少サンプル製品不良の解析に向けた KL divergence の推定方法*

勝 村 義 輝** 今 沢 慶** 杉 西 優 一***

Effectiveness Evaluation for a Estimator of Kullback-Leibler Divergence for Small Sample Product Failure

Yoshiteru KATSUMURA, Kei IMAZAWA, and Yuuichi SUGINISHI

This paper describes about a root cause analysis method in product quality control, especially, when failure data quantity is extremely small. In the root cause analysis, t-test, F-test and Kullback-Leibler (KL) divergence are possible indicators to quantify a difference between pass and failure data distribution. But those are not stable on small sample data such as less than 10 samples. To solve the problem, we have proposed a failure occurrence probability as an estimator of KL divergence. The failure occurrence probability is an intermediate state of KL divergence because the limit of occurrence probability becomes KL divergence.

In the evaluation of rank correlation with actual data, the proposed method shows a highest correlation on 5 sample and 50 sample test cases comparing with other methods. This results comes from the failure occurrence had a smaller variance than KL divergence. In conclusion, the occurrence probability has a possibility to estimate a distribution difference, stably on the small sample data.

Key words: kullback-leibler divergence, estimator, small sample data, variance, asymptotic equipartition property

1. は じ め に

製造業では、高信頼性の製品を低価格で提供することが重要な課題である。その為、製造における品質管理では、製造中や完成後に製品の検査を行い、一定の基準を満たすものを出荷している。この検査基準が甘いと不良品が出荷され信頼性が下がり、厳しいと良品のオーバーキル率が増えるため製造コストが上がる。したがって、高品質の製品を低価格で提供するためには、検査基準を適正化すると共に、不良品の発生要因を特定し早期に対策することが重要な課題となる。

この不良品の発生要因を特定する方法としては、2つの手法が一般的に使われている。1つ目は、対象製品を分解調査するなどの物理解析アプローチ、2つ目は、データマイニング等の統計技術を用いたデータ解析アプローチである。物理解析アプローチは、原因の特定を正確に行うことができるが、原因の特定までに時間を要することや分析に専門知識や設備が必要なことから、数量をこなすことが難しく、抜き取り検査や事故対策として行われることが多い。一方で、データ解析アプローチは品質と疑似相関を持つデータを探索してしまうなど、正確さに課題があるものの、原因の特定までに要する時間は短く、システム化も容易であることから、品質を日々管理する方法として広く普及している。本研究では、このデータ解析アプローチを対象とする。

このデータ解析アプローチでは、製品を製造した際の製造装置の温度や圧力などのモニタデータや製造中の検査データといった製造データに対して、不良原因との関連性に応じた順位づけを行う。順位が上位の製造データから、物理モデルによる仮説が立てられるかどうかを検証し、仮説が立てられた場合は、実験などによる仮説の検証を行う。そのため、データ解析アプロー

チでは、製造データを不良原因との関連性に応じて順位づけるための指標が必要である。

また、品質管理では不良点数の少ない初期段階で早期に対策することが効果的であるため、順位づけるための指標には、数点の少ない不良品データでも、大量の不良品データを用いて評価した順位と同じ結果が得られる順位の再現性が求められる。

ここで、製造データを不良原因との関連性に応じて順位づけるための指標には、不良品のデータ分布と良品のデータ分布の差を定量化する指標が用いられる。従来、このような2群の分布の差を定量化するための手法として、統計的仮説検定が用いられてきた¹⁾。統計的仮説検定は分布の代表値に対して、帰無仮説を設定し、仮説の妥当性をデータから評価する方法である。例えば、t検定は2群のデータの平均値が等しいという帰無仮説を設定し、その妥当性をデータから評価し、2つの分布の違いを定量化する²⁾。また、F検定は2群のデータの標準偏差が等しいという帰無仮説を設定し、その妥当性をデータから評価し、2つの分布の違いを定量化する²⁾。

また、2つの分布の代表値を比較する以外の方法で、分布の差を定量化する従来手法の一つに Kullback-Leibler (KL) divergence がある³⁾⁴⁾。KL divergence は2つの分布が等しいとき、またそのときに限り、0になる指標である。この KL divergence を用いれば、2つの分布の代表値を比較することなく、分布の差を定量化することができる。

