Ⅷ 参考：統計の基礎

「成長段階にある児童・生徒の体重は身長に比例する」と言われている。本当だろうか。こういった疑問が生じたら、児童・生徒の身長と体重に関する統計データを手に入れて、相関係数を計算すれば、その真偽の程を確かめることができる。

統計は、このように世の中のいろいろな現象について、その裏に潜む法則性や規則性を発見し、或いは確認するための道具となるものである。

以下に、その概念、使い方、表現方法などの基礎的な事項について説明する。

第１　統計の概念

統計の定義

統計とは、人、物、出来事などの集まり（統計集団）を対象とし、その集まりを構成する各個体（統計単位）を観察した結果に基づき、その集まりの大きさや内部の構成を具体的な数字で表したものである。

＜統計の作り方＞

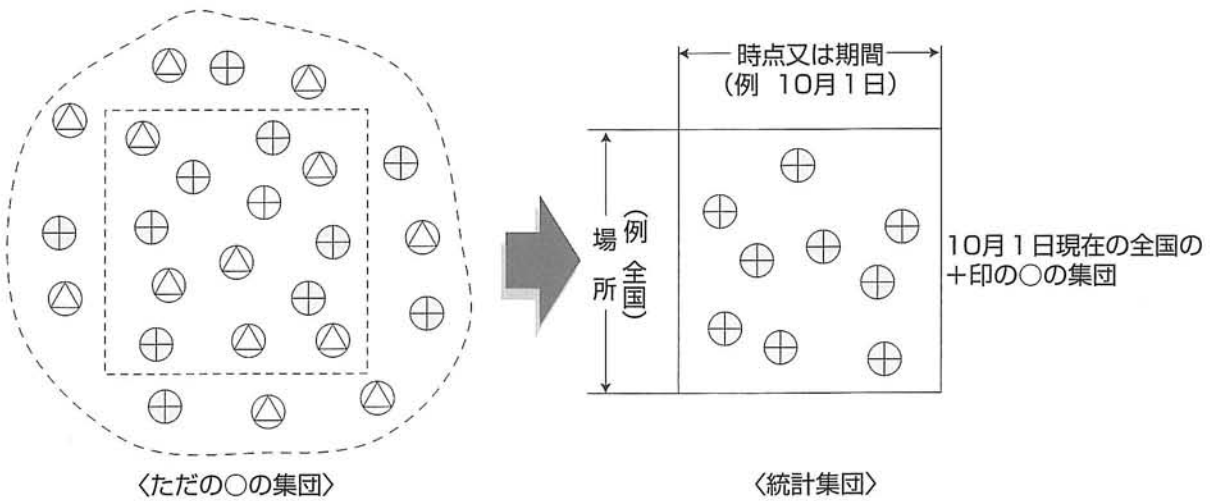
１．統計集団の規定

統計を作る場合、まずどのような集団について統計を作るのか、対象となる統計集団を確定する必要がある。

統計集団の規定に当たっては、それが人、物、出来事など具体的に何の集団であるのか、構成要素としての統計単位が具体的に何であるのかを客観的に定義しておく必要がある。

次いでその集団を観察するに当たって、いつの時点の集団であるのか、また、地理的にどの範囲のものであるのかといった、時間及び空間の両面からの規定が必要となる。

統計集団の規定



時間の規定の仕方

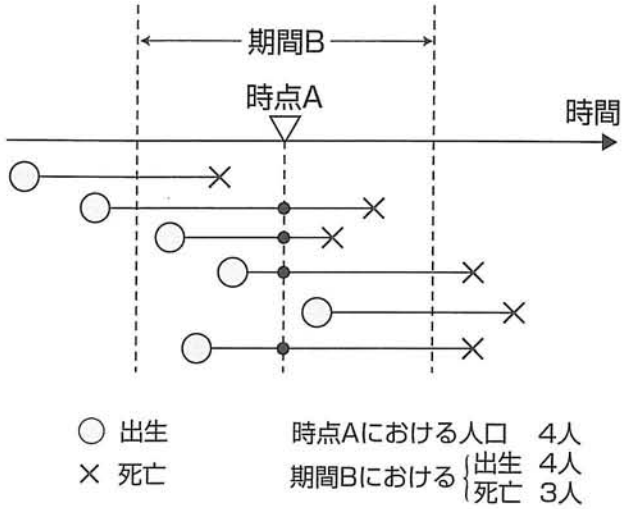
統計集団又は統計単位をどう規定するかは、どのような統計を必要とするかによって決まる。例えば、人は出生から死亡に至るまでの連続的な存在であり、同様に住宅も新築から滅失に至るまでの連続的な存在である。このような人又は住宅の状態について知りたい場合は、それぞれ人及び住宅の集団が対象となるが、構成要素としての一人ひとりの人及び各住宅は、時間軸の上ではそれぞれ異なった線分として表されるため、時点を定めて観察することとなる。

このような時間軸の上で線分として表される統計単位を構成要素とする統計集団を線集団又は静態統計集団と呼ぶ。

