EBNのための研究計画

データ処理とデータ解析

林　邦彦・高木廣文

EBNURSING Vol.3 No.3 2003〔別刷〕 中山書店

EBNのための統計の読み方 11

EBNのための研究計画
データ処理とデータ解析

林　邦彦 Hayashi Kunihiko
群馬大学医学部保健学科

高木廣文 Takagi Hirofumi
新潟大学医学部保健学科

嘘をつくもの：詐欺師，政治家，統計
ベンジャミン・ディズレイリ（英国政治家）

データ・マネジメント

調査や実験などで収集された情報は，数値，文字，記号，画像，音声といった様々な形態をとる．これら情報の記録保存には，電子媒体を利用すると便利である．近年のコンピュータ技術の発展や，SAS，HALWIN，SPSSといった統計パッケージの開発のおかげで，専門的なコンピュータ操作やプログラミングの技術がほとんどなくても，多くの人にとってデータ解析を簡単に行える環境となった．統計解析結果の再現性の点からも，データ量の多少にかかわらずコンピュータにデータ・セットを作成し，汎用されている統計パッケージを利用してデータ解析をする利点は大きい．

すぐれた疫学研究を実施するために作成された，米国GEPガイドライン（Guideline for Good Epidemiology Practice）[1]の研究実施の章における「データ収集と照合」と「解析」の項の記述を❶に示す．ここでは，電子媒体での記録保存の重要性やその管理責任者の必要性とともに，データの品質向上のためには，解析データ作成のため実施される全プロセスの概要を文書化しておくことが述べられている．

データ入力やデータ点検といったプロセスにおいて，標準作業手順書（Standard Operation Procedure：SOP）を準備することが，研究の質を上げるためには重要である．「研究結果」という最終製品の品質を上げるには，その作成プロセスを管理していくという品質管理（Quality Control：QC）の考え方が医学研究にも活かされている．これら品質向上のためのプロセスをデータ・マネジメント（Data Management：DM），それらを行う人をデータ・マネジャー（Data Manager）とよぶ．

Data collection and verification（データ収集と照合）
All data collected for the study should be recorded directly, accurately, promptly, and legibly. The individual(s) responsible for the integrity of the data, computerized and hard copy, shall be identified.
All procedure used to verify and promote the quality and integrity of the data shall be outlined in writing. An historical file of these procedures shall be maintained, including all revisions and dates of such revisions. Any changes in data entries shall be documented.

Analysis（解析）
All data management and statistical analysis programs and packages used in the analysis should be documented. Reasonable effort should be made to validate interim steps in the analysis.

❶米国GEPガイドラインでの「データ収集と照合」および「解析」の項（International Society for Pharmacoepidemiology：Pharmacoepidemiology and Drug Safety 1996；5：333-338.[1]より抜粋．下線は著者）

解析用データ・セット作成の手順とデータ入力

電子媒体に保存された情報が直接入手できる場合を除いて，何らかの記録からデータをコンピュータに入力して解析用データ・セットを作成する作業が必要となる．一般に，解析用データ・セットの完成までには，入力前処理，データ入力，データ点検，データ固定の手順となる（❷）．

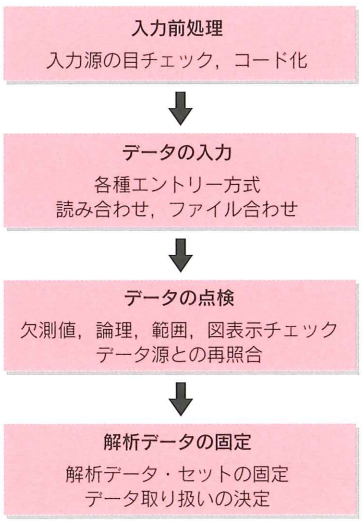
まず，入力前処理として，測定記録用紙，回収調査票，症例記録といった入力データ源の，どの情報をどのように入力するかを決定する．記録用紙をそのまま画像ファイルとして保存する場合や，記録の一部のみを統計解析する場合などには，必ずしも得られたすべての情報をコンピュータに入力する必要はないかもしれない．しかし，解析用データ・セットを作成した後に，「あっ，この項目も入力しておけばよかった」とならないように，統計解析に少しでも利用する可能性のある情報は一緒に入力しておく．