しかし、これらの指標では、不良品のデータが5点程度と少ない時点で評価した順位と、不良品データが数百点と十分に多くなった時点で評価した順位が大きく異なることがある。これは、大量の不良品データを用いて特定された順位を不良原因を特定可能な真の順位と仮定した場合、データ点数が少ない場合には真の順位を捉えられていないことを意味し、早期のデータ解析が難しい原因となっている。

そこで、本研究では、不良品のデータ数は10点以下であるという条件下でも、大量の不良データを用いて算出した値に基づ

* 原稿受付 平成25年4月25日

** 非会員 (株) 日立製作所 横浜研究所 生産技術研究センター
(横浜市戸塚区吉田町 292)

*** 正会員 (株) 日立製作所 横浜研究所 生産技術研究センター
(横浜市戸塚区吉田町 292)

く順位と近い順位が算出可能である指標として、良品データの分布から不良品データが生起する確率を用いた KL divergence の推定方式の活用を提案し、その有効性を評価した。以降、本論文では、不良品のデータ数が 10 点以下であるという条件を少サンプル条件と略記することとする。

2. Kullback-Leibler divergence の推定量

KL divergence とは 2 つの分布の差を測る指標である³⁾⁴⁾。以下に定義を紹介する。

定義 1 (Kullback-Leibler divergence for density function)

$X \in \mathbf{R}$, $p(x)$ と $q(x)$ を X 上の確率密度関数とする。このとき、 $p(x)$ と $q(x)$ の間の KL divergence を以下で定義する。

$$KL(p||q) = - \int_X p(x) \log \left(\frac{q(x)}{p(x)} \right) dx \quad (1)$$

□

この KL divergence は 2 つの確率密度関数について対称ではない。すなわち、 $KL(p||q) = KL(q||p)$ が成立しない。また、三角不等式も成立しない。したがって、KL divergence は距離の定義を満たさない。しかし、 $p = q \Leftrightarrow KL(p||q) = 0$ が成り立つ。そのため、KL divergence は確率密度関数間の距離としてしばしば用いられる。前述のように、この KL divergence を用いることで、2 つの分布の差を定量化することができる。

本研究では、この KL divergence を製造データを不良原因との関連性に応じて順位づけるための指標として利用することを検討した。しかし、KL divergence を計算するためには、2 群の確率密度関数²⁾をあらかじめ同定しておく必要がある。

一方、実データから分布を同定する場合、データを生成した確率分布の形状が不明のため、正規分布などの分布をあらかじめ設定することが難しい場合がある。このような場合に、確率分布を推定する方法として、ヒストグラム推定をはじめとするノンパラメトリック推定⁴⁾が用いられる。

そこで、本研究では、データを基数 r のデータに離散化し、ヒストグラム推定によって分布を同定し、定義 2 で定義する確率質量関数に対する KL divergence を計算することとした。

定義 2 (Kullback-Leibler divergence for mass function)

X : 有限集合, $p(x)$ と $q(x)$ を X 上の確率質量関数とする。このとき、 $p(x)$ と $q(x)$ の間の KL divergence を以下で定義する。

$$KL(p||q) = - \sum_{x \in X} p(x) \log \left(\frac{q(x)}{p(x)} \right) \quad (2)$$

□

ヒストグラム推定では、データを基数 r のデータに離散化することで、基数 r の離散分布からデータが生起したものと考えることができる。すなわち、 $\mathcal{X} := \{x_1, x_2, \dots, x_r\} \in \mathbf{R}^r$ とおくと、全てのデータは、下記に定義する $\mathcal{X}$ 上の分布族 S に属する確率分布から生起したものと考えることができる⁵⁾。

$$S := \{p_\xi = p(x; \xi) \mid \xi = [\xi^1, \xi^2, \dots, \xi^r] \in \Xi\} \quad (3)$$

但し、

$$\Xi := \left\{ [\xi^1, \dots, \xi^r] \mid \xi > 0 \quad \text{for } \forall i, \sum_{i=1}^r \xi^i = 1 \right\} \quad (4)$$

$$p(x_i; \xi) := \begin{cases} \xi^i & (1 \leq i \leq n) \\ 1 - \sum_{i=1}^n \xi^i & (i=0) \end{cases}$$

