

測定データの異常値除去方法 —基本統計量の歪度を用いたデータスクリーニング手法—

大川 眞一 増田 弘生

株式会社 ルネサス テクノロジ 〒187-8588 東京都小平市上水本町 5-20-1

E-mail: okawa.shinichi@renesas.com, masuda.hiroo@renesas.com

あらまし VLSI設計における素子ばらつきの考慮は、現在、そして未来における重大関心事である。しかも、素子測定、ばらつきのモデリング、設計への反映の各段階において、新たなる技術開発のステップが多数存在する。そしてその多くは独自の基本概念を必要とする革新的技術である。

そのような多数の課題の一つ、測定データの異常値除去の方法を新たに開発した。基本統計量の歪度を応用している。正常データに対する副作用は少なく、異常なデータを効率的に除去することができる。素子ばらつきの評価に必要な、大規模測定の信頼性を、著しく向上させることが可能になる。

キーワード LSI, 設計, ばらつき, 測定, 異常値, スクリーニング

A New Technique for Elimination of Irregular Data in Measured Values — A Data Screening Technique Applying Skewness of Basic Statistic —

Shin-ichi OHKAWA and Hiroo MASDA

Renesas Technology Corp., 20-1, Josuihon-cho, 5chome, Kodaira-shi, Tokyo 187-8588, Japan

E-mail: okawa.shinichi@renesas.com, masuda.hiroo@renesas.com

Abstract Consideration of device variation in VLSI design, at present or in future, is very important concerns. Moreover, in each stage of device measurement, variation modeling or reflection on design, there are many steps of developments of technology. And then, many technologies are novel one needing unique basic concept.

As one of such many problems, we have developed a new technique which eliminates irregular data in measured values. Skewness of basic statistic is applied to this. Side effects are small for regular data, while it eliminates irregular data efficiently. A Reliability of large scale measurements, which is needed for evaluation of device variation, is remarkably improved.

Keyword LSI, design, variation, measurement, irregular value, screening

1. 序論

数年前、LSIの素子ばらつきは、LSIの設計時における大いなる関心事であったが、45nm世代が実用化されつつある現在、極めて切実な問題となった。素子ばらつきの設計への取り込みは緊急課題である。

一言で素子ばらつきの取り込みと言ってしまうが、実際は多くの技術ステップが存在する。荒く考えても以下のステップが考えられる。

- 1) 測定パターンの策定
- 2) 評価チップの設計(アクティブマトリックスの評価チップ)
- 3) 高速・高精度の計測システム
- 4) 測定データのスクリーニング(不良データの除去)
- 5) 評価データのモデリング(素子ばらつきのモデル化)
- 6) 回路特性への反映方法
- 7) 遅延時間のばらつきモデルの作成
- 8) SSTAへの取り込み

基本的なステップだけでも8項目あり、さらにそれぞれの項目が多様な課題を抱える。素子ばらつきのLSI設計への反映は、多大な開発課題の解決を必要としている。しかもそれぞれの課題は、技術的、学術的に新たな領域の開拓であり、多大な尽力を必要とするものばかりである。

これらの課題の中から、本研究は、上記4項の“測定データのスクリーニング”

を取り扱うことにする。またその中でも、“異常値除去方法”という課題について詳述する。ここで言う“異常値除去方法”とは、一組の測定データにおける、主たる分布(ほぼ正規分布となる)から外れたデータ点を、適正に除去するためのデータ処理アルゴリズムである。

2. 異常値除去の必要性

ばらつきの評価は難しい。たとえば、あるプロセス条件のトランジスタと、それと異なる条件のトランジスタで、どちらの方のばらつきが少ないかを比較するとしよう。評価誤差1%(3 σ)の場合で各々9万点ものデータが、評価誤差を3%に下げた場合でも各々1万点のデータが必要である。これは1チップで256点の計測が可能なアクティブマトリックスの評価チップ(DMA[1])を80チップ測定した場合に相当する。このように、ある程度の水準の評価誤差でばらつきを正確に評価しようとすると、DMAのような、1つの素子に対して1チップで256点のデータを持つ評価チップを、ウェハ全面で測定をすることが必要になってくる。

DMAは1つのチップで350種の素子を搭載している。この全ての搭載素子を1ウェハ(ここでは80チップとする)について全面測定すると、 $256 \times 350 \times 80 = 716,8000$ 、すなわち700万点を超えるデータの収集が必要になる。これは1ウェハのデータであり、素子ばらつきの定量的評価を行おうとすれば数十枚(データ点数として数億点)に及ぶ評価が必要になる。大量のデータを適正に処理することが要求される。