これに対して、人の出生数や死亡数、住宅の新築、滅失数を知りたい場合は、それらの出来事の集団が対象集団として規定されるが、各出来事は時間軸の上では瞬間として点で表されるため、一定の長さをもった期間を定めて観察することになる。

このような時間軸の上で点で表される統計単位を構成要素とする集団を点集団又は動態統計集団と呼ぶ。

線集団と点集団



2．観察（調査）事項の決定

統計集団が決定された段階で、その集団について具体的に何を知りたいのかを検討することとなるが、この作業は一般的には調査票の設計を通じて進められる。

統計集団を構成する各要素は、その集団に関して共通の性質（集団の標識）をもっている。その要素を数え上げたものがその集団の大きさになる。

各統計単位は、集団の標識以外にも多様な性質をもっている。そのうちどの性質を調査するかは、その集団について何を知りたいかによって決定されるべきである。調査事項とした性質（調査標識）については、誰もが判断に迷いが生じないような客観的な定義づけ(注)を与えておくことが重要である。例えば「年」という場合、暦年であるのか、年度であるのかが明確に区別されていなければならない。

調査結果としての統計を利用する場合も同様で、ある事柄が具体的にどのような意味で使われているのか、一般常識とは異なる場合が多いので、報告書に当たって確認しておく必要がある。

（注）人を対象とする調査の中には、思想、信条、その他外観からは推し量ることができない主観的な事柄を調査事項とする調査がある（世論調査、アンケート調査など）。

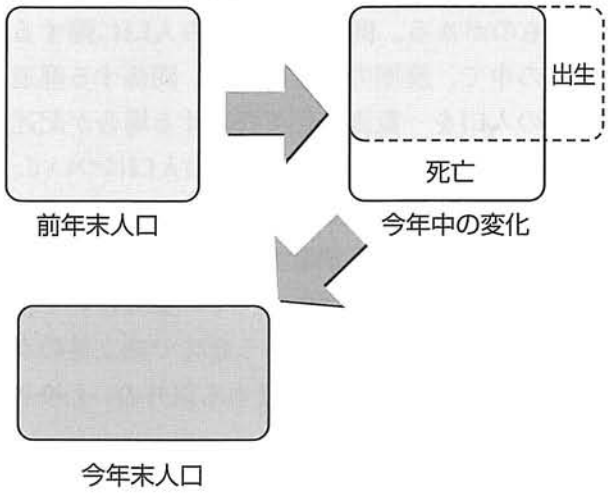
＜調査事項の類型 ― 静態事項と動態事項＞

統計集団において、静態集団と動態集団が区分されたのと同様、調査事項についても一定時点での出来事を調査する場合と連続的な状態を調査する場合とがある。前者を静態事項といい、調査に当たっては特定の時点が指定される。後者を動態事項と呼び、一定の長さを持った期間を定めて調査が行われる。

例えば商店について言えば、商品の在庫量は静態事項であり、どの時点における在庫量を調査するかが問題になるであろうし、販売量であれば、販売という行為の積み重ねであり、いつからいつまでのことか、その期間を定めておくことが必要である。

