統計研究彙報　第 72 号　2015 年 3 月　（41～54）

統計実務におけるレンジチェックのための外れ値検出方法

野呂　竜夫 †、和田　かず美 †

A Univariate Outlier Detection Manual for Tabulating Statistical Survey

NORO, Tatsuo
WADA, Kazumi

統計調査における調査データは、調査対象から得た情報が調査票等に記入される。その後、調査実施者が集計を行うが、集計前に調査票等の記入内容に誤りがないかなどの様々な審査が行われる。例えば数量項目の場合、記入内容の数値が調査単位ごとに大きすぎ（小さすぎ）ないかを確認する。政府が実施する統計調査はデータ量が多いので、確認すべきデータの基準（レンジ：人手による審査をしない正常値とみなす値の範囲）は、前もって定めておく必要がある。通常、レンジは実際の調査データから算出するが、調査データには外れ値が含まれるので、その影響を受けやすい平均値や標準偏差をもとにレンジを算出すると正常値とみなされる範囲が広がり検出漏れを起こす可能性がある。

本稿では、単峰で変換などによりある程度の対称性が確保されるようなデータについてのレンジを用いた外れ値検出の適切な方法を解説するとともに、サービス産業動向調査のデータを用いてレンジチェックのレンジの設定を行った事例を紹介する。

キーワード：中央値、四分位値、頑健、四分位範囲、レンジチェック

As for the official statistical surveys, information obtained from the survey units is filled in the questionnaires. Afterwards, the electronized information is examined to eliminate the erroneous data before tabulation. The examination includes detecting extreme values using editing bounds for the quantitative data. Since the data size of the official statistical surveys tends to be large, the editing bounds are often prepared in advance using the previous survey data in which outliers are inevitable. Under ordinary circumstances, outliers may not be properly detected if the bounds are set based on the mean and the standard deviation. The bounds will be grossly inflated and fail to detect the outliers as the mean and the standard deviation are not robust and easily affected by the extreme values.

In this paper, we explain how to prepare the editing bounds by means of robust statistics for data with a unimodal and symmetric distribution (after transformation, if necessary). We also show the case of monthly sales data for the Monthly Survey on Service Industries.

Keywords: Median, Quartile, Robust, Inter Quartile Range, Editing Bounds.

†独立行政法人統計センター統計情報・技術部統計技術研究課

1. はじめに

統計調査データにおいては、記入誤りや入力誤りあるいは間違いではないが特異な値など、外れ値の存在は避けることができない。ところが、調査実務において、レンジチェックのための方法論はこれまで整備されておらず、（算術）平均値と標準偏差を用いて上限値及び下限値を設定するような、検出漏れを起こしやすい方法が使用される場合もある。

このような状況を改善するため、数量項目データが単峰で必要により変換などを施しておおむね対称な分布とみなせるが、分布などが未知な外れ値が混入しているデータを想定し、このようなデータにおける外れ値検出のためのレンジの設定方法について、実務担当者に向けた手引きを提供することを本稿の目的とする。

本稿では、第２章において本研究にいたる背景を記し、第３章にて検出すべき外れ値の概念（定義）とその一般的な検出方法を示す。第４章では、第３章の外れ値検出方法を、サービス産業動向調査の調査票データを用いて実際に検証を行った結果について示している。最後に第５章では、本稿で述べる外れ値検出方法の可能性についてまとめている。

2. 背景

公的統計調査の製表業務においては、収集された調査票情報を電子化し、その後集計作業が行われる。まず、調査票の記入内容について審査（データチェック）を行い、何らかの誤りの可能性のあるデータを検出・修正して推計値が真値と乖離しないようにしている。数量項目については、極端に大きいあるいは小さい数値は、確認を必要とするデータとして検出し、個別に訂正の必要性を担当者が検討する。この確認すべきデータを抽出するためにあらかじめ許容される上限値や下限値を項目別に設定することがある。この上限値や下限値は過去のデータを基にして定められることが多い。

統計調査データは大抵の場合、正規分布などの理論上の分布には従わない。また、非対称な分布となることが多く、外れ値の存在は避けることが出来ない。一般に、製造現場での品質管理における、平均値から標準偏差の２倍・３倍といった管理基準の設定方法は広く知られており、これが統計調査データにおける外れ値検出に流用されるケースがよくみられるが、このような品質管理は個々の製品の異常というよりは管理された製造工程自体の異常を検出するためのものであり、最初から外れ値の存在が想定される統計調査データの外れ値検出基準には、次に述べるような理由で適さないことに留意が必要である。

平均値や標準偏差が外れ値の影響を受けやすく、正常値の範囲を決めるためにこれらを使用する場合に極端な値が存在すれば正常値とみなされる範囲が広がり、検出すべき外れ値を見逃してしまう現象が起きる。この現象はマスキングと呼ばれる（Wilcox, 2012）。信頼できる調査結果を公表するためには、外れ値の影響を受けにくい適切な外れ値検出手法を用いなければならない。

2.1 平均と標準偏差の問題点

正規分布が想定されるとき、データの位置を示す指標として平均値、バラツキの指標として標準偏差があり、データ分布の大部分を含む範囲を決めるために平均値から標準偏差の２倍あるいは３倍離れた値をレンジと考えることは、統計学的に自然な方法である。ただし、実際の統計調査において、そのような方法を適用するには二つの問題がある。

