医学情報処理演習第２回「基本的な図示」

中澤　港 (nminato@med.gunma-u.ac.jp)

2004 年 10 月 18 日

適切な図示の方法は，データの性質によって変わってくる。そのため，図示の方法に先立って，尺度と変数についてざっとさらっておく（以下，「R による統計解析の基礎」から抜粋して若干改変したものである）。

1　尺度と変数

尺度とは，研究対象として取り上げる操作的概念を数値として扱うときのモノサシの目盛り（の種類），言い換えると，「データに何らかの値を対応させる基準」である。尺度は，名義尺度，順序尺度，間隔尺度，比尺度（比例尺度ともいう）の４つに分類される。

研究対象として取り上げる操作的概念は，変数という形で具体化される。言い換えれば，変数とは，モノサシで測定された値につける名前である。変数は，それが表す尺度の水準によって分類されるが，一般には，名義尺度は定性的変数（カテゴリ変数），順序尺度，間隔尺度，比尺度は定量的変数に相当する。定量的変数には，整数値しかとらない離散変数と，実数値をとりうると考えられる連続変数がある。

前回説明したように，R の変数の型（str(変数名) とすれば表示される）には，int で表される整数型，num で表される数値型，Factor で表される因子型，character で表される文字列型などがあるが，整数型の変数は離散変数であり，数値型の変数は連続変数である。因子型や文字列型の変数はカテゴリ変数である。また，R では，as.*という形で，強制的に変数の型を変換することができる場合がある。変数の型によって，同じ関数でも動作が異なる場合が多いので，変数の型については注意が必要である。

順序尺度は離散変数，間隔尺度は離散の場合も連続の場合もあるが連続変数であることが多く，比尺度は連続変数である。定性的変数と離散変数の中には，１か０，あるいは１か２，のように，２種類の値しかとらない「２分変数 (dichotomous variable)」や，１か２か３，のように３種類の値しかとらない「３分変数 (trichotomous variable)」がある。変数がとり得る値の範囲を，その変数の**定義域**と呼ぶ。

変数は，被験者や研究対象のちがいによって，複数の異なったカテゴリあるいは数値に分かれるのでなければ意味がない。例えば，その研究のすべての対象者が男性であれば，性別という変数を作ることは無意味である。

対応する尺度の種類によって，変数は，図示の仕方も違うし，代表値も違うし，適用できる統計解析手法も違ってくる。尺度についてより詳しく知りたい方には，池田央『調査と測定』（新曜社）をお薦めする。

2 名義尺度 (nominal scale)

- 値の差も値の順序も意味をもたず，たんに質的データの分類基準を与える。
- 例えば，性別とか職業とか居住地とか病名は，名義尺度をもつカテゴリ変数である。
- 性別というカテゴリ変数は，例えば，男性なら”M”，女性なら”F”という具合に文字列値をとることもできるが，一般には男性なら 1，女性なら 2 というように，数値を対応させる。これは，第 1 回の講義で触れたとおり，コーディング（coding）と呼ばれる手続きである。
- 関心のある事象が，例えば血液中のヘモグロビン濃度のように，性別ばかりでなく，授乳や妊娠によって影響を受ける場合は，調査対象者を，男性なら 1，授乳も妊娠もしていない女性は 2，授乳中の女性は 3，妊娠中の女性は 4，という具合に，生殖状態（性別及び授乳，妊娠）という名義尺度をあらわす変数にコード化する場合もある。
- 名義尺度を表す値にはそれを他の値と識別する意味しかない。統計解析では，クロス集計表を作って解析する他には，グループ分けや層別化に用いられるのが普通である[*1]。

3 順序尺度 (ordinal scale)

- 値の差には意味がないが，値の順序には意味があるような尺度。
- 例えば，尿検査での潜血の程度について+++，++，+，±，− で表される尺度は，+の数を数値として，例えば 3, 2, 1, 0.5, 0 とコーディングしても，3 と 2 の差と 2 と 1 の差が等しいわけではなく，3 は 2 よりも潜血が高濃度に検出され，2 は 1 よりも高濃度だという順序にしか意味がないから，順序尺度である。
- 順序尺度を表す値は，順序の情報だけに意味があるので，変数の定義域が 3, 2, 1 であろうと，15, 3.14159265358979, 1 であろうと同じ意味をもつ。しかし，意味が同じなら単純な方がいいので，1 から連続した整数値を割り当てて，順位そのものを定義域にするのが通例である。同順位がある場合の扱いも何通りか提案されている。
- 本来は順序尺度であっても，もっともらしい仮定を導入して間隔尺度であるとみなす場合も多い。例えば，「好き」「普通」「嫌い」の 3, 2, 1 とか，「まったくその通り」「まあそう思う」「どちらともいえない」「たぶん違うと思う」「絶対に違う」の 5, 4, 3, 2, 1 などは本来は順序

