データの科学的な見方 2018

松本治彦

2017.4.11 読売　厚労省推計　2065 年人口 8808 万人に、出生率 1.44 減少やや鈍化

厚労省の国立社会保障・人口問題研究所は 10 日、2065 年までの日本の将来推計人口を公表した。1 人の女性が生涯に産む子どもの数を示す合計特殊出産率が、前回推計（12 年）の 1.35 から 1.44 に上方修正され、総人口が 1 億人を下回る時期は前回よりも 5 年遅い 53 年となった。ただ、人口減少の傾向は続く見通しで、15 年の 1 億 2709 万人は 65 年に 8808 万人まで減少する。推計の柱となる出生率を上方修正したのは、近年、30～40 歳代の出生率が上昇しているためで、前回推計で示された 60 年の出生率より 0.09 ポイント上回った。同研究所は出生率が上昇した背景について「保育の受け皿整備など子育て支援策の充実が影響した」と分析しているが、人口を維持するために必要な出生率 2.07（15 年時点）には遠く及ばなかった。推計によると、総人口は 53 年に 1 億人を割り込み、その 10 年後の 63 年には 9000 万人を下回る。65 歳以上の高齢者が人口に占める割合を示す「高齢化率」は、15 年の 26.6%から、65 年には 38.4%まで上昇、人口の 4 割に迫る。高齢者人口は、42 年に 3935 万人でピークを迎え、65 年には現在とほぼ同水準の 3381 万人に戻るが、それ以外の世代の減少幅が大きいため、高齢者の割合が増える。15 歳未満の年少人口は 15 年の 1595 万人から 65 年に 898 万人まで減り、人口に占める割合は 12.5%から 10.2%に低下する。経済活動の中心となる 15～64 歳の生産年齢人口は、7728 万人（人口割合 60.8%）から 4529 万人（同 51.4%）に減少する。平均寿命は、1 5年の男性 80.75 歳、女性 86.98 歳から 65 年には男性 84.95 歳、女性 91.35 歳まで延びる。

2017.4.11 読売　出生率なぜ改善？30～40 歳代の出産増

将来推計人口とは

約 50 年先までの日本の人口と年齢構成を、1 年毎に推計したデータ。直近の国勢調査が確定する日本の総人口が推計のベースとなり、対象には外国人も含まれる。1 人の女性が生涯生む子どもの数を意味する「合計特殊出生率」のほか、総人口に対する年間の志望者の割合を占した「死亡率」、「国際的な人口移動」など、人口の増加や減少に影響する数値を予測して、推計値をほぼ 5 年毎に公表する。出生率と死亡率は、最も可能性が高い標準的な「中位」に加え、「低位」、「高位」の 3 種類があるが、中位の数値が最も多く使用される。

出生率がさらに改善したのはなぜ

2011～15 年の 30～40 歳代の女性による出生率の実際の値が、前回 12 年時の推計を上回ったためだ。女性が初めて結婚する年齢の平均が高くなる一方で、50 歳時点で結婚していない女性の割合が低下する見通しであることも、プラスに影響したようだ。

データはどう活用されるのか

高齢者が受け取る公的年金の財政状況を確認したり、医療、介護などに必要な費用を計算したりするのに利用する。約 30 年先までの人口や、世代構成を市町村別に推計した「地域別将来推計人口」も後に公表され、地方自治体が町づくりに活用する。

2016.8.5 日経　AI、がん治療法助言、白血病のタイプ見抜く

膨大な医学論文を学習した人工知能（AI）が、診断が難しい 60 代の女性患者の白血病を 10 分ほどで見抜いて、東京大医科学研究所に適切な治療法を助言、女性の回復に貢献していたことが 4 日、わかっ

た。使われたのは米国のクイズ番組で人間のチャンピオンを破った米 IBM の「ワトソン」。東大は昨年からワトソンを使ったがん診断の研究を始めており、東條教授は「AI が患者の救命に役立ったのは国内初ではないか」と話している。他にもがん患者の診断に役立った例があるという。AI は物事を学習し、考える能力を持つコンピューターのプログラム。チェスや囲碁などで人間に勝つだけでなく、今後は医療への本格的応用が進みそうだ。

女性患者は昨年、血液がんの一種である「急性骨髄性白血病」と診断されて医科研に入院。2 種類の抗がん剤治療を半年続けたが回復が遅く、敗血症などの危険も出た。そこで、がんに関係する女性の遺伝子情報をワトソンに入力すると、急性骨髄性白血病のうち「二次性白血病」というタイプであるとの分析結果が出た。ワトソンは抗がん剤を別のものに変えるよう提案。女性は数カ月で回復して退院し、現在は通院治療を続けているという。**東大と IBM は昨年から、がん研究に関連する約 2 千万件の論文を学習させ、診断に役立てる臨床研究を行っている**。

2016.8.31 日経　受動喫煙　肺がん 1.3 倍、国立がんセンター　リスク評価、因果関係「確実」と指摘

国立がん研究センターは 30 日、家庭や職場など人が集まる場所で周りが吸ったたばこの煙にさらされる受動喫煙がある人は、肺がんにかかるリスクが約 1.3 倍に高まるとする研究結果を発表した。同センターはこれまで受動喫煙が招く肺がんのリスク評価を「ほぼ確実」としてきたが、科学的裏付けがとれたとして「確実」に引き上げた。予防対策に生かす。

国立がん研究センターがん対策情報センターの若尾センター長は「日本の受動喫煙対策は世界の中で最低レベルにある。東京五輪を契機に屋内完全禁煙を実施する必要がある」と訴えた。研究グループは受動喫煙と肺がんの関連を示した 420 本の論文の中から、1984 年から 2013 年に発表された 9 本の論文を選び、たばこを吸わない人が受動喫煙によって肺がんになるリスクを分析した。その結果、受動喫煙のある人はない人より肺がんにかかるリスクが 1.28 倍だった。受動喫煙と肺がんの関係は 80 年代から指摘されていたが、個別の研究では科学的な根拠が無く、リスクを高めるかどうかは確実とは言い切れなかった。複数の論文をそろえて分析したところ、受動喫煙が肺がんのリスクを高めることが確実となった。研究結果を踏まえて、同センターは喫煙、飲酒、食事など 6 項目で予防法を示している「日本人のためのがん予防法」でも現行の「他人のたばこの煙をできるだけ避ける」から「煙を避ける」と修正した。受動喫煙は肺がんだけでなく、循環器疾患や呼吸器疾患などにも影響する。厚生労働省研究班は受動喫煙が原因で死亡する人は、肺がんや脳卒中などを含めて国内で年間 1 万 5 千人に達するとの推計をまとめている。同センターの片野田統計室長は「遅きに失した感はあるが、対策を急ぐ必要がある」と話す。

統計学と医療・看護のかかわり

看護師は病棟で、看護記録や医師の診療録（カルテ）をみる機会が多い。そこには、当然のことながら身長や体重、体温、血液型、各種の検査結果などが患者ごとに記録されている。これらのデータの意味を正しく理解するためには、統計学の知識が必要となる。

例えば、全国の患者数を知りたい場合、患者数などのデータは国が公表しているデータベースを調べることで、その数値データを知ることができる。

受け持った患者が罹患している疾患について調べたいと思った場合、統計学の知識がなければ、データをさがしあてることは困難であるし、たとえみつけたとしても、そのデータがどのように集められて、どのように表されているのかを理解しなければ、データの示す意味をくみ取ることは難しい。

例えば、臨床に出た後、病棟で日常的に実施されているあるケアについて、それを患者に実施することは本当に患者の満足度を高くするのだろうか、という疑問が生じたとする。そのようなとき、その疑問をともに看護研究を行うことによって、そのケアの効果を確かめることができるかもしれない。また、研究によって得られた結果・結論を病院で共有したり、論文などにまとめて公表したりすれば、もっと多くの患者の満足度を高めることができるかもしれない。看護研究において、客観的なデータによって結果を出すことは、必要不可欠なことであり、そのために統計学は強力な道具となる。

（統計学が最強の学問である「西内啓著 2013.4」より「疫学」の部分を中心に抜粋）

統計学の６つの分野

統計学は数学的な理論に基づくが、それを現実に適用したときには必ずいくつかの仮定や、仮定の扱いに関する現実的な判断が必要になる。この現実的な判断は、分野ごとの哲学、目的、伝統や、扱おうとしているデータの性質によって左右される。

例えば、回帰モデルを使うとき複数の説明変数の間に相乗効果なしと仮定している。一方、その仮定をどう取り扱うか、という考え方は数学的な理論ではなく分野ごとの視点で異なる。

① 実態把握を行う社会調査法
② 原因究明のための疫学・生物統計学
③ 抽象的なものを測定する心理統計学
④ 機械的分類のためのデータマイニング（データの傾向、相関、等の情報を見つけ出す手法、技術）
⑤ 自然言語処理のためのテキストマイニング
⑥ 演繹（前提が正しいなら必ず結論も正しくなる）に関心をよせる計量経済学

正確さを追求する社会調査のプロたち

一般に「統計をとる」という表現は、単にデータを集めるという意味で使われる。社会調査に関わる統計家の「平均値やパーセンテージ」に対するこだわりは、「ただの集計」のレベルを大きく超える。ニューディール政策のころに実用化されたサンプリング調査を発展させ、可能な限り偏りなく、求められる誤差の範囲に収まる推定値を最も効率よく得るために、彼らは研究し続けている。

得られるべきデータが測定できなかったことを「欠測」と呼ぶが、社会調査の専門家は可能な限りこの欠測を減らすため調査員を訓練する。また調査方法の改善だけでは対処できない欠測を補完し、推定値の偏りを補正するための様々な手法を考案してきた。こうした統計家の関心は、議論の土台となる正確な数値を推定することにある。

ビジネスの領域では、マーケティング調査に社会調査の専門家がしばしば携わる。

「妥当な判断」を求める疫学・生物統計家

ものや人間以外の生物を対象にする限り、**ランダム（無作為の意味）化比較実験**は比較的容易。また、倫理や感情によってランダム化が許されない人間対象の領域では、疫学的な方法論を用いる。この両者に共通する考え方は、最終的に結果に与える影響の大きい「原因」を探すことである。逆に言えば、ｐ値に基づき「原因」がちゃんと見つけられるのであれば、推定値の「全国民におけるあてはまり」という社会調査分析の統計家が重視する点についてはそれほどこだわれない傾向にある。

（＊ｐ値；実際には何の差もないのに誤差や偶然によってたまたまデータのような差が生じる確率のこ

と）

もちろん、仮に「若者だけに限定すると逆に喫煙でも寿命が伸びる」という、結論を覆すレベルの強力な交互作用であれば問題になるが。どちらにせよ大きな影響があるなら、とりあえず喫煙率は下げた方がいいんじゃないか？という妥当な判断が下せれば、ある程度それで満足なのである。

そのため生物統計家や疫学者は「国全体からランダムサンプル」という点に関してはほとんどこだわりを見せない。「あくまでこの結果は医者という偏った集団のデータですがこういう関係が見られました」と注釈つきで普通に発表する。また、「他の集団でどうかは厳密にはわかっていませんので応用する際には注意してください」とか、「今後の課題として別の集団でも同じ関連性が見られるのか確認する必要があります」という文章が、誠実な論文には必ずといっていいほど記述されている。

こうした考え方は、疫学や生物統計学において十分な数の「全体からのサンプル」を得ようとすればとんでもないコストと手間がかかる、という現実的制約が影響している。

終わりのない言い争い

だから疫学者や生物統計家は、「ランダムサンプリングによる正確な推定値」よりも、「ランダム化による妥当な判断」を大事にする。

統計学は最善最速の正解を出す

統計学が最強の武器になるワケ

なぜ統計学は最強の武器になるのか、その答えは一言で言えば、どんな分野の議論においても、データを集めて分析することで最速で最善の答えを出すことができる。判断を誤れば 10 万人の命が失われる意思決定、「公衆衛生」「社会医学」「保健行政」などの領域では、今この瞬間にも慎重な議論を経てそう言った決定を下す。例えば、日本では毎年 35 万人ほどががんで亡くなり、19 万人ほどが心臓病で亡くなり、3 万人ほどが自殺している。適切な予防や治療の方策さえ取れれば、このうち何万人もの命が救われる。

こうした大量の人命がかかった間違いの許されない選択において最善の答えを出すために、人類は 19 世紀のロンドンで、史上初めて統計学の力を使って万単位の人命を奪う原因に戦いを挑んだ。

原因不明の疫病を防止するための学問を「疫学」と呼ぶ。世界初の疫学研究は 19 世紀のロンドンで、コレラに対して行われた。この中で統計学が大きな役割を果たした。当時、ロンドンは産業革命の真っ最中で、農業で食べていけなくなった人々が都会へ押し寄せ、工場で労働者として働くようになり始めた時期。急激な人口の増加に都市の発達が追い付かず、狭く不潔な地域に粗末な家がひしめき、その家の中に貧しい人が押しこめられ、下水も整備されないためにゴミや排せつ物が庭や地下室、道端、そこらじゅうに溜め込まれた。そした「臭い地域」に住む臭い労働者たちの多くがコレラで死亡したため、悪臭を取り除けばコレラもなくなるのではと考えた。

さらに、役人の中には町中の汚物を片っ端から清掃し、下水を整備し汚物を川へ流せるようにする政策を取った（1 度目と 2 度目の大流行期間の間の期間）。彼らの努力にかかわらず「二度目の大流行時（死亡者約 7 万人）は、むしろ最初の大流行時（死亡者約 2 万人）よりも大量の死亡者を出した。

「疫学の父」ジョン・スノウの活躍（**過去問；2005 年、第 91 回出題あり**）

コレラで亡くなった人の家を訪れ、話を聞いたり付近の環境をよく観察する

同じような状況下でコレラにかかった人と掛かっていない人の違いを比べる。

仮説が得られたら大規模にデータを集め、これらの発症/非発症と関連していると考えられる「違い」について、どの程度確からしいか検証する。

彼は調査した結果を詳細なレポートとして冊子の形にまとめ上げているが、その中で最も端的にこれらの予防法として示しているのは図表 3。

	家屋の数	コレラによる死亡者	1 万件当たりの死亡者数
水道会社 A を利用	40046	1263	315
水道会社 B を利用	26107	98	37

ただ、使う水道会社だけが異なる家々の間で 8.5 倍もリスクが違う。スノウの提案したコレラ流行の解決策はごくシンプルで「とりあえず、しばらく水道会社 A の水を使うのを止める。以上」。なおスノウがこの提案してから 30 年後、ドイツの細菌学者ロベルト・コッホはコレラの病原体である「コレラ菌」を発見、その結果、コレラが水中に生息すること、コレラ患者の排せつ物に含まれること、そしてコレラ菌の存在する水を飲むことでコレラに感染することが証明された。実は、水道会社 A と水道会社 B の違いは、前者がロンドンのテムズ川の下流から、後者がテムズ川の上流からそれぞれ採水している。当時のテムズ川に役人の努力によって大量にコレラ患者の排せつ物が流し込まれている。

残念なことにスノウの主張は「科学的ではない」あるいは「確実な証拠がない」として学会や行政からは退けられたが、彼の助言に従った町ではこれらの感染が止まった。

コレラの話からも分かるように、頭やセンスや行動力に優れた人たちを集めて話し合わせただけでは、こうしたシンプルかつ強力な解決策というのは出てこない。むしろ握りつぶされることが多い。代わりに出てくるやり方が、理屈としては一見正しくても、無益もしくは有害であることもしばしばである。

人類の寿命は疫学が伸ばした

スノウの提示した「疫学」という考え方は、徐々に医学全域において欠くことのできない重要なものとなった。たばこを吸えば肺癌をはじめとした癌になるリスクが上昇するということも、血圧が高ければ心臓病や脳卒中になるリスクが高まるということも、現代に生きる我々にとっては当たり前の常識である。しかし、ほんの 50 年前に、アメリカのフラミンガムという田舎町で行われた大規模な疫学研究の結果が公表されるまでは、全く明らかではなかった。それまでは医者や科学者の中でも、たばこが健康に悪いことなのかどうか、あるいは血圧が高いことが悪いことなのかどうか、様々な説があり侃々諤々の議論が重ねられてきた。

だが、「癌を減らしたければとりあえず喫煙率を下げろ。以上」とか「心臓病を減らしたければとりあえず血圧を下げろ。以上」といった疫学研究のシンプルな答えが侃々諤々の議論をブッ飛ばしたことで、医学研究と健康政策の方針は変わり、50 年前よりも我々の寿命は随分と伸びた。

もしこの判断を、今も、議論だけに基づいた誤ったものとしていたならば、一体どれだけの命が失われていたのか見当もつかいほど統計学は力を発揮したのである。

すべての学問は統計学のもとに

「エビデンス」が医療を変えた

「疫学」という、データと統計解析に基づき最善の判断を下そうという考え方は、スノウの発見から 100 年ほどかけて、医学の領域において欠くことのできないものとなった。現代の医療で最も重要な考え方として EBM（Evidence-Based Medicine)、日本語にすると「科学的根拠に基づく医療」というものがある。この科学的根拠のうち最も重視されるものの一つが、妥当な方法によって得られた統計データとその分析結果である。スノウの疫学は基本的なデータの集計によってコレラのリスク要因を明らかにしたが、疫学の方法論は、徐々に現代的な統計学の進歩を取り込み、より高度な正確なリ

スクの推定を可能にした。

人間の身体には不確実性が多く、データを取って分析すると、生理学的な理屈の上では正しいはずの治療法が効果を示さないケースや、経験と権威にあふれる大御所の医師たちがこれまで続けていた治療方法が実は全く誤りだった、という事例が少しずつ明らかになっている。

そのため、医師の経験と勘だけでなく、きちんとしたデータとその解析結果、すなわち、エビデンスに基づくことで最も適切な判断をすべきだ、というのが現在医学において主流の考え方。この EBM の考え方が世界的に広まったのは 1980 年代から 90 年代にかけてで、現在、責任ある立場で臨床を仕切っている医師たちの多くは「学生時代にはほとんど習っていなかったこと」。

医師に対する統計学の教育にはアメリカでも課題が多い。「研修医に基礎的な統計学のテストをした結果、大変に残念なものだった」。

医学において統計学的なエビデンスが最重要視されることに間違いない。例えば、製薬会社で新しい薬を作った時は、綿密に計画された研究方法で採取したデータに対して適切な統計解析を行う。その結果を厚生労働省に提出しなければ、新薬が許可されたり、保険適用が認められることはない。

エビデンスは議論をブッ飛ばして最善の答えを提示する。もちろん、データの取り方や解析方法によって、どれほどのレベルで正しいと言えるかとか、どこまでのことを正しいと主張して間違いないのかは異なってくる。しかしながら、エビデンスに反論しようとすれば理屈や経験などではなく、統計学的にデータや手法の限界を指摘するか、もしくは自説を裏付けるような新たなエビデンスを作るかといったやり方でなければ対抗できない。

一般的な知識としても統計学が重要になっています。例えば、テレビの特番、あるいはワイドショーで質問中の「人気」について考えるとき、あまり正確とは言えない方法でアンケートを取っています。「100 人中、60 人がイエスと答えました」で簡単に結論付けています。この結果が本当に使用できるかどうか、統計学の知識を使う必要があります。このような間違った情報や数字に騙されないためにも、最低限の統計学の知識が必要になっています。

また、看護の世界では根拠に基づいた看護 evidence-based nursing（EBN）が重要になっています。その根拠を導き出す手法が統計なので、間違った情報に基づかないように基礎的知識として統計学を身に付けることが必要です。

疫学の進歩が証明したタバコのリスク

タバコの箱を見てみよう

スノウ以後に進歩を遂げた疫学的方法に則り、喫煙と肺がんの関係性について統計学的な分析を試みた。

「ケースコントロール研究」の登場

彼らは 1948 年～1952 年にかけて、イギリス中の病院から 1465 名の肺がんによる入院患者を見つけ、彼らの性別・年代・社会階層や居住地域と喫煙歴の有無を調査した。そして同時に、喫煙以外の性別・年代・社会階層や今日中地域について同様の条件を満たす、肺がん以外の疾患で入院している患者を同数見つけ調査した。

