このドキメントは J14 論文の一部となる可能性がある日本語版下書きで、SonotaCo Network J14 非公開会議室の内部資料です。発表まで外部に漏れないようにお願いします。

この方法が有効かどうかは実際に流星データに適用するまで 本当に分かりません。まだ、この方法全体がボツになる可能性もあるので、よろしくお願いします。

流星群分類の統計学的有意性 (Statistical Significance for Meteor Shower Clustering)

Version: J02 2021/2/1

SonotaCo

Summary

集合内の要素の集中(クラスターと呼ぶ)の有意性を統計的に計算する方法を提案する。

本論文は流星群の決定に使用すること想定して、多くの人の直感に近い結果の数値を出来るだけ単純な数学によって誰でも計算できるようにすることを目標として検討したものである。

有意性の検定は、統計学における p 値を使用した帰無仮説による検定方法(Statistical null hypothesis testing method using p-value)を基本とするもので、あるクラスターが偶然の集まりでない確率をその有意性尺度(CSV(Cluster Significance Value)と呼ぶ) とするものである。

本論文における CSV は 1 つのインスタンスが偶然クラスターに含まれる確率 P1 と総インスタンス数とクラスター内インスタンス数のみから算出するものである。合成データによるシミュレーションでは例えば P1 が 0.25 の場合、p 値として 0.001 を採用すると主観的に肯定可能な結果であった。

CSV は集合の母数に依存して母数が大きいケースではより極端な有意性となる。CSV 演算においては、P1 の多数回のべき乗演算があるため、P1 が 0 または 1 に近い場合、または母数が大きい場合には演算上の有効数字の維持が困難な場合がある。CSV はクラスターの探索やクラスターの大きさの決定への利用が考えられるが、この 2 点についてさらなる検討が必要である。

Introduction

物事の確からしさの数値による表現としては、統計学上の p 値を使用した帰無仮説による検定方法がある。これは、ある命題が偶然によって成立してしまう確率を計算し、それが p 値以下の時に命題の成立が有意であると結論するものである。p 値は用途によって随時定められるもので、一般に社会学上は 0.05 または 0.01 が使用されることが多く、科学ではそれより遥かに小さい数値が使わることがある。例えば、重力子の発見などの素粒子物理学では観測を発見と認証する際に、偶然の結果である確率が 5 σ の外、すなわち、p=0.0000006 (発見が偶然でない確率が 99.99994%以上) を要求している。このように、帰無仮説における確率計算数値は、物事の有意性を数値により客観的に判断するかめに社会の中で広く使用されており、流星の群判定においても有用性が高いと考えられる。

流星群の決定における帰無仮説による検定とは、流星群に含まれない散在流星が一定の範囲でランダムに発生すると仮定して、あるクラスターが散在流星によって偶然できる確率 Q を計算し、Q が一定値以下であるときそのクラスターを流星群とみなし、1-Q を CSV とするという方法である。

CSV の計算のためには、1 つのインスタンスが偶然クラスターに入るか否かの確率 (P1) が算出されている必要がある。P1 はそれぞれの分野において直感に近い形で計算されることが望ましい。ここでは、多数の流星観測結果からの流星群を決定する場合を例として、空間内密度を基準とした尺度を用いて、有意性の計算結果を例示したい。

クラスター内出現確率 P1 の定義例

流星群の判定においては、図 1 に示すように散在流星がある空間内でランダムに発生すると仮定し、この空間内に流星群に含まれる流星がクラスターとして出現するモデルを考える。1 つの流星の情報を 2 次元空間内の 1 つの位置として表現する場合、1 つの散在流星が 1 つの流星群クラスターに偶然含まれてしまう確率 P1 は、クラスター領域の面積 c と c を含みその周辺で均等に散在流星が発生しうる背景領域の面積 a との比 c/a によって計算できる。背景領域 a の設定は任意であるが、出現の均等性と直感的な分かりやすさを考慮して、クラスター領域を同心円として含む 2 次元空間上の円を想定し、P1=0.25 となるように 2 倍の半径をもつ円として周辺領域を定義することを想定する。