しかし、このようにして計測したデータが、全て正常とは限らない。特にプロセスの開発初期には一定の不良が含まれることを避けることができない。また現れる不良も種々のケースがある。図1はある測定データの正規確率プロットの例である。十字マークが測定データで、実線はそれに対する近似直線である。(正規確率プロット上の直線は正規分布の累積分布関数を表す) 図1(a)は最も単純な例で、全測定データの内の1点が大きく外れているものである。図1(b)は分布が裾を引いている例、図1(c)は主な測定値(主分布)とは外れた値を含んでいる例である。

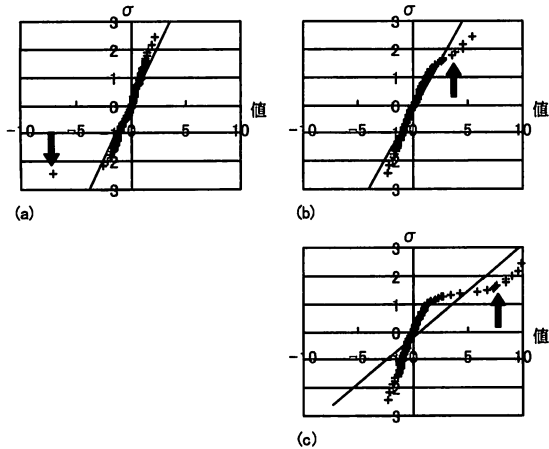


図1 異常値を含む測定データの例

このように、主たる分布(正規分布で近似できる成分)に対し、そこから外れた異常値が測定データに含まれる場合がある。この異常値をどのように処理するか。それが問題である。

この問題に対し、2つの解がある。第一には、異常値を含む測定データを無効データとして捨てる方法である。何らかの方法で異常値の有無を判定し、異常値を含む測定データを無効にする。この方法は、等価なデータが豊富にある場合、たとえばウェハ上に百個程度のチップがあるようなケースでは許容される。しかし測定がウェハ単位であるような場合、この方法は適さない。測定データが異常値を含むからと言って、その水準のウェハデータを破棄することは現実的ではない。またチップ単位の測定でも、歩留まりが厳しい場合には、この方法は不適切になる。

第二の方法は、何らかの方法により異常値成分を除去し、除去後のデータを有効な測定データとして扱うという方法である。この方法には2つのメリットがある。第一には、有効な測定データ数を増やせるということである。第二点には、異常値成分を除去できるということは、測定データを正常成分と異常成分に二分できるということである。このことは、正常成分だけでなく、異常成分も分析可能になるということである。

測定データは必ずしも正常値のみに価値があるのではない。異常値にも分析の価値がある。不良の原因分析や、歩留まりの分析に応用できる。その点においても、適正な異常値除去法の確立は有益である。

本研究における異常値処理の例を図2に示す。このデータは図1に示す測定データ(異常値を含む)の処理結果であり、適正に処理が施されている。この処理のアルゴリズムには基本統計量の歪度(Skewness)を利用している。これより本アルゴリズムの詳細を説明して行く。

3. 基本アルゴリズム

基本となるアルゴリズムは歪度を判定指標としたスクリーニングである。歪度について簡単に説明する。歪度 Sk は式1(s は標準偏差)によって表される。歪度は標本の分布が持つ左右の偏りを表現し(図3)、左に裾を引く(右に偏る)分布の場合は負の値、右に裾を引く(左に偏る)分布

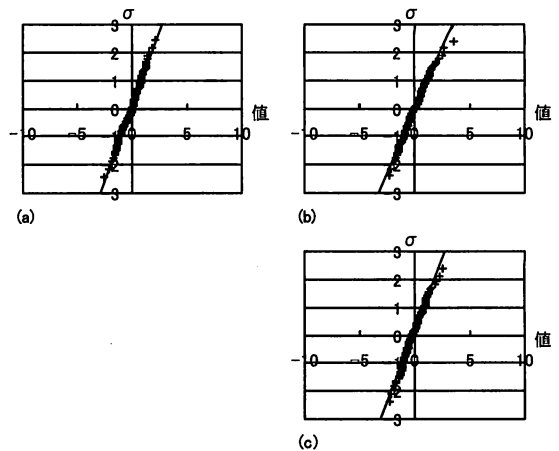


図2 異常値除去後の測定データ

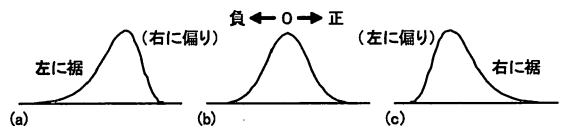


図3 歪度を示すサンプル

の場合は正の値を示す。左右対称な場合は0である。(ただし $Sk = 0$ が左右対称であるための十分条件ではない)

$$Sk = \frac{1}{n} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s} \right)^3 \quad (1)$$