情報が文字や画像といった場合には，解析しやすいように情報をコードに変えておく前処理を行う．入力前に，文字で記入された疾患名，症状名，薬剤名などにコード番号をふったり，自由記載で書かれた事柄を内容別に分類し直し，分類コード番号を付けておくといったコード化処理である．疾患名や死因名などでは，ICD（国際疾病分類，❸）[2] のように世界で標準的なコード体系を利用すると，ほかのデータとも互換しやすい．

また，データ入力前の早めの時期に，入力データ源の記載内容を目で見て点検することも大変重要である．データの抜けや誤りの修正は，観察・測定から時間が経てば経つほど困難となっていく．識別番号など基本的な情報に漏れや誤りがないかなど，目で見てチェックするための点検項目リストを前もって準備しておくと，点検者が違っても均質なチェックが行え，以降のデータ入力の効率も上がる．

入力前処理がすめば，いよいよデータ入力となる．

ずっと以前は，まずすべてのデータを英数字や半角カタカナに直してコーディング・シートとよばれ



❷ 解析用データ・セットの完成まで

脳血管疾患（I60 - I69）

コード	内容
I60	くも膜下出血
I61	脳内出血
I62	そのほかの非外傷性頭蓋内出血
I63	脳梗塞
I64	脳血管発作，脳出血または脳梗塞と明示されないもの
I65	脳実質外動脈の閉塞および狭窄，脳梗塞に至らなかったもの
I66	脳動脈の閉塞および狭窄，脳梗塞に至らなかったもの
I67	そのほかの脳血管疾患
I68	ほかに分類される疾患における脳血管障害
I69	脳血管疾患の続発・後遺症

❸ **国際疾病分類（ICD-10）の例**（厚生省大臣官房統計情報部編：疾病，傷害および死因統計分類提要－ICD-10準拠．第1巻総論．厚生統計協会；1995．[2] より）

る紙に記入し，それをカードにパンチして（実際に，穿孔機とよばれる機械でカードに穴をあけ），カード読み取り機でバコバコバコと音をたてながらコンピュータに読み込んで，磁気テープに保存したりした．現在では，マーク・シート式調査票などOCR（光学式文字読み取り）用に設計された記入用紙は，そのまま読み取り装置から読み込むことができる．より一般的には，表計算シートや簡易データ・ベースとよばれるコンピュータ・ソフトを利用して，入力データ源を目で見ながらキーボードでタイプして入力することが多いだろう．

データ入力プロセスの基本は，入力者が内容の判断をせずに前処理された情報をそのまま入力することである．1人の入力者がデータ入力した場合（シングル・エントリー方式）では，入力したデータの印刷出力と入力データ源との照合を行う．通常は，1人が目で見て照合するのではなく，2人がペアになり，片方が印刷出力（もしくは入力データ源）を見て声を出し読み上げ，もう一方の人がそれを聞いて入力データ源（もしくは印刷出力）と照合するので，この照合プロセスを「読み合わせ」とよぶ．データ解析者にとって，この作業で解析データの構造や特性への理解が進むという利点もあるが，データが大量の場合には，読み合わせ作業は大変な時間と人手と声を要する．

そこで，正確かつ効率よくデータ入力を行うために，2人の入力者が独立して入力を行う方式（ダブル・エントリー方式）をとることが多い（❹）．それぞれの入力者が作成したデータ・ファイルの全入力項目を照合して，2人が一致した箇所は正しいデータとみなす．不一致の箇所のみについてデータ源に戻って確認し，正しいデータを確定入力する．本当は両方の入力者とも同じように誤っているのに（一致した誤り），それを正しいとみなしてしまうことが心配であれば，独立した入力者をもう1人増やして3人とする（トリプル・エントリー方式）．3人とも同じように誤るのであれば，それはもう誤りではなく正しい入力データといってよいかもしれない．いずれの入力者も入力ミスが少ない場合には，大変効率よく正確なデータ・セットの作成ができるが，誰か1人でも入力ミスが多い者が混ざると，ほかの入力者がどんなに優秀でも不一致データが多くなり，か

