2017.7 第8版

6章 医療統計I

統計理論

6. 医療統計Ⅰ（統計理論）

医学・医療の分野において、統計解析が頻繁に用いられるようになった。たとえば、疾病統計、がん登録、病院経営、臨床研究、疫学調査、診療圏分析など、さまざまな用途で統計解析が行われ、診断や治療の支援、経営方針の決定、医療行政の資料などに活用されている。

診療情報管理士は診療情報を適切に管理することのみならず、その利活用も業務のひとつとなっている。したがって、診療情報管理士は医療統計学の実践的能力を身につけ基本的な統計解析を実行できること、ならびに、その理論的背景を理解することが必要である。

本章では、診療情報の利活用を実際に行うための実践的能力と解析結果の解釈、統計理論の習得をめざす。

6-1 実務で用いる統計解析

実務で統計解析を行う際のおおまかな手順と留意点を以下にまとめる。

1）統計解析の目的の明確化

統計解析の目的を明確に記述する。

〔具体例〕

（1）本院の 2018 年 1 月から 12 月までの月別外来患者数の増減の特徴を明らかにする。

（2）原発乳癌に対する術前化学療法 A と B との腫瘍縮小効果の優劣を評価する。

（3）クリニカルパス導入による在院日数短縮の効果を調べる。

2）目的の達成に必要となるデータ収集

（1）統計解析用データを紙カルテ、電子カルテ、アンケート調査（紙、Web など）、年報、白書、その他の資料から収集する。

（2）目的達成のための過不足のない変数を選定する。データ収集の前に、統計解析ではどういう集計・解析をするのかを緻密に計画することが必要である。

（3）欠損値の多い変数は収集しても利用できない可能性がある。

（4）収集する変数名だけではなく、どれくらいの症例数を収集できるのかの見積りを立てる。少数例では統計解析ができない場合がある。

3）データファイルの作成

収集したデータをエクセル（Microsoft Excel）などの表計算ソフトや統計解析ソフトウェアに入力する。電子カルテからの自動抽出ができる場合には、データ変換が正しく行われたかを目で見てチェックする必要がある。また、紙媒体からの手入力作業がとも

なう場合には、入力データと紙資料とのチェックが必要である。

4）統計解析ソフトウェアを用いた統計解析

エクセルもしくは統計解析ソフトウェアで統計解析を行う。医学や医療関係者が用いる統計解析ソフトウェアとしては、SPSS，JMP，R，EZR，STATA，SAS などがある。R や EZR は無料のソフトウェアである。EZR は R を対話型で処理できるように工夫されたソフトウェアで、マニュアル本も出ているので初心者でも利用しやすい。

5）統計解析結果の解釈

統計解析結果の正しい解釈を記述して、人に説明できるようにしておく。統計学のテキストの勉強のみでは解釈力はつきにくいので、実践的な解釈を数多く行う。

6-2 医療統計学の基本的な用語

医療統計学の実践的な統計解析の方法とその理論を解説する前に、基本的な用語を理解することがこの節の目的である。

【データファイル】

統計解析は数例から数万例の症例（患者もしくは健常者など）のデータを収集して行われる。統計解析に用いられるデータのまとまりのことをデータファイル（またはデータセット）と呼ぶ。データファイルとは複数の症例のデータをひとかたまりにしたファイルである。ここでは、F 市で行われた 1 ヶ月時乳幼児健診のデータファイルの例を表 1 に示し、データファイルの構造と基本的な用語を説明する。

表 1．F 市の 1 ヶ月時乳幼児健診のデータファイル

症例番号	母親年齢	父親年齢	何人目	性別	母親喫煙	父親喫煙	妊娠中異常	出生体重*（BW）
1	31	40	3	2	0	1	1	2,433
2	31	37	3	1	0	1	0	3,591
3	34	30	2	2	0	1	0	2,866
4	22	31	1	1	0	1	1	4,000
5	29	28	1	2	0	0	0	2,680
6	23	27	2	1	0	1	0	2,950
7	34	34	3	1	0	1	0	3,700
8	27	31	1	1	0	0	0	2,984
9	25	26	2	1	0	1	0	2,800

＊出生体重（BirthWeight：BW）

行方向に、症例番号1の症例の母親年齢31歳、父親年齢40歳、3人目の赤ちゃん、…とデータが並んでいる。また、列方向に母親の年齢、父親の年齢、何人目の赤ちゃんか、…のデータ項目が示されている。これを例に、いくつかの用語を定義する。

①**症例、個体、被験者**：行方向のひとりずつ（1行ずつ）。
②**変数（または変量）**：列方向の母親の年齢、父親の年齢などのデータ項目。
③**データ、観測値、測定値**：BW 2,866g、父親の年齢34歳などの個々の値。もしくは、複数個まとめてデータとよぶこともある。
④**コード（カテゴリー）**：性別の男と女をそれぞれ1と2に置き換えているが、この1，2のこと。喫煙のありの1、なしの0も同様である。統計解析では男、女などの文字より数値の方が統計処理しやすいため、コード変換が行われる。コードもデータである。

【変数の種類】

統計解析で扱われる変数は表2に示される4種類である。

表2．変数の種類

大分類	小分類	変数の例	データの呼び方
量的変数			量的データ
	連続変数	年齢、身長、体温	連続データ
	整数変数	家族の人数、投与回数	整数データ
質的変数			質的データ
	名義変数	性、診断名、血液型	名義データ
	順序変数	満足度、臨床進行期	順序データ

1）量的変数

量的変数には、連続的な値をとる変数と整数値をとる変数の2種類が含まれる。それぞれのデータを連続データ、整数データ、2つをまとめて量的データとよぶ。連続データとして、年齢は65歳、70歳が典型的な例であるが、場合によっては、65歳3ヶ月とか65歳110日とか、さらに連続的に詳しいデータを収集しうるものである。

〔補足1〕連続変数の例として年齢をあげた。年齢はふつう59歳、64歳などの整数で表されるが、月日まで考慮すると59.33歳などと表すことができる。一方、家族の人数は3.2人というデータは存在しない。整数より細かなデータを取得できるか、整数しかとらないかで連続変数と整数変数には違いがある。

2）質的変数

質的変数には、名義がデータとなる変数と順序のある値をとる変数がある。それぞれのデータのことを名義データ、順序データ、2つをまとめて質的データとよぶ。血液型という変数ではデータとして、A型、B型、AB型、O型という4種類のデータがある。

また、順序のある値をとる変数では満足度があるが、そのデータとして、大いに満足（5点）、少し満足（4点）、どちらでもない（3点）、やや不満足（2点）、非常に不満足（1点）のようにとられることがある。

〔補足１〕血液型はA型、B型、O型、AB型の名義が本来のデータである。統計解析ではこれらを1，2，3，4などとコード化するが、数値の大小に意味はない。

〔補足２〕満足度では、大いに満足、ほぼ満足、どちらでもない、やや不満、大いに不満の5つのカテゴリーが考えられる。これらを5，4，3，2，1とコード化したものは、数値が小さいほど不満度が増すという順序がある。このように順序、重症度、程度が加味されたデータである点が名義変数と異なる。

【統計量】

データから算出される統計処理で必要な量のことを統計量とよぶ。平均、標準偏差、割合、生存率などの量はすべて統計量である。

【分布】

分布とはデータのばらつきのことをいう。

〔例：分布〕

ある疾患の在院日数データ（単位は日）が以下のようにあるとする。

{11，12，12，13，13，13，13，14，14，15}

このデータの分布に関して以下のことが記述できる。

1）データの個数は10個、すなわち、10患者のデータである。

2）13日が度数最多で全体の40%を占める。

3）データは13日を中心に左右対称にばらついている。すなわち、分布の中心は13日で分布の形は左右対称である。

4）分布の範囲は11日から15日までである。

【度数分布】

度数分布とはデータのばらつきに、その度数（データの個数、人数、件数）を加えたものをいう。データの度数を「総数に対する度数の割合」に置き換えたものを相対度数分布という。度数分布は度数分布表、棒グラフ、ヒストグラムなどで表される。

〔例：度数分布表〕

ある年のN県の死因別死亡数のデータである。

表 3. N 県の死因別死亡数

死因	人数（人）	相対度数（%）
悪性新生物	7,876	27.3
心疾患	4,155	14.4
脳血管疾患	2,875	10.0
肺炎	2,278	7.9
自殺	496	1.7
その他	11,142	38.7
合計	28,822	100.0

度数分布表に相対度数も付け加えられた表である。この表からわかることは以下の 3 点である。

1）悪性新生物による死亡が全体の 27.3% を占めている。

2）3 大死因（悪性新生物、心疾患（高血圧性を除く）、脳血管疾患）による死亡割合は 51.7% である。

3）3 大死因に肺炎を含めた主要死因の割合は 59.6% であり、全体の 6 割を占める。

〔例：棒グラフ〕

日本胃癌学会の全国胃癌登録の集計結果に基づき、作成された占拠部位長軸の度数分布表と棒グラフである。棒グラフの横軸の変数は質的変数、縦軸は度数または割合である。

胃癌の占拠部位長軸で E，U，M，L，D はそれぞれ、食道、上部、中部、下部、十二指腸をさす（胃癌取扱い規約第 14 版）。UE は食道または上部に腫瘍があったことを意味する。占拠部位長軸は名義変数である。

表 4. 占拠部位長軸別の胃癌患者数

占拠部位長軸	人数	割合（%）
UE	4,720	21.9
M	8,446	39.2
LD	7,640	35.5
MUL	731	3.4

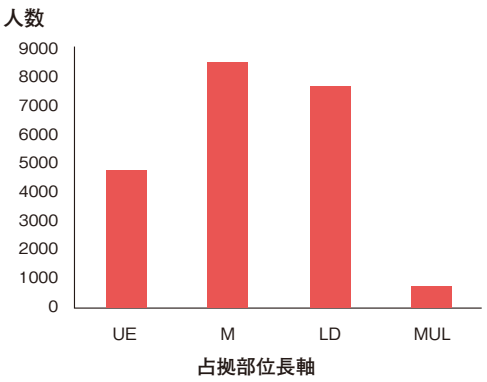


図 1. 占拠部位長軸別の胃癌患者数

この度数分布から、日本の胃癌患者は中部と下部または十二指腸に発症する頻度が高いことがわかる。これらの分布がわかっていると、自院での胃癌患者の占拠部位が全国の胃癌の占拠部位の分布とどこが違うのか、また、海外の胃癌患者と占拠部位ではどういう違いがあるのか、などを検討することができる。

〔例：ヒストグラム〕

診療情報管理士が1日に診療情報提供の業務に要した時間の度数分布表とヒストグラムを以下に示す。ヒストグラムの横軸が業務時間、縦軸が人数である。業務時間は量的変数（連続変数）であり、分布を調べる際にはデータをいくつかの階級（○以上△未満など）に分ける必要がある。このとき、階級の幅は等しくすること、端点がどちらに入るのか（以上、未満、より大きい、以下）をはっきり明示することが重要である。

表 5. 診療情報管理士の 1 日の診療情報提供業務時間

業務時間（分）	人数（人）
0	4
1 ～ 29	19
30 ～ 59	12
60 ～ 89	5
90 ～ 119	4
120 ～	6

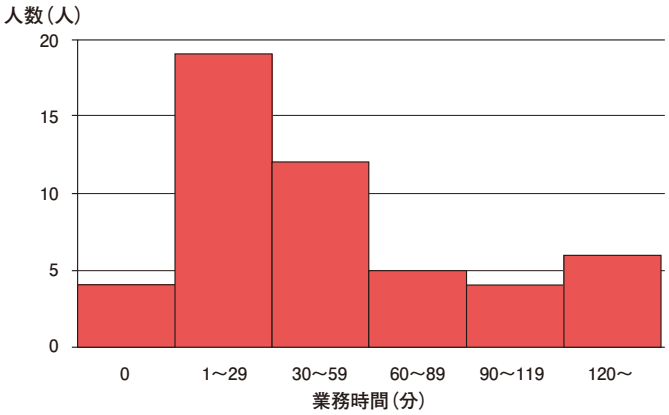


図 2. 診療情報管理士の 1 日の診療情報提供業務時間

ヒストグラムの横軸の階級は、30 分以上 59 分以下と読む。両側は 0 分と 120 分以上である。このヒストグラム（分布）からわかることは、30 分以内と 30 分以上 60 分未満の人数が多いこと、60 分以上は 3 つのカテゴリーで 5 人程度とほぼ一定であることがわかる。

【正規分布】

ひとつ山で左右対称のつり鐘型の度数分布のことを正規分布と呼ぶ。

正規分布がどういう性質をもち、実践的な統計解析でどのように使われるのかは、p.371「6-5 正規分布」で説明する。ここでは正規分布の例を示す。

〔例：正規分布〕

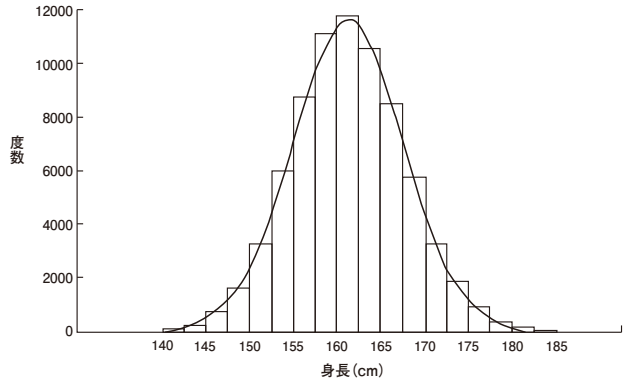


図 3. 身長のヒストグラムと正規分布曲線

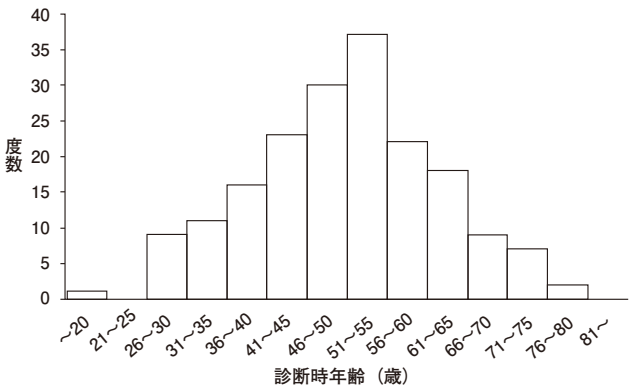


図 4. 肝硬変患者の診断時年齢のヒストグラム

〔例：非正規分布〕

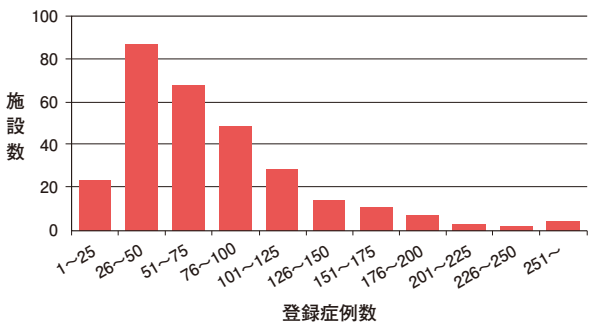


図 5. 胃癌登録事業の登録症例数別の施設数

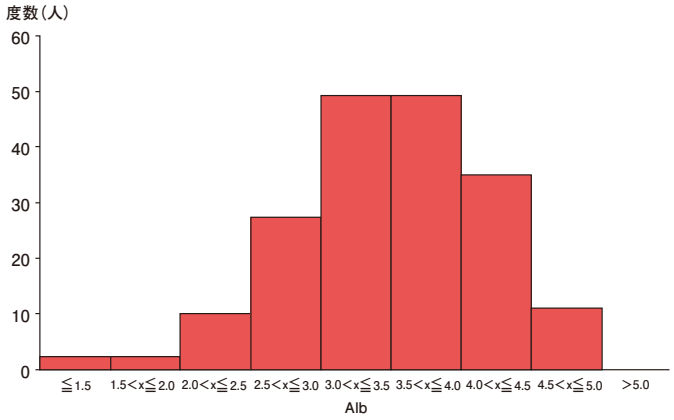


図 6.　肝硬変患者の血清アルブミン値（Alb）の分布

6-3　データを要約する

病院内のデータには診療データ、診療報酬データ、経営データ、がん登録データなどがある。また、インフルエンザ発生に関するデータは市内や県内で収集される。これらの数値や文字の羅列を眺めていてもその特性は把握できない。

そこで、収集したデータを要約することを考える。要約の仕方には次の 2 種類がある。

1）図表の作成

図表の作成はデータの分布（データのばらつき）をわかりやすく図や表で表すことである。データのばらつきを度数分布表、棒グラフ、ヒストグラムなどで示すことをいう。

2）記述統計量（Descriptive statistics）の算出

※ 要約統計量（Summary statistics）ともいう。

記述統計量は分布の特性を数値で表すために使われる。記述統計量には、後述する平均や標準偏差、割合などがある。p.353「変数の種類」で述べたように、変数は量的変数と質的変数に大別されたが、変数の種類によって用いる記述統計量は異なる。以下では、それぞれの変数について分布の特性を表す際に使われる統計量を説明する。

