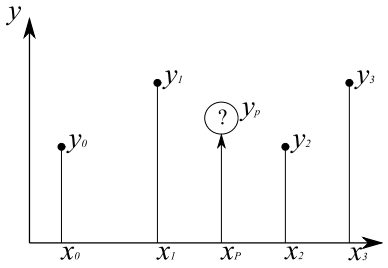
# 1　データ内挿

## 1.1　線内挿

既存のデータや計測等によって多くのデータが得られたとしても，目的となる場所のデータが存在しないことがしばしばある．このようなときに目的となる場所のデータを推定するのが**内挿** (interpolation) である．内挿は，**補間**と呼ばれることもある．

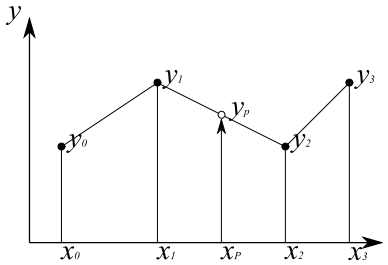
下図のように，データ $(x_0, y_0) \cdots (x_3, y_3)$ があるが，$(x_1, y_1)$ と $(x_2, y_2)$ との間にある $x_p$ での $y_p$ の値は，どのように推定するかという問題である．




ここでは，幾つかの代表的な内挿手法を紹介する．

### 1.1.1　線形内挿

線形内挿は，下図のように単純に点と点の間を直線で結んで内挿するという手法である．




データが $n$ 個，$(x_0, y_0) \cdots (x_n, y_n)$ とあり，$i$ 番目と $i+1$ 番目の間の $x_p$ における $y_p$ の値を推定するためには，$(x_i, y_i)$ と $(x_{i+1}, y_{i+1})$ とを結ぶ直線の式をまず求める．方程式型で表すと，次式で表すことができる．

$$y = \frac{y_{i+1} - y_i}{x_{i+1} - x_i} x + y_i - \frac{y_{i+1} - y_i}{x_{i+1} - x_i} x_i \tag{1}$$

したがって，$y_p$ は，$x = x_p$ を代入して計算すれば良い．

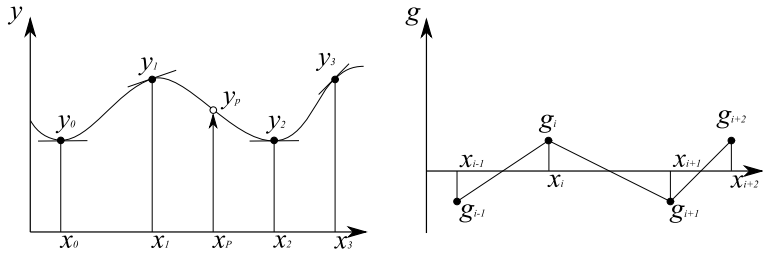
### 1.1.2　スプライン

点と点とを直線ではなく，下図の左のように，滑らかな曲線で結ぶ手法もある．その代表的な手法が**スプライン** (Spline) と呼ばれる手法である．スプラインは，各点と点の間を三次関数によって表現

する．例えば，$x_i$ と $x_{i+1}$ に囲まれる区間の三次関数を $f_i(x)$ とすると，次式で表される．

$$f_i(x) = a_i x^3 + b_i x^2 + c_i x + d_i \tag{2}$$

ここで，$a_i, b_i, c_i, d_i$ は，係数である．このとき，滑らかに繋ぐ必要があるので，各点においては，一階導関数及び二階導関数の値が，左右の曲線ともに同じ値をとる必要がある．この条件の元，三次式の係数を決定するわけだが，その計算は少々工夫が必要となる．




様々な方法が提案されているが，「科学技術計算ハンドブック」に解きやすい方法が掲載されていたので，その手法を紹介する．

$x_i$ と $x_{i+1}$ に囲まれる区間の三次関数を $f_i(x)$ の二階導関数は，一次関数となるが，上図の右のように $x_i$ における二階導関数の値を $g_i$，$x_{i+1}$ における二階導関数の値を $g_{i+1}$ とすると，この二階導関数は，次式で表すことができる．

$$\begin{aligned} f_i''(x) &= \frac{g_{i+1} - g_i}{x_{i+1} - x_i} x + g_i - \frac{g_{i+1} - g_i}{x_{i+1} - x_i} x_i \\ &= g_i + (x - x_i) \frac{g_{i+1} - g_i}{x_{i+1} - x_i} \end{aligned} \tag{3}$$