まず実際の調査データの分布はたとえ単峰であっても正規分布よりは裾が長い場合が多い。

次に、データの中に意図的に値が大きすぎるあるいは小さすぎるものを少量混入させたとき、平均値はその影響を直接受けて値が変わり、標準偏差は平均値よりもさらに大きな影響を受ける。レンジの設定は、いわばどのデータが外れ値であるかを判定するための「ものさし」を作る作業であるが、そのものさし自体に、検出すべき外れ値の影響を受けやすい、つまり頑健性のない統計量を使えば、その検出法は信頼できるものにはならない。

なお、分布が対称でない場合、位置の基準に平均値のみあるいは中央値（メディアン：Median）のみを用いても上限値及び下限値の設定には不都合なことが多い。ただし、何らかの変換を行うことにより対称な分布に近似する場合には変換後のデータの分布に応じて上限値及び下限値を設定することが適切である。それでも完全には対称にはならないので、上限値と下限値に異なる位置の基準を用いることによって分布の歪みを考慮したレンジの設定を行うことが出来る。数値データを用いた検証例を、第３章３節及び第４章１節で紹介する。

2.2 頑健な統計量

では、外れ値の影響を受けにくい、頑健性のある統計量にはどのようなものがあるだろうか。Tukey(1977)は、探索的データ解析（EDA：Exploratory Data Analysis）と呼ばれる手法を考案し、データの分布があらかじめ想定できない場合、順序統計量を基礎とした統計量（推定量）を分布の位置やバラツキの推定に用いた。代表的な統計量は、中央値、四分位値及び四分位範囲(IQR:Inter Quartile Range)である。中央値は平均値と同じくデータの位置を示す指標であり、四分位値から導かれる四分位範囲は標準偏差と同様にデータのバラツキを示す指標となる。これらの順序統計量は、文字通りデータの順序のみにより決まり、データ内にある程度の量（中央値ならば 50%、四分位値ならば 25%）の外れ値が存在しても統計量に対して影響を受けない。Tukey は、必ずしも正規性が保証されない観測データから外れ値を検出するためのレンジを、四分位範囲に対する四分位値と観測値の差（絶対値）の比率により定め、さらにそれを可視化する方法として箱ひげ図（box plot または box-and-whisker plot）を提案した。この箱ひげ図は、四分位値を基準として下限と上限で異なるレンジになるため、ある程度は非対称な分布のデータにも適用することができる。

特に経済系の統計調査については、データの分布が正規分布から乖離していることが多く、正規分布の前提を置くパラメトリックな方法は適用が限られる。本稿で提案する方法は、分布については単峰という比較的緩やかな仮定を置き、ある程度の対称性を確保するため、極端な分布の歪みには変数変換を施す。さらに時系列や産業などによる軽微な対称性の崩れには、非対称なレンジ設定ができて外れ値にも頑健な箱ひげ図の考え方を用いることにより、マスキングが起こりにくい柔軟性の高いレンジを設定する。サービス産業動向調査の月次売上高データについて、この方法の事例として第４章で紹介する。

3. 外れ値とその適切な検出方法

3.1 外れ値の定義

一般的に、外れ値とは、データの大部分の傾向と異なるもので、必ずしも誤りとは限らないが、データ集計や分析の際にその存在が結果の精度を悪化させる可能性があるものを指す。調査統計の分野においても、国連統計部（United Nations Statistical Commission: UNSC）と国連欧州経済委員会(United Nations Economic Commission for Europe: UNECE) が 2000 年に刊行した“Glossary of Terms on Statistical Data Editing”では、「データ値の集合の統計的分布の裾に在るデータ値を指す」と定義されている。また、測定ミス・記録ミス等に起因する「異常値（あるいは特異値）」とは概念的に異なるが、実用上は区別できないこともある。通常、外れ値は異常値を含む概念（異常値⊂外れ値）とされている。

ただし、外れ値はその検出目的などに応じて様々な定義がされる。本稿の目的は、誤りの可能性のあるデータを検出するデータチェックのプロセスに適用する手法を示すことにあるので、外れ値の定義を「統計値において同一項目の他の値から著しく離れた値であり、何らかの誤りを含む可能性が高いものを指す」こととする。

外れ値の対応方針として、数理統計学的な先行研究をとりまとめた Barnett and Lewis(1994) は、正しいデータを誤って外れ値として除外しないよう、一方で実際に外れ値である極端な値を採用することも避けながら、外れ値の影響を緩和することのできる頑健な推定手法を利用することを提唱しており、本稿においてもこの方針を踏襲している。

3.2 平均値と標準偏差により起こるマスキングの例

以下、外れ値と考えられるデータを例示する。

まず、外れ値が存在しないと考えられるデータとして、図１では、平均４、標準偏差１（分散１）の正規母集団から無作為に 50 個データを取り出したものをヒストグラムとして表示したものである。曲線は、平均４、分散１の正規分布の確率密度関数を表している。