6-3-1　データの要約の作業手順

データの要約は以下の手順で行われる。

1）記述統計量の算出や図表作成の目的を明確にする。

2）診療記録や医事会計データ、経営データを適切に抽出する。

3）エクセルや統計解析ソフトウェアを用いて集計や図表の作成を行う。

4）図表の特徴を適切に記述する統計量を算出する。

5）集計結果や図表の解釈を記述する。

【データの要約の例１：図表を作成して特性をつかむ】

データの要約のうち、図表を作成してデータの特徴を把握する例を上述の手順に従って説明する。

1）目的の明確化

診療情報管理士が診療情報開示請求に対応する業務量の変化を定量的に把握する。

2）データの適切な抽出

業務日誌に基づき、診療情報管理士が対応した開示請求の問い合わせと請求の件数を年次別に収集する。

3）図表の作成

エクセルに以下の表を作成して折れ線グラフを作成する。

表 6. 診療情報管理士が対応した開示請求の問い合わせと請求の件数の度数分布表

年度	問い合わせ件数	請求件数
2010	15	8
2011	18	10
2012	26	21
2013	37	30
2014	59	37
2015	71	39
2016	85	43
2017	89	44

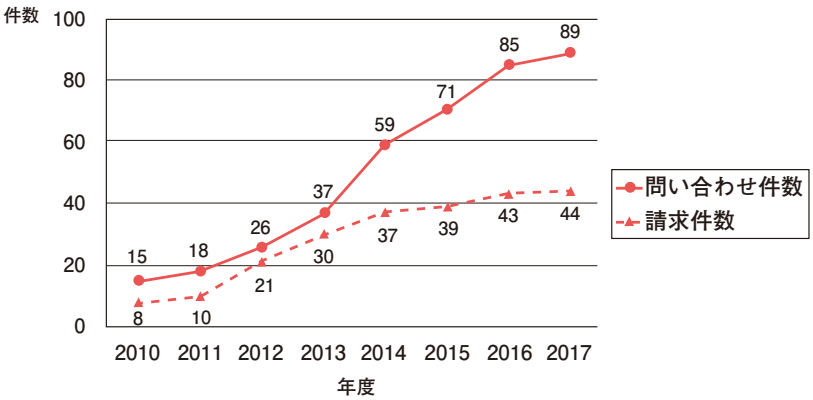


図 7. 診療情報管理士が対応した開示請求の問い合わせと請求の件数の年度別変化

4）図表の解釈

（1）表とグラフから、問い合わせ件数、請求件数ともに年々増加していることがわかる。

（2）グラフから、2013 年以前に比べて 2014 年以降は、問い合わせ件数が急に伸びているのに対して、請求件数の伸びは緩やかであることがわかる。

【データの要約の例 2：図表の作成と記述統計量の算出】

記述統計量によりデータの特徴を要約する例をサンプルデータ＊A：肝硬変データ 100 例から示す。

＊サンプルデータは A，B，C の 3 データあり、それぞれ通信教育 Web サイトからダウンロードできる。

1）目的の明確化

K 病院の肝硬変患者の診断時年齢の分布とその特徴を把握する。

2）データの適切な抽出

ある期間に K 病院の消化器内科を受診した肝硬変患者 100 例について、紙カルテからサンプルデータ A に示した変数のデータを収集した。一部、データが欠損しているところがある。

3）図表の作成と記述統計量の算出

エクセルを用いて以下のヒストグラムを作成し、記述統計量を算出した。

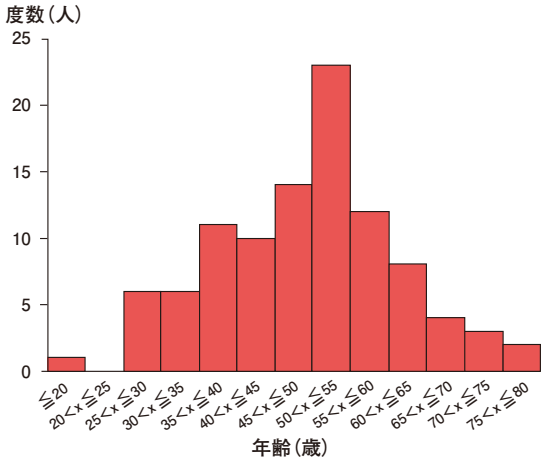


図 8. K 病院の肝硬変患者の診断時年齢の分布

図 8 は横軸に診断時年齢を 5 歳刻みの階級で表示して、縦軸を度数（人数）としたヒストグラムである。

4）図表の解釈と記述統計量

ヒストグラムはひとつ山の左右対称であり、正規分布に近い形となっている。この分布の中心位置と広がり具合を数値で示すために、平均と標準偏差を求めたところ、それぞれ 49.8 歳と 11.9 歳であった。このデータの最小値と最大値は 15 歳と 79 歳であり、最大値と最小値の差は 64 歳と広範囲であることがわかった。

6-3-2 データの要約に用いる記述統計量

データを要約する際には p.354 ～ 358「度数分布」「正規分布」で例示した図表を作成する方法がある。図表を見ると例示の解釈のように、データの特徴がつかみやすくなる。図表はデータ要約の方法として有用なツールであるが、図表の解釈をさらに数値で示すことを求められることが多い。

データの要約を数値で示すとき、その数値のことを記述統計量（もしくは要約統計量）とよぶ。記述統計量にはいろいろあるが、変数の種類別にその使い方を説明する。

1. 量的データの記述統計量

サンプルデータ* B：乳幼児健診データ 100 例の中で、母親の出産時年齢は量的データのひとつである。母親の出産時年齢のデータの要約を p.358「6-3-1 データの要約の作業手順」の手順に従って行ってみる。

*サンプルデータは A，B，C の 3 データあり、それぞれ通信教育 Web サイトからダウンロードできる。

データの要約の目的は、F 市の母親の出産時年齢の分布は全国の分布と比較してどういう違いがあるかを明確にすることである。全国の出産時年齢に比べて高齢出産であるのか、もしくは、出産時年齢のばらつきが大きいのか、などに興味があったとする。

用いるデータは F 市医師会が乳幼児健診で収集した 100 例分のデータである。

エクセルのヒストグラム作成機能を使って図 9 が得られた。このとき、要約されたヒストグラムを参照しつつ、データ要約の目的から次の点に興味がある。

①出産時年齢のヒストグラムの中心位置は何歳くらいか？

②何歳が一番多いか？

③出産時年齢のばらつきはどれくらいか？

④一番高い年齢は何歳か、また、一番低い年齢は何歳か？

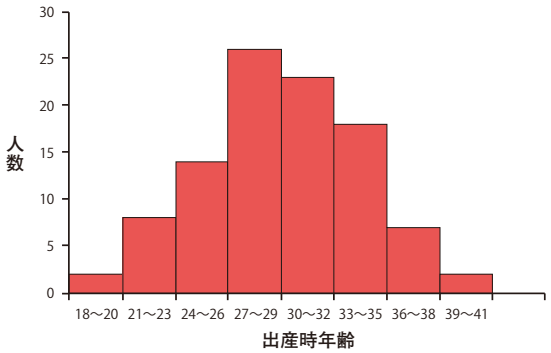


図 9. 母親の出産時年齢ヒストグラム

【代表値と散布度】

これらの興味は大きく2つに分けられる。ひとつは①と②であり、これらは分布（ヒストグラム）の位置を問題にしている。この位置を表す統計量のことを代表値（Measure of location）とよぶ。もうひとつは③と④であり、これらはデータのばらつき具合を問題にしている。分布のばらつきの統計量は散布度（Measure of Spread）とよぶ。

ひとつの分布（ヒストグラム）の特徴を記述する際に、代表値と散布度の2つを組み合わせて用いる。量的変数の分布を要約する際には、以下の記述統計量を用いる。なお、それぞれの記述統計量の求め方は次節で説明する。

1）変数の分布がひとつ山で左右対称（正規分布）に近いとき、代表値として平均（Mean）、散布度として標準偏差（Standard Deviation：S.D.）を用いる。
2）変数の分布がひとつ山ではない、もしくは、左右非対称のとき、代表値として中央値（Median）、散布度として［25％点（25th percentile），75％点（75th percentile）］を用いる。

ここでは、分布がひとつ山で左右対称のときとそうでないときで、代表値と散布度が異なるという点を覚えておく。

2. 代表値と散布度

量的データの分布の特徴づけを行う際に、代表値と散布度を組み合わせて用いる。ここでは、それらを詳しく述べる。

【代表値】

1）平均（Mean）

正規分布に近い形の分布の中心位置を示すときに用いられる。

〔求め方〕n 個の量的データ $\{x_1, \cdots, x_n\}$ があるとき、その平均 $\overline{x}$ は

$$\overline{x} = \frac{1}{n}\{x_1 + \cdots + x_n\}$$

で求める。すなわち、データを全部足し合わせてデータの個数で割った値が平均である。

〔例1〕健常者5名のBMI（体格指数）のデータが {20.0，21.3，21.5，22.2，23.0} とすると、平均は

$$\overline{x} = \frac{1}{5}\{20.0 + 21.3 + 21.5 + 22.2 + 23.0\} = 21.6$$

である。21.6は5つのデータのほぼ真ん中の値になっている。エクセルで平均を求

めるには、入力したデータを選択後、ホームタブ⇒編集リボン⇒Σ 記号の中の平均で求める。もしくは＝ AVERAGE（データ配列）を使う。

〔例２〕健常者 5 名の BMI データが {20.0，21.3，21.5，22.2，30.5} であるとすると、平均は

$$\overline{x} = \frac{1}{5}\{20.0+21.3+21.5+22.2+30.5\} = 23.1$$

となる。5 番目の 30.5 がきわめて高い値である。他の 4 つのデータは 23.0 未満であるにもかかわらず、30.5 があるために平均は 23.1 となり、5 つのデータの中心位置を示しているとは言い難い。このように、異常に大きな値（小さな値）がある場合、平均値はその値に引きずられて大きく（小さく）なり、分布の中心位置を示さないことが起こる。

〔解説〕例 2 の 30.5 は BMI のひとかたまりから著しく外れている。このようなデータを「外れ値」(outlier) とよぶ。外れ値を含むデータの記述統計量を求める場合には、次のいずれかを行う必要がある。

①平均でなく次に述べる中央値を用いる。

②外れ値を除外して平均を算出する。

2）中央値

分布が左右非対称であっても、また外れ値があっても、分布の中心位置を記述できる代表値である。

〔求め方〕データを小さい方から大きい方に順（昇順という）に並べたとき、真ん中の順位に当たる値を求める。具体的には

①得られたデータを小さい方から大きい方へ順に並べる

②観測値の個数 n が奇数のとき、

⇒ $\{(n+1)/2\}$ 番目の値

③観測値の個数 n が偶数のとき、

⇒ $\{n/2\}$ 番目と $\{(n/2)+1\}$ 番目の値の平均

〔例１〕A 病院における誤嚥性肺炎の在院日数が {14, 10, 12, 19, 15} であるとする。データは 5 個で奇数であり、昇順に並べ替えると

{10，12，14，15，19}

(5＋1) /2 ＝ 3 番目の値は 14 なので中央値は 14 である。エクセルでは＝ MEDIAN（データ配列）で求める。このデータでは平均も 14 である。

〔例２〕上述「平均」の例 2 のデータ {20.0，21.3，21.5，22.2，30.5} で中央値を求めてみよう。データは 5 個で奇数なので、順位が 3 番目の値 21.5 が中央値となる。

〔解説〕中央値はデータを順位に置き換えて、そのちょうど真ん中の順位に対応する元

のデータである。データの値そのものを使っていないので、平均のように外れ値の影響で著しく大きくなったりすることがない。「平均」と「中央値」の例2の比較でわかるように、中央値は外れ値の影響を受けにくく、5個のデータの中心は4個が固まっている中心付近の値を指し示している。

3）最頻値

データの度数が最も高い値のことを最頻値（Mode）という。

〔例1〕在院日数データ {7，8，9，9，10，10，10，11，12，12，13} の最頻値は10日。エクセルでは＝mode（データ配列）で求める。

〔例2〕在院日数データ {7，8，9，9，10，10，10，11，12，12，12，13} の最頻値は10日と12日。このように最頻値が2個ある場合を二峰性の分布とよぶ。

【散布度】

量的データの分布のばらつき（広がり）を記述するには、次の統計量が使われる。

1）標準偏差（Standard deviation：S.D.）

正規分布に近い場合にその分布のばらつきの程度を表す。

〔求め方〕n 個の量的データ $\{x_1, \cdots, x_n\}$ があり、その平均を $\overline{x}$ とする。標準偏差 $\overline{S}$ は以下の式で求められる。

$$\overline{S}=\sqrt{\frac{1}{n-1}\{(x_1-\overline{x})^2+\cdots+(x_n-\overline{x})^2\}}$$

分布のばらつきを記述するのに、各データと平均との差（偏差という）の2乗和を使っている。正規分布に従う場合には平均値が中心位置であるので、中心位置からの離れ具合を2乗値で求めている。標準偏差の2乗したものを分散とよぶ。

〔重要点〕分布の特徴を数値で記述するための統計量の説明をしているが、正規分布に従う分布の特徴を述べる際には平均と標準偏差をペアで用いる。すなわち、分布の中心位置とばらつき具合を一緒に記述する。報告の仕方として、平均 ± 標準偏差で表すことも多い。

〔注意点1〕上の求め方では（$n-1$）で割っているが、n で割り算する公式も広く使われている。統計学的には（$n-1$）で割る公式が正式な定義であり n で割る公式

$$\overline{S}=\sqrt{\frac{1}{n}\{(x_1-\overline{x})^2+\cdots+(x_n-\overline{x})^2\}}$$

よりも統計的に性質の良いものとされている。実践ではデータの個数がある程度多い場合には、どちらの公式を用いても標準偏差に大きな差はない。

〔注意点2〕標準偏差の単位は元のデータの単位を用いることができる。たとえば、$\{x_1,$

…，x_n｝が年齢のデータであれば、「標準偏差は 4.8 歳である」と記述できる。

〔例〕BMI のデータ {20.0，21.3，21.5，22.2，23.0} の標準偏差は、エクセル関数＝STDEV.S（データ配列）を使って 1.1 を得る。一方の BMI データ {20.0，21.3，21.5，22.2，30.5} の標準偏差は 4.2 となり、1.1 に比べてかなり大きくなっていることがわかる。ちなみに、外れ値 30.5 を除外して求めると 0.9 となり 1.1 に近づく。標準偏差を求める際には、必ずエクセルか統計解析ソフトウェアを用いる。

2）四分位範囲（Interquartile range：IQR）

右に尾を引く分布など正規分布に従わない分布のとき、データのばらつき具合を記述する際に用いる。通常、中央値（分布の中心位置）とともに用いられる。

〔求め方〕IQR ＝ 75 パーセント点－25 パーセント点

75 パーセント点は第 3 四分位数、25 パーセント点は第 1 四分位数ともいう。このテキストでは 75％点、25％点と記述する。それぞれの求め方は中央値の時と同様に以下のとおりである。

n 個の量的データ ｛x_1，…，x_n｝があるとき

① n 個のデータを昇順に並べ替える。

②小さい方から順位を 1，2，…，n とつける。

③小さい方から 25％目にあたる順位の元データが 25％点、小さい方から 75％目にあたる順位の元データが 75％点である。

〔重要点〕25％点、中央値（50％点）、75％点の 3 つをまとめて四分位点（四分位数）とよぶ。

〔注意点 1〕量的データが整数値だけであっても、四分位点は必ずしも整数とはならない。たとえば、10 個のデータが {1，2，3，4，5，6，7，8，9，10} のとき、中央値は順位が 5.5 番目の元データとなるので 5.5 となる。

〔注意点 2〕25％点、中央値、75％点以外にも、たとえば、10％点、90％点などを使うこともあり、一般に p％点とよぶ。

〔例〕エクセル関数で p％点を求めるには、＝ PERCENTILE.EXC（データ配列，p/100）を使う。p/100 は、たとえば、90％点を求める際には 0.9 を入力することを意味する。

3）範囲（Range）

データのばらつき具合を示す一番シンプルな方法である。

〔求め方〕範囲＝最大値－最小値

〔例〕p.362「代表値」の「平均」で示した例 1、例 2 の 5 例の BMI について範囲をもとめると、

例 1 の BMI データの範囲は 23.0－20.0 ＝ 3.0

例 2 の BMI データの範囲は 30.5－20.0 ＝ 10.5

となる。エクセル関数は最大値（= MAX（データ配列））から最小値（= MIN（データ配列））を引くことにより求めることができる。上のふたつの例でわかるように、範囲は外れ値があるときに大きな値となる。逆にいえば、外れ値の検出に用いることもある。

