは試行が必要で、４変数だと安定させるのに１兆回になる。必ずしもこの方法でできない場合があるようだが、できる範囲で行なって、残りは減った変数の数でシミュレーションすればいい。例えば、４変数が２変数に減れば、試行回数は100万回に減る。

●変数変換による分布の合成方法（一般）

変数変換によって統計分布の合成を行なう一般的な方法を紹介したい。モンテカルロ・シミュレーションとの関係についてもふれる。統計学の書籍やサイトには、変数変換のことがあまり載っていない。変数変換の中でも、統計分布の合成（演算）にあたる一変数に統合する方法はごく一般的な公式が書いてあるだけである。以下の本は例外的に、30頁ほどくわしく実例をあげて説明している。くわしいことを知りたい方は、ご覧いただきたい。なお、本文とは変数表記が違うが内容は同じである。

『１冊でマスター 大学の統計学』石井俊全　技術評論社　2018年

１）２つの確率密度関数から合成した確率密度関数を求める

２つの独立した確率密度関数

分布関数１　$f_{X_1}(x_1)$

分布関数２　$f_{X_2}(x_2)$

変数変換　２→２の一般的な場合

$Y_1 = g_1(X_1, X_2)$

$Y_2 = g_2(X_1, X_2)$

２つの分布の演算の場合は変換先は１変数なので

$Y_1 = g_1(X_1, X_2)$　　　X_1 について解くと $X_1 = g_0(Y_1, X_2) = g_0(Y_1, Y_2)$

$Y_2 = X_2$

２つの分布の合成（一般）

$$J = \begin{vmatrix} \partial x_1/\partial y_1 & \partial x_1/\partial y_2 \\ \partial x_2/\partial y_1 & \partial x_2/\partial y_2 \end{vmatrix}$$　ヤコビ行列式（ヤコビアン）

分布の合成の場合

$$J = \begin{vmatrix} \partial x_1/\partial y_1 & \partial x_1/\partial y_2 \\ 0 & 1 \end{vmatrix} = \frac{\partial x_1}{\partial y_1} = \frac{\partial g_0(y_1, y_2)}{\partial y_1}$$

変数変換は

$$f_{X_1}(x_1) f_{X_2}(x_2) dx_1 dx_2 = f_{X_1}\big(x_1 \to g_0(y_1, y_2)\big) f_{X_2}(x_2 \to y_2) |J| dy_1 dy_2$$

求めたい合成後の関数は、ダミーの y_2で積分した周辺分布である

$$h(y_1) = \int f_{X_1}\big(g_0(y_1, y_2)\big) f_{X_2}(y_2) |J| dy_2$$

一般的には、［−∞，＋∞］で積分

元の分布に限定的な積分区間があるときは、変数変換後に区間が変わるので注意が必要

例えば、一様分布は［０，１］で１であるが、その外側では０

$Y_1 = g_1(X_1, X_2) = X_1 / X_2$　　X_1 について解くと $X_1 = Y_1 X_2 = Y_1 Y_2$

$Y_2 = X_2$

ヤコビアンは

$$J = \begin{vmatrix} y_2 & y_1 \\ 0 & 1 \end{vmatrix} = y_2$$

求めたい合成分布

$$h(y_1) = \int f_{X_1}(y_1 y_2) f_{X_2}(y_2) |y_2| dy_2$$

具体的な確率密度関数を当てはめて、合成分布を求める。

【例2】標準正規分布どうしのわり算

分布関数１　$f_{X_1}(x_1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}{x_1}^2}$

分布関数２　$f_{X_2}(x_2) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}{x_2}^2}$

求めたい合成分布

$$h(y_1) = \int_{-\infty}^{\infty} f_{X_1}(y_1 y_2) f_{X_2}(y_2) |y_2| dy_2$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y_1 y_2)^2} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}{y_2}^2} |y_2| dy_2$$

$$= \frac{1}{\pi} \int_0^{\infty} e^{-\frac{1}{2}{y_2}^2({y_1}^2+1)} y_2 dy_2 \quad \text{左右対称なので２倍}$$

