原稿作成日：　2017 年 3 月 31 日

正しいデータの記述の仕方

〈教材提供〉
AMED 支援「国際誌プロジェクト」　提供

草案
新谷歩　大阪市立大学医学研究科医療統計学講座教授
加葉田大志朗　大阪市立大学医学研究科医療統計学講座特任助教

査読
大門貴志　兵庫医科大学医療統計学教授
角間辰之　久留米大学バイオ統計センター教授
市川家國　信州大学特任教授
山本紘司　大阪市立大学大学院医学研究科医療統計学講座准教授
石原拓磨　大阪市立大学大学院医学研究科医療統計学講座特任助教

目次

はじめに

人を対象とした研究において、収集したデータがどのような特性を持つ人から得たものかを知ることは、研究結果をどのような人に一般に当てはめることができるかを考える上でカギとなります。

研究結果をまとめた論文の多くは、その結果の項の冒頭部分に、研究に参加した「研究対象者の背景」を報告していますが、そこでは、研究で収集したデータを提供した研究対象者の年齢の平均値、男女の割合などの情報が整理されています。このようなデータの整理のことをデータの「記述」または「要約」と称し、そのような要約されたデータの値を「記述統計量」と称します。このデータの記述を間違えると、誤解を生む恐れがあります。せっかく集めた貴重なデータです。正しいデータの記述が統計解析の第一歩です。

本単元では、データの記述の方法を学んでいきましょう。

学習目標

本単元を通じてあなたが修得を目指すものは：

- データの種類を理解する
- 平均値と中央値の違いと利用方法を習得する
- 標準偏差と標準誤差の違いと利用方法を習得する
- 95%信頼区間の意味と特性を理解する

データの種類

データの種類によって記述の方法は異なります。したがって、データの種類に留意することが重要です。データの種類は、大雑把には、男性・女性のようなカテゴリー別の「カテゴリカルデータ（Categorical Data)」と、年齢、体重、血圧のように連続的な値をとる「連続データ（Continuous Data)」に分類することができます。カテゴリカルデータの記述には、頻度（Frequency）や割合（Proportion）を用います。例えば 50 名の研究対象者のうち 30 名が男性の場合、男性の頻度と割合はそれぞれ 30、60%となります。連続データの記述にはデータの代表値とばらつきを用います。データの代表値は平均値（Mean）や中央値(Median)で表わし､ばらつきは標準偏差(Standard Deviation)や四分位範囲(Inter-Quartile Range; IQR）で表わします。データの代表値に平均値を用いた場合は、ばらつきは標準偏差で表し、中央値を用いた場合は、四分位範囲で表すことが一般的です。

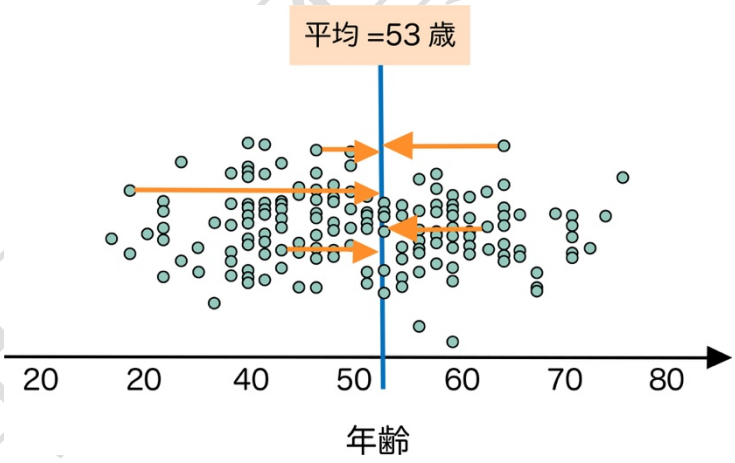
平均値（Mean）と標準偏差（Standard Deviation, SD）

データの代表値である平均値

平均値とは、収集したデータの合計をデータの個数で割った値のことです。たとえば 5 人の研究対象者の年齢をそれぞれ 10 歳、20 歳、30 歳、40 歳、50 歳とします。この 5 人の年齢の平均値は合計の 150 歳を 5 人というデータの個数で割ると 30 歳になります。この平均値のように、観測値の中心位置を表す値を「代表値」と呼びます。しかし代表値だけでは、5 人の年齢の観測値がどのようなパターンを取っていたかが、いま一つ、明確ではありません。そこでデータを記述する時には、代表値に加えて、データのばらつきを示す必要があります。

