超入門!

Rでできる ビジュアル統計学

学会・論文発表に役立つ データ可視化マニュアル

藤田医科大学
医療科学部

藤井亮輔
鈴木康司

Kinpodo

目次

1. 箱ひげ図（Box-whisker plot）

どんなグラフ？

箱ひげ図は、**量的な変数の分布を表すグラフ**として用いられています。その名前の通り、「箱」と呼ばれる四角形と「ひげ」と呼ばれる直線の2つの部分で構成されています（詳しくは後述します）。この「箱」と「ひげ」にデータの分布に関する情報が詰め込まれています[31)]。

箱ひげ図は、データを分析する前に対象とする変数の分布を確認するときに用い、この分布の確認によって、使用する分析手法を決めることがあります。また、複数の集団で連続値を比較する場合[1]にも箱ひげ図を描画して、それぞれの群における分布を表示することもあります。

論文のガイドラインやデータ可視化の論文でも、量的な変数については、箱ひげ図など分布の情報も含めて図示することが望ましいとされています[5,6,7,32,33)]。

実際にグラフを見てみよう！

SBP、TGは、いずれも量的な変数（連続変数）であることはPart 1で説明した通りです。実際に、これらを箱ひげ図として表現したものを見ていきます（**図5-1**、**図5-2**）。

図5-1 ｜ SBPの箱ひげ図

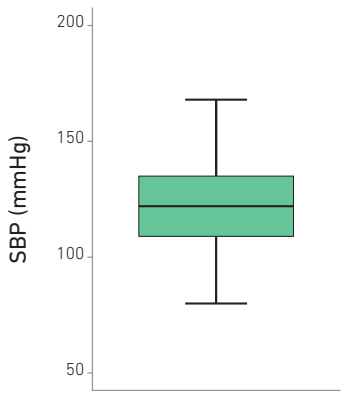
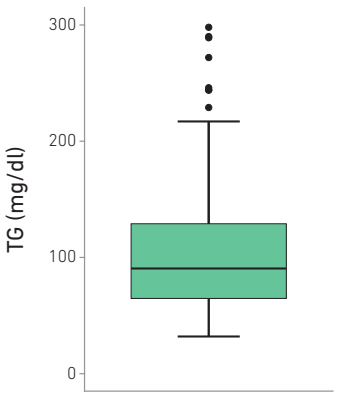


図5-2 ｜ TGの箱ひげ図



[1] あくまで、生データでの比較を想定していますので、結果の解釈については調整済み平均値（最小二乗平均値）などを活用した比較が必要になります。

第5章 データタイプ5（量的な変数・一変量）

【 このデータタイプに適したグラフ 】

- 箱ひげ図
- ヒストグラム
- 密度プロット

学　生　箱ひげ図とヒストグラムですね。これは、統計学やデータ分析を学ぶ上では必修のグラフですよね。ここから量的な変数を中心に扱っていくということですね。気持ちを切り替えてしっかりと勉強したいと思います。

Dr. グラフ　1つの量的なデータを扱う場合の基本的なグラフですが、複数の変数を扱う場合にも応用できることが多くあります。その基本をこの章で学習しましょう。さらに、これまでの質的データでは出てこなかった量的なデータに特有の「ばらつき」を扱います。それぞれのグラフでは、「ばらつき」をどのように表現しているか、注目してみていきましょう。

【 この章で使用するデータ 】

この章では、サンプルデータの一部を抜粋したものを使用していきます（**表5-1**）。

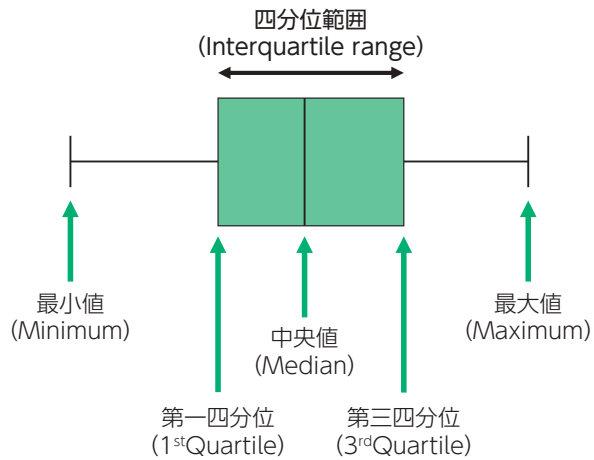
表5-1 ｜ サンプルデータの抜粋

ID	SBP	TG
1	136	140
2	172	88
3	107	391
4	132	90
5	154	61
…	…	…

SBP、TGの列には、それぞれ収縮期血圧（systolic blood pressure: SBP, mmHg）、中性脂肪（triglycerides: TG, mg/dl）の情報が格納されています。

