

データの科学的な見方 2015

松本治彦

(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、それを現実に応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の見せない。「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けませんので応用する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、そこからじゅうに溜め込まれた。そして「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が、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ツ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。

「ランダム化」という最強の武器

ミルクが先か、紅茶が先か

データの取り方自体を工夫すれば、あるいはより高度な解析手法を用いれば、かなりの部分明らかになる。近年「A/B テスト」と呼ばれ、統計家が長年「ランダム化比較実験」と呼ぶものが、どれだけ強力か、という話。

「科学」の対象を拡大したランダム化比較実験

「人間の制御しうる何物についても、その因果関係を分析できるから」。フィッシャーがほとんど独力で作り上げたランダム化比較実験は科学哲学を揺り動かし、科学で扱える対象の領域を爆発的に拡大させた。

1920 年代末のイギリスで、ある婦人がミルクティについて「紅茶を先に入れたミルクティ」か「ミルクを先に入れたミルクティ」で味が全然違うと答えた。この実験をやったのがフィッシャー。

なぜ、ランダムでなくてはならないのか？

両タイプのミルクティをランダムに飲ませ、どれほどあてられるのかを検証すればよい。これがランダム化比較実験の基本的な考え方。ミルクティはランダムに飲まされるのだから、見えない場所でミルクティを注がれた場合に順番を予測することは誰もできない。

「一杯の完璧な紅茶の淹れ方」

フィッシャーはさらに「実験計画法」の中で、婦人に実験のやり方をどの程度説明すべきか、何杯のミルクティでテストすべきか、といった詳細を検討し、また、想定される婦人の回答結果と「婦人がでたらめに回答してそれだけの正答率が偶然得られる確率」を計算している。

フィッシャーの考えた「科学的に実証するための手順」のうち最も重要なアイディアが、「ランダム化する」ということ。

婦人は出されたミルクティをすべて正確に言い当てた。つまり、彼女がランダムな 5 杯のミルクティを飲んでいたら、偶然すべて当てる確率は 2 の 5 乗分の 1、すなわち 32 分の 1 (約 3.1%)、もし 10 杯すべてを当てれば、1024 分の 1 (約 0.1%) になる。これほどの確率を示されれば、彼女が何

らかの形でミルクティを識別できていると考える方が自然である。

英国王立化学協会が 2003 年に発表した「1 杯の完璧な紅茶の淹れ方」について、「牛乳は紅茶の前に注がれるべきである。なぜなら牛乳蛋白の変性（変質）は、牛乳が摂氏 75℃になると生じることが確かだからである。もし牛乳がお湯の中に注がれると、それぞれの牛乳滴は牛乳としてのまとまりから外れ、確実に変性が生じるだけの時間を紅茶の高温に取り囲まれる。もしお湯が冷たい牛乳に注がれるならば、このような状況ははるかに起こりにくい。」

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に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。

保健師国家試験の疫学・保健統計出題傾向（2011 年度版 医学書院）より抜粋

『近年の疫学・保健統計に関する出題傾向には、以下のような特徴がある。

- 1) 疫学と保健統計を合わせると、国家試験問題の中で最も出題数が多い分野である。
- 2) 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。

- 3) 疫学・保健統計学は出題基準の「地域看護学Ⅱ・Ⅲ」との関連での出題も多く、それらの分野で、あるいは状況設定問題の中でも必然的に出題されている。
- 4) 内容として過去数年間に頻繁に出題されているキーワードを次に示す。実際のデータでの計算が求められるのは、b,c,g,h まれに i の項目である。
- a. 疫学の定義、疾病の発生要因
 - b. 疫学諸指標（罹患率、有病率、死亡率、寄与危険度、寄与危険度割合、人口寄与危険度割合）
 - c. 症例対照調査、コホート研究、流行調査（感染症集団発生時の疫学調査）、臨床試験（無作為対象試験）
 - d. 国民生活基礎調査、患者調査
 - e. 感染症対策の考え方、感染症分類、HIV 感染症、呼吸器感染症（結核、SARS、鳥インフルエンザ）、検疫感染症
 - f. 検診調査の結果と事後指導
 - g. スクリーニング法に関する指標など
 - h. 年齢調整死亡率、標準化死亡比
 - i. 人口動態統計、生命表
 - j. 因果関係の判断
 - k. 疾病登録と個人情報保護
 - l. 「健康日本 21」の数値目標、がん、心臓病、脳卒中、メタボリックシンドロームについての指標など
 - m. 保健医療の動向（栄養、母子保健、産業保健、精神保健など）と関連させた健康増進対策、ハイリスクストラテジーとポピュレーションストラテジー
 - n. 保健予防活動の評価など

対策：

出題傾向から考えると、対策としては次のような点が重要である。

- 1) 公式を丸暗記するのではなく、原理・原則を理解することが重要である。これは、将来において地域での活動に生かすことのできる学習方法でもある。
- 2) 問題の出典の多くは「国民衛生の動向」からのものである。ここでも数字を覚えるのではなく、読み物として読み、使われている用語などの解説を「国民衛生の動向」に加えて教科書で確認したり、興味のあることは専門書を紐解くことを勧めたい。
- 3) 疫学・保健統計は大きく分けると、学問的な理論の部分とそれらが健康を把握するための実践として使われる部分とがあり、両者ともに欠かせない部分である。したがって、母子・成人・高齢者・精神・障害者（児）・感染症・歯科保健指導など対象別保健指導との関連で、あるいは学校・産業など場面別保健指導との関連で出題される。その場面でも一度、どのように疫学や保健統計が関わっているかを考えてみることも学習になる。
- 4) 3) に述べたことの典型的なものが状況設定問題であり、問題を解くとともに、臨地実習のときに把握した地域の数値を当てはめて、自分で再考することも勧めたい。
- 5) 保健師国家試験の出題は以前に比較し、地域診断を取り入れた実践に基づいた方向に進んでいる。一次予防・二次予防について、統計を技術として身につけた、より疫学的評価能力を備えた保健師が求められている。明確に解答を得られる疫学・保健統計の分野を勉強することにより、確実に得

点につながると思われる。

- 6) 具体的な事例の中で、その状況に合った調査方法を考えたり、適切な分析方法を選んだり、与えられた数値表を読み、実施すべき対策を考える応用力も求められる。これらは基本的な考え方を身に付けておけば十分である。
- 7) 厚生労働省が国家試験問題を作成するので、その時々¹の厚生行政の優先課題に関係した統計が出題されるようになってきた。特に近年では、医療制度改革に関連して、国民医療費・高齢者医療・在院日数などが出題されている。日頃より新聞を読むことも重要である。
- 8) 今後は、統計学の問題が増加すると予想される。様々な確率分布の特徴、ばらつきの指標、適切な検定方法の選択、検定結果の解釈などをマスターしておきたい。』