4）変動係数（Coefficient of variation）

正規分布に従う分布に対して用いる散布度である。

〔求め方〕変動係数 C.V. は標準偏差$\overline{S}$を平均$\overline{x}$で割った値である。

$$C.V.=\frac{\overline{S}}{\overline{x}}\times 100(\%)$$

変動係数は 100 を掛けてパーセント（%）で表すことが多い。

〔解説〕変動係数は、単位の異なる 2 つ以上の変数について、データのばらつきを比較するときに用いられる。

たとえば、当院で新たに導入した生化学検査機器の精度を調べるために、同じ検体で γ-GTP と血清アルブミンを 8 回計測した結果が表 7 である。同じ検体を使っているので、本来は同一の数値が得られないといけないはずであるが、試薬の少量の違いやセンサーの誤差などにより表のように違った計測値が得られた。

表 7.　γ-GTP と血清アルブミン（Alb）の計測結果

	1	2	3	4	5	6	7	8	平均値	標準偏差	変動係数
γ-GTP	36	38	35	36	35	37	36	34	35.9	1.2	0.033
Alb	4.2	4.5	4.4	4.1	4.1	4.4	4.5	4.8	4.4	0.2	0.045

データのばらつきを調べたいので標準偏差を求めた。γ-GTP と血清アルブミン値の標準偏差はそれぞれ 1.2IU/L、0.23g/dL であるが、測定単位が異なるのでこれを比較しても精度の優劣はつけることができない。

変動係数は標準偏差を平均値で割るので、測定単位に依存しない値でパーセント（%）で表すことが多い。したがって、γ-GTP は 3.3%、血清アルブミンは 4.5% となり、血清アルブミンの方がデータの大きさに対する相対的なばらつきが大きいことがわかる。

3. 質的データの記述統計量

サンプルデータ A の K 病院における肝硬変データ 100 例の性別は、男性（コード 1）、女性（コード 2）のふたつのデータを持つ名義変数である。この場合、それぞれの人数と割合を記述する。すなわち、男性患者、女性患者の人数と割合はそれぞれ、87 人（87%）、13

人（13%）で性比は約 6.7 対 1 である、と記述される。

また、肝硬変の成因はデータがアルコール性、Hb（B 型肝炎ウイルス）陽性、その他の名義変数であり、肝シンチグラフィの正常、肥大、右葉萎縮、両葉萎縮は、この順番で肝硬変の重症度が増すので順序変数である。これらの場合も記述統計量として人数と割合を用いる。

4. 生存時間変数（Time-to-Event 変数）

サンプルデータ A：肝硬変データ 100 例において、生存時間（日数）が含まれている。この生存時間データは一見すると量的データに見えるが、データの性質上まったく別の種別のデータであることに注意する。

この生存時間データは肝硬変患者の診断日から死亡までの日数を計測したものである。肝硬変患者は定期的に来院してその治療にあたる。このように担当医師が患者を継続的に診察し、または、経過を観察することを統計学では「フォローアップ」とよぶ。患者のフォローアップを行っていくうちに、患者は死亡、もしくは、生存中のまま観察打ち切りのいずれかの結果となる。後者のことを打ち切り例とよび、その生存時間データのことを打ち切りデータとよぶ。観察打ち切りが発生する理由はいくつかあるが、典型的な打ち切りは以下の場合である。

1）フォローアップ中に急な転居もしくは何らかの理由で、患者が来院しなくなった。それまでの受診歴で最終の受診日が最終生存確認日であり観察打ち切り日となる。
2）フォローアップ中に報告書をまとめる、もしくは、学会報告するために、決められた年月日を締切日としてその時点が最終生存確認日で観察打ち切り日となる。

したがって、生存時間データの場合、性質の異なる 2 種類が混在する。すなわち、死亡までの生存時間とその時点までは生存していた時間の 2 種類である。死亡と生存中という最終観察時点での状態を「転帰（生存予後）」とよぶ。

図 10 は肝硬変患者のフォローアップの様子を表したものである。縦軸は患者 10 例、横軸は左から右へ時間が経過する。今回の肝硬変データでは 1975 年から 1985 年までの期間でフォローアップがなされている。各症例の横棒の長さが生存日数であり、棒の左端が診断時点、右端が転帰で●が死亡、○が打ち切りを表している。

このように、同じ生存日数でも 2 種類の転帰の異なるデータが混在する点が、体重、血圧などの連続データとは異なる点である。生存時間データの統計解析では、生存時間変数と転帰の変数をペアで用いる必要がある。量的データや質的データの記述統計量は使えない点に注意する。

生存時間データの記述統計量は生存率が使われる。たとえば、がん登録データでは 5 年

生存率などが、今回の肝硬変データでは 10 年生存率が用いられる。生存率を求める手法は Kaplan-Meier（カプラン－マイヤー）法を使うが、これは市販の統計解析ソフトウェアに入っている。

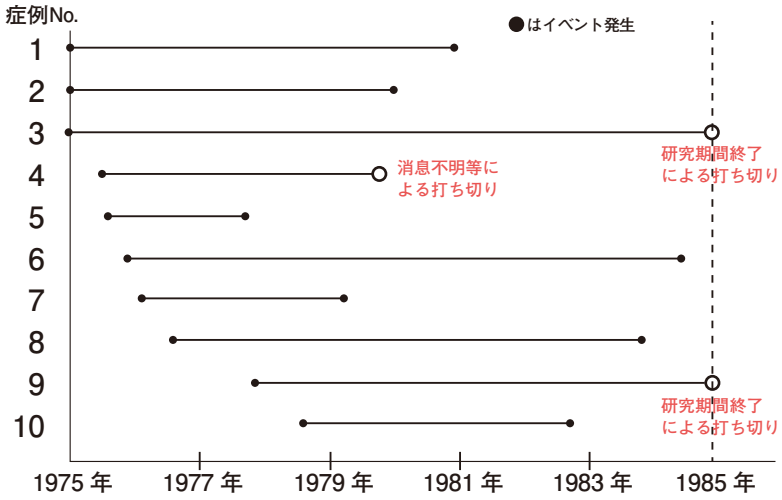


図 10.　肝硬変患者のフォローアップ

6-4　確率と確率分布

ここでは、確率とは何かを、いくつかの例とともに解説する。また、これまでの分布は量的変数や質的変数のデータとその度数で記述されたが、これと同じくデータとその確率を対応付けた確率分布を説明する。その中で最も大切な分布として、正規分布の諸性質と使い方を説明する。

【相対度数による確率】

確率は、われわれの生活の中でよく使われる。たとえば、降水確率、サイコロを投げて奇数が出る確率などである。医療においても確率が用いられており、消化器外科手術で手術部位感染（SSI）を起こす確率は 13.6%、日本の外科手術を受けた胃癌患者の 5 年生存率は 71.1%、脳卒中患者が 10 年以内に再発を起こす確率 50% などである。

一言で確率といっても専門的にはいくつかの分類があるが、このテキストでは確率を以下のように定義する。

〔定義〕事象 E の起こる確率とは、起こりうるすべての事象（全事象）の中で E が起こる相対度数をいう。

ここで事象（Event）とは、このテキストでは「物事が起こること」をいう。たとえば、消化器外科手術を受けた患者でSSIを起こす確率の場合、全事象は「消化器手術をうけること」をさし、事象Eは「SSIを起こすこと」をさす。

SSIを起こす確率は、全国サーベランスのデータなどから求めることができる。全事象である食道・胃・胆嚢・小腸・結腸・直腸・虫垂の外科手術を受けた患者は4,000例、このうちSSIを起こした患者（事象Eを起こした患者）は544例いたので、SSIを起こす確率は相対度数13.6%である。

〔記号〕事象Eを起こす確率（Probability）を$P(E)$と表す。したがって、SSIの確率の場合、
P（SSIを起こす）$= 0.136$

〔性質〕確率は次の性質をもつ。

①どのような事象Eに対しても、確率$P(E)$は0と1の間の値をとる。
すなわち、$0 \leqq P(E) \leqq 1$。
上では、確率を%に直して表記したが、正しくは0から1までの値である。

②全事象をSと書くと、$P(S) = 1$。

③同時には起こりえない事象E_1とE_2があるとき、E_1、E_2のいずれかが起こることを$E_1 \cup E_2$と表記する。この確率は、$P(E_1 \cup E_2) = P(E_1) + P(E_2)$。

〔注意〕上で述べた相対度数（相対頻度）による確率は、限られた症例数で得られる事象Eの起こる確率$P(E)$が、その症例数を非常に大きくしたときの相対度数と等しいものとして定義する。

【確率分布】

〔定義〕変数の値xに対して確率$P(x)$を対応させたものを確率分布という。

〔性質〕上述「相対度数による確率」の性質②により、xの取りうる範囲（全事象）で$P(x)$をすべて足し合わせると1となる。

〔例1：サイコロの目の確率分布〕

サイコロの目を横軸にとり、縦軸にはその目の確率を示した。それぞれの目が出る確率は1/6であり、確率を示す棒の高さはすべて等しい。このような等確率の分布を一様分布とよぶ。1から6までの目に対応する確率をすべて足し合わせると1となり、上の性質が確かめられる。

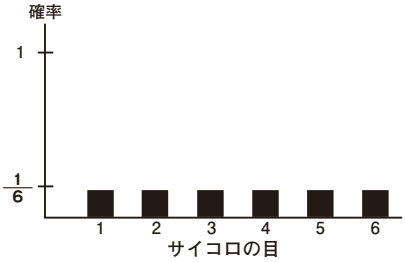


図 11. サイコロの目の確率分布

〔例 2：コイン投げの確率分布〕

コインを 3 回投げて表の出る回数と確率を対応付けると以下の確率分布となる。

3 回のコイン投げでは 1 回目、2 回目、3 回目で（表、表、表）、（表、表、裏）、…、で 8 通りの場合がある。これが全事象である。このうち、表の回数が 0 の事象は 1 回しか起こらないので確率は 1/8 である。表の回数が 1 回のみの場合は（表、裏、裏）、（裏、表、裏）、（裏、裏、表）の 3 回であり、確率は 3/8 となる。このようにして確率分布は以下の図となる。

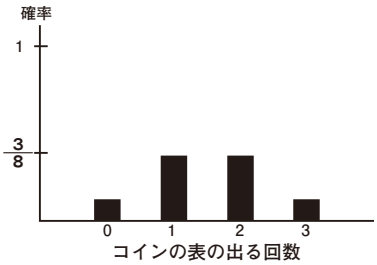


図 12. コイン投げの確率分布

〔例 3：サンプルデータ* C：脳卒中看護データ 65 例における発症時年齢、性別、部位の確率分布〕

脳卒中の発症部位は名義変数であるが、その確率分布を棒グラフで表した。この病院に入院する脳卒中患者において、発症部位別にみると大脳が他の部位に比べて発症確率（相対度数）が高いことがわかる。

*サンプルデータは A，B，C の 3 データあり、それぞれ通信教育 Web サイトからダウンロードできる。

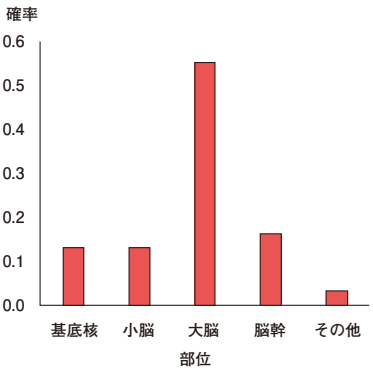


図 13. 脳卒中の発症部位の確率分布

〔例 4：肝硬変患者の成因と発症時年齢の相対度数分布〕

度数と相対度数を表にした確率分布の例である。

表 8. 肝硬変患者の成因と発症時年齢の相対度数分布表

成因	度数	相対度数
アルコール性	51	27.6%
Hb 陽性	34	18.4%
原因不明	65	35.1%
その他	35	18.9%
合計	185	100.0%

診断時年齢（歳）	度数	相対度数
～ 20	1	0.5%
21 ～ 25	0	0.0%
26 ～ 30	9	4.9%
31 ～ 35	11	5.9%
36 ～ 40	16	8.6%
41 ～ 45	23	12.4%
46 ～ 50	30	16.2%
51 ～ 55	37	20.0%
56 ～ 60	22	11.9%
61 ～ 65	18	9.7%
66 ～ 70	9	4.9%
71 ～ 75	7	3.8%
76 ～ 80	2	1.1%
81 ～	0	0.0%
合計	185	100.0%

6-5 正規分布

確率分布の中で特に重要な分布として正規分布がある。正規分布（Normal distribution）の横軸は連続変数である。正規分布は統計解析の実践において極めて重要な分布である。記述統計の集計方法、推測統計の推定方法や検定方法の選定において、連続変数が正規分布になっているか否かは重要な役割を果たす。

〔定義〕ひとつ山で左右対称のつり鐘型の度数分布のことを「正規分布」とよぶ。正規分布は平均と標準偏差を使って理論的に数式で表すことができるが、ここではその数式は割愛する。平均$\overline{x}$、標準偏差$\overline{S}$の正規分布を $N(\overline{x}, \overline{S}^2)$ と表す。

【実データにおける正規分布と非正規分布の例】

〔例 1：正規分布－身長の分布－〕

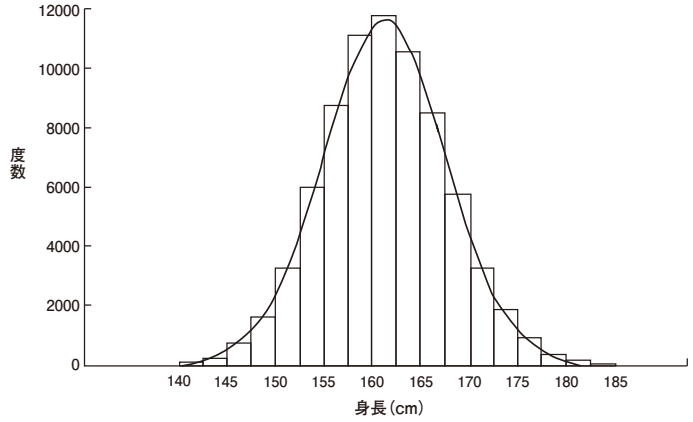


図 14. 身長のヒストグラムと正規分布曲線

図 14 は身長のヒストグラムと、その平均と標準偏差による正規分布の理論式に基づく正規分布曲線とを重ね合わせた図である。収集されたデータ数が数万人のとき、身長のヒストグラムはきれいな釣り鐘状の形となる。このようなとき、「身長は正規分布に従う」もしくは「身長は正規分布で近似できる」という。

〔例 2：正規分布－肝疾患患者 185 例の診断時年齢の分布－〕

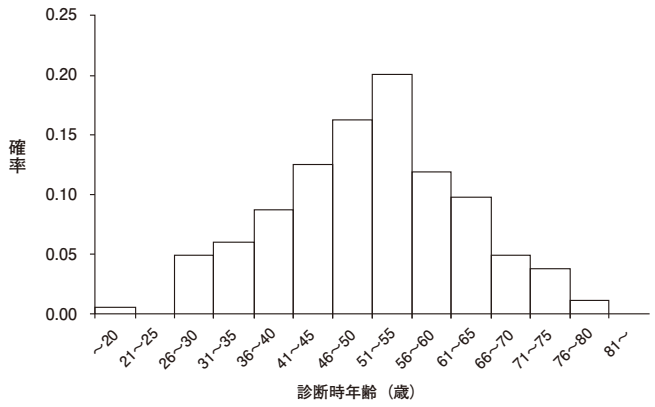


図 15. 肝疾患患者の診断時年齢のヒストグラム

K 病院の肝疾患患者 185 例の診断時年齢の分布（ヒストグラム）である。データの個数が小さいため完全に左右対称になっているとは言い難いが、実践的な統計解析では、このヒストグラムは正規分布に従っている（正規分布の形に近い）と解釈する。なお、20 歳以下の階級に外れ値（ひとかたまりの分布からは極端に外れた値）がある。

