2010

確率と最小二乗法による土地の境界復元

誤差の基礎Ｉ（１変数）

誤差について基準点測量の観点から見た場合と点の位置誤差から見た場合があります。

基準点測量では専門書が多くあります、ここでは点の位置誤差、つまり境界標の位置誤差の観点から、測量士ではなく土地家屋調査士の考えを解説してみます。

位置誤差の解説にはそれほど専門書がありませんので少しでも参考になればよいと思います。

これを機会にいろいろと考えて頂いて、中には今までいわれていることと違った解釈もあります、それについてはそれなりに検証しました。

誤差とは

誤差とは

測定値－真の値＝誤差

真値の代わりに平均値（最確値）を使う

最確値との差

測定値－最確値＝残差

平均値には算術平均値と最確値があります、算術平均は単純に総和をｎ（データ数）で割ったもの、最確値とは最も真値に近い値を指しますが実際の測量データは正規分布になることが知られていますので確率論によって選ばれたデータの平均値を最確値とします。

あるいは

期待値との差

間接距離－直接距離＝残差

たとえばa点とｂ点の距離で考えるとa点とｂ点を直接測った距離が期待値で角と距離の測定によって得られた座標値の間接距離が観測値といえます。直接距離が10.000で間接距離が10.005とすればこの点間の誤差は0．005といいます。

図面値－実測値＝残差

また、図面に書いてる距離（距離の数値でも座標値のST距離でも同じ）と実際に測量して得られた距離（直接距離の数値でも座標値のST距離でも同じ）との差も誤差と言います。

誤差とは真値と測定値の差であると言われますが真値は解りませんので真値の代わりに期待値又は平均値（最確値）を使うのが一般的です。

平均値には算術平均値と最確値があります、算術平均は単純に総和をｎ（データ数）で割ったもの、最確値とは最も真値に近い値を指しますが実際の測量データは正規分布になることが知られていますので確率論によって選ばれたデータのへ平均値といえます。

期待値との差、たとえばa点とｂ点の距離で考えるとa点とｂ点を直接測った距離が期待値で角と距離の測定によって得られた座標値の間接距離が観測値といえます。直接距離が10.000で間接距離が10.005とすればこの点間の誤差は0．005といいます。

また、図面に書いてる距離（距離の数値でも座標値のST距離でも同じ）と実際に測量して得られた距離（直接距離の数値でも座標値のST距離でも同じ）との差も誤差と言います。

標準偏差の詳しいことは別途説明しますが期待値との差又は最確値との差、この差のバラツキを表す指標が標準偏差です、標準偏差はバラツキの指標であると同時に精度を表しています。

確率論によって選ばれたデータ とは

確率論によって選ばれたデータとは

　確率論によって選ばれるとは具体的にどのような手法なのでしょうか、簡単にいえば得られたデータから正規分布になじまないデータ、異常なデータを除くということです。

得られたデータの中から異常に飛び出た数字のデータを除くということですが次のデータでどのデータが異常か判断できますか。

①

名	データ
1	15.15
2	15.12
3	15.11
4	15.09
5	15.08
6	15.07
7	15.05
8	15.05
9	15.05
10	15.04

②

名	データ
1	16.20
2	15.12
3	15.11
4	15.09
5	15.08
6	15.07
7	15.05
8	15.05
9	15.05
10	15.04

③

名	データ
1	15.13
2	15.12
3	15.11
4	15.09
5	15.08
6	15.07
7	15.05
8	15.05
9	15.05
10	15.04

一連のデータでデータ名1の数値を①15.15　②16.20　③15.13　にしてあります。

確率論によって選ばれるとは具体的にどのような手法なのでしょうか、簡単にいえば得られたデータから正規分布になじまないデータ、異常なデータを除くということです。

得られたデータの中から異常に飛び出た数字のデータを除くということですが次のデータでどのデータが異常か判断できますか。

②の１は異常に飛び出たデータであることは見た目にも解りますが①の１は判断に迷います、③の１は問題ないだろうといえます。

これを経験とか勘でなく統計的に判断することが確率論によって・・・ということです。

検定という手法 1

T検定体験プログラム

t 検定（t 検定の内容はχ²検定と t 検定のファイルで解説してあります）

統計では異常な値を除く検定という手法があります。この検定を使って異常な値を除いてみます。

データの平均値との差を検定する。

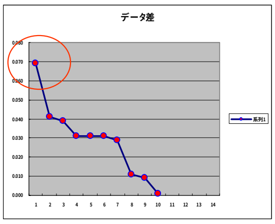
点名	データ	データ順
a1	15.15	1
a2	15.12	2
a3	15.11	3
a4	15.09	4
a5	15.08	5
a6	15.07	6
a7	15.05	7
a8	15.05	7
a9	15.05	7
a10	15.04	10