保健師模擬問題（医学書院 2015 年度版より抜粋）

午前

例 4 50 床の病棟で入院患者は 45 人である。この病棟の病床利用率を求めよ。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。（正解は 90）

問題 20 交絡因子の制御方法で正しいのはどれか。

1. 交絡因子の 1 つの状態のみを観察対象とすることを標準化といい、研究の計画段階での制御方法である（標準化ではなく、限定である）。
2. 曝露群と非曝露群において交絡因子の分布が等しくなるようにすることをマッチングといい、研究の解析段階での制御方法である（マッチングは計画段階での制御方法である）。
3. 対象集団より標本を抽出する際に、どの標本もまったく等しい確率で選ばれるようにすることを無作為抽出といい、研究の解析段階での制御方法である（無作為化は計画段階での制御法である）。
4. 母集団を同じ特徴の層に分けることを層化といい、研究の解析段階での制御方法である（○）。

問題 21 3000 人の住民を対象に、ある特定の疾患 X に関する検査 Y を行った。この集団を一定の診査基準に基づいて疾病の有無を判別したところ、以下の結果を得た。

- ・疾患 X があると診断された者が 800 人、うち 500 人が検査 Y で陽性と判別された。
- ・疾患 X がないと診断された者のうち、30 人が検査 Y により陽性と判別された。

敏感度、特異度、陽性反応的中度の正しい組み合わせはどれか。

1. 62.5%、98.6%、94.3%
2. 62.5%、94.3%、98.6%
3. 98.6%、94.3%、62.5%
4. 94.3%、98.6%、62.5%

（敏感度とは、疾患を持っているもののうち、陽性反応を呈したものの割合のことである。特異度とは、疾患を持っていない者のうち陰性反応を呈した者の割合である。陽性反応的中度はスクリーニングで陽性反応を呈した者のうち実際に疾患を持っている者の割合である。1 が○）

問題 37 飲酒が肺がんの危険因子になっているかどうか症例対照調査を実施したところ、「飲酒習慣がある」が「飲酒習慣がない」に比べ、2.5 倍のリスクのあることが示された。しかし、飲酒と肺がんの関係を「喫煙習慣がある」と「喫煙習慣がない」で分けて分析しなおしたところ、いずれの群も「飲酒

習慣がない」に比較して、「飲酒習慣がある」のオッズ比は 1.0 となった。この事例における喫煙の影響を取り除くために研究企画段階で実施するのが可能な対処方法はどれか、2 つ選べ。

1. 層化、
2. 制限、
3. 脱落、
4. 多変量解析、
5. マッチング

(2 と 5 が○。設問の事例は、飲酒と肺がんの関係において、喫煙が交絡因子になっている。喫煙は肺がんの原因として作用し、喫煙と飲酒は相互に関連が強い。交絡因子を研究デザインの段階で制御するのは、制限(喫煙者のみ、または非喫煙者のみで調査)、マッチング(症例の喫煙状況と同じ状況の対照を選ぶ)などであり、解析段階で制御する方法は層別分析(層化)や多変量解析である)

問題 40 下の表は、ある高齢者施設の入所者に集団発生した感染症の発生状況とその転帰(病気の進行)である。この場合の致死率(致命率)を求めよ。(答えは 2.5% ; 致死率(致命率)とは、おもに急性疾患で用いられる「発症からある一定の期間内の死亡の割合」を示す概念であり、以下のように算出する。
致死率=死亡者数÷発症者数)

入所者数	発症者数	入院者数	死亡者数
200	80	16	2

午後

問題 15 保健統計調査で正しいのはどれか。

1. 国民生活基礎調査は世帯について行う大規模な調査であり、5 年ごとに実施される。(3 年ごとに実施される)
2. 患者調査は 3 年ごとに実施される標本調査で、受療率などが含まれる。(○)
3. 死因を調査するのは医療施設調査である。(人口動態調査である)
4. 国勢調査は最も代表的な人口動態統計である。(人口静態統計である)

問題 22 陽性反応的中度を示す計算式はどれか。(1 が○。敏感度の計算式は 2、特異度の計算式は 4 である。)

	疾病あり	疾病なし	合計
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

1. $\frac{a}{a+b}$
2. $\frac{a}{a+c}$
3. $\frac{c}{c+d}$
4. $\frac{d}{b+d}$
5. $\frac{b}{b+d}$

問題 32 ある地域において循環器疾患の発生要因を検討するため、40 歳以上の男女 50000 人を対象に現在の生活習慣及び肥満度、血圧、血清脂質などの健康状態の調査を実施した。その後、この初回調査から 10 年間の対象者の循環器疾患による死亡の有無を評価し、初回調査時点での生活習慣及び健康状態との関連を検討した。このような調査のデザインはどれか。

1. 地区介入、
2. 横断研究、
3. 無作為割付け臨床試験、
4. コホート研究

(1、介入研究のうち、介入の対象が個人の場合は「無作為対照試験」であり、地域全体や特定のグループの場合は「地域介入」である。地区介入は地域住民などの集団を対象とし、疾患の予防効果を評価する研究である。2.「横断研究(断面研究)」とは、一時点における曝露と疾病の有無とを調査する研

究方法である。3、「無作為割付臨床試験」は、おもに新しい治療の評価や合併症の予防などを目的として行うものである。4. が○。「コホート研究」は、目的とする疾患（本設問では循環器疾患）に罹患していない者を対象に、曝露群と非曝露群とを設定して追跡し、疾病の発生状況を比較する。）

問題 33

問題 32 の調査において、初回調査の時点での高血圧であった者は 20000 人いた。10 年間の追跡期間中に、高血圧のなかから 50 人、非高血圧者から 50 人の脳卒中罹患を確認した。このときの高血圧の脳卒中に対する相対危険（リスク比）を求めよ。

（コホート研究における相対危険度（相対危険）とは、曝露群の罹患率と非曝露群の罹患率との比を示す。高血圧の脳卒中罹患に対する相対危険度＝高血圧者（曝露群）の罹患率÷非高血圧者（非曝露群）の罹患率である。設問の場合は、 $(50 \div 20000) \div (50 \div 30000) = 1.5$ ）

