2000-05-19
SQC 部会資料 No.2000-05-1

データ解析の手順について（変数変換の適用基準：試案）

稲葉太一（神戸大学）

はじめに

先月の SQC 部会で、正規分布からずれているデータの解析方法について、ご質問がありました。それに対するコメントを兼ねて、資料を作成してみました。

A. データ解析の手順

A.1. データの分類と従う分布の特定

1) 計量値データの場合
（正規分布かどうか。寿命データならば、指数分布やワイブル分布かどうか。）
2) 計数値データの場合
（二項分布かどうか。ポアソン分布かどうか。）

A.2. パラメトリック法を用いる場合

上記の分布のいずれかに従うならば、パラメトリック法（今まで習った手法）で解析するとよい。

A.3. 変数変換後、パラメトリック法を用いる場合

もし、どの分布にも従わない場合は、以下の変数変換を試みてパラメトリック法での解析に持ち込む。東京理科大学の芳賀敏郎先生が、このアイデアを下さった。

具体的には、Box-Cox 変換（B.3. 参照）のうち、

1) $\lambda = 0$ （対数変換）
2) $\lambda = 1/3$ （３乗根変換）
3) $\lambda = 1/2$ （平方根変換）
4) $\lambda = 2/3$ （３分の２乗根変換）
5) $\lambda = 1$ （そのまま変換せず）

の５つの場合を主たる候補として、変数変換を行い、正規性の検定（例えば、歪度と尖度）の結果を参考に変換を選ぶことを推薦したい。当初、この Box-Cox 変換は、データを最も正規分布に近づける λ を採用するように推奨された。ところが、このような方法だとデータに合わせ過ぎて却ってやり過ぎの変換を選んでしまうこととなり、広く利用者の賛同を得られなかった。上記の５つに限定することで、合わせ過ぎることを無くし、本当に必要なときにだけ変換するように考えた。

A.4. ノンパラメトリック法を用いる場合

いずれの変換でもうまく行かない場合には、ノンパラメトリック法の手法を用いる。

ノンパラメトリックな手法とは、データを順位に変換して、その情報だけから検定する「順位和検定」や、データの大小（勝敗）だけで検定する「符号検定」などがある。

A.5. 変数変換について

変数変換を考えるとき、以下の分散安定化変換（B を参照）が良い場合が多い。

B. 変数変換について（1999-9-2 の若干の改訂版）

変数変換を考えるとき、以下の分散安定化変換が良い場合が多い。

B.1. 分散の安定化変換とは

母平均 μ と、母分散 σ^2 に対して、ある変換 $h(\cdot)$ が

$$[h'(\mu)]^2\sigma^2 = const.$$

を満たすとき、その変換 $h(\mu)$ を、分散安定化変換 という。

B.2. なぜ、分散安定化変換が望ましい変換なのか？

この変換を施すと、変換後の平均 $h(\overline{X})$ の母分散が、推定したい母数 μ に（近似的に）無関係になる。このことは、母平均 μ を区間推定する際の区間幅が一定となり、区間推定の意味がはっきりするという大きなメリットがある。

（例えば、2 項分布の母比率 P を考える。標本比率 $p = \frac{X}{n}$ の母分散は $\frac{P(1-P)}{n}$ であり、P によって変化する。このまま、区間推定を行うと、単純な方法では、信頼区間の両端の P で考えたときの分散と、区間推定をしたときの p が違うため、不都合が生じる。これを是正するのが、分散安定化変換である。）

B.3. Box-Cox 変換

$$y^{(\lambda)} = \begin{cases} (y^\lambda - 1)/\lambda & (\lambda \neq 0) \\ \log y & (\lambda = 0) \end{cases}$$

この変換は、σ と $\mu^{1-\lambda}$ が比例するときは、分散安定化変換である。

また、$\lambda(\neq 0)$ を 0 に近づけると、$(y^\lambda - 1)/\lambda$ が $\log y$ に近づく。

B.4. 分散安定化変換の例

例１。データがポアソン分布に従うならば、平方根変換 が分散安定化変換である。

例２。また二項分布ならば 逆正弦変換 である。

例３。標準偏差が、平均値に比例している時は 対数変換 が分散安定化変換。

例４。標準偏差が、平均値の平方根に比例している時は 平方根変換 。

（例３、４の変換は、2. の Box-Cox 変換 の特別な場合と考えることができる。）

なお、二項分布の ロジット変換 は、逆正弦変換と良く似た性質を持っていて、分散を安定化する傾向がある。

B.5. 分散の安定化変換の仕組み（一般論）

標本平均 $\overline{X}$ の変換を行って、それの分散が母平均によらないようにしたい。

$$h(\overline{X}) = h(\mu) + (\overline{X} - \mu)h'(\mu) + R_2$$
$$E\{h(\overline{X})\} = h(\mu) + E(R_2)$$
$$V\{h(\overline{X})\} = [h'(\mu)]^2 E\{(\overline{X} - \mu)^2\} + \cdots = [h'(\mu)]^2 \frac{\sigma^2}{n} + \cdots = const.$$

このような、$V\{h(\overline{X})\}$ が母平均 μ によらないためには、一定値となればよい。そこで、上に挙げた

$$[h'(\mu)]^2\sigma^2 = const.$$

という式を満たす変換 $h(\mu)$ を求めればよいことが分かる。