| ID | KK300504 |
|---|---|
| 種別 | 解説 |
| タイトル | データの視覚化（5）<br>— SPSS のグラフ機能を利用して — |
| Title | Data Visualization (5):<br>Through the Use of the SPSS Graph Function |
| 著者 | 李　在鎬 |
| Author | LEE Jae-Ho |
| 掲載号 | 30巻5号 |
| 発行日 | 2016年6月20日 |
| 開始ページ | 292 |
| 終了ページ | 303 |
| 著作権者 | 計量国語学会 |

計量国語学 30 巻 5 号（2016 年 6 月）pp.292-303.

解説

# データの視覚化（5）
## ― SPSS のグラフ機能を利用して ―

李　在鎬（早稲田大学）

**要旨**

本稿では，SPSS によるデータの視覚化について説明する．特に，グラフ描画機能に注目する．言語の計量的研究において基礎的と思われる 3 つのグラフを取り上げる．ヒストグラム，箱ひげ図，散布図である．データとしては，李・長谷部・久保（2016）によるコーパスデータの文章難易度調査の成果物を利用する．利用データはウェブに掲載しているので，SPSS が利用可能な読者にはぜひとも試してみてほしい．

**キーワード**：SPSS，ヒストグラム，箱ひげ図，散布図，文章難易度（リーダビリティ）

### 1．はじめに

SPSS は世界的に定評のある統計分析用のソフトウェアである．1968 年に最初のバージョンが公開されてから，長年にわたって，色々な学問分野を支えてきたソフトウェアの一つである．バージョンによっては PASW Statistic と呼ばれたり，IBM SPSS と呼ばれたりするが，SPSS という名称がもっとも普及している．なお SPSS とは，Statistical Package for the Social Sciences の略称である．

断っておくが，SPSS は，あくまで統計分析用のソフトウェアであり，データの視覚化に特化したツールではないが，優れたグラフ描画機能が入っている．本稿では，SPSS によるグラフ機能を解説するもので，計量言語学の研究においてよく使われるヒストグラムと箱ひげ図と散布図を取り上げ，実例に基づいて説明していく．

説明用データとしては，李・長谷部・久保（2016）のものを利用するが，本稿の執筆のために 1000 件のデータを無作為で抽出して使用する．使用データは，http://jhlee.sakura.ne.jp/data/rb.sav として公開しているので，実際に試してみたい読者はダウンロードしてほしい．また，本稿は，Windows 8 環境および SPSS バージョン 22 に基づいて説明していく．なお，SPSS はほぼ 1 年単位でバージョンアップを行っており，2016 年 4 月時点での最新版はバージョン 24 である（http://www-01.ibm.com/software/jp/ analytics/spss/ 参照）．

## ２．SPSS 入門

説明の都合上，SPSS の基本構成について簡単に説明しておきたい．まず，rb.sav を SPSS バージョン 22 で起動させると，図 1 の画面が出る．


図 1．データファイルの例

SPSS のデータファイルは，表形式になっており，Excel の *.xls および *xlsx 形式や *.csv，*.txt，さらには Lotus の *.w* ファイルなどと互換性を持っているので，データのシームレスなインポートやエクスポートが可能である．図 1 からも分かるように，一見すると Excel に似ているが，操作方法や機能という意味では，全く別次元のソフトウェアと見るべきであろう．Excel と SPSS の操作でもっとも異なっている点としては，データファイルと分析の出力ファイルが別ファイルとして管理される点である．図 1 に示したのは，「*.sav」という拡張子がついているデータファイルであるが，SPSS で行うグラフ描画を含むすべての分析は図 2 の出力ドキュメント（拡張子は「*.spv」）の中に記録・管理されることになる．

