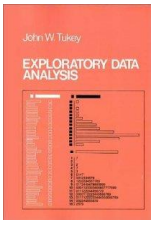
雑感　箱ひげ図の考案者とヒンジ

■　雑感 97「四分位数の誕生と考案者」の中で，「なるほど統計学園高等部」に，四分位数の誕生時期を尋ねたことを載せたが，実は，同時に「箱ひげ図はいつころ登場したのでしょう？誰が発案したのかも分かればいいなあと思います」と教えを請うた．

その結果，「J.W.Tukey（ジョン・テューキー　1915－2000）が，1977 年の著書「Exploratory　Data　Analysis」で箱ひげ図を導入したといわれています．参考：
http://ja.wikipedia.org/wiki/%E3%82%B8%E3%83%A7%E3%83%B3%E3%83%BB%E3%83%86%E3%83%A5%E3%83%BC%E3%82%AD%E3%83%BC」
との返事をいただいた．



■　その本を近くの図書館を通じて，某大学図書館から相互貸借で借りだした．500 ページもある大作である．苦手な英語でも，最近は翻訳サイトもあるから，その助けを借りて何とか概略を読み取れないかという企てである．

■　第 2 章　Easy summaries—numerical and graphical　の中の 39 ページ **2C. Box-and-whisker plots** に箱ひげ図が載っている．その最初の文章を載せると次のようである．

We always want to look at our results. We usually can. In particular,　we want to look at 5-number summaries. Exhibit 5 shows an easy way to do just this. We draw a long,　thinnish box that stretches from hinge to hinge,　crossing it with a bar at the median. Then we draw a "whisker" from each end of the box to the corresponding extreme.

Google 翻訳を元に，若干の手直しをすると，

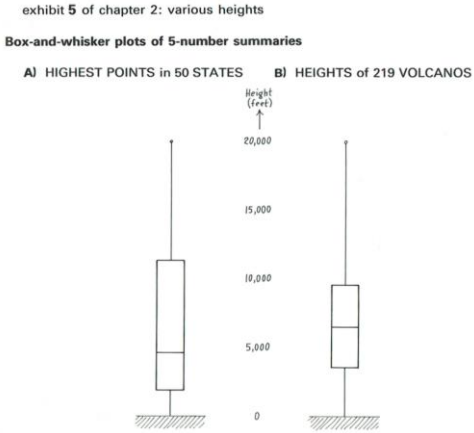
我々は常に我々の結果を見てみたい．我々は通常することができます．特に，我々は，5 数要約を見てみたい．図表 5 は，これを行うための簡単な方法を示しています．我々は，中央値のバーでそれを横断し，ヒンジ，ヒンジから伸びる長い，痩せ気味のボックスを描画します．その後，我々は，対応する最大値と最小値にボックスの各端から"ひげ"を描きます．

若干の語釈を付け加える．

* hinge…ここでは，第 1，第 3 四分位数 Q_1，Q_3 のこと

* extremes…the highest and lowest values

■　そして，これが Tukey による「最初の」箱ひげ図である．



扱われているデータは，A)は 50 州の最高標高，B)は 219 の火山の標高であろうか．

手書きであるところが，1977 年を彷彿とさせる．

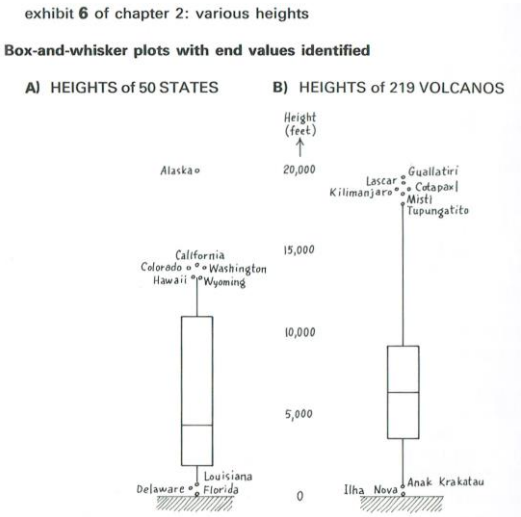
■　そして，箱ひげ図の項はさらに続く．

This process shows us the five-number summaries quite clearly, so clearly as to give us a clear idea of (some of) what we may have been missing. There is , inevitably more empty space in box-and-whisker plot than in a listing of a 5-number summary. There is more space for

identification. We can at least identify the extreme values, and might do well to identify a few more.

このプロセスは、私たちが見失っていたかもしれないものの明確なアイデアを（いくつかの）を得たので、はっきりと、非常にはっきりと私たち 5 数要約を示しています。5 数要約のリストよりも箱ひげプロットで必然的に複数の空のスペースがあります。識別のためのより多くのスペースがあります。我々は、少なくとも、極端な値を識別することができ、さらにいくつかを識別するためによくやるかもしれません。

う～む，よく分からないなあ．そこに載せられた図は，次の通りである．



これから察するに，いわゆる「はずれ値 outlier」のことを述べているのだろうと思われる．

なお，はずれ値に関して本書は，**2D.Fences, and outside values** に書いている．ここでは詳述しないが，四分位範囲の 1.5 倍を超えるデータをはずれ値とするアイデア(というか，1 つの見解というか)をすでに載せている．

■ さて，もう 1 つ，hinge のことについて触れよう．

統計用語では，次のように定義されるようである．

・下側ヒンジ(lower hinge)：中央値以下のデータの中央値
・上側ヒンジ(upper hinge)：中央値以上のデータの中央値
・ 5 数要約(five number summary)　最小値，下側ヒンジ，中央値，上側ヒンジ，最大値

つまり，下側ヒンジ＝Q_1，上側ヒンジ＝Q_3 である．

しかし，hinge とはそもそも蝶番のことである．なぜ，四分位数が蝶番なのか．

本書を見ていて，それが明確に分かったのである．

2B.Hinges and 5-number summaries
（本書 33 ページ）の記述を見よう．

If we have 9 values in all, the 5th from either end will be the median, since $\frac{1}{2}(1+9)=5$. Since $\frac{1}{2}(1+5)=3$, the third from either end will be a hinge. If we have 13 values, the 7th will be the median—and the 4th from each end a hinge. In folded form, a particular set of 13 values appears as follows:

−3.2　　1.5　　9.8
−1.7　　1.2　　1.8　　6.4
−0.4　　0.3　　2.4　　4.3
0.1　　3.0

The five summary numbers are, in order, −3.2, 0.1, 1.5, 3.0, and 9.8, at each folding point.

つまり，上のようにデータを順に 4 グループに分けて W のように並べたとき，その折れ曲がるところに現れる値がヒンジなのである．W は 2 つの蝶番を横から見た図ではないか．