0.1

要約

多分私は単純な心の持ち主なんですが，絵が好きなんです。新しい論文を書き始めるとき，まず私がすることはどっしり座ってどんな絵を描こうかなと考えるのです。頭の中で，その論文は実際にストーリーに沿った一連の図を思い浮かべています。残りは飾りに過ぎません。私がここで伝えたいことは，人間の視覚システムはとても強力なデータ分析ツールであるということです。図は正しい情報を与え，大量の情報を瞬時に読者に伝えることができます。“百聞は一見にしかず” という諺の通りです。そう考えると，この章はこの本の中で最も重要なものの一つではないかと思うのです，本章で扱ったのは次のとおりです。:

- *一般的なプロット*. この章のほとんどは統計学者が好んで使う標準的なグラフを紹介しました。: ヒストグラム (セクション 1.1) とボックスプロット (セクション 1.2) です。
- *画像の保存*. 重要なことで，あなたの描いた図を出力することについても言及しました (セクション 1.3)

最後にひとつ。JASP は美しいグラフを提供してくれますが，プロットの編集はできるようにはなっていません。もっと発展的なグラフやプロットの可能性を引き出すには，R におけるパッケージを使うことでさらに強力に進めることができます。最も有名な廟 g システムの一つは，`ggplot2` パッケージ (`http://ggplot2.org/` を参照) によって提供されています。これは “グラフィクスの文法"(**Wilkinson2006**). という考え方に基づいているのです。それは初心者向けではありません。それを使い始める前に，まず R の全体像を掴む必要がありますし，コツを掴むのには少し時間がかかります。ですが，準備ができたら学ぶ価値はあります。それは本当にパワフルでよりスッキリしたシステムなのですから。

1. グラフを描く

何よりもまず，データを見せろ
–Edward Tufte[*1]

データを可視化することは，データを分析しようとするものにとって最も重要な課題です。これが重要なのは，二つの異なる，しかし相互に関係し合う理由によります。まず，“提示するグラフ” を描くこととは，あなたのデータをスッキリと提示し，読者にとってあなたが言いたいことを簡単に理解させるために視覚的に訴えかけるようにすることです。同じぐらい，あるいはもっと重要なことは，グラフを描くことで*あなた自身*がデータを理解できるようになることです。そのために，“探索的なグラフ” を描くことは，あなたがいざ分析しようとしているデータについて理解するのを助けることになるのが重要なのです。このことは当たり前のようでもありますが，私はこれを人に何回言ったかわからないほどです。

この章の重要さを示すために，優れたグラフというものがいかに有用なのかを示す典型例から始めたいと思います。そのために，図 1.1 に最も有名なデータの可視化の例の一つを示しています。これは 1854 年，John Snow によるコロナの死亡者数の地図です。この図はその単純さにおいて，非常にエレガントだといえます。背景として，われわれは見る人の方向性を示すストリートマップを持っている，というのがあります。地図上には多数の小さな点があり，それぞれがコロナの発祥地点を表しています。大きな文字は水のポンプの位置を示していて，その名前ラベルがついています。この図をちょっと見ただけでも，アウトブレイクの源は Broad Street ポンプを中心にしていることが明らかです。このグラフを見て，Dr.Snow はポンプからハンドルを取り除き，500 人以上を殺したアウトブレイクを終わらせたのです。これが，良いデータの可視化の力です。

この章の目標は二つあります。まず，データを分析したり表示したりするとき，私たちがよく使うグラフについて説明し，続いてこれらのグラフを JASP で作成するにはどうすれば良いかを示します。このグラフそのものは，直接的なものなので，この章のある側面は非常にシンプルだと言えるでしょう。人がよく困惑するのは，グラフをどうやって作るかを学ぶとき，特に良いグラフをどうやっ