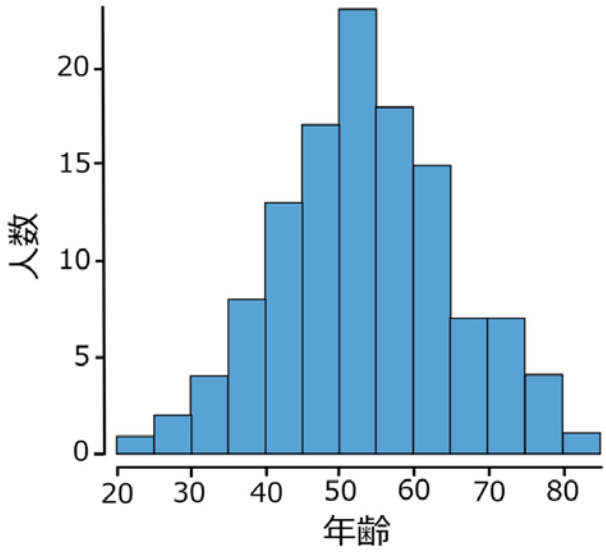
データのばらつきを表す標準偏差

先ほどの例でみてみると、年齢の平均値は 30 歳でしたが、データの中には年齢の高い人から低い人まで様々に存在しています。そのため目の前のデータが平均値を中心にどの程度ばらついているかを示す必要があります。データのばらつきの指標としてよく用いられるのが標準偏差（Standard Deviation: SD）です。標準偏差は概念的には「各観測値から平均値までの平均距離」と言えるものです。右の図では 100 人の年齢をグラフで示しています。63 歳の観測値から平均値の 53 歳までの距離は 10 歳です。こうした距離を全員について計算し、その平方和の平均（*注）の二乗根をとったものが標準偏差です。各データ値から平均値までの距離が短ければ標準偏差は小さくなり、それはばらつきが小さいことを意味します。反対に標準偏差が大きければばらつきが大きいことを意味します。



（*注）厳密には（標本平均-各観測値）の平方和を（n-1）で割った値を利用します。詳細についてはこの教材の範囲を超えますので省きます。

下の図は、100 人の年齢のデータをヒストグラムというグラフで表しています。

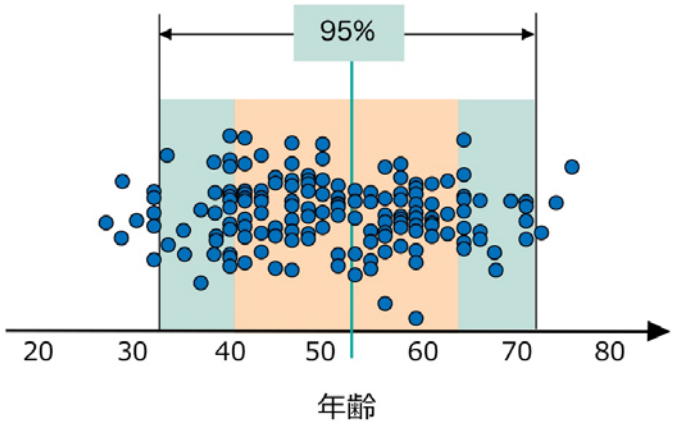


20 歳以上 25 歳未満の人が 1 人、25 歳以上 30 歳未満が 2 人、30 歳以上 35 歳未満が 3 人というように、年齢を 5 歳ごとの階級に分けたとき、各階級に何人の人がいるかを示したグラフです。このグラフでは平均年齢の 53 歳が含まれる 50 歳以上 55 歳未満の階級の人数が一番多く、その中心から遠ざかるほどおおよそ左右対称にその人数が減るのが見てとれます。このようにデータが平均値を中心に左右対称に同程度にばらついている場合、「データは正規分布（Normal Distribution）に従っている」と仮定できます。