点名	データ	abs偏差	差の順

点名	データ差

①のデータでは点a1が異常（t期待値とt観測地を比較して判断します）

個数	データ差	期待値t	観測値T	平均	標準偏差	τ	判定結果	点名
10	0.069	2.306004	2.963226	0.029	0.018340	2.170	×	a1
9	0.041	2.364624	1.258968	0.025	0.013348	1.215		a10
8	0.039	2.446912	1.3418	0.023	0.012784	1.271		a2
7	0.031	2.570582	0.862989	0.020	0.011986	0.882		a7
6	0.031	2.776445	1.02659	0.019	0.012078	1.021		a8
5	0.031	3.182446	1.400136	0.016	0.011771	1.257		a9
4	0.029	4.302653	3.600595	0.012	0.010235	1.612		a3
3	0.011	12.7062	0.866025	0.007	0.004320	0.926		a6
2	0.009			0.005				a4
1	0.001							a5



ここではｔ検定を使って異常値がないか確認する方法を紹介します、詳しくは「χ²検定とｔ検定」のファイルで解説してあります。

「t検定の体験」のプログラムをこのファイルのところにアップしてあります。

エクセル2000、2002、2003、2007とあって2007がダウンロードに注意が必要です。

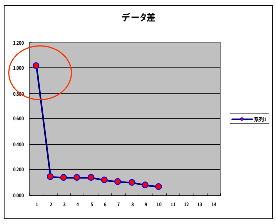
このプログラムを使ってｔ検定を体験してください。点名、データに入力して計算実行ボタンで計算されます。

①のデータではa1が異常と判断されます。

検定という手法　2

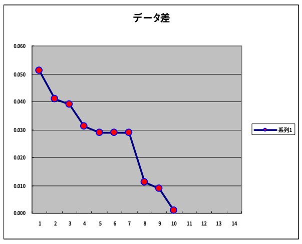
②のデータでは点a1が異常

個数	データ差	期待値t	観測値T	平均	標準偏差	τ	判定結果	点名
10	1.014	2.306004	29.77006	0.203	0.271618	2.987	×	a1
9	0.146	2.364624	1.278872	0.113	0.027080	1.231		a10
8	0.136	2.446912	1.075198	0.108	0.025860	1.063		a2
7	0.136	2.570582	1.314751	0.105	0.025314	1.242		a7
6	0.136	2.776445	1.936992	0.099	0.023570	1.556		a8
5	0.116	3.182446	1.469694	0.092	0.018547	1.294		a9
4	0.106	4.302653	1.511858	0.086	0.015811	1.265		a3
3	0.096	12.7062	2.886751	0.079	0.012472	1.336		a6
2	0.076			0.071				a4
1	0.066							a5



③のデータでは異常はなし

個数	データ差	期待値t	観測値T	平均	標準偏差	τ	判定結果	点名
10	0.051	2.306004	1.811643	0.027	0.014832	1.618		a1
9	0.041	2.364624	1.324201	0.024	0.013166	1.266		a10
8	0.039	2.446912	1.440721	0.022	0.012487	1.341		a2
7	0.031	2.570582	0.962391	0.020	0.011507	0.968		a7
6	0.029	2.776445	0.955018	0.018	0.011416	0.964		a8
5	0.029	3.182446	1.24877	0.016	0.011285	1.170		a9
4	0.029	4.302653	3.600595	0.012	0.010235	1.612		a3
3	0.011	12.7062	0.866025	0.007	0.004320	0.926		a6
2	0.009			0.005				a4
1	0.001							a5



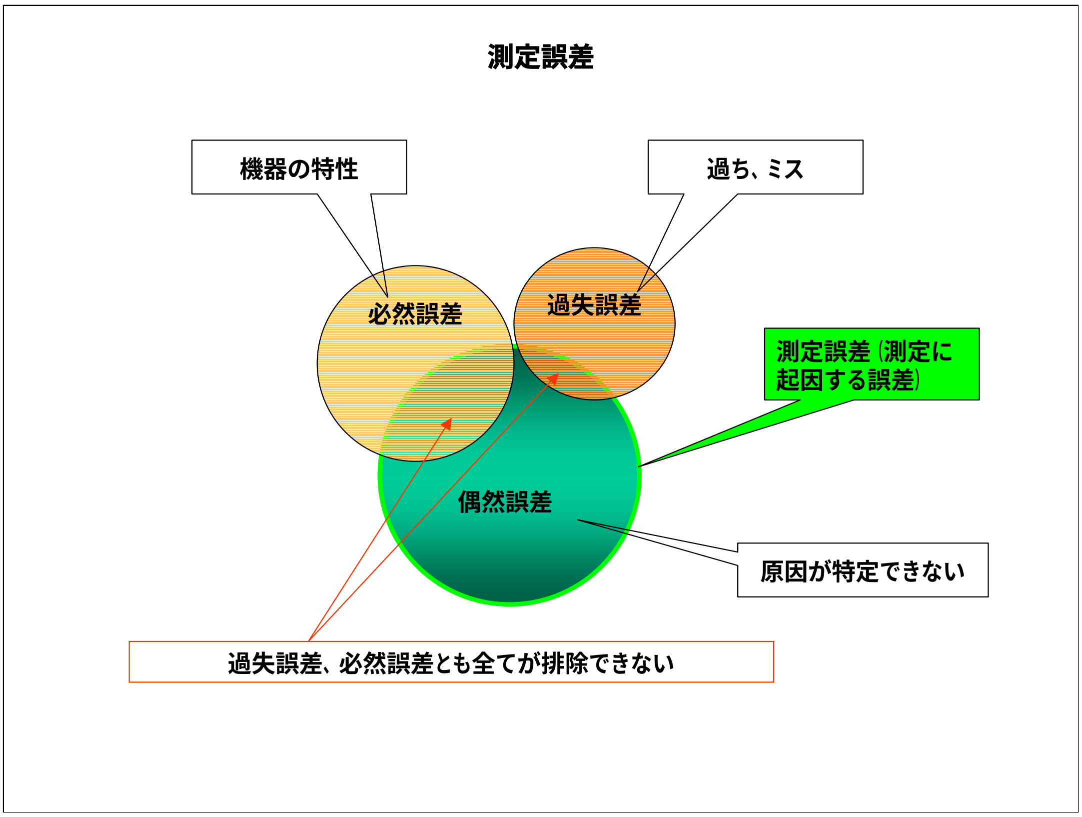
これと似通った方法に水平観測では観測差、倍角差、鉛直観測では高度常数のチェックをしている。