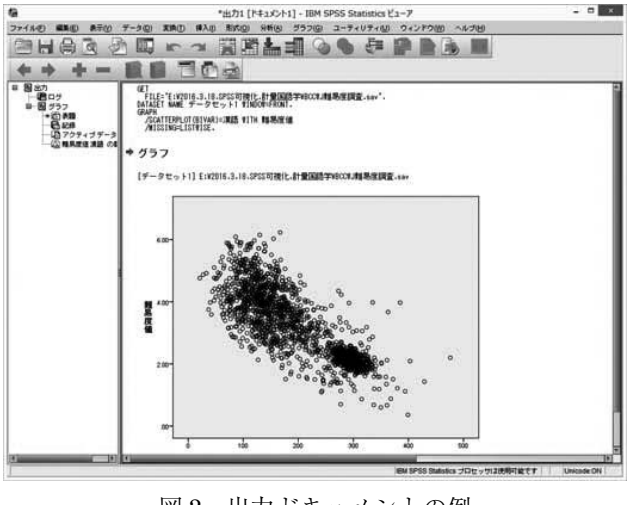


図 2．出力ドキュメントの例

図 1 の sav ファイルでは，データ入力および各種統計分析が実行できるほか，分析データに対する様々な操作が行える．例えば，尺度タイプ（順序，名義，スケール）の定義や欠損値の定義，ファイルの分割，データ変換など，データ分析をサポートする強力な機能が備えられている．図 2 の出力ドキュメントでは分析結果を確認するというのが主目的であるが，グラフ・エディターなどのツールを起動し，サイズや書式変更などの編集作業ができる．

## 3．データファイルの説明

説明に使用するデータについて説明する．李・長谷部・久保（2016）は日本語の生テキストの文章難易度を調査したもので，rb.sav には，ウェブのテキストとして，Yahoo! 知恵袋のデータ，書籍のテキストとして，BCCWJ の書籍コーパスのデータ，メディア系文章を代表するテキストとして，読売新聞記事のデータが入っている．すべてのデータは，1 テキスト，約 1,000 文字になるように調整してある．そして，rb.sav には，形態素解析がすんだテキストファイルに対して，Lee&Hasebe(in press) が提案する「日本語教育のリーダビリティ公式」（X={ 平均文長 *-0.056}+{ 漢語率 *-0.126}+{ 和語率 *-0.042}+{ 動詞率 *-0.145}+{ 助詞率 *-0.044}+11.724）（R2=.896）で計算した文章難易度が入っている．

文章難易度は「日本語文章難易度判別システム（http://jreadability.net/）」でも計算することができ，0.5 から 6.4 までの値をとる．0.5 に近ければ近いほど難しい文章，6.4 に近ければ近いほど易しい文章という解釈になる．具体的な対応は以下の表のようになる（http://jreadability.net/q_and_a から抜粋）．

表 1．文章難易度の日本語教育レベル

| 難易度値 | 難易度の目安 |
|---|---|
| 0.5～1.4 | 日本語学習者の上級後半レベルの文章 |
| 1.5～2.4 | 日本語学習者の上級前半レベルの文章 |
| 2.5～3.4 | 日本語学習者の中級後半レベルの文章 |
| 3.5～4.4 | 日本語学習者の中級前半レベルの文章 |
| 4.5～5.4 | 日本語学習者の初級後半レベルの文章 |
| 5.5～6.4 | 日本語学習者の初級前半レベルの文章 |

rb.sav の変数構成は，以下のとおりである．

表 2．rb.sav の変数構成

| No. | 変数名 | 尺度 | 説明 |
|---|---|---|---|
| 1 | 出典 | 名義 | 文章のジャンル情報を定義した変数 |
| 2 | 漢語 | スケール | 漢語の使用頻度 |
| 3 | 難易度値 | スケール | 文章の難易度値 |
| 4 | ガイドライン | 名義 | 文章の難易度値を解釈したもの |

rb.sav には 4 つの変数が入っている．ヒストグラムと箱ひげ図では，1 と 3 の変数を使用する．散布図では，2 と 3 の変数を使用する．4 は，3 の難易度値を表 1 にそって解釈したもので，例えば，難易度値が 4.59 に対して「初級後半」という情報が入っている．

## 4．SPSS でグラフ描画

計量国語学に限ったことではないが，統計分析にとって重要な目的の一つはデータの分布を見るという作業である．足し算や平均を求めるといった操作は，統計分析の世界では本質的な作業ではない．データの分布を見ることこそが重要とされるため，分散や標準偏差といった指標をもとに分布を見たり，さらには様々な統計モデルを使って分析したりするのである．SPSS では，こうしたデータの分布を視覚化し，より直感的に理解することを意図したグラフオプションが多数用意されている．