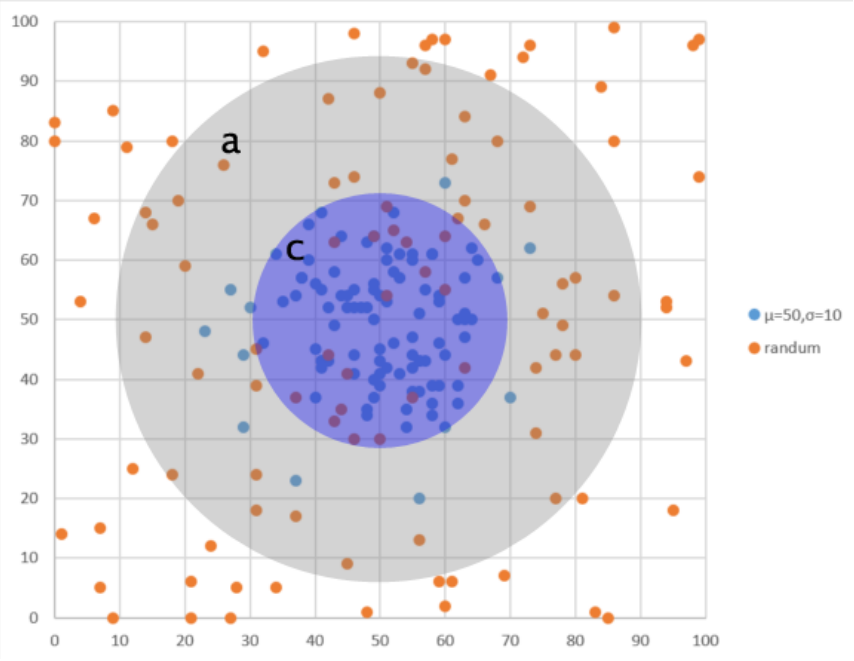


図 1 2 次元空間におけるクラスターの例

Area	Area size	All instance count	Shower instance count	Random instance count
a : all	1.0	150	100	50
b: outside	0.75	44	12	32
c: cluster	0.25 = P1	106	88	18

表 1 図 1 の例の面積比とインスタンスのプロット個数

偶然によるクラスター発生確率とクラスターの有意性

今、a 内の全インスタンス数を N, 1 つのインスタンスが特定の領域 c に入る確率を P1 とすれば、c 内に n 個のインスタンスが偶然入る確率p_n は以下の式で表される。

$$p_n = \binom{N}{n} \cdot {P_1}^n \cdot (1 - P_1)^{N-n} \quad \dots (1)$$

c 内に偶然 n 個以上が入る確率 Qn は

$$\mathrm{Qn} = \sum_{x=n}^{N} p_x \quad \dots (2)$$

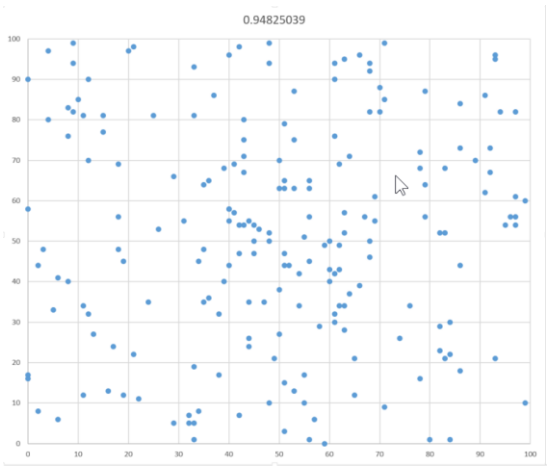
であり、このクラスターの有意性 Sn は(3)式となる。

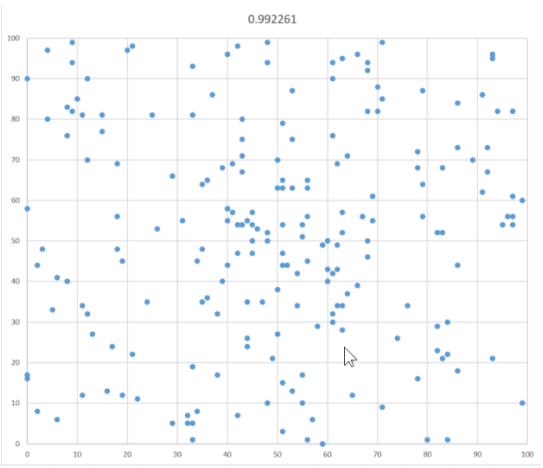
$$\mathrm{Sn} = 1 - \mathrm{Qn} = 1 - \sum_{x=n}^{N} \binom{N}{n} \cdot {P_1}^x \cdot (1 - P_1)^{N-x} \quad \dots (3)$$