問題 34

喫煙者と非喫煙者の 10 年間の脳卒中による死亡データを表に示す。このときの喫煙の肺がん死亡に対する寄与危険を求めよ。（寄与危険度（寄与危険）とは、曝露群の死亡率（あるいは、発生率・罹患率）から非曝露群の死亡率（あるいは発生率・罹患率）を減じたものである。本設問では、まず死亡率を求める。喫煙者の 10 万人年対の死亡率は $12 \div 40000 \times 100000 = 30$ であり、非喫煙者の 10 万人年対の死亡率は $6 \div 60000 \times 100000 = 10$ である。したがって、寄与危険度は $30 - 10 = 20$ である。相対危険度は曝露群の死亡率÷非曝露群の死亡率で求められ $30 \div 10 = 3$ となる。寄与危険度割合は寄与危険度÷曝露群の死亡率で求められ $20 \div 30 = 0.667$ （66.7%）となる。）

	観察年	肺がん死亡数（人）
喫煙者	40000	12
非喫煙者	60000	6

問題 35

ある集団に対して、コホート調査を実施したところ疾病 B の発生に対する要因 A の相対危険度は 5 であった。この集団の要因 A の曝露割合は 25% である。この集団全体において要因 A の曝露を完全になくすことができた時に、この集団において疾病 B の患者発生を予防できる割合を求めよ。（この設問は人口（集団）寄与危険度割合を問う問題である。集団全体の罹患率を I 、曝露群の罹患率を I_e 、非曝露群の罹患率を I_u とする。設問のデータから、相対危険度 $5 = I_e / I_u$ より、 $I_e = 5 I_u$ 。また、 $I = 0.25 I_e + (1 - 0.25) I_u$ より $I = 0.25 \times 5 I_u + 0.75 I_u = 2 I_u$ 、人口寄与危険割合 = $(I - I_u) / I = (2 I_u - I_u) / 2 I_u = 0.5$ （50%）すなわち、人口寄与危険度割合は 50% になる。設問の人口寄与危険度割合は、肺がんと喫煙の関係に似ており、喫煙率 25% の集団で喫煙者をすべて禁煙に導けば将来集団全体の肺がんの罹患率を半分にできるという計算となる。）

次の文を読み、問題 51、問題 52 に答えよ。

調査対象者である疾病 A の患者 200 人と健康診断で異常のなかった人 400 人にみずからの母子健康手帳を持ってきてもらい、対象者を妊娠した時の喫煙の有無を調べた。疾病 A の患者では、75 人の母親が喫煙しており、健康診断で異常のなかった人では、80 人の喫煙としていた。

問題 51 妊娠時の女性の喫煙による疾患 A 発症リスクについて、最も近い値はどれですか。

1. 0.4、 2. 0.6、 3. 1.7、 4. 2.4

（この設問の調査方法は、症例対照調査である。したがって、次の表からオッズ比を計算することにな

る。

	曝露あり＝喫煙	曝露なし＝喫煙せず	計
疾病発症	75	125	200
異常なし	80	320	400

すなわち、オッズ比は $(75 \times 320) \div (80 \times 125) = 2.4$ となる。コホート調査と間違えると、相対危険度として次のように計算し、3. を正答としてしまう。 $75/155 \div 125/445 = 1.72$ 。また、分子と分母を間違えると 1.、2. の解答になってしまう。）

問題 52 妊娠時の女性が 1 人も喫煙しなかったと仮定した場合、本調査の疾病 A 患者のうち何%が罹患を免れたと考えられるか。なお、この問題では、相対危険度とオッズ比は相互に近似できると仮定する。

1. 18%、2. 41%、3. 47%、4. 58%、5. 100%

（寄与危険度割合を問う設問である。曝露群の罹患率を I_e 、非曝露群の罹患率を I_u とすると、相対危険度は I_e / I_u 、寄与危険度割合 $= (I_e - I_u) / I_e$ であるから、この式を変形すると、 $= 1 - I_u / I_e = 1 - 1 / \text{相対危険度}$ 、相対危険度とオッズ比は近似できるので、問題 51 で求めたオッズ比を相対危険度として代入すると、 $1 - 1/2.4 \div 0.583 = 58.3\%$ になる）

次の文を読み、問題 53、問題 54、問題 55 に答えよ。

2008 年 4 月からメタボリックシンドロームの予防を主目的とした特定健康診査・特定保健指導が開始された。A 市に勤務する保健師は、特定健康診査でウエスト周囲径 85 cm 以上であった男性を対象に、腹囲減少に対する測定用具貸出しの効果を評価するため介入研究を実施した。介入期間は 6 か月で、参加協力の得られた 40 人からインフォームドコンセントを得た後、無作為に介入群（20 人）と対照群（20 人）に振り分けた。そして、対照群には通常の特定保健指導を実施し、介入群には特定保健指導に加えて万歩計の貸出を行った。介入群、対照群とも脱落者はおらず、また、介入研究開始時の介入群と対照群のウエスト周囲径に有意な差はなかった。介入研究開始時と終了時のウエスト周囲径は以下のとおりである。

	介入群（n＝20）	対照群（n＝20）
ウエスト周囲径；cm		
介入研究開始時	88.5±2.3	88.4±2.2
介入研究終了時	84.1±3.8	88.3±2.5

問題 53 このような介入研究の方法で適切なのはどれか。

1. 各被験者を介入群・対照群のいずれに振り分けるかは、本人の希望により決定した。
2. 特定保健指導では 6 か月後の面談を義務付けていないため、介入終了後のウエスト周囲径は自己測定とした。
3. 特定健康診断時のウエスト周囲径を介入前の値として用いた。
4. 対照群に割り当てられた被験者の不利益を減じるため、調査終了後、希望者に 6 か月間の万歩計と体重計の貸出を行った。

（1. は選択バイアスが生じやすく誤り。2. は介入前後の評価は同一の、標準化された方法で行う必要がある。3. 対象者選定に用いた特定健康診断時の値ではなく、介入前に再測定することが望ましい。4. ○。介入研究を実施する際には介入の有無による不利益に留意し、可能な限りそれを小さくする努

力が求められる。)

問題 54 個別指導の効果を検討するため、介入群と対照群の 2 群間のウエスト周囲径の変化量の平均を比較したいと考えた。最も適切と考えられる検定方法はどれか。

1. χ^2 検定、2. 対応のない t 検定、3. 対応のある t 検定、4. フィッシャーの正確確率検定
(1. χ^2 検定は頻度の比較の場合に用いられる。2. ○。対応のない t 検定では、設問の場合、両群の個々の被験者についてウエスト周囲径の変化量(介入後の測定値－介入前の測定値)を求めて、その平均の群間差を検討する。3. 対応のある t 検定は、同じ対象者における連続量の比較(反復測定)などの場合に用いる。4. フィッシャーの正確確率検定は、頻度の比較(クロス集計で期待値 5 未満のセルが存在する場合)に用いる。)

問題 55 介入群と対照群の 2 群間のウエスト周囲径の変化量の平均を比較した結果、介入群のウエスト周囲径の変化量は対照群よりも大きく、この差は 5%水準で有意であった。結果の解釈として正しいのはどれか。