（参考）静態統計と動態統計とは、あたかもストックとフローの関係に似ている。ある時点における静態統計にその後の一定期間における動態統計を加減すれば、その時点におけるストック量を推計することができる。

静態統計と動態統計の結合



3．観察（調査）の方法

統計は、対象とする統計集団を構成する各統計単位を観察（調査）することによって作成される。この観察（調査）のやり方には、直接、間接の別を含めていろいろな方法がある。

＜第一義統計と第二義統計＞

第一義統計と第二義統計の別は、それがどのようなデータに基づいて作成されたのかという作成過程の相違による区分である。

第一義統計は、統計の作成自体を目的とするいわゆる統計調査の結果によって作成されたもので、調査統計とも呼ばれる。

これに対して第二義統計は、特別な調査をするのではなく、別の目的で個人や団体等から収集された資料を利用して作成された統計を言う。統計調査の場合と同様の方法で統計集団の定義づけを行った上で、それらに適合するものを集計して作成される。統計の作成を直接の目的とはしない別途の業務資料に基づいて言わば副次的、第二義的に作成されるものであり、業務統計とも呼ばれる。

国の統計では、輸出入のための通関手続きの一環として提出される資料を利用して作成される貿易統計や交通事故統計などがある。小・中学校の場合、学科試験などの結果を集計して得られる平均点等も第二義統計である。

利用する上で、第一義統計と第二義統計とは基本的には差違はないが、第二義統計の場合、原資料の収集目的によっては一定の偏りが生ずることがある。

＜全数調査と一部調査＞

統計を作成するための調査は、必ずしも対象となった統計集団を構成するすべての統計単位について行われるとは限らない。一般的には対象となる集団がそれ程大きくはないので全数調査となるが、特に国の調査の場合は、対象集団が非常に大きいため、特定の目的を持った一部の調査を除き、一部調査が主流となっている。

全数調査：統計集団を構成するすべての統計単位を対象に実施される調査で、その代表的なものとして国勢調査や事業所・企業統計調査などがある。対象集団がそれ程大きくない調査については、原則として全数調査が行われることとなる。

国が実施する基本的な統計調査については、全数調査が基本であり、理想であるとされてきた。しかし、

① 調査の計画、実施から最終の統計作成までに相当の長期間が必要である
② 調査規模が巨大であることから調査誤差のコントロールが困難である

などの理由から、また、標本調査理論及び技術の進展もあって、全数調査の実施は、

① 対象集団の大きさや構造を正確に把握する必要があるもの

② 他の標本調査のための母集団枠を提供するもの

などに限定されている。

一部調査（標本調査）：統計集団を構成する統計単位の一部を無作為(注)に抽出し、それらを調査した結果に基づいて元の統計集団全体の姿を推定しようとする調査方法である。

(注)「無作為抽出」とは、抽出結果に偏りが生じないように、抽出に際して作為を加えず、各単位が抽出される確率が同一となる抽出をいう。

標本調査の対象となる統計集団を特に母集団と呼ぶ。集団を構成する統計単位がすべて確定されており、通常、一覧表の形で名簿が用意されている。母集団の枠であり、標本調査は、この母集団枠の中から一定の方法で調査対象となる統計単位（調査客体）が標本として無作為に抽出される。

標本の抽出方法としては、① 単純無作為抽出法、② 集落抽出法、③ 層別抽出法、④ 多段抽出法などがある。どの方法によるかは、対象集団の構成、調査目的、調査経費などによって決められる。いずれの場合も母集団を構成する統計単位のすべてを調査した場合に得られるであろう結果との差、即ち標本誤差の大きさを理論的に計算できるのが大きな特色である。

なお、標本誤差の大きさは、当然、標本の抽出数によっても大きな影響を受けることになる。このため、調査目的に照らしてどの程度の精度が必要であるかによって標本の抽出数が決められる。

第2　統計表

統計調査結果等に基づいて各種の統計表が作成されるが、統計表には幾つかの約束事や利用するに当たって知っておいた方がいい事柄が少なくない。以下にその主なものを紹介する。