図表 18　肺がんと喫煙の関連性

		総人数	喫煙者	非喫煙者
男性	肺がん患者	1357	1350	7
			99.5%	0.5%

	非肺がん患者	1357	1296	61
			95.5%	4.5%
女性	肺がん患者	108	68	40
			63.0%	37.0%
	非肺がん患者	108	49	59
			45.4%	54.6%

この結果についてカイ二乗検定を試みると、男性ではｐ値は 0.1%未満、女性でもｐ値が 1%未満となる。ともに誤差とは考えにくいレベルで肺がん患者の喫煙率が高かった。「ケースコントロール研究」と呼ばれるデータの取り方が重要。疫学におけるケースとは症例すなわち関心のある病気となった事例（患者）のこと。そしてコントロールとはその比較対照のことである（ちなみに「比較対照」は疫学の専門用語)。比較対照には「関心のある疾患とリスク要因の有無以外は条件がよく似た人」が選ばれる。だからドールとヒルは、喫煙というリスク要因以外の肺がんと関連しうる条件である、性別・年代・社会階層・居住地域といったものについて、調査対象とした患者と同様の人間を集めて男女別や年代別で区切ったグループごとに比較（専門用語ではこれを層別解析と呼ぶ）すれば、ランダム化しなくても「フェアな比較」ができるというものである。

「把握」と「予測」、そして「洞察」の統計学

ビジネスに必要なのは、人間を「洞察」するための統計学

人間の行動の「因果関係を洞察」する。因果関係の洞察以外の統計学の使用目的には「現状の把握」と「今後の予測」の２つがある。

「洞察」の統計学はどのように役立つのか

マーケティング部門などではしばしば予測よりも洞察のほうが重要になる。「どのようなプロモーションをすれば商品が売れるのか」「どのような商品を作ればヒットするのか」という洞察のほうが利益の源泉となる。購買という求める結果の背後にどのような原因が存在するのか、という因果関係を探り当てることが重要。

これは医学や公衆衛生学においてもまったく同じことが言える。**どうすればその人がより長く健康に生きられるか、という原因を発見することこそが医学で統計学を使う目的**。

「洞察」の統計学に必要な 3 つの知識

（１）平均値や割合など統計指標の本質的な意味の理解

（２）「データを点ではなく幅で捉える」という考え方

（３）「何の値を何ごとに集計すべきか」という考え方

「平均値」の本質が分かれば「割合」もわかる

平均値と割合と言うのは本質的にまったく同じ。「量的変数」は「平均値」の形で集計する。質的変数は「割合」を集計する。量的変数は「量として大きいか小さいか」という情報を示すものである。質的変数は「大きい小さいということではなく、そもそもの質が異なる」という情報を示す。

割合と平均値と言う全く別物の集計方法が存在しているわけではなく、数の形で表現できない量的変数については、それぞれの分類についての１か０かという形で表現される「該当するか度合い」という量的変数（二値変数）を考え、その平均値を計算している。

例として、100 人に対する調査で 60 人が男性と言うデータが得られたとき、男性の割合が 60%とい

う集計結果が得られたことになる。これを仮に「男性である度合い」という量的変数を考える。この「男性である度合い」調査の結果自分が男性であると回答した人なら１、そうでなければ０と言う値にあるものとする。この平均値は 0.6 となる。これは先ほどの 60%と言う割合と全く同じ値である。

データの存在する「幅」が重要

統計学は、平均値や割合を示す値の次に「おおよそデータはどこからどこまでの範囲に存在しているか」という幅を把握するための方法を生み出した。

「結果」と「原因」を絞り込め！

何の値を何ごとに集計すべきか、これは統計学を因果関係の洞察に使う上で最も重要な枠組み。因果関係とは、ある原因によってどのように結果が変わるのか、という関係。単純に平均値や割合での集計を行うにしても、適切な比較軸という考え方さえ適切であれば因果関係を見るための第一歩を踏み出せる。

せっかく経験と勘に反する新しい発見に出会うためのデータ分析と言う作業をしているのに、自分の経験や勘の検証しかおこなわないのではもったいない。データ分析を因果関係の洞察、すなわち、最終的にコントロールしたい結果とそれに影響を与えうる原因の候補、という観点で捉える。

授業概要：

コンピューターやインターネットが発達した現在、膨大な情報の中から自分の必要な情報を選別し、それを整理する能力が必要になっています。また統計処理した数値がどのような意味をもつかを判断する能力も必要になっています。この授業の到達目標は、**導き出した統計値の科学的な意味を理解する**ことです。そのためには、まず**基本的な統計値の意味**をしっかりと理解した上で、**統計図、統計表の見方**を学習することが重要です。そうして区間推定や検定を通じて**データの科学的な見方**を身につけていきましょう。

到達目標：

統計値の科学的意味を理解できる能力を身につけることです。**加えて、保健師国家試験の中で、統計学、疫学および保健統計に係る問題が解けるようになること。保健師になると業務のかなりの部分は統計処理。看護師になっても、データ分析で統計処理が必要になるので。なお看護研究でも統計学が必要になります。**

受講の心得：

毎回の授業内容の疑問点は質問カードに書き込み、次回の復習タイムで理解度を深めるように努力してください。

成績評価：毎回配布する出席・質問・感想カードの内容、小テスト、定期試験で総合評価します。

関連する科目：保健統計学、疫学

授業計画：

トピックス、統計学は最強の学問、統計学の考え方、データの科学的な見方、統計データのまとめ方１（度数分布）、統計データのまとめ方２（図示法）、集団を表す代表値（平均、分散、標準偏差など）正規分布、推定１（母集団、標本とは）、推定２（区間推定、ｔ分布）、検定１（帰無仮説、有意水準など）、検定 2（平均値の差の検定ｔ検定、χ^2検定、その他の検定）、相関と多変量解析、保健師国家試験対策、まとめ

１．統計学の考え方、データの科学的な見方

21 世紀は地球規模でものを考える情報の時代。
コンピューターやインターネットは日常生活になくてはならないもの。
コンピューターやインターネットに振り回されずに情報を使いこなすには、統計学が必要。
情報に強い人は、平均値の意味をよく知っている人。
図表が正確に描ける人、それは統計の知識がある人。
基本的な統計の出し方を理解することが必要。
コンピューターのソフトや使用書は日進月歩、古いのは使えない。
統計学の体系は大きくは変わらない。
人間はしばしば基本的なことを忘れる。
統計学は数学とは同じではないが、共通するのは数式を使って論理・推論を展開する学問。
情報科学の発展で仮説を立証する手段“統計学的検定”はますます重要。

例えば

医療の現場では肝臓ガンの手術を受ける患者に、この手術の 5 年生存率やエタノール注入療法などの内科的治療のメリットを数字で述べ、患者に選択権を与える必要があります。その治療成績などの根拠は、統計による判断が一般には最も客観的です。

また、薬の副作用の説明でも、副作用のあるなしは確率的統計の問題であり、科学的に説明できる人は統計に強い人です。

統計学とは

ある集団の状態を数量的に把握するための方法を統計的手法といい、これらを系統的に集大成したものが統計学です。

統計的手法には

どのような表をつくるのが最も適切か、どのように貯金をするのが最も有利か、など様々な内容が含まれ、日常生活と密着した分野です。

統計学は

生活や仕事のなかでの集団を対象とし、数量によってものごとの特性や規則あるいは法則を見出そうとする学問です。

1－1　統計学におけるものの考え方

（専門の科学には、その科学自体に根ざした独自のものの考え方や目的があります）

１．ありのままに観察し、正確に数え、数字を通して把握する

計算を「正確」にする、ありのままに観察することが「正確」につながる、面接調査結果から集計を行う場合、実際に面接した人についてだけ結果報告する。

２．その数字がなにを意味しているのか考える

なにを意味するのか粘り強く考えてみる。１枚の図・表が理解できると一挙に理解が進む。

３．その数字は真実を示しているのか、偶然の要素はないかどうかをよく吟味する
偶然か否かを検証していく（推測統計学）

４．使用されている分類や定義が適切かどうか検討し、妥当性を確認する
学会や専門書において用いられている定義・分類・用語に従うことが望ましい
（再現性の保証、他調査との比較が可能）

５．分類された数字に、一定の変化や傾向があるかどうかを考察する
表や図から一定の傾向や変化を読み取ることが重要、これは法則性の発見にもつながる方法

５つの視点に留意し、くり返しこの原点に立ち返ることが統計的なものの見方を身につける方法

日本経済新聞　2005.2.21　統計のウソより抜粋

社会の実態を浮き彫りにするはずの公式統計。公正中立なものとされるが、全てを鵜呑みにするのは禁物。算出方法に疑問符がつくものが少なくない。独り歩きする数字は、物事の本質を見誤らせ、必要な改革を遅らせることになる（政策擁護で変更）。

食料の自給率に関すること

供給された食料に占める国産品の割合を熱量（カロリー）換算で表すと「40%」
金額換算すると　「70%」

↓

こんなに開きがでる理由は、国産品が多い野菜や果物などは値段が高くてもカロリーが低いため

↓

これまで、食料政策をめぐる論議をするときは、カロリーベースが中心で、金額ベースは参考値でしかなかった。農林水産省は、「国内の生産活動を正当に評価するには、金額ベースも必要」として、今後両者を対等に扱う方針。

↓

農林水産省は、自給率の指標を自分たちの政策に合うように変えてきた。

↓

1987 年までは、自給率を金額ベースで表していた。ところが、88 年度からは「食料需給表」に金額とカロリーベースが併記されるようになった。95 年度版では、金額ベースが姿を消し、カロリーベースだけとなった。

↓

その結果、その年の政府が正式な自給率として発表した数字は、87 年度の 76%から 95 年度の 42%に急落した。

↓

なぜ、自ら進んで自給率を下げるようになったのか？
「背景には、コメ市場の自由化問題がある」
88 年には牛肉・オレンジ問題が決着。農業交渉の次の焦点は「コメ」
「低い自給率を示し、コメを守れと主張する論拠にした」

＊ カロリーベースは、国際基準ではない。公式な指標として使っているのは、日本と韓国ぐらい。**先進国では、穀物自給率を使うことが多い**。農水省が日本の自給率との比較で示す他国のカロリーベースの数値は、同省の推定値です。

＊ 金額ベースでごまかして自給率のハードルを下げている。

＊ 穀物：農作物のうち、種子を食用とするため栽培するもの。米、麦、粟、稗、豆、キビ、とうもろこしなど。多く、主食とされる。

食糧自給率、低くない？2014.8.8 読売新聞より抜粋

昨年度ｶﾛﾘｰﾍﾞｰｽの自給率は 39%、基準を変えれば 65%になる！

＊ ＊（ｶﾛﾘｰﾍﾞｰｽ総合食糧自給率）＝（国民１人１日当たりの国内生産ｶﾛﾘｰ）÷（国民１人１日当たりの供給ｶﾛﾘｰ）

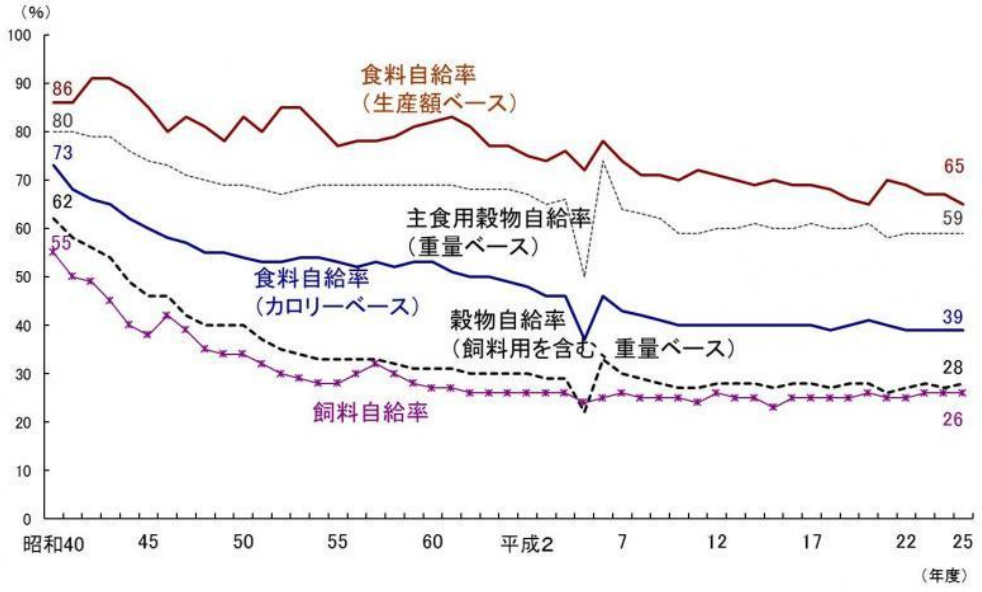
＊ ＊（国民１人１日当たりの供給ｶﾛﾘｰ）＝（国産供給ｶﾛﾘｰ）＋（輸入供給ｶﾛﾘｰ）＋（ロス廃棄ｶﾛﾘｰ）

＊ ＊畜産物は、国産であっても飼料を自給している部分しかｶﾛﾘｰﾍﾞｰｽの自給率に算入しない。

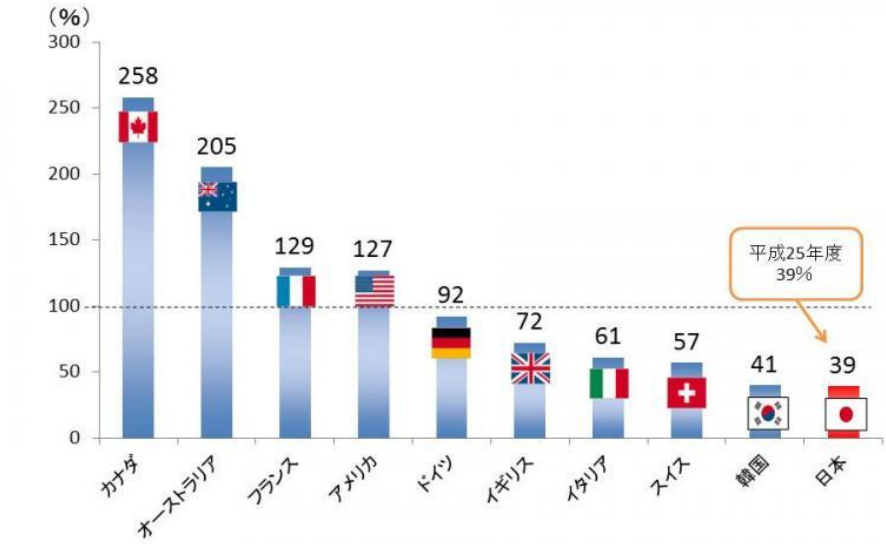
＊ 問題１；上記のことから考えると、輸入が減ると、ｶﾛﾘｰﾍﾞｰｽ総合食料自給率はどうなりますか？

＊ 問題２； 2005 年度の調査では、分母となる供給ｶﾛﾘｰは 2573 ㎉、日本人が１日に摂取する平均ｶﾛﾘｰは 1805 ㎉で、それ以外の（2573－1805＝768）㎉は廃棄されている。分母を摂取ｶﾛﾘｰ、国産供給ｶﾛﾘｰ 1013 ㎉で計算をし直すと、食糧自給率はどうなりますか？

＊ 問題３；卵の自給率は重量ﾍﾞｰｽでは 95%、ｶﾛﾘｰﾍﾞｰｽでは何%、鶏肉の重量ﾍﾞｰｽ自給率は 66%、ｶﾛﾘｰﾍﾞｰでは何%だと思いますか？



＊



資料：農林水産省「食料需給表」、FAO“Food Balance Sheets”等を基に農林水産省で試算（アルコール類等は含まない）。
（注）数値は平成23年（日本は平成23年度）。スイス、韓国のデータについては、各政府の公表値を掲載。

待機児童数の推移

2001 年に、厚生労働省がまとめる保育園の待機児童の算出方法が変わった。「通常の交通手段を使って 20～30 分未満で通える施設」に空きがあれば、待機児童とは見なさいといった項目が加わった。兄弟と同じ園に通うために空きを待っていた子どもなどが除外された。

その結果、**2001 年 4 月の待機児童数は 21,201 人と前年よりも 14,000 人も減った**。

この統計マジックは、「前年に始まった新エンゼルプランが掲げた**待機児童ゼロ**に近づける狙いではないか」といわれています。

育児休業取得率の推移

女性の場合、**2003 年度は 73.1%、前年度の 64.0%から大きく上昇。しかし、実は調査対象が違う**。2003 年度は「常用労働者 30 人以上の企業」に対して、2002 年度は「常用労働者 5 人以上の事業所」です。

小規模事業所が多くなるほど、育児休暇（育休）取得率は低くなる。調査は毎年実施されているが、統計には必ずしも連続性がない。

このように、どの統計も物事の一面でしかない。しかも、作成者の意図が隠されている場合もある。利用者は絶えず行間を読む姿勢が必要です。

＊行間を読む（文章に文字では書かれていない筆者の真意や意向を感じとる。）
＊科学的（論理的、客観的、実証的であるさま）
＊論理的（思考の形式・法則、議論や思考を進める道筋・論法に沿っていること）
＊客観的（個々の主観の恣意（勝手・きままの意）を離れて、普遍妥当性をもっているさま）
＊実証的（思考や推理によるのではなく、経験的な事実をもとにして明らかにされるさま）

1－2　統計学の方法

統計学は、**記述統計学**と**推測統計学**に分かれる。
記述統計学では、調査や実験で得られた多量のデータをまとめる（解析する）こと、すなわち、度数分布、平均値、分散、相関係数などの指標を用いてデータをまとめることで、データの背後に潜む何らかの特徴を探り、データから最大限の情報を獲得する。

記述統計学では

仮説を立てデータ収集
データ解析（度数分布、平均値、標準偏差など）
データに潜む何らかの特徴を探る
重要なことはデータの誤りをチェックすること（データ解析では必須）
整理段階の転記ミス・パソコン入力時のミス・外れ値（極端に大きいか、極端に小さい）

推測統計学では

データ解析で得られた特徴が本当に科学的立場から受け入れることができるか否か
仮説を立てて推論する
データ解析で得られた特徴
それらに関する全ての集団（母集団）の特徴と一致するかどうか
正規分布・推定・検定

例）；30%の効用があるといわれる新薬が開発されたとする。この薬を 10 人の患者に投与したところ、まったく効き目がなかった。このとき、この薬が同じ病気の患者に 30%の効き目があるという仮説は、はたしてどの程度信頼できるのか。検定結果は 5%の危険率で棄却。30%の患者に効果ありとの前宣伝はきわめて疑わしいものと推測される。

データとは何か

一定のルールに従って測定、あるいは観察された一連の数値、または文字の集合。いつ、何のために、どこで、誰を対象として、どのように収集され、何が記載されているのか、ということがわかると、これらの数値はそれぞれ意味を生じ、データとなる。

データの種類

量的データと質的データがある。

1．量的データ；「何らかの測定あるいは計測を行ってデータを得る」、身長、体重のように、何らかの単位をもち、数値そのものに意味があり、値の大小を比較できるデータのことを量的データと呼ぶ。
2．質的データ；単位のないデータ、例えば性、職業、好きな色のようなものが質的データと呼ぶ。

補遺；

量的データには、比率尺度（比尺度）と間隔尺度がある。間隔尺度は個々の値の間に等間隔が保証されている尺度である。比率尺度は等間隔性に加えてゼロを基点とすることができる尺度である。