図 3. SPSS のグラフオプション

図 3 では，SPSS バージョン 22 に収録されているグラフオプションを示しているが，グラフの種類そのものは，一般的な表計算ソフトに搭載されているものと大きな差はない．しかし，SPSS の場合，統計分析と連動する形で，グラフ表示が最適化される点と単一の操作ですべての視覚化作業が行える点が，大きなメリットと言える．

SPSS を使ってグラフを作成する方法は，2 つある．1）統計分析に付随するオプションとしてグラフ描画する方法，2）グラフ描画のための専用オプションを使う方法である．1）の方法の例としては，「分析＞記述統計＞度数分布表」コマンドで，「度数分布表」作成を作成すると同時に，円グラフやヒストグラムを作成することができる．または，「分析＞平均の比較＞一元配置分散分析」コマンドで，「一元配置分散分析」を行うのと同時に，平均値の線グラフを作ることができる．本稿は，データの視覚化に注目するものであるため，2）のグラフ専用のツールを使い，SPSS によるグラフ描画の特徴を説明する．

グラフ描画のための専用オプションを使う場合，すべての操作は規格化されている．具体的には以下の 3 ステップの操作を行う．

・ステップ 1:「グラフ＞レガシーダイアログ＞…」の中からグラフを選択する.
・ステップ 2: ダイアログボックスで，描画する変数を指定する.
・ステップ 3: 出力ドキュメント上に表示されるグラフを確認し，必要に応じて，グラフ・エディターで編集する.

上記の 3 ステップの例として，rb.sav の難易度値変数に対して，ヒストグラムを作成してみる.

図 4　(a) ステップ 1

図 4　(b) ステップ 2

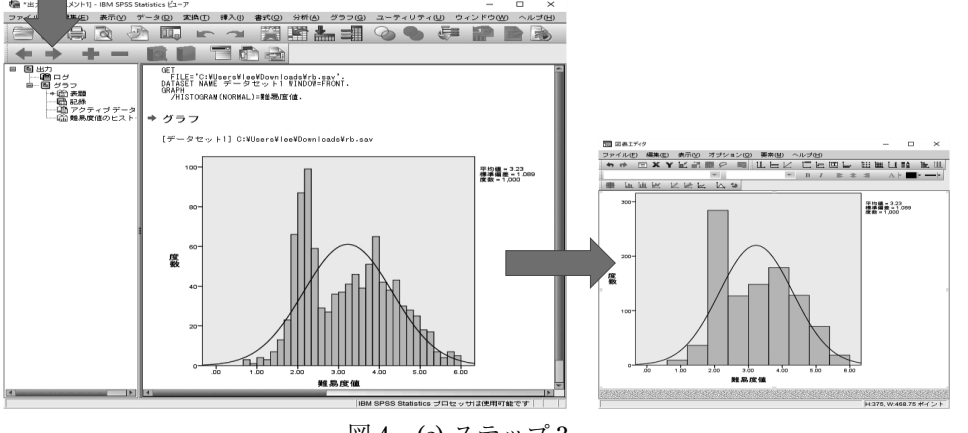

図 4　(c) ステップ 3

図 4 の (a) で「グラフ＞レガシーダイアログ＞ヒストグラム」を選択し，(b) で「難易度値」という変数を選択する．そして，「OK」ボタンをクリックすると，(c) のとおり，ヒストグラムが表示される．(c) に表示されているヒストグラムをダブルクリックするとグラフ・エディターが表示され，書式などの編集ができる．ステップ 3 まで完了したグラフは，コピー＆ペーストでオフィス系のソフトウェアに転写することもできるし，グラフ単体で *.PNG，*.JPG，*.BMP や *.EPS などの形式に変換・保存することもできる．

