

第3章

データの尺度・データの図示

3.1 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^{*1}

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。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。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い、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00、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、複数の異なる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。たとえば、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(1980)をお薦めする。

^{*1} これは Stevens が提唱した分類だが、絶対尺度を加えて5つとする分類もある。囲み解説参照。

尺度

より厳密には次の通り。非空の集合 A の要素間にいくつかの関係 R_1, R_2, \dots, R_n が成り立っているときに、これを $\alpha = < A, R_1, R_2, \dots, R_n >$ と書くことにし、数量的な要素からなる非空の集合 B の要素間に関係 S_1, S_2, \dots, S_n が成り立っているときに、これを $\beta = < B, S_1, S_2, \dots, S_n >$ と書くとき、もし B の中の要素が A の中のすべての要素 $x, y (x, y \in A)$ の写像 $f(x), f(y)$ からなり ($f(x), f(y) \in B$)、 x と y の間に関係 R_1, R_2, \dots, R_n が成り立っているときに B の中の $f(x)$ と $f(y)$ の間に関係 S_1, S_2, \dots, S_n が成り立っている（これを準同型という）ならば、関係系 α は関係系 β によって“表現される”，という。測定とは、経験的世界の関係系 α が数量的な形式関係系 β によって表現されることをいう。尺度とは、このような $< A, B, f >$ の組である。 B の各要素に変換 ϕ を施して得られる集合の要素を考えるとき、それがやはりもとの経験的関係系 α を表現しているなら、変換 ϕ はもとの表現 f に対して許容的であるといい、尺度は、許容的な変換の型が、それ自身のみであるか（絶対尺度）、正の実定数との積であるか（比例尺度）、正の実定数との積に定数を加えた一次変換であるか（間隔尺度）、単調関数なら何でも良いのか（順序尺度）、1 対 1 対応の写像なら何でも良いのか（名義尺度）、によって 5 種類に分けられる。

3.2 名義尺度 (nominal scale)

- 値の差も値の順序も意味をもたず、たんに質的なデータの分類基準を与える。
- たとえば、性別とか職業とか居住地といったものは、名義尺度である。
- 性別という名義尺度をあらわす変数は、たとえば、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 章で触れたとおり、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、たと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*2。

*2 このようにコーディングのやり方は一通りに限ったものではなく、分析の目的によって多様である。

- 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、クロス集計表を作つて解析する他には、グループ分けや層別に用いられるのが普通である。

ダミー変数

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。たとえば、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と数値を割り振つても、名義尺度であるには違ひない。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に当たる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。

3.3 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。
- たとえば、鉱物の強度、地震の震度、尿検査でのタンパクの検出の程度について+++、++、+、±、-で表される尺度^{*3}, 「好き」「普通」「嫌い」に3, 2, 1点の得点を割り付けた尺度などは、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、順序の情報としては、1から連続した整数値を割り当てるのが通例であり（同順位がある場合の扱いも何通りか提案されている）、その場合に使えるノンパラメトリックな統計手法が数多く開発されている^{*4}。順序尺度を表す変数の平

場合によっては再コーディングが必要となることもある。ここで注意すべきは、性別という名義尺度と、生殖状態という名義尺度は、別の尺度だということである。しかし、男性をM、女性をFと表しても、男性を1、女性を2と表しても、1対1変換である限り、それは同じ性別という名義尺度である。

^{*3} +の数を数値として、たとえば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よりも尿タンパクが高濃度で検出され、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

^{*4} ノンパラメトリックとは分析の際に母数を仮定しないという意味であり、順位相関や順位和検定などがその例である。詳しくは第9章を参照されたい。

均^{*5}を求めるこには意味がないが、中央値^{*6}には意味がある。

- ただし、**もつともらしい仮定**を導入して間隔尺度であるとみなし、平均や相関を計算することも多い。たとえば、「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などは本来は順序尺度なのだが、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る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。

3.4 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。^{*7}**
- たとえば、摂氏温度や西暦年は、間隔尺度である。気温が摂氏30度であることは、摂氏10度より摂氏20度分、温度が高いことを意味するが、3倍高いことは意味しない。しかし、平年なら最高気温が摂氏20度であるようなときに摂氏30度であれば、摂氏25度であるのに比べて、平年との差が2倍あるとは言って良い。
- 間隔尺度をもつ変数に対しては、平均や相関係数など、かなり多くの統計量が計算できるが、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。^{*8}**
- たとえば、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、

^{*5} 次章で説明するが、ここでは、全部の値を足し合わせて値の数で割ったもの、と普通に考えておけば良い。

^{*6} これも次章で説明するが、ここでは、小さいほうから順番に値を並べて、ちょうど中央にくるものと考えれば良い

^{*7} 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。

^{*8} 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。

間隔尺度と変動係数

たとえば、標準偏差を平均で割った値を%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は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。

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^{*9}。

3.6 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遙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。

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とする。

3.6.1 離散変数の場合

- 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。

- 帶グラフ：横棒を全体を100%として各値の割合に従って区切って塗り分けた

^{*9} 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。

図である。R では

```
px <- table(X)/NROW(X)
barplot(matrix(px,NROW(px)),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、各値の割合に従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^{*10}。

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、R で次のようにプログラムすれば