質的データには、順位尺度と名義尺度がある。順位尺度は順番で順位だてられるが、個々の値の間に

等間隔性が保証されない尺度である。名義尺度は、その順番に意味がないものである。

データの集め方

調査と実験

データの収集方法には、２つの方法がある。実験室等で行う化学、物理あるいは生物実験と、一般地域住民を対象として行うフィールド調査がある。自分で収集した生データを１次データ、既存の資料のように加工されたデータを２次データと呼ぶ。

断面調査、前向き調査、後向き調査（**過去問**）

１）断面調査（横断的調査）

ある任意の一時点（あるいは一期間）を設定し、その時点における現況や実態を把握しようとするもの。時間を追った情報を与えてくれるものではない。因果関係にまで議論を広げることは不可能である。

２）前向き調査

時間を追って変化を調べ、因果関係を調べようとする調査研究の代表的なものが前向き調査。観察研究、臨床試験など。前向き調査は事象の時間的関連を調べることができ、その観察も時間を追って自然な状況を追いかけていくため、因果関係を調べるのには理想的な手段である。欠点としては、場合によっては結果がでるのが数十年先になる。

３）後向き調査

結果の有無によるグループ間の比較により、原因の分析に差があるかを調べる。現在存在している特定の結果から、時間を遡って過去の原因を探索する。

標本抽出法

国勢調査は、全数調査（悉皆調査；しっかいちょうさ**（過去問）**）であるが、普通の調査は仮想する対象全員（母集団）を調査するのではなく、その一部を調査し、母集団の特性を推測する。母集団から取り出された一部を標本と呼び、標本を取り出すことをサンプリングと呼ぶ。また標本を用いて行う調査を標本調査と呼ぶ。統計学は、標本調査で得られた結果から母集団の状況を調べる手法を与える。

一部から全体を推測する場合、その一部が全体の縮図となっていなければ、正しい全体を推測できないことは直感的にも了解できる。標本を選ぶ場合、あるルールに基づいて対象者を選ぶ必要がある。一般によく用いられるものに無作為抽出方法（**過去問**）がある。文字どおり何の作為もなくという意味です。対象者が選ばれる確率が、どの人をとってみても等しくなるような抽出方法です。乱数表を利用する。

2　統計データのまとめ方

２－１　度数分布

度数分布について、下記データ（107 人の男子学生身長データ）で作成した表 4 を用いて、説明します。

〔身長データⅠ(単位：cm)〕

158.0	159.0	159.5	161.0	161.0	161.5	161.5	162.0	162.0	163.0
163.0	163.0	163.5	163.5	164.0	164.0	164.5	164.5	165.0	165.0
165.5	165.5	166.0	166.0	167.0	167.5	167.5	168.0	168.0	168.0
168.0	168.0	168.5	168.5	168.0	168.0	167.5	167.5	169.0	169.0
168.5	168.5	169.5	169.5	170.0	170.0	170.5	170.5	170.0	170.0
171.0	171.0	171.0	169.5	169.5	171.0	170.0	170.5	170.5	170.0
170.0	171.5	172.0	172.0	173.0	173.0	172.0	171.5	171.5	172.0
172.0	171.5	172.5	172.0	172.0	173.5	174.0	174.0	175.0	175.0
174.0	173.5	173.5	174.0	175.0	173.5	174.5	174.0	174.0	175.5
176.0	176.5	176.5	176.0	177.0	176.5	176.5	177.0	176.0	177.5
178.5	178.0	179.0	178.5	179.5	179.5	183.0			

度数分布表（表４）とは、測定データをある一定の間隔（ここでは２ｃｍ間隔）に区切り、その中にあてはまる人数を記載したもの。測定データを区切った 1 つの層を階級、区切られた層の数を階級数、階級のまんなかの値を階級値、区切る目安とした間隔を階級の幅という。各階級の測定値の頻度（ここでは何人）を度数という。度数分布表をもとに、これを図にしたものをヒストグラム（頻度分布図）という（図 15)。

表 4　身長の度数分布・相対度数・累積度数・累積相対度数(大学生，男)

身長階級	度数	相対度数	累積度数	累積相対度数	階級値
157.5～159.5	2	0.0187	2	0.0187	158.5
159.5～161.5	3	0.0280	5	0.0467	160.5
161.5～163.5	7	0.0654	12	0.1121	162.5
163.5～165.5	8	0.0748	20	0.1869	164.5
165.5～167.5	5	0.0467	25	0.2336	166.5
167.5～169.5	17	0.1589	42	0.3925	168.5
169.5～171.5	19	0.1776	61	0.5701	170.5
171.5～173.5	14	0.1308	75	0.7009	172.5
173.5～175.5	14	0.1308	89	0.8318	174.5
175.5～177.5	10	0.0935	99	0.9252	176.5
177.5～179.5	5	0.0467	104	0.9720	178.5
179.5～181.5	2	0.0187	106	0.9907	180.5
181.5～183.5	1	0.0093	107	1.0000	182.5
計	107	1.0000	107	1.0000	

ポイント；データをどのように区切るかを検討する　↓

階級数の決定

階級数が少なすぎても、多すぎても集団の性質はわからなくなる。目安としては、10～20 くらい。階級の幅の決定には、集団の最大値と最小値を見つける。上記の身長のデータの場合、最大値と最小値をそれぞれ見つけてください。最大値 183.0 ㎝、最小値 158.0 ㎝　最大と最小の差（25cm）を 10～20 で割った値（この場合、10 で割ると 2.5ｃｍ）が階級の幅となりますが、実際には階級の幅はその前後の整数 2cm あるいは 3cm を便宜上用います。この例の場合は、階級の幅を 2cm とすると身長の度数分布表は表 4 のようになります。階級数は 13、160cm 以下、180cm 以上が少なく、168～171cm に多くの人が集中していることがわかります。

度数分布を作るときの約束ごと；

1）階級幅は一定。

2）年齢階級が 0～4、5～9、…….　としてあるとき、階級の幅が 4 歳ではなく、5 歳になる前日までが 0～4 歳の階級に入るので、階級の幅は 5 歳である。

3）階級が 140～145、145～150、…..　としてあるときは、145 は“145～150”の階級に入れる。

4）年齢階級が 0、1、2、3、4、5～9、10～14、……としてあるときは、0～4 歳までの間、年齢ごとに著しく変化するため、他の階級と同じように比較できないので、それぞれの年齢で度数を求めて、比較しようとしているのである。

累積度数　→　ある階級以下の度数を合計したもの。最期の階級では、累積度数は測定値データの合計数と等しくなる。

相対度数・累積相対度数　→　度数・累積度数を測定データの合計数で割ったものが、相対度数・累積相対度数。それぞれ度数・累積度数の百分率に一致する。

問題 1

ある会社の従業員男女それぞれ 10 人を対象に貧血検査（ヘモグロビン濃度 mg/dl）を行い、下記の

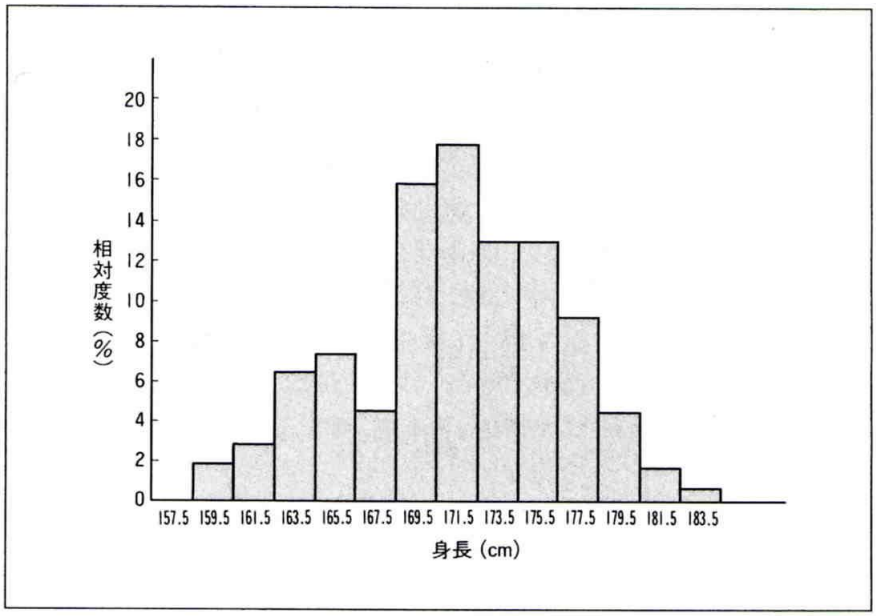
結果を得た。

男　15.5,15.0,14.0,14.5,13.5,10.0,16.0,16.5,17.0,15.0

女　10.0,15.0,14.5,12.5,14.0,17.5,8.5,10.0,11.0,14.5

度数分布表を男女別に作成してください。

図 15　身長のヒストグラム(相対度数，大学生，男子)



２－２　分割表（クロス集計表）

縦にある変数、横に別の変数をかいて、それぞれの項を分割集計（クロス集計）したものを分割表（クロス集計表あるいは単にクロス表（過去問））という。このように 2 つあるいは 2 つ以上の変数の間の関係をみる場合、分割表を作成するのが普通です。

事例）ある会社で 2 月 5 日午後 2 時 30 分から 8 時ごろにかけて、吐き気・嘔吐・下痢・腹痛を訴える患者が続出、当日、正午から昼食をとったが、給食を受けた 75 人中、50 人がスープを飲んだ。スープを飲んだ 50 人中、45 人が上記の症状を訴えた。スープを飲まなかった者のうち 2 人が上記の症状を訴えた。この場合、スープが食中毒の原因といえるか。

分割表の作り方）症状はすべて似ているので、症状のあり・なしを 1 つの変数、スープを飲んだ・飲まな

かったを別の 1 つの変数としてクロス表を作ると、表 5 のようになる。表 5 は 2×2 分割表（四分表）といい、食中毒関係の分野では、マスターテーブル（過去問）とも呼ばれる。ここでは、2 つのカテゴリー（グループ）に分けたが、症状を重症、軽症、症状なしの 3 つに分けることもできる。

表 5　事例に基づく分割表(クロス表)の作成

	スープを飲んだ（＋）	スープを飲まなかった(－)	合計
症状あり（＋）	45(95.7%) [90.0%]	2(4.3%) [8.0%]	47(100.0%)
症状なし（－）	5(17.9%) [10.0%]	23(82.1%) [92.1%]	28(100.0%)
合計	50[100.0%]	25[100.0%]	75

1)（　）内は行の変数(症状)の相対度数，［　］内は列の変数(スープ飲食)の相対度数をあらわしている。
2) 症状(＋)者 47 人中，45 人(95.7%)がスープを飲んでいることから，症状とスープの関係は濃厚であり，後述の χ^2 検定によって統計学的有意差を求めると，スープを飲んだことが有力な原因であることがみとめられる。さらに潜伏期・症状や疫学的な特徴から，ブドウ球菌性食中毒の可能性が示唆される。

問題２

男性の肺がん患者 100 名と、肺がんでない健常者 100 名について聞き取り調査を行った。その結果、喫煙歴のある者は、肺がん患者では 82 名、健常者では 60 名であった。また、食事調査で、「毎日必ず朝食をとる」者は肺がん患者で 74 名、健常者で 80 名であった。喫煙歴、朝食と肺がんの関係についてそれぞれクロス表を作成してください。

２－３　図示法

表 4 を図に描くと、その特徴が一層明らかになる。他の人に直観的に理解させるためには図による表現がよい。

① 棒グラフ（過去問）と折れ線グラフ（図 14）
　最も一般的な図示法である。作成に留意する点は、
　（１） 縦軸、横軸の変数（単位）を明示すること
　（２） 図代を明示し、出典、発表年を明記すること
　（３） 表題は図では下に書くのが正しい

② ヒストグラム（頻度分布図）（図 15）

身長、体重、年齢など連続的変数による度数分布表を図に描くときは、ヒストグラムが用いられる。図 15 より、多少の凹凸はあっても左右対称であること、すなわち、両すそ野は低く、中央が高くなっていることが容易に分かる。

③ 度数折れ線（度数多角形）

これは、ヒストグラムの長方形の柱の上辺の中点を直線で結んだ折れ線グラフ。図 16 は 1977 年の死亡者約 69 万人のうち、糖尿病との診断名で記載されたもの全てを抽出し、年齢別の百分率分布を示したもの。図 16 をみると、男女の年齢別の分布に違いのあることが分かる。

④ 円グラフ（扇形図表）

いくつかの項目について、相対的な大きさの比較を角度の大きさによって比較するものです。12 時の位置から大きい順に時計の針の動く方向に描く。文字の大きさは一定の方向に書くと見やすく、また、同心円のなかには対象者・調査数などを記入する。図 17 は、調査時の健康状態に関する回答の一例。一見して、数量的な違いが把握できる。

⑤ 帯グラフ（帯図表）

これは、円グラフと同様に、いくつかの項目についてその相対的な大きさを比較するときに用いる。帯グラフは、いくつかを並べて年次的な変化を見るときに便利である。図 18 は死因群別死亡割合の年次比較をしたもの。

⑥ レーダーチャート

各項目間で比較して、一定の傾向が一目で見られるようにしたのが、図 19 に示したレーダーチャートによる表現法である。図 19 は本学の実施した授業評価で一般学生と長期履修学生との違いを比較したものです。

⑦ 地図による表現

図 20 はポリオ（小児麻痺）の発生状況です。これを見るとサハラ以南のアフリカ諸国、インド、パキスタンなど南西アジアにポリオが分布していることがわかる。なお、現在ではワクチン接種によってポリオの発生はほとんどなくなっています。

⑧ 散布図

散布図は 2 つの変数の関連をみるときによく用いられる。図 21 は萩市の明神池で観測した密度の時間変動です。水面下 0.5m と水面下 3.0m での違いを示しています。

図示するときの注意点：

① 棒グラフは扱う資料が離散型の変数なので、0 と 1、1 と 2 の間をあけて描く。

② ヒストグラムで階級の幅とグラフの目盛りは必ずその比を一定とする。

③ 棒グラフで長方形を途中でカットし、上辺に値を記入してあるものは、図は直感的な大きさの相違を見るものなので、利用者が間違うおそれがある。この場合、むしろ表だけのほうがよい。

図 14　棒グラフと折れ線グラフの一例

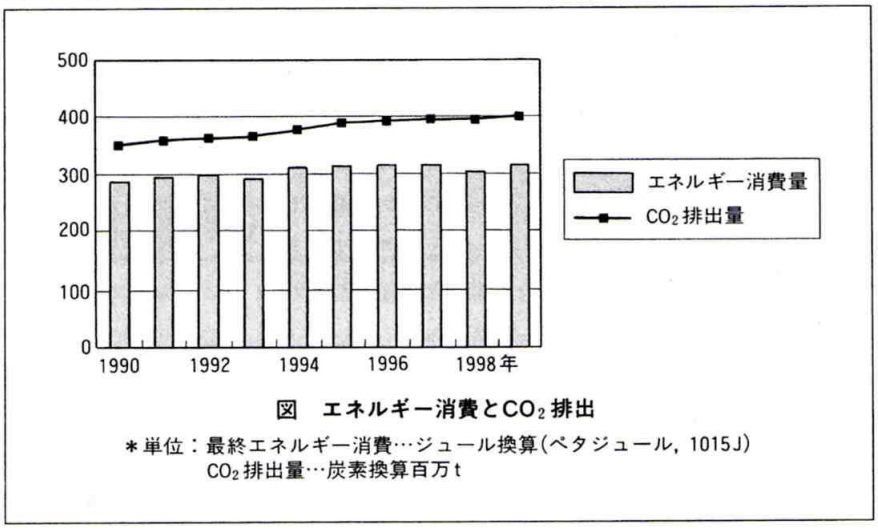


図　エネルギー消費とCO_2排出

＊単位：最終エネルギー消費…ジュール換算(ペタジュール，1015J)
CO_2排出量…炭素換算百万t

図 16　糖尿病有病死亡者の百分率分布(方波見ら)

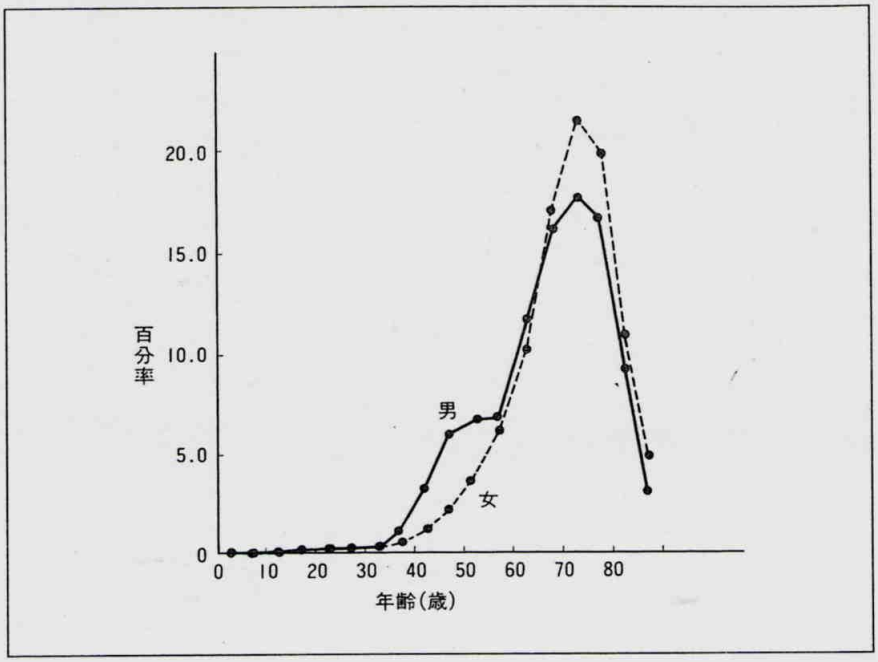


図 17　調査時の健康状態(未受診者)

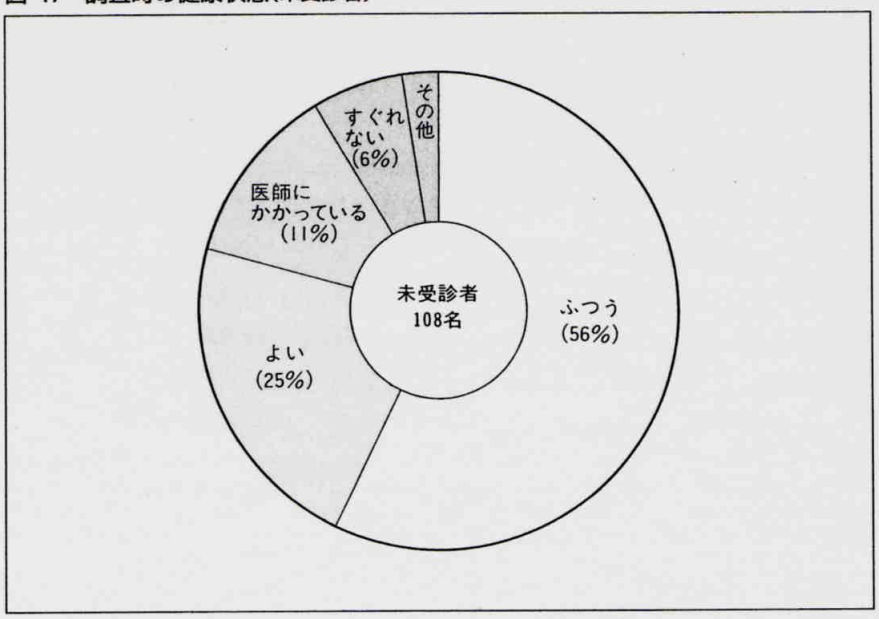


図 18　死因群別死亡割合の年次比較

図 20　1998 年における世界のポリオ発生状況(WHO 資料による)

