発表レポート

2011-10-10

箱ひげ図について

新「数学Ⅰ」統計分野に於ける「R言語」支援による教材作成の一例として

立命館宇治中学校・高等学校
数学科　稲葉芳成

1　はじめに

新「数学Ⅰ」の統計分野「データの分析」は従来の「数学B」の「統計とコンピュータ」の内容と一部重複する部分があるものの、これまで扱われなかった、四分位数や箱ひげ図などが新たに登場している。背景として、平成17年に出された日本学術会議統計学研究連絡委員会報告書「知識創造社会に向けた統計教育の推進について」に於いて学習モデルのとして箱ひげ図や幹葉プロットが取り上げられている。

ここでは、箱ひげ図を中心にとりあげることにする。また授業の教材作成の援用として「R言語」（または単に「R」）を用いての実例も紹介する。

2　要約統計量とは

要約統計量とは標本の分布の特徴を代表的に表す統計学上の値のことであり、いくつかの値がある。またこれは記述統計量、基本統計量、代表値とも呼ばれる。要約統計量の例としては、平均値、期待値、分散、歪度、尖度、中央値、四分位点、最小値、最大値、最頻値などがある。特に平均値についてはさらに細かく、相加平均、相乗平均などに細分化される。

標本の分布の特徴を上側、下側のヒンジ（四分位より上下の値のそれぞれの中央値）を含む最大値、最小値、中央値、上側・下側ヒンジの5つの値、つまり、0, 約 0.25, 0.5, 約 0.75, 1 分位数で要約することを五数要約という。これを図示したものが箱ひげ図であるとも言える。

ここで、若干の注意を要する。まず、最小値や最大値については、例えば一般の箱ひげ図では外れ値を除いた範囲でのものであるが、外れ値を考慮しない方式やあるいは外れ値の境界をどの程度にとるかによってそれぞれ値が変わってくる。

また、25%や 75%のパーセンタイル値（所謂、四分位点のうちそれぞれ Q1, Q3 と書く）についても厳密な 25%, 75%のパーセンタイル値を採用するか、上下のヒンジを採用するかは書籍や解析ソフトなどによっても異なる。いくつかの求め方があり、そのうちのどれを採用するかで求める値に若干の違いが出ることになる。

ちなみに東京書籍教科書では中央値を含まない下側と上側のデータのそれぞれの中央値を第1四分位点、第3四分位点としている。つまり上下のヒンジを採用しているようである。

3　箱ひげ図の概要と箱ひげ図について

(1) 箱ひげ図の作り方

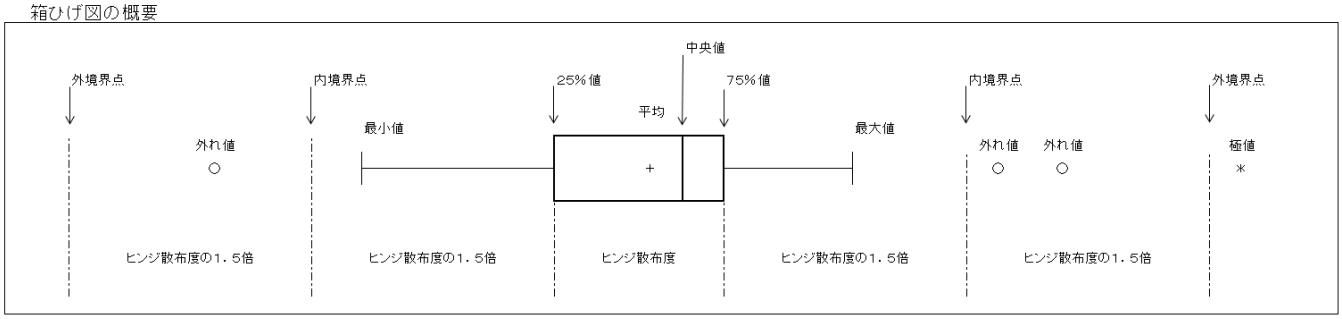
あらためて、箱ひげ図とは、標本の分布の様子を視覚的に表現するためのもので、長方形の箱の両側に延びる「ひげ」で構成される。ここでは、米国の統計学者ジョン・テューキーによる方式を説明する。この方式は最もポピュラーなものであるらしいが、後述するように箱ひげ図の作り方は外れ値の扱いなど重要なところで一通りではない。