測定データが異常値を含む場合、その影響が歪度に現れる。図1において、(a), (b), (c) それぞれの歪度は $-2.04, 1.22, 2.37$ である。(図3は (a), (b), (c) それぞれ $-0.95, 0, 0.95$) よって歪度を指標として測定データが異常値を含むか含まないかを判定することが可能である。

これを原理として、異常値を除去するアルゴリズムを考えることができる。このアルゴリズムを端的に言葉で表現すれば、「歪度の大きさ(絶対値)がある水準判定値を越えていた場合、歪度が示す裾側の、中心(平均値)から最も離れているデータ点を除去する」、ということになる。この操作を、歪度の大きさがある一定の水準判定値以下になるまで繰り返すことにより、異常値の除去された測定データを得ることができる。

このアルゴリズムのフローチャートを図4に示す。操作そのものは単純である。ここでポイントなのが水準判定値の選び方である。図5はそ

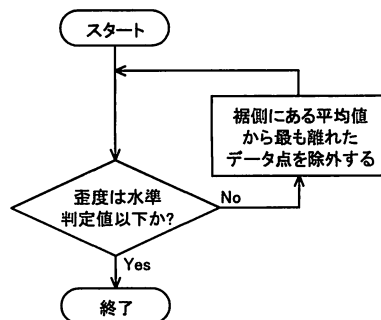


図4 異常値除去のフローチャート

の水準判定値を示すグラフである。詳細は後に説明するが、このグラフは有意水準5%での水準判定値を示しており、サンプルサイズ(測定デ

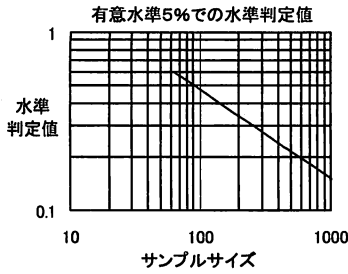


図5 水準判定値の例(有意水準5%)

ータのデータ点数)の関数になっている。(水準判定値は概ねサンプルサイズの平方根に反比例する) サンプルサイズが大きの場合に水準判定値は小さくなっており、歪度は厳しく(左右の歪みが小さく)判定されることになる。これについての詳細もこの後に述べる。

4. 水準判定値の定義

歪度を指標として、どこまでを正常値、どこまでを異常値と判定すれば良いのか。それについて説明する。

測定データが純粋に正規分布だったとしよう。サンプルサイズNの標本が多数あったとする。このとき、各標本の歪度は0にはならず、0を中心とした分布を持つ(図6)。標本毎に左右の揺らぎを持つからである(図7)。この時の揺らぎ分布はサンプルサイズに強く依存し、ほぼサンプルサイズの平方根に反比例する(図5)。サンプルサイズが小さいと左右の揺らぎを発生しやすく(図6(a),図7(a))、逆に大きい場合は純粋な正規分布に近づく(歪度が小さくなる)からである(図6(b),図7(b))。

補足事項であるが、図6の分布形状に触れる。図6(a),(b) いずれの分布も正規分布に似ているが正規分布ではない。しかし非常に正規分布に近い形状で、“N→大”において正規分布に漸近する。

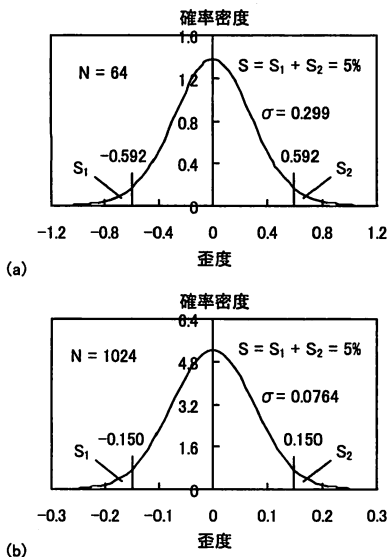


図6 歪度の分布のサンプルサイズ依存

このように、純粋に正規分布に従う標本においても、サンプルサイズに依存した歪度の分布を持つ。標本の中には、偶然の揺らぎによって大きい歪度を示すものがある。図6(a)において、 $|\text{歪度}| > 0.592$ となる面積が5%ある。図4のフローチャートにおいて、この0.592を水準判定値とするならば、純粋に正規分布に従う標本(N=64)でも、5%の標本が擬似的に異常値を含むとして処理される。この時の処理のグレードを

