原稿作成日：　2017 年 3 月 31 日

記述統計量とグラフの描き方

〈教材提供〉
AMED 支援「国際誌プロジェクト」　提供

草案
新谷歩　大阪市立大学医学研究科医療統計学講座教授
加葉田大志朗　大阪市立大学医学研究科医療統計学講座特任助教

査読
大門貴志　兵庫医科大学医療統計学教授
山中竹春　横浜市立大学医学部臨床統計学教授
市川家國　信州大学特任教授
山本紘司　大阪市立大学大学院医学研究科医療統計学講座准教授
石原拓磨　大阪市立大学大学院医学研究科医療統計学講座特任助教

目次

はじめに

本単元ではグラフを使って視覚的にデータを表す方法について学びます。グラフにすることで、数字でまとめただけではわからないような情報や、研究の論旨をより分かりやすく読者に伝えることができます。ソフトウェアを使って実際にグラフを描く方法はビデオで解説します。

学習目標

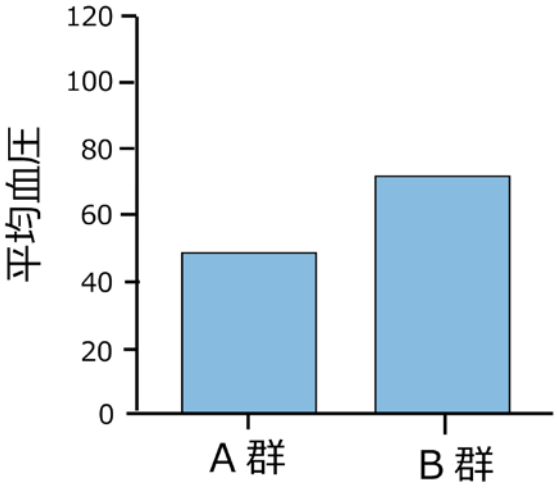
本単元を通じてあなたが修得を目指すものは：

- 各グラフの名称と意味を説明できる
- エラーバーの意味と使い方を説明できる

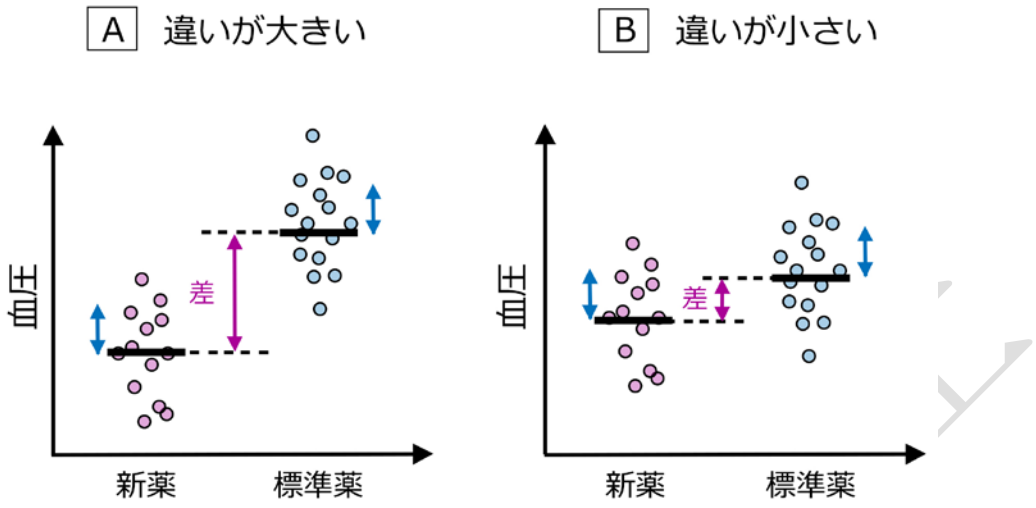
グラフの種類と読み方

棒グラフとエラーバー

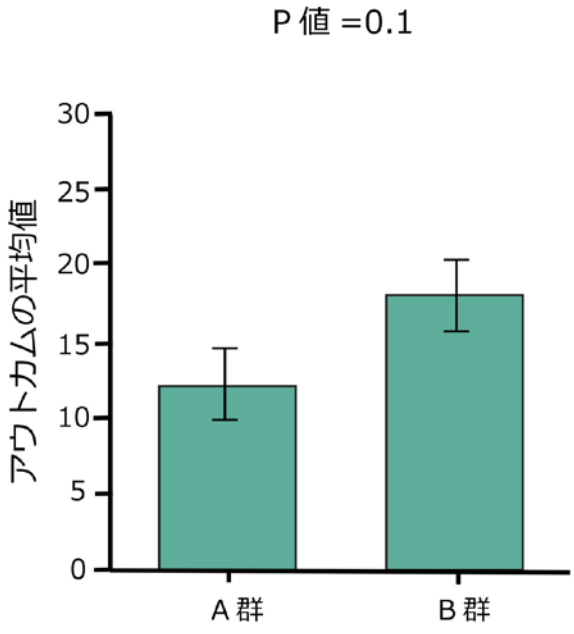
おそらく多くの方が一度は見たことがあるグラフは棒グラフでしょう。このグラフは、数値を群間で比べるときによく使われます。一番よく使われる棒グラフは棒の高さが平均値を表すグラフです。下のグラフではB群の平均血圧がA群より高いことが分かります。



しかし、この図からはB群の平均血圧がA群の平均血圧よりも統計的に有意な差を持って高いか否かはわかりません。検出したい群間のデータの違い（平均血圧の群間差など）がデータのばらつきよりも大きいとき、統計的に有意な差は得られ易くなります。したがって、アウトカムの平均値を複数の群で比較する場合は平均値を表す棒以外にデータのばらつきを示すことが大切です。データのばらつきを示すのによく用いられるのがエラーバーです。