この章では、これらを1つずつ独立した変数として扱っていきます。

図5-3 ｜ 箱ひげ図の説明



どのグラフも、縦軸にはそれぞれの量的な変数の分布が示されています。改めて、箱ひげ図の説明を**図5-3**に示しています。

箱ひげ図の「箱」を描くためには、データを4つに分割したときの3つの区切りの値である第一四分位数（データ全体の下から25%に当たる値）、第二四分位数（データ全体の下から50%に当たる値、いわゆる中央値）、第三四分位数（データ全体の下から75%に当たる値）を使用します。「箱」全体は、第一四分位数から第三四分位数までの範囲を示し、この範囲を四分位範囲（IQR: interquartile range）と呼びます。

また、「ひげ」には、最大値と最小値を用いる場合が多いですが、計算によって特定の値を求める場合など他にもさまざまなパターンがあります。

このグラフから読み取れる情報としては、上記のような基本的な統計量に加えて、次のようなものがあります。

①中央値の偏り＝分布の歪み

箱ひげ図では、中央値（データの50%に当たる値）が「箱」の中にある直線で表現されています。この中央値の位置によってデータの分布をおおよそ推測することができます。中央値が第一四分位に近い場合には右に歪んだ分布（**図5-2**）、第三四分位に近い場合には左に歪んだ分布であることが推測されます。箱の中心に中央値がある場合には、正規分布に近い分布であることが考えられます（**図5-1**）。

②四分位範囲（第一四分位から第三四分位までの距離）の広さ＝データのばらつき

四分位範囲、すなわち、箱ひげ図の「箱」の幅がデータのばらつきを示しています。四分位範囲が広い場合にはデータのばらつきが大きく、狭い場合にはデータのばらつきが小さいことが考えられます。

③外れ値（Outlier）の有無

図5-2のように、箱ひげ図の「ひげ」よりも外側にある点として表されているのが外れ値です。このような外れ値は、データ分析に大きな影響を与えうるので気をつけなくてはなりません。

まずは、外れ値がデータ収集やコーディング

のミスであるかどうか確認します。つまり、その値を取ることが医学的に可能かどうか考えるわけです。原因によって外れ値を除外し、その問題に対応できるあるいは影響しない統計解析の手法を選択する必要があります。いずれにしても、よく吟味することになります。

ここに注意！

量的なデータの分布について、前述のように図示する機会は比較的多くあると思います。分布を示す場合には、表現したい情報のレベルに応じて、箱ひげ図以外のグラフと併用し、使い分ける必要があります。使い分けるグラフの種類とその理由も含めて整理しましょう。

また、箱ひげ図の「ひげ」の描き方には種類がありますので、その点にも気をつけましょう。

1. 箱ひげ図も分布を知る上で、万能なグラフではない

箱ひげ図は、通常の棒グラフよりもデータの分布を示す場合にはより多くの情報を提供するグラフとされています。確かに**図5-4**を見ると、本来の分布を個別のデータを点として表現しているジッタープロット（Jitter plot）（一番右のカラム）はそれぞれ異なるにもかかわらず、棒グラフは全て同じ高さとして描画されています（エラーバーは、ばらつきをきちんと反映されています）。

しかし、箱ひげ図もジッタープロットに比べると、データの分布は隠されていることが分かります。

図5-4 ｜ 3つの連続量に関する棒グラフ、箱ひげ図、ジッタープロット（本来の分布）

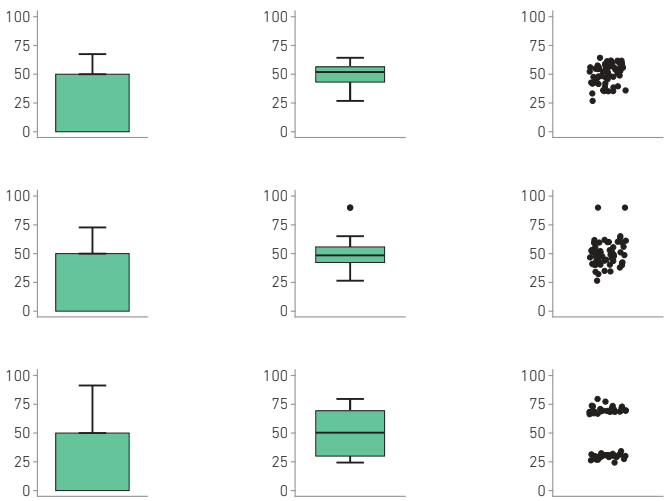
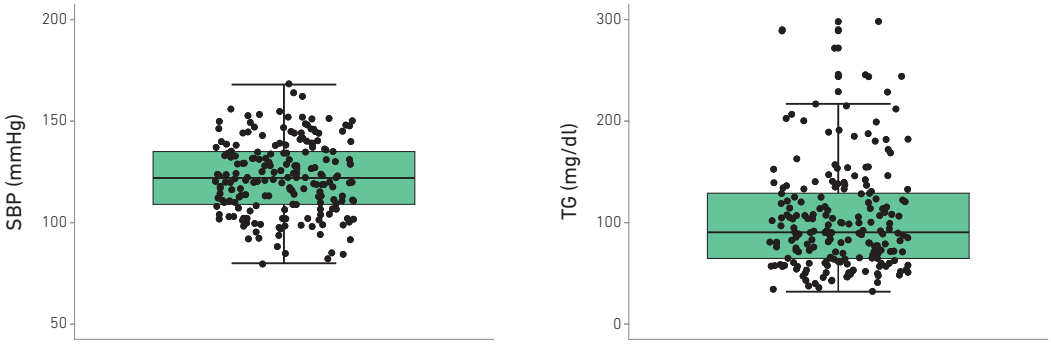


