医学情報処理演習第２回「基本的な図示」 2005年10月17日 中澤　港

前回の課題の回答例

http://phi.med.gunma-u.ac.jp/medstat/p01.xls を読み込み，タブ区切りテキスト形式 p01.txt として保存し，R で p01 <- read.delim("p01.txt") としてから，str(p01) とすると，以下の出力が得られるので，オブザーベーションの数は 100（つまりサンプルサイズは 100）であり，各変数は，pid が int（整数型），sex が”F”，”M”という 2 水準をもつ Factor（要因型），ht と wt が num（数値型）であることがわかる。課題の回答としては，ここまででも良い。ただし，R は case sensitive，つまり，大文字と小文字が区別されることに注意せねばならない。PID という変数は pid とは別のものを指す。

```
'data.frame': 100 obs. of  4 variables:
 $ pid: int  1 2 3 4 5 6 7 8 9 10 ...
 $ sex: Factor w/ 2 levels "F","M": 1 2 1 1 2 1 2 2 1 1 ...
 $ ht : num  165 169 160 163 172 ...
 $ wt : num  61.3 65.7 57.6 62.9 58.3 55.2 70.2 60.6 60.3 58.9 ...
```

ここでさらに summary(p01) とすれば，以下が得られるので，NA の数までわかり，有効なデータ数は，pid と sex が 100，ht が 99（欠損 1），wt が 98（欠損 2）であることがわかる。

```
      pid         sex          ht              wt
 Min.   :  1.00   F:50   Min.   :150.6   Min.   :45.60
 1st Qu.: 25.75   M:50   1st Qu.:160.2   1st Qu.:58.08
 Median : 50.50          Median :165.0   Median :61.95
 Mean   : 50.50          Mean   :165.1   Mean   :61.93
 3rd Qu.: 75.25          3rd Qu.:169.7   3rd Qu.:66.40
 Max.   :100.00          Max.   :181.3   Max.   :76.60
                         NA's   :  1.0   NA's   : 2.00
```

さらにいえば，このように欠損が少ない場合は，１つでも欠損があるケースは解析から除いてしまうことも考えられる。その場合，

```
p01s <- subset(p01,!is.na(pid)&!is.na(sex)&!is.na(ht)&!is.na(wt),drop=T)
```

とすれば，欠損値を含まないサブセット p01s が得られるので，これに対して str(p01s) を行うことで，有効なサンプルサイズが 97 であるとわかる。もちろん，各変数の型はサブセットにする前と同じである。

尺度と変数

今回の演習の習得目標は，入力済みのデータについて，適切な図示を行うことであるが，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。

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，実数値をとりうる連続変数がある。

前回説明したように，R の変数の型には，int で表される整数型，num で表される数値型，Factor で表される要因型，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[*1]。要因型や文字列型の変数はカテゴリ変数である[*2]。同じ関数でも，変数の型によって動作が異なる場合が多いので，変数の型（及びその変換）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[*3]。

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。ここでは簡単にまとめるが，より詳しく知りたい方は，池田央『調査と測定』（新曜社）等の専門書を参照されたい。

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。
- 変数の型としては，文字列型か要因型になる。
- 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，前回触れたとおり，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，カテゴリごとの度数を求めたり，クロス集計表を作って解析するほかには，グループ分けや層別化に用いられるのが普通である[*4]。

[*1] `as.numeric()` を使えば，離散変数を数値型扱いすることは可能である。しかし，整数でない実数に対して `as.integer()` を用いて整数型にすると，小数点以下が切り捨てられて値が変わってしまう。

[*2] `as.ordered()` を使って順序型にすることもできる。

[*3] ただし，後に別のデータと併合することを考えて，敢えて作っておく場合もある。

[*4] ３つ以上のカテゴリをもつ変数を，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有無を示す複数の２分変数群に

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。
- 変数の型は，順序型 (Ord.factor) になる。R では，読んだだけで順序型と自動判定されることはないので，順序型変数を用いたい場合は，数値型または整数型としてデータ入力しておき，`as.ordered()` を使って型変換する。
- 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。3, 2, 1, 0.5, 0 とコーディングしておき，`as.ordered(o.blood)` のようにして型変換すべきである。
- 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（上の例のように元データに近い形にすることもある)。同順位がある場合の扱いも何通りか提案されている。
- 注意しなければならないのは，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得点化し，間隔尺度であるとみなす場合も多い，ということである。例えば「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 などは本来は順序尺度だが，等間隔な得点として扱われる場合が多い。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（たとえ調査者が意識していなくても，尺度構成をしていることになる)。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。