[*1] より複雑な統計解析に使う場合は，ダミー変数として値ごとの 2 分変数に変えることもある。例えば，居住地という変数の定義域が {東京，長野，山口} であれば，この変数の尺度は名義尺度である。東京を 1，長野を 2，山口を 3 と数値を割り振っても，名義尺度であるには違いない。しかし，居住地という変数を無くして，代わりに，東京に住んでいるか（1）いないか（0），長野に住んでいるか（1）いないか（0），という 2 つのダミー変数を導入することによって，同じ情報を表現することができる。ダミー変数を平均すると，1 に当てはまるケースの割合になる性質をもつために，ダミー変数は多くの統計手法の対象になりうる。

尺度だが，等間隔であるという仮定をおいて間隔尺度として分析される場合が多い。質問紙調査などで，いくつかの質問から得られるこのような得点の合計によって何らかの傾向を表す合成得点を得ることが頻繁に行われるが，得点を合計する，という操作は各質問への回答がすべて等間隔であり，変数ごとの重みも等しいという仮定を置いているわけである。合成得点が示す尺度の信頼性を調べるためにクロンバックのα係数という統計量がよく使われるが，α係数の計算には平均や分散が使われていることからも，それが間隔尺度扱いされていることがわかる。

4 間隔尺度 (interval scale)

- 値の差に意味があるが，ゼロに意味がない尺度。[*2]
- 例えば，体温は間隔尺度である。体温が摂氏 39 度であることは，摂氏 36 度に比べて「平熱より 3 度高い」という意味をもつが，39/36 を計算して 1.083 倍といっても意味がない。
- 間隔尺度をもつ変数に対しては，平均や相関など，かなり多くの統計手法が適用できるが，意味をもたない統計量もある。[*3]

5 比尺度 (ratio scale)

- 値の差に意味があり，かつゼロに意味がある尺度。[*4]
- 例えば，cm 単位で表した身長とか，kg 単位で表した体重といったものは，比尺度である。予算額といったものも，0円に意味がある以上，比尺度である[*5]。

6 データの図示

データの大局的性質を把握するには，図示をするのが便利である。人間の視覚的認識能力は，パターン認識に関してはコンピュータより遥かに優れていると言われているから，それを生かさない手はない。統計解析は，いろいろな仮定をおいて理論構築されているので，ただソフトウェアの計算結果としての有意確率だけを妄信してしまうのは危険である。図示されたものをみれば，直感的

[*2] より正確に言えば，値の比に意味がない尺度ということになる。ただし，値の差の比には意味がある。

[*3] 例えば，標準偏差を平均値で割った値を%表示したものを変動係数というが，身長という変数でも，普通に cm 単位や m 単位やフィート単位で表した比尺度なら変動係数に意味があるが，100cm を基準とした cm 単位や，170cm を基準とした 2cm 単位のように間隔尺度にしてしまった場合の変動係数には意味がない。変動係数は，分布の位置に対する分布のばらつきの相対的な大きさを意味するので，分布の位置がゼロに対して固定されていないと意味がなくなってしまうのである。

[*4] より正確に言えば，値の比にも意味がある尺度ということになる。

[*5] 予算額には0円やマイナスが普通にありえるし，何%成長とか何%削減という扱いより絶対値の増減が問題にされる場合が多いので間隔尺度とすべきという見方もありうる。

なチェックができるので，仮定を満たしていない統計手法を使ってしまう危険が避けられる場合が多い。つまり，

統計解析前に図示は必須

であると心得よう。R で図示をした場合，最大の利点は，その図をベクトルグラフィックスとして加工したり再利用できることである。図を作った後で，pdf 形式あるいは jpg 形式，png 形式，tiff 形式などで画像として保存しておくことも可能だが，Windows 環境ならばメタファイル形式にしておくと再加工が容易である（Mac や Linux 環境なら postscript 形式がよいと思われる）。しかし，たくさんの図を作ったときは，ある程度まとめて管理できた方が便利だし，コメントもつけておく方が，再利用するときに役に立つと思われる。そこでお勧めしたいのは，図をメタファイルとしてプレゼンテーションソフトに貼り付けておくことである。Powerpoint があればそれでいいし，なくても，フリーソフトの OpenOffice.org を使えば，Impress という Powerpoint 互換（一部非互換）のソフトが使える。今回の演習でも，できた図は Powerpoint に貼り付けて管理することにしよう。

変数が表す尺度の種類によって，さまざまな図示の方法があるので，それをざっと示すことにする。

6.1 離散変数（名義尺度または順序尺度をもつ変数）の場合

- 度数分布図：値ごとの頻度を縦棒として，異なる値ごとに，この縦棒を横に並べた図である。離散変数の名前を X とすれば，R では `barplot(table(X))` で描画される。
- 積み上げ棒グラフ：値ごとの頻度の縦棒を積み上げた図である。R では

  ```
  fx <- table(X)
  barplot(matrix(fx,NROW(fx)),beside=F)
  ```