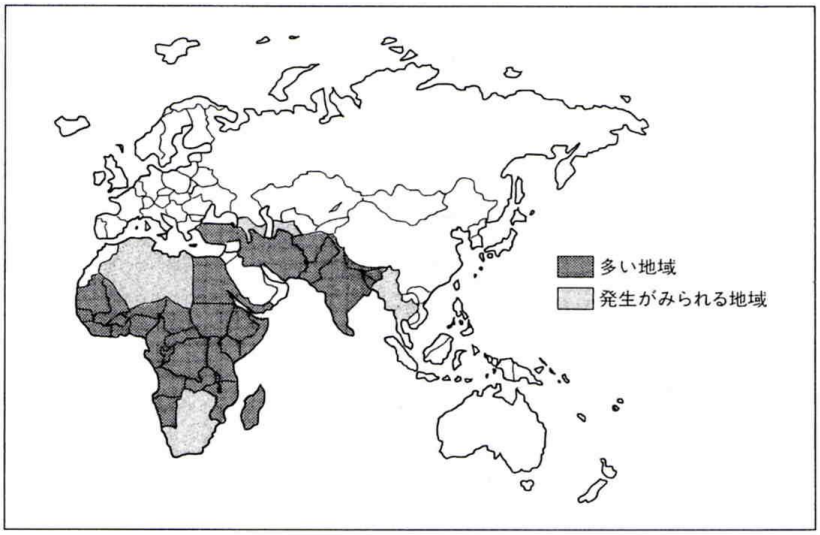
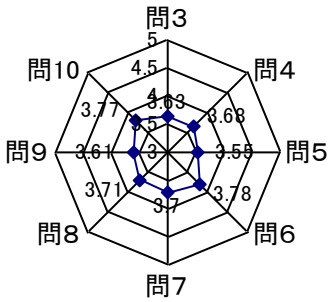


図 19　レーダーチャートの一例（一般学生と長期履修学生との授業評価の違い（平成 17 年度））

一般学生（前期）



長期履修学生（前期）

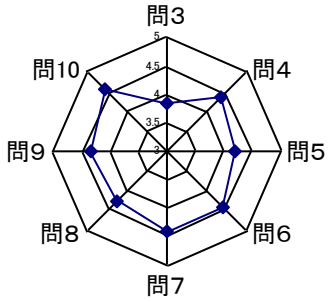
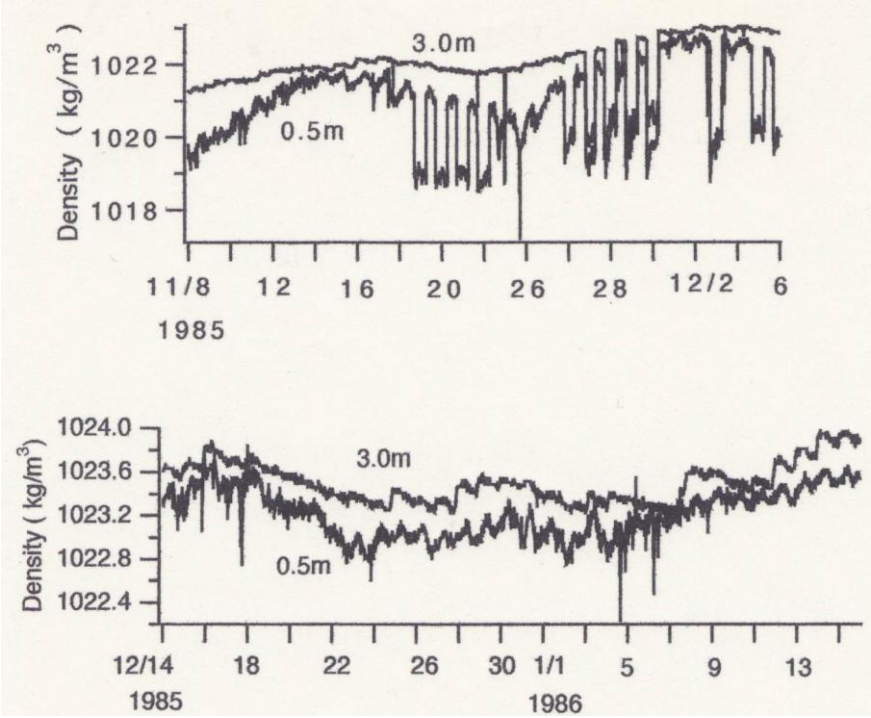


図 21　密度の時間変化（明神池 1985 年～1986 年）

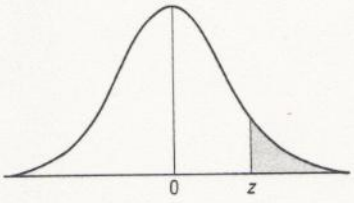


付表 1　ギリシア文字

大文字	小文字	読み方
A	α	アルファ
B	β	ベータ，ビータ
Γ	γ	ガンマ
Δ	δ	デルタ
E	ε	エプシロン
Z	ζ	ゼータ
H	η	エータ，イータ
Θ	θ	テータ，シータ
I	ι	イオータ
K	κ	カッパ
Λ	λ	ラムダ
M	μ	ミュー
N	ν	ニュー
Ξ	ξ	クシー，クサイ
O	o	オミクロン
Π	π	ピー，パイ
P	ρ	ロー
Σ	σ, ς	シグマ
T	τ	タウ
Y	υ	ユプシロン
Φ	ϕ, φ	フィー，ファイ
X	χ	キー，カイ
Ψ	ψ	プシー，プサイ
Ω	ω	オメガ

付表 4　正規分布表

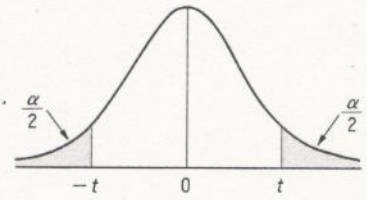
zから外側(色の部分)の面積である。
zが負の場合も，対称性から同面積である。



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
3.5	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002

付表 5　t 分布表

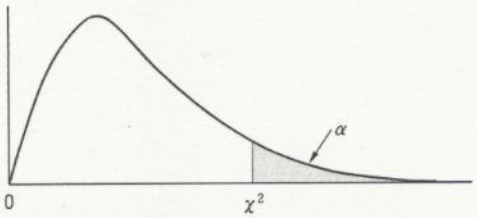
t，$-t$ の外側（色の部分）の面積を加えて α となるが，そのときの自由度（ν）と t の値である。



ν \ α	.30	.20	.10	.05	.02	.01	.001
1	1.963	3.078	6.314	12.706	31.821	63.657	636.619
2	1.386	1.886	2.920	4.303	6.965	9.925	31.598
3	1.250	1.638	2.353	3.182	4.541	5.841	12.941
4	1.190	1.533	2.132	2.776	3.747	4.604	8.610
5	1.156	1.476	2.015	2.571	3.365	4.032	6.859
6	1.134	1.440	1.943	2.447	3.143	3.707	5.959
7	1.119	1.415	1.895	2.365	2.998	3.499	5.405
8	1.108	1.397	1.860	2.306	2.896	3.355	5.041
9	1.100	1.383	1.833	2.262	2.821	3.250	4.781
10	1.093	1.372	1.812	2.228	2.764	3.169	4.587
11	1.088	1.363	1.796	2.201	2.718	3.106	4.437
12	1.083	1.356	1.782	2.179	2.681	3.055	4.318
13	1.079	1.350	1.771	2.160	2.650	3.012	4.221
14	1.076	1.345	1.761	2.145	2.624	2.977	4.140
15	1.074	1.341	1.753	2.131	2.602	2.947	4.073
16	1.071	1.337	1.746	2.120	2.583	2.921	4.015
17	1.069	1.333	1.740	2.110	2.567	2.898	3.965
18	1.067	1.330	1.734	2.101	2.552	2.878	3.922
19	1.066	1.328	1.729	2.093	2.539	2.861	3.883
20	1.064	1.325	1.725	2.086	2.528	2.845	3.850
21	1.063	1.323	1.721	2.080	2.518	2.831	3.819
22	1.061	1.321	1.717	2.074	2.508	2.819	3.792
23	1.060	1.319	1.714	2.069	2.500	2.807	3.767
24	1.059	1.318	1.711	2.064	2.492	2.797	3.745
25	1.058	1.316	1.708	2.060	2.485	2.787	3.725
26	1.058	1.315	1.706	2.056	2.479	2.779	3.707
27	1.057	1.314	1.703	2.052	2.473	2.771	3.690
28	1.056	1.313	1.701	2.048	2.467	2.763	3.674
29	1.055	1.311	1.699	2.045	2.462	2.756	3.659
30	1.055	1.310	1.697	2.042	2.457	2.750	3.646
40	1.050	1.303	1.684	2.021	2.423	2.704	3.551
60	1.046	1.296	1.671	2.000	2.390	2.660	3.460
120	1.041	1.289	1.658	1.980	2.358	2.617	3.373
∞	1.036	1.282	1.645	1.960	2.326	2.576	3.291

付表 6　χ^2 分布表

χ^2 の右側(色の部分)の面積が α となるときの自由度(ν)と χ^2 の値である。



ν \ α	0.990	0.950	0.100	0.050	0.010	0.001
1	……	……	2.71	3.84	6.63	10.83
2	0.02	0.10	4.61	5.99	9.21	13.82
3	0.11	0.35	6.25	7.81	11.34	16.27
4	0.30	0.71	7.78	9.49	13.28	18.47
5	0.55	1.15	9.24	11.07	15.09	20.52
6	0.87	1.64	10.64	12.59	16.81	22.46
7	1.24	2.17	12.02	14.07	18.48	24.32
8	1.65	2.73	13.36	15.51	20.09	26.12
9	2.09	3.33	14.68	16.92	21.67	27.88
10	2.56	3.94	15.99	18.31	23.21	29.59
11	3.05	4.57	17.28	19.68	24.72	31.26
12	3.57	5.23	18.55	21.03	26.22	32.91
13	4.11	5.89	19.81	22.36	27.69	34.53
14	4.66	6.57	21.06	23.68	29.14	36.12
15	5.23	7.26	22.31	25.00	30.58	37.70
16	5.81	7.96	23.54	26.30	32.00	39.25
17	6.41	8.67	24.77	27.59	33.41	40.79
18	7.01	9.39	25.99	28.87	34.81	42.31
19	7.63	10.12	27.20	30.14	36.19	43.82
20	8.26	10.85	28.41	31.41	37.57	45.31
21	8.90	11.59	29.62	32.67	38.93	46.80
22	9.54	12.34	30.81	33.92	40.29	48.27
23	10.20	13.09	32.01	35.17	41.64	49.73
24	10.86	13.85	33.20	36.42	42.98	51.18
25	11.52	14.61	34.38	37.65	44.31	52.62
26	12.20	15.38	35.56	38.89	45.64	54.05
27	12.88	16.15	36.74	40.11	46.96	55.48
28	13.56	16.93	37.92	41.34	48.28	56.89
29	14.26	17.71	39.09	42.56	49.59	58.30
30	14.95	18.49	40.26	43.77	50.89	59.70
40	22.16	26.51	51.80	55.76	63.69	73.40
50	29.71	34.76	63.17	67.50	76.15	86.66
60	37.48	43.19	74.40	79.08	88.38	99.61
70	45.44	51.74	85.53	90.53	100.42	112.32
80	53.54	60.39	96.58	101.88	112.33	124.84
90	61.75	69.13	107.56	113.14	124.12	137.21
100	70.06	77.93	118.50	124.34	135.81	149.45

２－４　集団を表す代表値（平均、分散、標準偏差など）

集団を表す代表的な数値を特性値という。

１）平均

標本からの算術平均はχの文字の上に―をつけて（下記のように）「エックス・バー」と読む。

また、母集団からの平均はμ（ギリシア文字）で「ミュー」と読む

今、A,B,C 3 人の大学生の体重 65kg、80kg、73kg のとき、

$$\bar{x}=\frac{65+80+73}{3}=72.7$$

一般に n 人の体重が測定され、それぞれ χ_1、χ_2、χ_3、…..、χ_n とすると、

$\bar{x}$ は

$$\bar{x}=\frac{x_1+x_2+x_3+\cdots\cdots\cdots\cdots+x_n}{n}=\frac{1}{n}\sum_{i=1}^{n}x_i$$

Σはギリシア文字シグマの大文字で$\Sigma\ \chi_i$ は χ_i がとりうるすべての値を加えることを意味しています。χ_i は i が 1 から n までの間の任意の整数をとる場合の値をさすので、

$$\sum_{i=1}^{n}\chi_i$$

は χ_i がとりうるすべての値を加算する（$\chi_1+\chi_2+\chi_3+....+\ \chi_n$）とう意味となります。

平均値の補遺

平均値には算術平均（または相加平均)、幾何平均（または相乗平均）（**過去問**)、調和平均、重みづけ平均（後述）などがある。

算術平均はデータの代表値として直感的にわかりやすく、数学的な取り扱いに適しているので、統計学で平均と言った場合には算術平均のことを示している。平均値を表示するには有効数字を考える。例えば、元のデータが整数の場合には小数点１桁で表示。

幾何平均は割合の平均値を求めたいとき、調和平均は逆数に意味のある変数のときに用いられることがある。今、変数を a、b、c とすると、

算術平均＝（a＋b＋c）÷3、幾何平均＝（a＋b＋c）$^{1/3}$、調和平均＝3/（1/a＋1/b＋1/c）となる。また、算術平均≧幾何平均≧調和平均の関係がある。

２）分散と標準偏差（過去問）（s^2、s）

度数分布の項で表 4 に示したように、集団の中の個々の値はすべて異なっています。しかも、集団の特性によってその大きさは異なるが、なんらかの変動（**ばらつき**）がみられます。その変動は平均からの隔たりの大きさ（偏差)、言い換えると平均の周囲に標本が密集する程度によってあらわすことができます。

ここで、「ばらつき」とは、集団の中の個々の数値が、一定の基準から離れてその周辺に不規則にちらばって存在（分布）することを意味しています。すなわち、平均値からの隔たりの大きさ「偏差」となります。この「ばらつき」の大きさを示すものとして、分散（variance)、標準偏差（standard deviation）

などがあります。計測値の分布の中心（平均値）からの「偏差」を二乗して足し合わせた値を「偏差平方和」と呼びます。分散はその平均値です。標準偏差は分散の平方根。計測値の分布の中心（平均値）からの平均的なゆらぎの幅を表す指標です。

$$\text{標本分散} \quad ; \quad s^2 = \frac{\text{偏差平方和}}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$$

$$\text{母分散} \quad ; \quad \sigma^2 = \frac{\text{偏差平方和}}{n} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

$$\text{標本標準偏差} \quad ; \quad s = \sqrt{\text{標本分散}} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$$

$$\text{母標準偏差} \quad ; \quad \sigma = \sqrt{\text{母分散}} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}}$$

＊ 一般には、n－1 ではなく n で除する。しかし、n で除するのはデータが母集団全部の場合なので、ここでは、n－1 を採用する。

<u>問題 3</u>

今、大学生 5 人の体重を 65kg、80kg、73kg、63kg、70kg とすると、その平均値、偏差平方和、標本分散および標本標準偏差を式で表してください。

標本集団の特性を表すのに、「(平均値) ± (標準偏差)」という形で表現すると、平均値と標準偏差が一目で見られる。

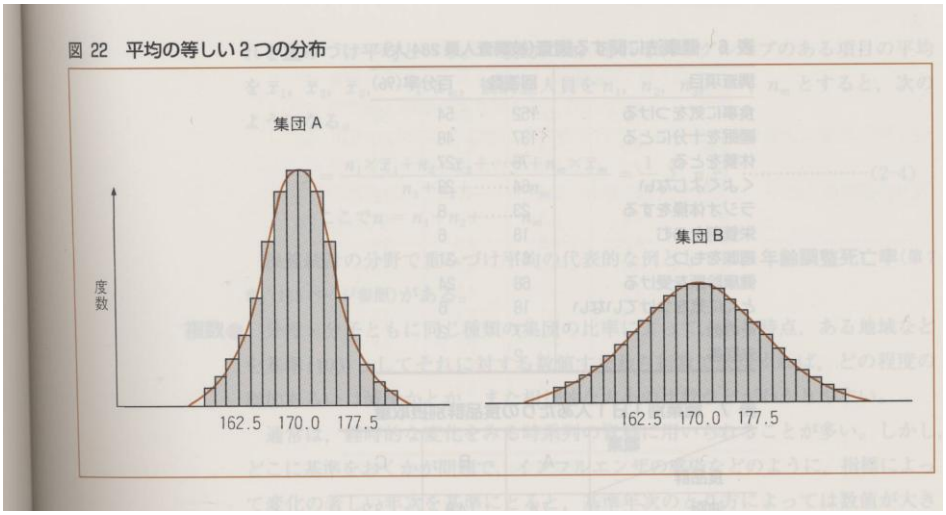


図 22 の場合、集団 A（500 人とする）「170.5±7.0」cm、集団 B（500 人とする）「170.5±15.0」cm 同じ平均値であるが、A は B よりもばらつきが小さい、A は B よりも標準偏差が小さい、A の個々の測定値が B よりも平均値に近いことを示しています。

３）その他の数値

（１）百分率（%）；

ある調査で A、B 両地区の結核住民検診の受診者数が

A 地区では、対象者　1,384 人のうち 98 人

B 地区では、対象者　8,011 人のうち 436 人　であった。

問題４

両者の百分率を式で表してください。

このように A 地区　7.1%、B 地区　5.4%となって、両者の相対的な大きさがわかります。

表 6　健康法に関する調査（被調査人員 284 人）

調査項目	回答数	百分率(%)
食事に気をつける	152	54
睡眠を十分にとる	137	48
休養をとる	76	27
くよくよしない	64	23
ラジオ体操をする	23	8
栄養剤をのむ	18	6
趣味をもつ	61	21
健康診断を受ける	68	24
とくに気をつけていない	18	6
その他	7	2
未回答	2	

表 7　職業別１日１人あたりの食品群別摂取量

職業 / 食品群	A	B	C
油脂	7.4	4.8	3.2
動物性タンパク	259.3	189.0	188.3
果実	113.2	83.4	96.7
被調査人員	300	100	50

表 6 は健康法に関する調査結果です。回答数を百分率で示すとわかりやすい。

（２）重みづけ平均；

表 7 はある地域の集団での A、B、C それぞれの職業における１日１人あたりの食品群別摂取量です。表 7 の条件の場合、この地域全体としての１日１人あたりの摂取量を知るには、どのようにしたらよいですか？たとえば、油脂摂取量について求めると、

$$\bar{x}=\frac{7.4+4.8+3.2}{3}=5.1$$

しかし、被調査員がそれぞれの職業について同数の場合はこの計算でよいが、この場合には異なるので

$$\bar{x} = \frac{7.4\times300 + 4.8\times100 + 3.2\times50}{300 + 100 + 50} = 6.4$$

A、B、Cの職業の摂取量とそれぞれの被調査人員の数とが平均値に影響します。これが、この場合には正確な平均を意味します。これを重みづけ平均といいます。一般には、それぞれのグループのある項目の平均を

$\bar{x}_1$、 $\bar{x}_2$、 $\bar{x}_3$、....， $\bar{x}_m$ 、 被調査員を $n_1,n_2,n_3...,n_m$ とすると、

$$\bar{x} = \frac{n_1\times\bar{x}_1 + n_2\times\bar{x}_2..... + n_m\times\bar{x}_m}{n_1 + n_2 +.....+ n_m} = \frac{1}{n}\sum_{j=1}^{m} n_j\bar{x}_j$$