〔例 3：非正規分布－胃癌登録事業の登録症例数別の施設数－〕

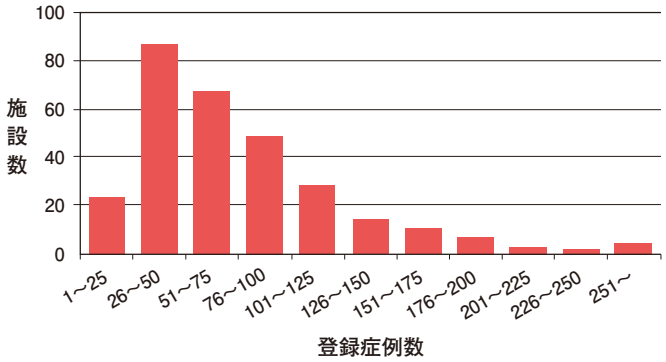


図 16. 胃癌登録事業の登録症例数別の施設数

左右非対称のヒストグラムであり、この場合を「右に尾をひく分布」という。登録症例数は正規分布に従っているとはいえない。

〔例 4：非正規分布－肝硬変患者の血清アルブミン値の分布－〕

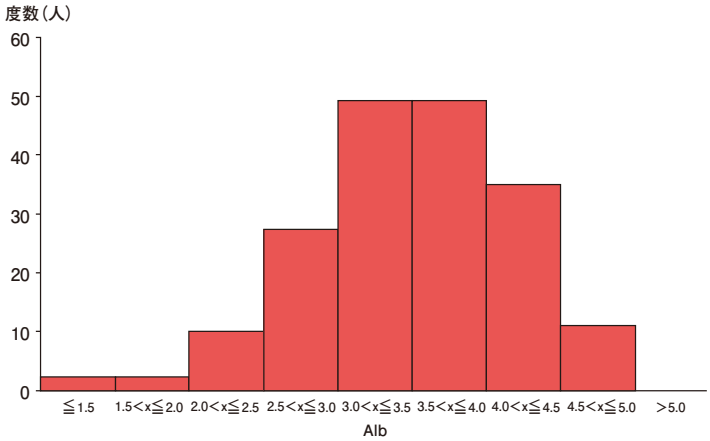


図 17. 肝硬変患者の血清アルブミン値（Alb）の分布

横軸に血清アルブミン値、縦軸に人数をとったヒストグラムである。ヒストグラムは左右非対称であり「左に尾を引くヒストグラム」の形となっている。

6-5-1 正規分布の性質

正規分布には実践的な統計解析に有用ないくつかの性質がある。

〔性質 1〕正規分布の形は、平均と標準偏差の 2 つの値によりひとつに決まる。

〔例 1〕体重を想定した平均 60kg、標準偏差 7kg の正規分布 $N(60,\ 7^2)$

平均 60kg、標準偏差 7kg の正規分布は以下の図となる。標準偏差 7 の 2 乗 7^2 は分散である。

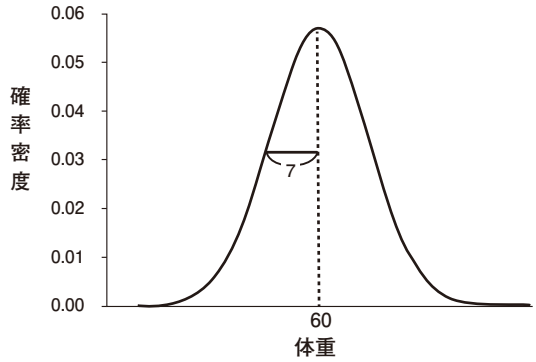


図 18. 体重を想定した平均 60kg、標準偏差 7kg の正規分布 $N(60,\ 7^2)$

ここで注意したいことは、縦軸は確率密度とよばれ確率ではない値である。体重 55kg の縦線の長さ 0.044 であるが、これは体重 55kg の人の確率（相対度数）ではない。

55.0kg，55.1kg，…，55.9kg などの 55kg 以上 56kg 未満という体重の幅に対しては正規分布曲線とこの幅で囲まれる面積が確率となる。

理論的な正規分布では、横軸のある幅と正規分布曲線で囲まれた面積が確率となる。

〔性質 2〕正規分布の中心位置（代表値）は平均、ばらつき（散布度）は標準偏差である。

〔例 2〕中心位置が異なるときの 3 つの正規分布

左から順に $N(48,\ 7^2)$，$N(60,\ 7^2)$，$N(70,\ 7^2)$ の正規分布が下図である。ばらつきを変えずに平均だけを変えているので、山の高さは変わらず同じ形の正規分布が横にシフトするだけである。

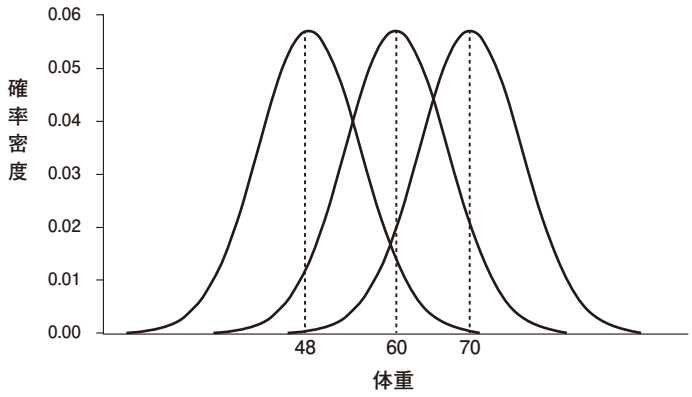


図 19. 中心位置が異なるときの 3 つの正規分布

〔例 3〕ばらつきが異なる 3 つの正規分布

山の高い方から順に、$N\,(60,\ 3^2)$，$N\,(60,\ 7^2)$，$N\,(60,\ 12^2)$ の正規分布である。標準偏差が 3 の正規分布では、山の最大値（確率密度）は 0.133 となる。このように標準偏差によって山のとがり具合が著しく変化することがわかる。

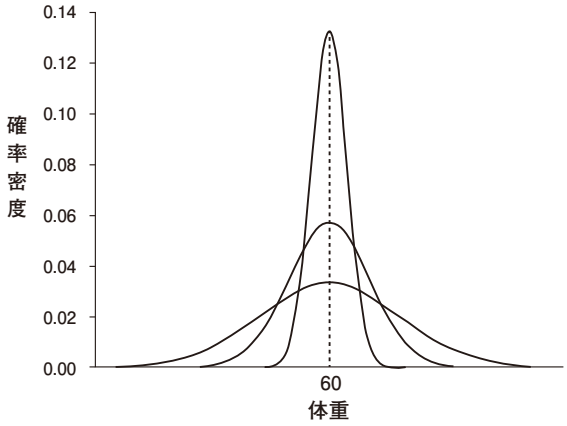


図 20.　ばらつきが異なる 3 つの正規分布

〔性質 3〕横軸と曲線で囲まれる面積は常に 1 である。

正規分布は確率分布のひとつであり、横軸と曲線とで囲まれる部分の面積が確率を表す。横軸と曲線全体で囲まれる部分の面積は 1 となる。

〔性質 4〕正規分布では、平均、中央値、最頻値が同じ値にある。

正規分布の理論式から、山が一番高くなるときの横軸の値が平均となり、左右対称であるので、平均より左側の面積は 0.5、右側の面積も 0.5 となる。このことから、中央値と最頻値も平均と同じ値となる。

6-5-2　標準正規分布　$N\,(0,\ 1^2)$

身長や年齢などの連続変数を横軸にとり、縦軸が確率密度である正規分布をこれまで説明した。ここでは、実務でもしばしば使われる特別な正規分布、標準正規分布について説明する。

〔定義〕平均 0、標準偏差 1 である正規分布 $N\,(0,\ 1^2)$ のことを「標準正規分布」とよぶ。

標準正規分布の横軸（連続変数）をこのテキストでは「Z」と表す。

標準正規分布の場合、Z 値に対する確率が本書巻末の付属資料「統計数値表」の付表 1（1）正規分布表（上側確率）、またはエクセル関数で簡単に求めることができる。付表 1 に標準正規分布における Z より大きい値をとるときの相対度数、すなわち面積（確率）が示されている。付表 1（1）では「Z がある正の値以上」の確率を示している。ある値以上の確率を上側確率とよぶ。

【付表 1（1）の見方】

1）付表 1(1) の一番左の列（ピンク色）は、Z 値の 1 の位と小数点第 1 位（○.○）を表す。

2）一番上にピンク色の行があるが、これは Z の小数点第 2 位の数を表す。

3）たとえば、$Z = 1.35$ のとき、左の列の 1.3 と上の行の 0.05 がクロスする値が標準正規分布における $Z = 1.35$ の上側確率である。すなわち、$P(Z \geqq 1.35) = 0.0885$、Z が 1.35 以上となる確率は 0.0885 である。

このようにして付表 1(1) を使う。いくつかの例をあげる。

〔例 1〕 $P(Z \geqq 1)$

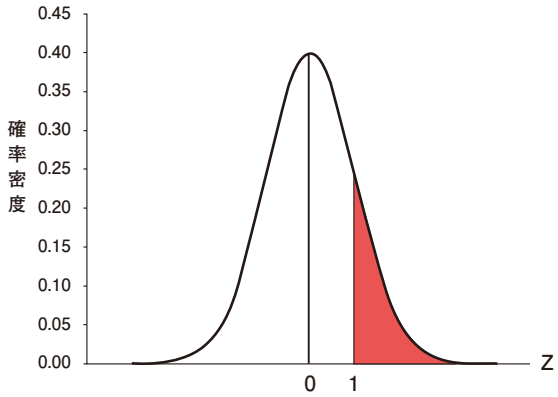


図 21. $P(Z \geqq 1)$

Z が 1 以上となる確率 $P(Z \geqq 1)$ は、1 を 1.00 と考え 1.0 と 0.00 がクロスする確率を読む。すなわち、$P(Z \geqq 1) = 0.1587$。なお、$Z = 1$ の確率密度の縦線は面積を持たないので、$P(Z \geqq 1) = P(Z > 1)$ である。

エクセル関数では　＝ 1－NORM.S.DIST（1.0，TRUE）

で求めることができる。エクセルでは下側確率を求める関数を用いる。

〔例 2〕　$P\,(Z \leqq -1)$

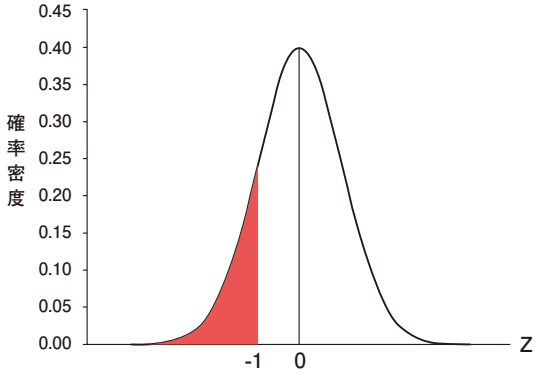


図 22.　$P\,(Z \leqq -1)$

付表 1(1)では正の数の Z 値しかでていない。正規分布が左右対称である性質を使って、$P\,(Z \leqq -1) = P\,(Z \geqq 1) = 0.1587$ を得る。

エクセル関数では　= NORM.S.DIST（−1，TRUE）= 0.1587

〔例 3〕　$P\,(-1 \leqq Z \leqq 1)$

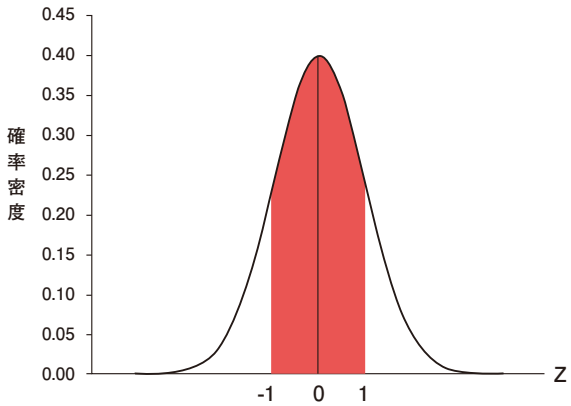


図 23.　$P\,(-1 \leqq Z \leqq 1)$

標準正規分布の中心部分の確率を求めてみる。Z 軸と曲線とで囲まれる全体の面積は 1、そこから $Z \leqq -1$ の確率と $Z \geqq 1$ の確率を引き算する。したがって、

$1 - 2 \times P\,(Z \geqq 1) = 0.6826$

エクセル関数では　= 1−（2 *（1−NORM.S.DIST（1，TRUE））= 0.682689

〔例 4〕　$P\,(Z \geqq -1)$

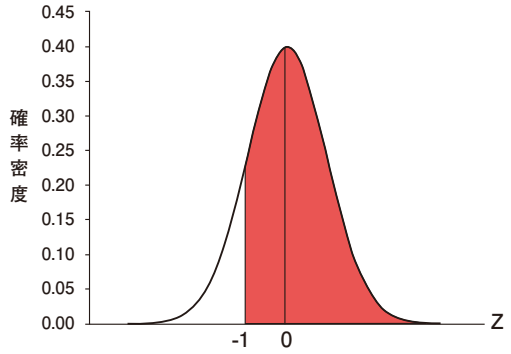


図 24.　$P\,(Z \geqq -1)$

正規分布の左右対称の性質、全体の面積（確率）が 1 であることを使って、

$P\,(Z \geqq -1) = 1 - P\,(Z \leqq -1) = 1 - P\,(Z \geqq 1) = 0.8413$

〔例 5〕　$P\,(Z \geqq 1.645)$

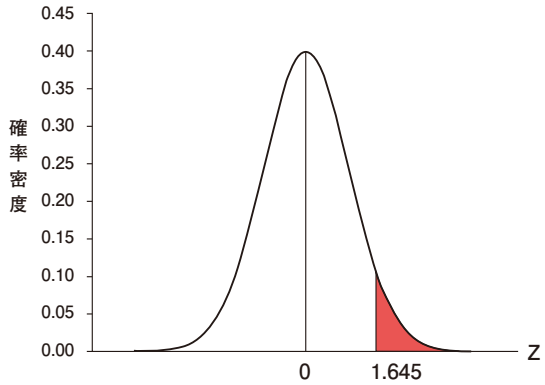


図 25.　$P\,(Z \geqq 1.645)$

表からは $P\,(Z \geqq 1.64)$ と $P\,(Z \geqq 1.65)$ しか求まらない。1.645 は両者の真ん中をとり 0.05 を得る。

エクセル関数では　= 1 − NORM.S.DIST（1.645，TRUE）= 0.04998

この片側確率 0.05 に対応する Z 値 1.645 は統計学的検定でよく出てくる数値である。

〔例 6〕 $P(Z \geqq 1.96) + P(Z \leqq -1.96)$

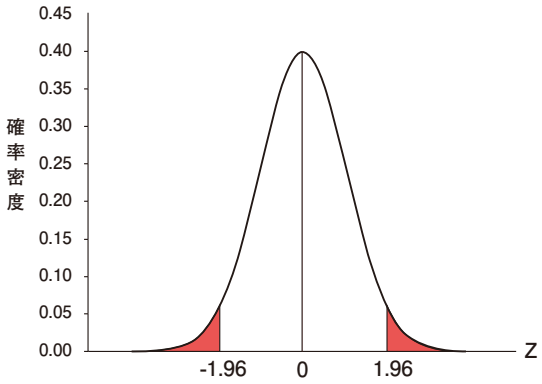


図 26. $P(Z \geqq 1.96) + P(Z \leqq 1.96)$

$P(Z \geqq 1.96) + P(Z \leqq -1.96) = 2 \times P(Z \geqq 1.96) = 0.05$

エクセル関数では　$= 2 * \text{NORM.S.DIST}(-1.96, \text{TRUE}) = 0.049996$

6-5-3 標準正規分布を使った一般の正規分布の確率計算

標準正規分布において、Z のいろいろな区間での確率を求めることを学んだ。

しかしながら、身長や血圧などの正規分布は平均 0、標準偏差 1 ではない。以下では、身長が平均 160cm、標準偏差 7cm の正規分布に従うとき、153cm から 167cm の確率を求める方法を述べる。

付表 1 の数表から一般の正規分布 $N(\mu, \sigma^2)$ における面積（確率）も求めることができる。横軸の変数 X とし、X は正規分布 $N(\mu, \sigma^2)$ にしたがうとする。このとき、正規分布 $N(\mu, \sigma^2)$ における X がある値 a より大きい部分の面積は以下の手順で計算できる。

【身長 X の正規分布 $N(160, 7^2)$ において $P(153 \leqq X \leqq 167)$ を求める方法】

〔手順 1〕 $N(160, 7^2)$ に従う身長 X を標準正規分布の Z に対応づけるための変換を行う。

この変換のことを標準化もしくは基準化とよぶ。具体的には、$X-160$ を 7 で割った値を標準正規分布 Z とすることである。

一般の正規分布 $N(a, b^2)$ の変数 X を $N(0, 1^2)$ の変数 Z に変換する標準化の公式

$$Z = \frac{X-a}{b}$$

$N(160, 7^2)$ の 153cm と 167cm を Z に変換すると

$$Z=\frac{153-160}{7}=-1.0\,,\quad Z=\frac{167-160}{7}=1.0$$

〔手順 2〕変数 X が正規分布 $N(\mu,\ \sigma^2)$ にしたがうとき、標準化公式で得られる Z は標準正規分布 $N(0,\ 1^2)$ にしたがう、という統計的な理論がある。したがって、$N(160,\ 7^2)$ における確率計算 $P(153 \leqq X \leqq 167)$ は $N(0,\ 1^2)$ における $P(-1 \leqq Z \leqq 1)$ と等しくなる。

〔手順 3〕p.377「6-5-2 標準正規分布 $N(0,\ 1^2)$」の例 3 により、$P(-1 \leqq Z \leqq 1) = 0.6826$

このようにして、身長や血圧が正規分布に従うとき、身長 160cm 未満の確率や、血圧 140mmHg 以上の確率などを求めることができる。

〔例 1〕健康的な成人男子のヘモグロビン濃度（g/L）が、およそ正規分布 $N(15.3,\ 1.8^2)$ にしたがうとする。このとき、ヘモグロビン濃度が 17g/L より高い確率はいくらか。