図１：外れ値のない分布

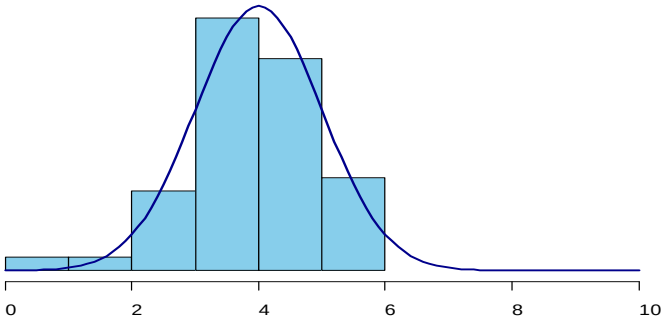
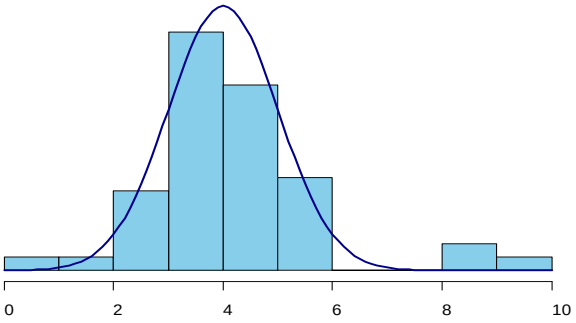


図２は、平均 4、分散 1 の正規母集団から無作為に 47 個、平均９、分散 0.25（標準偏差 0.5）の正規母集団から無作為に３個抽出し、合計 50 個のデータによりヒストグラムを描い

た。３個のデータが視覚的に外れ値と認識され得る。

図２：外れ値をもつ分布

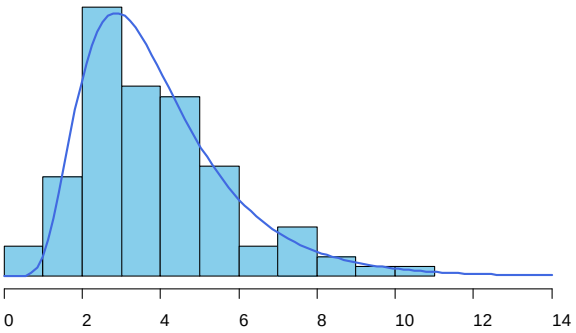


次に、外れ値と認識するのが難しい分布の例を挙げる。

図３は、分布の山が左側（０）に偏っている（右に歪んだ）分布の例である。このように非対称な分布では、右裾が長いのでこのままで計量的に外れ値を検出するのは困難である。こういった場合でも変数変換を施すことにより対称分布に近づくことで、変換後の分布に基づき外れ値検出が可能となる。

このような例は、世帯の所得金額などの経済データによく見られる。第４章で具体例を示すが、経済データでは、多くの場合対数変換により正規分布に近づく。

図３：右に歪んだ分布

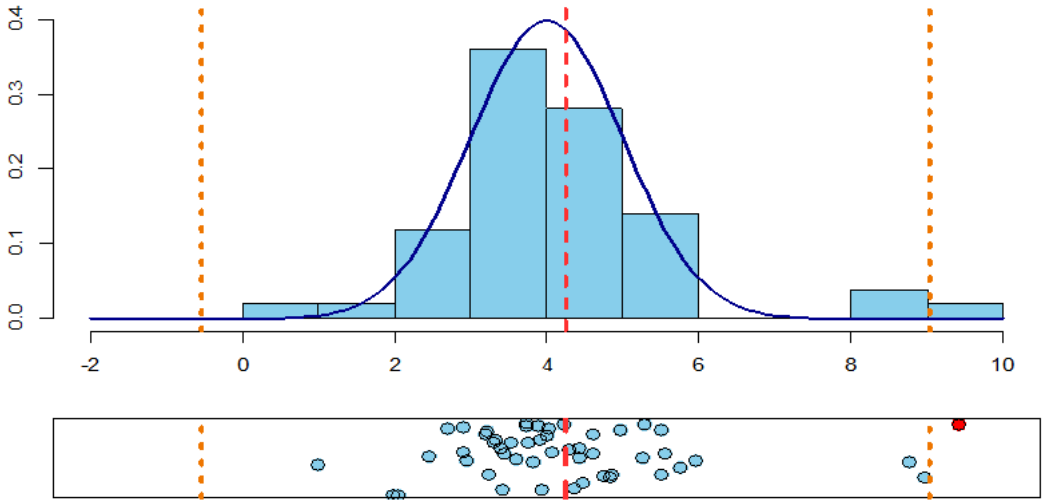


「外れ値」は画一的な判定方法が存在しないため、統計データの性質によりそれぞれの方法が行われている。平均値と標準偏差を用いた典型的な方法を以下に示し、その問題点を説明する。

母集団の分布を正規分布と想定する場合、観測されたデータをすべて用いてそのデータの平均値や標準偏差（分散）を求め、観測値と平均値の差が標準偏差の３倍を超えると外れ値とすることが製造業の品質管理において行われている（吉澤，2004）。正規母集団では平均を中心に標準偏差の３倍以内に入る確率が 0.9974 となるからである。

しかし、経済データにおいては、観測データから得られる平均値及び標準偏差が外れ値によって偏ったり拡大したりすると、それにより外れ値検出に誤りを生じやすい。図４は外れ値が存在しているため、平均値と標準偏差が大きくなり、両側のレンジ（橙破線）が広がってしまい、上側３個のデータのうち１個だけが外れ値とされた。