②のデータではa1が異常と誰もが考える。計算でもa1が異常とされます。

③のデータではどうか、異常はないだろう・・・これも統計的に証明できました。

異常な値を除いた平均値が最確値ということになります。

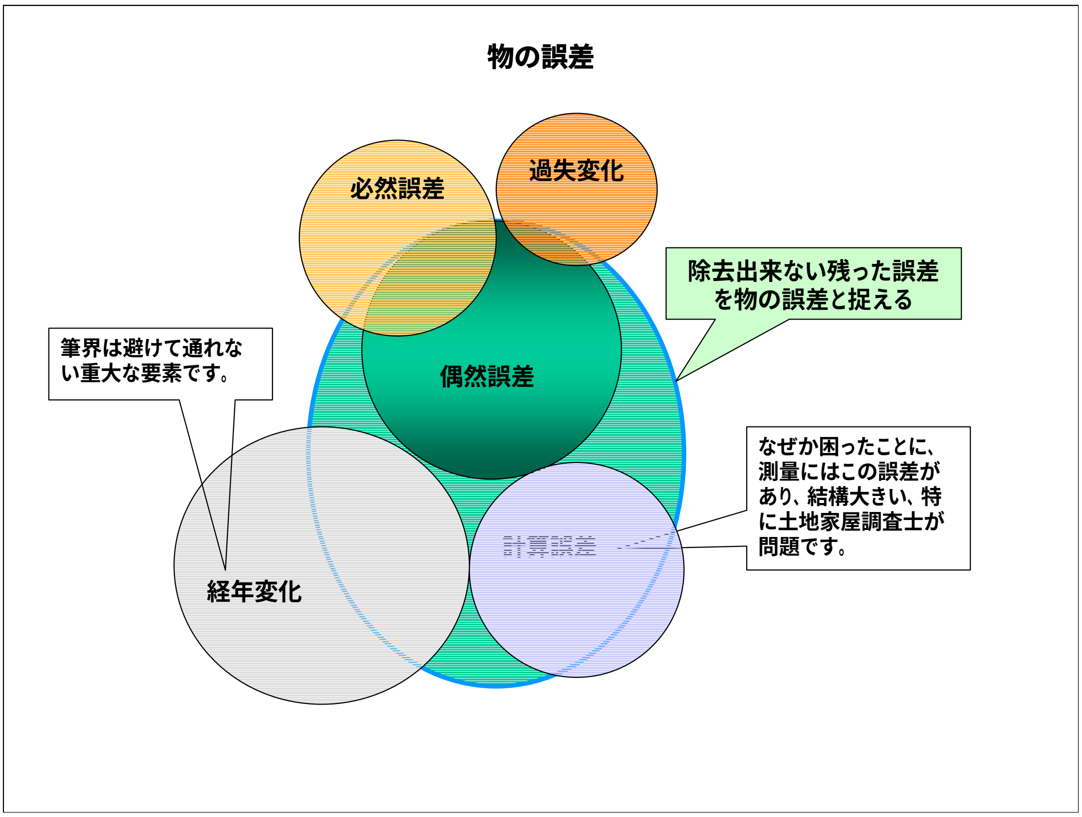
統計では他にも「χ^2検定、t検定」で解説してあるとおりχ^2検定という方法もあるし他にもいろいろあります。