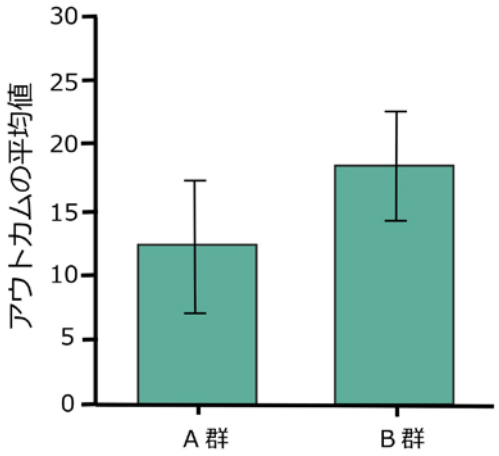
エラーバーとして「標準偏差」「標準誤差」「95%信頼区間」の 3 通りのうち一つが通常用いられます。



棒グラフと標準誤差のエラーバー

上の図は平均±標準誤差でエラーバーを付与した図です。これは非常に多くの論文で用いられますが、このグラフを用いて「二つのエラーバーが重ならないので統計的に有意な差がある」とよく言われますが、これは間違いです。実際のところ、上の図では二つのエラーバーは重なっていませんが、P 値は 0.1、つまり、統計的に有意な差は示されていません。二つのエラーバーが重ならないことでもって統計的に有意な差があると判断できる図は信頼区間をエラーバーとして用いた図のみです。この例の場合、95%信頼区間の上限および下限は平均値±2×標準誤差に対応します。

P 値 =0.1

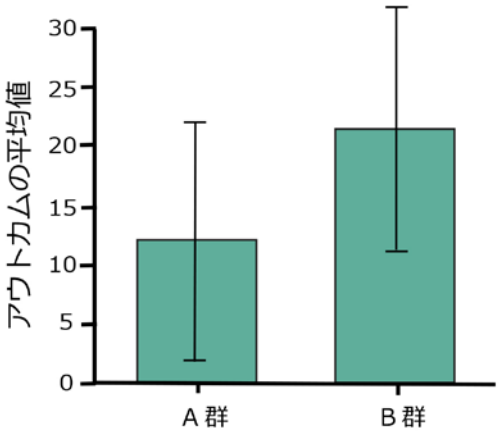


棒グラフと信頼区間のエラーバー

上の図は平均を表す棒グラフに 95%の信頼区間を付与したものです。この結果を見ると二つの信頼区間が重なっています。このような場合は有意水準を 5%で仮説検定を行った場合には、統計的に有意な差があるとはいえません。

下の図は平均値±標準偏差でエラーバーが描かれています。データが正規分布に従っているとき、平均値±標準偏差はデータのおよそ 3 分の 2 が入っている範囲を意味します。下の図では二つのエラーバーは重なっていますが、統計的に有意な差が出ています。標準偏差のエラーバーは収集したデータのばらつきを示す指標にはなりますが、統計的に有意な差を確認する上では役立ちません。

