盛岡三高数学科通信 How do you solve ? How do you teach ?

第 17 号
発行責任者
盛岡第三高等学校
下町壽男

続・箱ひげ図には気をつけろ

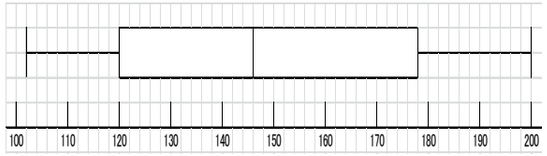
今回は，箱ひげ図の「良さ」と「問題点」について考えていきたいと思います．

【箱ひげ図の良さ】

箱ひげ図を用いる良さとして次の３つの視点から例示しておきましょう．

＜例１＞（分布の傾向を見る）

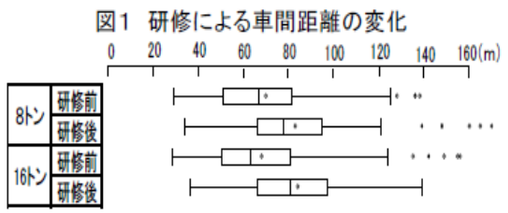
下の箱ひげ図は，ある店における１ヶ月間（30 日）の来客数のデータを表したものです．



（埼玉県教育委員会 HP より）

四分位数によって示される「箱」「ひげ」４つの領域に含まれる変量の割合がすべて 25%であると考えれば，例えば，来客数が 150 人以下の日は 15 日以上であること，180 人以上の来客数の日と 120 人以下の日がどちらもほぼ一カ月の１／４であること，一か月の半分は 120 人~180 人の来客数であったことなどが見て取ることができます．

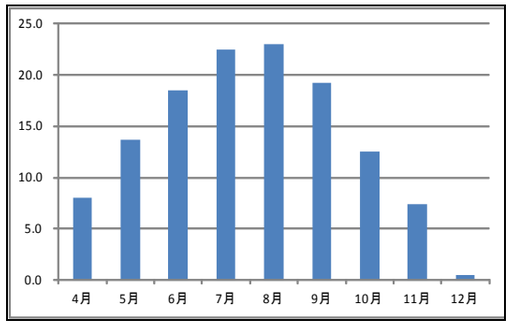
＜例２＞（分布の推移を見る）



（自動車安全運転センター（調査研究部）より）

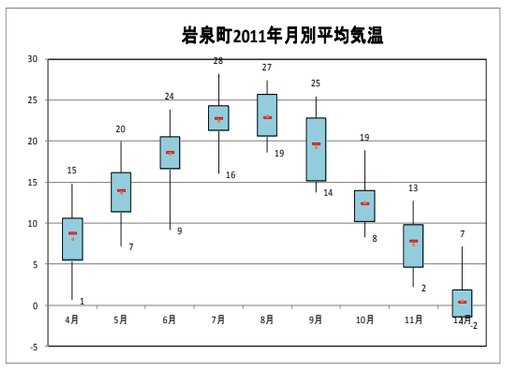
上の箱ひげ図は，トラック運転者教育についての実証実験の結果です．箱ひげ図の推移から，研修後に８トントラックも 16 トントラックも走行中の車間距離が大きくなるように推移していることがわかります．平均値の比較だけではく，箱ひげ図を用いることによって，研修による改善効果が大きいことの説得力ある検証データとして扱われています．

＜例３＞（経年変化などについて多次元的表現ができる）



上図は岩泉町の 2011 年の４月から 12 月までの月別平均気温を棒グラフにしたものです．

しかし，棒グラフで示される値は，それぞれの月に含まれる 30 日程度の日数の代表値に過ぎません．



そこで，平均気温に併せて，各月の状態を箱ひげ図で表してみます．例えば７月と８月の状態や，６月と９月を比較したとき，平均気温だけでは違いがわかりませんが，箱ひげ図にすると違いをイメージすることができます．

【箱ひげ図の問題点・限界・内包する矛盾】

箱ひげ図の問題点として，前回の内容と重なる部分もありますが，以下の点をあげておきます．

＜作成についての問題点＞

・箱ひげ図はシンプルな図であるが，データのソーティングが前提なので，作成はシンプルではない．（ヒストグラムを作った方が早いし分布の性質をよく表す．それに，データをソーティングすれば，更に箱ひげ図を作ることもない）

・コンピュータで作る方法もあるが，エクセルで用意されているパーセンタイル関数は文科省が定義する手法と異なる．

・文科省の定義する四分位数は，データの個数が 型によって算出法が異なるため，箱ひげ図の作成が，資料の傾向や性質を知ることよりも，データの個数のタイプに応じて四分位数を求める技能に矮小化されることが予想される．

＜箱ひげ図自体の問題点＞

・５つの要約統計量で決定される，「箱」，「ひげ」の４つの領域内の変量の分布はどうなっているかわからない．

・ヒストグラムから箱ひげ図は類推されるが，箱ひげ図からヒストグラムを（一意に）再現することができないため，間違った解釈をする可能性がある．

・特に，少ないデータの場合，変量の個数によって中央値のとり方が異なるので，変量が１個違うだけで，全く異なる箱ひげ図ができあがる恐れがある．

・データによっては，四分位範囲が大きいほど散らばり具合が大きいとは判断できない．

ひとつの例として，何森仁氏（神奈川大学）の提唱している例から引用しておきましょう（右図）．

19 個の変量からなる４つのデータ ABCD において，A と C では A の分散が大きいのですが，実際に箱ひげ図を作ると，C の四分位範囲が大きいことがわかります．

つまり，分散という散布度に従えば，A の方が散らばり具合が大きいデータなのに，四分位偏差に従えば，C の方が散らばり具合が大きいデータと判断されてしまいます．

また，A と B，A と D は全くことなる分布であるが，箱ひげ図にすると同じものに表されます．

逆に，C と B，C と D は視覚的に見て，類似の分布ですが，異なる箱ひげ図になっています．

順番	5数要約値	データ			
		A	B	C	D
1	最小値	0	0	0	0
2		0	20	20	20
3		0	20	20	20
4		0	20	20	20
5	第１四分位数Q_1	40	40	30	40
6		40	50	50	50
7		40	50	50	50
8		40	50	50	50
9		40	50	50	50
10	中央値Q_2	50	50	50	50
11		50	50	50	50
12		70	50	50	50
13		70	50	50	50
14		70	50	50	60
15	第3四分位数Q_3	70	70	70	70
16		90	70	70	70
17		90	80	70	80
18		90	80	100	80
19	最大値	100	100	100	100
合計		950	950	950	960
平均		50.0	50.0	50.0	50.5
分散		1021	547	621	552
標準偏差		31.953	23.388	24.92	23.495