1. 介入群でのみ、指導後にウエスト周囲径が 5%減少した。
2. 「介入の効果があつた」との判断が誤っている可能性は 5%よりも小さい。
3. 他の集団にこの介入方法を適用した場合、5%の確率でウエスト周囲径減少の効果がかる。
4. 介入群と対照群の介入後におけるウエスト周囲径の差が 5%であった。
(2. が○。統計学的に 5%水準で有意差があるとは、観測された 2 群間の頻度・測定値の差が帰無仮説(2 群間には差がないとする仮説)のもとで偶然に生じる確率が 5%未満であるという意味である)。

第 100 回保健師国家試験問題 (2014.02 実施)

午前

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。
(正解は 3。ヒストグラムは 1 つの連続量の度数分布を表した柱状のグラフであり、横軸に階級をとり、縦軸に階級ごとの頻度を示す。)

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
(正解は 3。1. 国勢調査は我が国の人口静態統計を得るために、日本に住むすべての人と世帯を対象として 5 年に 1 度実施される。調査内容は、性、年齢、職業、婚姻状況、国籍、居住状況、収入、教育

課程である。2. 国勢調査がある一時点での統計であるのに対し、人口動態調査は、「戸籍法」および「死産の届け出に関する規程」により届けられた出生、死亡、死産、婚姻、離婚に関する個票に基づき集計される業務統計（業務データを集計することにより作成される統計のこと）である。転入出は、人口動態調査で取り扱う業務データではない。3. ○。社会生活基本調査は、基幹統計（統計法に基づき、国が実施する基本的で重要な調査。本問題の選択肢1～4はすべて基幹統計として指定されている）の1つである。本調査の結果は、1日の生活時間の配分や余暇時間におけるおもな活動の状況などの国民の社会生活の実態を明らかにするための基礎資料となる。4. 国民生活基礎調査は保健、医療、福祉、年金、所得などの国民生活の基礎的事項を調査するものである。調査項目は傷病頻度、自覚症状の頻度、受診状況、就業状況、傷病の影響、医療保険加入状況などである。）

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。

（2つの連続量間直線的（比例的）関係を表したものが相関である。正解は3.）

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

（寄与危険度は、疫学における指標の1つで、曝露群と非曝露群における疾病の頻度の差である。曝露群の発生率から非曝露群の発生率を引いたものが寄与危険度となり、曝露効果の強さを示すことができる。正解は3.。 $0.6 - 0.3 = 0.3$ である。）

午後

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 型の人がある確率を知りたい。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ものはどれか。

1. t 分布、2. F 分布、3. 正規分布、4. 二項分布

（正解は4.。）

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。

(正解は1.、2.と4.。レコードリンケージとは、同一の対象者に対して実施された異なる統計調査のデータ同士をキーとなる共通項目を手掛かりに結合させ利用するもので、同じ形式のデータに限る必要はない。)

授業概要：

コンピュータ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。

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：

保健師過去問の提示、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、疫学と統計Ⅰ、疫学と統計Ⅱ、疫学と統計Ⅲ、保健師過去問の解答と解説、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ます。
- * 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。

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す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、どの人に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。

（補遺） 疫学研究のデザイン

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。

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（過去問）（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（過去問）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（過去問）；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（過去問）を経て）疾病をもたない人（対照；過去問）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（**過去問**）（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 **過去問**）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。

標本抽出法の補遺

全数調査、母集団、標本、サンプリング、標本調査、無作為抽出、乱数表、単純無作為抽出法（**過去問**）、系統抽出法（**過去問**）（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）、多段抽出法（**過去問**）（たとえば、まず都道府県を無作為に抽出し、次に市区町村を無作為に抽出し、その選ばれた集団単位の中から無作為に標本を選ぶ。）、層化抽出法（**過去問**）（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）

ところで、断面調査、前向き調査、後ろ向き調査は、標本調査（母集団から取り出された一部が標本、標本を取り出す操作をサンプリングという）なのか？

断面調査は通常、標本調査となる。前向き調査、後ろ向き調査は、無作為抽出した標本からの調査ではない場合が多い。前向き調査は、特定地域住民などの全数調査であることがある。また後ろ向き調査は無作為ではなく、恣意的に選んだ対象者での調査である場合がある。そのため、後ろ向き調査では母集団を特定することが難しい場合があり、推定や検定などを行っても、妥当でないこともありうる。

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 2 cm 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 身長の数値分布・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 cm、最小値 158.0 cm 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 2.5 c m) 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 とすると身長の数値分布表は表 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

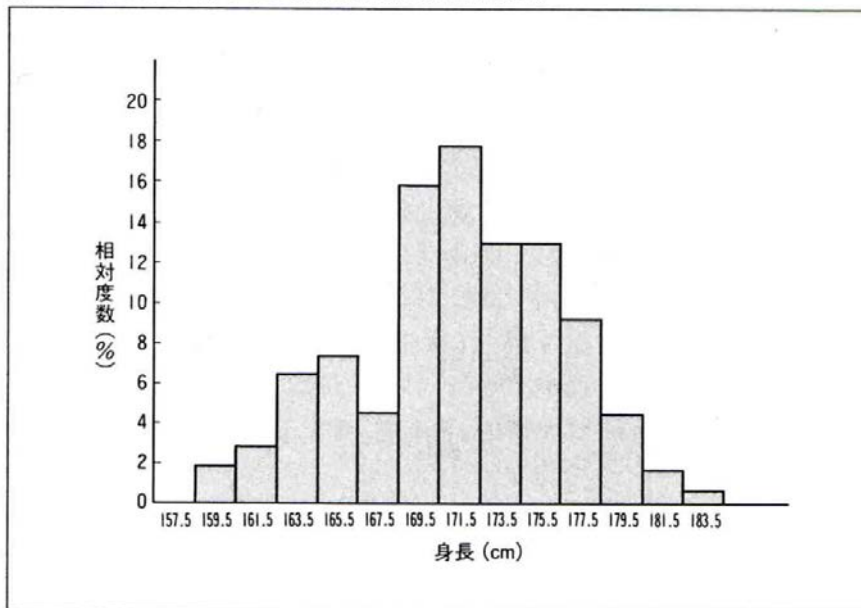
ある会社の従業員男女それぞれ 10 人を対象に貧血検査（ヘモグロビン濃度 mg/dl）を行い、下記の
結果を得た。

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 身長の高ヒストグラム(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、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、利用者が間違えおそれがある。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