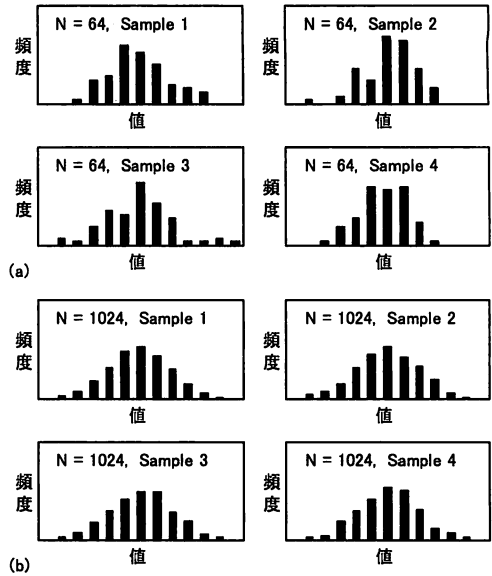


図7 サンプルサイズ = 64, 1024 での分布サンプル

“有意水準5%”と呼ぶことにする。よって、たとえば有意水準20%と云えば、標本全体の20%が異常値処理の対象にされ、厳しいスクリーニングをすることになる。逆に有意水準1%ではほとんどの標本が処理されず、ゆるいスクリーニングをしていることになる。

有意水準とは、純粋な正規分布の標本の何%が処理対象にされるかという意味で、水準判定値はその時の歪度の大きさの境界値である。

5. 水準判定値の近似式の導出

本研究による異常値除去方法(図4)で要となるのは水準判定値である。水準判定値は“有意水準”と“サンプルサイズ”の関数になる。図4のフローチャートにおいて、“有意水準”はスクリーニングの強度を表す定数である。一方の“サンプルサイズ”はスタート時のサイズが初期値として与えられ、除去作業が進むにつれて減少していく変数である。

本異常値除去のフローチャートの実行のためには、“有意水準”と“サンプルサイズ”を変数とした水準判定値の関数が必要である。最も良いのは数学的に導出された関数であるが、導出は困難であると考えられた。ここではモンテカルロシミュレーションによる方法を選択した。

“サンプルサイズ”(以下Nと記す)をパラメータとした大量のモンテカルロシミュレーションを実行し、“有意水準”(以下Lsと記す)と水準判定値(以下Lskewと記す)との関係を求め、そこから近似式を導出する。モンテカルロシミュレーションで精度の良い近似式を得ようとすると、試行回数は膨大なものになる。今回のシミュレーションではNに32, 45, 64, 90, 128, 180, 256, 360, 512, 720, 1024の11種を選び、各々に対して800万回のモンテカルロシミュレーションを行った。一回の試行内容は「正規分布に従うサイズNの標本を発生させ、その時の歪度を求める」である。これを繰り返すことにより800万個の歪度の分布が求まる。

求めた水準判定値を図8に示す。Lsをパラメータとし、Nの平方根にほぼ反比例する特性となっていることがわかる。この特性に対し近似式を求めた。(式2、表1) 近似誤差は0.5%以下である。(N=32~1024)

$$Lskew = \exp\{a \ln(N)^2 + b \ln(N) + c\} \quad (2)$$

Ls	a	b	c
0.02	-0.0022	-0.4772	1.6894
0.05	-0.0042	-0.4491	1.4137
0.1	-0.0056	-0.4283	1.1622
0.2	-0.0068	-0.4113	0.8480

表1

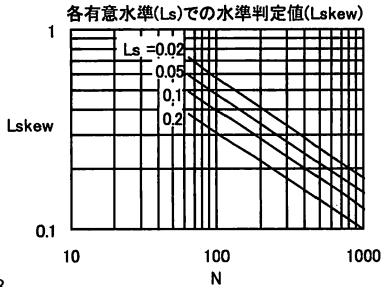


図8

以上にて、水準判定値の近似式が得られた。この関数を図4のフローチャートに適用することにより、本研究の異常値除去方法は完成する。これよりこの異常値除去方法の基本特性と、除去特性について調べる。