データが正規分布に従うと仮定できる場合、平均値から標準偏差の 2 倍の範囲（53 歳±11 歳×2）、つまり 31 歳から 75 歳の範囲内にはおよそ 95%の研究対象者の値が含まれていると統計的に考えることができます。

95% の患者の年齢が平均値 ± 2SD の間にある。

53±2×11=(31.75)

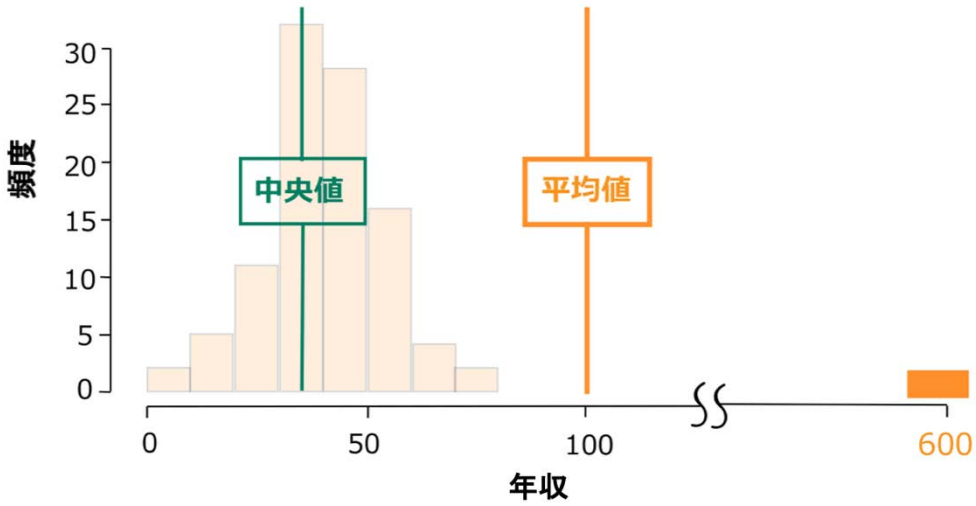


中央値（Median）と四分位範囲　（Inter-Quartile Range）

次にデータが正規分布に従っていない場合を考えてみましょう。

平均値の問題点と中央値

街角調査で 101 人から年収についてのデータを収集したとしましょう。100 人までの年収は 100 万円から 700 万円の間で、ごく一般にみられる値でしたが、101 人目に年収 6 億円の野球選手に出会いました。その結果 101 人の中心を表す平均値は約 1000 万円になりました。このとき、この平均値は本当にこのデータ全体の代表値といえるのでしょうか？



右の図はこのデータのヒストグラムです。野球選手以外の人の年収は 100 万円から 700 万円の中に収まっていました。したがって、この 100 人の研究対象者は多くても 700 万円しかもらっていないことになります。つまり、この調査対象集団は平均で 1000 万円もらっていると解釈すると、事実を間違って解釈するはめになります。

この例に示されるように、データの中心位置に当たる値は、本当はもっと低いのにも関わらず、平均値（1000 万円）は極端な値（6 億円）に引っ張られて算出されています。このような場合では、平均値はデータの中心を表す代表値として適切とは言えません。

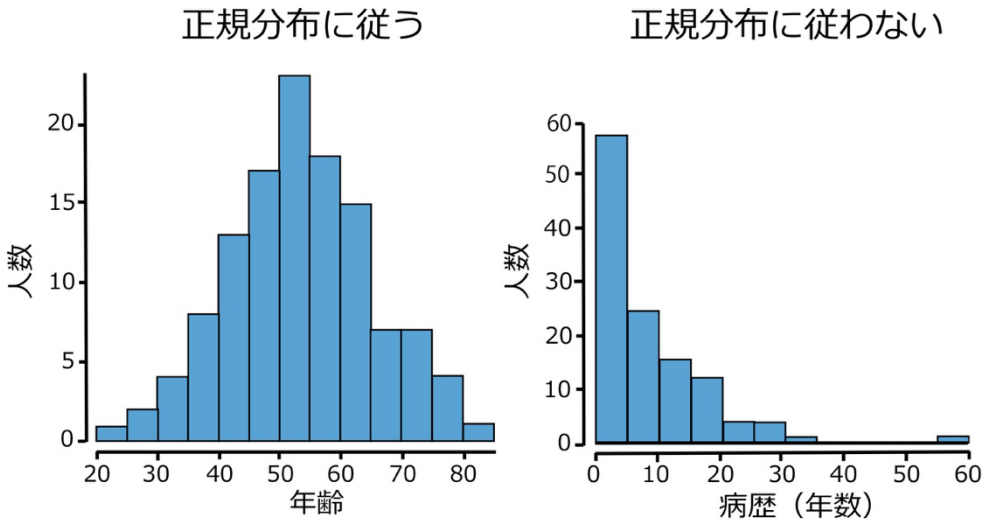
上述したような極端に離れている値（外れ値）が存在したり、データが正規分布に従っていないときに、データの代表値として使われるのが「中央値」です。このデータの場合、中央値は年収の一番低い人から高い人まで並べたときのちょうど真ん中の人の年収である 400 万円になります。中央値を使えば、たとえ 6 億円という極端に大きな値を持つ人がいた場合でも 101 人の中でちょうど真ん中の 51 番目の人の年収がこのデータの中心位置の値になるので、平均値 1000 万円のように外れ値に引っ張られた値をデータの代表値にしてしまうことを避けることができます。