入力者A作成のデータ・ファイル

ID	性別	年齢	身長	体重	BMI
1	2	28	156	68.5	28.1
2	1	35	160	65.5	25.6
3	2	45	142	40.5	20.1
4	2	58	**153**	52.0	22.2
・	・	・	・	・	・

入力者B作成のデータ・ファイル

ID	性別	年齢	身長	体重	BMI
1	2	28	156	68.5	28.1
2	1	35	160	65.5	25.6
3	2	45	142	40.5	20.1
4	2	58	**158**	52.0	22.2
・	・	・	・	・	・

データの照合 ↓

ID	性別	年齢	身長	体重	BMI
○	○	○	○	○	○
○	○	○	○	○	○
○	○	○	○	○	○
○	○	○	×	○	○
・	・	・	・	・	・

↓

照合者Cによって不一致箇所を確定したデータ・ファイル

ID	性別	年齢	身長	体重	BMI
1	2	28	156	68.5	28.1
2	1	35	160	65.5	25.6
3	2	45	142	40.5	20.1
4	2	58	**153**	52.0	22.2
・	・	・	・	・	・

❹ ダブル・エントリーでの入力と照合の例

えって効率は悪くなるので注意を要する．

データ点検とデータ固定

データ入力プロセスが無事に終了しても，すぐに統計解析を行うのではなく，作成したデータ・セットを系統的に一度点検する．点検の方法には，「欠測値チェック」，「論理チェック」，「範囲チェック」，「図表示チェック」などがある．

「欠測値チェック」では，欠測値を取りえない項目で欠測値がみられないか，空欄となっているが実はほかの内容を示していないか，などをチェックする．❺は，データ・セットから欠測値を探して，データ源を点検した例である．おそらく高血圧と高コレステロール血症以外の疾患では，「2．いいえ」に○を付け忘れただけで，既往はないと判断してよいだろう．問診票などではよく起こる事例である．しかし，全項目が空欄未記入の場合には，どの疾患も既往がないのか，それとも「不明」や「回答せず」なのかは判断できない．

「論理チェック」では，論理的に起こりえないデータがないかを探す．性別（男性）で出産回数（3）回とか，高血圧（なし）で降圧剤（服用中）とか，入院（平成15年4月）で死亡（平成12年1月）などは，どちらかの情報が誤っていると考え，データ源に戻って確認修正しなくてはならない．

「範囲チェック」では，取りうる値の範囲を設定して，それから逸脱したデータを拾い出し，データ源で誤りがないかを確認する．拾い出しの範囲設定では，必ずしも，正常・異常を区別する値や，今まで報告されている最大値・最小値といった値を使うのではなく，あくまでデータ源に戻って確認すべき値はどの範囲かを考える．範囲設定が困難な場合には，値の順にデータを並べて，大きな値から何個，小さな値から何個と決めて，それらを見直すといった方法もある．

ほかのデータと大きくかけ離れている特殊なデータを拾い出すには，図表示を用いると便利である．

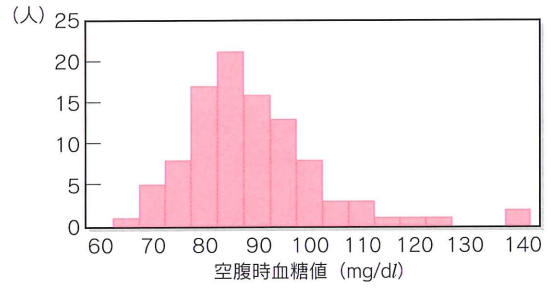
❻－aは100人の対象者の空腹時血糖値をヒストグラムで表示した例である．140～144.9 mg/dlの階級の2例については，本当に空腹時で採血したか不安であるため，記録用紙に戻り同時に測定したHbA1c値や糖尿病診断治療歴とともに測定値の点検を行った．