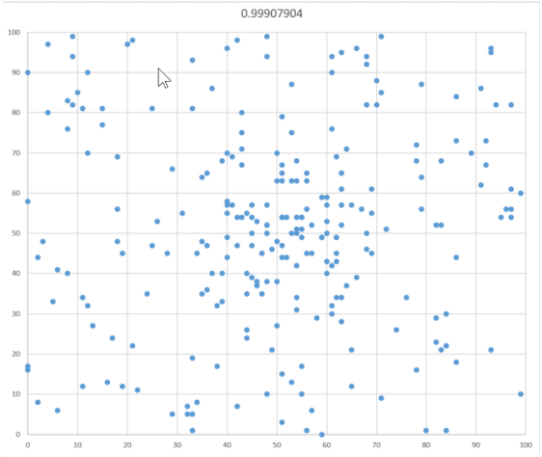
図 1 の例は N=150, n = 106, P1=0.25 であるから、これらを (3) 式に代入すると、この分布が偶然できる確率が 10 の-15 乗以下であり、このクラスターとしての有意性 S_{106} は 1.0 すなわち疑う余地が全くない有意なクラスターであることがわかる。

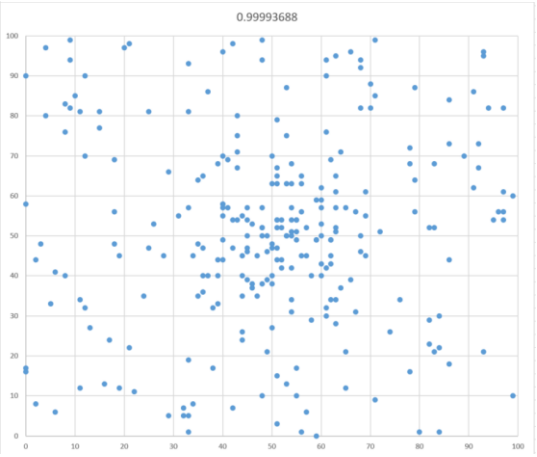
ちなみに、N=150、P1=0.25 の場合、偶然による c 内のインスタンス数 n の期待値は 37.5 であり、Sn が 0.95 以上 となる n は 46、0.99 以上 となる n は 55、0.9999999 以上となる n は 68 である。直感的な理解のため、CSV=0.95 から 0.999999 を超えるまでの分布の例を図 2 に示す。

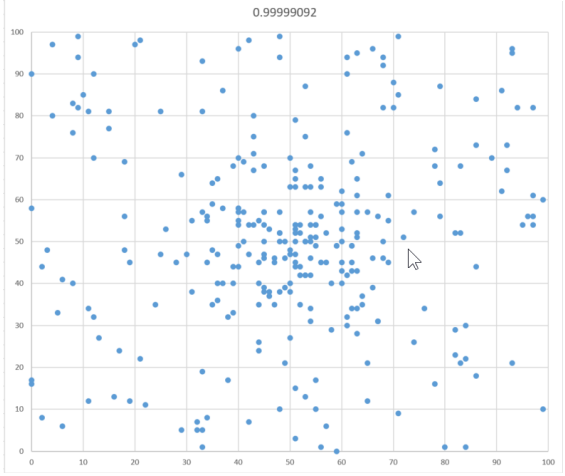
特定の流星群カタログを作成する際には採用する流星群について CSV の基準値を設定し全体の有意性を統一することが期待される。この際に使用する CSV は任意であり、今後実観測結果を元にした検討が期待されるが、図 2 からの主観的印象では 明確なクラスターとして認証するためには、例えば、p 値にして 0.001 すなわち CSV=0.999 以上(スリーナイン)が 1 つの案として考えられる。

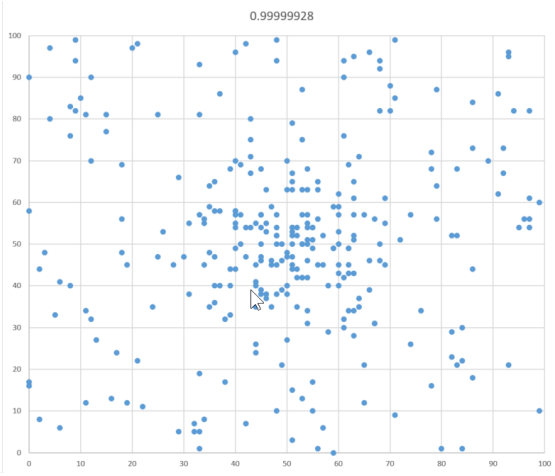

0.94825039


0.992261


0.99907904


0.99993688


0.99999092


0.99999928

クラスター強度と母数と有意性の関係

インスタンス出現密度 Dc=n/c と Da=N/a の比 Rd = Dc/Da はクラスターの出現の強さを表している。この Rd が大きいことは一般に高い集中度合いと考えられるが、必ずしも統計的な有意性が高いことを表さない点に注意が必要である。同じ母数に対しては Rd が大きければ、有意性 CSV も大きい。しかしながら、同じ Rd であっても、母数が小さい場合には CSV は小さい。これは全体の数が少ないと Rd が大きくてもそれが偶然である確率が大きいためである。直感的にも数個が集中してもそれが意味ある集中かどうか疑わしいが、数千個が同様に集中したら、それは偶然ではないと感じられる。逆にいうと、同じ CSV となるためには、母数が小さい場合にはより強い集中が必要である。P1=0.25, CSV=0.99 を例にとると、母数が 1000 個の場合にはその 26% が集中すればよいが，母数が 100 個の場合には 28%, 母数が 10 個の場合には 70%が集中することが必要となる。必要な集中の度合いは P1 により異なる、P1=0.5 の場合を図 3 に示す。

