

「四分位数」に関する 2, 3 の知見

〔はじめに〕

(社)日本分析学会主催、(社)日本環境測定分析協会共催で行っている環境分析分野の技能試験においては、各試験所からの報告値の「ロバストな統計計算手法」として、「四分位数法」を採用しております(クォータリー No.48,49,50 参照)

その手順について簡単に復習しますと、

- (1) 全報告値を小さいものから順(昇順)に並べる。
- (2) 小さい方からの順位が 1/4, 2/4, 3/4 の試験所の報告値を、それぞれ、第 1 四分位数 (Q_1 : 25 百分位数)、第 2 四分位数 (Q_2 : 50 百分位数)、第 3 四分位数 (Q_3 : 75 百分位数) とする。

(* : パーセンタイルとも呼ぶ)

- (3) 求められた Q_1 , Q_3 を 式に代入して正規四分位数範囲(ロバストな標準偏差(S))を求める。

$$s = (Q_3 - Q_1) \times 0.7413 \quad \dots\dots$$

- (4) 試験所(i)の報告値 (X_i) について、式により当該試験所の z スコア (z_i) を求める。

$$z_i = \frac{X_i - Q_2}{s} \quad \dots\dots$$

- (5) $|z_i|$ について評価を行う。

となります。

ここで、四分位数は、全報告値が N 個の場合、それぞれ次の順位に該当する報告値とされており、

$$\text{第 1 四分位数 } (Q_1) = \{(N-1) \times 1/4\} + 1 \dots\dots$$

$$\text{第 2 四分位数 } (Q_2) = \{(N-1) \times 2/4\} + 1 \dots\dots$$

$$\text{第 3 四分位数 } (Q_3) = \{(N-1) \times 3/4\} + 1 \dots\dots$$

具体的に $N = 101$ の場合で計算すると、 $Q_1 = 26$ 、 $Q_2 = 51$ 、 $Q_3 = 76$ 番目の報告値になります。ここで、単純に $101 \times 1/4 = 25.25$ 、 $101 \times 2/4 = 50.5$ 、 $101 \times 3/4 = 75.75$ 番目としないうのは何か? という疑問が生じます。

今まで、この点に関しましては、あまり深く考えずに「四分位数法」を使っておりましたので、ここで原点に戻って、この式の意味を考えることにします。

1. 13 本の杉の木の高さの四分位数

数が多いと判りにくいので、13 本の杉の木の場合について、その四分位数がどうなっているかを考えてみることにしましょう。

まず、杉の木を低い方から高さの順に等間隔に植えかえて、その木の天辺を紐で結びます(図 1)。

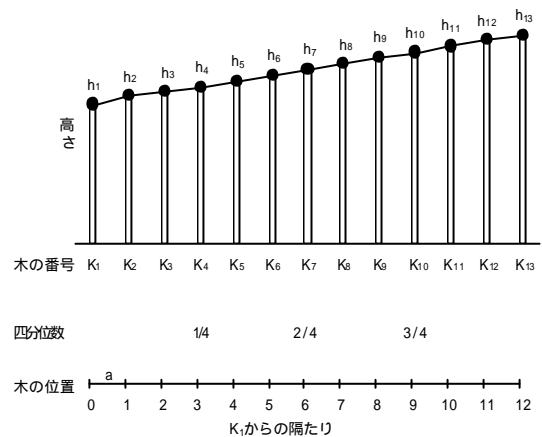


図 1 13 本の杉の木とその四分位数(1)

この 13 本の杉の木について、上記の式を適用して四分位数を求めると、

$$Q_1 = 4 \text{ 番目の木 } (K_4) \text{ の高さ} : h_4$$

$$Q_2 = 7 \text{ 番目の木 } (K_7) \text{ の高さ} : h_7$$

$$Q_3 = 10 \text{ 番目の木 } (K_{10}) \text{ の高さ} : h_{10}$$

に該当することになります。

今、この木の位置を黒く塗り潰してみると K_7 が 13 本の木の中央ということは判りますが、 K_4 が 4 分の 1 番目、 K_{10} が 4 分の 3 番目というのがピンときません。