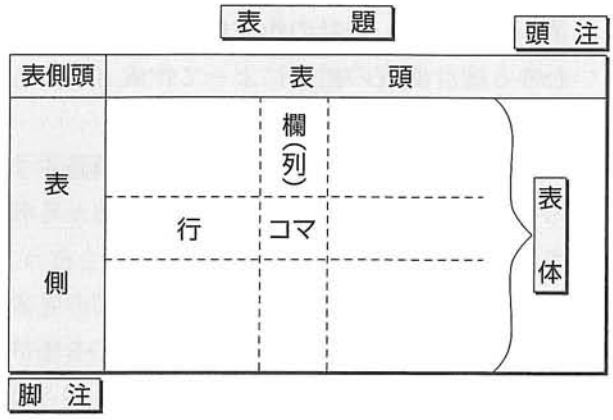
なお、統計表の中には、文章の一部として位置づけられる記述統計表及び挿入統計表と呼ばれるものがある。例えば我が国の人口に関する文章の中で、説明の便宜のため、関係する都道府県の人口を一覧表の形で掲載する場合が記述統計表であり、特定の都道府県の人口について、その経年推移の状況を文章で書く代わりに表で掲載する場合が挿入統計表である。いずれの場合も、文章を読まなければ、その意味合いや位置づけが分からない、という意味で独立性のないものである。ここでは、それら以外のいわゆる正式統計表を対象にする。

表　題

表題は、表番号と表名とからなる。

表番号は、単独の統計表にはつけられないが、幾つかの系列からなる複数の統計表である場合は、その系列の中での位置関係を示すとともに、索引としての機能を与えるために、第1表、第2表、……のように一連番号又は第1－1表、第1－2表、……のように前置又は後置番号を用いて表示される。

═正式統計表の構造と各部位の名称═



表名は、その統計表の内容を示す目録に相当するものであり、原則としてその統計表が対象としている統計集団の範囲、分類に関する事項及びその統計表が表している基本事項の三つの要素で構成されている。

① 平成12年10月1日現在の　② 全国の
③ 年齢別男女別の　④ 人口

①、②、④　：　集団の範囲
③　：　分類事項
④　：　基本事項

(注) ①及び ②は、一連の統計表に共通する場合は、一連の統計表全体の名称に含まれるため、省略されることが多い。

頭　注

その統計表の全体を理解する上で必要とされる事項が記載される。表名を補うという意味で表名の脚注と言われることもある。その統計表の全体の表示単位（例：人、トン、円など）が代表的な例である。

脚　注

特定のコマの数値に関する説明、表頭又は表側に関する定義の補足、資料の出所などが記載される。

表頭及び表側

何を表頭とし、何を表側とするかについては、必ずしも統一的な基準はない。一般的にはどのようにすれば利用者の注意を惹きつけ、その伝える統計数値の意味・内容を分かり易く示して利用価値が高いかという観点から通常、表頭には分類項目数が少ないもの、項目名が簡潔なもの、表章単位が他と異なるものが選ばれる。

表　体

表頭と表側に囲まれた統計数値の配置される部分をいう。

表体の縦の並びを欄又は列という。また、横の並びを行という。欄と行が交差する統計数値が配置される一つひとつの部分をコマ（セル）という。

第3　統計の解析

統計集団に対する統計的観察の結果から直接得られた統計数字を統計基礎数という。統計集団の大きさと内部構造の実態を統計単位の数という絶対数の形で表現したものである。第一次統計と呼ぶこともある。

これに対して、統計集団の諸特性を端的に表現するために、一定の方法を用いて統計基礎数から計算誘導された比率などの統計数字を統計誘導数という。

一般的には統計基礎数を主体とした統計表からだけでは、統計集団の内部構造や統計系列の変化の様子、さまざまな関係やある種の傾向などを読み取ることは困難なことが多く、また、誤解したり、見逃したりすることもある。