ここで、 $n=n_1+n_2+.....+n_m$

問題５

表 7 で動物性タンパク、果実について、それぞれ重みづけ平均を式で表してください。

（３）中央値（過去問）（メディアン、median、Ｍｅ）；

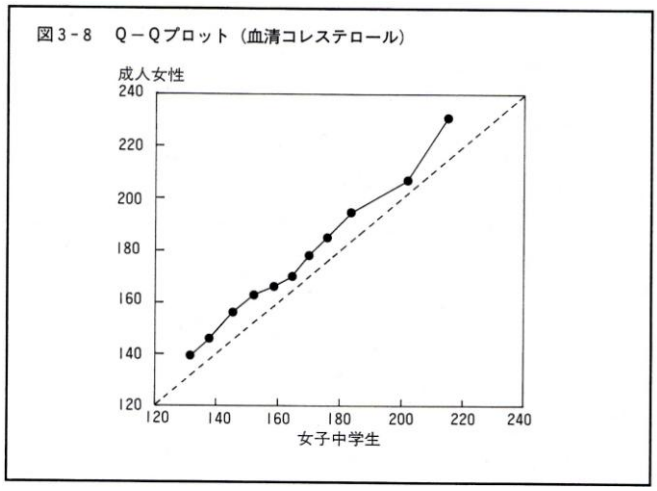
資料を大きさの順に並べたときの中央の測定値で標本数（n）が奇数のときは、（n+1）/2 番目の測定値、n が偶数であれば、（n/2）番目と（n/2）+1 番目の測定値の和を 2 で割った値。

問題６

資料が２,5,6,9,12,1,3 のとき、中央値はどれですか、計算式を立てて求めてください。

問題７

資料が 2,5,8,6,8,10,15,8 のとき、中央値はどれですか、計算式を立てて求めてください。



（４）モード（mode、最頻値）；

最も度数の多い値を代表させる。

（５）平均偏差（過去問）（ＭＡＤ：mean absolute deviation、ＭＤ：mean deviation）；

$$\text{平均偏差} = \frac{1}{n}\sum_{i=1}^{n}\left| x_i - \bar{x} \right|$$

（６）変動係数（過去問）（ＣＶ：coefficient of variation　）

長さの単位で標準偏差が 5cm といっても、平均値が 100cm のものと 50cm のものとでは意味が異なる。このように、平均値、標準偏差がともに変化するとき、その変動を変動係数で表すことがある。

$$CV = \frac{s}{\bar{x}} \times 100$$

ただし、平均値がゼロに近いときには変動係数は使うべきではない。

（７）四分位数、四分位偏差 IQR（過去問）、パーセンタイル値（百分位数）（過去問）、範囲（過去問）

数字の小さい方から数えて全体の 10%に当たる値を 10 パーセンタイル値、全体の 90%に相当する値を 90 パーセンタイル値という。

データの度数を 4 つ分するときの値のことを「四分位数」という。小さい方から順に、第 1 四分位数、第 2 四分位数（中央値でもある）、第 3 四分位数となる。

データの分布する幅が範囲（range）で範囲＝最大値－最小値となる。範囲は、すべてのデータを含むので分布の左右のすそ野、一般にはデータの少ない部分の影響を受けてしまう。そこで、中央値の前後で全体のデータ数の 50%を含む幅を使うことがある。中央値を挟んで 50%とは、第 1 四分位数と第 3 四分位数の間であり、この幅を四分位範囲（interquartile range）と呼ぶ。

四分位範囲＝（3 四分位数）－（第 1 四分位数）

また、中央値からの第 1、第 3 四分位数の偏差を求めて平均した四分位偏差（quartile deviation）も散布度の指標となる。

四分位偏差＝｛(3 四分位数）－（第 1 四分位数)｝÷2

Q－Q プロット、P－P プロット

パーセンタイル値を用いて２つの分布の形を比較する、もしくはある分布が確率分布のどれかにどれほど一致しているかを知る方法として、確率プロットの１つの方法は、２つの分布で同じパーセントに対するパーセンタイル値をそれぞれ求め、XY 平面にプロットしていく Q－Q プロットである。

また、Q－Q プロットは異なる変数どうしで作成することもできる。例えば大学入試の理解や社会科で選択科目があった場合、選択科目 A の８０点が、選択科目 B の７７点に相当するといったように、２つの異なる科目の得点を等化（equating）する場合に、利用可能である。Q－Q プロットとは逆に同じ変数値に対するパーセンタイルをそれぞれ求めて XY 軸に作図したものを P－P プロットという。

また、データの散らばり方を比較するために平均値と標準偏差からは変動係数を導いたが、これと同様に中央値と四分位偏差から単位や値の異なる分布の散らばり方を比較するための係数を求めることができる。それが四分位偏差係数（coefficient of quartile deviation）で、

四分位偏差係数＝四分位偏差÷中央値×100

で求まる。また、ヒストグラムに対して、箱ヒゲ図がある。図 3－14 は、コレステロールのデータを箱ヒゲ図に作成したもので、箱ヒゲ図の作り方は、まず各四分位数を求め、第 1 四分位と第 3 四分位の間に箱を描く。その箱に中央値のところに線を引く。箱の左右に延びている直線がヒゲであるが、これは四分位偏差の 1.5 倍の長さを限度とし、実際にはその範囲の内側の点までを結んでいる。ヒゲより外側にあるデータは、1 個ずつ丸印で示す。

単独の分布にも使うが、箱ヒゲ図が最も有効なのは図 3－15 のように複数の分布を同時に観察したい場合である。この図は、小学 4 年生から中学 3 年生までの男児の身長の図であるが、身長の伸びの様子と、中学 1、2 年でばらつきが最大になることがわかる。

図 3－14　箱ヒゲ図

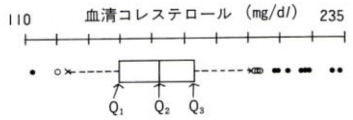


図 3－15　多層の箱ヒゲ図（小 4～中 3　男児の身長）

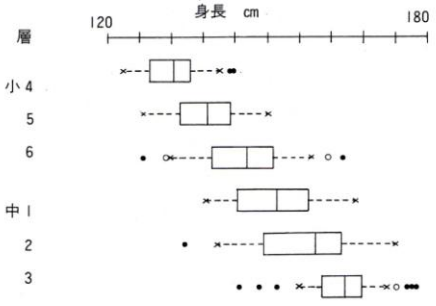


図 3－9　Q－Q プロット（架空例）

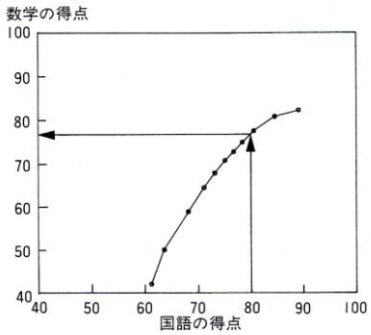
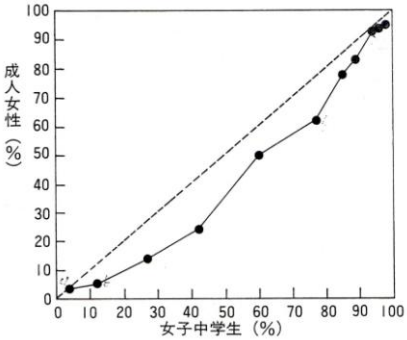


図 3－10　コレステロール値の P－P プロット



問題 8

1mの規格で作られた鉄棒の長さの平均 100cm、標準偏差 3.2cm、6 個のボールベアリングの長さの平均 2cm、標準偏差 0.029cm のとき、それぞれの変動係数を求めて、どちらがばらつきが小さいか判断してください。

3　正規分布と推定

3－1　正規分布

1）正規分布とは（過去問）

生物現象など自然界で観察される多くの計測値は、何であれ平均値に近いほどその出現率が高く、平均値からその両側に値が遠ざかるにしたがって出現頻度は少なくなる。

このうち、同じものを何度も繰り返し計測し、平均値からのずれ（誤差）の大きさを求め、その出現度数を描いてみると、平均値を中心として左右対称の釣鐘状の分布型になることが多い。

1812 年に数学者ガウスは、この純粋な条件で繰り返し計測したときに、一貫して現れる分布型を発見し、それを正規分布（normal distribution）と名付けた。発見者の名前をとってガウス分布（Gaussian distribution）ともいわれる。

この曲線は誤差曲線とも呼ばれます。その理由は、ある寸法を目標にして何かを作るとき、ちょっとした手のはずみなどで、目標の寸法よりもわずかばかり大きくなってしまったり、逆に小さくなってしまったりする‘誤差’が生じます。この誤差の大きさは、正規分布に従うことが知られているからです。つまり、物を作るときには、人によって、あるいは機械によって、目標より平均して大きめの物を作ったり、あるいは小さめの物を作ったりする‘くせ’があります。この誤差の平均値は 0 でないのが普通ですが、誤差の大きさは、その平均値を中心にして左右対称な正規分布にしたがいます。

正規分布は、その分布に従うあるグループの平均値と標準偏差が分かっていれば、その分布に関する全てが分かります。例えば、

正規分布曲線下の面積は、－∞～＋∞　で　1

$\mu-\sigma$ と $\mu+\sigma$ の間の正規分布曲線の面積は全体の約 68%

$\mu-2\sigma$ と $\mu+2\sigma$ の間の正規分布曲線の面積は全体の約 95%

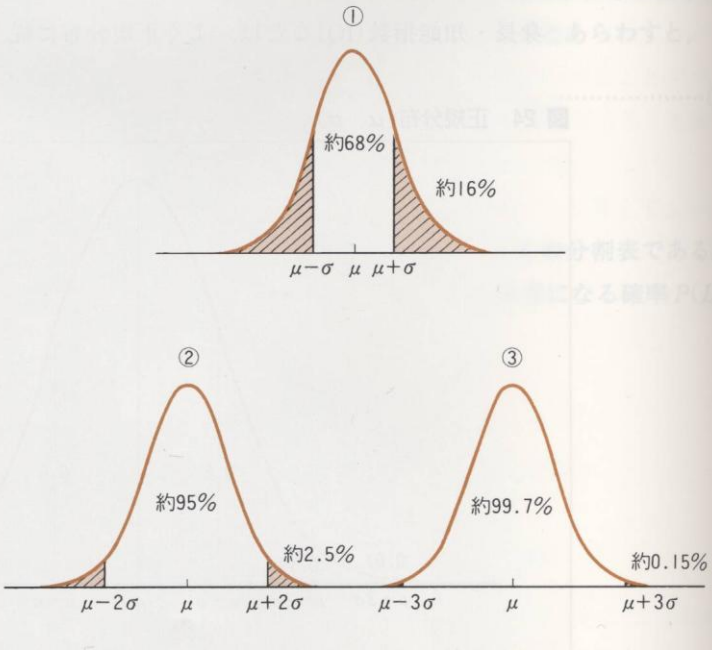
$\mu-3\sigma$ と $\mu+3\sigma$ の間の正規分布曲線の面積は全体の約 99.7%

$\mu-4\sigma$ と $\mu+4\sigma$ の間の正規分布曲線の面積は全体の約 99.99%

という具合となります。「平均値が μ、標準偏差が σ である正規分布」を N（μ、σ^2）と略して記号で表す習慣があります。N は normal distribution の頭文字です。

例）東京オリンピックで優勝した日本女子バレーボールチームの平均身長は 171cm、この当時の女子の平均身長（μ）を 156cm、標準偏差（σ）を 5cm と仮定すると、図 25 から、171cm 以上は③図の $\mu+3\sigma$ 以上にあたる。すなわち約 0.15%である。わが国の昭和 20 年代に約 200 万人の出生数があるのですが、女子がその約半分とすると 100 万人です。したがって、171cm 以上は 1,500 人しかいないことになります。その中から運動神経もある程度発達していて訓練に耐えることができ、なおかつバレーボールが好きな人をさがすのは、大変であったことがわかります。

図 25　正規分布曲線下の面積と標準偏差との関係

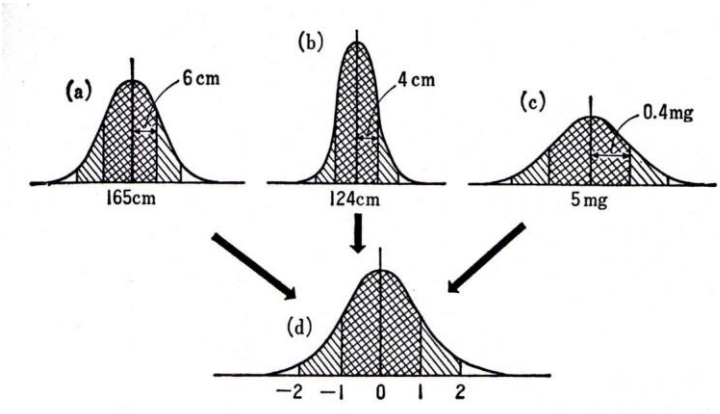


2）正規分布の標準化

先程の例で平均値よりも大きく、175 ㎝よりも小さい人は何%でしょう。これに答えるには、日本の当時の女子の身長について、細かい数表を作っておく必要があります。また、平均値や標準偏差は、各測定値によって単位が異なっています。例えば、身長は cm、体重は kg、ヘモグロビン濃度は mg/dl などです。これらについて、片っ端から数表を作ることが必要になります。しかし、実際にそのような作業をすることは現実的ではありません。そこで、正規分布に従うものならどんなものにでも適用できる数表があれば便利ですよね。それでは、そのような便利な数表を作るにはどうしたらよいのでしょうか。

それには「正規分布は平均値と標準偏差によって決まる」という性質を利用するのです。つまり、平均値を固定して、標準偏差をものさしにしてばらつきの大きさを表してやれば、数表は 1 つで済むことになります。

次の図を見てください。



（a）は青年男子の身長で平均値が 165 ㎝、標準偏差は 6 ㎝、（b）は小学 3 年生の身長で平均値は 124 ㎝、標準偏差は 4 ㎝、（c）は 1 錠の胃腸薬に含まれるパントテン酸カルシウムの量で平均値は 5 ㎎、標準偏差は 0.4 ㎎です。この 3 つの分布は、どれも正規分布なのですが、平均値も違うし標準偏差も異なります。単位も異なります。しかし、ともに正規分布である共通点を利用して、1 つの数表が使えるように工夫できそうです。正規分布であるという共通の点は、平均値の両側に標準偏差だけの幅をとると、その幅の中の面積（図で 2 重斜線の部分）は 0.6826 です。というように、平均値から標準偏差を単位としてある幅をとると、その範囲の面積がどんな正規分布の場合にも等しい値になるのです。そこで、（d）のような正規分布を考えます。この正規分布は平均が 0、標準偏差が 1 です。つまり N（0、1^2）です。この正規分布と他の 3 つの正規分布と比べてみると、例えば、（a）では 165 ㎝のところを 0 とみなし、横軸の目盛を 6 ㎝を単位（1 とする）にして書き直すと、（d）と全く同じになります。つまり、165 ㎝のところが 0 になり、171 ㎝のところが 1 になり、177 ㎝のところが 2 に、159 ㎝のところが－1 になるわけです。

したがって、（d）の正規分布 N（0、1^2）についての詳しい数表があれば、（a）の図形のどの部分の面積もわかることになります。（b）の場合も（c）の場合も全く同じことです。

このように平均値が 0、標準偏差が 1 になるように統計量を考えて、各測定値が平均 0、標準偏差 1 になるような正規分布を作成すると、各測定値の分布上の位置が、比較できて便利です。

問題 9

（a）の分布で 180 ㎝は（d）の分布に置き換えると、どのような値になるのですか、答えてください。

このような平均 0、標準偏差が 1 の正規分布を規準正規分布あるいは、標準正規分布といいます。

z を使って、

$$z = \frac{x - \mu}{\sigma}$$

このようにすると、各測定値の単位に関係なく分布上の位置を示したり、検定に利用できます。

z 分布表を使った例）

今、μ＝156ｃｍ、σ＝5ｃｍの身長分布で 169ｃｍ以上の人の割合を求めたいとします。

この場合まず、169ｃｍに対応する z を求めます。

z＝（169－156）/5＝2.6

付表 4 の正規分布表より、z＝2.6 より右側の面積は 0.0047、したがって、割合は 0.47%となります。また、同じμ、σの集団で 150ｃｍ以下の割合は、

z＝（150－156）/5＝－1.2

付表 4 より 0.1151 となります。したがって、割合は 11.51%となります。

＊ 注意；正規分布は左右対称なので、z が負でも面積は±1.2 と同じことなります。

（この部分に、付表の見方を記した部分を記載してください。）

3）偏差値から正規化を考えてみる！

偏差値とは、ある集団におけるある個体のある種の測定値 x_i を次のような式で標準化した値 y_i です。

$$\frac{y_i - 50}{10} = \frac{x_i - \bar{x}}{s}$$

$$y_i = \frac{10}{s}(x_i - \bar{x}) + 50$$

ここで、 s は標準偏差。偏差値 y_i はもとの測定値 x_i を平均 50、標準偏差 10 になるように標準化したものです。これから考えると基準正規分布は、

$$\frac{z - 0}{1} = \frac{x_i - \bar{x}}{s} \rightarrow \frac{x - \mu}{\sigma}$$

とおけるのです。つまり、みなさんはすでに、正規分の基準化よりも難しい計算を使っているのです。

4）標本平均の分布（中心極限定理）

身長が正規分布に従うことは以前に述べたとおりですが､その身長の集団から n 人を選び出し(抽出)、その標本平均を求めます。この作業を繰り返し行うと、標本平均の集団が出来上がりますが、その集団はどんな集団になるのでしょうか。この答えは重要な定理となっているので、以下に紹介します。

平均値 μ 、分散 σ^2 の任意の分布型（どんな分布でもよい）をした母集団から、大きさ n の標本 χ_1、χ_2、 χ_3、.....、 χ_n を選んだとき、標本平均

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots\cdots\cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

の分布は、n が大きくなると正規分布 N（μ 、 σ^2/n）に近づく。

この定理は、後述の区間推定や検定で使うので重要です。

問題 10

この正規分布の標準偏差（標準誤差）について答えてください。

ここで、もう一度、平均値、偏差、偏差平方和、分散、標準偏差および標準誤差の関係を整理しておいてください。

なお、標本平均の分布は n が 25 より大きいときはよく正規分布に近似し、x がどんなにゆがんでいても、n が 50 より大きいと正規分布によく近似することが知られています。

正規分布以外の確率分布

二項分布、ポアソン分布、指数関数型分布、カイ 2 乗分布（χ^2 分布）、ベイズの定理

3－2　推定

１）推定とは

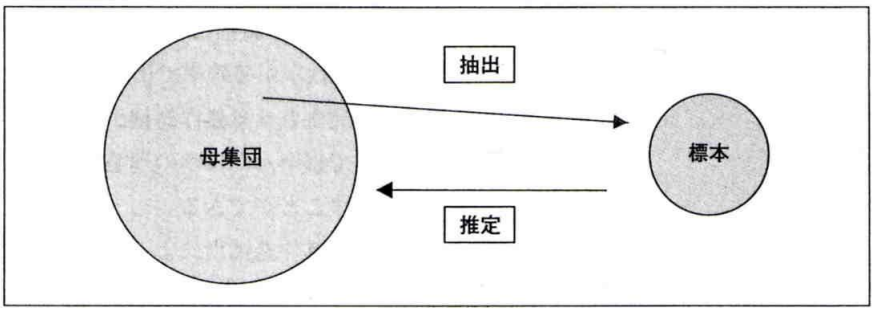
たとえば、インフルエンザにかかった子どもが脳症をおこして死亡する確率（何%とか）は、どれくらいか、知りたいことがあります。その確率はすべての子どもをインフルエンザに罹患させなければわからない。これは、実際には不可能です。

そこで、この場合には全体の一部のみから全体を知ろうとする。これが推定です。

２）母集団と標本

知りたいことは確率だけとは限らない。平均値を知りたいこともある。これらの知りたい値を**母数 parameter** といいます。母数を知ろうとする対象が**母集団 population**、そこから抽出する一部を**標本 sample** と呼びます。