これを積分すると，一階導関数が導ける．

$$f_i'(x) = f_i'(x_i) + g_i(x - x_i) + \frac{1}{2}(x - x_i)^2 \frac{g_{i+1} - g_i}{x_{i+1} - x_i} \tag{4}$$

さらに積分することで，求めたい三次関数 $f_i(x)$ となる．つまり，$g_i, g_{i+1}$ が計算できれば，三次関数が求まることになる．

$$f_i(x) = f_i(x_i) + f_i'(x_i)(x - x_i) + \frac{1}{2} g_i (x - x_i)^2 + \frac{1}{6}(x - x_i)^3 \frac{g_{i+1} - g_i}{x_{i+1} - x_i} \tag{5}$$

ここで，$x = x_{i+1}$ を代入すると，次式を得る．

$$f_i(x_{i+1}) = f_i(x_i) + f_i'(x_i)(x_{i+1} - x_i) + \frac{1}{2} g_i (x_{i+1} - x_i)^2 + \frac{1}{6}(x_{i+1} - x_i)^3 \frac{g_{i+1} - g_i}{x_{i+1} - x_i} \tag{6}$$

この式を用いて，右辺の第二項にある一階導関数を求めるべく整理すると，次式となる．

$$\begin{aligned} f_i'(x_i) &= \frac{f_i(x_{i+1}) - f_i(x_i)}{x_{i+1} - x_i} - \frac{1}{2} g_i (x_{i+1} - x_i) + \frac{1}{6}(x_{i+1} - x_i)(g_{i+1} - g_i) \\ &= \frac{y_{i+1} - y_i}{x_{i+1} - x_i} - \frac{1}{6}(x_{i+1} - x_i)(g_{i+1} - g_i) \end{aligned} \tag{7}$$

なお，$f_i(x_i), f_i(x_{i+1})$ については，データの値が利用でき，それぞれ $y_i, y_{i+1}$ なので，それを用いた．

次に，もともとの一階導関数，式??においても，同様に $x = x_{i+1}$ を代入して整理すると，次式のようになる．

$$\begin{aligned} f_i'(x_{i+1}) &= f_i'(x_i) + g_i(x_{i+1} - x_i) + \frac{1}{2}(x_{i+1} - x_i)^2 \frac{g_{i+1} - g_i}{x_{i+1} - x_i} \\ &= f_i'(x_i) + \frac{1}{2}(x_{i+1} - x_i)(g_{i+1} + g_i) \qquad \text{ここで，} f_i'(x_i) \text{ に式??を代入} \\ &= \frac{y_{i+1} - y_i}{x_{i+1} - x_i} - \frac{1}{6}(x_{i+1} - x_i)(g_{i+1} - g_i) + \frac{1}{2}(x_{i+1} - x_i)(g_{i+1} + g_i) \\ &= \frac{y_{i+1} - y_i}{x_{i+1} - x_i} - \frac{1}{6}(x_{i+1} - x_i)(2g_{i+1} + g_i) \end{aligned} \tag{8}$$

この式を用いて，変数の添字を 1 小さくすることで，左区間での一階導関数を求める．

$$f_i'(x_i) = \frac{y_i - y_{i-1}}{x_i - x_{i-1}} - \frac{1}{6}(x_i - x_{i-1})(2g_i + g_{i-1}) \tag{9}$$