このため、理解や考察に便利なように相対数や中心的な代表値が求められ、また、数学的に解析し、関数的な関係を検討するために、相関係数を計算するなどの操作が行われる。このようにして統計基礎数を整理・加工したものが統計誘導数である。第二次統計又は加工統計ともいうが、ここでは、統計を解析する上で用いられる主な統計誘導数を紹介することとする。

1．分布の中心 — 平均値

算術平均：統計数値の総和を数値の個数で割った値である。分布の中心的な位置を把握する上で最も重要な数値である。各数値と平均値との差を偏差というが、偏差の総和が0となる点が算術平均の大きな特長である。

＜計算式＞

n個の統計データをx_1、x_2、……、x_nとし、その平均値を$\bar{x}$（エックス・バー）とする。

$$\bar{x}=\frac{x_1+x_2+\cdots\cdots+x_n}{n}$$

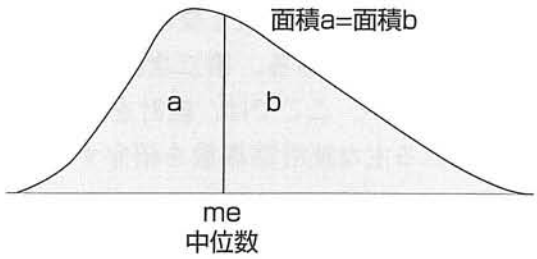
幾何平均：比率や成長率などの平均を求める場合に用いる。各比率の積を比率の個数で開いた根である。

$$G=\sqrt[n]{x_1\cdot x_2\cdots\cdots x_n}$$

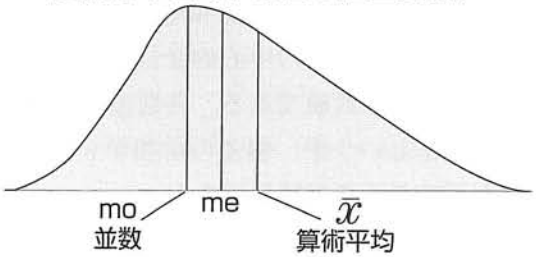
並数　Mode（モード）：度数分布表で最も度数の多い変数の値をいう。最頻値ともいうが、例えばあるテストの成績で、60点をとった生徒が最も多かったとすれば、60点が並数ということになる。

中位数　Median（メジアン）：ある系列の統計数値を大小の順に並べたとき、その中央に位置する数値をいう。中央値ともいう。統計数値が偶数の場合は、中央の２つの数値の平均が中位数である。算術平均は、極端に大きい値があるとその影響を受けて、中心的な位置の判断が困難となることがあるので、算術平均を補完する指標として用いられる。

度数分布図（中位数）



度数分布図（中心値の位置関係）



2．分布の広がり―分散及び標準偏差

幾つかの集団を比較する場合、中心の位置を比較するだけでは不十分であり、各データが中心の位置（平均値）からどれだけ散らばっているか、その度合いを示す指標が必要となる。分散及びその平方根である標準偏差がその指標である。

分　散

各データの中心からの散らばりの程度を端的に示すのは偏差である。偏差の総和は０となるので、偏差を平方した上でその総和を求め、データの個数で割って平均を計算したのが、散らばりの指標としての分散である。

n個の統計データx_1、x_2、……x_nの平均を$\bar{x}$とすると、分散Vは次式で与えられる。

$$V=\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}$$

上式または

$$V=\frac{(x_1-\bar{x})^2+(x_2-\bar{x})^2+\cdots+(x_n-\bar{x})^2}{n}$$

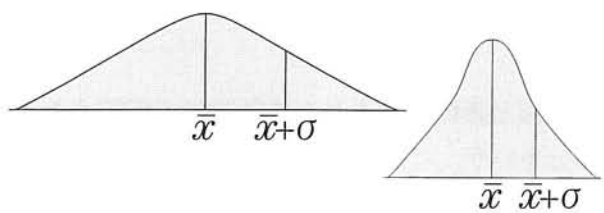
標準偏差

分散は、偏差の平方和で計算されているため、単位は各データの単位の２乗となっているので、各データの単位と同一の単位とするため、分散の正の平方根を計算したのが標準偏差である。言わば偏差の平均であり、標準偏差が大きい程、算術平均値からの散らばりの大きい分布となる。