図４：マスキングの例



3.3 頑健な外れ値検出方法の例

外れ値が存在し得るデータを扱う場合には、レンジ設定には平均値、標準偏差を用いるべきではないことを前節で実例を挙げて示した。このような場合において、外れ値による影響を受けにくい（頑健な）統計量としては分位数がある。最もよく利用されるのが、中央値及び四分位値である。まず、中央値はデータを大きさの順に並べたとき、大きいグループと小さいグループに同数ずつに２分する位置にあるデータの値をいう。データ数が偶数のときは２分する２つのデータの平均を中央値とする(Everitt, 1998)。すなわち、n 個のデータ $X_1, X_2, \cdots, X_n$ を値の小さい方から順に並べて

$$X_{(1)},\ X_{(2)},\ \cdots,\ X_{(n)}$$

としたとき、中央値 $\tilde{X}$ を

奇数$(n = 2m-1)$のとき　$\tilde{X} = X_{(m)}$

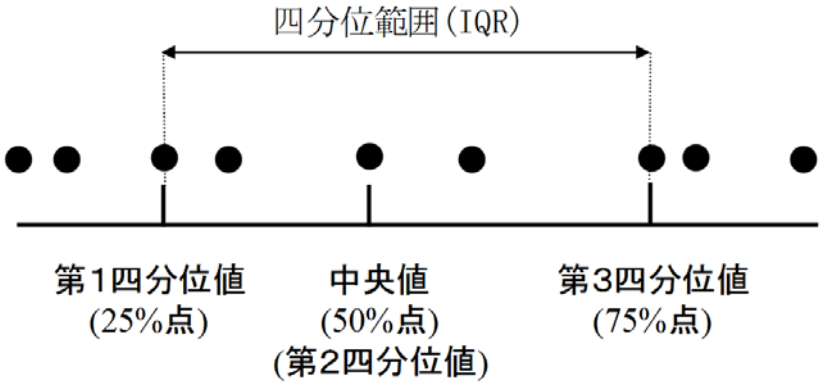
偶数$(n = 2m)$のとき　$\tilde{X} = \frac{X_{(m)}+X_{(m+1)}}{2}$

で定義する。

次に、第１四分位値と第３四分位値については、いくつかの定義が存在する(Frigge *et al.*, 1989)が、一般的なのは、中央値以下のデータの中央値を第１四分位値、中央値以上のデータの中央値を第３四分位値とするものである。そして、第１四分位値と第３四分位値の差を四分位範囲と定義し、データのバラツキの尺度とする。

例）　観測データが１から９までの９つの整数の場合、中央値は５であり、第１四分位値は３（１から５までの整数の中央値)、第３四分位値は７（５から９までの整数の中央値）である。これを図示すると、図５のようになる。

図５：中央値、四分位値及び四分位範囲



観測データの中央値及び四分位値は、それらを定める１個または２個のデータ以外の影響を受けないので、外れ値に強い（頑健)。例えば、上の９つのデータでは、６，７，８，９の４個の値がどんなに大きい値を取っても中央値５は変わらない。

外れ値に頑健な中央値と四分位範囲を用いて、平均値と標準偏差を推定してレンジ設定する方法を考えてみる。母集団分布を正規分布と仮定すると、母集団中央値は母集団平均値に等しく、標準正規分布の 75%点は約 0.6745 であるので、四分位範囲＝標準偏差×0.6745×2 より

$$3 \times \text{標準偏差} = 2.224 \times \text{四分位範囲} \qquad (1)$$

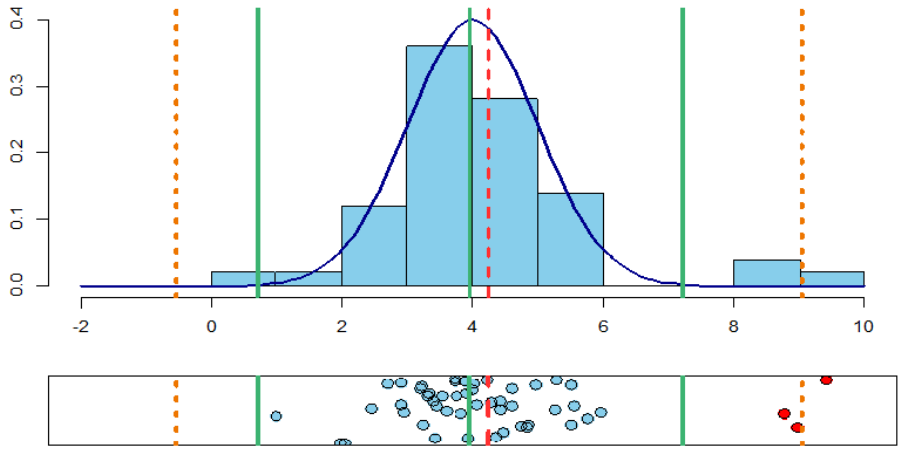
となるので、標準偏差の３倍を四分位範囲の 2.224 倍に、平均値を中央値に置き換えて

$$\text{下限値} = \text{中央値} - 2.224 \times \text{四分位範囲}$$
$$\text{上限値} = \text{中央値} + 2.224 \times \text{四分位範囲}$$