4.1 節ではヒストグラム，4.2 節では箱ひげ図，4.3 節では散布図を取り上げる．

## 4.1　ヒストグラム

ヒストグラムは，縦軸に度数，横軸に階級をとり，データの全体的な分布を把握するためのグラフである．本稿のサンプルデータの rb.sav を利用し，文章の難易度のヒストグラムを作成する．

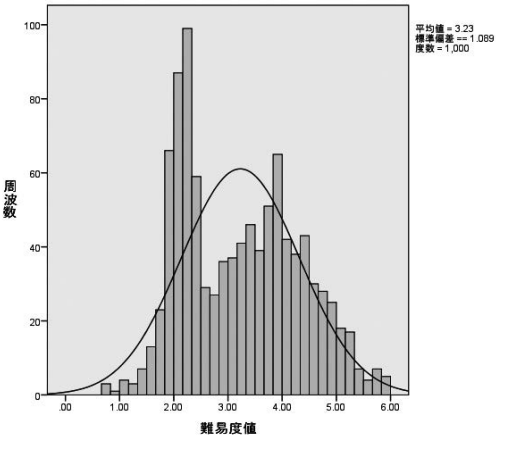


図 5．難易度値ヒストグラム

図 5 は，rb.sav の全データの難易度値変数に対する正規曲線付きのヒストグラムである．グラフの右上には平均値と標準偏差と度数が表示されている．図 5 の棒が高くなっているところ（峰）の値にデータが集中していることを示し，棒が低い部分の値をとったデータはあまりないことを示している．

SPSS では，ヒストグラムの階級幅や度数の間隔は入力データに応じて最適化がなされるため，デフォルトの出力においても十分に妥当なものになるが，カスタマイズをすることもできるようになっている．図 5 に関して言えば，階級幅がやや狭く，全体の特徴を捉えにくい部分があるので，図 6 のように階級幅を調整することもできる．