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。[*5]
- 変数の型は数値型か整数型である。
- 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。[*6]

変換することもある。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。ダミー変数は，平均値をとると，「1 に当てはまるケースの割合」と一致するため，本来なら量的な変数にしか使えないような多くの統計手法の対象になりうる。

[*5] 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。

[*6] 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

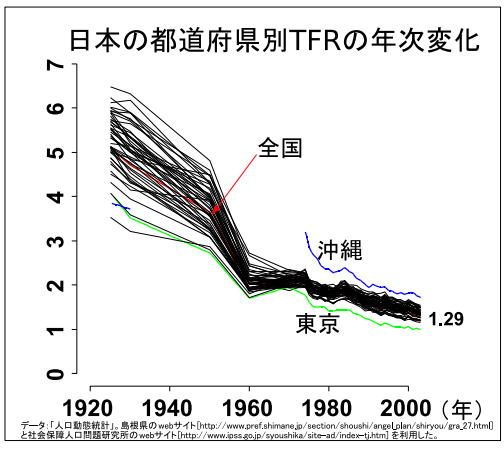
比尺度 (ratio scale)

- 値の差に意味があり，かつゼロに意味がある尺度。[*7]
- 変数の型は数値型または整数型であるが，数値型としておくべきである。
- 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[*8]。

データの図示

データの図示の目的は大別して２つある。１つは見せるためであり，もう１つは考えるためである。もちろん，両者の機能を併せもつグラフも存在するが，重視すべきポイントが変わってくるので，一般には，この２つは別のグラフになる。

見せるためのグラフでも，プレゼンテーションやポスターに使うグラフと，投稿論文に載せるグラフは，一般に別物である。前者は，１つのグラフに１つのことだけを語らせる必要があり，とにかくわかりやすさが最大のポイントであるのに対して，後者は複数の内容を語らせることも可能である。これは，見る人が１枚のグラフを見るために使える時間からくる制約である。例えば，日本の都道府県別 TFR（合計出生率）の年次変化のグラフを示すのに，プレゼンテーションならば左図のようにした方が見やすいが，論文に載せる場合は右図のようにする方が良い。いずれの場合も統計ソフトだけで仕上げるのは（不可能ではないが）面倒だし，管理上も不都合なので，プレゼンテーションソフト[*9]か描画ソフト[*10]に貼り付けて仕上げるのが普通である。



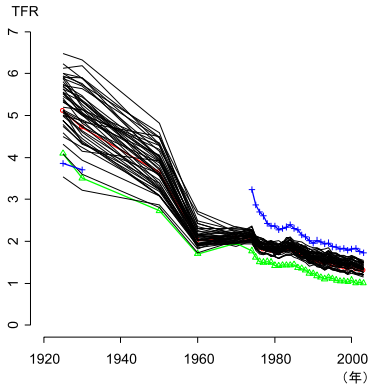


図2. 日本の都道府県別合計出生率(TFR)の年次変化
（△：東京，＋：沖縄，マークなし：他道府県，○：全国）．
データは「人口動態統計」に基づく．

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。

[*7] 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。

[*8] ただし，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る。

[*9] PowerPoint のほかには，フリーソフトの OpenOffice.org に含まれている Impress というものが有名であり，概ね PowerPoint と互換である。

[*10] OpenOffice.org に含まれている Draw も使える。もし使える環境にあれば，Adobe の Illustrator がよい（高価だが）。

見せるためのグラフについて詳しく知りたい方は，山本義郎 (2005)『レポート・プレゼンに強くなるグラフの表現術』講談社現代新書 (ISBN4-06-149773-1) を一読されることをお勧めする。時間的な制約もあるので，この演習では，今回は考えるためのグラフに絞って説明する。考えるためのグラフに必要なのは，データの性質に忠実に作るということである。データの大局的性質を把握するために，ともかくたくさんのグラフを作って多角的に眺めてみよう。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の数値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 (Macintosh 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。そのためにも，前述の通り，作った図は，メタファイルとしてプレゼンテーションソフトまたは描画ソフトに貼り付けておくことをお勧めする。