[*1] この言葉の原典は，Tufte の本『量的情報を可視化する』です。

て作れば良いかを学ぶときです。幸い，JASP でのグラフの書き方は，あなたがグラフの見え方にそれほどこだわらなければ，かなりシンプルなものです。私がこれをいうことの意味は，JASP のデフォルトのグラフがかなり良いものだということで，ほとんどの場合すっきりとクオリティの高いグラフィックを提供できるということです。しかし，標準的でない図を描きたいとあなたが思ったとき，あるいは図にかなり特殊な変更を加える必要があるとき，JASP のグラフィック関数は発展的な仕事や詳細な編集にはまだ向いていないということはあります。

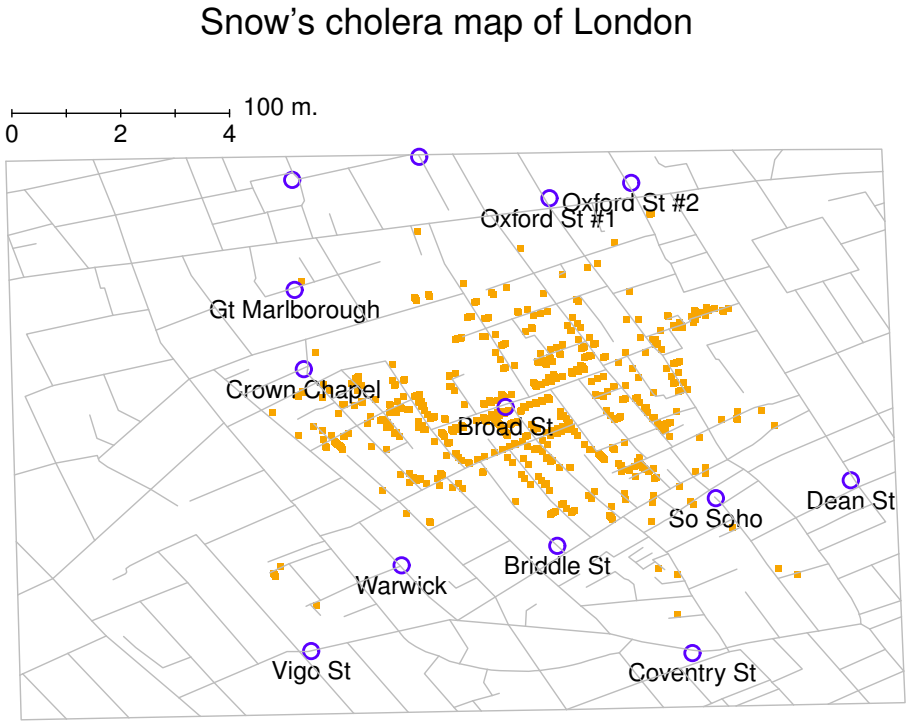

Figure1.1　John Snow のスタイリッシュなコロナマップのオリジナル。小さな各点はコロナ発生点で，大きな円は井戸の位置を示しています。このプロットが明らかにしたように，コロナのアウトブレイクは Broad St のポンプを中心にしていることがわかります。