P 値 =0.01



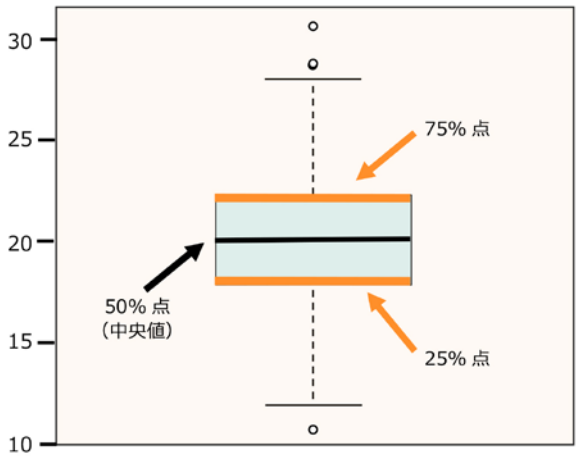
棒グラフと標準偏差のエラーバー

エラーバーが何を示しているかを示す

ここまでで述べたように、どのばらつきの指標をエラーバーとして利用するかによって、その図が示す内容が異なります。そのため、論文でデータの平均を棒グラフを利用して示す場合には、棒グラフとともに用いられるエラーバーが何を示しているのかを明記しておくことが重要になります。Nature などの国際誌のガイドラインでは、グラフにはできるだけエラーバーを付与し、そのエラーバーの名称を記載するように指摘しています。

様々な情報をもつ箱ひげ図

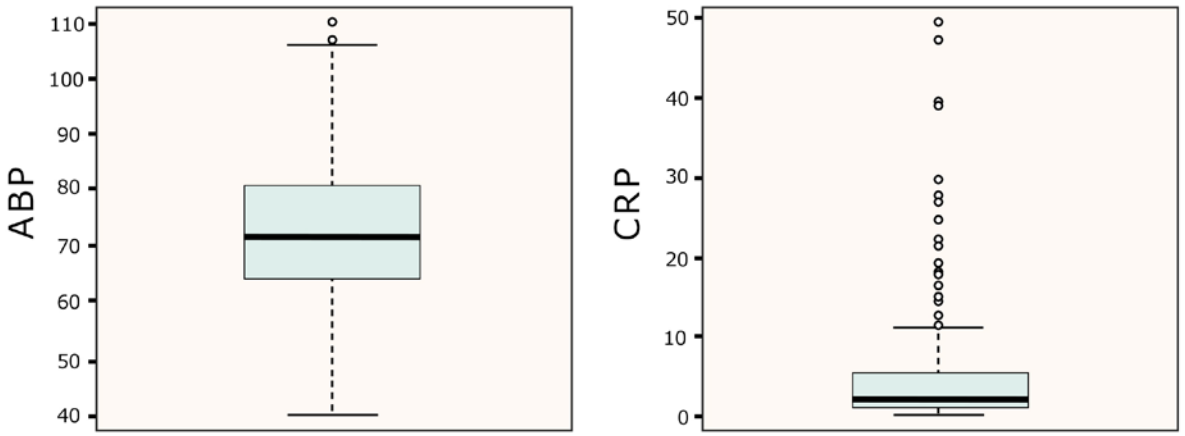
ここまでは平均値の持つ情報をプロットする方法として棒グラフを説明してきました。しかしデータの要約の仕方には、それ以外に中央値と四分位範囲があることも説明しました（単元 8 を参照）。中央値や四分位範囲を図示する方法として箱ひげ図があります。