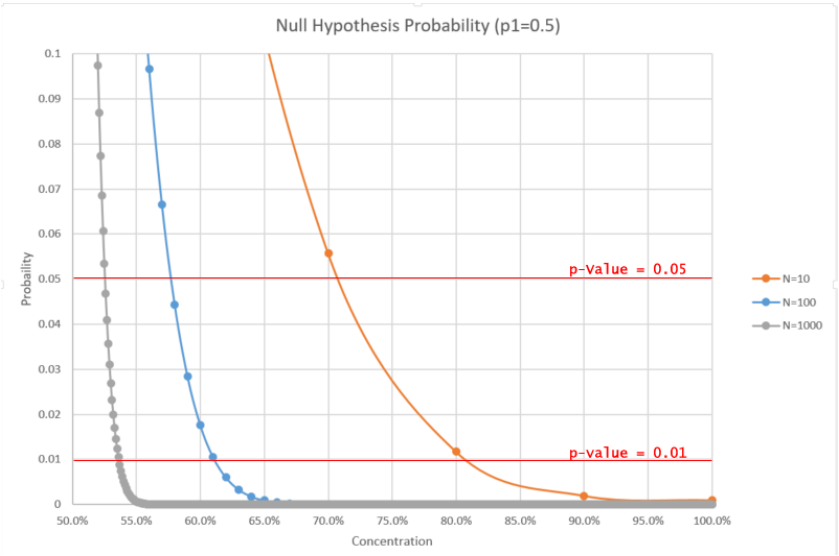


図 3 同一の p 値に対するクラスター内比率の母数による違い

クラスター半径決定における CSV の利用

任意のクラスターの有意性または強さが計算できる場合、これを用いた繰り返しの試行により、クラスターの発見やクラスター半径の最適値の決定を行うことができる。しかしながら、本論文で述べる CSV は母数に依存するため、母数の違いにより、直感的な集中度とは一致しない結果を生むことがある。クラスターが存在する場合にその半径を大きくとると母数が増加して CSV が上昇する例などである。また、P1 および(1-P1)の母数に応じたべき乗演算と階乗演算を必要とするため、実用上有効数字が不足する場合がある。P1 が 0 または 1 に近い場合や集中が明確なクラスターなどではどのように計算しても CSV は 1.0 または 0.0 に極めて近くなり、p 値による判定は可能だが、値の比較は困難になるわけで

ある。クラスターの発見やその半径の探索にはより単純なクラスターの強度の利用も有望であり、今後の検討が必要である。

流星群判定への適用にむけて

クラスターの有意性とある要素がクラスターに属するか否かの判定は独立した問題である。クラスター間関係の発見もまた別の命題である。

1 つの事象は多数の属性をもっている。例えば、流星の類似性は、輻射点方向(RaDc)、地心速度(Vg)、出現太陽黄経(Ls)という直接的な観測結果、あるいはそこから算出された日心軌道要素(a, q, e, peri, node, incl)など利用可能な多数の値をもっている。一般に多面的な性質をもつ事象のクラスターの発見においては、事象のもつすべての面で同時に集中することは要求されない。一部の次元の特定の評価方法において有意に集中していることがクラスターであることの十分条件である。もちろん他の次元併用することにより、偶然紛れ込むノイズを減らすことができる場合もある。流星群の例では、RaDc で有意に集中していれは、Ls や Vg が十分に集中していなくてもクラスターと判定できる。Ls は軌道の進化による広がりをもち、Vg は測定誤差により大きく分散していることがあるからである。しかし、L-Ls として知られる座標変換や、太陽黄経と速度を特定の範囲に限定した母集団上でクラスターの評価をすれば、より淡い集中も発見できる可能性がある。さらには、RaDc 距離と Vg の差の 2 次元上で評価すれば、より精度の高い判定ができる可能性がある。また D’尺度などの 2 軌道間類似性をもとにしたクラスタリングを行えば、別の集中を発見することや、クラスター間の関係を発見できる可能性がある。

今後、本 CSV 算出方法を実際の流星観測データに適用し、その有効性と問題点を検討したい。