1.1

ヒストグラム

普通の**ヒストグラム**の話から始めましょう。ヒストグラムは最もシンプルで最も一般的な，データ可視化手法の一つです。あなたが間隔尺度水準，あるいは比率尺度水準のデータ (例えば，第 **??** 章の `afl.margins` データなど) を持っていて，その辺図宇野全体的な印象を掴みたいと思った時に，ヒストグラムは有効です。ヒストグラムがどんなものかは，ほとんどの人が知っていると思います。広く使われていますからね。でも完璧を期すために，しっかり説明しておきます。あなたがすべきことは，あり得る値を**ビン**幅に分割し，各区間に入る観測度数の数を数え上げることだけです。この数のことを頻度とかビンの密度といい，それが垂直に伸びるバーとして表示されます。AFL の勝利数データでは，得点が 10 点未満だったゲームが 33 ゲームあり，これが以前示した第 **??**の 図 **??** 中，左端のバーの高さとして表されています。以前のグラフは JASP の能力を超えた，R の発展的プロットパッケージの力を使って描かれていました。しかし JASP もそれに近いことをしてくれます。JASP でのヒストグラムの描画はとても簡単です。'記述' - '記述統計' メニューの下にある 'プロット' をひらき，'分布のプロット' チェックボックスをクリックしたのが，図 1.2 に示されています。JASP のデフォルトでは，y 軸が '度数' とラベルされていて，x 軸が変数名になっています。**ビン**は自動的に選択されます。度数が表示されますが，実際の数字はそれほど問題にならないことに注意してください。むしろ，われわれが本当に興味を持っているのは，分布の形状からくる印象なのです。それがあ正規分布しているのか，それが尖っていたり歪んでいたりしないか?私たちの第一印象は，**ヒストグラム**から作られるのです。

JASP の特徴を一つ付け加えるなら，'密度' 曲線をこのヒストグラムの上に書き加えられるというところです。これをするには 'プロット' の下にある '密度を表示' のチェックボックスをクリックしてください。これが図 1.3 に示されているプロットです。密度プロットは連続した区間や時系列全体をカバーする分布を可視化します。この図は，プロットされた値に**カーネルスムージング**を使ったヒストグラムの一種で，ノイズを除去した平滑化によって分布をよりスムーズにしたものです。密度プロットのピークは，区間中の値がどこに集中しているかを示してくれています。ヒストグラムの上に密度プロットを描くことの利点は，分布の形をわかりやすくすることにあります。なぜならこれはビン (ヒストグラムで使われている各バー) の数に影響されないからです。たった 4 つのビンしかないヒストグラムは，20 のビンをもつヒストグラムに比べて分布の形をうまく表現できません。でも密度プロットでは，そういう問題が生じません。

この画像はプレゼンテーション用のグラフィック (例えばレポートに入れるもの) にするには，かなり修正する必要がありますが，データを描画する分にはかなりいい仕事をしてくれます。実際，ヒストグラムや密度プロットの強みは (適切に使えば)，データの全体的な広がりを表示し，それがどんな形をしているのかについてかなり良い直感を与えてくれることです。ヒストグラムの欠点は，コンパクトさに欠けるところです。他のプロットと違って，20 から 30 ものヒストグラムを一つの図に詰め込んで人に説明するのはとても難しいのです。そしてもちろん，データが名義尺度水準であれば