6. 異常値除去アルゴリズムの基本特性

図4のアルゴリズムの基本特性を調査する。ここで言う基本特性とは、標本が純粋な正規分布に従う(異常値成分を持たない)場合のアルゴリズムの振る舞いを言う。3項および4項で説明したように、このアルゴリズムは、標本が純粋な正規分布に従う場合でも、ある一定の確率でデータ点の除去が発生する。その確率を指定するのが、Lskew(水準判定値)の関数に与えられる変数の Ls(有意水準)である。たとえば Ls = 0.05 とすれば、試験した標本の95%には何も操作が発生しないが、5%の標本には除去操作が発生する。ここで問題とするのは、この操作によって観測したいデータ(たとえば標本の σ や平均値)にどの程度の影響が生じるかである。基本的に、この操作によって、平均値はばらつきが広がり、 σ は値が小さい方へシフトする。この影響の詳細を調べたので以下に記す。(標本のサイズは256とした)

図9、図10、表2、表3にその結果を示す。標本(サイズ256、 $\sigma = 1$ の正規分布)を100万個発生させ、その時の影響を見ている。図9は除去数の分布である。図10は左列が σ の分布、右列は平均値の分布を示している。表2は図9に対応するデータである。

図9および表2において、除外数0の度数は、(a)、(b)、(c)それぞれ、95万個、90万個、80万個になっている。これは Ls の5%、10%、20% に対応している。除去数の度数は指数関数的に減少して行き、除去数10 ~ 25辺りで ppm オーダーの発生率になる。

除去操作を行うと、ある程度の副作用が発生する。 σ 値は小さくなり(わずかに負側にシフト)、平均値の分布(ばらつき)は大きくなる。その様子を図10に示す。 σ 値の変動も平均値の分布の変動も小さく、実用上の障害にはならない(表3)。あえて難点を言うならば、 σ 値の分布が負側で裾を引く傾向があることである。ppm オーダーの分布を問題にする場合は Ls = 20% の水準は、やや裾引きが大きい。(ただしこの裾引きは N が大になると軽減される傾向がある)

以上と同様な観察を N = 64, 128, 256, 512, 1024 について行った。結果を図11(符号注意)、図12に示す。これは表3と同様のデータをグラフ化したものである。N が大きいほど値の変動が少ないことが分かる。

7. 異常値除去アルゴリズムの除去特性

図4のアルゴリズムの除去特性を調査する。異常値成分を人工的に含ませて、その除去効果を検証する。図13は異常値成分の混合に仕方の説明である。図13(a)が混合方法を示している。主分布はおおよそ面積1(正確には1から異常値の面積を減じたもの)の正規分布($\sigma = 1$)である。それに対し、異常値は、主分布から距離 X_s 分シフトした位置に分布する $\sigma = 1$ の正規分布である。図13(b)~(g)は $S_2 = 0.1$ で X_s を1 ~ 6 まで振ったときのグラフである。図13(b)~(g)を調査範囲の基本とし、表4の○印のついた水準を調査した。