誤差には偶然誤差、必然誤差、過失誤差があります、必然誤差、過失誤差は様々な点検を行えば防げるといわれていますがそれでも小さな必然誤差、過失誤差は残ります。

こういったスライドは何処にもあります、こんなことをしてるとついついつくってしまうのですが基本の考えですから勘弁ください。

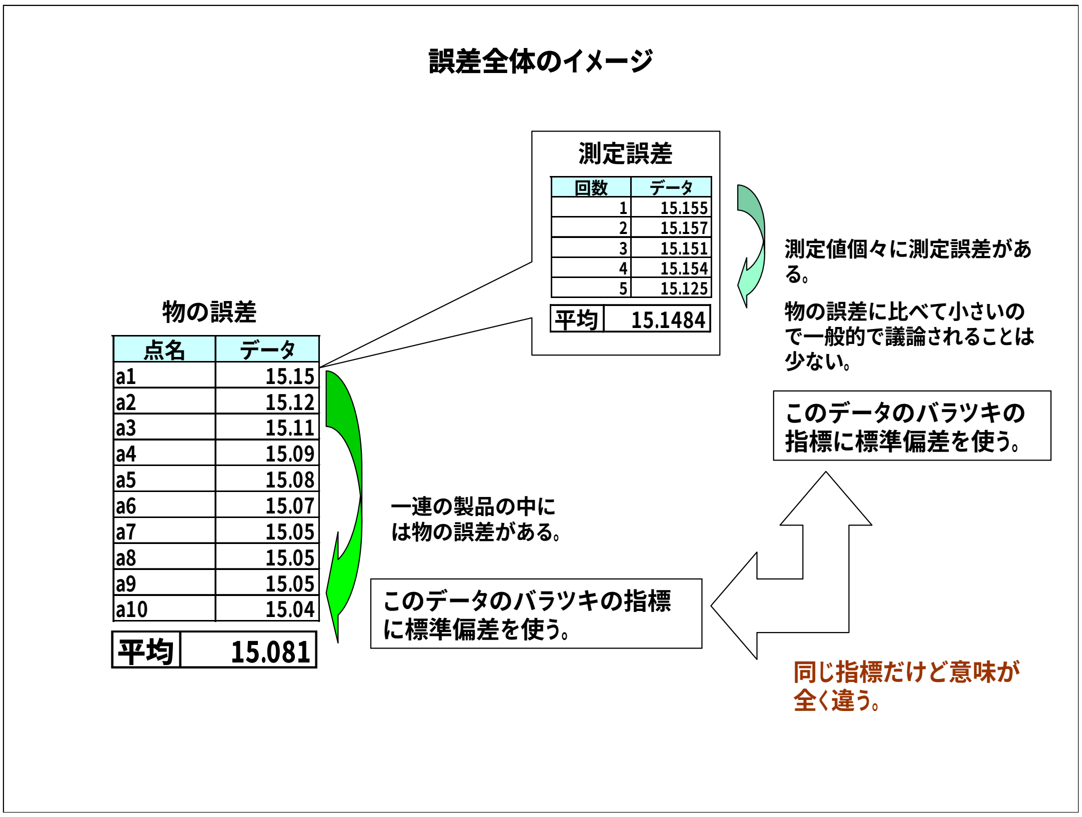
次のスライドにいきますと「ちっとは勉強してるな」と思って頂けるのではないかと思います。



筆界の場合は経年変化という誤差があります、これは一般の測定では考えられない事ですが筆界は筆界が始めて形成された時に遡る、筆界が始めて形成され時に遡って復元するため経年変化が重要になってきます。

それと、測量の場合は計算誤差が大きなウエイトを占めることがあります、とくに作成された年月の古い測量図面を相手にする場合は測量に使われた機械の種類、機械の性能、各種測定値の補正の有無、測量の骨格となる多角点の計算方法が問題になります。

古い図面ではそれを予測しながらデータ解析を要求される為過去に使われた器機、技術の知識も要求されます。



誤差を測定対象で見れば測定対象が一つのものは同じところを繰り返し測ることになりますからこれは測定のバラツキを測っているものであり測定誤差といいます。

同じ境界標を数十回も測ることはないでしょうから測定誤差での説明は現実的ではありません、例外もありますが。

測定対象が一つでない誤差、これはある製品を大量に作っものを測る場合です、これはモノのバラツキを測っている事になります、実際には測定誤差も含まれている訳です、ですからモノのバラツキを言う場合は測定誤差がモノの誤差に対して充分に小さい事が必要です。