標準偏差の問題点と四分位範囲

代表値に中央値を使う場合には、データのばらつきとして四分位範囲を用います。

四分位範囲とはデータを小さい順に数えて前から 4 分の 1 のところに位置する値である 25%点（第 1 四分位点）と前から 4 分の 3 に位置する値である 75%点（第 3 四分位点）の範囲のことを指します。したがって、四分位範囲にはデータの 50%の値が含まれると言えます。なお、前から 4 分の 2 のところ、すなわち、ちょうど真ん中の 50%点（第 2 四分位点）が中央値となります。上の例では年収の中央値が 400 万円であるのに対し、四分位範囲は 300 万円～600 万円、つまり調査対象集団の半分の人の年収が 300 万円から 600 万円の間に存在すると解釈できます。

正規分布と医学データ



先ほど簡単に触れましたが、正規分布とは、左右対称の釣鐘型を描く分布のことを称します（上の左図）。平均値近傍のデータの個数が最も多く、平均値から離れるほど左右均等に徐々に少なくなっていく、といったデータが抽出したサンプルの中で見られた場合、データが集められたもととなる集団（母集団）において変数の値が正規分布に従っていると仮定できます。例えば年齢のデータは正規分布に従っていると仮定できますが、その他のデータは必ずしもそうではありません。とくに医学研究では正規分布に従うと仮定できないデータ（上の右図）を扱うことが多く、平均値や標準偏差を用いると誤解を招くことが多いものです。

平均値（標準偏差）と中央値（四分位範囲）の選び方

データが正規分布に従っていると仮定できる場合にはデータの記述として平均値と標準偏差を用い、そうでない場合には中央値と四分位範囲を用いるということを学習しました。では、それぞれのデータについて、平均値、中央値のどちらを使うかは正規分布に従っていると仮定できるか否かを確認して決めるのでしょうか？もちろんデータによって平均値と中央値を使い分けている論文もありますが、データが正規分布に従っていると仮定できる場合には、平均値と中央値とはほぼ等しい値をとりますので、最近ではデータによらず中央値と四分位範囲だけを使ってデータを記述している報告も少なくないようです。