（従って、以下の記述について若干不正確な部分があるかも知れないがお許し頂きたい）

はじめに五数要約の値をあらかじめ求めておく。以下次のようにする。

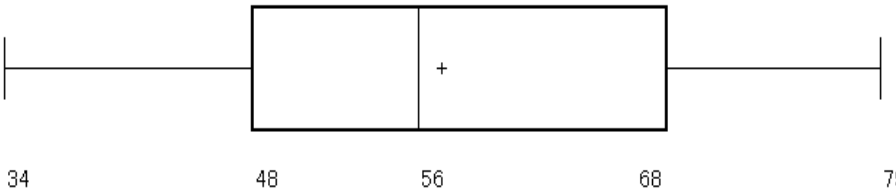
① データの第1四分位点 Q1 と第3四分位点 Q3 により、全データの半数が含まれる箱を描く。
② 中央値 Q2 を縦線で描く。
③ 平均値を「+」で描く（省略されることもあり）。
④ 箱の長さ（これをヒンジ散布度という）の 1.5 倍を箱の左右にとり（内境界）、それを超えない内側のデータの最大値と最小値まで「ひげ」を引く（内側すべてに「ひげ」を引く方法もある）。
⑤ 内境界点の左右にヒンジ散布度の 1.5 倍の長さをとり（外境界）、その範囲にあるデータをはずれ値として「○」でプロットする（最小値と最大値まで「ひげ」を引く方法ではこれは描かない）。
⑥ 内境界点の外側にあるデータを極値（極外値）として「＊」でプロットする。

(2) 一般的な箱ひげ図の概要（スケッチ）



(3) 箱ひげ図の実際（例）

ここで仮にデータを用意して箱ひげ図のひとつの例を見ておこう。いまデータとして 15,34,38,42,43,48,50,51,51,53,56,59,64,65,67,68,71,72,72,79,99 を用意する。ここでは「R」と呼ばれるソフト（後述）を用いて簡単に箱ひげ図を作成することにする。実際に、授業でも生徒に特に手作業で箱ひげ図を書かせる時間がとれるかどうかは疑わしいところである。そこでこの「R」であるが、これによると、簡単な操作で五数要約の値が得られる。いまここでは中央値 56、Q1　48、Q3　68、平均値 57、最小値 15、最大値 99 である。この値までわかれば手作業でも箱ひげ図はかける。Q3-Q1=20 であるから、ヒンジ散布度 20 の 1.5 倍の 30 を箱の両側にとると、内境界の範囲は 18 ～ 98 ということになる。この範囲での最小値は 34、最大値は 79 であるからこの範囲に「ひげ」をとると、以下のような箱ひげ図ができあがる。

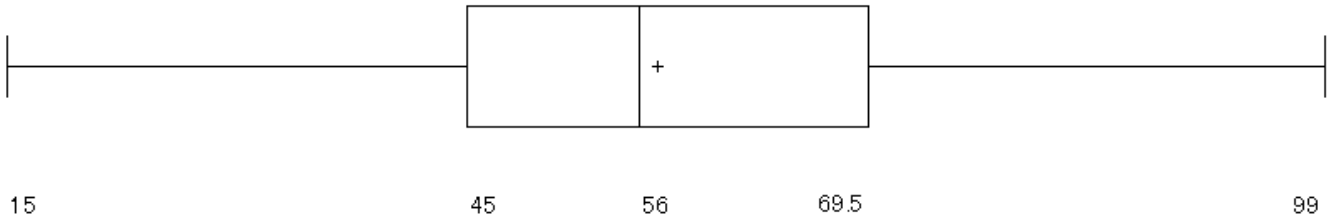


ただし、ここでは外れ値については省略してある。

(4) 箱ひげ図の作り方（教科書）

教科書ではまず、四分位数の求め方に若干の注意を要する。教科書では、まず中央値を求め、その後、その中央値を除いた下側データと、上側データのそれぞれについての中央値を求め、それらを Q1、Q3 としている。さらに、最大値や最小値についても外れ値を考慮しないものとしている。

例として先のデータ 15,34,38,42,43,48,50,51,51,53,56,59,64,65,67,68,71,72,72,79,99 を用いると、教科書では、これの中央値は 56 であるが、Q1、Q3 についてはそれぞれ、45 と 69.5 となる。
また、最小値は 15、最大値は 99 であるので、箱ひげ図はひげが横に長くなる形となる。



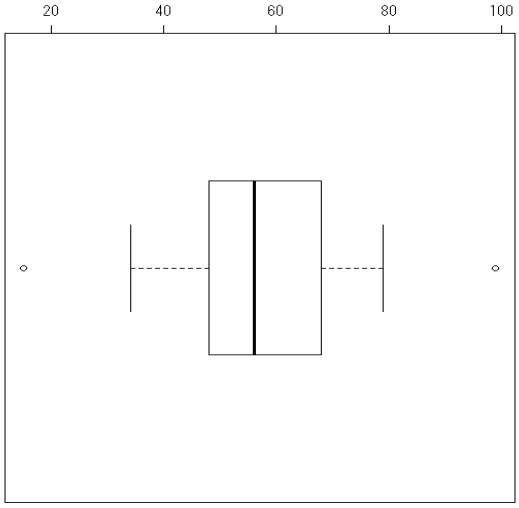
教科書がこのような方式を採用しているのは、外れ値の扱いを省くことで、できるだけ簡便なものをという配慮と、あえて外れ値を含むデータの分布の様子に視点を据えているためだと思われるが本当かどうかは著者に確認が必要かも知れない。