測量の場合は同じ測点を多数回測るわけではありませんので測定対象が1つの誤差を計算することはありません。

多数の測点を１回測って個々のバラツキ、精度を全体で示しているので図面対現地にを1つの製品とみているわけですから、各測点の位置誤差も計算されます。

標準偏差（一変数）

変数Xiの標準偏差の式

標準偏差　σx＝√（Σ（（Xi－Xiav）² /（n-1））

エクセルの関数「STDEV（X_1：Xn）」で求められます。

標本に基づいて予測した標準偏差（エクセルから）

Xi　：　変数・測定データ

n　：　データ数

Σ　：　n個の総計

Xiav：　最確値（平均値）

変数Xiの標準偏差の式（通常この式は使わない）

標準偏差　σx＝√（Σ（（Xi－Xiav）² /（n））

エクセルの関数「STDEVP（X_1：Xn）」で求められます。

母集団全体に基づくある母集団の標準偏差（エクセルから）

標準偏差の計算式は上の箱の中の式です、何処にでもありますのでご存じでしょう。
データのバラツキを表す場合、精度をいう場合はこの式からの数値を一般的に言います。
下の式は母集団、データ全ての場合とかXiavが真値の場合ですが真値はないので全てのデータの場合です。

標準偏差

標準偏差の計算

データ

標準偏差	8.7369	幅の中心	-39.3	-30.6	-21.8	-13.1	-4.4	4.4	13.1	21.8	30.6	39.3
		標準偏差幅	-4.5	-3.5	-2.5	-1.5	-0.5	0.5	1.5	2.5	3.5	4.5
		%	0	0	8	12	24	40	16	0	0	0
№	データXi	合計 25	0	0	2	3	6	10	4	0	0	0
№1	-3						1					
№2	3							1				
№3	1							1				
№4	-3						1					
№5	-18				1							
№6	-6						1					
№7	10								1			
№8	8							1				
№9	-18				1							
№10	-12					1						
№11	7							1				
№12	-5						1					
№13	5							1				
№14	-3						1					
№15	-9					1						
№16	5							1				
№17	1							1				
№18	9								1			
№19	5							1				
№20	6							1				
№21	15								1			
№22	-2						1					
№23	12								1			
№24	1							1				
№25	-9					1						
№26												
№27												
№28												
№29												
№30												
		枠内確率	0	0.13	2.14	13.59	34.14	34.14	13.59	2.14	0.13	0