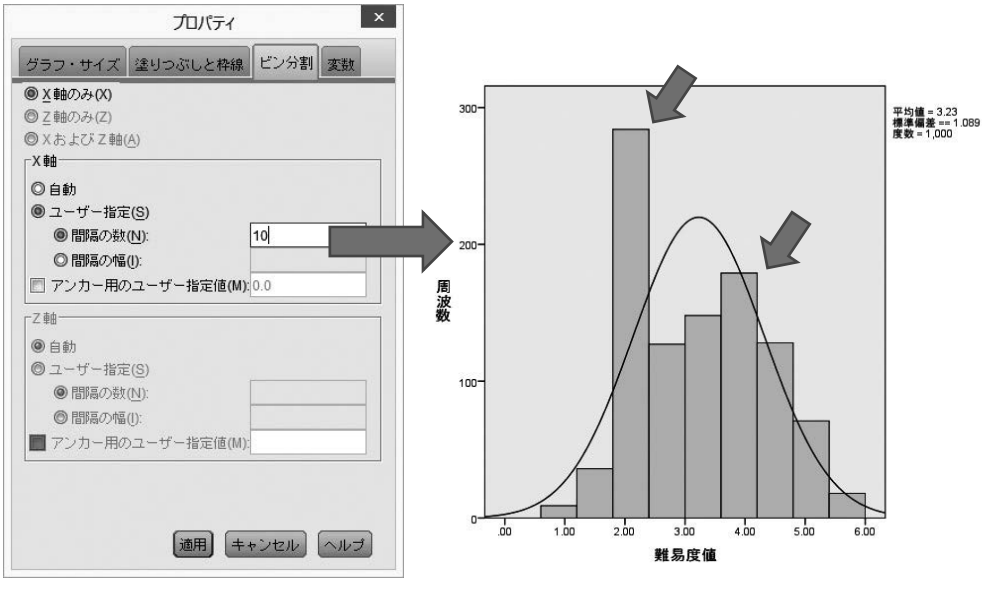


図 6. 間隔の数を 10 に設定した場合のヒストグラム

ヒストグラムの解釈ポイントとして，1) ピークがいくつあるか，2) 対称的がどうかである．1) として，ヒストグラムをみて，ピークが 1 つか，複数のピークが存在するかをチェックする．1 つのみのピークで構成されたものを「単峰な分布」とよび，複数のピークで構成されたものを「多峰な分布」と呼ぶ．平均値などの代表値を考えることに実質的な意味があるのは，山がひとつである場合に限られる．「多峰な分布」の場合は，質の違うものが混じっていることを意味するので，データを分けて集計するといった作業を行う．次に，2）の対称性の問題は，ピークの部分を真ん中にした場合，左右が対称かどうかにかかわる問題である．左右が対称ということは，データの大きくなるなり方と小さくなるなり方が同じであることを意味し，分散や標準偏差をばらつきの尺度として利用して問題ないということになる．一方，対称でない分布の場合には，ずれ方を考慮した分析が必要になるが，統計分析の議論になるため，詳細は涌井（2013）などを参照してほしい．

さて，図 5 および図 6 のデータに戻った場合，ピークが 2 か所において観察されるため，「多峰な分布」であることが明らかになった．そこで，データをジャンル別に分けて集計する．具体的には表 2 の出典変数を使って，ウェブのデータと書籍のデータと新聞のデータを分けてグラフ描画する．Excel の場合，範囲を指定し，ジャンルの数だけ，グラフ描画をすることになるが，SPSS では，ジャンルを定義した変数を使い，データを分割し，分析するという流れになる．ここでは，「データ＞ファイルの分割」オプションを使用し，「出典」変数をもとにデータを分ける．その上で，同じ手順でヒストグラムを作成する（図 7）．

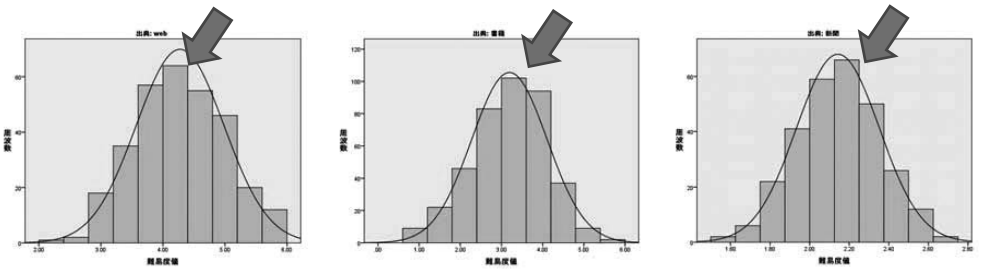

図 7. ジャンル別難易度値のヒストグラム

一度の操作で，3 つのジャンルにおける難易度値のヒストグラムが表示される．ジャンル情報は，ヒストグラムの上段に表示されている（例：出典：新聞）．図 7 からいずれのデータも単峰型であることが確認できる．

### 4.2　箱ひげ図の描画

前節のヒストグラムは基本的には正規分布のように左右対称の分布を想定しているが，実際のデータでは必ずしも左右対称でない場合がある．こうした時に役立つのが箱ひげ図である．特に集団同士を比較したり，極値を検出したりする時に，箱ひげ図は便利である．

SPSS では，カテゴリ軸と変量を指定するだけで図 8 のような箱ひげ図が描画できる．

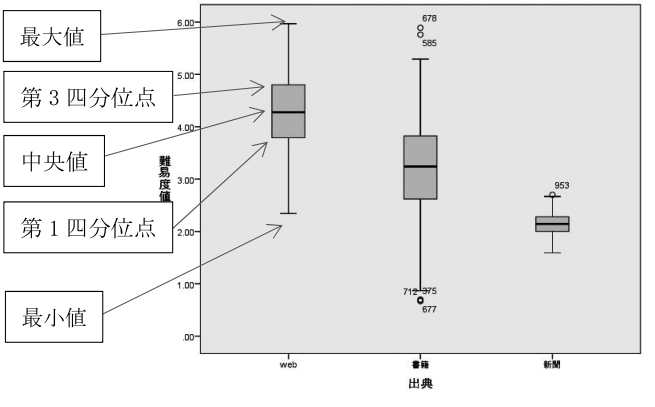


図 8. ジャンル別の箱ひげ図

図 8 は，rb.sav の表 2 の出典をカテゴリ軸に，難易度値を変数にした箱ひげ図である．箱ひげ図では，5 種の要約統計量「最小値，第 1 四分位点，中央値，第 3 四分位点，最大値」を同時に表現できる．図 8 の web のデータの場合，最大値が 5.97 であり，最小値は 2.34，中央値は 4.27，第 1 四分位点は 3.78，第 3 四分位点は 4.8 になる（数値の解釈は表 1 参照）．