(5) 「R 言語」について

「R 言語」は「R」とも呼ばれるオープンソースでフリーソフトウェアの統計解析向けプログラミング言語であり、またその開発実行環境でもある。データの互換性として EXCEL などで作成した CSV ファイルをインポートすることができる。

簡単な例として、先のデータの五数要約を求めてみよう。操作はわずか 2 行である。1 行目はデータの入力、そして 2 行目の summary(x)が五数要約の値を求めるコマンドである。また、計算結果は直下に表示されている。

```
> x<- c(15,34,38,42,43,48,50,51,51,53,56,59,64,65,67,68,71,72,72,79,99)
> summary(x)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
     15      48      56      57      68      99
```

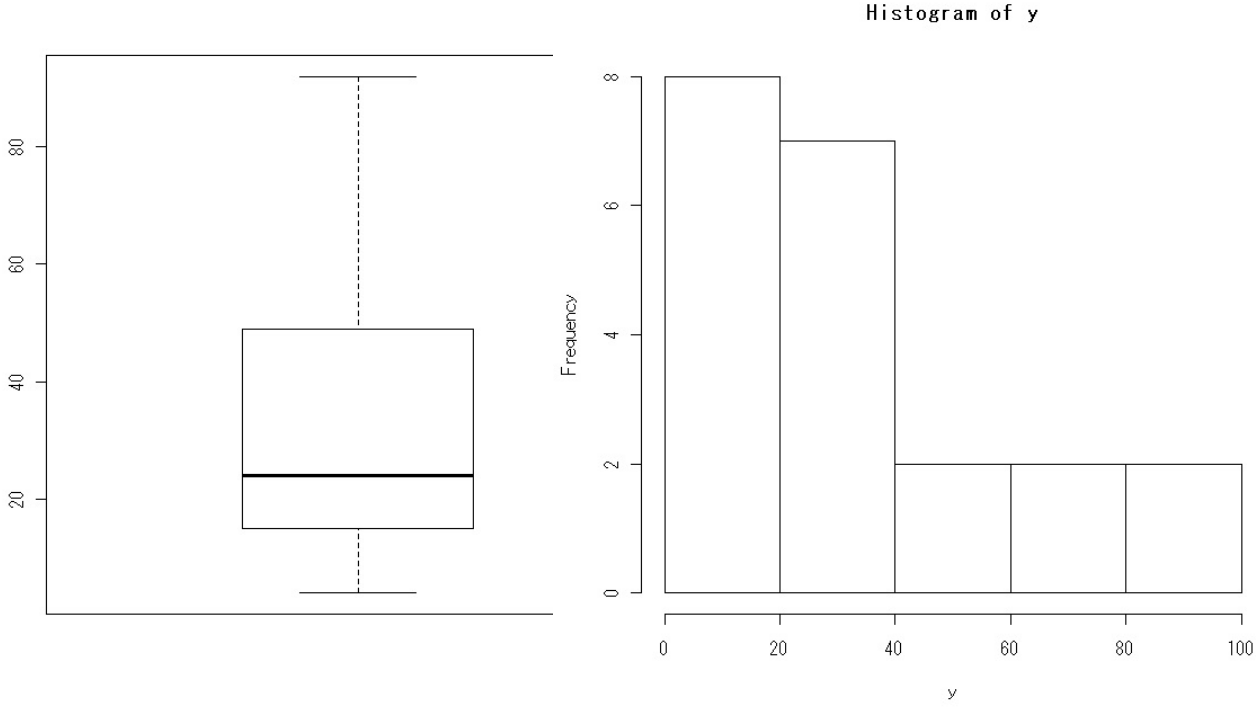


左の図は、はじめの箱ひげ図と同じものを「R」で作ったものである。操作は以下の通りである。こちらは外れ値をこめて作成してある。操作は簡単であるから、いろいろなデータについて箱ひげ図を作成して比較する際には便利なものである。

```
> x <- c(15,34,38,42,43,48,50,51,51,53,56,59,64,65,67,68,71,72,72,79,99)
> boxplot(x)
```