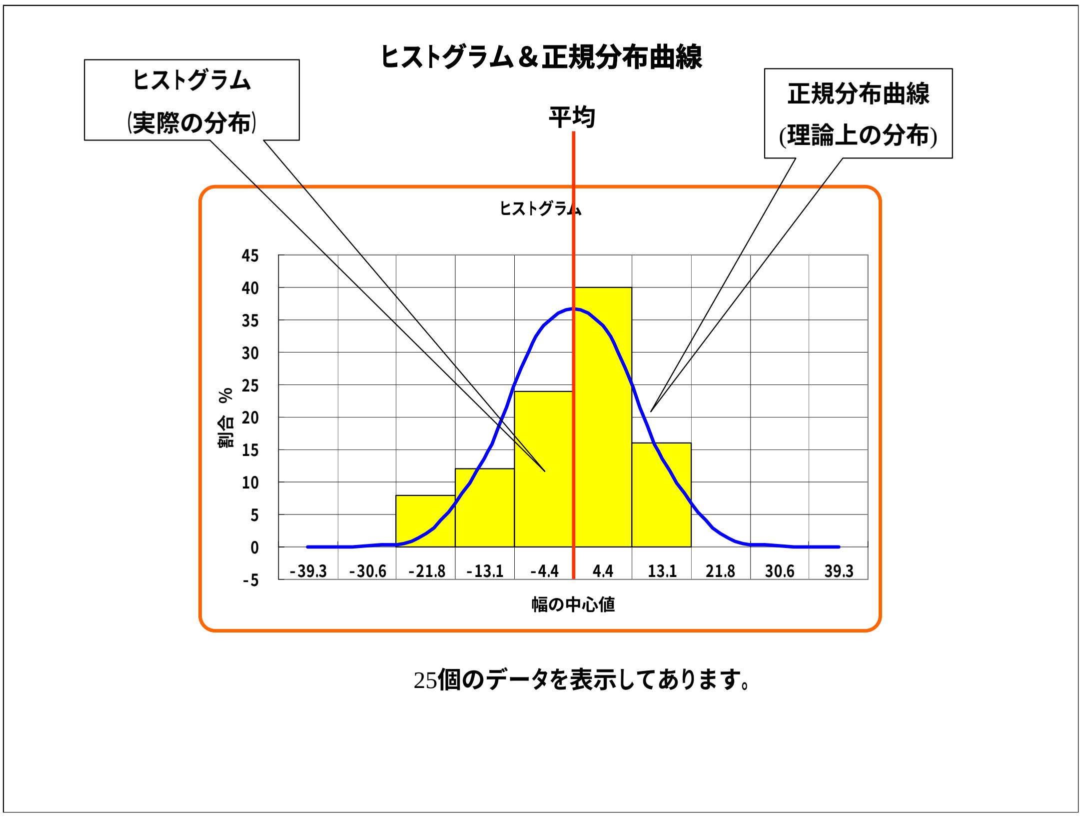
標準偏差=STDEV(B7:B36)　**エクセル関数で計算しています。**

「標準偏差の体験」プログラムを（このファイルのところにアップしてあります）　使って標準偏差を計算して見てください。

水色枠にデータを入力すれば標準偏差が計算されます。

身近なデータを入力して頂くと分布の状態が解ります、それによってデータは正規分布になるのだということを実感して頂ければこのプログラムの役目は終わりです。

次のスライドにグラフになった形を表しました。



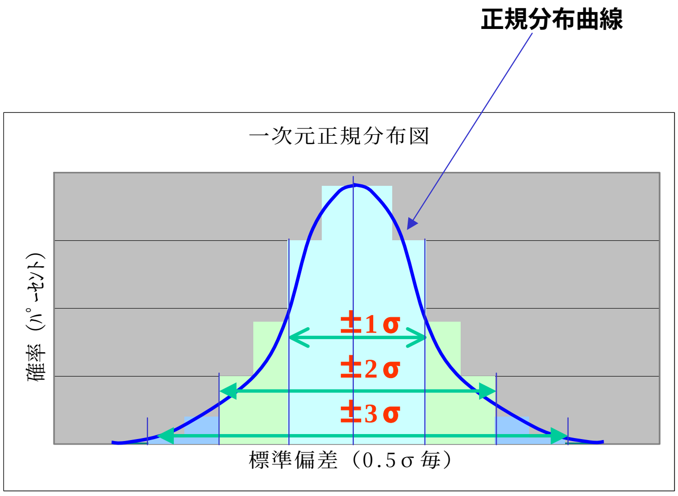
実際の分布をヒストグラム（%で表した棒グラフ）といいます、青の曲線は理論上の分布です。

データ数が多くなるとこの青の線に棒グラフが近づいてきます。

土地家屋調査士が扱うデータは数が少ないのでなかなか実感出来ませんが100個程度のデータを表示させるとよく解ります。

ここでは25個のデータを表示してありますが結構近い形に分布しています。

正規分布曲線と確率



確率表 標準偏差	1σ	2σ	3σ
確率	68.3	95.5	99.7

標準偏差	確率		
0.05	0.0399	2.05	0.9596
0.1	0.0797	2.1	0.9643
0.15	0.1192	2.15	0.9684
0.2	0.1585	2.2	0.9722
0.25	0.1974	2.25	0.9756
0.3	0.2358	2.3	0.9786
0.35	0.2737	2.35	0.9812
0.4	0.3108	2.4	0.9836
0.45	0.3473	2.45	0.9857
0.5	0.3829	2.5	0.9876
0.55	0.4177	2.55	0.9892
0.6	0.4515	2.6	0.9907
0.65	0.4843	2.65	0.9920
0.7	0.5161	2.7	0.9931
0.75	0.5467	2.75	0.9940
0.8	0.5763	2.8	0.9949
0.85	0.6047	2.85	0.9956
0.9	0.6319	2.9	0.9963
0.95	0.6579	2.95	0.9968
1	0.6827	3	0.9973
1.05	0.7063	3.05	0.9977
1.1	0.7287	3.1	0.9981
1.15	0.7499	3.15	0.9984
1.2	0.7699	3.2	0.9986
1.25	0.7887	3.25	0.9988
1.3	0.8064	3.3	0.9990
1.35	0.8230	3.35	0.9992
1.4	0.8385	3.4	0.9993
1.45	0.8529	3.45	0.9994
1.5	0.8664	3.5	0.9995
1.55	0.8789	3.55	0.9996
1.6	0.8904	3.6	0.9997
1.65	0.9011	3.65	0.9997
1.7	0.9109	3.7	0.9998
1.75	0.9199	3.75	0.9998
1.8	0.9281	3.8	0.9999
1.85	0.9357	3.85	0.9999
1.9	0.9426	3.9	0.9999
1.95	0.9488	3.95	0.9999
2	0.9545	4	0.9999
		4.05	0.9999
		4.1	1.0000

理論的な分布になったときの棒グラフのしめる割合を標準偏差の幅で見たのが確率です。

変数が1つの時の正規分布の形と確率は図のようになります。

全体100としたときの各幅の割合を確率といい、±１標準偏差に含まれる確率は68.3パーセント、 ±２標準偏差に含まれる確率は95.5パーセント、 ±3標準偏差に含まれる確率は99.7パーセントです。

この説明は変数が角度だけとか長さだけの場合です。

この変数が1つの一次元正規分布の形、確率と変数が１つの一次元正規分布の形と確率の関係は是非、覚えておいてください。

右に細かい表を載せましたので参考にしてください。

標準偏差は何のために必要か

抜き取りで調べたとき全データの精度を予測する

10,000個の物を作ったとき、全部を検査せずに数%（何%でもいいですが）のデータをランダムに抜き取り検査します。

この時に標準偏差と分布の中心を計算して確率から不良率推定し、全体が合格か不良かを判断する場合に使います。

但し国土調査では標準偏差ではなく平均二乗誤差という指数を使っています。これは与点が同じという条件の場合に使える合否の判定指標として考えられたものと考えられ標準偏差とは違った意味合いがあるようです。