ヘモグロビン濃度を X と表すと、求める確率は $X > 17$ となる部分の正規分布曲線の面積で表される。この確率は、X を Z の値に変換すると、

$$Z=\frac{17-15.3}{1.8}=0.94$$

より、付表 1 の Z の値 0.94 の上側確率を読んで、0.17361 と求まる。

〔例 2〕健康的な成人男子における、ある検査値の分布が正規分布 $N(a,\ b^2)$ にしたがうとき、検査値が $a-2b$ 以上、$a+2b$ 未満に入るのは健康的な成人の何％くらいといえるか。

求める割合は検査値 X が、$a-2b < X < a+2b$ となる範囲に入る確率として表される。X の値 $a-2b$、$a+2b$ を標準正規分布に変換すると、それぞれ

$$Z=\frac{(a-2b)-a}{b}=-2$$

$$Z=\frac{(a+2b)-a}{b}=2$$

より、標準正規分布の -2 から 2 までの確率であり、先にみたように 0.95445 と求まる。

一般に X が $N(a,\ b^2)$ にしたがうとき、X の値と、Z の値の間には

$X = a+b,\ a+2b,\ a+3b \ \rightarrow\ Z = 1,\ 2,\ 3$

$X = a-b,\ a-2b,\ a-3b \ \rightarrow\ Z = -1,\ -2,\ -3$

となる対応関係がある。X を変換した Z の値は、「X が a から、標準偏差の何倍、正の方向に、または負の方向に離れているか」を表している。

6-6 2つの変数の相関を調べる

医学・医療の変数が X と Y の2つあるときを考える。X が増加すると Y も増加する関係が認められる場合があるし、その逆で、X が増加すると Y は減少する場合もある。また、X の増加にともない Y もある直線に近いところを増加する場合もあるし、Y がかなり幅をもって増加する場合もある。このような関係を視覚的に、あるいは、定量的に示すための統計的手法が相関分析である。

6-6-1 相関とは？

〔定義〕相関とは2つの変数の関係をいう。相関分析とは、2つの変数の関係を図や表で表したり、関係の強さや向きを数値で表すための分析手法をいう。

【例1：正の相関がある散布図（相関図）】

2つの変数が量的変数の場合、変数 X と Y の関係を視覚的にとらえるために散布図が用いられる。以下は、サンプルデータC：脳卒中看護データ65例の体重と身長の関係を散布図で表したものである。体重を X（横軸）にとり、身長を Y（縦軸）にとって、座標平面上に、一人ひとりの観測値 x と y の値の組を打点したものである。

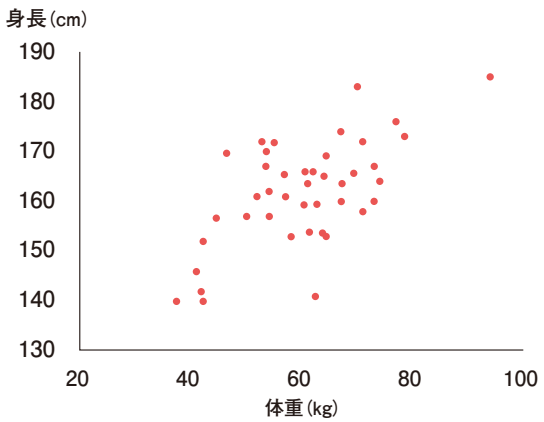


図27. 脳卒中データの体重と身長の関係

散布図から脳卒中患者の体重と身長の関係について次の2点を読み取ることができる。

1）体重の増加とともに身長も高くなる傾向がある。

2）その傾向はほぼ直線的である。

このように、X の増加にともない Y も増加するとき、「X と Y は正の相関がある」という。

【例 2：負の相関がある散布図（相関図）】

肝硬変データにおいて、X に診断時年齢、Y に血清アルブミン値をとり作成した散布図である。

※ 実際の散布図では集団から外れた点があり、それらの点はここでは削除した。

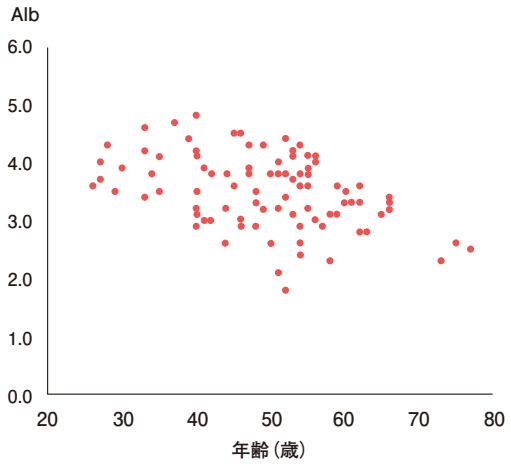


図 28. 肝硬変データの診断時年齢と血清アルブミン値（Alb）の関係

この散布図より以下の 2 点が読み取れる。

1）診断時年齢の増加とともに、血清アルブミン値は低下する傾向がある。

2）その傾向はほぼ直線的である。

このように、X の増加にともない Y が減少するとき、「X と Y は負の相関がある」という。

【例 3：順序変数のクロス集計表】

上述の例 1、例 2 は 2 つの変数がともに量的変数であり、散布図により相関の特徴を把握した。2 つの変数が順序変数のとき、以下のようなクロス集計表を作成することにより、相関関係を特徴づけることができる。

表 9. 泌尿器科のある疾患患者の尿失禁と ADL の重症度

		ADL の重症度					
		G0	G1	G2	G3	G4	計
尿失禁の重症度	G0	0	2	1	0	0	3
	G1	1	5	3	0	0	9
	G2	0	1	2	2	0	5
	G3	0	0	1	12	0	13
	G4	0	0	0	8	62	70
	計	1	8	7	22	62	100

表 9 では、泌尿器科のある疾患患者のクロス集計表の結果である。行に尿失禁の重症度、列に ADL の重症度をとっている。それぞれの変数のカテゴリーでは G0 から G4 に向かうと重症度は増す。このクロス集計表からわかることは、ADL の重症度が増すと尿失禁も重症化する傾向が顕著にみられる。

6-6-2 相関係数

2 つの変数の関係を図表に表して、その相関関係を視覚的にとらえる方法を上で述べた。ここでは、その相関関係の強さと向きを数値で表す相関係数について説明する。

〔定義〕相関係数とは 2 つの変数の相関の強さと向きを −1 以上、+1 以下で表した統計量のことである。このテキストでは、データから計算される相関係数を r で表すことにする。散布図の作成、相関係数の算出およびそれらの医学的な解釈をつける統計解析のことを「相関分析」とよぶ。

〔性質〕相関係数 r は次の性質をもつ。

1）$-1 \leqq r \leqq +1$

2）$r > 0$ のときは「正の相関」、$r < 0$ のときは「負の相関」

3）r の絶対値が 1 に近いほど変数 X と Y の相関は強い。0 に近いほど変数 X と Y の相関が弱い。

4）$r = 0$ のとき、このテキストでは無相関とよぶ。

【例：正の相関と負の相関】

以下の図で示すように、点が右上がりの分布の場合は正の相関、右下がりの分布の場合は負の相関とみなす。

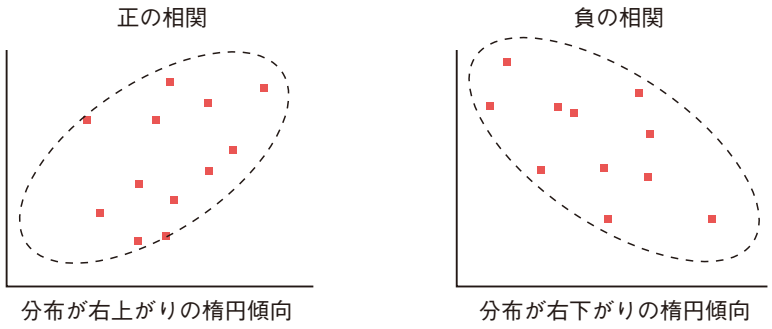


図 29.　正の相関と負の相関

【例：すべての点が直線上に並んでいる散布図】

散布図ですべての点が右上がりの直線に並んでいる場合、相関係数は＋1 である。散布図ですべての点が右下がりの直線に並んでいる場合、相関係数は－1 である。

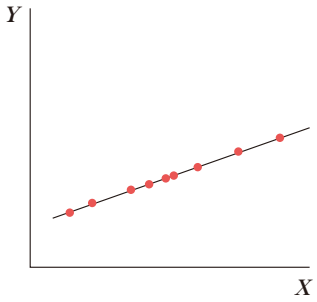


図 30.　すべての点が直線上に並んでいる散布図（相関係数＝ 1）

X の増加にともない Y がひとつの値に定まり直線的に増加。相関係数＝ 1

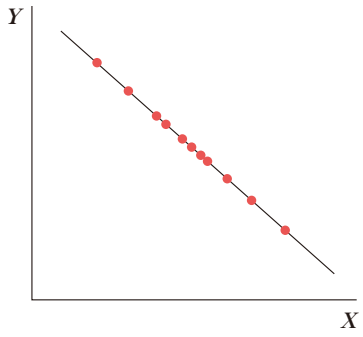


図 31.　すべての点が直線上に並んでいる散布図（相関係数＝－1）

X の増加にともない Y はひとつに値が定まり直線的に減少。相関係数＝－1

【例：無相関】

以下の散布図は無相関の例である。X が増加しても Y は増減がなく、ほぼ一定の範囲を変動している。このような X と Y の関係を無相関とよぶ。以下の散布図のように無相関であれば相関係数は 0 である。

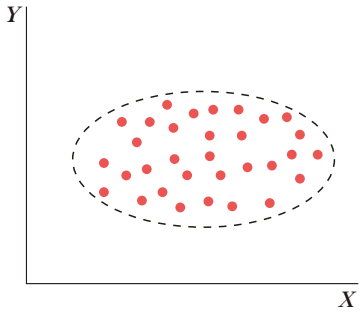


図 32.　無相関

6-6-3　強い相関、弱い相関

相関係数は −1 以上 +1 以下の値しかとらないが、値の大きさにより「強い相関」、「中程度の相関」、「弱い相関」と表現することがある。これらの表現はあくまでも医学論文や学会報告での呼び方であり、統計学的な定義ではないことに注意する。以下では、相関係数を r と表す。

1）r が 0.7 以上、もしくは、−0.7 以下のとき、強い相関とよぶ。

2）r が 0.4 以上 0.7 未満、もしくは、−0.4 以下 −0.7 より大きいとき、中程度の相関とよぶ。

3）r が −0.4 より大きく 0.4 より小さいとき、弱い相関とよぶ。ただし、$r = 0$ の時は無相関であるとよぶことが多い。

※p.387「6-6-5 相関係数の解釈における注意点」の注意点 2 を参照。

6-6-4　相関分析の例

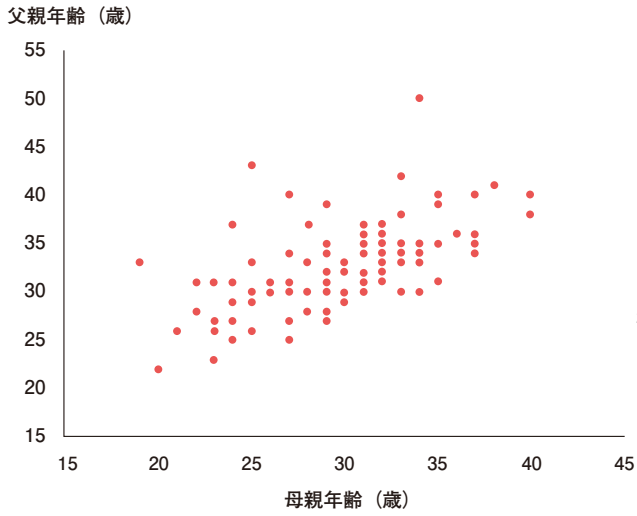
サンプルデータ A：肝硬変データ 100 例、サンプルデータ B：乳幼児健診データ 100 例に基づき、相関分析の実例をいくつか示す。

【例 1：乳幼児健診データの出産時の母親年齢と父親年齢】

横軸に母親の年齢、縦軸に父親の年齢をとり散布図を作ると図 33 となる。

散布図より、母親年齢が高いほど父親年齢も高くなる傾向が見て取れる。相関係数を求

めると 0.598 であり、正の中程度の相関があることがわかる。



相関係数：0.598

【散布図の作成方法】
①横軸のデータと縦軸のデータを範囲選択する。
②挿入タブの散布図をクリックする。

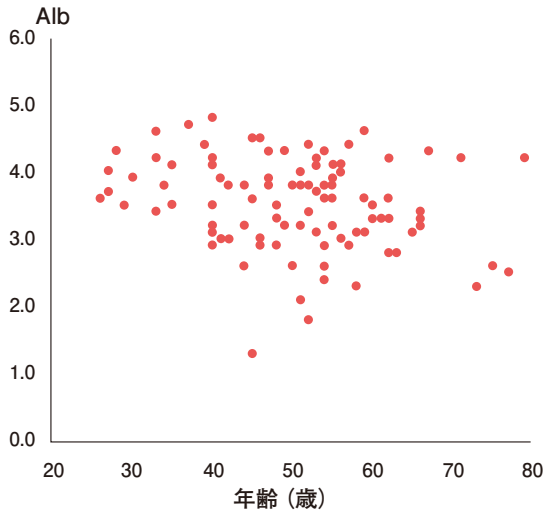
【相関係数の求め方】
1. CORREL（配列 1, 配列 2） コリレーション関数
配列 1，2 にデータのセル範囲を入力する。

2. データ分析の「相関」を使う場合
①データタブのデータ分析から、「相関」を
クリックする。
②入力範囲にデータのセル範囲を入力する。

図 33. 乳幼児健診データ（母親年齢・父親年齢）の散布図と相関係数

【例 2：肝硬変データの診断時年齢と血清アルブミン値の相関係数】

肝硬変データを用いて、横軸に診断時年齢、縦軸に血清アルブミン値（Alb）をとった散布図 34 が示されている。p.382「6-6-1 相関とは？」の例 2 とは違い、今回はすべてのデータを用いている。この相関係数を求めると－0.264 であり、負の弱い相関が認められた。



相関係数：-0.264

図 34. 肝硬変データにおける年齢と血清アルブミン値（Alb）の散布図と相関係数

6-6-5 相関係数の解釈における注意点

〔注意点 1〕相関係数は 2 つの変量の間の因果関係を表しているわけではない。たとえば、無作為に抽出した社会人 1,000 人の「年収と体重」の散布図を描くと、体重の多い人ほど年収も高くなる関係があるかもしれない。この場合、相関係数 r は正の値となるが、体重が重い人ほど年収も多いとの結論づけはできない。体重が原因で年収が結果なのではなく、第 3 の因子、年齢が高いほど体重も年収も高くなる傾向にあるため、あたかも年収と体重に相関があるようにみえるだけである。この場合の「年齢」を交絡要因という。統計解析では交絡要因の存在を意識しながら解析を進める必要がある。

〔注意点 2〕図 32 で示した無相関の散布図では相関係数が 0 となる。無相関ならば相関係数は 0 である。一方、相関係数が 0 だからといって図 32 のような無相関の図とならない場合がある。このことから、散布図と相関係数を同時に求めることが大切である。

6-7 一方の変数からもう一方の変数の値を予測する（回帰分析）

散布図において X と Y が直線的な関係にあるとき、その関係を最もよく表す直線をあてはめ、X の値が与えられたときに、Y の値を予測することを考える。

【定義】

1）回帰直線

変数 X からもう一方の変数 Y を予測するための直線式

$$Y = a + b \times X$$

を回帰直線という。この直線において、予測の元となる観測値 x に付いている b のことを傾きといい、定数 a のことを切片とよぶ。理論的な説明は省略するが、X, Y それぞれの平均の点（$\overline{X}$, $\overline{Y}$）を通る直線となる。

2）説明変数と予測変数

回帰直線では X を説明変数、Y を予測変数（目的変数）という。または X を独立変数、Y を従属変数ともいう。

3）回帰分析

直線的な関係にある X と Y の観測値の n 個の組（$x1$, $Y1$），（$x2$, $Y2$），…，（xn, yn）から回帰直線を求め、その予測精度の評価などを分析することを回帰分析とよぶ。

6-7-1 回帰分析の例

【例 1：脳卒中看護データにおける体重と身長】

サンプルデータ C：脳卒中看護データ 65 例に基づき、体重から身長を予測する回帰分析を行う。

散布図は図 35 となる。回帰分析はエクセルのデータ分析を使って以下の手順で行うことができる。

65 例の体重と身長のデータに対して、エクセルのフィルター機能などを使って欠損値のある症例を除外する。回帰分析はペアでデータがそろっている 42 例で行う。42 例のデータセットに対して

データタブ⇒分析リボン⇒データ分析⇒分析ツール⇒回帰分析

を選択する。回帰分析のウィンドウで以下の操作を行う。

入力元　入力 Y 範囲：身長 42 例分の数値をドラッグして選択
　　入力 X 範囲：体重 42 例分の数値をドラッグして選択
　　ラベル　　：身長、体重のラベルを選択に含めている場合にはチェックを入れる
出力オプション　　：出力先を指定する。何も指定しなければ新シートに出力される
残差や正規確率　　：必要に応じて指定する