また，散布図を利用すると，1変数の分布だけでなく2変数の分布の様子から，点検候補データを拾い出すことが可能となる．❻－bは，ある仮想集団でのウエスト周とヒップ周を散布図で示したものである．肥満の指標としてはBMI（body mass index．体重（kg）を身長（m）の2乗で割る）がよく用いられているが，同じBMIの値でも腹部の脂肪が多い体型のほうが各種疾患発生リスクが高いといわれる．そこで，体型の指標としてウエスト周/ヒップ周の比をみるために，ウエスト周

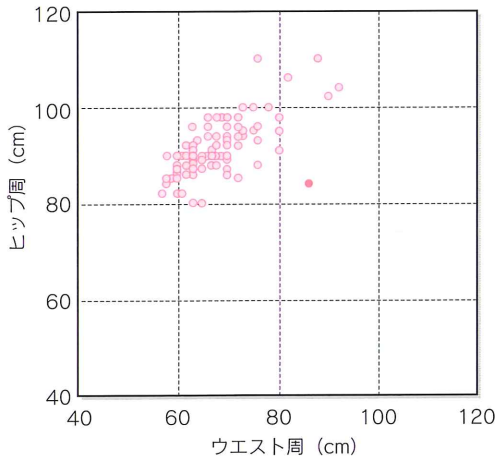
以下の病気と診断されたことがありますか．

疾患名	診断されたことがありますか「はい」の場合は右欄もお答えください		初めて診断された年齢
1．高血圧	1．はい	2．いいえ	（ 62 ）歳
2．心筋梗塞	1．はい	2．いいえ	（ ）歳
3．狭心症	1．はい	2．いいえ	（ ）歳
4．くも膜下出血	1．はい	2．いいえ	（ ）歳
5．脳出血	1．はい	2．いいえ	（ ）歳
6．高コレステロール血症	1．はい	2．いいえ	（ 58 ）歳
7．乳癌	1．はい	2．いいえ	（ ）歳
8．胃癌	1．はい	2．いいえ	（ ）歳

❺ 欠測値データの入力データ源の例



❻-a ヒストグラムを利用した図表示チェックの例



❻-b 散布図を利用した図表示チェックの例

とヒップ周を自己記入式調査票で調査した例である．

図中の●印のデータは，それぞれの変数では範囲内にあるが，2変数の関係からはほかのデータと少しかけ離れている．そこで，再度，ウエスト周とヒップ周の誤記入や誤入力がないかを調査票に戻って確認することなる．

いずれの点検法でも，決してそれらから誤ったデータがわかるのではなく，その可能性がある候補データを拾い出していることを十分に認識しなければならない．設定範囲を超えた特殊なデータは，必ずしもそのことだけで誤ったデータを意味しているのではなく，内容の点検をせずに一律に統計解析から除いたりすると，研究結果に影響を与えてしまうだろう．また，その特殊なデータが研究者に語りかけていた重要な本質をも，見逃してしまうかもしれない．

データ点検やクリーニング作業は際限ないものだが，点検終了と考えた時点で解析用データ・セットの固定を行う．データ固定を宣言した後は，よほどのことがないかぎりデータ修正は行わない．データ固定を英語ではフリーズ（freeze．凍らせ固め動けなくする感じか）というようだが，米国映画で警官が犯人に銃を向けながらいう台詞と同じである．フリーズの後に動くことは，それなりの危険を覚悟しなくてはならない．