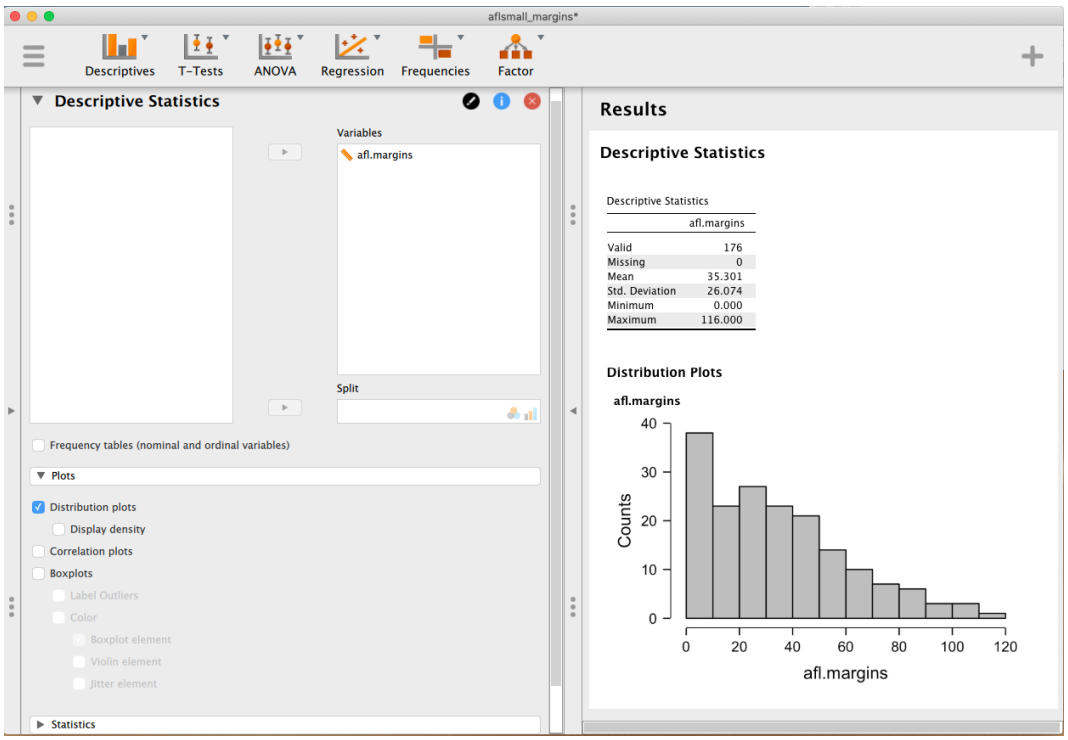

Figure1.2 ‘分布のプロット’ オプションによって作られたヒストグラムを描いた JASP のスクリーンショット

. . . . . .

ヒストグラムは適用できません。

1.2 ボックスプロット

ヒストグラムの代わりになるのは，**ボックスプロット**，別名 “箱ヒゲ図” と呼ばれるものです。ヒストグラムのように，間隔あるいは比率尺度水準のデータに適しています。ボックスプロットの背後にある考え方は，中央値，四分位範囲，データの幅を単純に示して見せようというものです。ボックスプロットによる表現は非常にコンパクトで，特にデータ分析の探索的な段階でデータがどんなものかを理解しようとする時の手法としてとてもポピュラーなものになっています。ではそれがどういうものか，`afl.margins` のデータを例にしてみていきましょう。

ボックスプロットがどんなものかを見るために，まず描いてみるのがいいでしょう。‘ボックスプロット’ をクリックすれば，右下に図 1.4 のようなものが示されると思います。デフォルトでは，JASP は最も基本的なボックスプロットを示します。このプロットを見れば，そこから何がわかるか