また、エクセルの分析ツールを使った回帰分析の結果を表 10 に示す。

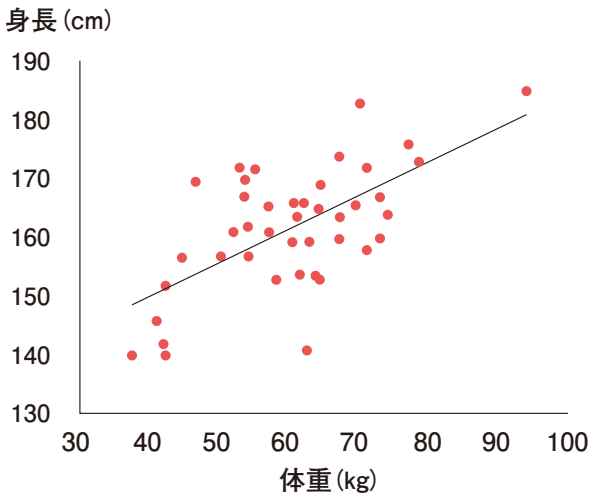


図 35.　脳卒中データにおける体重と身長の散布図

表 10. 脳卒中データにおける体重と身長の回帰分析の結果

概要

回帰統計	
重相関 R	0.627175
重決定 R2	0.393349
補正 R2	0.378182
標準誤差	8.360475
観測数	42

分散分析表

	自由度	変動	分散	観測された分散比	有意 F
回帰	1	1812.843	1812.843	25.9357158	8.81E-06
残差	40	2795.902	69.89755		
合計	41	4608.745			

	係数	標準誤差	*t*	*P* 値	下限 95%	上限 95%	下限 95.0%	上限 95.0%
切片	127.7706	6.801692	18.78513	1.88942E-21	114.0239	141.5174	114.0239	141.5174
*X*値1	0.564531	0.110851	5.092712	8.8073E-06	0.340493	0.788569	0.340493	0.788569

上の結果で 8.81E-06 は $8.81 \times 10^{-6} = 0.00000881$ の意味である。エクセルの出力結果から、次のことがわかる。

1）この散布図にあてはまる回帰直線は以下である。

身長＝ 0.56× 体重＋127.77

2）回帰直線の傾きから、体重が 1kg 増すと身長は平均で 0.57cm 増す。

【例 2：肝硬変データにおける年齢と血清アルブミン値】

横軸が診断時年齢、縦軸が血清アルブミン値（Alb）の散布図と相関分析の結果は、p.386「6-6-4 相関分析の例」の例 2 で示した。ここでは、その回帰分析を行ってみる。

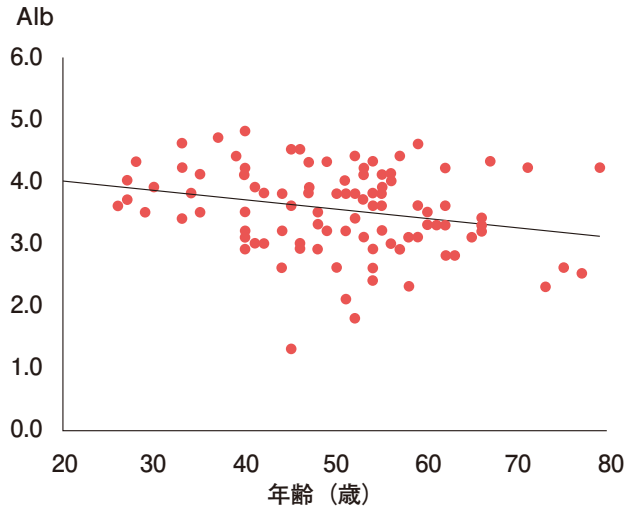


図 36. 肝硬変データにおける年齢と血清アルブミン値（Alb）の散布図

表 11．肝硬変データにおける年齢と血清アルブミン値（Alb）の回帰分析

概要

回帰統計	
重相関 R	0.264
重決定 R2	0.070
補正 R2	0.060
標準誤差	0.665
観測数	100

分散分析表

	自由度	変動	分散	観測された分散比	有意 F
回帰	1	3.262	3.262	7.369	0.008
残差	98	43.383	0.443		
合計	99	46.645			

	係数	標準誤差	*t*	*P* 値	下限 95%	上限 95%	下限 95.0%	上限 95.0%
切片	4.3160	0.2874	15.0166	0.0000	3.7457	4.8864	3.7457	4.8864
*X*値1	−0.0152	0.0056	−2.7146	0.0078	−0.0264	−0.0041	−0.0264	−0.0041

この回帰分析からわかることは以下のとおりである。

1）回帰直線は

$$\text{Alb} = -0.15 \times \text{年齢} + 4.32$$

2）回帰直線の傾きから、年齢が 1 歳増加すると血清アルブミン値は平均で 0.15 減少する。

6-8 推測統計の基礎

これまでは、病院あるいは地域で収集したデータを要約して、それらの特性を明らかにする記述統計の手法を説明してきた。医療統計学にはこれ以外に推測統計という、別の大きな柱がある。

推測統計を比喩的に表現すると、「木を見て森を知ること」となるだろう。「森を知る」とは森の樹木の種類と割合、樹齢、樹木の健康状態を知ることを意味する。森のすべての木々を計測するのは本数が多く、時間と費用がかかり過ぎるので不可能である。そこで、一部の木々を抽出してその小集団での樹の種類、樹齢などを計測する。これが「木を見て」の意味である。

小集団で得られた結果を、森全体でもこの小集団と同じ特性であろうという仮定のもとで「木を見て得られた特性を森全体に適用して、森の特性を知る」のである。

医療の場合も同様で、たとえば、胃癌患者は世界中に数百万人いるだろうが、その数百万人の年齢、性別、生活習慣、癌の進行度、病理学的分類を収集するのは不可能である。とてつもない大きな集団のデータをすべて収集するのは費用も時間もかかるので、小集団

を抽出して、そこから得られる統計的特性を元の大集団に適用すること、これが推測統計である。

6-8-1 推測統計で使われる用語の定義

推測統計では、これまでに出てこなかった統計用語がいくつか出てくるので、その定義を行うことにする。

〔定義 1〕母集団

知りたい対象全体、研究や調査の対象となる個と対の全体集合をいう。上述での森の木すべて、世界中の胃癌患者全員が母集団にあたる。母集団は大きな集団であるので、母集団を調査して、森の樹木の割合や胃癌患者の診断時年齢を求めることは、一般には不可能である。すなわち、母集団での樹木の割合や診断時年齢の平均は未知である。

〔定義 2〕無作為抽出

母集団の各個体の抽出される確率がすべて等しくなる抽出方法のことをいう。無作為抽出では、通常、乱数（規則性がまったくない数の並び）を用いる。すなわち、母集団が世界中の胃癌患者 1,000 万例でその中から 100 例を無作為抽出したい場合には、各患者に 1 番から 1,000 万番まで付与しておき、8 ケタの乱数を 100 個作り、その番号に当たった患者を抽出する。

〔定義 3〕標本

母集団から抽出した一部の症例（個体、要素）の集合、もしくは、そのデータの集合をいう。母集団の特性を持ち合わせた小さな集合のことである。母集団から無作為抽出により標本を抽出する、という表現を使う。気をつけるべきことは、標本が母集団のミニチュアであり、母集団の特性を引き継いでいることである。逆にいうと、母集団特性を継承するような標本抽出を心がける、ということである。

〔定義 4〕標本の大きさ（標本サイズ）

母集団から抽出された標本の個体数のことをいう。

〔定義 5〕母集団の特性

母集団が持つさまざまな性質、特徴であり、統計量で表すことができるものをいう。たとえば、母集団が世界中の胃癌患者であるとき、診断時年齢の分布の中心位置やばらつきの程度、早期胃癌の割合などのことである。

〔定義 6〕全数調査と標本調査

全数調査とは、母集団のすべての個体からデータを収集する調査のことである。日本人全員を母集団とすると、5 年に 1 回行われる国勢調査は全数調査のひとつの例である。全数調査はセンサスともよぶ。

6-8-2 身近で使われている推測統計

推測統計は一般社会でよく使われている。いくつかの例を示す。

〔例 1〕**世論調査**

日本人の世論を把握するために、有権者全員（母集団）の中から標本を可能な限り無作為抽出して調査が行われる。母集団は住民基本台帳だったり、電話帳に記載のある全員であり、この中から年齢や性で層別したのちに、各層で無作為に標本抽出が行われる。

〔例 2〕**感染症の発症数の推測**

感染症サーベイランス事業では、全数報告対象疾患（エボラ出血熱、結核、コレラなど）と定点広告対象疾患（小児の特定の疾患やインフルエンザ、眼科疾患など）がある。推測統計的には、前者はそれぞれの疾患に関する全数調査にあたり、後者は標本調査にあたる。インフルエンザの場合、2018 年 2 月現在、全国約 5,000 ヶ所の内科・小児科医療機関が標本となっている。

6-8-3 母集団と標本の関係

図 37 では母集団と標本の関係が示されている。標本は母集団から抽出されたミニチュアであり、統計解析可能なデータ量である。標本を使って行われる統計解析には推定と検定があり、これらの解析結果は母集団に対して適用される。

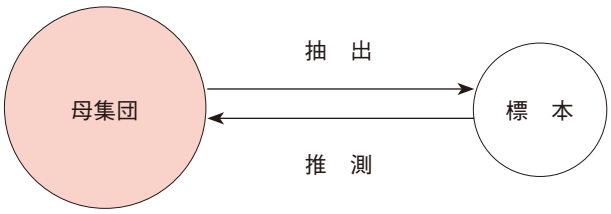


図 37. 母集団と標本

〔注意〕上の図 37 の説明では「標本は母集団から抽出されたミニチュア…」と記述した。本来であれば、「標本は母集団から無作為抽出されたミニチュア…」と書くべきであるが、無作為抽出は実際には難しい抽出法といえる。日本の胃癌患者（母集団）の特性を見出したいときに、A 病院の胃癌患者 200 例（標本）を使って解析した、という学会発表や論文記述がある。

この場合、A 病院の胃癌患者は日本全体からの無作為抽出になっておらず、母集団のミニチュアとは言い難いであろう。このような発表では、「無作為抽出はされていないけ

れども、母集団のミニチュアとほぼ同じ特性を持つだろう」という仮定がなされていると考えるべきである。この仮定が正しいかどうかは全数調査ができないので検証不可能である。

6-8-4 母集団と標本の記述に関する約束ごと

日本の胃癌患者全員を母集団として、そこから無作為に 1,000 例を抽出し標本を作る。エクセルなどによる統計解析は標本を使って行い、その解析結果は母集団に適用される。

母集団における平均、分散、標準偏差などの統計量のことを

母平均、母分散、母標準偏差

などという。また、標本から求められる平均、分散、標準偏差などを

標本平均、標本分散、標本標準偏差

などとよび、統計学を学ぶ際には母集団のそれらとは区別する。

一般に、母集団の統計量はギリシャ文学で表される。たとえば、母平均は μ（ミュー）、母分散は σ^2（シグマ 2 乗）、母標準偏差は σ で表す。母集団分布の特徴を記述する μ や σ などを総称して母数（parameter）とよぶ。

〔母集団の統計量（母数）〕		〔標本の統計量〕	
母平均	μ	標本平均	$\overline{X}$
母分散	σ^2	標本分散	V
母標準偏差	$\sigma = \sqrt{\sigma^2}$	標本標準偏差	$S = \sqrt{V}$

6-9 推定

標本に基づく推測には、推定と検定があることを述べた。検定に関しては後で詳しく説明することにして、ここでは推定の定義と大まかな説明を行う。

6-9-1 推定の定義

〔定義 1〕推定とは、母集団の分布の特性を標本に基づき数値化することをいう。
〔定義 2〕推定値とは、推定において特性を表す数値（統計量）のことをいう。

推定とは、μ や σ の未知の母数を標本から算出される推定値$\overline{X}$や$\overline{S}$で言い当てることである。

〔説明〕母集団が日本全体の胃癌患者であるとして、その発症時年齢の分布（ヒストグラム）を考える。その特性とは、記述統計でも述べたように、分布の中心位置とばらつきを平均と標準偏差などで表すことである。また、日本全体での胃癌患者において、男女の患者数の分布の特性を知りたいとき、標本からその割合を求めることが行われる。上述の平均と標準偏差や男女の割合は推定値である。

つまり、記述統計も推測統計も用いられる統計量、代表値や散布度は同じである。2つの違いは、記述統計では「収集された自院の胃癌患者の発症時年齢の分布は平均 67.7 歳、標準偏差 11.6 歳である」と結論づけるのに対して、推測統計では、「標本から推定される日本の胃癌患者の発症時年齢の平均と標準偏差は、それぞれ 67.7 歳、11.6 歳である」と結論付けられる。推測統計では、67.7 歳や 11.6 歳は母集団分布の推定値となっている。

6-9-2 推定の利用事例

母集団の特性を標本から推定して新たな知見を見出す例を以下に示す。

母集団を 2017 年に出産した日本の妊婦全員とする。この母集団から無作為に近い形で標本を抽出し、サンプルデータ B：乳幼児健診データ 100 例での変数に関するデータを収集したとする。これにより、母集団の特性として次の推定値が得られる。

1）妊娠中の妊婦の喫煙割合
2）妊娠中の異常有りの割合
3）出産時年齢の平均と標準偏差
4）出生体重*の四分位数

*出生体重（BirthWeight：BW）、以降 BW と表す。

これらの推定値により、たとえば、次のような知見を得ることができる。

a）妊娠中の妊婦の喫煙や妊娠中の異常の有りの割合は、2007 年、1997 年に比べて 2017 年では高くなっている。

b）出産時年齢平均は 2007 年、1997 年に比べて○歳上昇している。
c）F 市の BW 平均は、全国平均に比べて○グラム少ない。
d）妊娠中の喫煙割合は欧米に比べて△%低い。
e）BW の四分位範囲は過去 20 年間で小さくなってきた。

このように、経時的な変化や、他の地域に比べて出産時年齢や BW の代表値や散布度が大きい（小さい）ことが定量的に把握できる。

6-9-3 点推定と区間推定

母集団の特性を数値で表す推定には、点推定と区間推定の 2 種類ある。それぞれを詳しく説明する。

【点推定】

〔定義〕点推定とは、母集団のひとつの特性をひとつの数値（推定値）で表す推定のことをいう。

〔例：点推定〕

1）母集団での出産時年齢の分布の中心位置を平均で表す。
2）胃癌患者母集団（母集団が胃癌患者全員）における腫瘍長径の分布の中心位置を中央値で表す。
3）母集団の出産時年齢分布のばらつきを標準偏差で表す。
4）胃癌患者母集団における腫瘍長径分布のばらつきを四分位範囲で表す。
5）妊婦母集団の妊娠中の喫煙ありの母割合を標本割合で表す。
6）乳癌患者母集団の 5 年生存率を標本による 5 年生存率で表す。
7）日本の新生児の BW の母平均 μ を標本平均 $\overline{X}$ で推定する。
8）N 県の成人男性の HbA1c の母中央値を標本中央値で推定する。
9）関西地方の喫煙する人の母割合を標本による喫煙率で推定する。
10）日本の胃癌患者の母 5 年生存率を、学会のがん登録データによる標本 5 年生存率で推定する。

点推定における推定値は、「6-3-2 データの要約に用いる記述統計量」で述べた p.362「代表値」や p.364「散布度」の計算方法と同じである。

【区間推定】

〔定義〕区間推定とは、母集団の母数が、ある確率で区間（a，b）にあるような a，b を標本から求めることをいう。母平均、母中央値、母割合、母生存率などを区間で言い当てる推定法である。

上の定義における用語の説明を次に示す。

1）母集団の母数

母平均 μ、母分散 σ^2、母比率（母割合）などである。

2）ある確率

0.95（95%）や 0.99（99%）が用いられ、これを信頼係数という。

3）区間（a，b）

a を上側信頼限界、b を下側信頼限界という。

〔例：区間推定〕

日本の BW の母平均の 95%信頼区間は（2,750g，3,160g）と推定される。

6-9-4 母集団の分布が正規分布に従う時の母平均の区間推定の手順

母集団のある変数が正規分布 $N(\mu, \sigma^2)$ に従うと仮定する。たとえば、胃癌患者母集団の診断時年齢の分布が母平均 μ、分散 σ^2 であるとする。母集団は膨大なのでデータ収集は不可能であり、μ や σ^2 は未知の値であることは前述のとおりである。したがって、正規分布に従うかどうかはあくまでも仮定である。標本のデータ $\{x_1, \cdots, x_n\}$ があるとして、母平均 μ の 95%信頼区間の推定手順を述べる。

【母平均 μ の 95%信頼区間の推定手順】

1）標本平均 $\overline{X}$ と標本分散 S^2 を求める

$$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}x_i \qquad S^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{X})^2$$

2）母平均 μ の 95%信頼区間は次の公式で計算できる

$$\left[\overline{X}-t_{0.025}(n-1)\times\frac{S}{\sqrt{n}},\ \overline{X}+t_{0.025}(n-1)\times\frac{S}{\sqrt{n}}\right]$$