では，具体的な図示の方法に入ろう。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。

名義尺度または順序尺度をもつ変数の場合

要因型または順序型の変数についての作図は，カテゴリごとの度数を情報として使うことになる。そのため，作図関数に渡す値は一般にデータそのものではなく，その集計結果になる[*11]。もちろん，既に表の形になっている場合は，そのまま作図関数に渡すことができる。なお，以下の例の多くは，R のプロンプトで `source()` 関数を使って実行可能である[*12]。

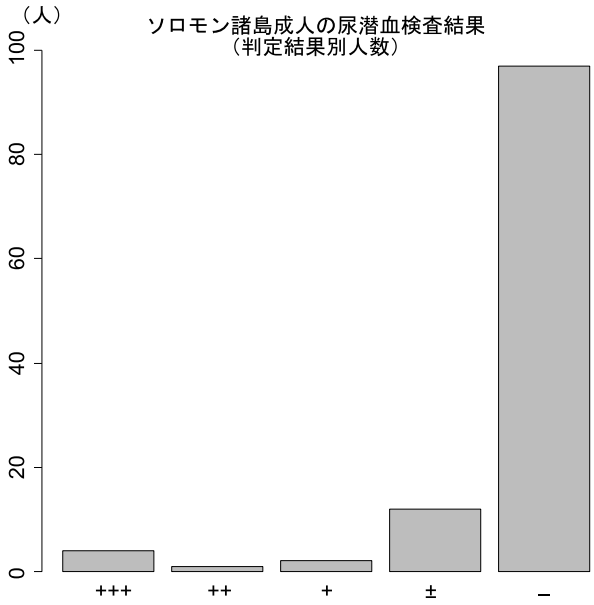
- 度数分布図：値ごとの頻度を縦棒として，異なる値ごとに，この縦棒を横に並べた図である。離散変数の名前を X とすれば，R では `barplot(table(X))` で描画される。例えば，ソロモン諸島の M 村の成人について尿検査をして，潜血の結果が，＋＋＋ が 4 人，＋＋ が 1 人，＋ が 2 人，± が 12 人，− が 97 人だったとしよう。これを度数分布図として棒グラフを作成するには，どうしたらいいだろうか。この例では，`table(X)` に当たる部分が既に与えられているので，下枠内のように，まずカテゴリ別度数を `c()` で与え，`names()` を使ってカテゴリに名前を付けてから，`barplot()` 関数で棒グラフを描画すればよい（以下の例では OpenOffice.org の Draw を使って加工済みのグラフを載せておく)。

[*11] `table()` 関数を使って度数分布を求め，その結果を作図関数に与えるのが普通である。

[*12] `source("http://phi.med.gunma-u.ac.jp/medstat/it02-3.R")` などとする。または，R Console の「ファイル」で「スクリプトを開く」を選び，ファイル名を入力するところに URL を打つと R Editor にコードを読み込めるので，その後，「編集」の「全て実行」を選んで実行させてもよい。

it02-3.R

```
ob <- c(4,1,2,12,97)
names(ob) <- c("+++","++","+","±","-")
barplot(ob,ylim=c(0,100),main="ソロモン諸島成人の尿潜血検査結果\n(判定結果別人数)")
```



なお，合計で割って縦軸を割合にした方が見やすい場合もある。

- 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。上のデータで積み上げ棒グラフを描くには以下のようにする。最初の 2 行は変わらない。残りは，`barplot()` の引数を変えることと，カテゴリ名を適切な位置に書き込むために必要な部分である。各自実行されたい。

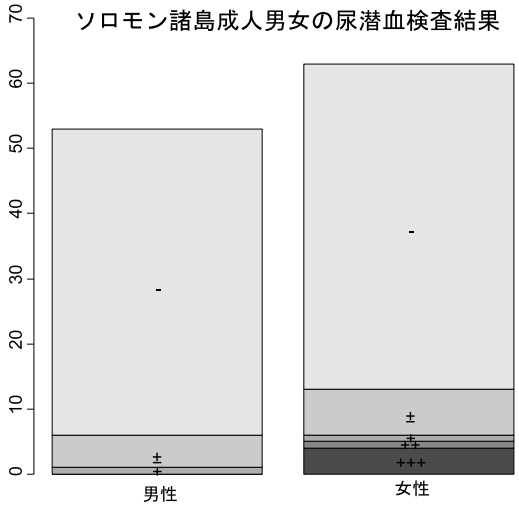
it02-4.R

```
ob <- c(4,1,2,12,97)
names(ob) <- c("+++","++","+","±","-")
ii <- barplot(matrix(ob,NROW(ob)),beside=F,ylim=c(0,120),main="ソロモン諸島成人
の尿潜血検査結果")
oc <- ob
for (i in 1:length(ob)) { oc[i] <- sum(ob[1:i])-ob[i]/2 }
text(ii,oc,paste(names(ob)))
```

