あとがき

本書では、データ分析のおおまかなイメージをつかんでいただくため、細かな説明を省略しているところがあります。あとがきに代えまして、本書の内容について少しばかり補足をしておきたいと思います。

第1章で箱ヒゲ図を紹介しました。中央値と四分位範囲という数値を使ってデータのバラツキを検討する方法でした。箱ヒゲ図では箱のフタと底から直線のヒゲがのびていますが、その長さは四分位範囲の1・5倍になっています。この数値について、文太は、七面鳥みたいな名前の統計学者が決めたように話していました。

文太が思い浮かべていたのは、ジョン・テューキー（John Tukey）という統計学者です。七面鳥（turkey）ではありません。ところでテューキーが四分位範囲の1・5倍を超えるデータを外れ値と考えたのは、正規分布との関係からのようです。

テューキーは、データの分布を正規分布と考えた場合に、極端に大きな、あるいは小さなデータの出現する確率を検討する基準として「四分位範囲×1・5」を提案したようです。四分位範囲に1・5をかけた長さのヒゲを箱に足した範囲は、だいたいデータの理論的な範囲の99％強に相当します。この範囲を超えるデータが出てくる可能性は1％もないことになります。したがって、実際のデータにこうした異常に大きい、あるいは小さい数値が含まれている場合、外れ値として検証が必要だと考えられるのです。

第2章においては分散の説明がありました。分散とは平均値を中心としたバラツキを表す数字でした。本文では計算方法を以下のようにまとめています。

- 平均値を求める
- 個々のデータから平均値を引く
- 引き算の結果をそれぞれ自乗する
- それぞれ自乗した結果を合計する
- 合計値をデータの個数で割る

この最後の「データ数で割る」という部分ですが、他の統計入門書では「データ数から1を引いた数で割る」と説明されている場合が多いです。そして特に「不偏分散」と呼ばれています。通常はデータ数で割った分散を利用して構いませんし、またデータの数が多くなると、どちらの分散の値もほとんど同じになります。

2つの分散には使い分けがあります。たとえば母集団全体を調査した結果であれば、データ数で割る分散を使えばいいのです。ただ通常は母集団全体を調べることはできず、その標本から母集団の平均値や分散を推測することになります。

特に標本平均値から母集団の平均値の範囲を推測する場合、理論的には母集団の分散を使う必要がありますが、実際上は標本から求めた分散で代用します。この際には「不偏分散」を使うほうが適切です。標本の分散は、母集団の分散より小さくなることが知られているからです。それを補正して、母集団の分散に近づけるために、データ数からわざわざ1を引いた数で割るのです。

統計ソフトの多くは、分散という場合、無条件に「不偏分散」を計算します。Ｅｘｃｅｌの場合、ＶＡＲ関数が出力するのが不偏分散です。データ数で割った分散のほうはＶＡＲＰ関数で求めます。Ｒというソフトウェアで標準偏差を求めるｓｄ関数でも、分母はデータ数マイナス1となっています。本書の92ページの標準化の説明では、このｓｄ関数を使った計算結果を示していました。

なお分散の分子はバラツキの自乗を合計した数値ですが、これを特に「偏差平方和」といいます。そして分散は偏差平方和をデータ数ないしデータ数から1を引いた値で割った結果です。後者を特

に不偏分散というわけです。また、標準偏差は分散の平方根ですが、こちらも多くの統計ソフトでは不偏分散の平方根が出力されます。

第2章では確率分布が取り上げられました。そこでは65ページでコイン投げの確率のグラフを、また60ページで正規分布のグラフを紹介しました。

コイン投げの確率グラフではY軸が確率を表しています。表の枚数が0から10枚までの11通りあり、11個の確率を合計すると1、つまり100％となります。

一方、60ページ上の正規分布のグラフは、平均値が0で標準偏差が1の分布を表しています。これを特に「標準正規分布」と呼びます。ところで、このY軸は確率ではありません。離散値とは違い、連続量の場合、X軸に無数の数値が集まっています。

たとえば筆者の体重は厳密にはかるならば65.132783623198748…となるかもしれません。このような数値が無数にX軸に密集しているのが連続量の特徴です。したがって、ある特定の数値が生じる確率が仮に0.00000000000000001だとしても、このような数値が無限にありますから、確率の合計は1を超えてしまうことになります。