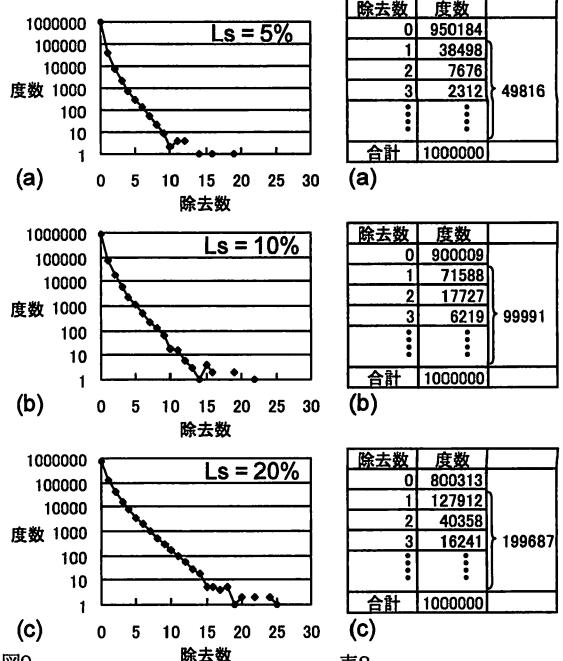


図9

表2

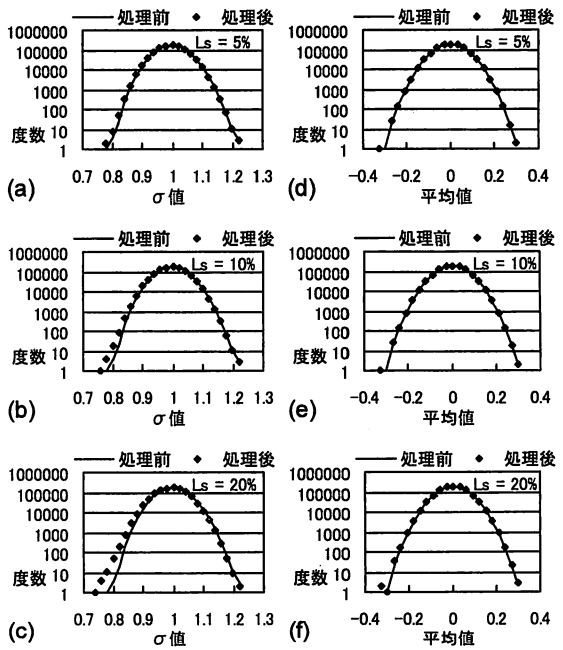


図10

Ls	σ 値の変動*	平均値の σ 値の変動*
5%	-0.13%	0.24%
10%	-0.27%	0.55%
20%	-0.55%	1.38%

表3

* 処理前を1とする

		異常値混合率(S_2)		
		2%	5%	10%
Ls	5%	○	○	○
	10%	×	×	○
	20%	×	×	○

表4

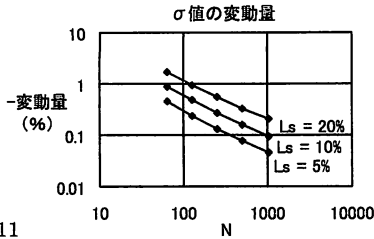


図11

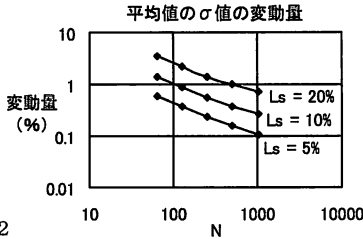


図12

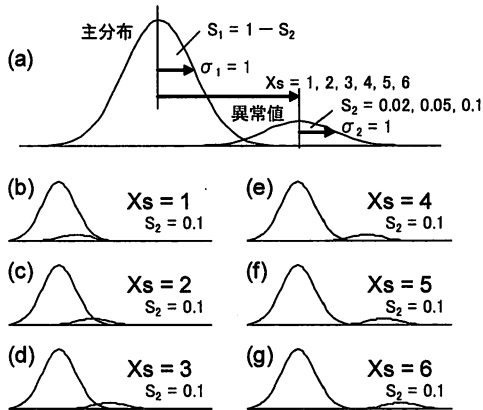


図13

以上のように調査方法を設定し、モンテカルロシミュレーションにより異常値除去の特性を調査した(試行回数10万回)。標本数は 64, 128, 256, 512, 1024 を使用した。全調査量はグラフ数にして550個にも及ぶので、ここでは代表して「異常値混合率5%、 $L_s = 0.05$ 、 $N = 256$ 」の場合のグラフ(図14)を用いて説明する。以下に観測のポイントを列記する。

- 図(a)~(f): X_s が大きくなるにつれて除去数の分布が大きい方へシフトする。
- 図(g)~(i): X_s が大きくなるにつれて処理前と処理後の σ の差が大きくなる。処理後の分布は大きく変化していない。
- 図(m)~(r): X_s が大きくなるにつれて処理前と処理後の平均値の差が大きくなる。処理後の分布は大きく変化していない。
- 図(s): 図(a)~(f)のまとめ。 $X_s = 3$ あたりで除去効果が立ち上がる。
- 図(t): 図(g)~(i)のまとめ。 $X_s = 3$ あたりで σ の平均の変化が最大(+8.41%)になる。除去操作が効果的に作用している。
- 図(u): 図(m)~(r)のまとめ1。全平均値を示している。全平均値のシフト量は元の $\sigma = 1$ の分布幅を基準として+8.92%である。
- 図(v): 図(m)~(r)のまとめ2。平均値分布の幅を示している。分布はやや広がっている ($X_s = 3$ で+17.0%の増加)

以上の評価結果をさらに集約して要点となる値を図15~17にまとめる。図15~17は、図14(t)~(v)における評価値①②③をグラフ化したものである。(表4の条件で $N = 64 \sim 1024$ について集計) この特性を見ることによって、この異常値除去方法の効果を見渡すことができる。