この式が，式??と等しくなるので，整理すると次式を得る．

$$\begin{aligned} \frac{y_{i+1} - y_i}{x_{i+1} - x_i} - \frac{1}{6}(x_{i+1} - x_i)(g_{i+1} - g_i) &= \frac{y_i - y_{i-1}}{x_i - x_{i-1}} - \frac{1}{6}(x_i - x_{i-1})(2g_i + g_{i-1}) \\ 6\left(\frac{y_{i+1} - y_i}{x_{i+1} - x_i} - \frac{y_i - y_{i-1}}{x_i - x_{i-1}}\right) &= (x_{i+1} - x_i)(g_{i+1} - g_i) - (x_i - x_{i-1})(2g_i + g_{i-1}) \\ &= (x_i - x_{i-1})g_{i-1} + 2(x_{i+1} - x_{i-1})g_i + (x_{i+1} - x_i)g_{i+1} \end{aligned} \tag{10}$$

最終的には，$g_{i-1}, g_i, g_{i+1}$ の値を求めることになる．$i = 1, 2, 3$ を代入すれば，式が 3 つできるのに対して未知変数は，$g_0, g_1, g_2, g_3, g_4$ の 5 つとなり，解くことができない．これを解決するのに，両端の二階導関数については，情報がもともと不足しているので，未知数の $g_0, g_4$ の値を適当に与えることで対応する．例えば，$g_0 = g_4 = 0$ とおけば，未知数は 3 つとなり，連立方程式を解くことができる．

解いて得られた二階導関数の値 $g_1, g_2, g_3$ を式??に代入して整理すれば，三次関数を導くことができる．この例は，3 点のデータなので，3 つの式を連立させて解けば良いので簡単だが，点数が多くなると膨大な式を連立させて解く必要がでてくる．しかし，区間を分けて計算しても構わないので，プログラムを書くときには，コンピュータのメモリに応じて区間を設定すれば良い．

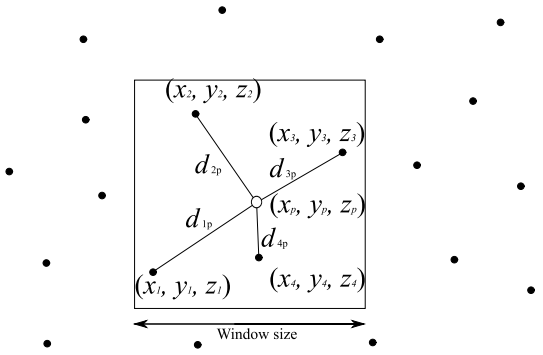
## 1.2 面内挿

### 1.2.1 重み付き平均

これまで解説した線形内挿，スプラインは，線を対象とした内挿であった．地理情報は，面的な広がりを持つデータであることが普通なので，ここでは面的な内挿について解説する．下図のようにランダムな $n$ 個の三次元のデータが $(x_1, y_1, z_1), \cdots, (x_n, y_n, z_n)$ あるとき，任意の点 $(x_p, y_p, z_p)$ の値を推定することを目的とする．例えば，$z$ が標高データで，そのデータがランダムポイントのデータでしかないとき，これをグリッド型に変換するときに内挿が必要となる．つまり，$(x_p, y_p)$ が与えられたとき，$z_p$ はどんな値をとるかを推算することとなる．これは，画像処理の章で解説した再配列の

手法と似ている．最近隣法により最も近い値で代表できる場合もあるであろうし，共一次内挿法のように平均値を利用する場合もあるであろう．ただ，共一次内挿法は，元のデータが等間隔で並んでいるようなグリッドデータのときには適用できるが，ランダムに配置されたデータには適用できない．そこで，重み付き平均による内挿が必要となる．

下図を用いて，重み付き平均による内挿手法を解説する．この図は，ランダムポイントデータ点データがたくさんあり，内挿したい点 $(x_p, y_p, z_p)$ を中心にある一定範囲のウィンドウの中に，4 個のデータが存在している様子を表している．