あるデータの精度を予測する

上記の場合で標準偏差が解っていれば、ある点のデータの精度はその標準偏差と同程度の精度の中にあると予想できます。

境界復元では復元された点の精度は不明ですが復元とした点の標準偏差が解っていれば復元点の位置誤差を予測できます。

① 最小二乗法による境界復元では標準偏差が3mmであれば復元された点も3mmの標準偏差の中にある推定される（復元点の復元精度が予測出来る）。

② 交点計算では位置誤差5mmの二つの点を使えば$\sqrt{(5^2+5^2)}=7.1$と予測出来ます（但し交点計算の場合使った点の位置誤差が解らないのが問題）。

標準偏差は何の指標として必要なのかを知って頂いてから使うことが大事です。

ものを生産する場合や国土調査などで使う場合は「抜き取りで調べたとき全データの精度を予測する」為ですが

境界復元などでは「あるデータの精度を予測する」場合に使います。

このことは境界復元の基礎で説明します。

異なった方法から結果を知りたい

隣接する図面に同じ点があり、片方はヘルマート変換で復元し、もう片方はアフィン変換で復元したときに、分散（標準偏差²）の逆数を重量として平均を計算するときそのまま分散を使って計算すると（パラメータ）の多い方程式を使ったアフィン変換の標準偏差が小さく計算されるとアフィン変換が優位になる。

この場合パラメータを考慮した標準偏差を求めてやればその不公平が回避出来る。

ヘルマート変換の標準偏差

$\sigma=\sqrt{(\Sigma\delta^2/(2n-m))}$ ・・・・・・・・・ δ=残差、m=4（14条地図活用マニュアルから）

アフィン変換の標準偏差

$\sigma=\sqrt{(\Sigma\delta^2/(2n-m))}$ ・・・・・・・・・・ δ=残差、m=6（14条地図活用マニュアルから）

mは変換方程式の未知数の数

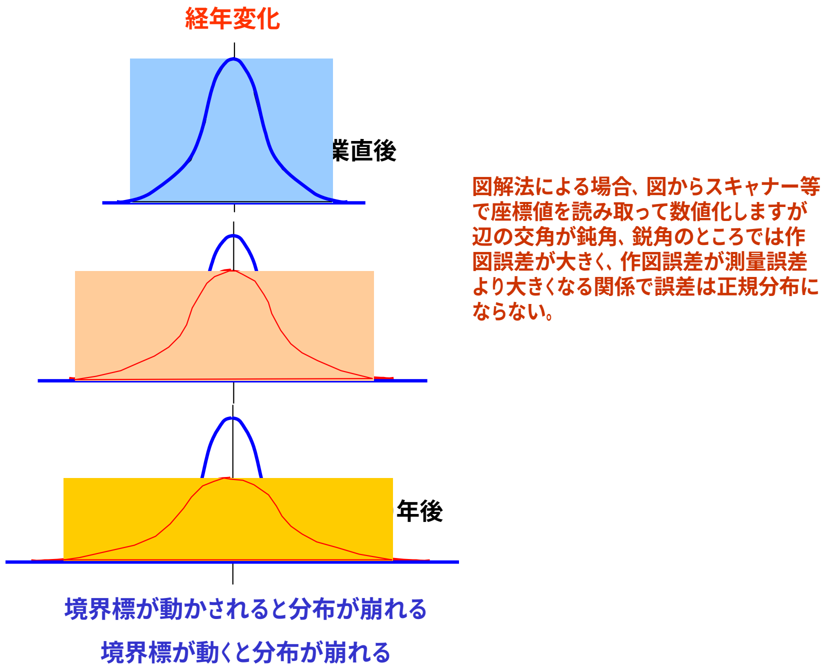
これは計算過程で使われるものでバラツキを表す標準偏差とは異なるので参考に書いてみた、また網平均計算では角度と距離の単位重量の標準偏差を求め計算に使っている例もある。

標準偏差といわれる計算式はいろいろあります、ここに示したのは座標変換の方法によって計算する標準偏差です。

同じデータでヘルマート変換とアフィン変換して実測値と変換値の差を見ればアフィン変換が小さく出ますのでそのまま計算しては常にアフィン優位になります。

データとしての正確性はヘルマート変換が優位になる場合もありますのでその不公平を回避するためにパラメータmをいれて計算し、その時の分散を逆数にしてやれば不公平が解消されると考えたものでしょう。