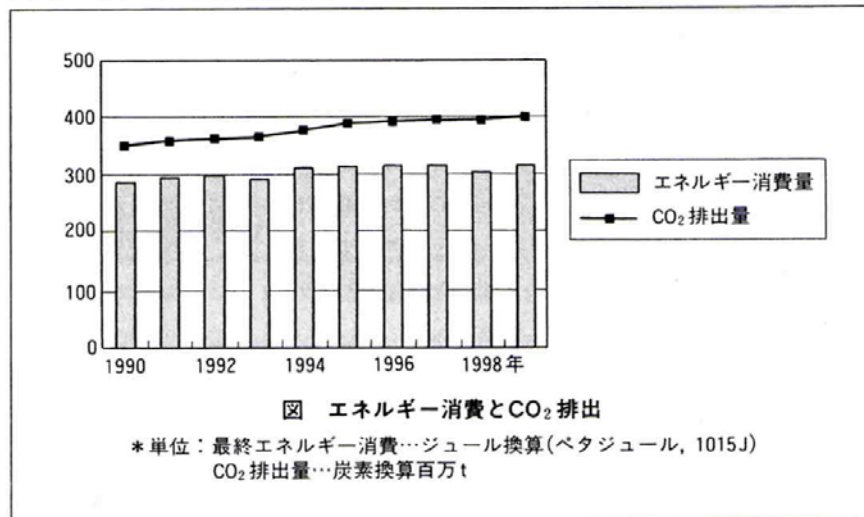


図 16 糖尿病有病死亡者の百分率分布(方波見ら)

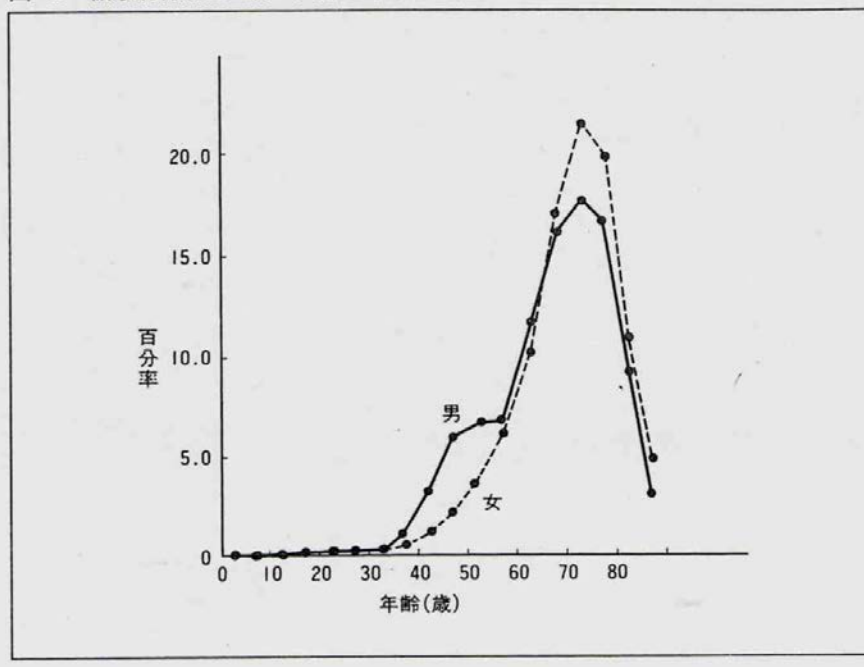


図 17 調査時の健康状態(未受診者)

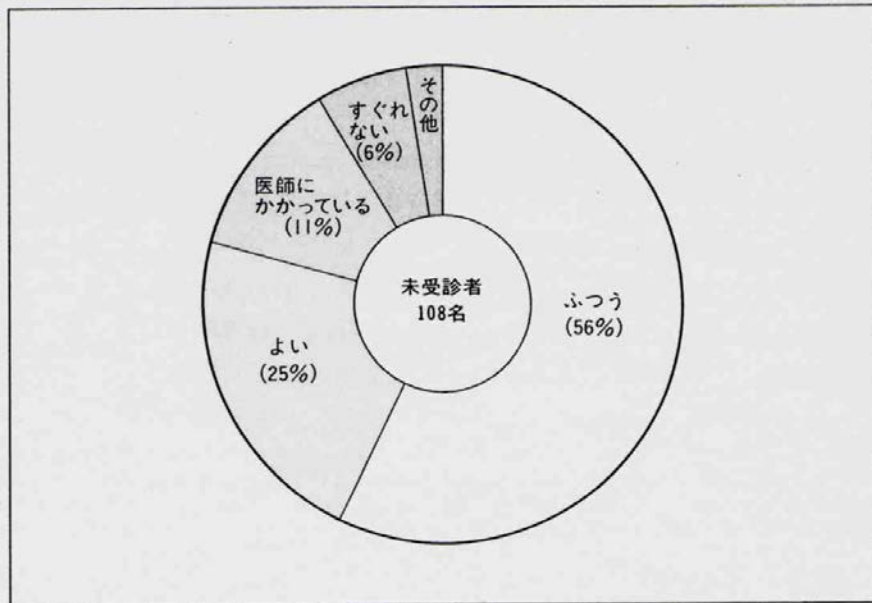


図 18 死因群別死亡割合の年次比較

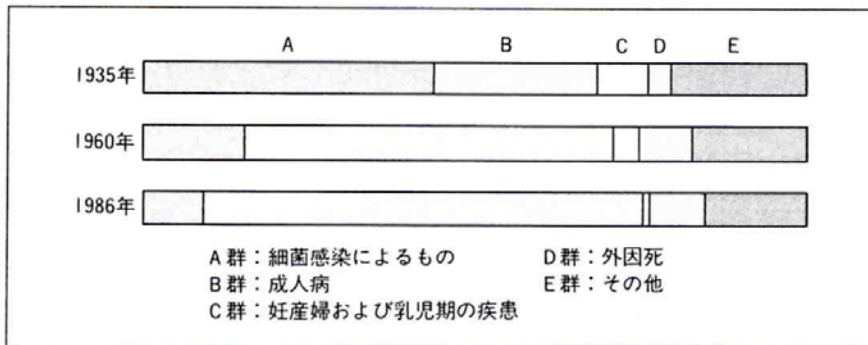


図 20 1998 年における世界のポリオ発生状況(WHO 資料による)