図 33　母集団と標本



3）平均値μの区間推定

中心極限定理より、大きさnの標本を無作為抽出して得られる $\bar{x}$ は、平均 μ 、標準偏差 $\frac{\sigma}{\sqrt{n}}$ の正規分布に近似する。このことを用いると、図34のように、$\bar{x}$ が $\mu - 1.96\frac{\sigma}{\sqrt{n}}$ と $\mu + 1.96\frac{\sigma}{\sqrt{n}}$ との間に入る確率は95％となる。これを式で表すと $P\left\{\mu - 1.96\frac{\sigma}{\sqrt{n}} < \bar{x} < \mu + 1.96\frac{\sigma}{\sqrt{n}}\right\} = 0.95$ なる。この式を変形して $\bar{x}$ と μ を入れかえると、$P\left\{\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right\} = 0.95$ となる。

ここで σ をsで置き換えると、$\bar{x} - 1.96\frac{s}{\sqrt{n}} < \mu < \bar{x} + 1.96\frac{s}{\sqrt{n}}$

この式が μ の95％信頼区間となる。

＊正規分布の項では、$\mu \pm 2\sigma$ で約 95％としていましたが、ここではより 95％に近づけるために、
2 →1.96 を用います。

この式の意味；
標本平均を求めるため標本を抽出する作業を何回も繰り返して、そのたびごとに信頼区間をつくると、100 回のうち 95 回はその区間に母平均（μ）が含まれているということです。
けっして、1 回だけ求めた信頼区間に母平均（μ）が入る確率が 95％という意味ではありません。

問題 11
上記で、標準偏差が s/√n になっているのはなぜですか、答えてください。

図 34　標準平均 $\bar{x}$ の分布

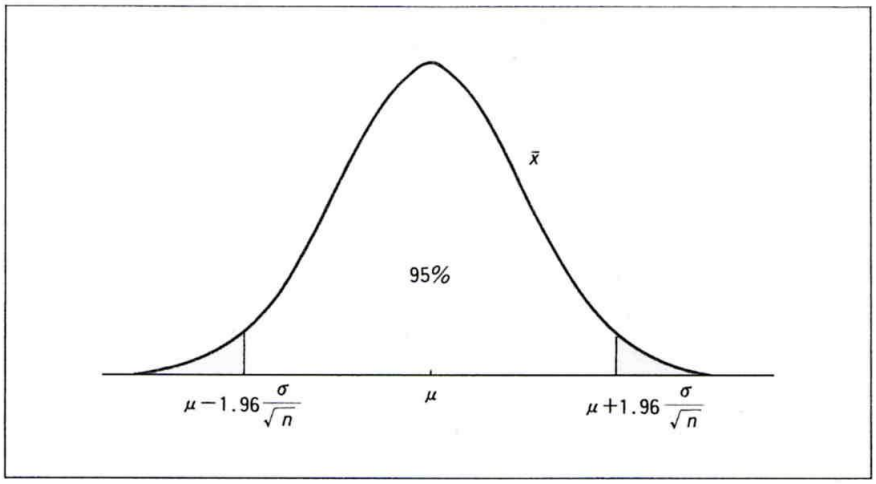


図 34 の標準平均と記載されていますが、これは誤りで、正解は標本平均です。

4）t 分布（スチューデント分布ともいう）による区間推定

先ほど、μ の区間推定で標準偏差 σ のかわりに標本推定値 s を使用しました。これは、標本が大きいからです。標本の大きさ n が 25 より大きいときは s を使用して差し支えありません。

しかし、標本の大きさが5とか10の時には、σ をsで置き換えることはできない。このときには、t分布を用いる。t分布は平均が0の正規分布に似た左右対称の分布で、自由度（ν ＝n－1）によって曲線が異なる。

$$t = \frac{\bar{x}-\mu}{\frac{s}{\sqrt{n}}}$$

自由度は独立な変数の個数で、ここでは調査対象をnとしたとき、n−1で与えられる。

図35は基準正規分布とt分布（自由度　ν ＝5　）の曲線を描いたものです。t分布で左右の面積が5％となるtの値は2.571である。nが大きくなるにつれて、t分布の曲線は次第に正規分布に近づき、t＝2.571は基準正規分布で左右の面積が5％となるzの値1.96に近づく。

図35から、ν ＝5の場合、μ の95％信頼区間は、

$$\bar{x}-2.571\frac{s}{\sqrt{n}}<\mu<\bar{x}+2.571\frac{s}{\sqrt{n}}$$ で

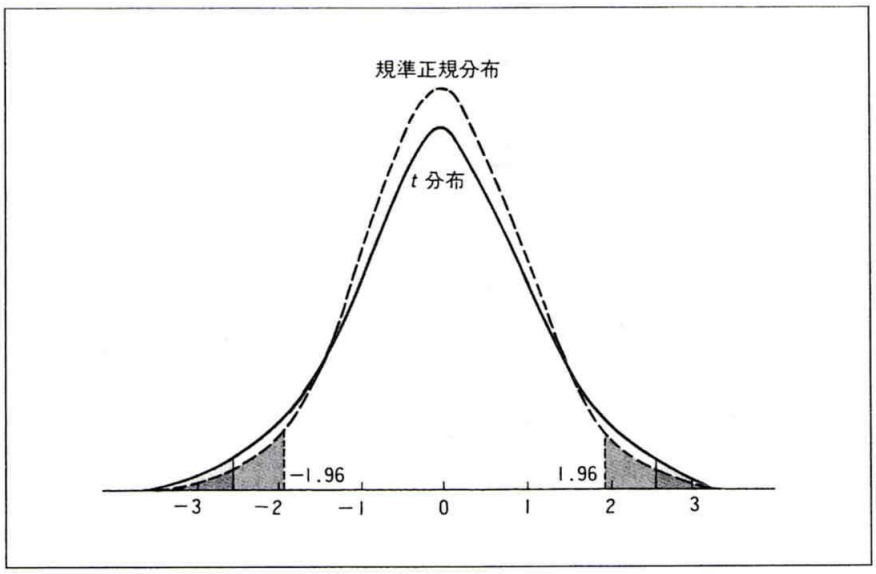
一般的な式は $\bar{x}-t_0\frac{s}{\sqrt{n}}<\mu<\bar{x}+t_0\frac{s}{\sqrt{n}}$ である。ここで、t_0 は自由度 ν と目的とする信頼区間によって決まるもので、付表5のt分布表から決定される。

ν ＝5、信頼区間が95％では　t_0＝2.571、　ν ＝10、信頼区間が95％では　t_0＝2.228

ν ＝5、信頼区間が99％では　t_0＝4.032、　ν ＝10、信頼区間が99％では　t_0＝3.169

ここで、付表 5 の t 分布表について説明します。

図 35　規準正規分布と t 分布($\nu=5$)



自由度が 1 の t 分布はコーシー分布とも言われる。

問題 12

今、10 人の男子学生の身長が、166、167、170、175、173、169、177、171、175、178（ｃｍ）とすると、標本平均は 172.1ｃｍ、標本標準偏差は 4.15 ㎝になります。このときに、ｔ分布を使って μ に対する 95%の信頼区間を求めください。

自由度の説明：

今、正規分布から取り出された 2 つのデータがあるとします。データが 2 つあれば、ｔ分布を利用して母平均 μ の区間推定ができます。区間推定をするには、まず 2 つのデータから標本平均を求め、それを使って標本標準偏差ｓを計算するのが第一段階です。しかし、よく考えてみると、ｓを計算するに際して、本当の平均値 μ がわからないので、データから架空の平均値を作り出して、その平均値で後の計

算をやっています。本来なら、標本平均が母平均と同じになるように標本を選ばなければなりません。ということは、2 つの標本を選ぶ場合には、2 つめの標本は 1 つめとの平均がちょうど母平均と同じになるように選ばなければならないので、勝手に選ぶことができません。自動的に決まってしまうということです。すなわち、2 つの標本を選ぶ場合には、自由が許されるのは最初の 1 つだけで、２つめは自由が許されないのが本来の姿ということになります。そういう意味で、n が 2 のときには、自由度は 1 になります。n が 3 以上のときも同じことです。したがって自由度はν＝n－1 になるのです。

自由度という考え方は「データの数から、そのデータで作り出して使用した平均値の数を差し引いたもの」と考えておけば間違いありません。

問題 13

1977 年の 15 歳男子 118 人の座高の平均値は 88.8 ㎝、標準偏差は 3.50 ㎝であった。μ に対する 95％信頼区間を求めてください。

問題 14

上記でμ に対する 99%信頼区間を求めてください。

４　統計的検定

４－１　検定とは（過去問）

「差がある」という仮説を検定する場合、差の程度が不明なため、そのままでは検定できない。そこで考え出されたのが、その逆の「差がない」という仮説を検定して、それに何らかの矛盾が見つかれば、もとの「差がある」という仮説を採用する。逆に、明らかな矛盾がないときにはその判定を保留する。このような論法で検定を行います。

ここで、「差がない」という仮説は本来「無」に帰すべきものとして「帰無仮説」（**過去問**）（null hypothesis）と呼び、H_0 と略す。また、もとの「差がある」という仮説は、「対立仮説」（alternative hypothesis）と呼び、H_1 と略す。

仮説の設定としては、

例えば、2 つのグループ（A,B）間で「差がある」という仮説について検定する場合、まず「2 つのグループ間には差がない；A=B」とう仮説（帰無仮説）を設定し、元々証明したい「差がある；A≠B」

という仮説（対立仮説）は、一旦伏せておく。

例題）さて、とりあえず今までに何度も事例として紹介した男子学生の集団 107 人の身長データを使って、検定を行ってみます。この集団の標本平均は 170.5cm、標本標準偏差は 5.1cm です。今、この世代の身長の全国平均が 169cm であるとすると、この集団の身長の平均が全国平均と異なるかどうかを検定する。

（この続きは、自分で以下に記入してください。）

有意水準（危険率）

実際の検定には、帰無仮説を誤って捨てる確率α（第 1 種の過誤の確率で危険率とも言う）を示し、そのときの検定統計量から確率を出して、

その値がαよりも小さいときに

↓

「有意水準αで統計的に有意である」といい、帰無仮説を棄却する。

検定統計量より得られる**実際の確率をＰ値（過去問）**（P-value）と呼び、

Ｐ値がα以下となるような統計量の範囲を

棄却域という。

一般に有意水準αには 5%（0.05）、1%（0.01）が用いられる。

統計的判断における 2 種類のエラー（過誤）

第１種の過誤（α）→ 有意差があると判断した場合におこる。
本当は、帰無仮説が正しいのに、実際の確率Ｐがα以下のため、帰無仮説が違っているとして棄却する誤りを第１種過誤または、αエラーと呼ぶ。

第２種の過誤（β）→ 有意差がない（判定保留）と判断した場合におこる。
本当は帰無仮説が誤っているのに、実際の確率Ｐがαより大きいため、帰無仮説を棄却しない、すなわち、対立仮説を採用しない誤りを第２種過誤または、βエラーと呼ぶ。
その誤りを起こす確率は、データ数に依存する。

片側検定と両側検定

正規分布やｔ分布を利用する場合、片側、両側検定の区別が問題なります。

両側検定の方が検定は厳密である。

例）正規分布（付表４）を用いた片側検定で有意水準 0.05 に対するｚは 1.65 である。これは両側検定の場合の有意水準 0.1 の位置（ｚの値）に相当する。(両側検定で有意水準 0.05 に対するｚは 1.96)
したがって片側検定では、検定統計量の偏りより少なくても「統計的に有意」となるので、判定が甘くなる。

片側検定は、あらかじめ変化（差）の向きが理論的に片側にだけ起こると想定される場合に行う。例えば、降圧剤の効果を調べる実験で、投与後の血圧の上昇を想定しなくてよい場合に用いる。
両側検定は、処理効果がどちら向きの変化をもたらすかを予想できないときに用いる。通常は、変化の向きを予め予測できないとして、両側検定を用いる。

母平均の検定①

全国10歳女子の身長の平均と標準偏差は、それぞれ μ ＝140cm、σ ＝5cmの正規分布となる。

今、ある10歳のクラス25人の身長の分布は、正規分布になるものとし、その平均値 $\bar{x}$ は137cmであった。このクラスの身長は全国水準と違うと言えるか。

（1） 仮説の設定;「全国水準と違う」という仮説は、違いの程度を特定できない。そこで、その逆の「全国水準と同じ」という仮説（帰無仮説 H_0 ; μ ＝140）を採用し、もとの仮説（対立仮説 H_1 ; $\mu \neq 140$）を一旦伏せておく。

（2） 検定統計量を求める;この場合、25人の身長の平均 $\bar{x}$ ＝137cmを統計検定量とする。

（3） 確率Pを求める; H_0 が正しい場合、標本平均 $\bar{x}$ は平均値 μ 、標準偏差 $\sigma/\sqrt{n}$ の正規分布に従うことを利用する。 $z=\dfrac{\bar{x}-\mu}{\sigma/\sqrt{n}}=\dfrac{137-140}{5/\sqrt{25}}=-3$ 付表4からz＝3.0の面積は0.0013となる。

両側の面積は0.0026となる。つまり、確率Pは0.26％となる。

（4）判定; H_0 が起こる確率は0.26％である。このようなまれな現象が実際に起こったと考えるよりは H_0 が正しくないと考えるほうが妥当である。したがって、H_0 を棄却して、対立仮説 H_1 を採用する。すなわち、そのクラスの身長は全国平均と比べて差があるのは、有意水準0.1で統計的に有意であると判定する。

（ここで、付表４の見方を以下に記入してください。）

母平均の検定②

ある看護学校の学生15人の身長を計測し、平均$\bar{x}$が162.0cm、標準偏差sが5.0cmであった。20歳の全国平均は157.5cmである。この学生の集団は、有意水準5%(0.05)で、一般よりも身長が高いといえるでしょうか。

(1)仮説の設定;帰無仮説H_0; μ =157.5、対立仮説H_1; μ ≠157.5

(2)検定統計量を求める;この場合、15人の身長の平均$\bar{x}$=162.0cmを統計検定量とする。

(3)tの値を求める;この場合、t分布表から有意性の検定を行う。

$$t=\frac{\bar{x}-\mu}{s/\sqrt{n}}=\frac{162.0-157.5}{5.0/\sqrt{15}}\fallingdotseq 3.49$$

自由度 ν =n-1=15-1=14なので、付表5より面積が0.05(5%)のt値は2.145である。

(4)判定;計算で求めたt=3.49の方がt分布表より求めたt値2.145よりも大きい。したがって、帰無仮説は棄却され、この集団の平均身長は全国平均よりも高いと結論付けることができる。ただし、有意水準0.05(5%)で。

(ここで、付表5の見方を以下に記入してください。)

2 つの集団における平均値の差の検定

対応のある場合とない場合とは？

t検定には対応のあるデータに対する検定と､対応のないデータに対する検定の2種類があります。ここで､”対応のある“とか”対応のない“とはどういうことなのでしょうか？

その答えは、同一個体（人でも動物でも機械でもよい）におけるある処置の前と後の状態（前後のある測定値）を比較するのが対応のある場合となります。一方、異なる個体間である測定値を比較するのが対応のない場合となります。

例えば、手術の前後で体温が変化するかどうかを検定するため、5 人の患者の手術直前と直後の体温を測定したとします。この場合、1 人の患者に対して術前の体温と術後の体温の 2 つのデータが対になって得られます。このように 2 つのデータ集団が同一個体から得られる場合を対応があるといいます。

対応のある場合

例題）表のように、手術前後の体温の比較をしてみます。

患者	術前の体温	術後の体温	差	差の偏差	差の偏差平方
A	36.5	35.5	-1.0	-0.2	0.04
B	36.7	36.0	-0.7	0.1	0.01
C	36.4	35.4	-1.0	-0.2	0.04
D	36.5	35.0	-1.5	-0.7	0.49
E	36.2	36.4	0.2	1.0	1.00

この場合、表のように同一個体の差を求め、その平均値、偏差、偏差平方、偏差平方和、標準誤差を求める。そして、ｔ＝差の平均値/差の標準誤差を求めて、t 分布表（付表 5）のｔ値と比較する。危険率 5%で有意差があるかどうか検定してください。下記に、差の平均、差の偏差平方和について、式を立てて、求めてみてください。その後、差の標準誤差について、式立てをしてください。なお、差の標準誤差の答えは 0.281、ｔは 2.84 です。ところで、自由度はこの場合、いくつですか？

対応のある場合の一般式

先程も説明しましたが、ある集団について、冬と夏の血圧を測定して両者を比較する場合や、同一人の左右の握力を測定して比較する場合、同集団で夏と冬の食事の内容（タンパク質、脂肪、摂取エネルギーなど）を比較する場合なども、2 変数は独立ではなく、対応のあるデータとなります。この場合に用いる t を求めるための一般式と例題を下記に示します。

このような場合に2つの平均値$\bar{x}$、$\bar{y}$を比較する時には、次のようにする。

冬の血圧値;x_1、x_2、x_3、・・・・・、x_n　　夏の血圧値;y_1、y_2、y_3、・・・・・、y_n

各個人の冬と夏の血圧値の差;$z_1=x_1-y_1$、$z_2=x_2-y_2$、$z_3=x_3-y_3$、・・・・・、$z_n=x_n-y_n$

これから平均 $\bar{z}$ 、標準偏差 s_z を求めると、次のようになる。

$$t=\frac{\bar{z}-\mu_z}{\frac{s_z}{\sqrt{n}}}$$

ここで、t は自由度（n－1）のt分布である。

例）患者10人に対し、A、Bの両睡眠薬を与え、下記のようにその睡眠の増加時間を調査した。睡眠の効果は同じであるかどうかという問題である。

患者	1	2	3	4	5	6	7	8	9	10
A薬(x)	1.9	0.8	1.1	0.1	-0.1	4.4	5.5	1.6	4.6	3.4
B薬(y)	0.7	-1.6	-0.2	-1.2	-0.1	3.4	3.7	0.8	0.0	2.0
x - y = z	1.2	2.4	1.3	1.3	0	1.0	1.8	0.8	4.6	1.4

$\bar{x}=2.33$、 $\bar{y}=0.75$、 $\bar{z}=1.58$、　　$s_x=2.002$、 $s_y=2.789$、 $s_z=1.230$

H_0： $\mu_x=\mu_y$ （$\mu_z=0$） 以上の値から計算すると、

$$t = \frac{1.58 - 0}{\frac{1.230}{\sqrt{10}}} = 4.06$$

$t_9(0.05)= 2.262$ で t = 4.06 の方が大きい。 したがって、有意な差があり、 $\mu_x=\mu_y$ （ $\mu_z=0$ ）という仮説は棄却される。この結果、対立仮説である　$\mu_x\neq\mu_y$ （ $\mu_z\neq0$ ） が採択され、A薬の方がB薬に比べて睡眠に対する効果があると考えた方が妥当である。

また、仮説を示すために文章の終わりに 「 P ＜ 0.05 」 と表すこともある。

対応のない場合

対応のない場合は 2 つのデータ集団の個体は同一ではなく、対応のある場合のように、2 つの測定値の差というものはありません。したがって、2 つの集団それぞれの平均値を比較することになります。

平均値の差の検定

2つの学校における体重の差のように、2つの集団における平均値の差の検定を必要とする場合がしばしば生じる。この場合、通常は2つの母集団の平均・標準偏差とも不明なことが多いので、基本的にはt検定を行う。ここでは、問題としている変数x、y、その母標準偏差をそれぞれσ_x、σ_yとする次のような条件がある。

(1) x、yが独立であること

(2) x、yがともに正規分布に従うこと

(3) $\sigma_x = \sigma_y = \sigma$ であること

以上の条件のもとでt検定を行うが、例えば母集団と標本が次の表のようだとすると、次の式は
自由度$=n_1+n_2-2$のt分布に従う。

	母集団		標本	
	平均	標準偏差	平均	標準偏差
Aグループ	μ_1	σ	$\bar{x}$	s_x
Bグループ	μ_2	σ	$\bar{y}$	s_y

$$t=\frac{\left(\bar{x}-\bar{y}\right)-\left(\mu_1-\mu_2\right)}{\sqrt{\frac{(n_1-1)s_x^2+(n_2-1)s_y^2}{n_1+n_2-2}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