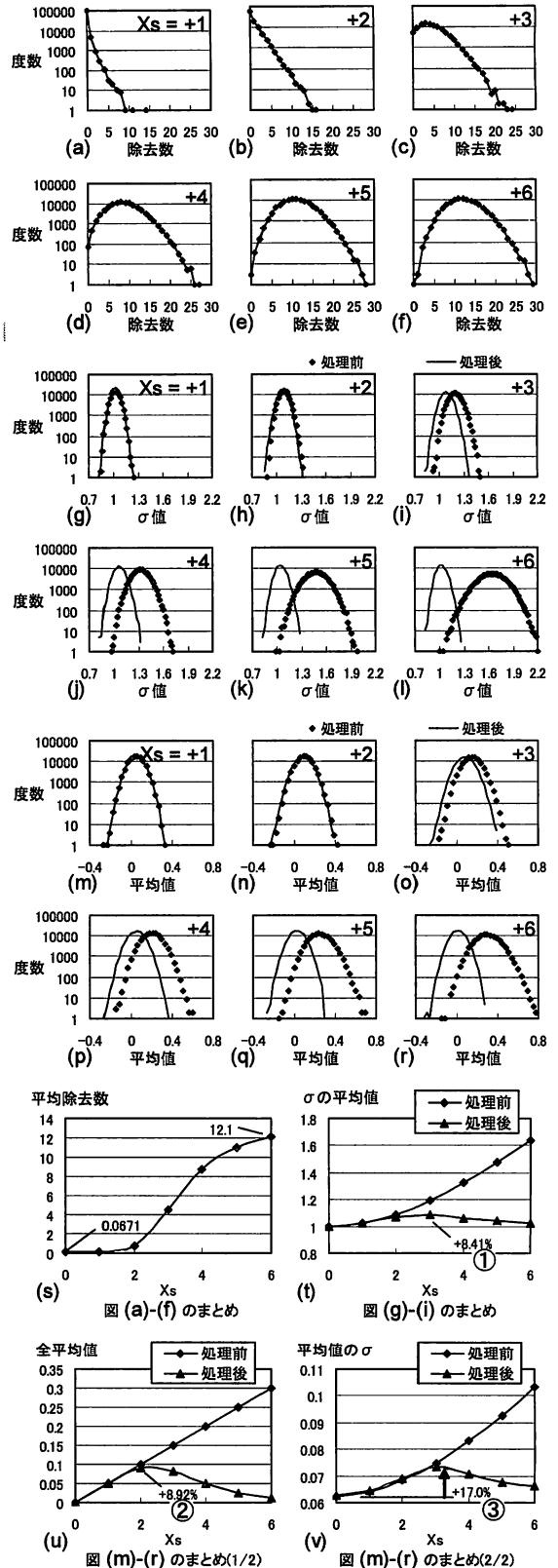


図14

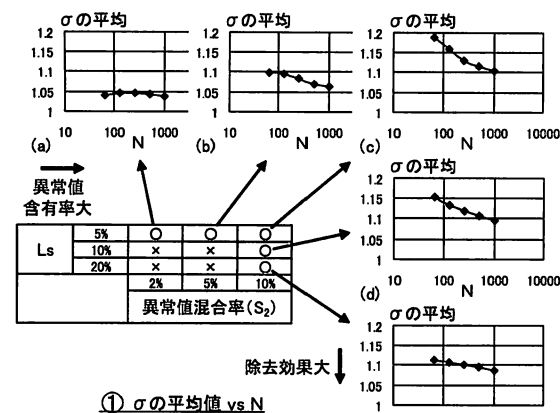


図15

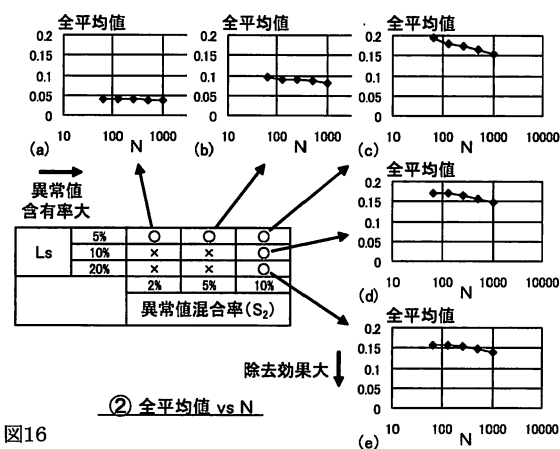


図16

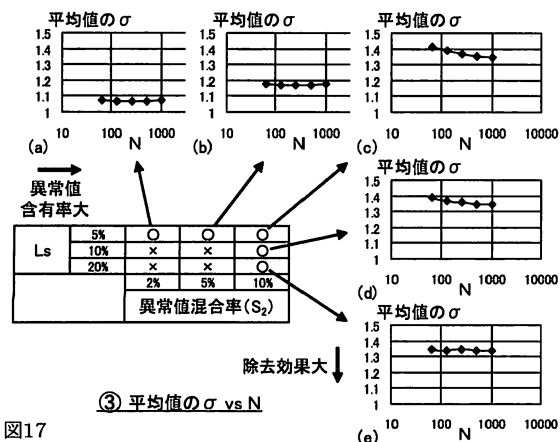


図17

図15(a)において、有意水準5%、異常値混合率2%の場合は、 σ の平均の変化は5%以下である。この数値が大きいか小さいかについては次のようなケーススタディを考えるとすることができる。