積み上げ棒グラフは単独で用いるよりも，複数の積み上げ棒グラフを並べて比較するのに向いている。例えば，上の結果を男女別に見ると，男性では + + + と ++ が 0 人，+ が 1 人，± が 5 人，− が 47 人，女性では + + + が 4 人，++ が 1 人，+ が 1 人，± が 7 人，− が 50 人だったとき，男女別々に積み上げ棒グラフを描いて並べると，内訳を男女で比較することができる。実行するための R のコードは以下の通りである。

it02-5.R

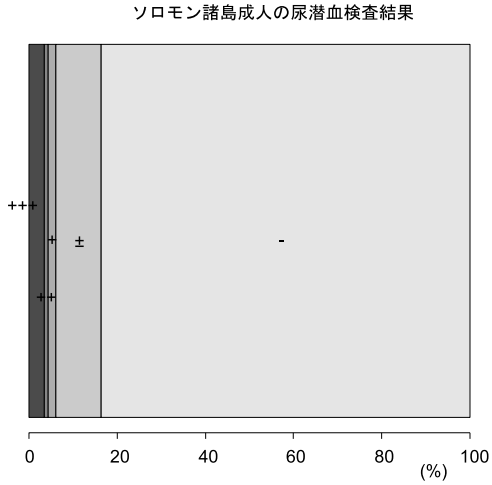
```
obm <- c(0,0,1,5,47)
obf <- c(4,1,1,7,50)
obx <- cbind(obm,obf)
rownames(obx) <- c("+++","++","+","±","-")
colnames(obx) <- c("男性","女性")
ii <- barplot(obx,beside=F,ylim=c(0,70),main="ソロモン諸島成人男女の尿潜血検査結果")
oc <- obx
for (i in 1:length(obx[,1])) { oc[i,1] <- sum(obx[1:i,1])-obx[i,1]/2 }
for (i in 1:length(obx[,2])) { oc[i,2] <- sum(obx[1:i,2])-obx[i,2]/2 }
text(ii[1],oc[,1],paste(rownames(obx)))
text(ii[2],oc[,2],paste(rownames(obx)))
```



- 帯グラフ：横棒を全体を 100%として各カテゴリの割合にしたがって区切って塗り分けた図である。内訳を見るのに向いている。複数並べて構成比を比べたいときに効果を発揮する。ソロモン諸島成人の尿潜血検査結果について帯グラフを描くための R のコードは下枠内の通り。2 行目で人数を構成割合 (%) に変える計算をしているのと，4 行目の `horiz=T` が重要である。

it02-6.R

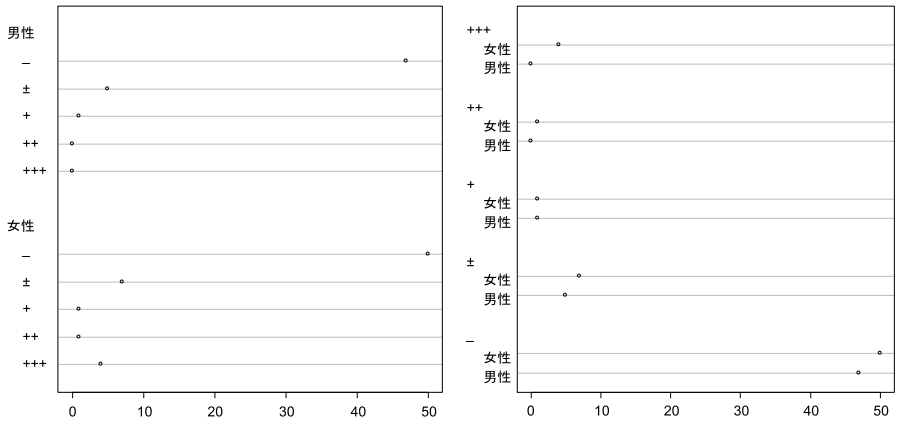
```
ob <- c(4,1,2,12,97)
obp <- ob/sum(ob)*100
names(obp) <- c("+++","++","+","±","-")
ii <- barplot(matrix(obp,NROW(obp)),horiz=T,beside=F,xlim=c(0,100),
 xlab="(%)",main="ソロモン諸島成人の尿潜血検査結果")
oc <- obp
for (i in 1:length(obp)) { oc[i] <- sum(obp[1:i])-obp[i]/2 }
text(oc,ii,paste(names(obp)))
```



