

人間行動システム統計演習 A

第 3 回：図表の作成（クロス集計、ヒストグラム他） 担当：中川、寺井、和嶋*

2015 年 4 月 23 日

東工大ポータルにログインし、Tokyo Tech OCWi から `shintai.csv` をダウンロードし、実習用フォルダの中に保存してください。

第 2 回授業で配布した `ten1.csv`、`ten2.csv` を読み込み、2 つのデータセットを縦につないだデータセット `ten` を作成してください。

1 度数分布・円グラフ・棒グラフ

1.1 度数分布表の表示

1 つのカテゴリ変数に関し、分布状況を調べる際に度数分布を作成します。「統計量」「要約」「頻度分布」をクリックし、対象となるカテゴリ変数を選択してください。table 関数を用いて、度数分布表と相対度数分布表が作成されます。

データセット `ten` の血液型に関し、度数分布表、相対度数分布表を作成してください。

1.2 円グラフを描く

1 つのカテゴリ変数に関し、分布状況をグラフ化する際に円グラフが用いられます。「グラフ」「円グラフ」をクリックし、対象となるカテゴリ変数を指定して下さい。pie 関数を用いて、棒グラフが作成されます。

1.3 棒グラフを描く

カテゴリの度数分布を棒グラフによって、図示することができます。「グラフ」「棒グラフ」をクリックし、対象となるカテゴリ変数を選択してください。必要に応じて、「X 軸ラベル」「Y 軸ラベル」「グラフタイトル」を指定してください。barplot 関数を用いて、棒グラフが作成されます。

R コマンドでは細かなオプションの指定は行えませんが、以下のようにスクリプトウィンドで `col` オプションを指定したコマンドの実行を行うことで、バーの色などを変更することができます。

* 御質問などは寺井 (terai.a.aa@m.titech.ac.jp) にメールにてご連絡下さい。

す。色の名前は、`colors()` で確認できます。数字や RGB で指定することもできます。ちなみに、デフォルトでは色の数字での指定は 1 から順に、黒、赤、緑、青、水色、紫、黄、灰となっています。「グラフ」「色パレット」で各数字と対応する色を変更することが可能です。

```
> barplot(table(ten$Blood), #血液型に関する棒グラフ
+ col = "pink", # バーの色
+ border = "red")
```

R のグラフィックス環境は強力で、その自由度も極めて高くなっていますが、R コマンドーでは限られたオプションの指定しかできません。演習中にすべてのオプションを網羅して取り扱うことは困難ですので、各自試してみてください。詳細は `help` が最も有益な参考になります。

層別平均値の棒グラフを表示する機能は R コマンドーにはありません。以下のように平均値に関するベクトルまたは行列を作成し、`barplot` 関数を用いて、棒グラフが作成することが可能です。以下の例では、エラーバーをつけるため別途標準偏差からなるベクトルを作成し、`arrows` 関数を用いて、棒グラフにエラーバーを追加しています。

```
> data <- c( #血液型ごとの英語の平均点からなるベクトル
+ mean(ten$Eigo[ten$Blood=="A"], na.rm=T),
+ mean(ten$Eigo[ten$Blood=="B"], na.rm=T),
+ mean(ten$Eigo[ten$Blood=="O"], na.rm=T),
+ mean(ten$Eigo[ten$Blood=="AB"], na.rm=T))
> B <- barplot(data, ylab="Eigo", #座標情報を変数 B に保存
+ xlab="Blood",ylim=c(0,100),names.arg = c("A", "B", "O", "AB"))
> data.sd <- c( #血液型ごとの英語の標準偏差からなるベクトル
+ sd(ten$Eigo[ten$Blood=="A"], na.rm=T),
+ sd(ten$Eigo[ten$Blood=="B"], na.rm=T),
+ sd(ten$Eigo[ten$Blood=="O"], na.rm=T),
+ sd(ten$Eigo[ten$Blood=="AB"], na.rm=T))
> arrows(B, data-data.sd, B, data+data.sd, angle = 90, length = 0.1)
> arrows(B, data+data.sd, B, data-data.sd, angle = 90, length = 0.1)
```