  で描画される。
- 帯グラフ：横棒を全体を 100 %として各値の割合にしたがって区切って塗り分けた図である。R では

  ```
  pc <- table(X)/NROW(X)
  barplot(matrix(pc,NROW(pc)),horiz=T,beside=F)
  ```

  で描画される。
- 円グラフ（ドーナツグラフ・パイチャート）：円全体を 100 %として，各値の割合にしたがって中心から区切り線を引き，塗り分けた図である。ドーナツグラフでは 2 つの同心円にして，内側の円内を空白にする。R では `pie()` 関数を用いる[*6]。

[*6] R-1.5 以前は `piechart()` 関数だったが置き換えられた

6.2 連続変数の場合

- ヒストグラム：変数値を適当に区切って度数分布を求め，分布の様子を見るものである。R では `hist()` 関数を用いる。
- 正規確率プロット：連続変数が正規分布しているかどうかを見るものである（正規分布に当てはまっていれば点が直線上に並ぶ)。R では `qqnorm()` 関数を用いる。
- 幹葉表示 (stem and leaf plot)：大体の概数（整数区切りとか 5 の倍数とか 10 の倍数にすることが多い）を縦に並べて幹とし，それぞれの概数に相当する値の細かい部分を葉として横に並べて作成する図。R では `stem()` 関数を用いる。ただしテキスト出力画面に出力されるため，グラフィックとして扱うには少々工夫が必要である。
- 箱ヒゲ図 (box and whisker plot)：データを小さい方から順番に並べて，ちょうど真中にくる値を中央値 (median) といい，小さい方から 1/4 の位置の値を第 1 四分位 (first quartile)，大きいほうから 1/4 の位置の値を第 3 四分位 (third quartile) という。縦軸に変数値をとって，第 1 四分位を下に，第 3 四分位を上にした箱を書き，中央値の位置にも線を引いて，さらに第 1 四分位と第 3 四分位の差（四分位範囲）を 1.5 倍した線分をヒゲとして第 1 四分位の下と第 3 四分位の上に伸ばし，ヒゲの先より外れた値を外れ値として○をプロットした図である。カテゴリによって層別した箱ヒゲ図を横に並べて描くと，大体の分布の様子と外れ値の様子が同時に比較できるので便利である。R では `boxplot()` 関数を用いる。
- ストリップチャート (stripchart)：2 群間で平均値を比較する場合などに，群ごとに大まかに縦軸での位置を決め，横軸には各データ点の正確な値をプロットした図（群の数によって縦軸と横軸は入れ換えた方が見やすいこともある)。R では `stripchart()` 関数を用いる（縦軸と横軸を入れ換えるには，`vert=T` オプションをつける)。
- 散布図 (scatter plot)：2 つの連続変数の関係を 2 次元の平面上の点として示した図である。R では `plot()` 関数を用いる。異なる群ごとに別々のプロットをしたい場合は `plot()` の `pch` オプションで塗り分けたり，`points()` 関数を使って重ね打ちしたりできる。点ごとに異なる情報を示したい場合は `symbols()` 関数を用いることができるし，複数の連続変数間の関係を調べるために，重ね描きしたい場合は `matplot()` 関数と `matpoints()` 関数を，別々のグラフとして並べて同時に示したい場合は `pairs()` 関数を用いることができる。データ点に文字列を付記したい場合は `text()` 関数が使えるし，マウスで選んだデータ点にだけ文字列を付記したい場合は `identify()` 関数が使える。
- レーダーチャート：複数の連続変数を中心点から放射状に数直線としてとり，データ点をつないで表される図である。それら複数の変数によって特徴付けられる性質のバランスをみるのに役立つ。1 つのケースについて 1 つのレーダーチャートができるので，他のケースと比較するには，並べて描画するか，重ね描きする。R では `stars()` 関数を用いる。