```
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); op <- par(mfrow=c(2,2))
barplot(fc, main="frequency bar plot")
barplot(matrix(fc,NROW(fc)), beside=F,
main="cumulative bar plot", col=rainbow(NROW(fc)))
barplot(matrix(pc,NROW(pc)), horiz=T, beside=F,
main="OBI-graph", col=rainbow(NROW(pc)))
pie(pc, main="Pie chart", col=rainbow(NROW(pc)))
par(op)
```

図 3.1 のように表示ができる。

3.6.2 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つ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} R-1.5 以前は `piechart()` 関数だったが、置き換えられた。

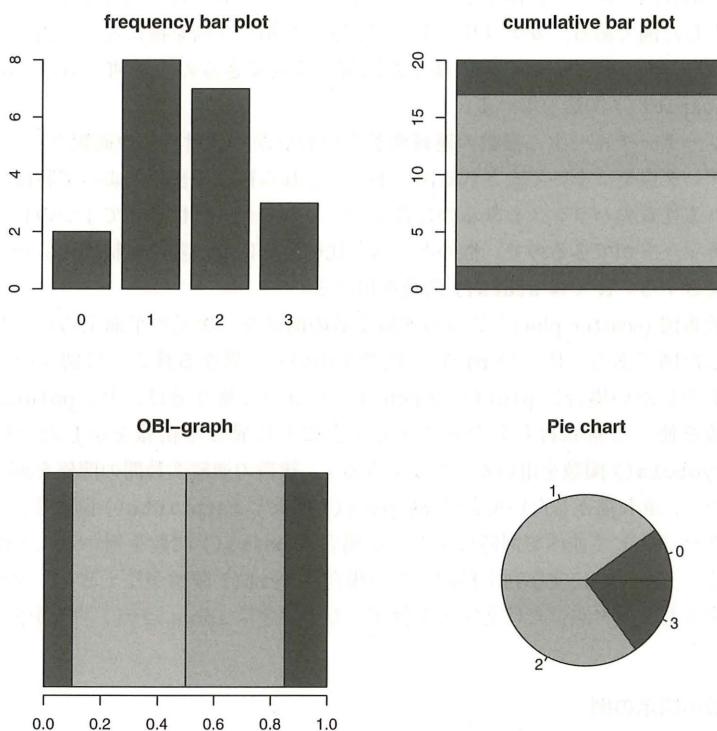


図 3.1 離散変数の図示の例

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。

- 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。縦軸に変数值をとって、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。

- 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。
- 散布図 (scatter plot)：2つの連続変数の関係を二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。

連続変数の図示の例

平成元年3月9日から4月2日の東京地区の最低気温(℃)が次のようであったとする。

3.2, 3.1, 5.1, 4.8, 8.3, 9.8, 8.3, 6.6, 5.1, 3.8, 5.2, 5.6, 6.5, 5.7, 5.7, 7.4, 6.2, 7.0, 6.7, 5.7, 6.2, 6.0, 8.8, 10.7, 8.5

このとき、次のRのプログラム

```
prec <- c(3.2, 3.1, 5.1, 4.8, 8.3, 9.8, 8.3, 6.6, 5.1, 3.8,
5.2, 5.6, 6.5, 5.7, 5.7, 7.4, 6.2, 7.0, 6.7, 5.7, 6.2,
6.0, 8.8, 10.7, 8.5)
op <- par(mfrow=c(2,2))
hist(prec, main="Histogram")
qqnorm(prec, main="Normal QQplot")
boxplot(prec,main="Box and whisker plot")
stem(prec,2)
par(op)
```

を使えば、図 3.2 のような図示ができる^{*11}。

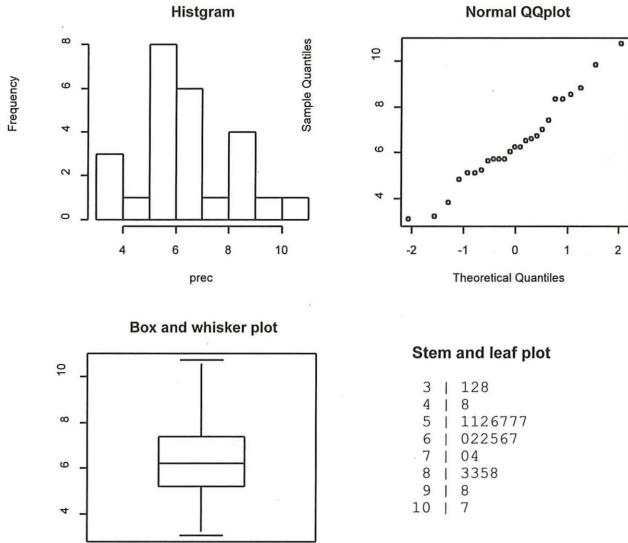


図 3.2 連続変数の図示の例

^{*11} ただし、右下の幹葉表示はテキスト表示しかされないので、それをコピーして OpenOffice.org の draw 上でグラフィック画面に貼り付けた。