標準誤差（Standard Error: SE）と信頼区間（Confidence Interval: CI）

標準誤差とは

標準偏差はデータ自体のばらつきであると説明しましたが、次に標準誤差について説明します。

例えば、ある降圧薬の効果を確認する研究を実施するとしましょう。施設 A では 100 人を対象として、降圧薬を服用した際に血圧が平均で 20mmHg 下がりました。さて、この結果をもって、「降圧薬は研究対象となった人と同様な人の血圧を 20mmHg 引き下げる効果がある」と言えるでしょうか？あくまでもこの結果は手元のデータから得られた値であり、世の中すべての人を対象にした場合でも同じことが言えるかどうかはそれらすべての人の血圧のデータを収集してみないと分かりません。このように手元のデータから、世の中に存在するすべてのデータに基づく降圧薬の効果の値を推し測ることを「統計的推定」と呼びます。この例では、降圧薬の効果「20mmHg 減圧」が手元のデータから推定されました。さて、ここで他の施設で似たような研究をした場合、同じような値が得られるでしょうか？例えば施設 B でも同様に 100 人集めて研究を行い、施設 C でも、施設 D でも…とあわせて 100 施設で同様の研究をした場合、すべての施設で施設 A とまったく同じ平均 20mmHg の降圧効果が出るとは考えにくいでしょう。つまりこの「平均的な降圧効果」として推定された値（推定値）も研究をたくさん行えば、ばらつきが起こります。しかしながら、100 施設で同様の研究を行うことは、不可能に近く非現実的です。しかし、「もし同じ研究を何回も行ったら推定値（ここでは血圧の変化量）はどのくらいばらつくのか？」ということを理論上計算することは可能です。このような理論上の推定値のばらつきを「標準誤差」と呼びます。標準誤差は、数学的な根拠に基づいてデータのばらつきである標準偏差を研究対象者数の二乗根で割った数で計算できます。

$$\text{標準誤差} = \text{標準偏差} / \sqrt{\text{研究対象者数}}$$

例えば、100 人から収集した血圧データの平均値が 80mmHg で、標準偏差が 10mmHg だったとします。このときの標準誤差は、10÷√100 で 1 と計算できます。この式からわかるように、研究対象者数が分母に表されているので、研究対象者数が多い研究ほど標準誤差は小さくなります。

標準誤差と真の値

標準偏差は概念的に「平均値からの平均的な距離」ということは既に学びました。標準誤差は同じような研究が同じ数の研究対象者からのデータを用いて無数に行われた場合、その無数に存在する研究のそれぞれで得られたデータの平均値のばらつきを表しています。標準誤差で表されたばらつきは、それぞれの平均値と「真の値」の平均的な距離を意味しています。この「真の値」とは今回の例であれば、「世界中の人がこの降圧薬を使ったときの効果」のことを示しており、神のみぞ知る値といえます。

これは抽象的な話なので、理解しづらいかもしれません。そこで、あなたが神様になったことを想像してみてください。あなただけがこの降圧薬の降圧効果が 15mmHg であることを知っています。人間界をのぞいていると、施設 A で「降圧効果は 20mmHg だ」と言っているところを見てあなたは「惜しいなあ」と思っているかもしれません。次の施設 B では「降圧効果は 30mmHg だ」と結論づけているところを見て「ダメ、ダメじゃないか」と呆れているかもしれません。あなた（神様）だけは、中心となる本当の値を知っているので、研究者たちの推定した結果がどのくらい真の値から外れているかがわかります。そのため真の値を中心とした推定値の分布を描くことができ、真の値（±標準誤差）などというデータの記述をすることも可能かもしれません。しかし研究者たちは真の値が分からないので、神様のようにそのような分布を描くことができません。そのため、研究者たちはわからない真の平均値を表現することを諦め、その代わりに、自分のデータの平均値そのものの代わりに、「信頼区間(Confidence Interval: CI)」というものを利用するのです。

信頼区間とは

これまで説明したように、データから得られた推定値（例えば平均値）と標準誤差を用いて示された範囲のことを信頼区間と呼び、推定値が正規分布に従っていると見なすこと（仮定すること）ができるとき、特に「推定値±2×標準誤差」（*注）で計算される範囲を「95%信頼区間」と呼びます。信頼区間は手元のデータから推定した値の精度を示すことになります。信頼区間が広ければ推定の精度は低く、反対に信頼区間が狭ければ推定の精度が高いことになります。

（*注）厳密には「推定値±1.96×標準誤差」

例えば、100 人から収集した血圧データの平均が 80mmHg で、標準偏差が 30mmHg だったとします。その場合、標準誤差は、30÷√100 で 3 となります。平均血圧 80mmHg を中心とした 95%信頼区間を計算すると下限値が 80－2×3=74、上限値は　80＋2×3=86 なので、この 2 つの数字で表される範囲として信頼区間は［74、86］と推定されます。標準誤差の計算式に示されるとおり、研究対象者数が分母に含まれているので、研究対象者数が多い研究ほど標準誤差は小さくなります。平均と標準偏差が同じあっても研究対象者数が 1 万人の場合、信頼区間は下限値が 80－2×（30÷100）=79.4、上限値が　80＋2×（30÷100）=80.6 ということで、［79.4、80.6］と推定されます。このように研究対象者数が多くなればなるほど推定の精度が上がることが分かります。