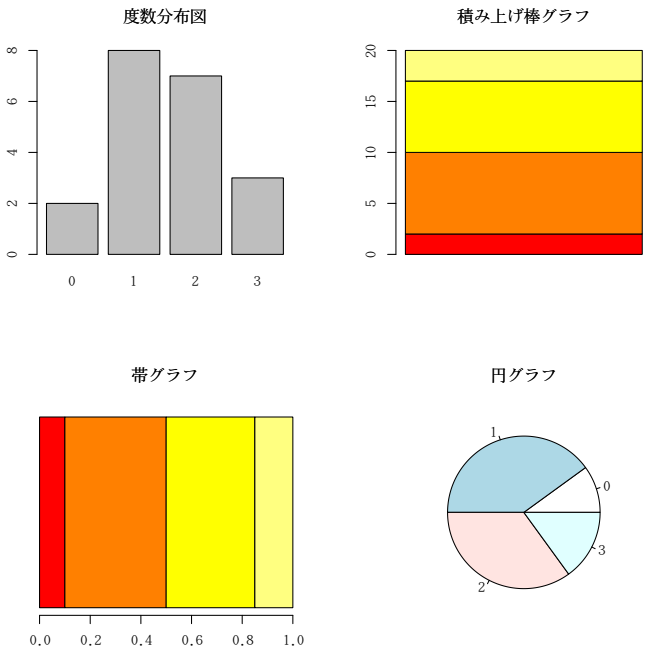

図 1: 離散変数の図示の例

7　離散変数の図示の例

20 組の夫婦について，その子ども数が，2, 3, 1, 0, 3, 2, 2, 1, 1, 1, 2, 2, 1, 3, 2, 1, 0, 2, 1, 1 だった場合，上で説明した図を描くための R のプログラムは下記の通り。

```
child <- c(2, 3, 1, 0, 3, 2, 2, 1, 1, 1, 2, 2, 1, 3, 2, 1, 0, 2, 1, 1)
fc <- table(child)
pc <- fc/sum(fc)
```

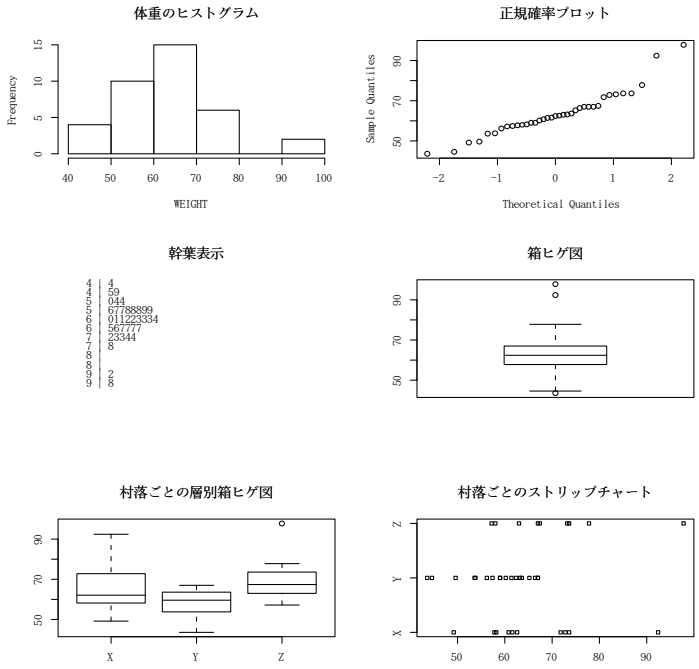

図 2: 連続変数の図示の例

```
par(mfrow=c(2,2))
barplot(fc, main="度数分布図")
barplot(matrix(fc,NROW(fc)), beside=F, main="積み上げ棒グラフ")
barplot(matrix(pc,NROW(pc)), horiz=T, beside=F, main="帯グラフ")
pie(pc, main="円グラフ")
```

8　連続変数の図示の例

http://phi.med.gunma-u.ac.jp/master/sample2.dat にある南太平洋の 3 つの村（X, Y, Z）の男性の身長と体重のデータを使えば，データが作業ディレクトリにある状態にしておいてから，

```
dat <- read.delim("sample2.dat")
attach(dat)
par(mfrow=c(3,2))
hist(WEIGHT, main="体重のヒストグラム")
qqnorm(WEIGHT, main="正規確率プロット")
stem.out<-capture.output(stem(WEIGHT,2))
plot(c(1,2),c(1,length(stem.out)),type="n",axes=F,xlab="",ylab="")
text(rep(1,length(stem.out)),length(stem.out):1,stem.out,pos=4)
title("幹葉表示")
boxplot(WEIGHT,main="箱ヒゲ図")
boxplot(WEIGHT~VG,main="村落ごとの層別箱ヒゲ図")
stripchart(WEIGHT~VG,main="村落ごとのストリップチャート")
```

とすれば，上で説明したそれぞれのグラフが描かれる（散布図とレーダーチャートは適用すべきデータの種類が違うので省略するが，別の回で説明する）。

課題

http://phi.med.gunma-u.ac.jp/medstat/p01.txt は，男性 50 人女性 50 人の（性別はカテゴリ変数 SEX で示されている），身長 (HT) と体重 (WT）のデータである。体重の分布を男女別に図示せよ。

結果は学籍番号を含め，A4 の紙に印刷し，氏名を自筆して提出すること。結果の提出をもって出席確認とする。