これら 4 つのデータから $z_p$ の値を求めるものであるが，単に 4 つのデータの平均値を求めるのは問題となる．それは，x-y 平面上で，$(x_p, y_p)$ に近い $(x_4, y_4)$ は $z_p$ に対して大きな影響を及ぼし，逆に最も遠い $(x_2, y_2)$ はあまり影響を及ぼしていないので，これを考慮する必要があるのである．そこで，平均値を計算するときに，重みを距離に応じて設定した上で，計算することで，より現実に近い内挿ができるわけである．重み付き平均の計算は，内挿に用いるデータ数を $n$，各データに対する重みを $w_i$ とすると，次式で表すことができる．

$$z_p = \frac{\sum_{i=1}^{n} w_i z_i}{\sum_{i=1}^{n} w_i} \tag{11}$$

ここで，問題となるのは $w_i$ をどのように計算するかである．今回，距離が近いほど大きい重みにする必要がある．ある点 $(x_i, y_i)$ と $(x_p, y_p)$ との距離 $d_i$ は，簡単に計算できるので，それを重みにするときは，その逆数を用いれば良い．式で表すと，以下のようになる．

$$w_i = \frac{1}{\sqrt{(x_i - x_p)^2 + (y_i - y_p)^2}} \tag{12}$$

この重みの計算法は，様々なものが提案できる．例えば，近い距離のデータに，より大きい重みを与えたい場合には，距離計算における平方根を外せばよい．次項で解説するクリッギングという方法で，データのばらつきを考慮した重みを計算することもできる．いずれにしても，内挿する対象によって重みの計算法は，慎重に選択すべきである．

一方，重みの計算法だけでなく，内挿計算のためのデータ数も考慮しなければならない．上図の例

では，４つのデータから内挿するものであった．しかし，ウィンドウの範囲を大きくして，内挿計算の範囲を広げれば，多くのデータを用いて内挿計算することになり，なだらかな変化をする内挿結果となる．逆に，ウィンドウの範囲を小さくすれば，少ないデータを用いて内挿計算することになり，最近隣法に近い結果となる．このように，ウィンドウサイズを変更することでも内挿結果は変わってくるので，適したウィンドウサイズを設定しなければならない．データの空間分布が均等でなく，偏りがある場合は，ウィンドウサイズを固定して内挿計算を行うよりも，データ数を固定してウィンドウサイズを変化させながら内挿計算を行う方が現実的である．

### 1.2.2 クリッギング

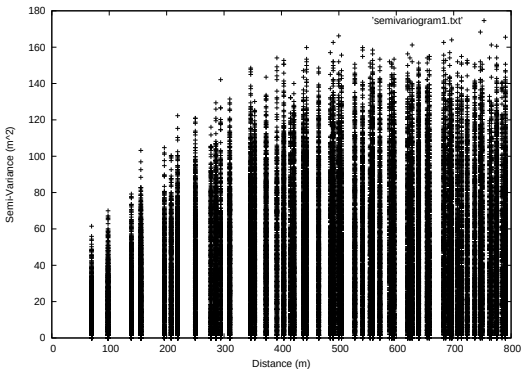
**クリッギング**は，地質学の分野で発達した重み付き平均の一種の内挿手法である．内挿する点において，周辺のデータのばらつきを考慮して重みを決定し，内挿するもので，近年様々な分野で利用されている．

クリッギングでは，データのばらつきを判断するのに，分散の値ではなく，**半分散**が利用される．分散は，データが正規分布しているものを対象に，平均値と各データとの差の二乗和で計算されるが，様々なデータにおいて，ある一定範囲のデータは，正規分布しているとは限らない．そこで，半分散という指標でばらつきを表している．二つのデータ $(x_i, y_i, z_i)$ と $(x_j y_j, z_j)$ とがあるとき，半分散 $\Gamma$ は，データ間の距離 $d_{ij}$ の関数で表され，次式により計算される．

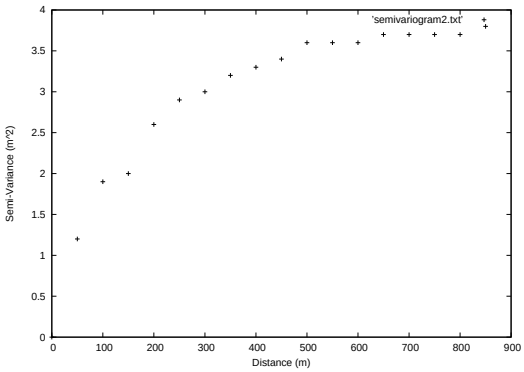
$$\Gamma(d_{ij}) = \frac{(z_i - z_j)^2}{2} \tag{13}$$