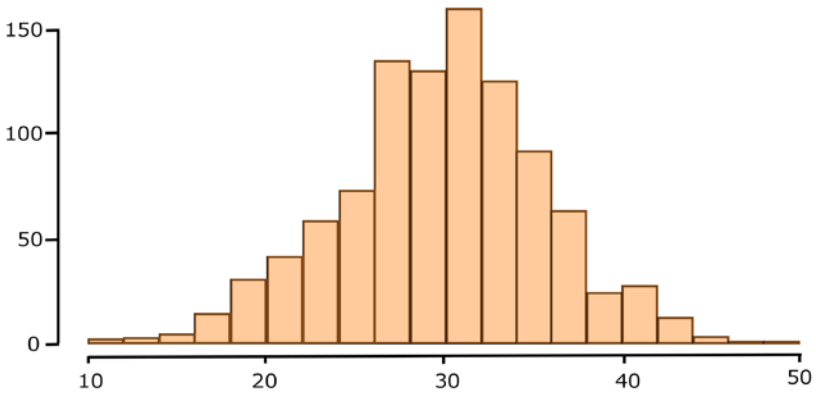
箱ひげ図は日の字のような「箱」の上下に、エラーバーのような「ひげ」がついた図になります。まず箱の方から説明すると、箱の一番下のラインが 25%点、一番上のラインが 75%点、箱の中のラインが 50%点（中央値）となります。つまり箱の上下の値が四分位範囲を示すことになります。上のひげは通常、「箱のふた部分にあたる赤のライン上部から箱の縦の長さの 1.5 倍の範囲」の中に存在するデータの最大値のところに描かれています。下のひげも同様に、「箱の底の赤のラインから下の範囲で、赤のラインから箱の縦の長さの 1.5 倍の範囲」に存在するデータの最小値のところに描かれています。この「1.5」という値はソフトウェアによって設定が変わってくるので、利用の前には確認が必要です。ひげの外にくるデータを外れ値としてドットとして加えて記入し、極端な値をもつデータがどこにあるかを箱ひげ図の上に表示することも可能です。

箱の真ん中あたりに中央値のラインがあり、外れ値も左右対称である場合は、データは正規分布に近い分布を持っていると仮定することができます。しかし中央値のラインが極端に箱の上下のどちらかに寄っていれば、それは歪んだ分布を持つデータであることが示唆されます。このようなケースでは平均値を用いると誤解を生じることは単元 8 で説明した通りです。このように箱ひげ図は、中央値と四分位範囲を示すことに加えて、そのデータがどのくらい歪んだ分布を持つデータなのかを視覚的に把握させてくれるので、より情報量の多いグラフとして好まれる傾向があります。下に正規分布を仮定できる分布（左）と、歪んだ分布（右）を例示しておきます。

正規分布を仮定できる分布　　正規分布を仮定できない歪んだ分布

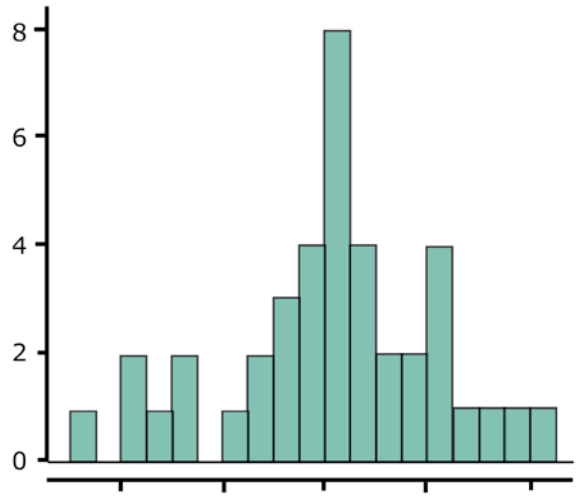


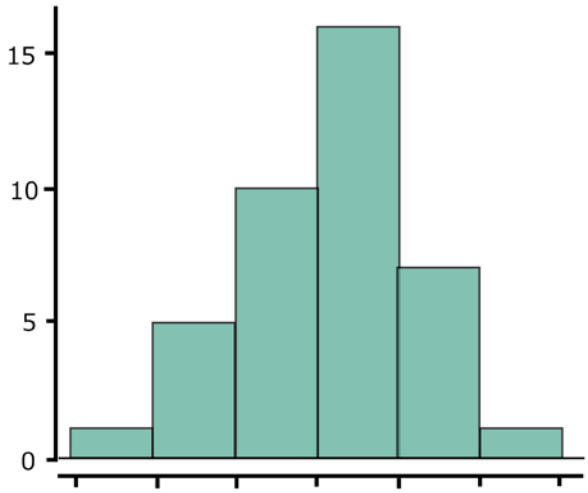
ヒストグラム



ヒストグラムは年齢など連続変数の分布を調べるのに適したグラフです。横軸が興味のあるデータのとる値、縦軸がデータのとる値、あるいは一定の幅内にある値の頻度になります。右の図は500人の年齢の分布を表しています。12歳（すなわち、生後12年0日 - 12年364日の人）が30人、15歳が70人、20歳が160人というように、その年齢に何人いるかがカウントされて縦軸の高さになります。ヒストグラムは連続データの分布を視覚的にとらえることができるというメリットがあります。右上の図のようにヒストグラムの頂点を中心に両側に均等に釣鐘型をとるように広がっている場合は、正規分布に従っていると判断します。