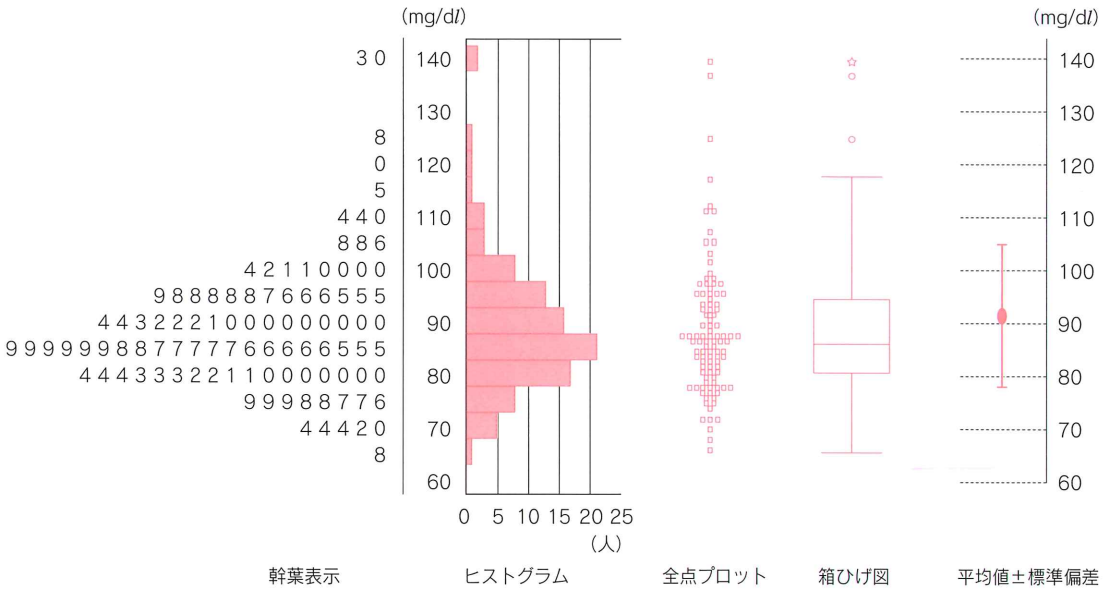
また，データ・セットの固定と同時に，データの取り扱いの細部を決めることも必要となる．すでに研究計画書に，主要エンドポイントでの統計解析法や大まかな統計解析方針は定めてはある．しかし，毎週測定するはずの血圧値が2週後だけ測定忘れの例が出た，観察開始後3か月目に転勤のため対象者が引っ越したので以降，測定ができなくなったなど，生じた個々の問題データの取り扱いを決める必要が出てくる．特に欠測値については，欠測値がない対象者のみを統計解析の対象とする方法，一時点前のデータで補填する方法（Last Observation Carries Forward：LOCF法）など様々な方法があり，データ固定時までに取り扱いを決定しておく．

データの表示

解析用に入力したデータは，その性質によって大きく2つに分けることができる．

「性別」というデータは，男性35人に女性65人といったように数を数え上げることに意味があるので，計数データ（質的データ）とよぶ．計数データの表示では，棒グラフ，棒内訳グラフ，円グラフなどで，各分類での度数が全体に占める割合を示す．

一方，「身長」(cm)や「体重」(kg)など，数え上げるのではなく量を計るデータを計量データ（量的データ）という．計量データの分布や構造の記述に有用な手法に，探索的データ解析法（Exploratory Data



❼いろいろなデータの図表示法

Analysis：EDA）がある[3-5]．この手法の考え方は，データの構造をできるだけ視覚的にとらえることができる手法を用いる，飛び離れたデータ（はずれ値，outlier）の影響を受けにくい手法を用いる，分布の仮定など統計学的前提から逸脱してもその影響を受けにくい手法を用いる，といった特徴をもつ．

前述の100人の空腹時血糖値データを用いて，探索的データ解析法でよく用いられる，幹葉表示，ヒストグラム，箱ひげ図でデータ表示してみよう．❼の左端が「幹葉表示」（stem-and-leaf display）だが，空腹時血糖値を5きざみの階級を幹にして，1の桁の数字を葉として表示している．たとえば，最小値は，幹で65～69.9の位置の葉に8という数字があるので68とわかる．同様に，幹で70～74.9の位置には，0 2 4 4 4と5個の葉の数字があるので，70，72，74，74，74と5つの測定値があったことがわかる．列車の発車時刻の「時」を幹に「分」を葉にして幹葉表示にすれば，駅で見慣れた時刻表ができあがる．データ数が少なければ，幹葉表示は手書きでも簡単に作成できる．データを視覚的に把握できると同時に，後述する中央値や四分位点などを表示から拾い出すこともできる．