ただし、S は標本標準偏差、n は標本数、$t_{0.025}(n-1)$ は自由度（$n-1$）の t 分布の上側確率 0.025 に対する t 値である。この t 分布における上側確率 0.025 と t 値との関係は、p.375「6-5-2 標準正規分布 $N(0, 1^2)$」で述べた標準正規分布の上側確率と Z 値の関係と類

似している。正規分布を自由度（$n-1$）の t 分布に置き換え、上側確率は両者で同じ意味であり、Z 値を t 値に置き換えればよい。t 分布の上側確率から t 値を求めるための数表は付表 2 である。

【母集団の分布が正規分布に従う時の母平均の95%信頼区間をエクセルで求める方法】

上で説明したように、正規分布の母平均の 95%信頼区間の下限値と上限値を求める公式は以下である。

下限値：標本平均 $-t$（0.025，$n-1$）×（標本標準偏差 / √標本数）
上限値：標本平均 $+t$（0.025，$n-1$）×（標本標準偏差 / √標本数）

この各値を次のエクセル関数で求めればよい。

1）標本平均　　：　＝ AVERAGE（データ）
2）標本標準偏差：　n でわる公式に対して　＝ STDEVP（データ）
　　$n-1$ でわる公式に対して　＝ STDEV（データ）
3）標本数　　：　＝ COUNTA（データ）
4）t（0.025，$n-1$）×（標本標準偏差 / √標本数）：
　　＝ CONFIDENCE.T（0.05，標準偏差，標本数）
5）下限値と上限値を 1 と 4 から計算する。

6-9-5　95%信頼区間（CI）の意味

95%の信頼区間の解釈について考えてみる。

母平均 μ の母集団から大きさ n の無作為標本を抽出して標本平均を求め、これを $\overline{X}_1$ と表す。次にまた大きさ n の標本を抽出して標本平均を求め、これを $\overline{X}_2$ と表す。この操作を何回も繰り返し、n 個の標本平均 $\overline{X}_1$，$\overline{X}_2$，…，$\overline{X}_n$ を求める（図 38）。図 38 ではこれらの標本平均は、標本の右横の黒丸で示した。

ここで μ は未知ではあるが固定された値で、そのまわりをそれぞれで得られた平均値 $\overline{X}_i$，（$i = 1, 2, \cdots, n$）が変動する。各標本における母平均の信頼区間を標本の右横の線分で描かれている。前述の 95%信頼区間の公式により、標本ごとで線分の長さは異なるが、ひとつの 95%信頼区間では標本平均を中心にして左右同じ長さである。

多くの信頼区間では μ を含むが、中には μ を含まない信頼区間もある。95%信頼区間とは、仮に母集団から標本を 100 回無作為に抽出したとき、5 回程度は μ を含まない信頼区間が生じうるような確からしさで作成された区間である。

95％信頼区間より区間幅が短くなると、その分だけ μ を含まない確率が高くなる。すなわち、同じ 100 個の標本を用いても μ を含む回数は 95 回を下回ることになる。この区間幅が短いときというのが、たとえば 90％信頼区間である。

信頼係数を 95％から 90％に下げると、信頼区間の公式で $t_{0.025}(n-1)$ を $t_{0.05}(n-1)$ に変えることになるが、同じ自由度（$n-1$）では上側確率が小さいほうが t 値は小さくなる。すなわち、$t_{0.05}(n-1) < t_{0.025}(n-1)$ となる。したがって、90％信頼区間の方が 95％信頼区間に比べて区間の幅が狭くなり、はずれの起こる可能性は 100 回につき 10 回程度に上がる。信頼係数が低いほど、信頼区間の幅は狭く、はずれの可能性は高くなる。逆に信頼係数を上げると、それだけ信頼区間の幅が広くなり、はずれが起こりにくくなる。同じ標本を用いて 99％信頼区間を作ると、95％信頼区間の長さに比べて長い区間が作られる。

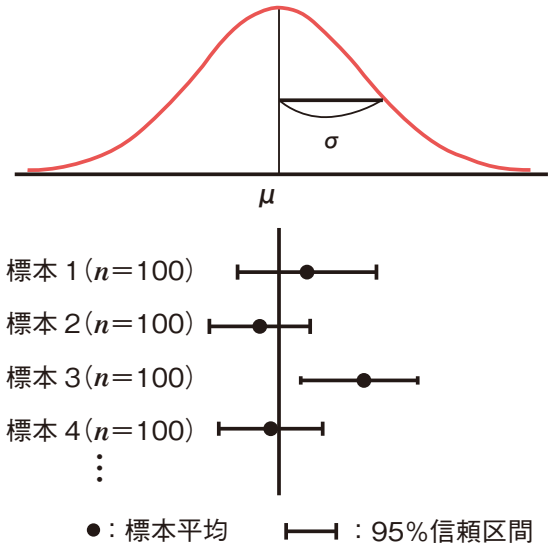


図 38.　95％の信頼区間の解釈

6-9-6　実践的な推定方法のまとめ

推定についてまとめると以下のようになる。

1）推定には点推定と区間推定がある。母集団の分布が正規分布に従うと仮定した場合、母平均を標本平均で推定するのが点推定であり、95％信頼区間などの区間で推定するのが区間推定である。

2）母集団の分布が正規分布と仮定できる場合、次のいずれかで推定される場合が多い。

（1）標本平均と標本標準偏差

（2）標本平均と 95％信頼区間

3）母集団の分布が量的変数に対する非正規分布と仮定できるとき、

「中央値と（25%点，75%点）」

を標本から求めて推定値とする。

4）母集団の分布が質的変数に対する分布のとき、

「それぞれのカテゴリーに対する標本割合とその 95%信頼区間」

を推定値として求める。母割合に対する 95%信頼区間の求め方は割愛する。

5）母集団の分布が Time-to-Event 変数に対する分布であるとき、

「生存率とその 95%信頼区間など」

を標本から求めて推定値とする。

6-9-7 推定の例

【例 1：乳幼児データに基づく日本の BW の推定】

日本で生まれる新生児全体を母集団と考えて、その BW の分布の特徴を推定する。サンプルデータ B：乳幼児健診データを標本と考える。標本の大きさは 100 例である。

標本が正規分布に従っていれば、母集団も正規分布に従っていると仮定できる。そこでまず、BW のヒストグラムと正規確率紙を作成する。正規確率紙はヒストグラムが正規分布に従うとき、グラフの直線上にプロットが並ぶように工夫された特別な方眼紙である。いずれもエクセルで作成可能であるが、正規確率紙のプロットはこのテキストのレベルを超えるので、インターネット等で調べていただきたい。

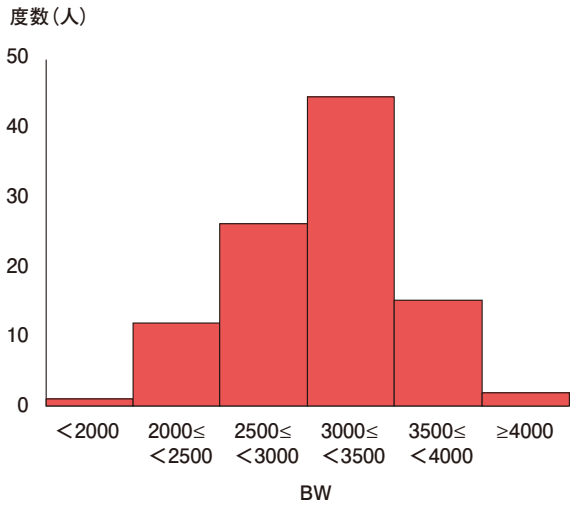


図 39. 出生体重（BW）のヒストグラム

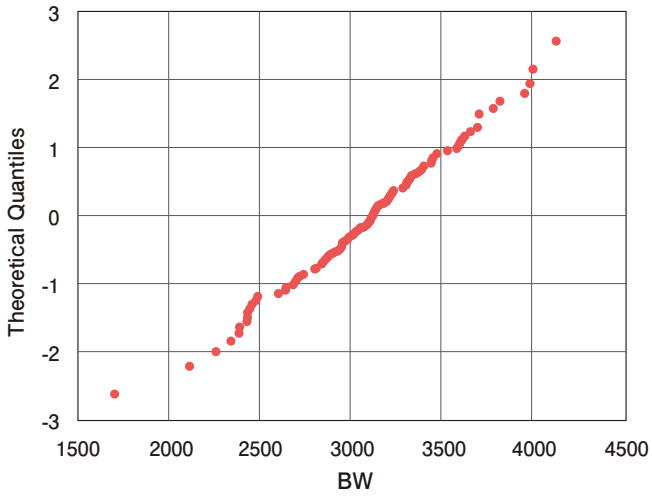


図 40. 出生体重（BW）の正規確率紙

図 39 から、ヒストグラムは症例数が少ないので棒の高さに凸凹があるが、ひとつ山で左右対称であることがわかる。したがって、量的変数である BW の分布の特徴づけは、標本平均と標本標準偏差で行うとする。すなわち、

「日本の新生児の BW の母平均と母標準偏差は 3,091 ± 443g であると推定される」

というまとめ方となる。

ちなみに、厚生労働省の「人口動態調査」によると、修正体重の平均は 2015 年が 3.00kg、2000 年が 3.03kg、1990 年が 3.12kg、1980 年が 3.19kg であり、年々、平均 BW は減少傾向にあることがわかる。この資料は厚生労働省の e-Stat というウェブサイトからダウンロードできる。その方法は以下のとおりである。

1）「e-Stat」https://www.e-stat.go.jp/ にアクセスする。

2）キーワードで探す。

　欄に「人口動態調査　年次別　出生時　平均体重」を入力し検索する。

3）検索結果を確認し、該当のデータの「DB」アイコンをクリックする。

4）「レイアウト設定」で表のレイアウトが変更できる。

5）「ダウンロード」をクリックし、表をダウンロードする。

【例 2：妊娠中の喫煙割合の推定】

例 1 と同様に、乳幼児データを標本として日本の妊婦の喫煙割合を推定する。喫煙の有無は質的変数なので、推定値として喫煙ありの割合を求める。その結果、日本の妊婦の喫煙する母割合は 14%であると推定される。

ちなみに、厚生労働省の「乳幼児身体発育調査　平成 22 年」によると、日本全体での妊

娠中の母親の喫煙割合は、平成 2 年が 5.6%、平成 12 年が 10.0%、平成 22 年が 5.0%である。

今回のサンプルデータ B：乳幼児健診データ 100 例の喫煙割合は、母集団の喫煙割合に比べて非常に高い値であるといえる。

【例 3：肝硬変患者の肝シンチグラフィの分布の特性】

日本の肝硬変患者全員を母集団として、サンプルデータ A：肝硬変データ 100 例を標本とする。肝シンチグラフィは、「1　正常」「2　肥大」「3　右葉萎縮」「4　両葉萎縮」の 4 つのカテゴリーがあり、コードが大きくなるに従って重症度も増す順序変数である。エクセルの COUNTIF 関数などを用いると、推定結果として表 12 を得る。

表 12. 肝硬変患者の肝シンチグラフィの分布の特性

肝シンチグラフィー	人数（人）	割合（%）
正常	6	6
肥大	46	46
右葉萎縮	28	28
両葉萎縮	9	9
不明	11	11
計	100	100

1）コード定義より、各コードが下記の場合、COUNTIF 関数を使って度数を集計する。

COUNTIF（肝シンチグラフィーのデータ範囲，コード）

※ 肝シンチグラフィー：1…正常　2…肥大　3…右葉萎縮　4…両葉萎縮　空白 …不明

正常　　：COUNTIF（肝シンチグラフィーのデータ範囲，1）

肥大　　：COUNTIF（肝シンチグラフィーのデータ範囲，2）

右葉萎縮：COUNTIF（肝シンチグラフィーのデータ範囲，3）

両葉萎縮：COUNTIF（肝シンチグラフィーのデータ範囲，4）

不明　　：COUNTIF（肝シンチグラフィーのデータ範囲，””）

2）各人数を合計で割り、割合を出す。（各人数 / 人数の合計）* 100

6-10 仮説検定

推測統計は推定と検定に大別されるが、後者の検定、正式には仮説検定について説明する。

〔定義〕仮説検定とは、臨床的な仮説を統計学的な 2 つの仮説に置き換えて、どちらが正しいかを標本に基づき判定することをいう。

たとえば、母集団を世界中の乳癌患者全員とする。乳癌患者の抗がん剤治療における嘔吐の改善に使われるAとBの治療薬のうち、どちらが嘔吐をより改善するのかを調べたいとする。臨床医は「新しく開発されたAの方がよく効くはずだ」という臨床的仮説を持っている。この仮説を2つの統計学的仮説、「A法とB法の改善割合は等しい」と「A法とB法の改善割合は異なる」に置き換える。そして、嘔吐のある乳癌患者200例を抽出して無作為に治療法A，Bに100例ずつ割り付けた。この200例が標本である。どちらの治療法が優れているのかを改善割合によって判定するが、その判定方法は後で説明する。

標本における嘔吐の改善割合はAが55%、Bが45%という推定結果を得た。この10%の割合の差を考えるとき、2群における治療法以外での違い、たとえば、年齢、乳癌の重症度、抗がん剤の感受性などがA群の改善割合に有利に働き、A群の改善割合が高めになったかもしれない。これらの個体や腫瘍などの既知または未知の要因による改善割合の違いを「偶然要因による差」とよぶ。

「A法がB法に比べて改善割合が高い」と科学的に結論づけるためには、2群の改善割合にみられた差は「偶然には生じえないほどに大きな（意味のある）差」なのか、それとも治療効果以外の偶然要因による差なのかを、統計的に検討する必要がある。このような問題に適用されるのが仮説検定である。

6-10-1 有意差検定

統計学的な仮説検定にはいくつかの種類があるが、このテキストで扱う仮説検定は有意差検定に限定する。

〔**定義**〕有意差検定とは、2群または3群以上の母平均、母中央値、母割合、母生存率などに、「意味のある差」があるか否かを調べる検定手法をいう。

意味のある差ではない差とは、前述の「偶然要因による差」のことである。有意差検定は群間の母平均などの差を積極的に立証しようとする検定であるのに対して、非劣性検定や同等性の検定などが医学・医療の研究や実務で使われる。

非劣性検定は、たとえば、「新薬の効果が従来の効果に比べて劣ることはない」ことを立証しようとするものである。また、同等性の検定は、文字通り群間の母平均などに大きな差がないことを示すための検定方法である。

次の節では、有意差検定の手順と検定で使われる統計学の用語を説明する。

6-10-2 仮説検定の手順

仮説検定には、母平均の有意差検定、母中央値の有意差検定、母割合の有意差検定、母生存率の有意差検定などがあるが、ここでは2群間の母平均の有意差検定を例にとり説明する。

【2群間の母平均の有意差検定の手順】

1）帰無仮説 H_0、対立仮説 H_1（または H_A）を設定する。
2）有意水準を定める。
3）得られたデータ（標本）から検定統計量を計算する。
4）検定統計量に基づき、H_0 が正しいと仮定したときに標本の得られる確率、P 値を求める。
5）P 値が有意水準より小さいとき、対立仮説を採択（帰無仮説を棄却）する。
6）P 値が有意水準より大きいとき、帰無仮説を保留する。すなわち、帰無仮説を棄却することができなかったとする。

【有意差検定の手順を示すための例】

上述の手順を乳幼児データにおける喫煙群と非喫煙群の BW の有意差検定を例にとりながら説明する。

1）「臨床的な仮説」として、「妊娠中に喫煙した群（喫煙群）は喫煙をしなかった群（非喫煙群）に比べて BW は低下する」を調べたい。
2）母集団は日本の妊婦全員で喫煙群と非喫煙群の2群に分けられる。
3）BW の母集団分布は2群とも正規分布に従う。
4）正規分布の代表値として平均を用いる。

6-10-3 帰無仮説と対立仮説

上述の「臨床的な仮説」を2つの統計的仮説に置き換える。すなわち下記である。

帰無仮説 H_0：喫煙群の BW の母平均 μ_1 ＝ 非喫煙群の BW の母平均 μ_2
対立仮説 H_1：喫煙群の BW の母平均 μ_1 ≠ 非喫煙群の BW の母平均 μ_2

H_0 は2群の母平均は等しいという仮説であり、H_1 は2群の母平均は異なるという仮説である。次のようにまとめられる。

1）仮説は常に 2 つ
2）H_0 は＝を用いて表される仮説
3）H_1 は≠を用いて表される仮説

【留意点】

1）仮説は常に 2 つ。
2）ふたつの仮説を帰無仮説と対立仮説とよぶ。
3）母集団に対する仮説であることを忘れずに。
4）群間比較の場合、群を明確に記述する。
5）検定の対象となる統計量を明記する。