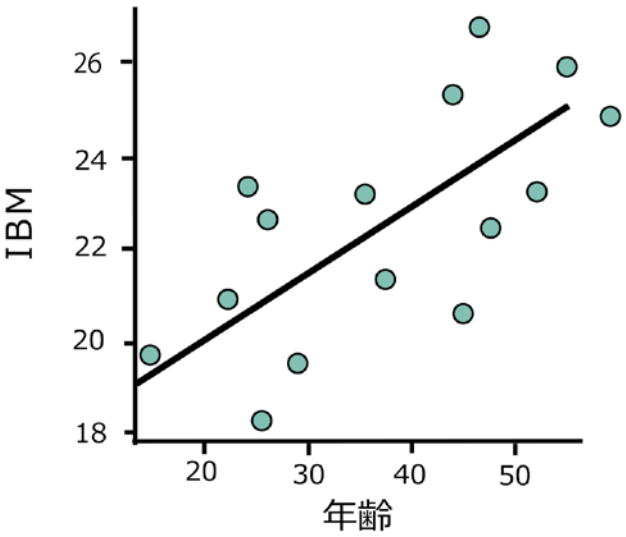
一方、ヒストグラムは個々の変数の値が含まれる幅を広くも狭くも設定して描くことができます。例えば左下の図では1歳ごとに幅が分割してカウントされていますが、これが右下の図のように16-20歳、21-25歳、26-30歳と5歳ずつに幅を分割し直した場合、見え方はかなり異なってきますので、幅を設定する際は注意が必要です。





散布図

散布図というのは字のとおり、各データが散布されているように点で描写されるグラフになります。散布図は、2つの連続変数の関連性を見るために便利なグラフです。例えば、右の図は、年齢とBMIの散布図です。年齢が高くなるにつれ、BMIも増えるという関連性をみてとれます。



この単元に関係するビデオ教材

棒グラフ
箱ひげ図
ヒストグラム
標本（患者）背景表

本単元は日本医療研究開発機構:研究公正高度化モデルである「医系国際誌が規範とする研究の信頼性にかかる倫理教育プログラム」（略称：AMED 国際誌プロジェクト）によって作成された教材です。作成および査読等に参加した専門家の方々の氏名は、冒頭に掲載されています。

この単元に関係する国際誌におけるチェックポイントをいくつか紹介します。
（内容は解釈を助けるために一部意訳している部分もあります）

①Nature
（http://image.sciencenet.cn/olddata/kexue.com.cn/upload/blog/file/2010/12/2010128212513557501.pdf; visited on 2018.02.11）
②New England Journal of Medicine （http://www.nejm.org/page/author-center/manuscript-submission#electronic; visited on 2018.02.11）
③Science （http://www.sciencemag.org/authors/science-editorial-policies; visited 2018.02.11）
④The EMBO Journal （http://emboj.embopress.org/authorguide#embargopolicy; visited on 2018.02.11）
⑤JAMA （http://jamanetwork.com/journals/jama/pages/instructions-for-authors; visited on 2018.02.11）

①Nature

- ■ 効果量の提示方法を編集した際（グラフＹ軸の短縮など）には、その旨と理由を記載すること
- ■ エラーバーは可能な限り各グラフにつけること
- ■ エラーバーが何を示すか明記すること

③Science

- ■ 連続変数については散布図や、箱ひげ図、ヒストグラムなどのグラフを用いて、または平均や中央値および標準偏差や四分位範囲などデータの中心位置やばらつきを表す指標を用いて記述すること。

④The EMBO Journal

- ■ 統計手法に関する記載は基本的に materials と methods に記載するが、図表においても解析手法・症例数・P 値などの基本的な情報は記載しておく。
- ■ グラフにはエラーバーとそのエラーバーの名称を記載しておく。さらに±標準偏差や±標準誤差の表記が必要。