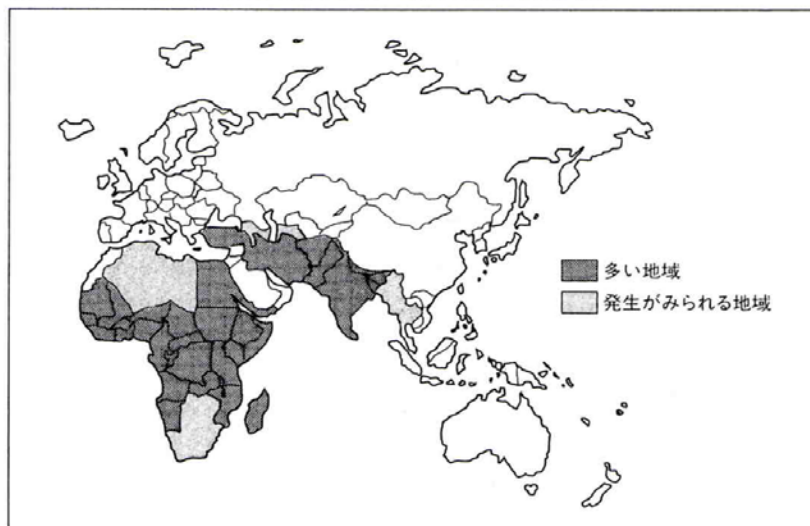
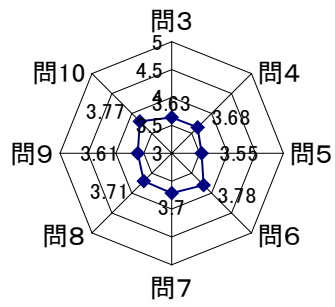


図 19 レーダーチャートの一例 (一般学生と長期履修学生との授業評価の違い (平成 17 年度))

一般学生(前期)



長期履修学生(前期)

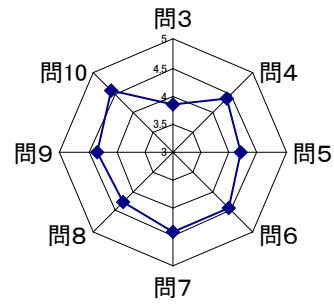
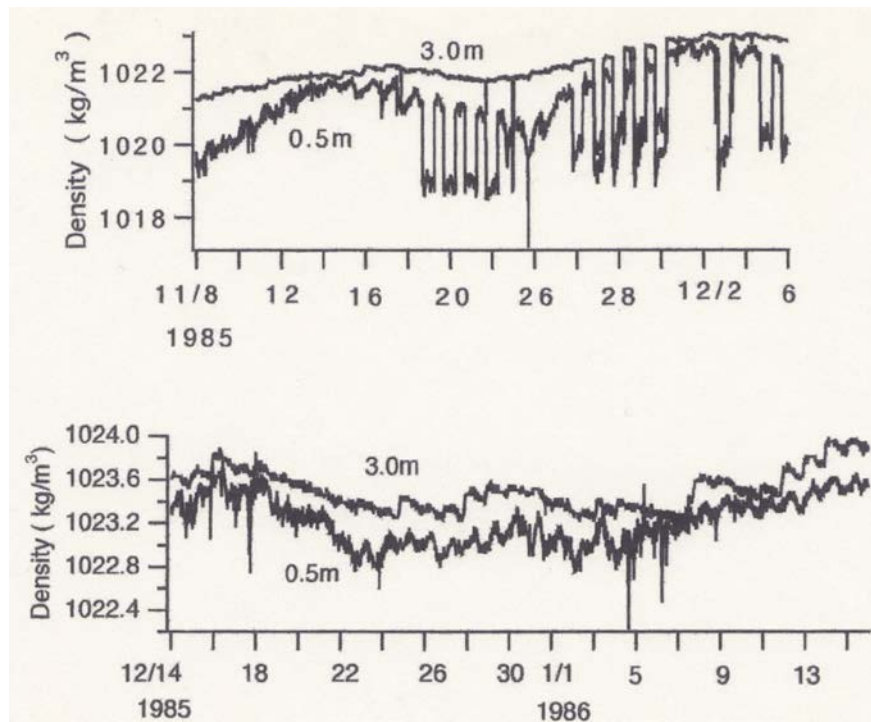


図 21 密度の時間変化 (明神池 1985 年～1986 年)

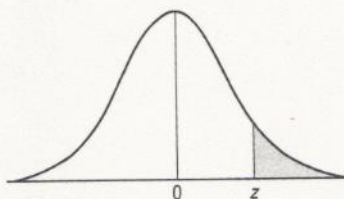


付表 1 ギリシア文字

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, s	シグマ
T	τ	タウ
Y	υ	ユブシロン
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

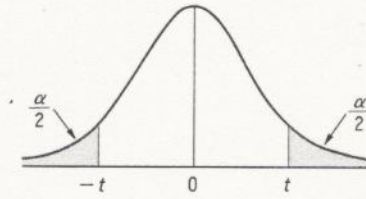
z から外側(色の部分)の面積である。
 z が負の場合も、対称性から同面積である。



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

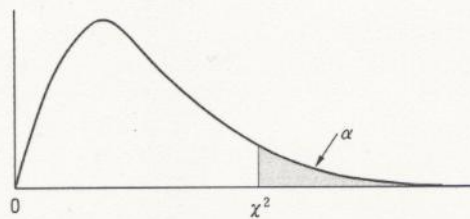
t , $-t$ の外側 (色の部分) の面積を加えて α となるが, そのときの自由度 (ν) と t の値である。



$\nu \backslash \alpha$	.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。



α	0.990	0.950	0.100	0.050	0.010	0.001
ν						
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 $\bar{x}$ の文字の上に一をつけて（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 $x_1, x_2, x_3, \dots, x_n$ とすると、

$\bar{x}$ は

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

Σ はギリシア文字シグマの大文字で $\sum x_i$ は x_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。

x_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、

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

は x_i がとりうるすべての値を加算する ($x_1 + x_2 + x_3 + \dots + x_n$) という意味となります。

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 $= (a+b+c) \div 3$ 、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 $\geq$ 幾何平均 $\geq$ 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。

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

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

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

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

＊ 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。

問題 3

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。

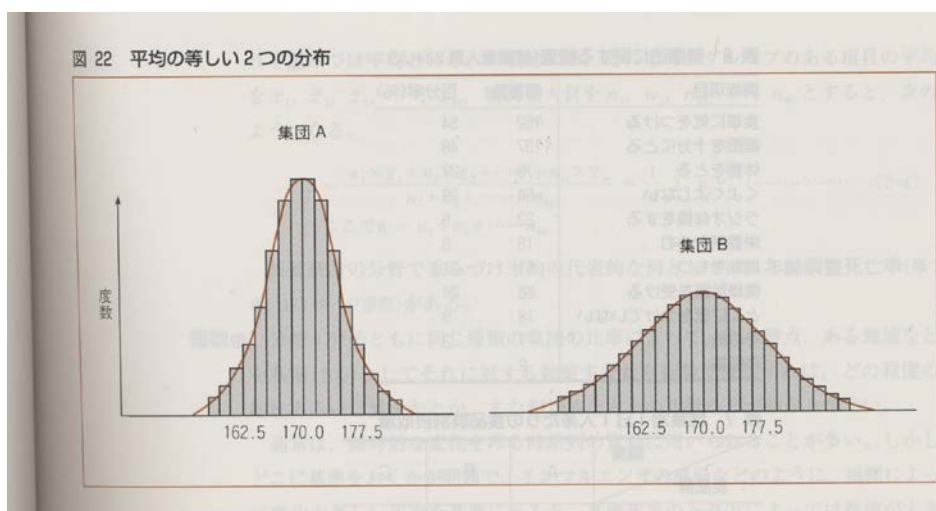


図 22 の場合、集団 A (500 人とする)「 170.5 ± 7.0 」cm、集団 B (500 人とする)「 170.5 ± 15.0 」cm 同じ平均値であるが、A は B よりもばらつきが小さい、A は B よりも標準偏差が小さい、A の個々の測定値が B よりも平均値に近いことを示しています。