また、描画された図は、「グラフ」「グラフをファイルで保存」を選択することで保存可能です。(または、R のメニューから図を選択した状態で「ファイル」「別名で保存」)

演習

血液型ごとの社会の平均点からなる棒グラフを、バーの色を”skyblue”、Y軸の範囲を0~80点にて作成し、血液型ごとの社会の標準偏差をエラーバーとして追加してください。

2 分布の描画

すでに、numeric 型の変数に関する基本統計量（平均値、分散等）の求め方は紹介いたしました（「統計量」「要約」「アクティブデータセット」または「数値による要約」）。ここでは、データの分布状況を図示する方法について紹介します。

2.1 ヒストグラム

Rでヒストグラムを描画する際は、階級を事前に指定しなくとも階級を自動で設定してくれます。任意の区切りを設定することもできます。「グラフ」「ヒストグラム」をクリックし、データタブにて対象となる変数の一つを選択します。次に、オプションタブにて「頻度の計算」「パーセント」「密度」のいずれかを選択します（「頻度の計算」：各区間の度数（データ数）「パーセント」：各区間の相対度数（各区間に属するデータ数が全体に対して占める割合）「密度」：各区間の確率密度）。任意の区切りを設定する場合は、「区切りの数」で区関数を設定してください。

層別にヒストグラムを表示する場合は、「層別でプロット」をクリックし変数を選択してください。

R コマンドの設定で描画すると、階級幅の定義は、“~より大きい, ~以下”になります。階級幅の定義を“~以上, ~未満”にしたい場合は `right=FALSE` オプションを用いてください。`barplot` 関数と同様に、バーの色などを変更することも可能です。また、区切り幅はデフォルトでは『データの範囲を $\log_2 n + 1$ (n はデータの個数) 個の階級に分割して各階級に属するデータの数を棒グラフとして作図する』という Sturges (1926) の方法を用いています。

データに関し、関数 `density` 関数を用いて密度推定を行い、図示することもできます。「グラフ」「密度推定」をクリックし、データタブで対象となる変数を選択してください（オプションタブで推定方法を選択することができます）。

演習

国語の得点に関し「区切りの数」を5、10、15と変化させてヒストグラムを作成し、ヒストグラムの形状が異なることを確認してください。

2.2 箱ひげ図

「グラフ」「箱ひげ図」をクリックし、対象となるデータを選択します。層別に箱ひげ図を作成する場合は、「層別でプロット」をクリックし変数を選択してください。オプションタブで、外

れ値情報の表示方法を選択できます。

箱ひげ図の箱の中の太い線は中央値、箱の底辺は 25% 点 (Q1)、箱の上辺は 75%(Q3) の値です。箱の長さは、75% 点と 25% 点の差、すなわち四分位範囲の値を表します。箱の上下から延びる「ひげ」の長さは、箱から四分位範囲の 1.5 倍に設定されています。最小値、最大値が 1.5 倍範囲内に収まるときには、それらの値が上下のひげの端となります。上下のひげの範囲から外にはずれる値を「はずれ値」(異常値)といい、それらは (数字はオブザベーション番号)で示されます。

```
> boxplot(ten$Kokugo~ten$Blood,  
+ main = "Kokugo",           #図のタイトル  
+ range = 1,                 #ひげの設定  
+ horizontal = TRUE)  #箱ひげの向き
```

スクリプトウィンドで上記のようにオプションを指定して実行することでひげの長さなどを任意に変更したり、箱ひげの向きを変更することができます。range = は、ひげの設定をしています。ここでは、内側四分位範囲の 1 倍を指定しています。