そこで今度は、両端の木を黒く塗ってみましょう(図2)。



図2 13本の杉の木とその四分位数(2)

このようにしてみると、 K_4 が4分の1の位置の木、 K_{10} が4分の3の位置の木ということが判ると思います。

ここでもう一度図2を見てみると、1/4、2/4、3/4と呼んでいるのは、 K_1 から K_{13} までの木の間隔(12a)を、それぞれの割合に分割した位置であることが判ります。更に、木の位置(K_1 からの隔たり) = 木の番号 - 1 となっていることが判ります。

2. 四分位数の計算式の意味

以上、1で判ったことを踏まえて、各四分位数を求める式(式)の意味を考えてみましょう。

$$Q_r \text{に該当する木の番号} = \{(N-1) \times r/4\} + 1 \dots\dots$$

(rは1,2又は3のいずれか)

N: 全体の木の数

$N-1$: K_1 から K_{13} までの隔たりの数

$r/4$: 求める位置までの隔たりの割合

$(N-1) \times r/4$: 求める位置までの隔たりの数

$\{(N-1) \times r/4\} + 1$: 求める位置の木の番号

従って、式から求めた番号の木の高さが Q_r に該当する高さということになります。

なお、式は、四分位数の場合です。百分位数(パーセントイル)の場合は、

$$(\text{パーセントイル}) = \{(N-1) \times r'/100\} + 1 \dots\dots$$

($r' = 1 \sim 100$ のいずれかの値)

の順位の値となりますので、

第1四分位数 = 25 百分位数

第2四分位数 = 50 百分位数

第3四分位数 = 75 百分位数

であることがわかると思います。(単純に割合を乗ずる

のは誤りです)

3. Q_1, Q_2, Q_3 に小数部分が出る場合

報告値の総数(N)が「4の倍数+1」の場合は、各四分位数に該当する報告値が存在しますが、それ以外は、式の計算結果に小数部分が出ます。このような場合は、その値に近接する報告値の間を、その小数部分の割合に補間して求めます。図1の例でいえば、該当する木の間に張り渡した紐までの高さが、その位置における「求める値」となるわけです。

4. ISO/IEC 13582 7.3 記載の百分率順位について

たまたま、ISO/IEC 13582をばらばらとめくっていたところ、その7.3にpercentage rank(百分率順位)というのが載っていました。その計算式は、「順位を $i = 1, 2, 3 \dots p$ 」とすると、百分率順位は、

$$(\text{百分率順位}) = 100 \times (i - 1/2) / p \dots\dots$$

として計算される、と記載されております。

これを乃至式の形及び同じ記号で書き直しますと、

$$i \text{ (該当する報告値の順位)} = N \times r'/100 + 0.5 \dots\dots$$

(r' : 百分率順位、 $N = P$ 、 i : 報告値の順位)

となり、式と異なることが判ります。

この式に $r' = 0$ を入れると $i = 0.5$ となり、 $r' = 100$ を入れると $i = N + 0.5$ になりますので、この式は、0.5から $N + 0.5$ までをベースにして100等分していることになりました。これに対して、式は1からNまでを100等分していることになりました。

式の意味はよく判りません。経済的な統計量として、何かの指標に使われているのではないかと思います。ご存知の方がおられましたらお教え下さい。(なお、式の場合、 $i = 0.5$ 又は $i = N + 0.5$ に該当する値をどのようにして算出したらよいのでしょうか。 $i = 1$ と2の値、または $i = N - 1$ とNの値から外挿して計算するのもおかしいと思いますが...))

〔おわりに〕

以上、四分位数についての検討結果をまとめました。参考にして戴ければと存じます。