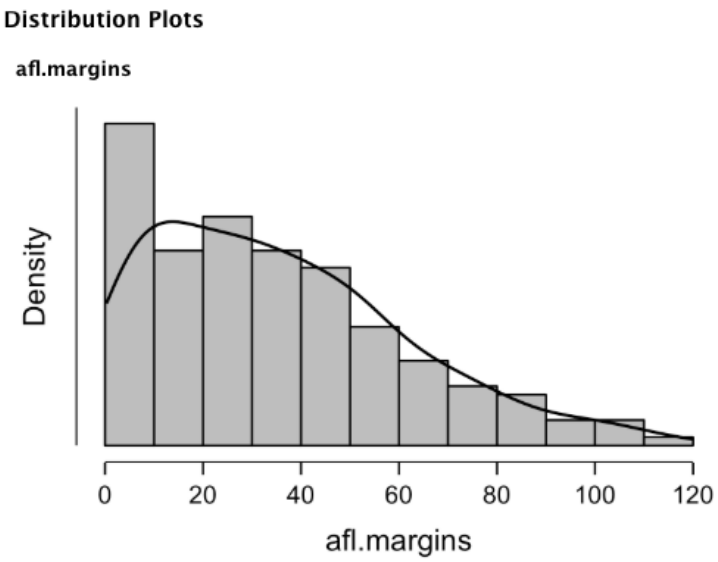

Figure1.3　`afl.margins` 変数の JASP による密度プロット

一目瞭然です。箱の中心にある太い線が中央値です。箱の幅は 25 パーセンタイルと 75 パーセンタイルの幅になっています。そして “ひげ” の部分はある限界値を超えない最も極端なデータポイントです。デフォルトでは，この限界値は四分位範囲 (IQR) の 1.5 倍で，下限は `25 パーセンタイル点 - (1.5*IQR)`，上限は `75 パーセンタイル点 + (1.5*IQR)` になっています。この範囲の外に入る点は，髭でカバーできないので円あるいは点で示され，これは一般的に**外れ値**とよばれます。私たちの AFL 勝率データでは，二つの観測点がこの範囲の外に落ちており，この観測データは点で表されています (上限は 107 で，スプレッドシートのデータをみると 2 件これより大きいもの，108 と 116 があり，それぞれの点が打たれています)。

1.2.1 Violin plots

伝統的なボックスプロットのバリエーションとして，バイオリンプロットというのがあります。バイオリンプロットはボックスプロットに似ていますが，異なる値におけるデータのカーネル確率密度も表示してくれます。典型的には，バイオリンプロットはデータの中央値と，標準的なボックスプロットと同じような四分位範囲を示すボックスも同時に示します。JASP では，この種の機能は ‘バイオリンの要素’ と ‘ボックスプロット要素’ のチェックボックスをチェックすることでできます。図 1.5 では，データ点もプロットしました (これは ‘Jitter 要素’ のチェックボックスを選択することで，