標準偏差をσ（シグマ）とする。

$$\sigma=\sqrt{V}=\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}}$$

度数分布図（標準偏差）



標準偏差等の計算例

区分	成績 x	$(x-\bar{x})$	$(x-\bar{x})^2$
①	98	19	361
②	97	18	324
③	94	15	225
④	93	14	196
⑤	90	11	121
⑥	83	4	16
⑦	80	1	1
⑧	75	−4	16
⑨	73	−6	36
⑩	60	−19	361
⑪	55	−24	576
⑫	50	−29	841
計	948	0	3074

平均値$\bar{x}=\frac{948}{12}=79.0$　分散$V=\frac{3074}{12}=256.16\cdots$

標準偏差$=\sqrt{256.2}=16.0$

変動係数

標準偏差は、言わば偏差の平均であることから、その大きさは平均値の大きさによって影響を受ける。また、単位の異なる集団との間では比較はできない。標準偏差を平均値で割ることによって無名数化したものが変動係数である。

変動係数を計算することによって、平均値からの相対的な散らばりの程度が把握でき、例えば平均賃金が大きく異なる戦前と戦後の賃金格差の状況を比較することができる。また、円とドルという表示単位の異なる日米の賃金分布の状況を直接比較することもできる。

変動係数をＣＶとする。

$$CV=\frac{\sigma}{\bar{x}}$$

3．統計比例数

統計比例数とは、統計的関係を表現する一つの方法であり、単一の数値をもって表された二つの集団の大きさを比の形で関係づけた統計誘導数である。

統計比例数の本質は、分母と分子の形で二つの集団を関係づけ、それによって統計基礎数のままでの比較では明確に把握できない統計的関係を明らかにすることである。統計比例数は、分母となる数値と分子となる数値の性格によって、次の3種に区分される。

(1) 構成比率

一つの統計集団は、統計分類によっていくつかの部分集団に分割される。部分集団を分子とし、集団全体を分母とする比率を構成比率という。通常、百分率（%）で表される。

構成比率の例

$$\text{高齢化率}=\frac{\text{65歳以上人口}}{\text{総人口}}\times 100\,(\%)$$

$$\text{エンゲル係数}=\frac{\text{うち食料費}}{\text{家計消費支出}}\times 100\,(\%)$$

(2) 発生比率

ある集団を母体として別種の集団が生ずることがある。例えばある特定地域における人口を母体として、１年間に出生した新生児の集団がそれである。分子を新生児の集団とし、分母を人口とする比率を発生比率という。通常、百分率％又は千分率‰（パーミル）で表される。

出生数などの発生集団の大きさは、当然、母体となる集団の大きさによって異なるので、母体となる集団の大きさを百又は千とする比率を計算することによって、地域間の比較や時系列変化の状況の把握が可能となる。

発生比率の例

$$出生率=\frac{1年間の出生数}{年央人口}\times 1000\,(‰)$$

$$自動車千台当たり事故件数=\frac{自動車事故件数}{自動車保有台数}\times 1000\,(‰)$$

(3) 指　数

同じ種類の統計数値を分母と分子にする統計比率である。分母となる統計数値を固定し、比較の対象となる数値を分子として相互の大小等の関係を比較・吟味する場合に用いられる。分母となる数値を指数の基準という。

指数の分母と分子は、同じ単位の統計数値であるため、無名数となる。そのため、例えば生産数量と生産金額のように単位の異なる統計数値であっても、それぞれを指数化することによって、時系列変化の状況等を相互に比較することができるようになる。

指数が用いられるのは、主として地域間の相違の実態を比較する場合と時系列変化の状況を把握する場合である。前者を静態指数、後者を動態指数というが、消費者物価指数や鉱工業生産指数などのように一般的には後者の動態指数が多く用いられる。