- ドットチャート：棒グラフの棒を描く代わりに上端に点を打ったグラフである。複数のドットチャートを並列することもできる。基本的に `barplot()` の代わりに `dotchart()` を使えばよい。ソロモン諸島成人男女の尿潜血の例では，下枠内のコードを用いればよい。

it02-7.R

```
obm <- c(0,0,1,5,47)
obf <- c(4,1,1,7,50)
obx <- cbind(obm,obf)
rownames(obx) <- c("+++","++","+","±","-")
colnames(obx) <- c("男性","女性")
dotchart(obx)
dotchart(t(obx))
```



- 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，各カテゴリ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。構成比を見るには，帯グラフよりも直感的にわかりやす

い場合も多い[*13]。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[*14]。ソロモン諸島成人の尿潜血検査結果について円グラフを描かせる R のコードは下枠内の通りである。各自試されたい。

it02-8.R

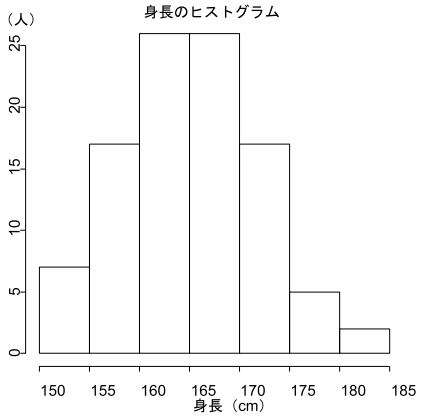
```
ob <- c(4,1,2,12,97)
names(ob) <- c("+++","++","+","±","-")
pie(ob)
```

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。Excel ではツールのアドインの分析ツールに含まれているヒストグラム作成機能で区切りも与えてやらないと作成できず，非常に面倒だが，R では hist() 関数にデータベクトルを与えるだけである，棒グラフとの違いは，横軸（人口ピラミッド[*15]のように 90 度回転して縦軸になることもある）が，連続していることである（区切りに隙間があってはいけない）。基本的に，区切りにはアプリオリな意味はないので，分布の形を見やすくするとか，10 進法で切りのいい数字にするとかでよい。対数軸にする場合も同様である。第 1 回に使用した身長と体重のデータ[*16]で，身長のヒストグラムを描かせるコードは下枠内の通りである。

it02-9.R

```
dat <- read.delim("http://phi.med.gunma-u.ac.jp/medstat/p01.txt")
hist(dat$HT,main="身長のヒストグラム")
```



[*13] ただし，認知心理学者の Cleveland は，その著書【Cleveland WS (1985) The elements of graphing data. Wadsworth, Monterey, CA, USA.】の p.264 で，「円グラフで示すことができるデータは，常にドットチャートでも示すことができる。このことは，共通した軸上の位置の判定が，正確度の低い角度の判定の代わりに使えることを意味する。」と，実験研究の結果から述べているし，R の help ファイルは，構成比を示すためにも円グラフよりもドットチャートや帯グラフあるいは積み上げ棒グラフを使うことを薦めているので，円グラフはむしろ「見せるためのグラフ」として使う際に価値が高いといえよう。

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

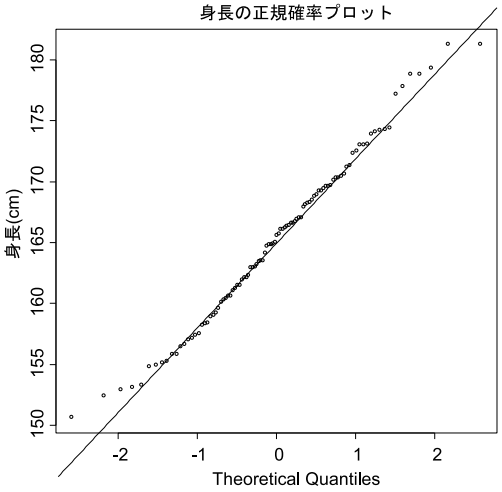
[*15] 詳しくは http://phi.med.gunma-u.ac.jp/demography/makepyramid.html を参照。