とした。

このことを利用し、データから式 (4) における ξ を推定することで、2 つの確率質量関数 $p(x)$ と $q(x)$ を推定する。

しかし、不良品のデータは少サンプル条件のため、 ξ の推定結果が、大量データから推定した場合と、少量データから推定した場合で大きく異なることがある。そこで、少サンプル条件下における KL divergence の推定量として、以下の指標の活用を提案する。

定義 3 (Kullback-Leibler divergence の推定量)

$\forall p, q \in S$ を任意に固定する。確率分布 $p(x)$ から生起した n 点のデータ、確率分布 $q(x)$ から生起した m 点のデータをそれぞれ

$$\begin{aligned} G_p &= \{x_{p1}, x_{p2}, \dots, x_{pn}\} \\ G_q &= \{x_{q1}, x_{q2}, \dots, x_{qm}\} \end{aligned} \quad (5)$$

とおく。このとき、 $KL(p \parallel q)$ の推定量 $\widehat{KL}_n(G_p \parallel G_q)$ を以下で定義する。

$$\begin{aligned} \widehat{KL}_n(G_p \parallel G_q) &:= - \frac{1}{n} \cdot \log \left(\frac{n!}{n_{p1}! \cdot n_{p2}! \cdot \dots \cdot n_{pr}!} \cdot q_1^{n_{p1}} \cdot q_2^{n_{p2}} \cdot \dots \cdot q_r^{n_{pk}} \right) \\ &= - \frac{1}{n} \cdot \sum_{i=1}^n \log(i) \\ &\quad + \frac{1}{n} \cdot \sum_{j=1}^r \sum_{i=1}^{n_{pj}} \log(i) - \frac{1}{n} \cdot \sum_{j=1}^r n_{pj} \cdot \log(q_j) \end{aligned} \quad (6)$$

但し、 $\#$ で集合の濃度を表すとするとき、

$$\begin{aligned} n_{pi} &:= \# \{a \in \{1, \dots, m\} \mid x_{pa} = x_i\} \\ n_{qi} &:= \# \{a \in \{1, \dots, n\} \mid x_{qa} = x_i\} \\ q_j &:= \frac{n_{qj}}{m} \end{aligned} \quad (7)$$

とした。

□

次に本推定式の意味について説明する。本推定式は下記式の対数を取ってサンプル数で割ったものである。

$$P = \frac{n!}{n_{p1}! \cdot n_{p2}! \cdot \dots \cdot n_{pr}!} \cdot q_1^{n_{p1}} \cdot q_2^{n_{p2}} \cdot \dots \cdot q_r^{n_{pk}} \quad (8)$$

本研究では、式 (5) における G_p に不良品データ、 G_q に良品データを当てはめることで、不良品データと良品データの KL divergence を推定する方式を検討した。

上記の式 (8) において、 q_i は G_q のデータ点数が十分大きいとき、 $q(x_i)$ に等しいとみなすことができる。よって、上記式における $q_1^{n_{p1}} \cdot q_2^{n_{p2}} \cdot \dots \cdot q_r^{n_{pk}}$ は、確率分布 q から、独立にデータをサンプ

リングしたとき, $G_p = \{x_{p1}, x_{p2}, \dots, x_{pn}\}$ が (特定の整列順位で) 生起する確率を示している. また, $n = n_{p1} + n_{p2} + \dots + n_{pr}$ であるため, $\frac{n!}{n_{p1}! \cdot n_{p2}! \cdot \dots \cdot n_{pr}!}$ は多項係数である. したがって, 上記式は確率分布 q から, 集合 G_p が得られる確率を表している.

この推定量は KL divergence に確率収束する. すなわち, 次の定理が成り立つ.

定理 4 (漸近一致性)

$\forall p, q \in S$ を任意に固定する. 確率分布 $p(x)$ から生起した m 点のデータ, 確率分布 $q(x)$ から生起した n 点のデータをそれぞれ, 式 (5) に示すように, G_p, G_q とおく. このとき, 以下の等式が成り立つ.

$$\widehat{KL}_n(G_p \parallel G_q) \xrightarrow{p} KL(p \parallel q) \quad (m \rightarrow \infty)$$

但し, $X_i \xrightarrow{p} X \quad (n \rightarrow \infty)$ は, 確率変数列 X_i が X に確率収束することを表すこととした. $\square$

以上から, $\widehat{KL}_n(G_p \parallel G_q)$ は KL divergence に収束するため, $\widehat{KL}_n(G_p \parallel G_q)$ が KL divergence の推定量であることがわかる. 本研究では, 不良解析に対して, KL divergence の推定量である $\widehat{KL}_n(G_p \parallel G_q)$ を用いることを提案する.