近年，箱ひげ図が広く使われるようになり，統計ソフトウェアによっては特徴をつかみやすくするため，様々な工夫を行っている．とりわけ SPSS では，外れ値を考慮したグラ

フ作成ができる．というのは，箱ひげ図は外れ値があると，最大値と最小値が大幅に伸びてしまい，ひげの形が変わってしまうからである．SPSS では，次の方法で，外れ値を定義している．それは，中央の箱の長さを $H$ とした場合，箱の上限と下限から $H$ の 1.5 倍以上，離れているデータはすべて外れ値として描画する．そして，この外れ値を除いたデータを使って，ひげの部分を決めるという方法がとられている．図 8 で言えば，○で表現されているケース（書籍で 5 件，新聞で 1 件）がそれに該当する．○に書いてある番号は，データの行番号であるため，どのデータが外れ値であるかが分かるようになっている．なお，図 8 にはないが SPSS では，2 段階の外れ値を設定している．$H$ の 1.5 倍から 3 倍までの外れ値は「○」で，3 倍以上，離れているデータは「＊」で表示する仕様になっている．

図 8 の箱ひげ図を使うことで，全体の値のばらつき具合がひと目で判断でき，データ同士の比較も容易になるという利点がある．具体的には，新聞の場合，第 1 四分位点と第 3 四分位点のいずれも中央値の 2.1 に非常に近いため，箱の長さすなわち四分位範囲が短く，最大値と最小値の幅を示すひげも短い．すなわち，データのばらつきが少ないことが分かる．一方，書籍の場合，中央値 3.2 に対して，第 1 四分位点は 3.8，第 3 四分位点は 2.6 で，箱も長い上に，ひげも長く，データのばらつきが大きいことが分かる．また中央値を比較することでウェブ，書籍，新聞の順で難易度が上がっていくことが確認できる．

### 4.3　散布図の描画

縦軸と横軸にデータをプロットしたグラフのことを散布図というが，2 変量間の相関関係を分析する目的でよく用いられる．特に多変量解析をする前段階のデータ分析として散布図描画は基本的な作業ステップの一つであり，言語の計量的研究においても多く用いられる．

SPSS の散布図オプションでは，Y 軸と X 軸，マーカーの設定（任意）をするだけで図 9 のような散布図が作成できる．

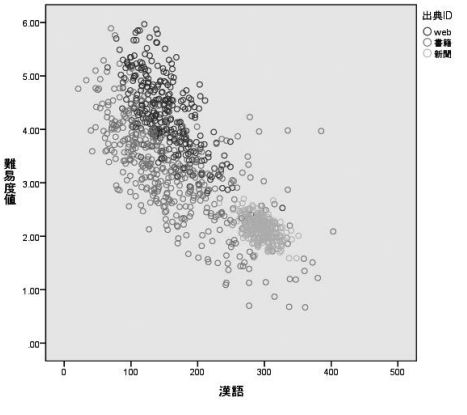


図 9．難易度値と漢語の使用頻度の散布図

図 9 は，縦軸に難易度値，横軸に漢語の出現頻度，マーカーとして出典を配置した散布図である．SPSS のデフォルトの表示では，すべてのケースは色分けされた◯記号で表示されるが，編集ツール「グラフ・エディター」を使って変えることもできる（図 10）．