今、男女各10人の身長を測定すると、次のようになった。

男:178,168,170,174,164, 171,169,171,170,180

女:155,162,159,147,162, 151,160,151,162,161

男:$\bar{x}=171.5$　$s_x=4.72$

女:$\bar{y}=157.0$　$s_y=5.58$

ここで、帰無仮説H_0; $\mu_1=\mu_2$、対立仮説H_1; $\mu_1\neq\mu_2$として検定すると、

$$t=\frac{(171.5-157.0)}{\sqrt{\frac{(10-1)4.72^2+(10-1)5.58^2}{10+10-2}\left(\frac{1}{10}+\frac{1}{10}\right)}}=6.28$$

となる。大学生では男は女より体格がよいので片側検定で行う。t分布の自由度18、有意水準が$\alpha=0.05$のとき、付表5より片側検定なのでαは0.1のところをみればよい。自由度νが18でみるとtは1.734となり、計算で求めたtの値6.28の方が大きい。したがって帰無仮説は棄却され、男の身長は女の身長よりも大きいと結論付けられる。

演習問題

１．次の資料は、男女学生の体重です。体重が正規分布に従うと仮定して、t 分布により、男女の体重（μ、μ）が等しいという仮説を検定してください。

ポイント：片側検定で行う。有意水準は 0.05 とする。

男：65，58，53，63，70，68，48，62，70，56

女：51，53，47，42，49，51，55，51，55，52

男の平均＝61.3、女の平均＝50.6、男の標準偏差＝7.44、女の標準偏差＝3.89

t＝4.03

２．次の資料は妊娠前と妊娠後の体重です。妊娠前後の体重に差がないという仮説を検定してください。

妊娠前：52.0，51.0，42.0，51.0，42.0，49.0，51.0，44.0，46.1，47.0

妊娠後：61.5，57.3，50.7，60.2，52.5，53.6，57.1，56.0，55.6，54.7

差の平均は-8.41、差の標準偏差は 2.24、ｔ＝-11.87

もっと難しい統計学

割合の検定、χ^2検定（例えば、分割表の独立性の検定)、フィッシャーの直接確率計算法、適合度の検定、等分散性の検定（F 検定)、独立試行の定理、ノンパラメトリックな検定、U テスト（マン―ホイットニー法)、コロモゴロフースミルノフの検定、

関連、回帰直線、相関関係、相関係数、単回帰分析、独立変数、従属変数、一般化線形モデル、回帰直線、回帰係数、重回帰分析、重相関係数、偏相関係数、ロジスティック回帰分析、多変量解析

もう一度、検定とは

ｔ検定は、1 群の平均値について、母集団の平均との比較をするとき、あるいは 2 群の平均値の差について検定するときに使う。なお、2 群の平均値の差の検定の場合、母分散が未知であり、等しいと考えられない場合には、ウェルチの検定を行う。

分散分析は 3 群以上の平均値の差の検定に使う。F 検定は等分散性の検定に使う。これは 2 群の母平均が等しいとみなせるかどうか判断するときに使う。カイ二乗（χ^2）検定は、クロス表（マスターテーブル)、ある食品を摂取した人としない人とで症状を有する人の割合が相違しているかどうかを確かめる（分割表の独立性の検定）ときに使う。この検定は、ほかに適合度、一様性の検定などに使う。なお、観測度数が小さいとき、とくに 4 以下の時は、フィッシャーの直接確率計算法を利用する。

2018年度　保健師国家試験対策

（統計の基本的な事項と保健統計・疫学の計算を要する部分）

松本治彦

疫学と保健統計を合わせると、保健師の国家試験問題の中で最も出題数が多い分野、出題の内容は単に数値を問うものではなく、基本的な理論や考え方を問う傾向が見られる！

保健統計

１．統計学の基礎

・正規分布、代表値（平均値、中央値、最頻値）、ばらつき（分散、標準偏差）
・ｔ分布、F分布、二項分布（結果が成功か失敗のいずれかであるｎ回の試行で、成功数で表される離散確率分布）
・度数分布、四分位数、四分位偏差、変動係数、範囲、パーセンタイル値、百分率
・量的データ（比尺度、間隔尺度）、質的データ（順序尺度、名義尺度）
・円グラフ、棒グラフ、ヒストグラム、帯グラフ、折れ線グラフ、散布図
・検定、対立仮説、帰無仮説、有意水準（危険率）、帰無仮説を棄却
・第1種のエラー（αエラー）、第2種のエラー（βエラー）、マスク化（盲検法）

２．人口統計

・出生数＝出産数－死産数、出生率（人口千対）＝（1年間の出生数）/（その年の人口）×1000
・生産年齢人口（15歳～64歳）、年少人口（0～14歳）、老年人口（65歳以上）
・高齢化の定義；高齢化率＝65歳以上の人口÷総人口
　高齢化社会（高齢化率7～14%）、高齢社会（高齢化率14～21%）、超高齢化社会（高齢化率21%以上）
・合計特殊出生率（粗再生産率）；15～49歳までの女性の年齢別出生率を合計したもの、その年の年齢別出生率で1人の女性が一生の間に産む平均子供数に相当する。≧2.07で将来人口は増加、＜2.07で将来人口が減少。＝（母の年齢別出生数）/（年齢別女性人口）の15～49歳の合計
・純再生産率＝（母の年齢別女児出生数）/（年齢別女性人口）の15～49歳の合計
・総再生産率＝（(母の年齢別女児出生数）/（年齢別女性人口））×（(女性の生命表の同年齢の定常人口）/（10万人））の15～49歳の合計

3.パラメトリックな検定（正規分布を仮定している）

ｔ検定は、1群の平均値について、母集団の平均との比較をするとき、あるいは2群の平均値の差について検定するときに使う。なお、2群の平均値の差の検定の場合、母分散が未知であり、等しいと考えられない場合には、ウェルチの検定を行う。

分散分析は3群以上の平均値の差の検定に使う。F検定は等分散性の検定に使う。これは2群の母

平均が等しいとみなせるかどうか判断するときに使う。カイ二乗（χ^2）検定は、クロス表（マスターテーブル)、ある食品を摂取した人としない人とで症状を有する人の割合が相違しているかどうかを確かめる（分割表の独立性の検定）ときに使う。この検定は、ほかに適合度、一様性の検定などに使う。なお、観測度数が小さいとき、とくに 4 以下の時は、フィッシャーの直接確率計算法を利用する。

4. ノンパラメトリックな検定

平均値の差の検定や t 検定は、正規分布を仮定して成り立つもの（パラメトリック法)。しかし、統計的な検定を行う場合、このよう母部集団の分布を仮定して検定することができる場合がすべてではない。例えば、質問紙調査の回答としてしばしばみられる、「１．とてもそう思う」「２．ややそう思う」「３．あまりそう思わない」「４．まったくそう思わない」といった順位尺度データの場合には、2 つの質問項目の回答に有意な差があるかどうかについて知る方法として、母集団の分布形を仮定しないノンパラメトルック法を用いる。独立 2 標本の代表値の差の検定は、マンホイットニー法（U テスト)、対応のある 2 群の代表値の差の検定はウィルコクソン（の符号付順位和）検定を使う。2 群以上の代表値の差の検定にはクラスカルウォリス検定を使う。抽出した標本の元となる母集団と特定の理論分布が一致するかどうかを検定するにはコルモゴロフスミルノフ法(K－S 法)を使う。このとき、累積相対度数を比較する。また、正規分布からのずれの大きさによって外れ値の有無を判定するには、グラッブス・スミルノフ棄却検定を使う。

5. 相関

相関係数は 2 つの変量ｘ、ｙの関連の強さを表す指標で－1 から 1 の間の値を取る（負の相関、正の相関、相関なし)。また、両者に直線的相関が想定できるとき、2 変数の関係は、直線の式で表すことができる。 ｙ＝bx＋a　　（この式の場合、ｙが従属変数、ｘが独立変数になる）
この直線を回帰直線、直線の式を回帰式という。また、この式の係数の決定方法が最小 2 乗法である。

6. 多変量解析

複数個の変量をもつ多変量データを変量間の相互関係を考慮に入れて分析する一連の統計的手法が多変量解析である。医学、看護学分野では、1 人の人間がもつデータは、体格、血圧をはじめ、血圧データ、心理データなど、単に 1 時点をとってみても、非常に多数あり、そのほとんどは相互に関連しあっているといえるもの。したがって、人間に関する医学、看護学データの分析にとって、多変量解析は必要不可欠なもの。

主成分分析は、多くの変量が与えられたとき、それらの変量のもつ情報をできるだけ多く表現できるような合成得点を求める分析方法。例えば、人の「身長」「体重」という 2 つの変量を考えると、身長も体重も大きい人は「体格」がよいとされる。このことから、これら 2 変量をもとに「体格（の良さ、悪さ)」という主成分を求め、「体格」だけで、「身長」「体重」の 2 変量がもつ情報をある程度表現しようとしたもの。

因子分析は、数多くの変量の中に潜む共通の因子を探り出す分析方法。例えば、上記の例で人の「身長」「体重」「胸囲」などの変異から、「体格増大の素因」「肥満傾向」などの基礎として隠れた因子を探り出すもの。典型的なのものには、性格特性の分析がある。

重回帰分析は、ある変数を、いくつかの変数によって予測する式を作成する。

判別分析は、いくつかの変量をもつ個体が、いくつかの母集団のうちのどれかに属しているかを判

別する分析方法。変量間の相関関係をもとに、母集団ごとの差が明確に現れるような値が得られるように各変量に重みをつけ、そうして得られた重みづき得点により母集団を判別する。

クラスター分析は個体のもつ変量の類似度をもとに、個体をいくつかの集団（クラスター）にまとめ上げていこうという手法の総称。類似度は、個体間の距離の算出。

疫学（語句の意味と計算手順）

1. 疫学の概念

・疫学 epidemiology とは、人間の集団を対象に、健康の状態を把握するとともに、異常の原因を見つけ、疾病対策を講じ、その効果を測定するまでの一連の研究のことを言う。疫とは「はやり病」の意味、感染症が公衆衛生上の最大の問題であった時代のネーミングがそのまま残っている。しかし、現在では感染症にとどまらず、悪性新生物や循環器疾患などの生活習慣病まで、その範囲は広がっている。

・1849 年　ジョン・スノウ、1949 年　フラミンガム研究、明治時代　高木兼寛（かねひろ）の脚気（かっけ）に関する研究

・断面調査、前向き調査、後向き調査

断面調査（横断的調査）

ある任意の一時点（あるいは一期間）を設定し、その時点における現況や実態を把握しようとするもの。時間を追った情報を与えてくれるものではない。因果関係にまで議論を広げることは不可能である。

前向き調査

時間を追って変化を調べ、因果関係を調べようとする調査研究の代表的なものが前向き調査。観察研究、臨床試験など。前向き調査は事象の時間的関連を調べることができ、その観察も時間を追って自然な状況を追いかけていくため、因果関係を調べるのには理想的な手段である。欠点としては、場合によっては結果がでるのが数十年先になる。

後向き調査

結果の有無によるグループ間の比較により、原因の分析に差があるかを調べる。現在存在している特定の結果から、時間を遡って過去の原因を探索する。

・疫学的研究方法の分類；

疫学研究は介入の有無により大きくは介入研究 intervention study（介入疫学；実験的研究）と観察研究とに分かれる。観察研究はさらに記述研究（記述疫学；仮説の設定）と分析研究（分析疫学；仮説を分析、検証する）に分かれる。分析研究は生態学的研究、横断的研究、症例対照研究 (case-control study（患者対照研究、後ろ向き研究)）およびコホート研究（追跡研究 follow-up study；前向き研究、縦断研究）に分かれる。一方、介入研究（仮説を介入実験して確かめる）は臨床試験（対照群があるか否か、対照群がある場合にはクロスオーバーと無作為対照比較試験などに分かれる；患者に対する治療の効果を判定するのに用いられる)、野外試験（ワクチン効果の評価など疾病をもっていない人に対する介入研究）および地区介入（地域研究；集団を対象）に分かれる。

介入研究は疾病の治療などについて、介入の効果を測定する研究。ここで介入とは、研究のために対象者の生活習慣・保健行動・治療法などを意図的に変えたり、意図的に固定したりすることで

ある。

一方、観察研究は介入を一切行わずに対象集団の自然の姿を観察する研究。介入研究も観察研究にも、それぞれ特徴の異なる調査方法がいくつかあり、それらは上記のように研究デザインと呼ばれて分類されている。

コホートは古代ローマ軍の歩兵集団のことで、最初からある大隊のメンバーは決まっているため、戦死などによるメンバー個々の推移を実際に観測できることからこの名前が用いられている。

症例対照研究は、まず研究対象となる疾病をもつ人（症例）を集め、症例に対応して（多くの場合マッチングを経て）疾病をもたない人（対照）を集める（マッチングは、一般的には症例と性、年齢、居住地区などが同じ対照を選ぶ作業をマッチングとよぶ。これにより、統計的な効率がよくなる「同じ大きさの標本をとっても、偶然の誤差によって結果が真の値からずれる程度を小さくできる。」）。この 2 群で疾病の原因となりそうな要因に相違がないかを調べる研究である。

生態学的研究（エコロジカル研究 ecological study）は、一般には既存の資料から、地域ごとの曝露情報と健康情報を用いてその関連をみるものである。たとえば胃がんの死亡率と食塩摂取量との関連で、食塩摂取量の少ない地域ほど胃がんの死亡率が低くなっている、という関連をみる。安価でしかもすばやくできるのがこの研究の利点ではあるが、この研究では真の原因でない要因でも関連があるという結果が起こり得る（ecological fallacy）ため、この研究で関連が認められただけでは原因であると強く言えない。

横断的研究（cross-sectional study）、有病率研究、断面調査（世論調査）はある一時点の調査で、疾病とその要因に関する情報を収集する研究である。

・流行曲線；発症日時別に発症者数の経過を示したヒストグラムで、発症日時から感染源への暴露時期の特定や感染様式の推定に有用である。
・量－反応関係；感染源への暴露の度合い（量）と発症者数（反応）の関連を検討することで、両者の関連の強さから感染源を推定する際に用いられる。
・マスターテーブル（これは統計学でやったので知っていますね）
・スクリーニング検査（敏感度、特異度、偽陰性率、偽陽性率（100－特異度）、陽性反応的中度、陰性反応的中度、有病率）
・ROC 曲線（receiver operating characteristic curve）集団を対象にしてスクリーニング検査を評価する際に用いる。いくつかのスクリーニング・レベルごとに、敏感度を縦軸に、偽陽性率を横軸にプロットして描く。敏感度と特異度は、一方を上げれば他方が下がるという関係（トレード・オフ関係）にある。ROC 曲線では、敏感度・特異度ともに 1 である点（左上）により近づく曲線を示すものが、より正確な検査であることを意味している。
・因果関係判定のための 5 因子（関連の時間性、一致性、特異性、整合性、強固性）
・因果推論（必要条件、十分条件）；特定の疾患が起こるためには、必ず特定の要因に曝露している（必要条件）、特定の要因に曝露すると、必ず特定の疾患が起こる（十分条件）、必要条件と十分条件を満たすもの（必要十分条件）
・疫学研究で患者に説明する事項；研究の目的、意義、方法、研究の参加によって期待される効果、不利益、同意を撤回できること、同意しなくても不利益を被らないこと、個人情報の保護（匿名化など）

2. 疾病頻度の指標

・曝露；有害物質や病原菌などにさらされること。

・疫学指標（罹患率、累積罹患率、有病率、死亡率、致命率、生存率、死因別死亡割合、PMI、相対危険度（リスク比)、寄与危険割合、人年法）

・有病率＝（ある一時点における有病者数）/（ある一時点における観察対象者集団の数）

・罹患率（リスク）＝（一定期間内に新発生した患者数）/（健康な状態でいる観察期間の総和）

例えば、会食した 100 人で 4 月 1 日～4 日の期間に食中毒になったのが 40 人（2 日 20 人、3 日 20 人、3 日 3 人死亡、4 日 1 人死亡）の場合、（食中毒になっていない 60 人）×4 日＋（2 日に食中毒になった 20 人）×（健康な期間 1 日＋（3 日に食中毒になった 20 人）×（健康な期間 2 日）から 4 月 1 日～4 日の期間の罹患率は 40/（60×4+20×1+20×2）

4 月 3 日夜の時点の有病率は、（40－3）/100－3）

・累積罹患率＝（一定期間内に新発生した患者数）/（観察開始時点での人数）

4 月 1 日～4 日の期間の累積罹患率は 40/100

・死亡率＝（一定期間内での死亡者数）/（観察対象者集団の数）

・粗死亡率（人口千対）＝（1 年間の死亡数）/（その年の人口）×1000

・致命率（致死率）＝（一定期間内でのある疾病による死亡者数）/（一定期間内でのある疾病の罹患者数）

・生存率＝（一定期間内でのある疾病に罹患したが死亡はしていない人数）/（一定期間内でのある疾病の罹患者数）

4 月 1 日～4 日の期間の死亡率は 4/100

4 月 1 日～4 日の期間の致命率は 4/40

4 月 1 日～4 日の期間の生存率は（40－4）/40

・死因別死亡割合は全死因の死亡者数のうち、ある特定死因による死亡者数の占める割合

・PMI は全死亡者の中での 50 歳以上の死亡者の割合

・人年法による死亡率、人口 1000 人で 10 年間での死亡数が 20 人の時、人年法による死亡率（10 万人年対）は、20/（1000×10)、分母を 10 万人年にすると、200 人/100000 人年

・リスク比（相対危険（度））＝（暴露群での疾病発症のリスク）/（非暴露群での疾病発症のリスク）

・リスク差（寄与危険（度））＝（暴露群での疾病発症のリスク）－（非暴露群での疾病発症のリスク）

・寄与危険割合（%）＝（(暴露群での疾病発症のリスク）－（非暴露群での疾病発症のリスク））/（暴露群での疾病発症のリスク）×100

＝1－(非暴露群での疾病発症のリスク)/(暴露群での疾病発症のリスク）×100

＝（リスク比－1）/リスク比×100

＝（1－1/相対危険）×100

・人口寄与危険割合＝（(人口集団の罹患率）－（非暴露群の罹患率））/（人口集団の罹患率）×100

3. オッズ比

・オッズ比（Odds ratio）；ある事象の起こりやすさを 2 つの群で比較して示す尺度で、医学の臨床試験の結果を示す指標としてよく用いられる。例えば、2 つの群の起こる確率をそれぞれｐ、ｑとする

とオッズ比は、$\frac{\frac{p}{(1-p)}}{\frac{q}{(1-q)}}=\frac{p(1-q)}{(1-p)q}$

例題；男女 100 人にビールを飲むか飲まないかについて聞いた結果、男は 80 人、女は 20 人が飲むと答えた。このときのオッズ比は $\frac{\frac{0.8}{0.2}}{\frac{0.2}{0.8}}=\frac{0.8\times0.8}{0.2\times0.2}=1.6$ となる。

4. 疫学調査法

・偶然誤差と系統誤差（バイアス）

調査や実験により、標本から収集したデータに含まれる個々の測定値（観測値）には、本当に知りたい真の値に対して、必ず偶然誤差あるいは系統誤差（バイアス）と呼ばれる 2 種類の誤差が含まれている。

測定値（観測値）＝真の値+誤差（偶然誤差+系統誤差）

偶然誤差と系統誤差は測定（観測）ごとにばらつくものであるが、それらが集合した誤差全体においても一定の性質を持つ。

①偶然誤差

偶然誤差とは、同じ条件で測定をしたときに同じような測定値が出る程度であり、測定値のばらつきをあらわす。偶然誤差の身近な例には、同じ条件で複数回測定したときの、血圧値のばらつきなどがある。偶然誤差は、「データの各測定値がどの程度の精度で測定されたか」ということ（信頼性）の指標であり、誤差が大きいほどデータの信頼性が低いことを示す。このことは、偶然誤差が大きいほど、データに含まれるどの測定値が真の値に近いのかが分らないために測定値の信頼性が低いと考えるとイメージがしやすい。なお、理想的な条件下で収集されたデータの場合、偶然誤差は標準偏差と同じであるため、データの標準偏差が小さいほど信頼性が高いと言える。