3. 実データによる製品不良解析の評価

3.1 評価方法

実際の製品不良事例において, 提案指標である $\widehat{KL}_n(p \parallel q)$ と KL divergence, t 検定, f 検定を比較した結果を述べる.

ここで, t 検定と f 検定はデータから直接計算することができるが, KL divergence を算出するためには, まず, 不良品データの分布 $p(x)$ と良品データの分布 $q(x)$ を同定する必要がある. そこで, $\widehat{KL}_n(p \parallel q)$ を計算する際と同様に, データを離散化し, ヒストグラムを用いた方法で 2 つの分布を同定し, 確率質量関数に対する KL divergence (定義 2 参照) 算出することとした.

このようにして算出した KL divergence は, 離散化したデータに対する KL divergence ではあるが, 情報理論で知られる漸近等分割性⁶⁾を用いることで, サンプル数が十分大きいとき, 不良品データと良品データの真の分布 (離散化する前のデータの真の分布) から算出した KL divergence (定義 1 参照) に収束する指標であることを示すことができ, KL divergence の推定量として利用することが出来る.

これを説明するために, 確率収束の概念を用いて, 漸近等分割性を定義する.

定理 5 (漸近等分割性)

確率変数列 $X_1, X_2, \dots, X_n$ が互いに独立に X 上の確率分布 $p(x)$ に従うとする. データ列 $X_1, X_2, \dots, X_n$ が確率分布 $p(x)$ から観測される確率を $p(X_1, X_2, \dots, X_n)$ とするとき, 以下が成り立つ.

$$\frac{1}{n} \log p(X_1, X_2, \dots, X_n) \xrightarrow{p} \int_{x \in X} p(x) \log(p(x)) dx \quad (n \rightarrow \infty) \quad (9)$$

$\square$

漸近等分割性から, 次の定理が成り立つ.

Table 1 Experimental Data

item	value
Parameter Counts	468
Pass Data Counts	67175
Fail Data Counts	270

定理 6 $X_1, X_2, \dots, X_n$ が互いに独立に X 上の確率分布 $p(x)$ に従うとする. このとき, 以下の等式が成り立つ.

$$\frac{1}{n} \log \left(\frac{q(X_1, X_2, \dots, X_n)}{p(X_1, X_2, \dots, X_n)} \right) \xrightarrow{p} -KL(p \parallel q) \quad (n \rightarrow \infty) \quad (10)$$

$\square$

この定理から, $X_1, X_2, \dots, X_n$ が互いに独立に X 上の確率分布 $p(x)$ に従うとき,

$$AEP_n(p \parallel q) := -\frac{1}{n} \log \left(\frac{p(X_1, X_2, \dots, X_n)}{q(X_1, X_2, \dots, X_n)} \right) \quad (11)$$

とおくと, $AEP_n(p \parallel q)$ は KL divergence の有限個のデータによる近似になっていることがわかる.

また, この $AEP_n(p \parallel q)$ は, $X_1, X_2, \dots, X_n$ が離散化されたデータ, $p(x)$ が $X_1, X_2, \dots, X_n$ からヒストグラム推定によって推定した離散確率分布, $q(x)$ が $p(x)$ と同じ有限集合上で定義された離散確率分布である場合, $p(x)$ と $q(x)$ から算出した KL divergence に一致することを確認することができる.

以上から, データを離散化して計算した KL divergence は, サンプル数が十分大きいとき, 不良品データと良品データの真の分布から算出した KL divergence に収束する指標であることが分かる.

そこで, 本研究では, 比較対象の KL divergence の計算方法として, $p(x)$ を不良品データ, $q(x)$ を良品データの従う確率質量関数とし, データを離散化して計算した KL divergence である, $AEP_n(p \parallel q)$ を用いることとした.