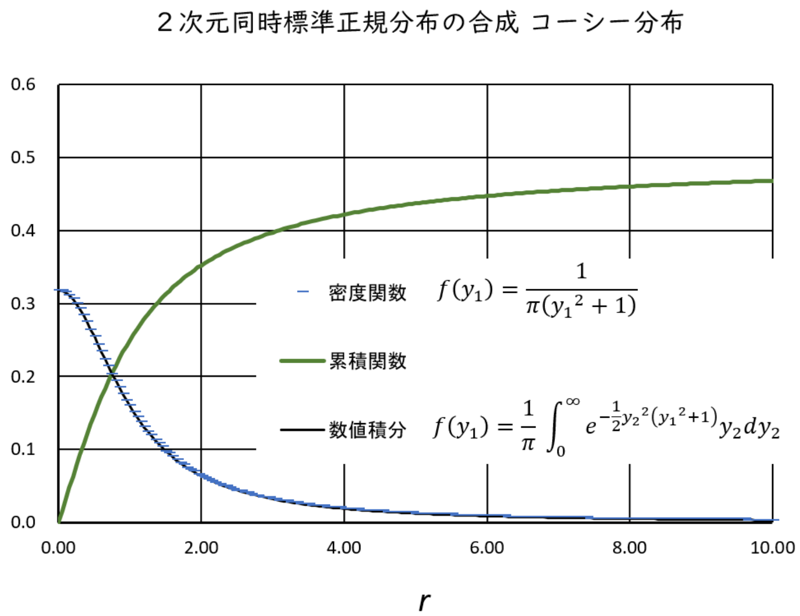
$$= \frac{1}{\pi({y_1}^2+1)} \int_0^{\infty} t e^{-\frac{1}{2}t^2} dt \quad t\ \text{で置換積分}\quad t = y_2\sqrt{{y_1}^2+1}$$

$$= \frac{1}{\pi({y_1}^2+1)} \left[-e^{-\frac{1}{2}t^2}\right]_0^{\infty} \quad \text{定積分は１}$$

$$= \frac{1}{\pi({y_1}^2+1)} \quad \text{コーシー分布＝自由度１の t 分布}$$

標準正規分布どうしの割り算は、コーシー分布になる。この場合は、上記一番下の式のように関数形が求まるが、数値積分するしかないときもある。あるいは、すべて数値積分ですませたほうが手間がかからない場合もある。以下のグラフで、最終的な関数形（黒い実線）と下から３行目の式で数値積分（青い横棒）を実行した場合を示した。グラフでは見分けがつかないほど、２つは一致した。参考に、累積分布も載せた。

図 17－13　分布の合成　コーシー分布



２）ｎ個の確率密度関数から合成した確率密度関数を求める

複雑なシミュレーションに対応する確率密度関数の合成は、一般的に３つ以上の変数をあつかうことも多い。シミュレーションでは安定的な関数形を得るには膨大な計算が必要であるが、関数の合成でも解釈がむずかしくなるのでむやみに変数は増やさないほうがいい。関数の合成の場合の計算量は、最大30^n程度である。積分はｎ-1回だが密度関数を求めるのにそれを30回程度必要。それほど複雑な密度関数でなければ、数値積分は30区間程度で十分であると思える。また、ｎ-1個の変数がすべて積分にからむとは限らないので、計算量はもっと少なくなる。

ヤコビアンは、変数が増えても行列式の一番左上を１回偏微分するだけである。X_1が偏微分しにくいときは、偏微分の分母と分子を入れ替えてY_1を偏微分すればいい。ただし、積分のときに変数を代入（これが変数変換）するのでX_1について解く必要はなくならない。

ｎ個の独立した確率密度関数

分布関数１　$f_{X_1}(x_1)$

分布関数２　$f_{X_2}(x_2)$

：

分布関数ｎ　$f_{X_n}(x_n)$

分布合成の場合は変換先は１変数なので

$Y_1 = g_1(X_1, X_2, \cdots, X_n)$　　　X_1 について解くと　$X_1 = g_0(Y_1, Y_2, \cdots, Y_n)$

$Y_n = X_n \quad n \geq 2$

ヤコビアン

$$J = \begin{vmatrix} \partial x_1/\partial y_1 & \partial x_1/\partial y_2 & \cdots & \partial x_1/\partial y_n \\ 0 & 1 & \cdots & \\ \vdots & \vdots & \ddots & \vdots \\ 0 & & \cdots & 1 \end{vmatrix} = \frac{\partial x_1}{\partial y_1} = \frac{\partial g_0(y_1, y_2, \cdots, y_n)}{\partial y_1}$$

変数変換は

$$f_{X_1}(x_1) f_{X_2}(x_2) \cdots f_{X_n}(x_n) dx_1 dx_2 \cdots dx_n$$
$$= f_{X_1}\big(g_0(y_1, y_2, \cdots, y_n)\big) f_{X_2}(y_2) \cdots f_{X_n}(y_n) |J| dy_1 dy_2 \cdots dy_n$$

求めたい合成後の関数は、ダミーで積分した周辺分布である

$$h(y_1) = \iint \cdots \int f_{X_1}\big(g_0(y_1, y_2, \cdots, y_n)\big) f_{X_2}(y_2) \cdots f_{X_n}(y_n) |J| dy_2 \cdots dy_n$$

３）３つの方法

関数の合成も途中で数値積分する場合があるので、シミュレーションを含めて３つの方法になる。

ａ）シミュレーションによる合成

ｂ）変数変換による密度関数の解析的合成

ｃ）変換変換による密度関数の数値積分を使った合成

経済に関するシミュレーションでは、出てきた分布を積分するなどして、１つの数値を求める事が多い。この場合、分布形の再現ではないので、試行回数は少なくできる。■