次に，❼では，同じデータ・セットを用いて作成したヒストグラムを示した．データ数n＝100なので，通常のヒストグラムでは階級数は7段階程度（Stugesの式：$1+\log_2 n$）となり，空腹時血糖値を10きざみの階級で作成するかもしれないが，探索的データ解析法では20（$=10\cdot\log_{10} n$）程度の階級数が適切とされ，例では空腹時血糖値5きざみの階級を作った．このように，探索的データ解析法でのヒストグラムは，階級数を多くしてデータの構造を視覚的に把握するのが特徴である．

ヒストグラムの右に，全測定値をプロットした図を表示した．これは最も素直な表示法といえ，すべてのデータの位置関係がそのまま把握できる．ただし，データ数が多くなり同じような値が増えるに従い，図はだんだん見にくいものになってしまう．そこで，データ数が増えてもその構造を把握できる表示法として，「箱ひげ図」（box-whisker plot）が考案された．データを小さな順に並べ直して，小さな

ほうから25％の位置にある値を第1四分位数（25パーセンタイル），50％の位置にある値を中央値（第2四分位数，50パーセンタイル），75％の位置にある値を第3四分位数（75パーセンタイル）という．

まず，第1四分位数と第3四分位数で箱を作り，箱中の中央値の位置に線を入れる．ついでに，平均値の位置に＋や◇を加えることもある．25パーセンタイルと75パーセンタイルで箱を作ったので，全データの半数はこの箱の範囲内に散らばっていることがわかる．第3四分位数と第1四分位数の差（つまり箱の長さ）をヒンジとよぶが，ひげはヒンジの1.5倍を箱の上下それぞれにとり，その内側で最も近いデータまで描く．ひげの範囲から外れたデータ点を○や☆で表示する．ひげの描き方には，5パーセンタイルと95パーセンタイル，最小値と最大値として表示することもある．❼の箱ひげ図は1.5倍ヒンジを用いて作図したが，箱もひげも値の大きいほうに長く，高値のほうに裾が長いデータ分布の様子が読み取れる．また，高値のほうのひげの外側にはずれた値が3つあることもわかる．

参考までに，❼の右端に平均値±標準偏差での表示をしてみた．平均値を真中に，（平均値－標準偏差）から（平均値＋標準偏差）までを線分で表したものである．この表示法では，どのようなデータ構造でも線分の両端はいつも平均値から等距離となる．❼の平均値±標準偏差からは，高値のほうに裾をひいている分布の特徴や，140mg/dl以上の測定値があったことなどわかる由もない．データの構造を把握するという観点からは，決してよい表示法とはいえないだろう．

データの要約

◆平均値，中央値，最頻値

次に，計量データにおいて，中心的なデータの位置を一つの値として要約することを考えてみよう．

これは代表値とよばれる指標だが，「平均値」（mean），「中央値」（median），「最頻値」（mode）が考えられる．

前述の空腹時血糖値データで算出してみる．データの個数n＝100で，1番目の人の測定値をx_1，2番目の人の測定値をx_2，i番目の人の測定値をx_i，そして最後n番目の人の測定値をx_n，またiが1からnまでの測定値の和を$\sum_{i=1}^{n} x_i$という記号で表すとしよう．平均値$\bar{x}$（より正確には算術平均）は，

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_i + \cdots + x_n}{n}$$

となる．空腹時血糖値データの例では，91.0 mg/dlであった．この平均値は，たとえ少数であってもほかのデータから大きくかけ離れたはずれ値があると，その影響を受けやすい．例題でも，❻－aや❼をみるとわかるように，140 mg/dl以上の例などのためか，平均値は視覚的に中心と思われる位置から若干高値を示す感じがある．そこで，このようなはずれ値の影響を受けにくい指標として，中央値や最頻値を計算してみる．

「中央値」は，すべての測定値を値の順に並べてちょうど中央の順位にある値を指す．データの個数nが奇数であれば，ちょうど中央となる順位は$\frac{n+1}{2}$番目となる．たとえば，9個のデータであれば小さなほうから（もしくは大きなほうから）5番目のデータが中央値となる．データの個数が偶数であれば，$\frac{n}{2}$番目と$\left(\frac{n}{2}+1\right)$番目の平均をとる．例題ではn＝100であったので，50番目と51番目の平均値をとることになるが，両方ともその値は89であったので，中央値は89.0 mg/dlとなる．