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

$$\bar{X} = \frac{7.4 \times 300 + 4.8 \times 100 + 3.2 \times 50}{300 + 100 + 50} = 6.4$$

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

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

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

ここで、 $n = n_1 + n_2 + \dots + n_m$

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) ;

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

- (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)

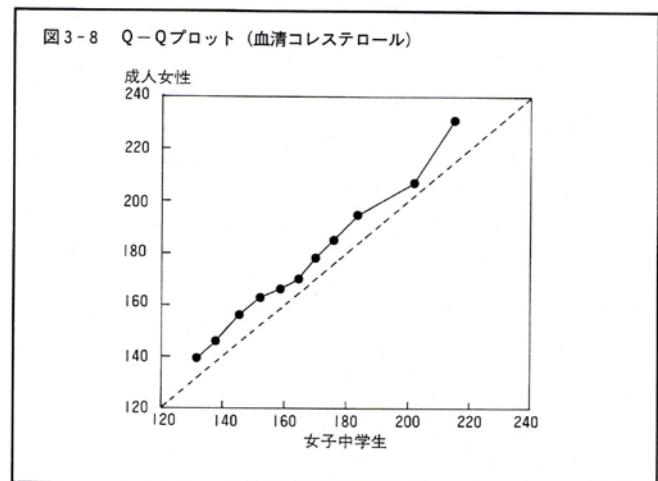
また、中央値からの第 1、第 3 四分位数の偏差を求めて平均した四分位偏差 (quartile deviation) も散布度の指標となる。

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を図であるが、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。

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。

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)

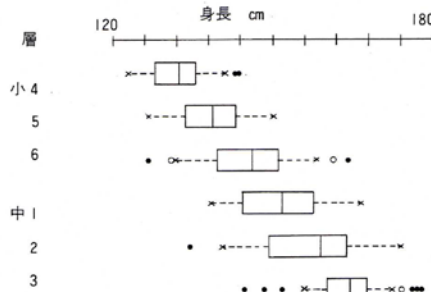


図 3-9 Q-Qプロット (架空例)

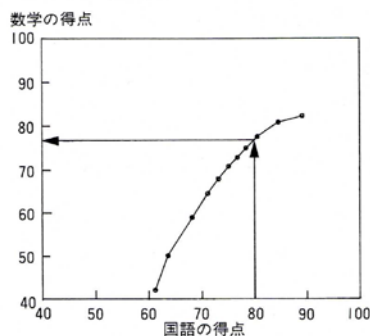
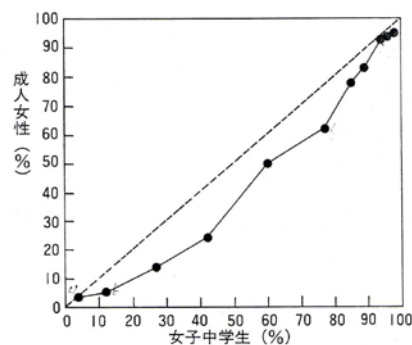


図 3-10 コレステロール値のP-Pプロット



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るならば、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、 $-\infty \sim +\infty$ で 1

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

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

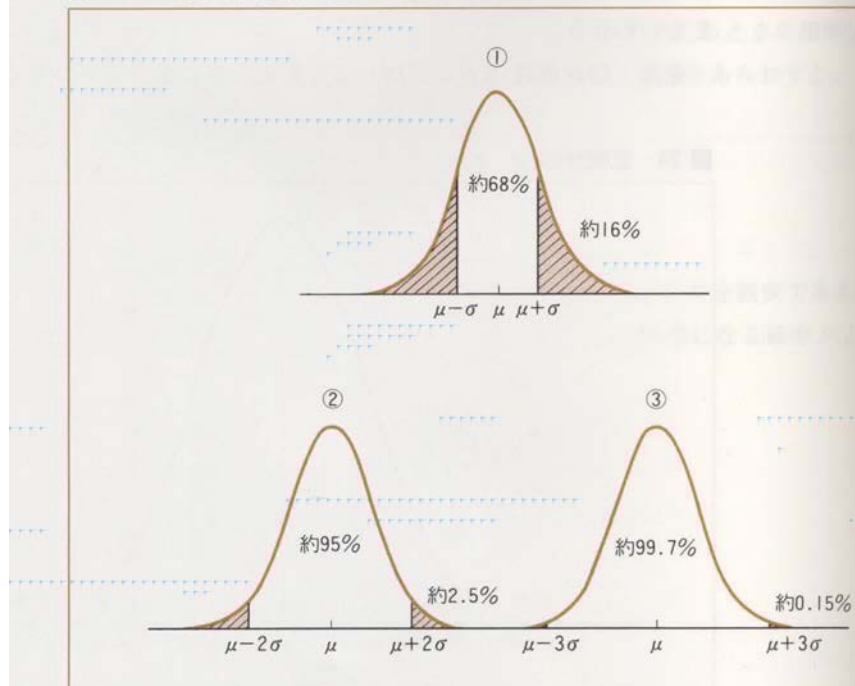
$\mu - 3\sigma$ と $\mu + 3\sigma$ の間の正規分布曲線的面積は全体の約 99.7%

$\mu - 4\sigma$ と $\mu + 4\sigma$ の間の正規分布曲線的面積は全体の約 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 $N(\mu, \sigma^2)$ 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