そこで連続量の場合、特定の数値1点に確率を対応させるのではなく、X軸のある数値から別の数値までの範囲を確率と考えます。つまり面積が確率になります。個別の点に対応する確率は0とみなします。

たとえば標準正規分布では－1.96から1.96の範囲で曲線の下の面積が0.95となります。X軸が

−1.96の位置で山の形をした曲線と交差する位置をY軸で確認すると約0.058となります。くどいようですが、連続量では個別の点に対応する確率は0です。これは確率そのものではありません。そこで区別するため、Y軸の値に相当する数値を「確率密度」と呼んでいます。

第3章では相関について述べられています。気温が高くなるとビールの消費量が増えるという例がありました。

ところで、これは因果関係でしょうか？　気温が高くなるからビールの消費量が増えるのでしょうか？　一見するとアタリマエの関係かもしれませんが、こう考えるとどうでしょう。「気温が高くなると喉が乾くからビールを飲みたくなる」と。さらにいうと、喉が乾いたら水を飲めばいいだけなので、あえてビールを飲む必要はありません。むしろ飲みすぎると、かえって逆効果でしょう。

何をいいたいのかというと、気温とビールの消費量が相関していることは間違いない事実ですが、これが原因と結果になっているとは即断できない、ということです。子供の身長が高くなるほど、覚えている漢字の数が増えているからといって、身長と漢字に因果関係があるわけではありません。子供が成長して学年が進むほど、習う漢字が増えているだけの話です。相関係数が高いからといって、それが因果関係の証明にはならないということは忘れないようにしてください。

第6章でカイ二乗検定についての解説があります。そこでは時間帯や年齢ごとに、2種類のお弁当の売上に違いがあるかどうかを分析する方法が紹介されています。その際、データは分割表にまとめられていますが、じつは分割表の中に5未満の数値が1つでも含まれている場合、検定の結果

が不正確になることがあります。そのため、データ分析ソフトの出力に警告が表示されることがあります。詳細は省きますが、このようなときは、カイ二乗検定に代えて、フィッシャーの正確確率という方法を検討すると良いでしょう。

なお2行2列の分割表の場合、やはり統計ソフトによってはイェーツの補正が行われるため、本書で掲載したカイ自乗値とは少し異なる結果が出力されます。カイ自乗検定では、分割表から求めた数値に、カイ自乗分布という確率分布をあてはめます。ところが分割表の数値が離散値（1とか2、3といった整数）であるのに対して、カイ自乗分布は連続量（小数点を含む数値）の分布であるため、ズレが生じます。このズレを修正するのが、イェーツの連続性補正です。イェーツの補正の必要性については議論のあるところですが、いずれにせよ、統計ソフトウェアの出力が補正された数値なのかは確認が必要です。

なお本書で掲載しているデータの数値計算やグラフ作成にはRというソフトウェアを利用しました。Rは無料ながら非常に高機能な統計解析ソフトウェアです。そこで筆者は、本書に登場する統計量やグラフをRで確認できるソフトウェア（パッケージ）を作成し、サイトで公開しています。

読者に余裕があれば http://rmecab.jp/ranko にアクセスいただき、サイトの説明にしたがって本書の内容をもう一度おさらいしていただくとうれしく思います。

最後に本書は、共立出版(株)の稲沢会さんから「読みやすい統計入門書を」という依頼を受けて執筆を始めました。しかし「読みやすい」というベクトル（方向）が稲沢さん（そして共立出版さ

ん）の本来の意図とは異なる「専門書」になってしまったようです。小説風に仕上げていますが、筆者は小説を執筆した経験はありません。ただ、物語風に説明することで、読者がデータ分析の目的や方法をイメージしやすくなるのではないかと考えました。文章の未熟な点は、りんと氏によるイラストが援護してくれたのではないかと思います。

さらに編集をご担当くださった赤城圭氏をはじめ、共立出版の編集部の皆さん、そして南條光章社長からも、さまざまな助言をいただくことができました。また最後まで読んでくださった読者の方々に、この場を借りまして、深く御礼申し上げます。

2013年7月

石田基広