とすると、外れ値の影響を受けにくいレンジ設定を行うことができる。Schwertman *et al.*(2004) は有限標本の場合の倍率の調整について検討している。

図４と同一のデータについて、中央値と四分位範囲を基準にレンジを設定すると図６のとおりとなる。緑線がレンジ（上限値及び下限値）となり赤点の３つのデータを外れ値として検出している。

図６　中央値及び四分位範囲を用いたレンジチェック



経済データにおいては、非対称な分布をもつことが経験的によく知られているので、分布の位置の基準に中央値を用いるより、分布の歪みを反映した第１四分位値及び第３四分位値を用いる方がレンジの設定に有用と考えられる。正規母集団との整合性を考慮すると、対称な母集団分布において中央値と四分位値との差は四分位範囲の２分の１であるから
2.224－0.5＝1.724 より

$$\text{下限値} = \text{第１四分位値} - 1.724 \times \text{四分位範囲}$$
$$\text{上限値} = \text{第３四分位値} + 1.724 \times \text{四分位範囲}$$

となる。

ところで上記の 1.724 を 1.5 に換えると四分位範囲を箱型（長方形）に、第１四分位値と第３四分位値からそれぞれ四分位範囲の 1.5 倍の長さの線をひげ（髭）としてデータの分布状況を図示した Tukey(1977)の箱ひげ図の考え方に該当する（吉澤，1990）。

4.　調査データによる分析

本章では、総務省が実施するサービス産業動向調査について、月次の売上高データを用いて、第３章３節で示したレンジチェック方法の適用と検証を行う。

レンジは、産業別に売上高については上限のみ、事業従事者１人当たり売上高については上限及び下限を設定する。レンジ算出に用いる月次データは当該月前 24 か月分のデータのうち、資本金１億円以下の事業所で実際に調査された売上高を用いた。なお、売上高が０あるいは負数となっている事業所は除外している。

4.1　検証方法

ここでは、頑健性がない平均値と標準偏差によるため外れ値検出法としては適さない方法を含め、第３章で検討した次の①～③に示す３つの方法によりレンジを設定する。①のレンジは、参照用に平均値から標準偏差の３倍に設定する。②のレンジは①の方法を頑健化し、

平均値を中央値、標準偏差を四分位範囲に置き替え、その範囲が正規分布の場合に標準偏差の３倍に相当するよう四分位範囲に 2.224 を乗じている。②は①よりも頑健であるが、①と同様に分布は対称であることを前提としている。一方で、中央値ではなく第１及び第３四分位値を基点とする③のレンジは、ある程度の分布の歪みに対応することができる。ただし、大きな歪みに対応する能力はない。

① 平均値及び標準偏差に基づくレンジ（対称分布用）

下限値 ＝ 平均値 － 3 ×標準偏差
上限値 ＝ 平均値 ＋ 3 ×標準偏差

② 頑健な推定量（中央値及び四分位範囲）に基づくレンジ（対称分布用）

下限値 ＝ 中央値 － 2.224 × 四分位範囲
上限値 ＝ 中央値 ＋ 2.224 × 四分位範囲

③ 分布の歪み（非対称性）を考慮したレンジ

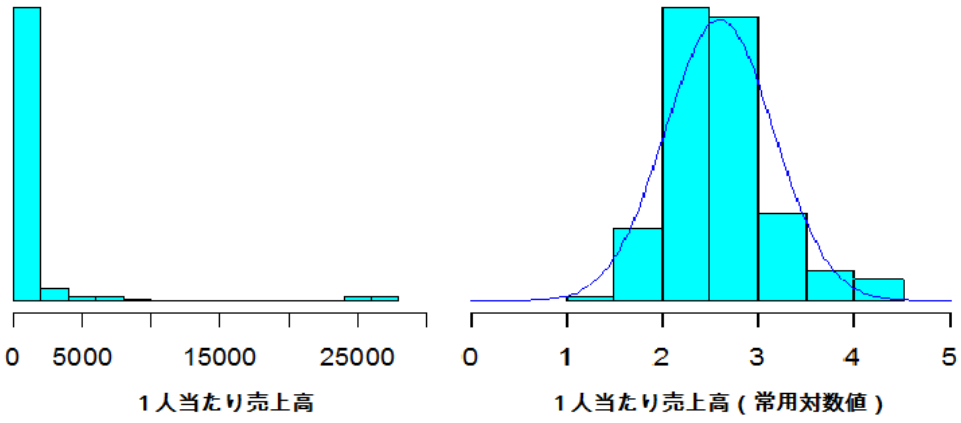
下限値 ＝ 第１四分位値 － 1.724 × 四分位範囲
上限値 ＝ 第３四分位値 ＋ 1.724 × 四分位範囲

正規分布想定で標準偏差の３倍という基準にそろえるならば、四分位範囲の 1.724 倍になる。ただし、この数字は実際データの分布の裾の長さに応じて外れ値の検出量を調整することにより調査担当者が任意で設定する。

実際の統計調査データ、例えば売上高や資本金等の金額データは、右裾が長い歪んだ分布となりやすい。このようなデータは、平方根変換や対数変換などの変数変換を施すことにより、その分布を正規分布に近づけることができる場合が多い。