3 クロス集計表

2 つのカテゴリ変数の度数分布表をクロス集計表と呼び、2 つのカテゴリ変数間の関係性を調べる際に用います。まずはじめに、KokugoC として、国語の点数に関し、D : 60 点未満、C : 60 点以上 70 点未満、B : 70 点以上 80 点未満、A : 80 点以上の 4 つの水準を持つカテゴリ変数を作成してください。新しく作成した変数を認識させるため「データ」「アクティブデータセット」「アクティブデータセットを新しくする」をクリックしてください。

```
> ten$KokugoC <- cut(ten$Kokugo, breaks = c(-Inf, 60, 70, 80, Inf),  
+ right = FALSE, labels = c("D", "C", "B", "A"),  
+ ordered_result = TRUE) 
```

KokugoC と Blood に関するクロス集計表を作成します。「統計量」「分割表」「2 元表」をクリックしてください。行の変数、列の変数をそれぞれ選択します。オプションタブで、パーセントの計算方法、仮説検定について選択することができます(仮説検定については今後の授業で説明します)。

演習

SansuC として、算数の点数に関し、L : 平均点未満、H : 平均点以上の 2 つの水準を持つカテゴリ変数を作成してください。その後、血液型ごとに KokugoC と SansuC のクロス集計表を作成してください(「統計量」「分割表」「多元表」を用いて下さい)。

4 二変数の関係をグラフ化する

2つの連続変数の関係をグラフ化する際は、散布図が用いられます。shintai.csvのデータは、生徒の身長 (height)、生徒の体重 (weight)、母親の身長 (MaHeight)、父親の身長 (PaHeight)、性別 (sex)、クラス (class) となっています。データセット shintai としてデータを読み込んでください。

生徒の身長と体重の関係を図示するため、それらをプロットし散布図を scatterplot 関数を用いて作成します。「グラフ」「散布図」をクリックしてください。データタブで X 変数 (weight)、Y 変数 (height) をそれぞれ指定します。オプションタブでオプションを設定できます。軸の目盛りは、対象としたデータの分布から自動的に調整されます。「層別でプロット」をクリックし、変数を指定することで層別でのプロットも可能です。

オプションについてですが、プロットする記号は 0~25 の数字で指定することが可能です (0 : 、1 : 、2 : 、3 : + 等と対応しています。詳しくは、(help(pch)) で確認してください。”文字”とすれば、文字そのものをシンボルとすることもできます。

シンボルの色は col オプションで指定できますが、コマンドではそのオプションには対応していません。指定したい場合は、スクリプトウィンドからコマンドを実行してください。層別でのプロットにおいて複数の記号を指定する場合は col=c(1,2) または col=c("red","black") といったように c 関数を用いて指定してください。

x (y) の値にゆらぎを与えて表示オプションは、jitter 関数で変数の値にゆらぎ (ノイズ) を加え、少しずつ点の位置をずらすためのものです。

5 その他のグラフ

- インデックスプロット：Y 軸を目的変数の値、X 軸をケース番号にしたグラフ
- 幹葉表示：簡易的なヒストグラム
- QQ プロット：データの分布が正規分布しているかどうかを目視で判断するためのグラフ (標準正規分布とデータの分位点との関係を図示したもので、これが、直線上に乗っているとデータが正規分布によくあてはまっている)
- 散布図行列：3つ以上の変数を選択する。各2変数に関する散布図を同時に表示可能
- 折れ線グラフ
- 条件付き散布図：層別に散布図を作成
- 平均のプロット：層別の平均値をプロット (オプションでエラーバーの種類を選択可能)
- ドットチャート：層別に値をプロット
- 3Dプロット

6 本日の課題

1. ten1.csv、ten2.csv を縦につなぎ合わせたデータを用いて、社会の得点 30 点未満を D、30 点以上 40 点未満を C、40 点以上 50 点未満を B、50 点以上を A とする評価をした上で、社会の評価と血液型の関係を示す表を作成する。
2. shintai.csv を用いて父親の身長 ("PaHeight") と生徒本人の身長 ("height") の関係を表すグラフを作図する。その際に、生徒の性別によりデータを区別して表示する。

作成したスクリプトファイルを OCWi から提出してください。課題の締め切りは次回の講義の前日です (5 月 6 日 23:59:59)。