なお, $p(x)$ と $q(x)$ の役割を入れ替え, $p(x)$ を良品データ, $q(x)$ を不良品データの従う確率質量関数とし, $AEP_n(p \parallel q)$ を計算する方式を利用することも考えられる. その場合, 式 (11) における $X_1, X_2, \dots, X_n$ に良品データからサンプリングされたデータを用いることになる. 一方, $q(x)$ は少サンプルである不良品データから推定することになり, $q(x) = 0$ となる x が生じる可能性があり, $q(X_1, X_2, \dots, X_n) = 0$ となり, 結果として, $AEP_n(p \parallel q)$ が発散してしまう恐れがある. そこで, $p(x)$ と $q(x)$ の役割を入れ替えて計算した $AEP_n(p \parallel q)$ は用いないこととした.

3.2 実験方法

実験方法について述べる. まず, 評価に用いたデータを表 1 に示す. 対象パラメータ数 (Parameter Counts) が 468 パラメータ, 良品サンプル数 (Pass Data Counts) が 67175 サンプル, 不良品サンプル数 (Fail Data Counts) が 270 サンプルのデータである.

次に, 評価実験の手順について述べる. 最初に全不良品データと全良品データからなるデータセット 1 と, 乱数を用いて不良品データからランダムに 5 点選び出した不良品データと全良品データからなるデータセット 2 を準備する. まず, データセット 1 を用いて, 各パラメータごとに $\widehat{KL}_n(p \parallel q)$ を計算し,

Table 2 Experimental Result of Median Value

Method	5pt	50pt
$\widehat{KL}_n(p \parallel q)$	0.72	0.95
$AEP_n(p \parallel q)$	0.69	0.92
t-test	0.53	0.89
f-test	0.24	0.68

$\widehat{KL}_n(p \parallel q)$ により算出した製造データの順位を推定すべき真の順位と考える。次に、データセット 2 を用いて、各パラメータごとに $\widehat{KL}_n(p \parallel q)$ を計算し、 $\widehat{KL}_n(p \parallel q)$ により算出した製造データの順位を少サンプル条件下で推定した順位と考える。上記の手順により算出した 2 つの順位が類似していればよいほど、良い指標と考えることができるため、2 つの順位の相関係数を算出し、それを評価指標とした。同様の手順による相関係数の算出を $AEP_n(p \parallel q)$ 、t 検定の p 値、f 検定の p 値に対しても行い、指標の比較を実施した。

ここで算出される相関係数はデータセット 2 を生成する際の乱数に依存して値が変動することが予想されるため、同様の実験を 1000 回繰り返し、相関係数の分布の違いを比較することとした。また、比較のため、同様の実験を選出する不良データの数 50 点にしたケースを評価した。

3.3 実験結果

不良データ数が 5 点の実験結果を図 1 に、不良データ数が 50 点の実験結果を図 2 に示す。各図ともに、横軸が指標、縦軸が順位相関を取っており、箱ひげ図を用いて 1000 回の実験結果の分布を可視化してある。箱の上と下の線がそれぞれ、確率分布の 75% 点、25% 点に対応し、箱の中央の線が中央値、上方向に伸びたひげの上端が分布の最大値、上方向に伸びたひげの下端が分布の最小値である。各図における (a) 図は全指標を比較したものであり、(b) 図は $\widehat{KL}_n(p \parallel q)$ と $AEP_n(p \parallel q)$ の結果のみを拡大した図である。図 1、2 に示した箱ひげ図における中央値の値を、表 2 に示す。

図 1 において負の値が存在するのは、サンプル数が少なく評価の安定性を欠いた結果、順位相関が負の値となってしまった実験結果が存在することを意味する。この現象は、サンプル数を増やした図 2 では解消されており、サンプル数の増加により、評価が安定することを示している。

1000 回の繰り返し実験の結果、 $\widehat{KL}_n(G_p \parallel G_q)$ は、 $AEP_n(p \parallel q)$ や t 検定、f 検定と比較し、順位相関の分布位置が高く。これは、不良品サンプル数が 5 点の場合に、不良品データ数が 270 点のデータで算出した場合にもっとも近い解析結果を導くことが出来ると考えられ、今回評価した指標の中では最も少サンプル条件での評価において安定した方式であるという結果を得た。

また、図 2 に示す不良品サンプル数が 50 点の場合においても、 $\widehat{KL}_n(G_p \parallel G_q)$ の有意性は保たれているが、他の指標との差は小さくなっており、これはサンプル数の増加と共に評価が安定した結果と考える。