実例として、事業従事者１人当たりの売上高データを図７に示す。変換していない元のデータのヒストグラムが左側、同じデータを常用対数変換した後のものが右側である。右側の変換後のヒストグラムには、正規分布の密度関数を青い曲線で表示した。対数変換後の事業従事者１人当たりの売上高データが、この正規分布に近似していることがわかる。

図７：実データと対数変換後データの分布



そこで、このように大きな歪みのある売上高データのレンジ設定は、まずデータ分布を対称に近づけるために対数変換を行い、その後で③の方法を適用することにより、わずかな歪みにも対応できるレンジを設定する。

4.2 検証結果

本稿では、まず上述の②の方法を基本として、データにあわせてカスタマイズを行い、三つの方法について、まずAの方法、次いでB・Cと３通りの比較を行い、それぞれの方法の問題点や特徴を明らかにする。

A. ①と②の方法の比較

図８及び図９では、「貸家業、貸間業」の平成 23 年２月分（154 データ）を使用している。両図とも縦軸に売上高、横軸に事業従事者数をとった散布図を常用対数目盛で描き、24 か月分（平成 21 年２月～23 年１月）データを使用して、図８では①の平均値及び標準偏差を用いた方法によりレンジを求め、レンジとなるその境界線を描画している。これに対して、図９では②に示した中央値と四分位範囲（2.224 倍）から算出したレンジを示している。

図８：平均値及び標準偏差に基づくレンジ

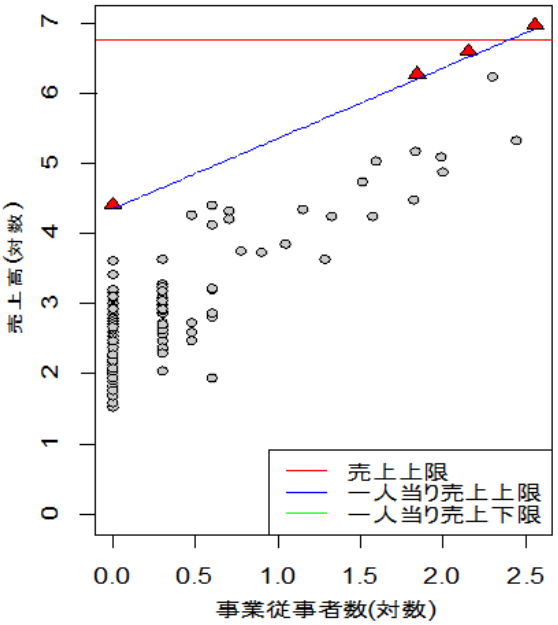
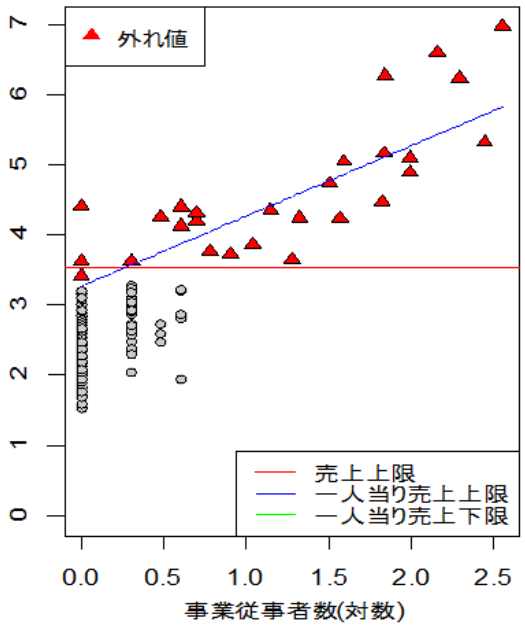


図９：中央値及び四分位範囲に基づくレンジ



レンジとなる境界線として、赤線は売上高の上限、青線は事業従事者１人当たり売上高の上限、そして緑線は事業従事者１人当たり売上高の下限を示している。外れ値とされたデータを赤三角（▲）でプロットしている。事業従事者１人当たり売上高の下限は計算上負の数となって、下限は設定されない。

この二つの結果を比較すると、第３章２節で説明したように、図８に示す①の方法では、外れ値の影響を受けてレンジの幅が広がってしまうので、検出される外れ値が非常に少なくなる。ここで検出された外れ値の数は、４個であった。

一方で、図 9 に示す②の方法では、売上高のレンジ（赤線）が下方に移動して、検出された外れ値は 27 個で、今度はその数が多すぎるように見える。これは、実データが正の値を取る変数で分布の右裾が長い（図 7 左のヒストグラム参照）ため、頑健な統計量を使用しても、対称な分布を前提とした方法によりレンジを定めると、裾の長い側、ここでは上限のレンジを超えるデータが多く出現してしまうからである。

B．②の方法での分布の対称化による効果の検証

ここでは上述Aの結果を踏まえ、同じデータを使用して、分布の対称性を確保するために変数を対数化し、レンジを対数値データから算出した。図 9（再掲）は、レンジの設定に対数化を行っていない図 9 の再掲、その右の図 10 は対数化したデータから得られるレンジを表示している。