偶然誤差の分布

データの偶然誤差は、一般に、正の値と負の値のものが混在しており、全体としては正規分布することが知られている。このことは、偶然誤差に対しては、統計処理やデータ数を増やすことで対応できることを示している。血圧の例でいえば、血圧値を 3 回測定して平均を計算するという統計処理によって、真の血圧値により近い数値を入手することができる。

②系統誤差（バイアス）

臨床研究の基本的な目標は、exposure 暴露と outcome 結果の関係の強さを検討する事です。バイアスとは、exposure と outcome の関係の強さを系統的に歪めてしまうものであり、一般的に選択バイアス、情報バイアス、交絡に分類されます。それぞれのバイアスが、実際に臨床研究が進められているステップのどこに深く関与しているかに注目すると、理解しやすいです。

系統誤差はバイアスとも呼ばれ、何らかの理由によって、データに生じる偏りの程度を表す。上述の血圧の例でいうと、「あるメーカーの全自動血圧計で測定した血圧値は、手動式のアネロイド型血圧計で想定した血圧値に対してつねに数㎜ Hg 高くなる」などの場合があてはまる。

バイアスは「データからどの程度納得できる分析結果を出せるのか」「データはどの程度母集団の性質を代表しているのか」などの性質（妥当性）を表す指標である。バイアスが大きいという場合、データの妥当性が低いことを示す。

バイアスは、事前にかたよりの程度（数 mmHg 測定数値が高くなるなど）が明らかであれば、数値を修正して対応できるが、通常は、収集してしまったデータの偏りを調べる方法はない。バイアスは気付かない間に結果をゆがめてしまうため、データ収集の前に細心の注意が必要である。バイアスには、多くの種類がある。

選択バイアス

選択バイアスは、標本として抽出した対象が、研究対象として想定した母集団（概念的母集団）を反映していないことから生じるバイアス。選択バイアスには、母集団から標本として対象を抽出する際に問題が生じる場合と、概念的な母集団と操作的な母集団の間にかい離がある場合とがある。前者の場合、操作的母集団の全員の名簿があれば、理想通りに無作為抽出をすることで回避できる。一方、後者の場合、外的妥当性を慎重に検討する必要がある。

選択バイアスは以下のような具体例を考えるイメージがつかみやすい。

1）ある村全体のことを調査したいときに、検診受診者から標本を選ぶと、健康な人の比率が多くなる。

2）ある地域の待機児童の現状を調べたいときに、すでに地域の幼稚園・保育園に通園させている両親から標本を選ぶと、現在待機児童をかかえている人は調査の対象外になる。

3）胃がんに関する患者対照研究で、対象者の標本を別の疾患 A で入院している患者から選ぶと、胃がんと疾患 A に共通する危険因子はみつけられない。

情報バイアス

情報バイアスとは、標本から実際にデータを入手する際に、入手した情報が間違っている（あるいはかたよっている）ことから生じるバイアスである。

情報バイアスは、上述した測定機器（全自動血圧計）の不具合などの例はイメージをつかみやすい。ただし、それ以外に、以下のような場合も情報バイアスが大きい例として考えられるため、注意が必要である。

1）アンケートの質問が 2 つの意味を持っていたり、誘導質問になっていたりするため、回答がかたよってしまう。

2）標本に問題があって虚偽の回答が発生しやすいなど、データとして意味を持たない欠損値の比率が多い。

3）特殊な疾患の患者などは同じようなアンケートが繰り返されることにより、対象の記憶状況に偏りが生じる。

4）アンケート項目に対し、対象が先入観で判断してしまう。

交絡バイアス

交絡バイアスとは、2 つの要因間の関係を調べようとするときに、着目していない別の要因（交絡因子）が両者に影響を与えるために、本当の関係とは異なった見かけ上の関係（偏った関係）が生じてしまうバイアスである。

例えば、大学による喫煙率の調査をした時に、「看護大学と総合大学で比較すると、看護大学は総合大学に比べて喫煙率が有意に低い」という結果が出たとする。このような場合、結果からそのまま「看護では、喫煙に関するリスクを勉強しているので、喫煙率が低くなる」などと安易に考察してしまいがちである。しかし、ここで看護大学と総合大学では男女比が大きく異なることに注目するとどうだろうか。一般に、喫煙率は男子大学生のほうが女子大学生よりも高い。そのため、大学の種類によって喫煙率が異なるという結果は、禁煙教育の影響ではなく、学生の男女比の影響かも

しれないのである。その後、データをさらに詳しく分析し、女性だけを取り出してみると喫煙率に有意な違いがなかったとしよう。この場合、「大学の種類により教育の違いと喫煙率の関係を調べたところ、性別（男女比の違い）が交絡因子として影響した」とうことになる。

上述のように、性別や年齢などは一般的に交絡因子になりやすいため、これらの因子にはつねに注意する必要がある。これらに対しては、標本抽出の段階であらかじめ怪しいと思う因子をカテゴリーに分けておけば、ある程度の予防ができる。また、既に抽出された標本についても、あとから性別や年代別のようにカテゴリー別に分析することができれるのであれば、それらが交絡因子であった場合の影響を除外することができる。これを層別分析（層化分析）という。そのほか、性別・年齢など、交絡因子になりそうな条件が等しくなるように標本を抽出したり、分析したりすることができれば、層別分析と同様に影響を回避することができる。このような処理をマッチングという。

マッチング；マッチングは比較する 2 群の対象者の交絡因子の分布が等しくなるように対象者を選び出す方法で通常、対象者選定の段階で用いられる方法。

・標本抽出

単純ランダム（単純無作為抽出法）；母集団の各構成員に一連の番号を割り当て、乱数表（数値がランダムに並んでいる表）等で希望する標本数までランダムに抽出する。最も理想的な標本抽出法であるが、母集団全員の名簿が必要であり、全員を乱数表で順番に選ぶのは手間がかかる。名簿がエクセルなどに入力されている場合は、関数によって乱数を発生させることができる。エクセルの RAND 関数や IF 関数を組み合わせると、簡単に標本を抽出することができる。

系統抽出法；母集団の全個体に通し番号を付ける。標本の最初の個体（抽出開始番号）だけは乱数表などでランダムに選ぶ。それ以降の個体は、その数字から始めて一定間隔で順に抽出する。例えば、1000 人の母集団から 100 人の標本を選ぶとき、抽出率は 100/1000 の 10%で 10 人に 1 人を選ぶことになる。この場合、最初の 1 人を乱数表などで「3 番の人」と決めたら、以後は 10 人ごとに「13 番、23 番、33 番・・・」と標本を選んでいくことになる。

層化ランダム（層化抽出法、層別抽出法）；母集団の各構成員を属性（性別、年齢、職業、居住地など）によっていくつかの層（カテゴリー）に分け、それぞれの層から一定の抽出率で抽出する方法。特定の変数の比率が偏らないようにする方法。

多段ランダム（多段抽出法）；母集団をまずいくつかの集団に区分し、その集団をランダムに抽出する。次に、選定された当該集団の中から個人を標本としてランダムに抽出する方法。たとえば、総合病院で個々の診療科の中から無作為に標本を選ぶ方法。

集落抽出法；多段抽出法を行う場合、通常は、最後の段階で名簿とともに無作為抽出を行って標本を決める。しかし、調査する標本が大きい場合は、最後の段階で選ばれた地域全体を標本とする場合がある。この場合、調査対象として個人を抽出するのではなく、調査対象として地域を抽出することになるので、集落抽出法と呼ばれる。例えば、国民生活基礎調査や国民健康・栄養調査の場合、調査区をもとにした集落抽出法が用いられている。

有意抽出法；有意抽出法とは無作為抽出ではない抽出法の総称であり以下のようなものがある。

（1）自分で調査対象として典型的だと思う人に調査する典型法

（2）知り合いに頼む縁故法（機縁法）

（3）テレビのインタビューのように道に通りかかった人に調査する偶然法

（4）調査に協力してくれる人を募集する募集法

（5）調査対象に次の調査対象を紹介してもらう雪だるま法（スノーボール法）

これらの方法は、少人数であれば比較的手軽に対象を選定できるが、統計的な手法で母集団を類推することはできないため、質的研究や予備調査などに利用されることが多い。なお、有意抽出法を用いる場合、思わぬ誤差が隠れている場合があるため、標本の分析結果が母集団の様子を本当に反映しているのかどうかをしっかり検討する必要がある。

5. スクリーニング

ふるいわけ。適格審査、特に健康な人も含めた集団から、目的とする疾患に関する発症者や発症が予測される人を選別する医学的手法。

・スクリーニングの条件；

罹患率が高い、または発見・治療が遅れると重篤化する。

疾患の自然史が判明している（疾患の経過が判明している）。

最初の微候が現われてから顕在化するまでの期間がある程度長いこと。

スクリーニング陽性者に対して、精密検査で確定診断できる疾患であること。

簡便で低侵襲で信頼性の高い、比較的安価な検査方法が存在すること。

敏感度、特異度が高いこと。

被検者に肉体的・精神的負担を与えないこと。

疾患に対する治療法が確立されていること。

		疾病		合計
		あり	なし	
検査結果	陽性	a	b	a+b
	陰性	c	d	c+d
合計		a+c	b+d	a+b+c+d

・敏感度＝a/（a+c）×100

・特異度＝d/（b+d）×100

・偽陰性率（100－敏感度）＝c/（a+c）×100

・偽陽性率（100－特異度）＝b/（b+d）×100

・陽性反応的中度＝a/（a+b）×100

・陰性反応的中度＝d/（c+d）×100

・有病率＝（a+c）/ （a+b+c+d）×100

第 103 回（2017.2）保健師国家試験

午前問題

問題 27

人口 10 万人の市において、ある一定期間の結核患者の発生頻度を表現する指標として適切なのはおどれか。

1. 罹患率、2. 有病率、3. 被患率、4. 受療率、5. 相対頻度

問題 28

生態学的研究によって、世界各国の 1 人当たりの食塩摂取量と高血圧症有病率との関連の程度を評価するために計算するのはどれか。

1. 寄与危険、2. 変動係数、3. 相対危険、4. 相対頻度、5. 相関係数

問題 29

ある集団の特定健康診査で得られたヘモグロビン A1c 値の頻度の分布を確認するのに最も優れているのはどれか。

1. 散布図、2. 円グラフ、3. 帯グラフ、4. ヒストグラム、5. 折れ線グラフ

次の文を読み問題 44、問題 45 の問いに答えよ。

人口 5 万人の市、市の人口は平成 20 年度以降は変化していない。市は、A、B および C の 3 つの地区からなり、肺がん対策として検診の受診率の向上に取り組んでいる。市の肺がん検診は、平成 26 年度までは A 地区の保健センターで行う集団検診のみであったが、平成 27 年度からは B 地区にある病院でも検診を行っている。各地区の肺がん検診の受診者数および対象者数を表に示す。

表　市内各地区の肺がん検診の受診者数および対象者数（人）

	A 地区		B 地区		C 地区	
	受診者数	対象者数	受診者数	対象者数	受診者数	対象者数
平成 20 年度	2400	8000	3100	10000	500	2000
平成 27 年度	3100	8000	4900	12000	600	2000

問題 44

市内の肺がん検診の状況について正しいのはどれか。

1. 検診実施施設が増えた後、市全体の検診受診率は増加した。
2. 検診を実施している病院がある地区の対象者数に変化はない。
3. 平成 27 年度の市全体の人口に占める検診対象者の割合は 50%以上である。
4. 市全体の検診受診率は平成 24 年のがん対策推進基本計画の目標値を上回っている。

午後問題

問題 34

因果関係を推測することのできる研究デザインはどれか、２つ選べ。

1. 横断研究、2. 記述疫学、3. 生態学的研究、4. コホート研究、5. 症例対照研究

問題 35

心筋梗塞発症者 100 人と性･年齢をマッチングした心筋梗塞非発症者 100 人の 5 年前の健康診査の結果を調査し、糖尿病の有無を確認した。その結果、心筋梗塞発症者で 20 人、心筋梗塞非発症者で 15 人が糖尿病であった。

糖尿病であることの心筋梗塞発症に対するオッズ比を求めよ。ただし、小数点以下第 2 位を四捨五入すること。

次の文を読み問題 45、問題 46、問題 47 の問いに答えよ。

運動習慣の死亡率に対する影響の調査のために 1 万人を対象とした 10 年間のコホート研究を行った。本研究では、年齢、食習慣及び経済状況など参加者の基礎的背景も併せて調査した。脱落者を除いた結果を表に示す。

	疾患別の死亡数（人）				観察人年
	疾患 A	疾患 B	疾患 C	その他の疾患	
運動習慣あり	100	20	5	200	40000
運動習慣なし	140	100	10	400	40000
合計	240	120	15	600	80000

問題 45

運動習慣があることに対して、運動習慣がないことの疾患 A による死亡に関する 1 万人対の寄与危険はどれか。

第 102 回（2016.2）保健師国家試験

午前問題

問題 17

疫学研究に関する記述で正しいのはどれか。

1．記述疫学には介入研究が含まれる。2．横断研究によって因果関係を証明できる。3．分析疫学は記述疫学よりも疾病と要因との関連を示しやすい。4．前向きコホート研究は稀少疾病の罹患リスクを検討するのに優れている。

問題 38

直接法による年齢調整死亡率の特徴はどれか。

1．小規模な集団の観察に適している。2．高齢者の多い集団では高くなりやすい。3．値は標準化死亡比（SMR）として示される。4．異なる観察集団の死亡率を直接比較できる。5．計算には観察集団の年齢階級別死亡率が必要である。

問題 39

割合の差の検定について正しいのはどれか。2 つ選べ。

1．横断研究が必要である。2．t 検定で有意差を検定する。3．クロス集計表は有用である。4．ハザード比を求めることができる。5．χ 二乗検定で有意差を検定する。

問題 40

疾病 A の新しいスクリーニング検査の性能を評価するために､疾病 A の患者 100 人と疾病でない者 100 人に対して検査を実施した。疾病 A の患者のうち 60 人と、疾病でない者のうち 10 人とが検査の結果、陽性であった。特異度を求めよ。ただし、小数点以下の数値が得られた場合には、小数点以下第 1 位を四捨五入すること。

午後問題

問題 21

特定健康診査を受診した 100 人の胸囲と HbA1c 値について、個人ごとの 2 つのデータを一度に示し両者の関連を表現するのに優れているのはどれか。

1．折れ線グラフ。2．ヒストグラム。3．円グラフ。4．散布図。

次の文を読み問題 51、問題 52、問題 53 の問いに答えよ。

出生 10000 対 6 の発症率と言われている先天性神経疾患 A について、その発症要因に関する症例対照研究を計画している。

問題 52

先天性神経疾患 A の発症に、母親の出産時年齢が有意に関連することが既に分かっている。そのため、症例群の母親と同じ出産時年齢の母親を対照群として選定し、ペアを作成して調査対象者を集めた。この制御法はどれか。1．層化。2．限定。3．無作為化。4．マッチング。

第 101 回（2015.2）保健師国家試験

午前問題

問題 24　統計分析について正しいものはどれか。

1. 割合に関する検定には、ｔ検定を用いる。
2. 点推計値での偶然性の判定はできない。
3. 多変量解析は情報バイアスを補正する。
4. 帰無仮説を採択して統計学的有意性を確定する。

午後問題

問題 19.　差をとって寄与危険を計算できる指標はどれか。

1. 罹患率　2. 有病率　3. オッズ比　4. 致命率（致死率）

問題 20.　スクリーニングについて正しいのはどれか。

1. 確定診断を目的とする検査である。
2. 敏感度 100%の検査で陽性結果であれば疾患がある。
3. 陽性反応的中度は有病率の影響を受けにくい指標である。
4. 同一検査で敏感度と特異度の両方を改善することはできない。

問題 21.　分布の指標について正しいのはどれか。

1. ヒストグラムで最も頻度が高い値は中央値である。

2. 広く散らばった分布は標準偏差が小さい。
3. 対象数が増えると標準偏差は大きくなる。
4. 平均値は外れ値の影響を受けやすい。

次の文を読み 50〜52 の問いに答えよ。

保健所で肥満防止を目的とした教室の参加者を対象に、運動と体重変化の関連を調べることにした。対象者は軽度肥満の 40 歳代の女性 300 人であり、本人の希望で軽い体操をする群 100 人（体操群）と中等度の運動をする群 200 人（運動群）とに分かれ、同じ保健師による集団指導を受けた。教室開始時に体重測定を行い半年後にどれだけ体重が変化したかを調べた。

問題 50. この調査の研究デザインはどれか、2 つ選べ。

1. 横断研究　2. 介入研究　3. 前向き研究　4. 症例対照研究　5. 生態学的研究

問題 51. 半年間で体重が 3.0 kg以上減少した参加者を減量ありと判定した。体操群で減量ありは 10 人、なしは 90 人、運動群での減量ありは 50 人、なしは 150 人であった。脱落者はいなかった。運動群の減量効果についてのオッズ比を求めよ。ただし、小数点以下第 2 位を四捨五入すること。

問題 52. この調査は体操群と運動群とを本人の希望によって分けている。くじで割り付ける場合と比べて、仮説検証の妥当性からみたこの調査の問題点で最も大きいのはどれか。

1. 二重盲検ができない。　2. 相対危険が計算できない。　3. 2 群の人数を等しくできない。　4. 2 群の属性の分布が制御できない。

第 100 回（2014.2）保健師国家試験

午前問題

問題 23　ヒストグラムについて正しいのはどれか。

1．連続量や度数の経時的変化を折れ線で示す。
2．名義尺度の度数の分布を棒の高さとして示す。
3．ある範囲にある連続量の度数を面積の大きさとして示す。
4．標本のもつ２つの連続量をプロットしてその関連を示す。

問題 24　統計調査と調査内容の組み合わせで正しいのはどれか。

1．国勢調査・・・健康保険の種別
2．人口動態統計・・・転出入
3．社会生活基本調査・・・生活時間の配分
4．国民生活基礎調査・・・栄養摂取状況

問題 35　相関について正しいのはどれか。

1．因果関係の必須項目である。
2．相関係数が大きいほど相関関係は強い。
3．相関が全くないときの相関係数は 0 である。

4．相関係数は 0 から 100 までの数値で表される。
5．２つの連続量の一方を使用して他方を推計することをいう。

次の文を読み、問題 44、問題 45、問題 46 に答えよ。
山間部にある人口 8000 人、高齢化率 40%の A 町、高齢者のうち独居者の割合 35%。町に急な坂が多く、電車やバスが運行していないため、高齢者は買い物に不便を感じている。
問題 45　保健師が高齢者を対象に健康に関するアンケートを実施した結果、食事回数が 1 日 2 回と回答した割合が、独居高齢者では 60%、同居者がいる高齢者では 30%であった。独居高齢者の食事回数が 1 日 2 回であることに対する寄与危険はどれか。
1．2.0、2．0.50、3．0.30、4．0.26

午後問題

問題 16　日本の血液型のうち AB 型の割合が 10%であるとする。無作為に選んだ 100 人の日本人集団の中に AB 型の人が 20 人以上いる確率を知りたい。この集団の中に含まれる AB 型の人数が従う分布として最も適切なのはどれか。
1．t 分布、2．F 分布、3．正規分布、4．二項分布

問題 18　情報処理について誤っているのはどれか。
1．データをコンピューターで使用可能な形にすることをデータの電子化と言う。
2．体系づけられたデータやファイルの集まりのことをデータベースという。
3．同じ形式のデータを連結することをレコードリンケージという。
4．氏名の削除や番号・記号への置き換えのことを匿名化という。