「最頻値」は，その名のとおり最も頻度が高く表れた値を探すことになるが，例題としてヒストグラム（❻－a）から計算してみる．85～89.9の階級に最も度数が多いが，その位置は階級幅（いずれの階級も幅は5 mg/dl）を両隣の階級の度数で按分する．

1つ下の階級80～84.9に17人，1つ上の階級90～94.9に16人なので，85から16：17に分ける位置とする．つまり，最頻値は

$85.0+\frac{16}{17+16}\times 5=87.4$ となる．

このように，中央値も最頻値も視覚的な中心的位置とほぼ一致する値を得ることができた．しかし，最頻値は，階級幅を変えると値が変わる，最も大きな度数が最下位や最高位といった端の階級にあると上記の按分ができない，また，同じ度数のピークが2つあると（二峰性），どちらが最頻か決められないという問題がある．そこで，はずれ値の影響を受けない指標として，通常は中央値を用いている．

◆データのばらつきの度合いを示す指標

次に，データのばらつきの度合いを示す指標を考えてみよう．最小値や最大値もデータのばらつきを示しており，最大値と最小値の差をとった範囲（range）が指標となる．例題では，最小値が68 mg/dl，最大値が143 mg/dlで，範囲は75（143－68）となる．また，データの大きさの順序を使ってばらつきの程度を考えるのであれば，前述の四分位数（quartile）を用いることが多い．第1四分位数と第3四分位数の差を四分位範囲（箱ひげ図における箱の長さに相当）といい，その半分の値を四分位偏差（quartile deviation）という．例題では，第1四分位数が82.5，第3四分位数が97.5であったので，四分位範囲が15(97.5－82.5)，四分位偏差が7.5(15/2)となる．これらは，同様に順序をもとに算出される中央値と一緒に用いられ，データのばらつきの度合いを示す．

一方，平均値を代表値としたときの，ばらつきの程度を表す指標を考えてみる．個々のデータと平均値との差$(x_i-\bar{x})$をとって，その平均的な値をみるのが素直であろう．しかし，この差のままであれば負もあれば正もあるので，絶対値をとるか2乗してから平均してみる．差の絶対値の平均$\frac{1}{n}\sum_{i=1}^{n}|x_i-\bar{x}|$を「平均偏差」(mean deviation)，差の2乗の平均$\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$を「分散」(variance）とよぶ．手元にあるデータから想定する母集団での分散を推定するには，差の2乗の和をnのかわりに（n－1）で割る[6,7]．この$\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$を「不偏分散」(unbiased estimate of variance）とよぶ．分散，不偏分散とも，元の測定値の2乗の単位となっているので，平方根をとって元の単位に戻す．それが「標準偏差」(standard deviation：SD）である．通常は不偏分散を使った標準偏差$\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}$が，個々のデータの平均値からのばらつき程度を示す指標として使われている．例題では，平均偏差が9.43，分散が169.2，不偏分散が170.9，そして標準偏差が13.07と算出された．

文献

1) International Society for Pharmacoepidemiology：Guidelines for good epidemiology practices for drug, device, and vaccine research in the United States. Pharmacoepidemiology and Drug Safety 1996；5：333-338.
2) 厚生省大臣官房統計情報部編：疾病，傷害および死因統計分類提要－ICD-10準拠．第1巻総論．厚生統計協会；1995.
3) Tukey JW：Exploratory Data Analysis. Addison-Wesley, Reading MA；1997.
4) 渡辺洋，鈴木則夫，山田文康ほか：探索的データ解析入門．朝倉書店；1985.
5) Hartwing and Dearing：Exploratory Data Analysis. 柳井晴夫，高木廣文共訳：探索的データ解析の方法．朝倉書店；1981.
6) 柳井晴夫，高木廣文編：新版看護学全書基礎科目 統計学．メヂカルフレンド社；1995.
7) 高木廣文：ナースのための統計学－データのとり方・いかし方．医学書院；1984.