- ・ 図15(a)は σ の平均の変動が最大になるXsにおける値である。実際にはXsは大きくばらつくと予想される。よって実効的な値としては数%以下になると考えられる。
- ・ 常にどのサンプルにも異常値が含まれるとは限らない。たとえば異常値を含むサンプルの発生率が数十%とするならば数% \times 数十% $<$ 1%と試算できる。

よって、異常値混合率2%程度のサンプルがサンプル全体の数十%に含まれている場合でも、その全サンプルから算出される σ の平均値の付加的な誤差は1%未満であると言える。

次に図15(b),(c)を見てみる。異常値混合率5%、10%の場合である。Nが小さい領域でのフィルタ効果が弱まっているのが分かる。あるいは逆にNが大きければほどフィルタ効果が効率的になるとも言える。

図15(c),(d),(e)は有意水準(Ls)を5% \rightarrow 10% \rightarrow 20%と高めていった場合のフィルタ効果の変化を示している。Nが小さいところでのフィルタ効果の改善が著しい。

さて、ここで困るのはLsをいくつに選ぶのが適正かという問題である。Ls = 20% におけば効果は大である。しかしその場合、過剰に処理してしまう可能性がある。標本が純粋に正規分布に従うならば、問題ないことを7項で示した。しかし標本が必ずしも正規分布に従うとは限らない。いくらかの偏りを持っている場合も考えられる。このような場合にLs = 20% という水準を適用した場合(特にNが大の場合で)、元の分布の偏りを異常値と判定してしまい、過剰な処理を行ってしまう危険がある。

この問題を定量的に扱うためには、多大な検討と議論を必要とするので、別の機会に詳細を検討したい。現段階では以下のようにガイドラインを定めておく。

- ・ サイズが100程度の一つの標本に、異常値の発生が数個程度以下であるならば、Lsは5%を使用する。
- ・ サイズが100程度の一つの標本に、異常値の発生が数十個程度になる場合はLs = 10%~20%の使用を検討する。
- ・ N = 1024 のように標本のサイズが大きい場合はLs = 2%~5%を使用する。

図16について説明する。この図は試行した10万個の標本の全平均値である。理想値は0であるが、ここでは異常値成分により値がシフトして見える。図15と比較して、LsやNに対する依存性が小さい。これらの値のシフトが測定値評価に対して与える影響はあまり大きくない。たとえばこのグラフでの値0.1が測定値に与える誤差を概算してみる。測定値が100%として、そこに $\sigma = 5%$ のばらつきが乗っていたとしよう。値0.1は $\sigma = 1$ に対する比率なので、測定値での影響は5% \times 0.1 = 0.5%となる。よってそれほど大きい影響ではないと考えられる。

図17について説明する。この図は試行した10万個の標本の平均値の σ の増大を見ている。図16と同様に、LsやNに対する依存性は小さい。グラフの値が1.2であった場合について影響を試算してみる。サイズ100、 $\sigma = 5%$ の標本があったとする。するとこの標本の平均値のばらつきは5% $\div\sqrt{100} = 0.5%$ である。これが1.2倍されるので、0.5% \times 1.2 = 0.6%となる。よって影響は非常に軽微である。

8. まとめ

LSIの素子ばらつきの扱いは、LSI設計における最重要事項になりつつある。はらつきの適正な扱いのためには多くの技術項目の開発が必要である。本研究においては素子ばらつきの測定に関する一項目である「測定データの異常値除去方法」について検討した。その結果として、基礎統計量の歪度を応用した異常値除去方法を開発した。この手法は、正常なデータ(異常値を含まない正規分布のデータ)に対しては副作用が少なく、異常値を含むデータに対しては効果的なフィルタとして作用した。この技術により測定データの信頼性を極めて高めることが可能となった。

文献

- [1] S. Ohkawa, M. Aoki and H. Masuda, "Analysis and Characterization of Device Variation in an LSI Chip Using an Integrated Device Matrix Array," IEEE Trans. on Semiconductor Manufacturing, vol. 17, pp. 155-165, May 2004.