二点間の差の二乗をデータ数で割ったものであり，データ間の距離 $d$ に対する変化量と見なすことができる．距離 $d_{ij}$ は，$d_{ij} = \sqrt{(x_i - x_p)^2 + (y_i - y_p)^2}$ で計算できるが，多数のランダムポイントデータの場合，$d_{ij}$ は様々な値をとり，同じ値となるものはほとんどない．また，$d_{ij}$ が似たような値であっても $\Gamma$ は様々な値を示す．ただ，$d$ が小さい場合は，$\Gamma$ も小さくなり，$d$ が多きくなりと，$\Gamma$ も大きくなる傾向は想像できる．この $\Gamma$ の変化量は対象によって様々で，標高を例にとると，平野では変化量は少なく，山間部では変化量が大きい．また，起伏が周期的に変化しているような場所では，最初は急に変換するものの，やがて変化量が少なくなる．このような状況を考慮して内挿を行うのがクリッギングである．

半分散は，２つのデータより計算するものなので，データ数が多くなると，すべての組み合わせ分の半分散 $\Gamma(d_{ij})$ を計算しなければならない．データ数が $n$ のときは，${}_nC_2$ 組の半分散が計算できる．下図は，四国の数値地図 50m メッシュ標高データにおいて，松山市付近の 1km 四方のデータをもとに半分散をプロットしたものである．横軸が距離，縦軸が半分散である．グリッド型のデータを用いて計算したため，距離の値は飛び飛びの値をとなっている．このグラフは，**セミバリオグラム**と呼ばれている．このグラフには，非常に多くの点がプロットされているが，各距離における半分散の最大値は，距離が 500m より大きくなると，ほぼ一定になっていることがわかる．距離が長くなると，二点間の標高差は，非常に近いものもあるが，標高差の最大値は，あまり変化しないことが言える．




このセミバリオグラムを解りやすく表現するために，ある一定距離の範囲ごとに半分散を平均化したものが下図である．




もとが 50m メッシュのグリッドデータなので，50m ごとに区切って平均値を求めている．この平均値を各距離における半分散の値として利用する．重み $w_i$ に関しては，４つの点を用いて内挿する場合，セミバリオグラムから各点間の距離における半分散の値を求めると，次式が成り立つ．

$$
\begin{aligned}
\Gamma(d_{1p}) &= w_1\Gamma(d_{11}) + w_2\Gamma(d_{12})w_3\Gamma(d_{13}) + w_4\Gamma(d_{14}) \\
\Gamma(d_{2p}) &= w_1\Gamma(d_{21}) + w_2\Gamma(d_{22})w_3\Gamma(d_{23}) + w_4\Gamma(d_{24}) \\
\Gamma(d_{3p}) &= w_1\Gamma(d_{31}) + w_2\Gamma(d_{32})w_3\Gamma(d_{33}) + w_4\Gamma(d_{34}) \\
\Gamma(d_{4p}) &= w_1\Gamma(d_{41}) + w_2\Gamma(d_{42})w_3\Gamma(d_{43}) + w_4\Gamma(d_{44})
\end{aligned} \tag{14}
$$

それぞれの $\Gamma(d_{ij})$ の値は，上のグラフより求めて代入すると，未知数が $w_1, \cdots, w_4$ の連立方程式となる．この連立方程式を解いて，重み $w_i$ を求め，式**??**を用いれば，$z_p$ の値が求まる．

さて，半分散を計算するとき，対象範囲全体にわたってデータが同じようなばらつきを持つものは，対象範囲全体でセミバリオグラムを作成する．しかし，対象範囲が非常に広いときには，場所によってばらつきが異なる場合もあると思われる．そのようなときは，あるウィンドウサイズを設定し，場所に応じて，その内部にあるデータに限定してセミバリオグラムを作成することで対応する．例えば，標高データを内挿する際，地形の形状によってばらつきが異なる．同じ山間部でもなだらかな場合と急峻な場合とで区別する必要がある．このような場合には，地形分類結果をもとに区域ごとにセミバ

リオグラムを用意すべきであろう.