この単元に関係するビデオ教材

EZR のインストール
データセットの作り方
平均値と標準偏差
中央値と四分位範囲
標準誤差と信頼区間

本単元は、日本医療研究開発機構：研究公正高度化モデルである 「医系国際誌が規範とする研究の信頼性にかかる倫理教育プログラム」（略称：AMED 国際誌プロジェクト）によって作成された教材です。作成および査読等に参加した専門家の方々の氏名は、冒頭に掲載されています。

この単元に関係する国際誌におけるチェックポイントをいくつか紹介します。
（内容は解釈を助けるために一部意訳している部分もあります）

①Nature
（http://image.sciencenet.cn/olddata/kexue.com.cn/upload/blog/file/2010/12/2010128212513557501.pdf; visited on 2018.02.11）
②New England Journal of Medicine （http://www.nejm.org/page/author-center/manuscript-submission#electronic; visited on 2018.02.11）
③Science （http://www.sciencemag.org/authors/science-editorial-policies; visited 2018.02.11）
④The EMBO Journal （http://emboj.embopress.org/authorguide#embargopolicy; visited on 2018.02.11）
⑤JAMA （http://jamanetwork.com/journals/jama/pages/instructions-for-authors; visited on 2018.02.11）

①Nature
- 用いたデータセットの症例数を記載すること
- 平均や中央値などデータの代表値に何が使用されたかを示すこと
- 標準偏差や四分位範囲などデータのばらつきを表すためにどの指標が用いられたかを記載すること
- 標準偏差や標準誤差については a ± 標準偏差 や a ± 標準誤差 のように記載すること

②New England Journal of Medicine
- 信頼区間などのデータの不確実性を示す指標は首尾一貫して用いられるべきである。これは結果を包括的に図示する図についても同様。

③Science
- データの事前編集（データの変換、再コード可、再スケール化、標準化、一定値以上のデータの切り捨て（Truncation）、測定限界値以下の観測値や外れ値の取り扱い、またいかなる観測値の削除や編集も含む）については、正しい知識と正当性のある理由付けが必要である。
- 研究結果の理解に不可欠な変数については、記述統計量を記載すること。記述統計量には症例数や平均や中央値など、どの統計量が使用されたかについても記載すること。
- 正規分布に従う連続変数には、平均や標準偏差を用いること。また分布が正規分布のように左右対称でない連続変数については中央値や最小値・最大値の範囲または四分位範囲を用いること。いかなる場合においても平均が用いられたのか、中央値が用いられたのか、どのばらつきの指標が用いられたのかなどを明記すること。
- 症例数が 20 未満のようなデータにおいては、個人情報保護の観点から倫理的に問題がない限りすべてのデータ値を表にして記載することが望まれる。すべての測定値は単位を記載すること。
- 平均、相関係数、回帰係数などの点推計値や、平均差、オッズ比、ハザード比などの比較の定量化に用いた指標については標準誤差、信頼区間など不確実性を表す指標を併せて記載すること。

④The EMBO Journal
- 記述統計ではデータの中心を示す値（平均や中央値）と散布度を示す値（標準偏差や範囲）を示す必要がある。データの数が少ないような場合には、標準偏差よりも範囲など指

標を用いることが望ましい。標準誤差や信頼区間は群間の比較を行う際に記載するのが適当である。

⑤JAMA

■ Result（結果）の章では、可能な限り結果を数量的に表し、それらを信頼区間などの不確かさの指標とともに記載すること。可能な限り計量的な結果（頻度、率）などを信頼区間など不確実性の度合い（測定誤差など）をあらわす指標とともに報告すること。

■ 要約統計量を計算するために利用した一般的に利用される解析手法については詳細を記載する必要はないが、Method の章で簡単に説明しておくべきである。

■ 可能な限り計量的な結果（頻度、率）などを信頼区間など不確実性の度合い（測定誤差など）をあらわす指標とともに報告すること。

■ Mean（平均）や SD（標準偏差）はデータが正規分布に従うときに使用し、そうでない場合は Median（中央値）や Interquartile ranges（IQRs）を用いること。