4. 実験結果に対する考察

4.1 サンプルングに対するばらつきの評価方法

3.3 節で述べた結果より、 $\widehat{KL}_n(p \parallel q)$ は、今回評価した指標の中では最も少サンプル条件での評価に安定した方式であるという結果を得た。本章では、この結果に対して、特に近い結果

Fig. 1 Comparison Result of 5pt

を得た $AEP_n(p \parallel q)$ と比較することで、論理的な考察を行う。

少サンプル条件下で、順位を決定するための指標を算出する場合、指標の計算結果は 5 点のデータセットに影響される。3 章の評価では、270 点の不良データから、5 点と 50 点のサンプルリングをランダムに 1000 回繰り返し評価することで、指標に基づく順位の再現性を順位相関で評価した。ここで、サンプルリングに依存して値が大きくばらつく指標を用いると、順位決定においても結果がばらつくことが予想される。

そこで、 $\widehat{KL}_n(p \parallel q)$ と $AEP_n(p \parallel q)$ のサンプルリングに依存したばらつきを評価した。

定義 7 (サンプルングに対するばらつきの評価指標)

順位づけのための指標を $\hat{\theta}$ とおくと、サンプルングに対するばらつきの評価指標 $R(\hat{\theta})$ を以下で定義する。

$$R(\hat{\theta}) := E[\|\hat{\theta} - E[\hat{\theta}]\|^2] \quad (12)$$

但し、 $E[\cdot]$ は期待値を表すとする。

□

サンプルングに対するばらつきの評価指標として $R(\hat{\theta})$ を用いた理由について説明する。理由を述べるために、まず推定量の期待二乗誤差のバイアス・バリアンズ分解を紹介する^{4,7)}。

定理 8 (バイアス・バリアンズ分解)

$\hat{\theta}$ を推定量、 θ^* を真の値とすると、推定量と真の値の期待二乗誤差は次のように分解できる。

$$\begin{aligned} E[\|\hat{\theta} - \theta^*\|^2] \\ = E[\|\hat{\theta} - E[\hat{\theta}] + E[\hat{\theta}] - \theta^*\|^2] \end{aligned}$$

Fig. 2 Comparison Result of 50pt

Fig. 3 Flow Chart of Variance Evaluation

□

バイアス・バリエーション分解における第1項をバイアス項、第2項はバリエーション項とよぶ。バイアス項は推定量の期待値と真の値の距離を示す。したがって、バイアス項は推定量の真の値に対する近似精度を表現していると考えられる。バリエーション項は推定量の期待値に対する推定量のばらつきを示す。したがって、バリエーション項が推定量のサンプリングによるばらつきを表現していると考えられる。そのため、一般に、バリエーション項は不偏推定量の有効性の評価指標として用いられる⁷⁾。

本研究の目的は少サンプル条件のデータを用いて、複数の製造データの中から不良品と良品のデータの分布が異なる製造データを特定することで、製品不良の原因特定を行うことである。不良品と良品のデータの分布が異なる製造データを特定するために、KL divergenceの大きさを推定し、複数の製造データ間で比較する。この際、推定量のサンプリングによるばらつきが大きいと、サンプリングに応じて、比較結果が異なってしまう。そのため、サンプリングによるばらつきがより少ない推定量が優れた推定量と考えることができる。前述のバリエーション項は推定量のサンプリングによるばらつきを評価する評価指標として適している。

そこで本研究では、バイアス・バリエーション分解におけるバリエーション成分を $R(\hat{\theta})$ で表し、サンプリングに対するばらつきの評価指標とした。

4.2 シミュレーション実験によるバリエーションの評価方法

本章では、シミュレーションを用いて、順位づけのための指標のバリエーションを評価した結果について説明する。3章の評価結果から、t検定とf検定は他の2方式と比べて、性能が劣るため、本章では、 $\widehat{KL}_n(p || q)$ と $AEP_n(p||q)$ のみを比較対象と

した。実施したシミュレーション実験の方法を説明する。

本研究では、2つの群のうち、一方の群のサンプル数が少ないデータ (≤ 10) を対象としている。そこで、サンプル数が少ない群のサンプル数を n とおくと、 $n = 5, 10$ の場合と、参考として、 $n = 50$ の場合について実験を行った。