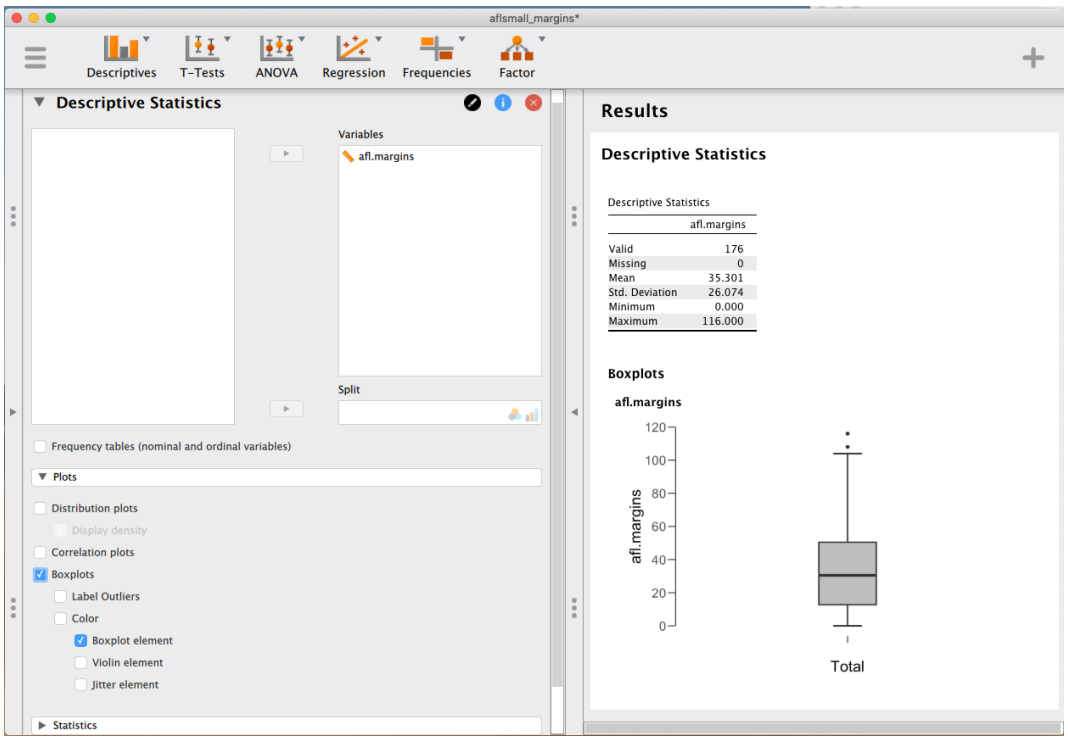

Figure1.4　JASP による `afl.margins` 変数のボックスプロット

プロットに実際のデータ点を追加します)。

1.2.2　複数のボックスプロットを描画する

最後にもう一つだけ。複数のボックスプロットを一度に書くにはどうしたらいいでしょう？ 例えば，2010 年の AFL 勝率データだけでなく，1987 年から 2010 年までの各年度のボックスプロットを個別に描きたいと思ったとしましょう。これをするためには，まずデータを見つけなければなりません。このデータは `aflsmall2.csv` ファイルにあります。では JASP に読み込んで，みてみましょう。これはちょっと大きなデータセットであることがわかると思います。ここには 4296 ゲームとその変数が含まれています。JASP で **勝率** 変数についてのボックスプロットを描く時に，**年度**ごとに分けたいですね。それをするためには，**年度**変数を名義尺度水準の変数に変換し，**年度**にわたってボックスの ‘分割’ をします。

その結果が図 1.6 です。このバージョンのボックスプロットは，年度ごとに分割されており，ヒストグラムよりボックスプロットを使った方がいいこともあるのはなぜか，ということがすぐにわかりますね。これを見ると，データの詳細に入り込まなくても年度ごとにどうなっているか，わかりやすくなっています。もしこのスペースに 24 個のヒストグラムを詰め込もうとしたら，何が起こるか

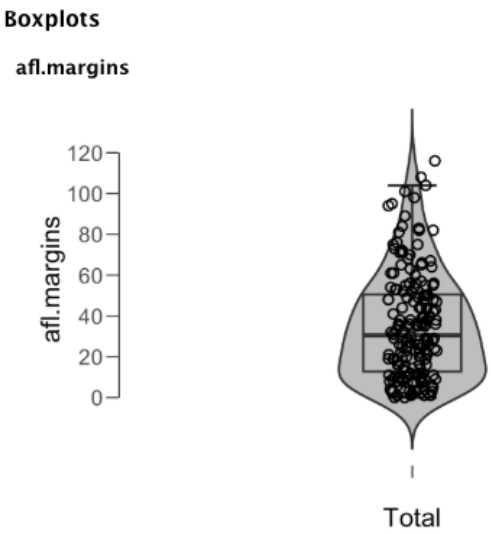

Figure1.5　JASP における `afl.margins` 変数のバイオリンプロットにボックスプロットとデータ点も重ねてみました

考えてみてください。そんなことをしても，読者が何かを学べるとは思いませんけどね。

1.3

JASP で画像を保存する方法

ちょっと待って，と思ってるかもしれませんね。JASP でいい図が欠けてもそれを保存したり友達に送り，私のデータがいかに素晴らしいかを語れないようでは意味がありません。図を保存するにはどうしたらいいでしょう？ 簡単です。プロットの上部，横についている三角形をクリックして，‘名前をつけて画像を保存’ を選ぶだけです。いくつかのフォーマットを選んで保存することができ，選択できる形式は ‘png’,‘pdf’,‘eps’,‘tif’ があります。これらのフォーマットで友達に画像を送ったり，(もしかするとさらに重要なことには) それらを課題や論文に含めることができます。