〔参考〕医学の仮説検定における H_0、H_1 の例

①「A 群と B 群の治療では血糖値に差があるか」の 2 つの仮説

H_0：A 群の血糖値の母平均＝ B 群の血糖値の母平均

H_A：A 群の血糖値の母平均≠ B 群の血糖値の母平均

②「C 群と D 群の有効割合に差があるか」の 2 つの仮説

H_0：C 群の（母）有効割合＝ D 群の（母）有効割合

H_A：C 群の（母）有効割合≠ D 群の（母）有効割合

③「E 群と F 群の 5 年生存率に差があるか」の 2 つの仮説

H_0：E 群の（母）5 年生存率＝ F 群の（母）5 年生存率

H_A：E 群の（母）5 年生存率≠ F 群の（母）5 年生存率

6-10-4 有意水準

H_0、H_1 のどちらをとるのかを決める際の基準値である。有意水準は α（アルファ）で表し、臨床研究ではたいていの場合、$\alpha = 0.05$ と設定される。これは検定前に解析者が決めておく数値である。

6-10-5 検定統計量

【2 群間の母平均の有意差検定における検定統計量とその解釈】

標本に基づき、母集団の喫煙群と非喫煙群の分布の離れ具合を統計量で表す。この統計量のことを検定統計量とよぶ。2 群の母分布の離れ具合を表すための検定統計量は、以下の数式で表される。

$$T=\frac{X_1-X_2}{\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\left(\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}\right)}}$$

上の検定統計量は手計算やエクセルで数式を作って計算する必要はない。生データさえあればエクセル関数などで簡単に求めることができる。

この数式を提示したのは、母分布の離れ具合を数値で表す仕組みを説明するためのものである。式 T の分子は標本平均の差であり、これは分布の離れ具合を表すことは直感的に理解できるであろう。標本平均の差の絶対値（差が正の数のときはその値、差が負の数のときには符号を置き換えて正の数に直したもの）が大きければ大きいほど、分布は離れていることになる。

式 T の分母は一見すると複雑な式に見えるが、分母で使われている数値は標本分散（標本標準偏差の 2 乗した値）と標本の大きさのみである。詳しい説明は省略するが、この分母は「標本平均の差のばらつき量」を表している。BW の標本平均の差が 100g であっても、2 群の標準偏差が大きいときと小さいときでは、2 群の正規分布の重なりが大きかったりほとんど重ならなかったりする。つまり、離れ具合が違ってくる。ばらつきの程度による分布の重なりを考慮するために「標本平均の差のばらつき」で割り算している。

要約すると、式 T は 2 群の母分布の離れ具合を表す統計量である。式 T は 2 群間の母平均の有意差検定で用いられる検定統計量であるが、2 群間の母割合の有意差検定などでも同様に、標本割合の差をそのばらつきで割り算して検定統計量を算出している。

【H_0 の下での T 検定統計量の分布】

前節では、標本の大きさ n_1 と n_2 の BW データを抽出したときの 1 個の検定統計量の解釈を述べた。そこで次のことを考える。

H_0 が真実であると仮定する。統計学のテキストでは、このことを「H_0 の下で…」と表現する。H_0 の下で標本の大きさが同じ BW データを抽出して T を計算する。これを T_1 とする。同様に 2 回目のデータを抽出してその T を T_2 とする。このような操作を 1 万回繰り返して $\{T_1, \cdots, T_{10000}\}$ を作る。

この 10,000 個の T 値のヒストグラムを作ると次のようになる。正規分布に似ているが、

標本の大きさ n_1 と n_2 により分布の形が少し異なるひとつ山で左右対称の「t 分布」を得る（図 41）。

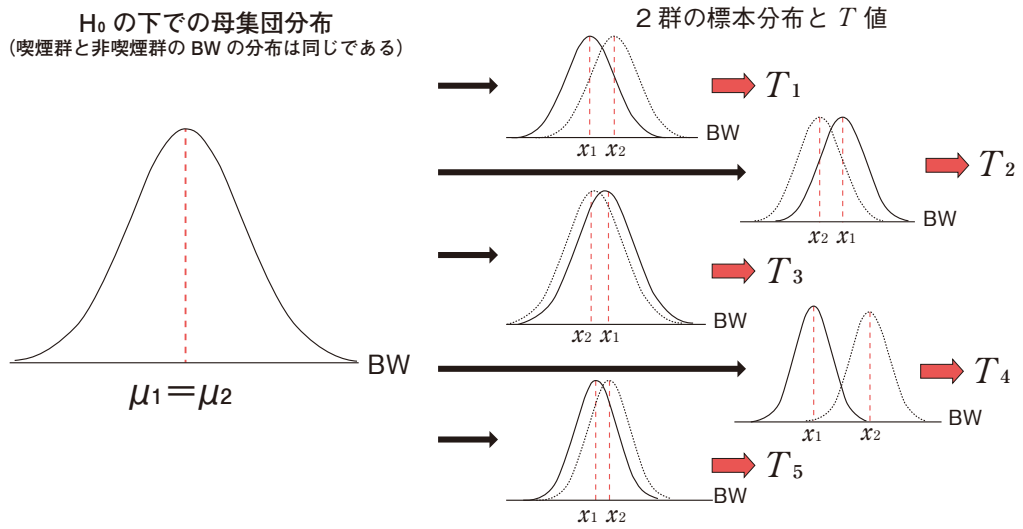


図 41. H_0 の下での 2 群標本とそれらから求めた T 値

元々はひとつの正規分布からデータを抽出してきたので、本来は 2 つの標本は同一になるはずである。つまり、T 検定統計量の分子は 0 となるはずである。しかしながら、無作為抽出なので、母平均より大きいところから多めにとってきたり少なめに取ってきたりするので、標本平均 X_1, X_2 は必ずしも一致しない。それでも T 検定統計量の分子の絶対値は、0 に近い値になることが多いはずである。その一方で、確率は小さいが極端に離れた標本分布になることもある。この場合、$X_1 - X_2$ の絶対値は大きくなるので T 検定統計量も非常に大きくなる（もしくは小さくなる）。これらの分布が図 42 で示された t 分布である。0 付近では確率密度が高く（起こる確率が高く）、T 値の絶対値が大きくなるほどその確率は小さくなる。

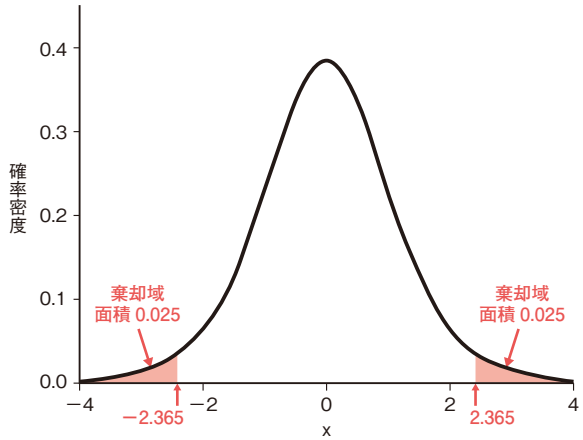


図 42. 自由度 $\nu = 7$ の t 分布

6-10-6 P 値

上述では、H_0 の下で T 検定統計量の分布、n_1 と n_2 に依存して決まる t 分布を説明した。t 分布の 0 付近の T 値は H_0 の下で比較的起こりやすいことを示し、絶対値の大きい T 値（たとえば、T の絶対値が 2 以上）は H_0 の下ではまれにしか起こらないことを示している。

標本の大きさ n_1、n_2 の一組の標本があり、その T 値が正の数のとき、P 値を次のように定義する。

P 値 = {t 分布の $t \geqq T$ となる確率} + {t 分布の $t \leqq T$ となる確率}
= {t 分布の $t \geqq T$ の横軸と対応する曲線で囲まれる部分の面積}
+ {t 分布の $t \leqq T$ の横軸と対応する曲線で囲まれる部分の面積}

P 値も n_1, n_2, T 値をエクセル関数に入力すれば容易に求めることができる。多少乱暴ではあるが、P 値を解釈すると以下のようになる。

P 値とは、H_0 の下でのその標本の得られやすさを示す指標

P 値は次の性質をもつ。

1）$0 \leqq P \leqq 1$

2）T 検定統計量が 0 のとき、$P = 1$

3）BW のような量的変数のとき、$P > 0$

6-10-7 有意水準と P 値の関係

p.404「6-10-4 有意水準」では、有意水準 α は 0.05 とするのが通例であることを述べた。この α と P 値の大小関係により対立仮説を採択するか否かを決定する。

【$P < \alpha$ のとき】

収集された標本の H_0 の下での得られやすさは 0.05 未満である。2 群の母平均が等しい（母集団分布が同一である）と仮定したときに、得られた標本はあまりにも起こりにくい、100 回標本抽出をしたときに 5 回程度しか起こらない珍しい標本である。

したがって、H_0 が正しいと考えるよりは、2 つの母集団の平均が異なる 2 つの正規分布から抽出されたものである可能性が高い。結論として、H_0 を棄却して H_1 を採択する。

【$P \geqq \alpha$ のとき】

収集された標本の H_0 の下での得られやすさは 0.05 以上である。2 群の母平均が等しい（母集団分布が同一である）と仮定したときに、得られた標本は比較的起こりやすいものである。

したがって、H_0 が正しいという仮説は棄却することができず保留する。

以上、仮説検定の手順をまとめると次のようになる。

【手順】

1）帰無仮説（H_0）と対立仮説（H_A）を立てる

2）有意水準 α を決める。通常、両側検定で $\alpha = 0.05$

3）標本を抽出する

4）データの性質を確認して検定手法を決める

5）検定統計量を算出する

6）検定統計量が棄却域に入る　⇒対立仮説を採択（帰無仮説を棄却）

検定統計量が棄却域に入らない（採択域に入る）

⇒帰無仮説を保留（とる）

6-10-8　乳幼児データにおける喫煙群と非喫煙群の BW の母平均の有意差検定

これまでの説明に基づき、妊娠中の喫煙が BW を低下させるかどうかの有意差検定の手順と結果をまとめる。

1）H_0 と H_1

p.403「6-10-3 帰無仮説と対立仮説」で述べたように、以下である。

H_0：喫煙群の BW の母平均 μ_1 ＝ 非喫煙群の BW の母平均 μ_2

H_1：喫煙群の BW の母平均 μ_1 ≠ 非喫煙群の BW の母平均 μ_2

2）有意水準の設定

6-10-4 で述べたように、$\alpha = 0.05$ とする。

3）*T* 検定統計量と対応する *P* 値の算出

T 検定統計量と *P* 値をまとめて算出するためにエクセルの「データ分析」を用いる。手順は以下のとおりである。

（1）乳幼児データの母親の喫煙の有無、BW の 2 変数のみを使う。並べ替えツールを用いて喫煙群、非喫煙群ごとに BW をひとまとめにする。

並べ替えは以下の手順で行う。

データタブ⇒並べ替えとフィルターグループ⇒並べ替え

（2）データタブ⇒分析グループ⇒データ分析を選択する。

データ分析が表示されない場合には

ファイルタブ⇒オプション⇒基本設定のアドイン⇒分析ツール選択⇒設定⇒分析ツールにチェック

（3）データ分析において

データ分析⇒t検定：等分散を仮定した2標本による検定

（4）t検定のウィンドウにおいて

変数1の入力範囲（1）⇒　喫煙群のBWを指定

変数2の入力範囲（2）⇒　非喫煙群のBWを指定

ラベル　⇒　1行目に変数名が入っているときはこれにチェックを入れる

出力オプション　⇒　適宜入力

OK

（5）以下の出力結果を得る。

表13

	変数1	変数2
平均	3130.302	2851.5
分散	186291.3	203926.6
観測数	86	14
プールされた分散	188630.7	
仮説平均との差異	0	
自由度	98	
t	2.227427	
P（$T<=t$）片側	0.014103	
t境界値　片側	1.660551	
P（$T<=t$）両側	0.028207	
t境界値　両側	1.984467	

読み取り方を説明する。検定結果を読み取りに必要なところだけを読む。

①変数1は非喫煙群、変数2は喫煙群で、標本平均、標本分散、観測数（標本の大きさ）が算出されている。

② T検定統計量が$t = 2.227$。

③ T検定統計量に対応するP値は、P（$T<=t$）両側$= 0.028$。

④検定結果は次のようになる。

喫煙群と非喫煙群のBWの母平均について有意差検定を行ったところ、$P = 0.028$であり有意水準0.05より小さい。したがって、喫煙群のBWの母平均は非喫煙群のそれに比べて有意に低い。

6-11 医学でよく使う仮説検定

前節では、仮説検定の手順とその特別な用語について説明した。また、2群の母平均が等しいかそうでないかの検定を例示した。調べたいこととデータの性質により、それに合った検定手法を用いる必要がある。この節では、医学データの検定でよく用いられる検定について説明する。

【2群の母平均の差の検定】

量的変数における2群の母分布を考えて、その代表値である母平均に差があるか否かを判定する検定である。前節で説明した検定である。

〔例：妊婦の受動喫煙のあり群となし群における BW の母平均の差の検定〕

喫煙しない女性が妊娠して夫の喫煙に暴露（さらされるの意味）した群と、夫も喫煙しない群とで、BW の平均に差があるかを判定する。標本は受動喫煙の影響を調べるために、妊娠中に喫煙のある母親14例は除外する。したがって、乳幼児データ100例の中の母親の喫煙なしの86例である。

1）臨床的仮説

受動喫煙群の BW の母平均は非受動喫煙群の母平均に比べて低い。

2）統計的仮説

H_0：受動喫煙群の BW の母平均 μ_3 ＝ 非受動喫煙群の BW の母平均 μ_4

H_1：受動喫煙群の BW の母平均 μ_3 ≠ 非受動喫煙群の BW の母平均 μ_4

3）条件と検定手法

（1）2群の母分散が等しいならば

Student の t 検定

（2）2群の母分散が等しくないならば

Welch の t 検定

t 検定には2種類あり、その使い分けを行う必要がある。

4）データ分析の「t 検定：等分散を仮定した2標本による検定」で解析した結果、次のようになる。

表 14. 非受動喫煙群と受動喫煙群における BW の母平均の差の検定

t 検定：等分散を仮定した 2 標本による検定

	変数 1	変数 2
平均	3186.611	3089.76
分散	165763.6	200749.166
観測数	36	50
プールされた分散	186171.9	
仮説平均との差異	0	
自由度	84	
t	1.026915	
P（$T <= t$）片側	0.153704	
t 境界値　片側	1.663197	
P（$T <= t$）両側	0.307408	
t 境界値　両側	1.98861	

推定結果と検定結果は、以下のようにまとめられる。

（1）母親の喫煙なしの 86 例のうち、非受動喫煙群は 36 例、受動喫煙群は 50 例であった。

（2）非受動喫煙群の平均と標準偏差は 3,186g と 407g、受動喫煙群のそれは 3,089g と 448g であった。

※標準偏差は分散の平方根、別途、p.365「標準偏差」の例で説明した＝ STDEV.S（データ配列）で求める。

（3）分散は異ならないと仮定して、Student の t 検定を行った結果、P 値 = P（$T <= t$）両側 = 0.307 であり有意水準 0.05 より小さい。したがって、H_0 を棄却できないので、2 群の母平均に有意差はない。

【2 群の母割合の差の検定】

〔例〕質的変数、特に 2 値変数（喫煙あり・なしのように 2 つの値しかとらない変数）の喫煙あり群と喫煙なし群における、結果変数（妊娠中の異常あり、なし）の異常ありの母割合の差の検定である。得られる標本は、以下のクロス集計表のようにまとめられる。

表 15. 妊娠中の喫煙あり群と喫煙なし群の異常あり母割合の差

	妊娠中の異常		
喫煙	なし	あり（割合）	計
なし	n_{11}	n_{12}（p_1）	a
あり	n_{21}	n_{22}（p_2）	b
計	c	d	N

ここで、n_{11}，n_{12}，n_{21}，n_{22} はそれぞれの人数、a，b，c，d は行方向、列方向の合計数、N は総数である。p_1，p_2 は妊娠中の異常ありの標本割合である。

1）臨床的仮説

喫煙群の妊娠中の異常の母割合は非喫煙群の母割合に比べて高い。

2）統計的仮説

H_0：喫煙群の異常の母割合 γ_1 ＝ 非喫煙群の異常の母割合 γ_2

H_1：喫煙群の異常の母割合 γ_1 ≠ 非喫煙群の異常の母割合 γ_2

※ 母割合の γ はガンマと読む。

3）条件と検定手法

（1）症例数が多いとき（クロス集計表の n がすべて 5 より大きいとき）

Pearson のカイ 2 乗検定（本書では、カイ 2 乗検定とよぶ）

（2）症例数が少ないとき

Fisher の正確検定

4）エクセルでの検定が可能である。

n_{11}，n_{12}，n_{21}，n_{22} とそれらに対応する期待度数 e_{11}，e_{12}，e_{21}，e_{22} を chisq.test 関数または chitest 関数に入力してカイ 2 乗検定の結果である P 値を求める。なお、期待度数は、たとえば、n_{11} については e_{11} ＝ （c × a）/N で求める。掛け算の指定はエクセルでは「*」を用いる。

【その他の仮説検定手法】

上で示した検定手法以外に、検定の目的、変数の種類、データの性質によりいくつかの検定手法の使い分けが必要となる。ただし、エクセルでデータ分析や関数を使って処理できるものとそうではないものとがある。