図3はシミュレーション実験のフローチャートである。まず、ディリクレ分布からランダムサンプリングすることで、2つの離散分布 $p(x)$ と $q(x)$ を生成する (図3(1))。なお、生成する離散分布の基数は5、ディリクレ分布のパラメータは、条件1: [1 1 1 1 1]、条件2: [2 3 4 5 6] の2通りの場合について実験を行った。条件1は、良品と不良品の分布が大きく異なる傾向にあり、条件2は、良品と不良品の分布が似る傾向を持つ。次に離散分布 $p(x)$ から n 点のデータを、 $q(x)$ から 10000 点のデータをランダムサンプリングし (図3(2)(3))、 $\widehat{KL}_n(p || q)$ と $AEP_n(p||q)$ を計算する (図3(4))。この操作を100回繰り返し、 $\widehat{KL}_n(p || q)$ と $AEP_n(p||q)$ のバリエーションを計算する。

この操作を100回繰り返すことで、 $\widehat{KL}_n(p || q)$ と $AEP_n(p||q)$ のバリエーションを100回計算し、比較をした (図3(6))。

4.3 シミュレーション実験結果及び考察

シミュレーション実験の結果を図4、図5に示す。図4、図5はそれぞれ、 $n = 5, 50$ のときの結果である。横軸が推定方式を示している。左は $\widehat{KL}_n(p || q)$ を用いた推定方式、右は $AEP_n(p||q)$ を用いた推定方式を示している。縦軸がバリエーションを示している。

本実験の結果、少サンプル条件下 ($n = 5, 10$) において、 $\widehat{KL}_n(G_p || G_q)$ のバリエーションの方が $AEP_n(p||q)$ のバリエーションよりも小さいことから、 $\widehat{KL}_n(G_p || G_q)$ の方が、 $AEP_n(p||q)$ に比べ、サンプリングの影響を受けない指標であることが証明

Fig. 4 Variance Evaluation Result ($n = 5$)

Fig. 5 Variance Evaluation Result ($n = 50$)

出来た。

したがって、 $\widehat{KL}_n(p || q)$ の方が、サンプリングに依存したばらつきが小さいため、少サンプル条件下でも大量の不良データを用いて算出した値に基づく順位と近い順位が算出可能な、安定した指標であると考えられる。

5. ま と め

本研究ではデータ解析アプローチにおける製品不良要因の特定方法を検討した。製品不良の原因特定方法をデータ解析で行う上での課題は、不良品データ数が 10 点以下と少ない条件下でも、大量の不良品データを用いて算出した評価指標に基づく各製造データの順位とより近い順位を算出することである。

この課題を解決するため、本研究では不良データの集合が生起する確率に基づく Kullback-Leibler divergence の推定量を用いた。本方式を実データを用いて評価したところ、比較方式と比べ、不良品サンプル数が 5 サンプルしか入手できなかったとしても、不良品サンプル数が 270 サンプル以上あるデータで算出した場合により近い結果を導くことができることを確認した。

また、シミュレーション評価した結果、漸近等分割性に基づく Kullback-Leibler divergence の推定との比較実験によりバリエーションが小さいことを確認し、不良データ数が少ない条件下でも、算出結果が安定した指標であることを確認した。これらの結果より、Kullback-Leibler divergence の推定量は、少サンプル条件下での不良原因の特定に適していると考えられる。

6. 今後の展望

本報告では、 $\widehat{KL}_n(G_p || G_q)$ のバリエーション項をシミュレーション実験を用いて評価した。推定量のバリエーション項の大きさはクラメル・ラオの不等式により下限値が評価可能である。したがって、本報告による推定量がバリエーションの観点で最適な推定量であるかどうかを検証するため、クラメル・ラオによる下限値の関係を明らかにすることが今後の課題である。

参 考 文 献

- 1) 永田靖：品質管理のための統計手法，日本経済新聞社（2006）。
- 2) 稲垣宣生：数理統計学，裳華房（1990）。
- 3) S.Kullback, R.A.Leibler: On Information and Sufficiency, Ann. Math. Statist. **22**, 1 (1951), 79.
- 4) C.M.Bishop: Pattern Recognition and Machine Learning, Springer (2006)。
- 5) 有本卓，甘利俊一，長岡浩司：情報幾何の方法，岩波書店（1998）。
- 6) T.M.Cover, J.A.Thomas: Elements of Information Theory, John Wiley & Sons (2006)。
- 7) 杉山将：統計的機械学習，オーム社（2009）。