図 9（再掲）：中央値及び四分位範囲に基づくレンジ

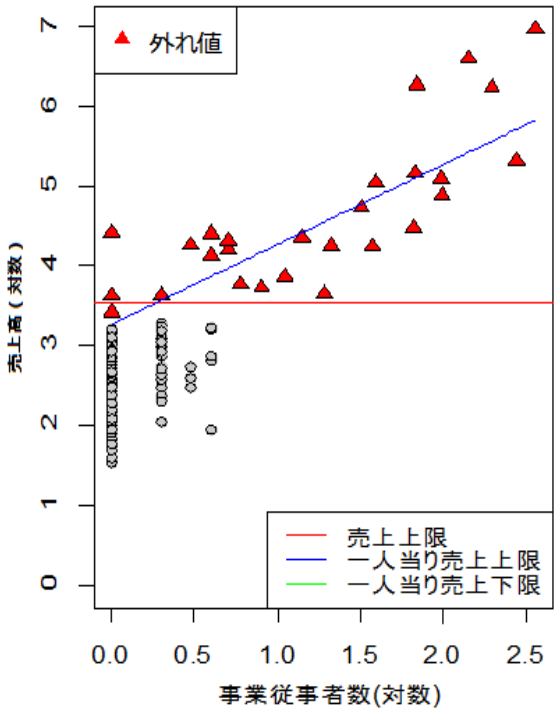


図 10：対数変換後データの中央値及び四分位範囲に基づくレンジ

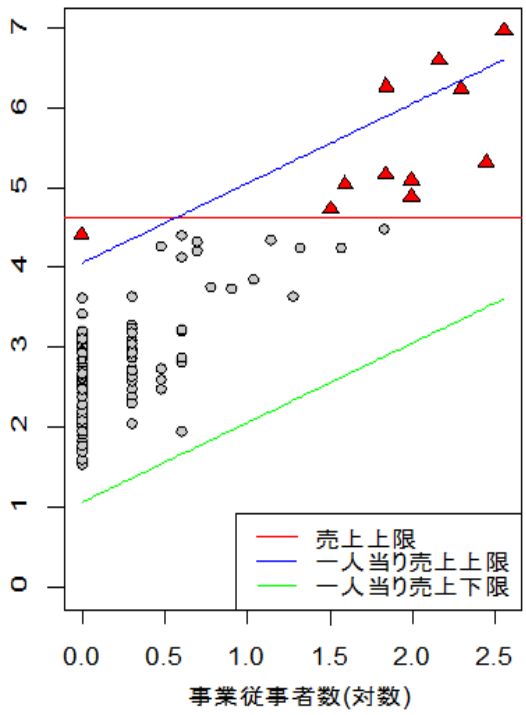


図 10 の緑線は事業従事者 1 人当たり売上高の下側レンジであり、図 9（再掲）の対数変換しないデータに基づくレンジ設定では、下側レンジが負数となりレンジが存在しないが、対数変換を施してレンジを設定した図 10 の場合、下側も対数で得られたレンジを指数変換して実金額に戻せば正の値をとるので、有効な下限レンジを設定することができる。

C．③の方法による分布の歪みの微調整の効果

ここではデータは対数変換した後レンジを算出している。図 10（再掲）は、中央値基準による②の方法のレンジで、図 11 は③の非対称な分布に対応できる方法によるものである。図 11 の点線は、図 10（再掲）の実線と同じレンジであり、レンジを示す直線の位置が全てわずかに上方に移動している。

このデータは、特に事業従事者数が大きいデータについて分布が上方に偏っており、そのために対称分布を前提とするレンジ（中央値基準）では下側よりも上側のデータが検出されやすくなってしまう。一方で、非対称分布に対応できる③の方法の効果により、その偏りが修正された結果、このデータの場合は検出される外れ値は変わらないが、レンジの位置が上方修正され、上側と下側の検出され易さの差異が緩和されることがわかる。

図 10（再掲）：対数変換後データの中央値及び四分位範囲に基づくレンジ

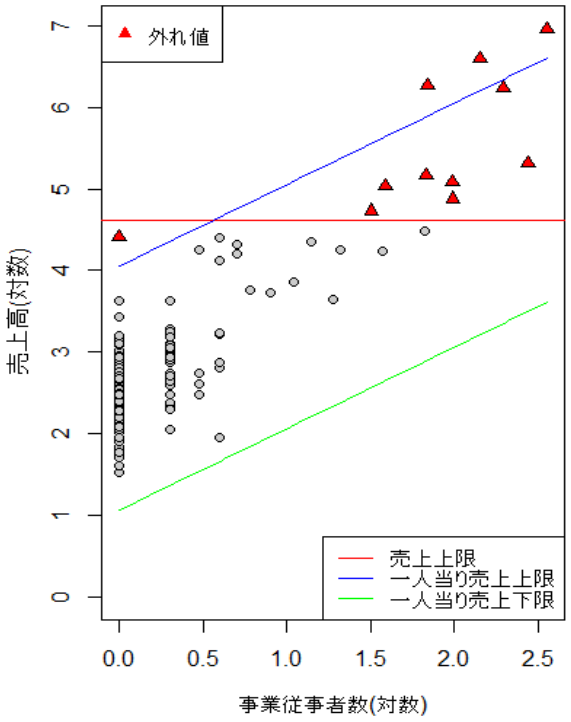
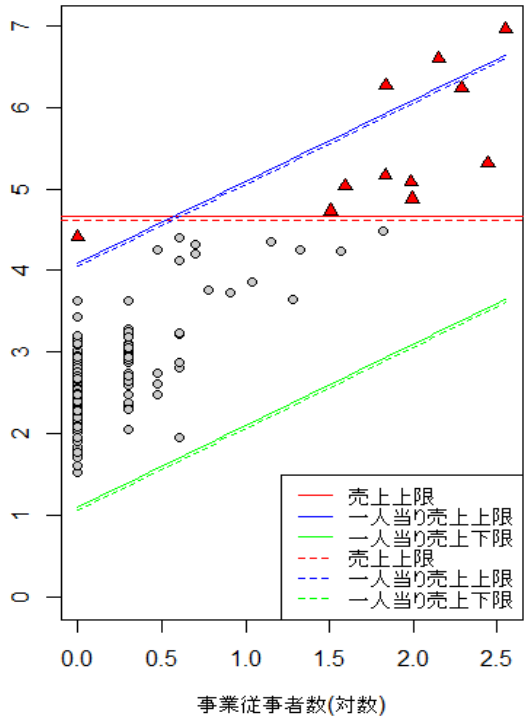