[*16] http://phi.med.gunma-u.ac.jp/medstat/p01.txt として公開してある。

- 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グラフである。正規分布に当てはまっていれば点が直線上に並ぶし，ずれていればどのようにずれているかを読み取ることができる。ヒストグラムに比べると，正規確率プロットから分布の様子を把握するには熟練を要するが，区切りの恣意性によって分布の様子が違って見える危険がないので，ヒストグラムと両方実施すべきである。ヒストグラムで示したのと同じデータについて正規確率プロットを描かせるには，下枠内を打てばよい。`qqnorm()` で正規確率プロットが描かれ，`qqline()` で，もしデータが正規分布していればこの直線上に点がプロットされるはず，という直線が描かれる。

it02-10.R

```
dat <- read.delim("http://phi.med.gunma-u.ac.jp/medstat/p01.txt")
qqnorm(dat$HT,main="身長の正規確率プロット",ylab="身長 (cm)")
qqline(dat$HT,lty=2)
```



- 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。ヒストグラムを 90 度回転して数字で作るようなものだが，各階級の内訳がわかる利点がある。身長の例では，下枠内のように打てばテキスト画面に幹葉表示が得られる[*17]。

```
dat <- read.delim("http://phi.med.gunma-u.ac.jp/medstat/p01.txt")
stem(dat$HT)
```

- 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

[*17] `source("http://phi.med.gunma-u.ac.jp/swtips/gstem.R")` としてから `stem()` の代わりに `gstem()` を使えば，図形としての出力が得られる。

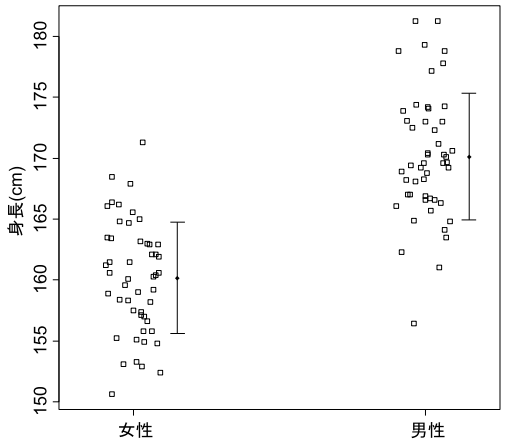
値を外れ値として○をプロットした図である。カテゴリによって層別した箱ヒゲ図を横に並べて描くと，大体の分布の様子と外れ値の様子が同時に比較できるので便利である。R では boxplot() 関数を用いる。身長データだと下枠内を打てばよい。

```
dat <- read.delim("http://phi.med.gunma-u.ac.jp/medstat/p01.txt")
boxplot(dat$HT)
```

- 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)。横に平均値と標準偏差の棒を付加することも多い。身長データを男女間で比較するためのコードの例を下枠内に示す。

it02-11.R

```
dat <- read.delim("http://phi.med.gunma-u.ac.jp/medstat/p01.txt")
attach(dat)
mHT <- tapply(HT,SEX,mean)
sHT <- tapply(HT,SEX,sd)
IS <- c(1,2)+0.15
stripchart(HT~SEX,method="jitter",vert=T,ylab="身長 (cm)")
points(IS,mHT,pch=18)
arrows(IS,mHT-sHT,IS,mHT+sHT,code=3,angle=90,length=.1)
detach(dat)
```



- 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[*18]，複数の連続変数間の関係を調べるために，重ね描きした

[*18] 使用例は，http://phi.med.gunma-u.ac.jp/medstat/semen.R を試されたい。

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。もっとも基本的な使い方として，身長と体重の関係を男女別にマークを変えてプロットするなら，下枠内のようにする。

```
dat <- read.delim("http://phi.med.gunma-u.ac.jp/medstat/p01.txt")
plot(dat$HT,dat$WT,pch=paste(dat$SEX),xlab="身長 (cm)",ylab="体重 (kg)")
```

- 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() 関数を用いるが，詳細は省略する。

その他のグラフ

以上説明した基本的なグラフ以外にも，maptools ライブラリと ESRI 社が公開しているデータを使えば，GIS のように統計情報によって地図を塗り分けることができるし[*19]，遺伝子データなどを使った系統関係を示す樹状図（デンドログラム）は，クラスタ分析として描くこともできるし，生存時間データについては生存関数を描くこともできるなど，目的に応じて多様なグラフがある。

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。R 上でできた図をパワーポイント，ワード，ペイントなどに貼り付けてから学籍番号を記入し，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。

[*19] 詳細は http://phi.med.gunma-u.ac.jp/swtips/EpiMap.html 参照。