経年変化という誤差



境界復元では境界標の経年変化を考慮しなければなりません、経年変化とは自然に動くこともありますが、どちらかと言うと人為的に動かされることが多いようです。

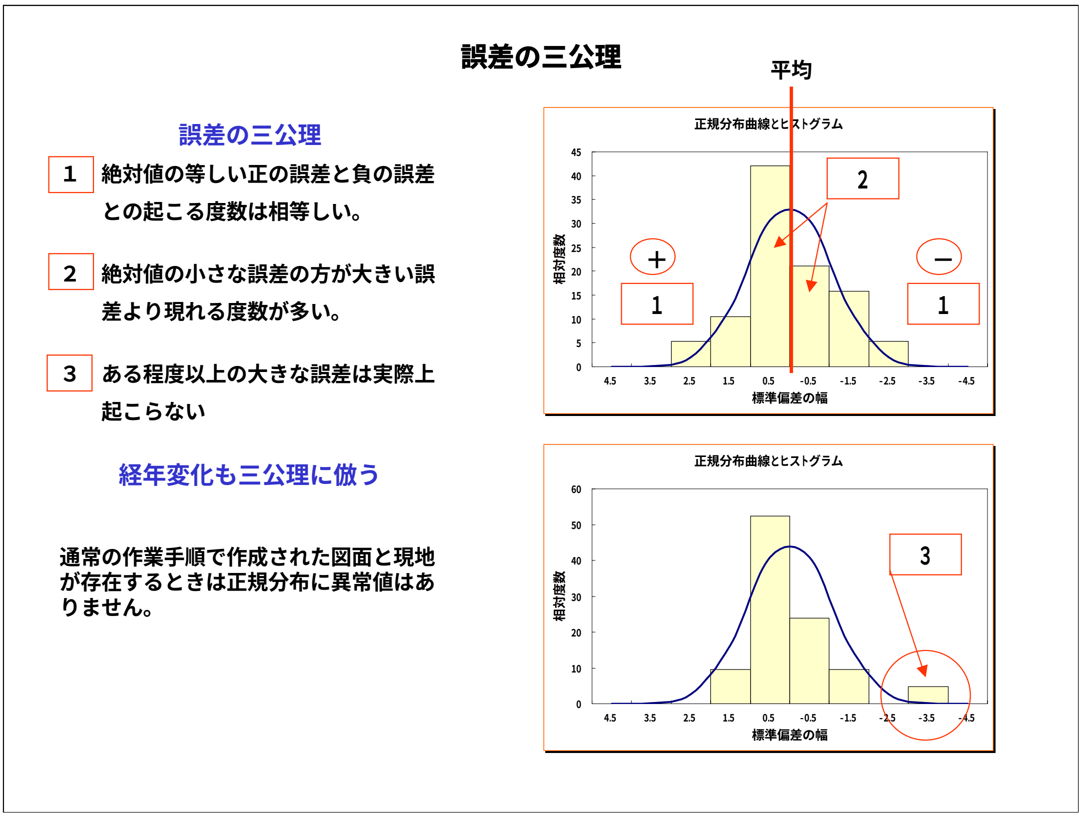
昔であれば筆界の認識が薄く土地を使い勝手のいいように境界標を動かしたこともあったでしょう、特に同じ所有者が何筆も所有していた場合はこの傾向が見られます。

図面によらず、筆界復元を記憶によって行われた場合は勘違いなどで動いてしまったとか、隣地が承諾しないので少し譲って確認書をもらうなどもあったでしょう。

いずれにしても図面が古いほど経年変化が問題になります、図面作成から時間が経っていればいるほど復元は難しくなります、数値が表示されていない公図、初期の地籍図は作図誤差が測量誤差より大きくなる関係で復元がさらに難しくなります。

計算では経年変化、偶然誤差、過失誤差、必然誤差をひっくるめて相対的な関係を計算することになります。

また、図解法による場合、図からスキャナー等で座標値を読み取って数値化しますが辺の交角が鈍角、鋭角のところでは作図誤差が大きく、作図誤差が測量誤差より大きくなる関係で誤差は正規分布にならないなどの問題点もあります。



今までの説明を踏まえて誤差の三公理ですがこの辺も当たり前のことですが図を使って説明してみました。

１．絶対値の等しい正の誤差と負の誤差との起こる度数は相等しい。
２．絶対値の小さな誤差の方が大きい誤差より現れる度数が多い。
３．ある程度以上の大きな誤差は実際上起こらない　　の三つです。

これらの誤差は作業の標準化を測る事により小さくできます、標準化と言いますのは作業書のようなものです、使う測量機器、作業と作業の間に行う点検方法とか点検基準などを決めたものです。

これに熟練が進むことと相まって必然誤差、過失誤差は小さくなり、その後に残った偶然誤差と小さな必然誤差、過失誤差は正規分布に従います。

一時代に正規分布神話というのがあったそうです、測量の原則を守って測量された結果が正規分布になるのであってなんでも正規分布になるわけではないようです。

さらに境界の復元では境界標の設置から復元までに長い時間経過があり、これに伴う経年変化を考慮しなければなりません。

経年変化、計算誤差も誤差の三公理と同じ性質を持つと考えられますので経年変化も含めて図面対現地の相対的誤差を考えます。

計算誤差については各調査士の技術レベルの問題であり、調査士会の指導の善し悪しの問題でもあると考えます。

この後「誤差の基礎Ⅱ」を見てください。