図 11：対数変換後データの四分位値及び四分位範囲に基づくレンジ



4.3 結論及び考察

この調査データでは、産業別に各変数を対数変換した後に③の方法を用いることにより、月次のデータ数が少ない産業を除いて適切なレンジを設定できることが確認できた。

サービス産業動向調査は前述のとおり、毎月の経済動向を把握する調査である。したがって、月内の営業日数や年次イベントなどにより季節変動を有すると考えるのが妥当である。ところが、実際のレンジチェックでは、レンジの算出においても季節性の影響を受けないよう２年（24 か月）分のデータを用いてすべての月に共通のレンジを使用している。このことは季節変動が激しい産業においては、適切とはいえない可能性がある。月毎の分布の変動が著しい場合は、月別にレンジを設定するのが望ましいと考えられる。ただし、月次のデータ数が多くない安定性に欠ける場合は前後の月次データを加えて複数月のデータで算定あるいは複数年の同一月データを用いてレンジを算定する方法も考えられる。1 年分（12 か月分）のデータでは季節変動以外の影響を受ける可能性があるので、今後データが十分に蓄積された時点で検討を行うべきと考える。

5. まとめ

本研究では、単峰であるが正規分布を前提とすることができず、分布に未知の外れ値が混入しているという条件下において、外れ値に頑健な順序統計量を用いて正常値のレンジを設定することで、適切なレンジチェックを行うことができることを示した。金額データなど極端な歪みのあるデータは、対数化などの変数変換を施した後に、さらにわずかな歪みに対応できる Tukey の方法を応用したレンジ設定が有効である。実際の調査データを用いて、この方法が、分布が歪むことの多い経済系の統計調査においても適用が可能であることを示した。

本稿で提案している方法は、サービス産業動向調査を事例として検証を行ったが、この方法は経済系調査に限るものではなく、調査実施者のニーズや調査項目データの分布に応じて四分位範囲に対する適切な倍率を設定することで、幅広い分野の調査に適用することができる。

参考文献

（欧文）

Arnold, B. C. Balakrishnan, N. and Nagaraja, H. N. (2008) "*A First Course in Order Statistics*", SIAM.

Barnett, V. and Lewis T. (1994) "*Outliers in Statistical Data*", 3rd ed., Wiley.

Everitt, B. S. (1998) The Cambridge Dictionary of Statistics, Cambridge University Press; 清水良一訳（2002）「統計科学辞典」，朝倉書店.

Frigge, M., Hoaglin, D.C. and Iglewicz, B. (1989) "Some Implementations of the Boxplot", *The American Statistician*, Vol.43, pp.50-54.

Lloyd, E. H. (1952) "Least-Squares Estimation of Location and Scale Parameters Using Order Statistics" *Biometrika*, Vol. 39, pp.88-95.

Schwertman, N. C., Margaret A. O. and Robiah A. (2003) "A simple more general boxplot method for identifying outliers", *ELSEVIER Computational Statistics and Data Analysis*, Vol.47, pp.165-174.

Tukey, J. W. (1977) "*Exploratory Data Analysis*" Addison-Wesley, Reading, MA.

United Nations Statistical Commission and Economic Commission for Europe (2000) "Glossary of Terms on Statistical Data Editing" United Nations.

Wilcox, R. (2012) "*Introduction to Robust Estimation and Hypothesis Testing*" 3rd ed., ELSEVIA.

（和文）

石川篤史，遠藤峻介，白鳥哲哉（2010）「ビジネスサーベイにおける外れ値対応」，日本銀行ワーキングペーパーシリーズ，日本銀行.

高橋将宜（2012）「諸外国のデータエディティング及び混淆正規分布モデルによる多変量外れ値検出法についての研究」，製表技術参考資料 17，統計センター.

和田かず美（2010）「多変量外れ値の検出～MSD 法とその改良手法について～」，統計研究彙報第 67 号，pp.89-157，総務省統計研修所.

吉澤正（1990）「情報処理入門コース　統計処理」，岩波書店.

吉澤正（2004）「クォリティマネジメント用語辞典」，財団法人日本規格協会

.