4. 寄与度と寄与率

幾つかの複数の要素からなる統計事象が増加(減少)した場合、どの要素によってどれだけ増加(減少)したのか、その程度を示すために寄与度と寄与率が用いられる。例えば各月の物価が上昇した場合、生鮮食品の上昇がどの程度物価全体を押し上げたのか、その内訳を分析する場合に用いられる。

寄与度は、その要素によって全体がどれだけ増加(減少)したのか、その内訳としての増加(減少)率を表す。各要素の寄与度の合計は、全体の増加(減少)率に等しくなる。

寄与率は、各要素の寄与度の全体の増加(減少)率に占める構成比率を百分率で示したものである。

5. 相関係数と回帰係数

2つの変数、例えば成長期にある児童・生徒の身長と体重のように、一方が増加すると他方も同じように増加する場合、2つの変数の間には正の相関があるという。逆に一方が増加すると他方が同じ程度に減少する場合、負の相関があるという。両者を合わせて相関関係という。

また、2つの変数の間の相関関係が直線的である場合であって、その関係の強さを測る指標を相関係数rという。相関係数は1から−1の間の値をとり、それが±1に近い程、2つの変数の関係がより強いことがわかる。

正の値の相関係数は正の相関に対応し、負の値の相関係数は負の相関に対応する。

相関係数の計算表

区分	平均身長 x	平均体重 y	x^2	xy	y^2
5歳	110.8	19.3	12276.64	2138.44	372.49
6歳	116.8	21.6	13642.24	2522.88	466.56
7歳	122.5	24.3	15006.25	2976.75	590.49
8歳	128.1	27.3	16409.61	3497.13	745.29
9歳	133.4	30.8	17795.56	4108.72	948.64
10歳	138.8	34.8	19265.44	4830.24	1211.04
11歳	145.0	39.1	21025.00	5669.50	1528.81
12歳	152.7	44.9	23317.29	6856.23	2016.01
13歳	160.0	50.2	25600.00	8032.00	2520.04
14歳	165.5	55.4	27390.25	9168.70	3069.16
15歳	168.6	60.1	28425.96	10132.86	3612.01
16歳	169.9	61.9	28866.01	10516.81	3831.61
17歳	170.7	63.5	29138.49	10839.45	4032.25
合計	1882.8	533.2	278158.74	81289.71	24944.40

(注) 身長と体重は、昭和61(1986)年誕生の男子の各歳の平均値
出所：文部科学省「学校保健統計」より作成

身長の平均値　$\bar{x}=\frac{\Sigma x}{n}=\frac{1882.8}{13}=144.83$

体重の平均値　$\bar{y}=\frac{\Sigma y}{n}=\frac{533.2}{13}=41.02$

相関係数 $=\frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2}\sqrt{\Sigma(y-\bar{y})^2}}$

$=\frac{\Sigma xy-n\bar{x}\bar{y}}{\sqrt{(\Sigma x^2-n\bar{x}^2)(\Sigma y^2-n\bar{y}^2)}}$

$=\frac{81289.71-13\times144.83\times41.02}{\sqrt{(278158.74-13\times144.83^2)\,(24944.40-13\times41.02^2)}}$

$=\frac{4057.66}{\sqrt{5474.26\times3070.07}}=0.990$

身長と体重との間には、相関係数0.990という非常に強い正の相関関係が認められることから、両者の間には、

Y＝a X＋b

という一次式（回帰式）を想定することができる。a とb は回帰係数と呼ばれ、、次式により計算される。

$a=\frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(x-\bar{x})^2}=\frac{\Sigma xy-n\bar{x}\bar{y}}{\Sigma x^2-n\bar{x}^2}$

$=\frac{4057.66}{5474.26}=0.741$

$b=\bar{y}-a\bar{x}=41.02-0.741\times144.83=-66.30$

したがって、男子児童の身長Xと体重Yの関係は、次の一次式で表される。

Y＝0.741 X－66.30

これによると、身長150cmの時の児童の平均体重は、44.9kgであると理論的には想定される。

（注）小数点以下の部分で四捨五入したため、パソコンによる計算結果とは多少異なる。