図5-5 ｜ 箱ひげ図にジッタープロットを重ね合わせたグラフ(左：収縮期血圧、右：中性脂肪)



箱ひげ図よりもさらに詳細な分布を必要とする場合には、箱ひげ図の上にジッタープロットを重ねて描画することで、解決できます[13]（**図5-5**）。また、後ほど説明するバイオリンプロット（Violin plot）（P. 90）もデータの分布について提示できるグラフとして使用されることがあります。

2．「ひげ」の定義に注意する

「ひげ」の範囲として最も一般的なものが最大値と最低値です。高校生で学習する「ひげ」も最大値と最小値になっています。

しかし、R の ggplot2 のデフォルト設定では、「ひげ」が第三四分位＋1.5×四分位範囲と第一四分位－1.5×四分位範囲で表現されるようになっています[2]。このように、「ひげ」の範囲は異なる場合があることを念頭において、十分注意を払いグラフを解釈するように心がけましょう！

[2] ただし、実際の上端・下端は、最大値が第三四分位＋1.5×四分位範囲を超えない場合には最大値、最低値が第一四分位－1.5×四分位範囲を下回らない場合には最低値で描くようになっています。

R で実行するコード

今回は、サンプルデータの SBP について箱ひげ図のみのグラフ（図 5-1）とそこにジッタープロットを重ね合わせたグラフ（図 5-5）を描いた場合のコードを下に示しています。

箱ひげ図の描画には、基本的に geom_boxplot を使用します。今回は、一つの変数（一変量）の描画になるので、ggplot の aes 内の x には、何もないことを "" として表示しています（2 行目）。

```
1  ggplot(data) +
2    aes(x = "", y = SBP) +
3    stat_boxplot(geom = "errorbar", width = 0.3) +
4    geom_boxplot(fill = "grey") +
5    theme_classic() +
6    labs(x = "", y = "SBP (mmHg)") +
7    scale_y_continuous(limits = c(50, 200))
```

次に、図 5-5 のようなジッタープロットを重ね合わせた箱ひげ図を描画していきます。基本的なコマンドは変わりませんが、geom_jitter で各データ（黒い点）を上乗せしています（5 行目）。

```
1  ggplot(data) +
2    aes(x = "", y = Age) +
3    stat_boxplot(geom = "errorbar", width = 0.3) +
4    geom_boxplot(fill = "grey") +
5    geom_jitter(color = "black", size = 1.5, width = 0.2) +
6    theme_classic() +
7    labs(x = "", y = "SBP (mmHg)") +
8    scale_y_continuous(limits = c(50, 200))
```

こんなグラフもあるよ！

バイオリンプロット（Violin plot）

バイオリンプロットとはその名の通り、バイオリンのような形を描くグラフです（分布によっては、そう見えないパターンもあります）（**図5-6**）。箱ひげ図と同じく、量的な変数の分布を示すために用いられます[34]。

簡単なデータのばらつきについて情報を得たい場合には箱ひげ図で十分ですし、詳しくそのデータの分布について知りたい場合にはバイオリンプロットやジッタープロットが推奨されます。一方で、バイオリンプロットやジッタープロットから要約統計量を実際の値として読み取ることは難しくなります。

最終的な解決策としては、それぞれの長所と短所を把握して、表現したい情報に適したグラフを選択することが必要になるのです。

こんな風に使うのも良いね！

データを収集し入力する段階で、量的なデータに欠測値がある場合には「99999」や「9」、「999」、「0」と入力することがあります。このような情報を持った個人については除外したり、特定の値を代入する前処理を行った上で統計解析を行います。

図5-7では、999を含んだ事例を示しています。このような状況ではすぐに統計解析に移ることもできません。解析する前に記述的なグラフを描くことで、データのコーディングやクリーニングのミスに気づく確率も高くなります。

図5-6 ｜ バイオリンプロット
（左：収縮期血圧、右：中性脂肪）

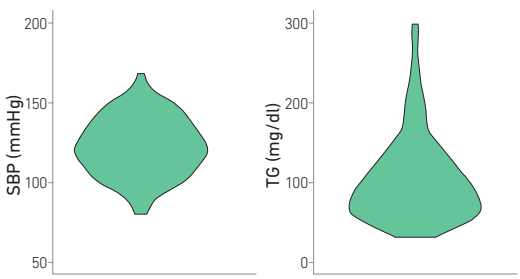


図5-7 ｜ 誤ったコーディングを含んだ状態での箱ひげ図