1.4

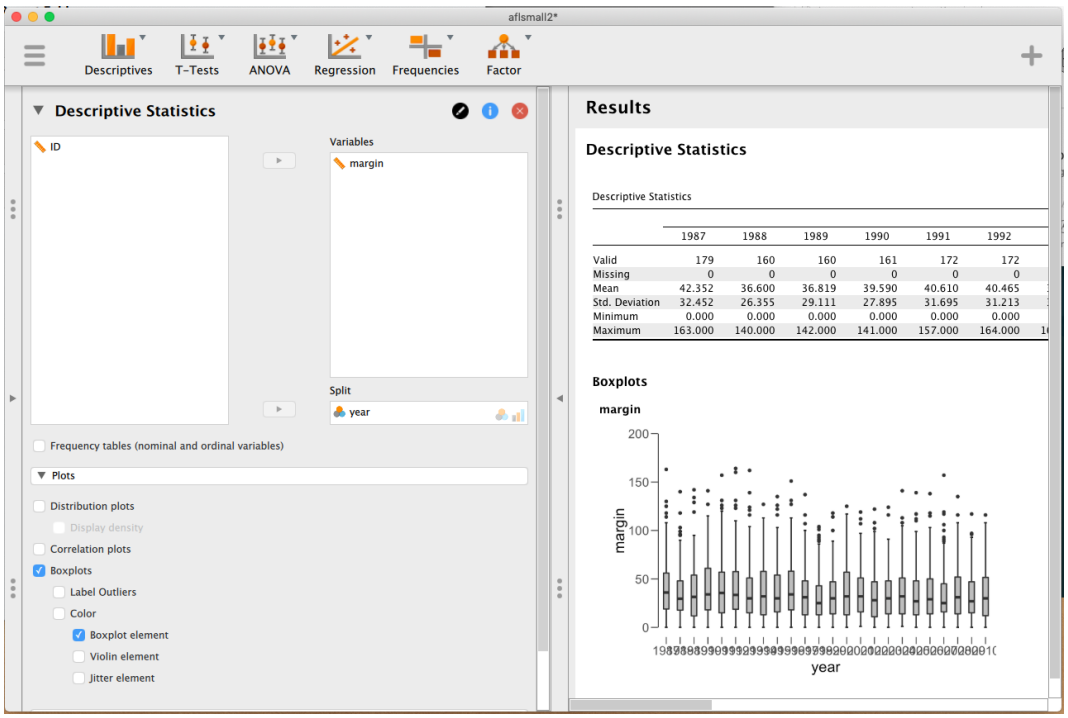

Figure1.6　JASP における複数ボックスプロット。`aflsmall2` データセットにおける，**年度**変数ごとの **勝率**

要約

多分私は単純な心の持ち主なんですが，絵が好きなんです。新しい論文を書き始めるとき，まず私がすることはどっしり座ってどんな絵を描こうかなと考えるのです。頭の中で，その論文は実際にストーリーに沿った一連の図を思い浮かべています。残りは飾りに過ぎません。私がここで伝えたいことは，人間の視覚システムはとても強力なデータ分析ツールであるということです。図は正しい情報を与え，大量の情報を瞬時に読者に伝えることができます。“百聞は一見にしかず” という諺の通りです。そう考えると，この章はこの本の中で最も重要なものの一つではないかと思うのです，本章で扱ったのは次のとおりです。:

- *一般的なプロット*. この章のほとんどは統計学者が好んで使う標準的なグラフを紹介しました。: ヒストグラム (セクション 1.1) とボックスプロット (セクション 1.2) です。
- *画像の保存*. 重要なことで，あなたの描いた図を出力することについても言及しました (セクション 1.3)

最後にひとつ。JASP は美しいグラフを提供してくれますが，プロットの編集はできるようにはなっていません。もっと発展的なグラフやプロットの可能性を引き出すには，R におけるパッケージ

を使うことでさらに強力に進めることができます。最も有名な廟 g システムの一つは，ggplot2 パッケージ (http://ggplot2.org/ を参照) によって提供されています。これは “グラフィクスの文法”(**Wilkinson2006**). という考え方に基づいているのです。それは初心者向けではありません。それを使い始める前に，まず R の全体像を掴む必要がありますし，コツを掴むのには少し時間がかかります。ですが，準備ができたら学ぶ価値はあります。それは本当にパワフルでよりスッキリしたシステムなのですから。