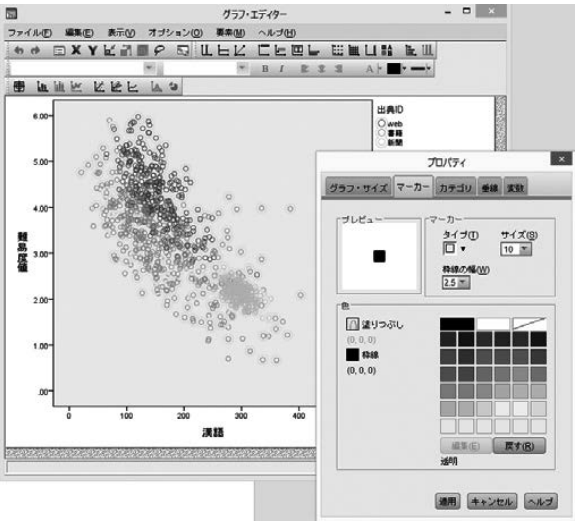


図 10. グラフ・エディターを使った散布図の編集

図 10 のグラフ・エディターでは，サイズや文字の配置，マーカーの書式，カテゴリーの調整などができ，図 11 のような編集ができる．

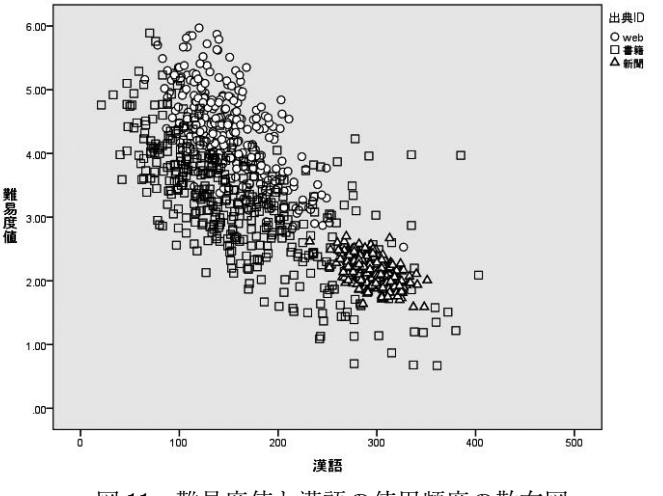


図 11. 難易度値と漢語の使用頻度の散布図

図 11 で，易しい文章（難易度値 6）から難しい文章（難易度値 1）に進むにつれ，漢語の使用頻度が増えている傾向が確認できる．また，テキストジャンルに注目した場合，ウ

ェブ（凡例の ”◯”）や書籍の文章（凡例の ”□”）は漢語の使用率が広範囲に散らばっているのに対して，新聞の場合（凡例の ”△”），250〜350 の間に集中して分布していることが確認できる．

## ５．最後に

本稿では，SPSS を使ったデータの視覚化として，3 つのグラフ描画機能について解説した．計量言語学の研究においてニーズの高いヒストグラム，箱ひげ図，散布図である．紙面の都合上，SPSS そのものについての説明は最小限にせざるを得なかったが，SPSS については，優れた解説書が多数，出版されているので，そちらを参考にして自分のものにしてほしい．

## 文献

李在鎬・長谷部陽一郎・久保圭「日本語コーパスの文章難易度に関する大規模調査の報告」（日本語教育学会ハンドアウト /2016.5.21）（http://jhlee.sakura.ne.jp/papers/nkg2016.pdf）

Lee, Jae-ho & Yoichiro Hasebe（in press）Readability Measurement for Japanese Text Based on Leveled Corpora, *Papers on Japanese Language from Empirical Perspective*. Znanstvena založba（http://jhlee.sakura.ne.jp/papers/lee-et-al2016rb.pdf）

涌井貞美（2013）『意味が分かる統計解析』ベル出版

（2016 年 3 月 30 日受付）

Mathematical Linguistics, Vol.30 No.5 (June 2016) pp.292-303.

*Tutorial*

# Data Visualization (5): Through the Use of the SPSS Graph Function

LEE Jae-Ho (Waseda University)

Abstract:
This paper gives a description of data visualization by SPSS and specifically focuses on the graph drawing function. It deals with three graphs, which are believed to be fundamental in the quantitative study of language. They are the histogram, boxplot, and scatter plot. The outcome of the research on text difficulty of corpus data conducted by Lee, Hasebe, and Kubo (2016) is used as data. The data used is posted on the Web so we expect SPSS users to experiment with it.

Keywords: SPSS, histogram, boxplot, scatter plot, readability