(6) 箱ひげ図の見方

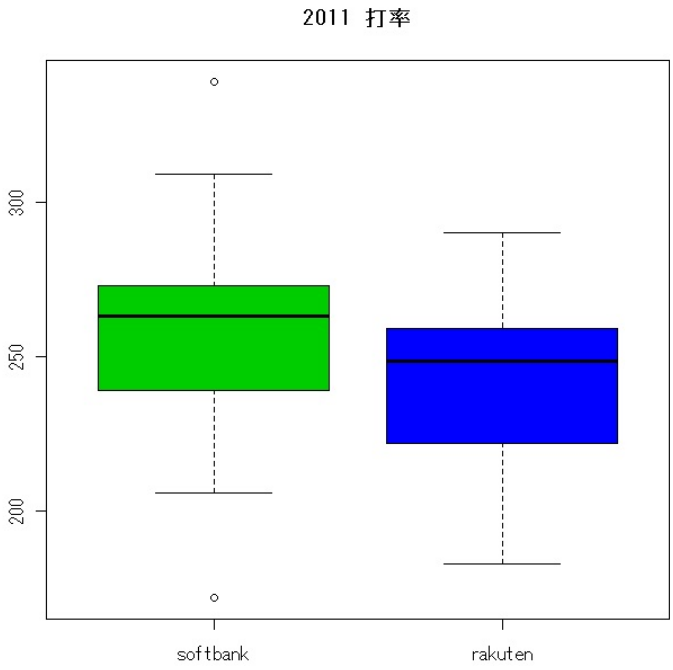
箱ひげ図は分布の様子をみるものであるから、例えば次のようなデータで考えてみるとよい。




```
> y <- c(4,8,10,13,14,15,17,17,21,22,24,25,25,30,35,49,58,70,78,85,92)
> hist(y)
> boxplot(y)
```

右のヒストグラムからもわかるように低い数値にデータが集中している場合である、このときの箱ひげ図は箱の部分が低い数値に寄っており、ひげの部分は高値へ長く伸びている。ヒストグラムも分布の様子を伝えるものであるが、箱ひげ図は縦や横にスリムに表現できることなどから、いくつかのデータ系列の分布の比較に用いられることがある。

（7）複数データの分布の比較



左の箱ひげ図は 2011 年プロ野球パ・リーグの、この原稿作成時点、つまり９月２日までの選手打率（打席数 100 以上が対象）を２つの球団で比較したものである。このように見てみると、その時点で首位のソフトバンクと３位の楽天の打撃力の差がある意味視覚的にわかるのではないだろうか？

この箱ひげ図も「R」によりわずか３行で作成されたものである。

またデータソースは「プロ野球フリーク http://baseball-freak.com/」による。

```
> x<- c(273,304,266,268,339,263,235,239,309,206,243,239,172)
> y<- c(290,259,288,247,230,227,250,260,258,254,183,222,195,184)
> boxplot(x,y,col=3:4,names=c("softbank","rakuten"),main="2011 打率")
```

４　まとめと補足事項

以上見てきたように、箱ひげ図は標本の分布を視覚的に表現する道具のひとつとして今時改訂の教科書に導入されている。これまで以上に、「統計的なものの見方」という点に視点が置かれているのが今般の学習指導要領改訂における統計教育分野の特徴である。箱ひげ図を指導するにあたっては、実際に箱ひげ図をかかせる時間が充分にとれるとは考えられないが、余裕があれば実際に手を動かすことも重要である。最低いくつかの標本（データ）のパターンでヒストグラムと箱ひげ図の対応などは押さえておきたいところでもある。さらに複数データ群の箱ひげ図による比較なども生徒にとっては視覚的には面白いものかも知れない。また、統計授業を支援するところの教育工学についてであるが、ここではエクセルをあえて用いずに説明してみた。エクセルでの箱ひげ図の作成については、適当なアドインソフトやマクロパッケージがあれば容易であるが、それを無くしてはなかなかプロセスも多く教室で説明しながらというのには難がある。勿論それに代わって「R」がすぐに役立つとは到底思えないが、ひとつの手段としてあえて紹介してみたところである。「R」はまだまだ一部の人々以外には認知度は低く、中等教育レベルで利用されている例もあまり聞かない。今般の事例で少しでも「R」の認識が高まることも個人的には期待するところである。

５　参考資料

・統計学研究連絡委員会報告「知識創造社会に向けた統計教育の推進について」
http://www.scj.go.jp/ja/info/kohyo/pdf/kohyo-19-t1031-10.pdf
・東京書籍版高等学校教科書「数学Ⅰ」
・「プロ野球フリーク」 http://baseball-freak.com/
・「R の基本操作入門」 http://www.e.okayama-u.ac.jp/~nagahata/bstat/R-kihon.pdf
・青山和裕「統計教育を取り巻く環境の変化と今後の推進に向けて」イプシロン,2009,vol51,37-42
http://repository.aichi-edu.ac.jp/dspace/bitstream/10424/3053/1/epsilon513742.pdf