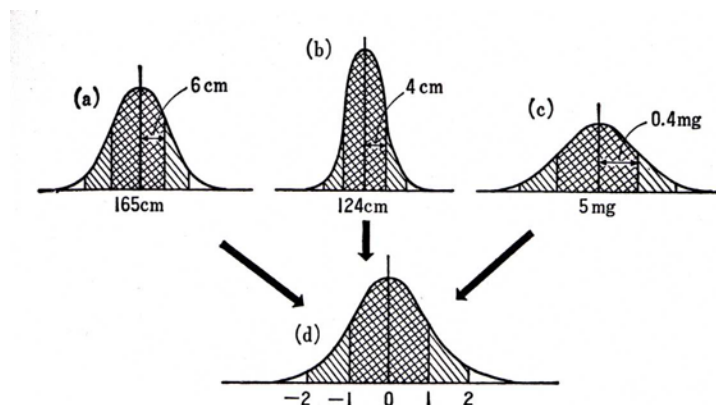


2) 正規分布の標準化

先程の例で平均値よりも大きく、175 cm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。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 cm、標準偏差は 6 cm、(b) は小学 3 年生の身長で平均値は 124 cm、標準偏差は 4 cm、(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 mg、標準偏差は 0.4 mg 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 cm のところを 0 とみなし、横軸の目盛を 6 cm 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 cm のところが 0 になり、171 cm のところが 1 になり、177 cm のところが 2 に、159 cm 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 cm 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、教えてください。

このような平均 0、標準偏差が 1 の正規分布を標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、 $\mu = 156$ cm、 $\sigma = 5$ cm の身長分布で 169 cm 以上の人の割合を求めたいとします。

この場合まず、169 cm 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 150 cm 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 と同じことになります。

(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、正規分布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くと、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 x_1 、 x_2 、 x_3 、.....、 x_n を選んだとき、標本平均

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

の分布は、 n が大きくなると正規分布 $N(\mu, \sigma^2/n)$ 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）について教えてください。

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、 $\bar{x}$ 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。

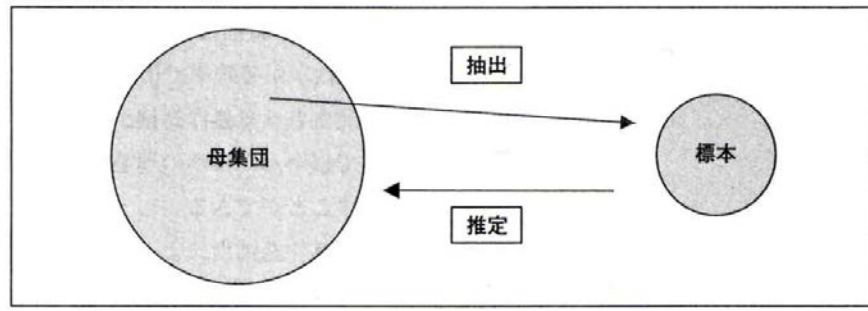
そこで、この場合には全体の一部のみから全体を知ろうとする。これが推定です。

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

図 34 標準平均 $\bar{x}$ の分布

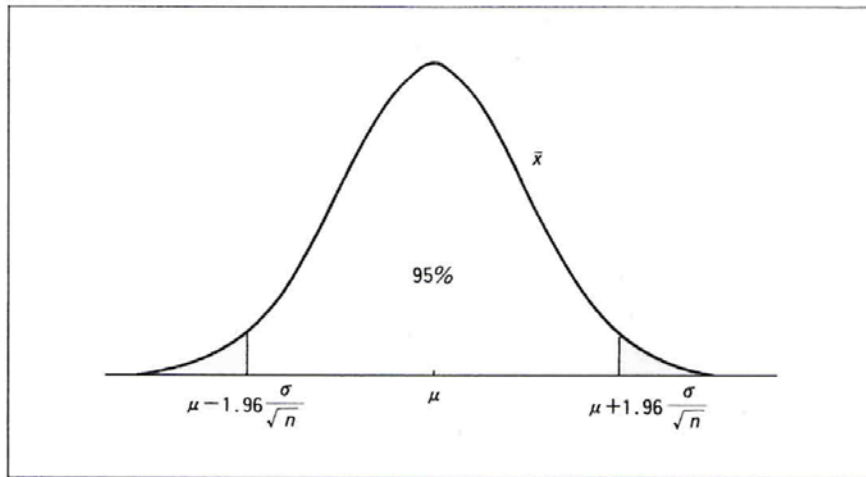


図 34 の標準平均と記載されていますが、これは誤りで、正解は標本平均です。

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 ($\nu = n - 1$) 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 $\nu = 5$) 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、 $\nu = 5$ の場合、 μ の95%信頼区間は、

$$\bar{X} - 2.571 \frac{S}{\sqrt{n}} < \mu < \bar{X} + 2.571 \frac{S}{\sqrt{n}} \quad \text{で}$$

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

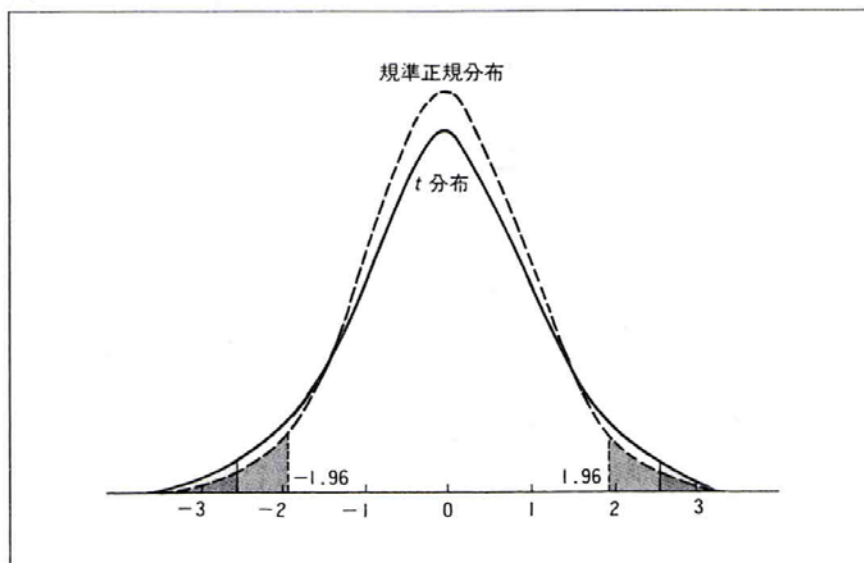
によって決まるもので、付表5の t 分布表から決定される。

$\nu = 5$ 、信頼区間が95%では $t_0 = 2.571$ 、 $\nu = 10$ 、信頼区間が95%では $t_0 = 2.228$

$\nu = 5$ 、信頼区間が99%では $t_0 = 4.032$ 、 $\nu = 10$ 、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 標準正規分布と 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 172.1 cm、標本標準偏差は 4.15 cm 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 $\nu = n - 1$ 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 cm、標準偏差は3.50 cm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ければ、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； $A \neq B$ 」

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、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になります。

両側検定の方が検定が厳密である。

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とも「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歳女子の身長平均と標準偏差は、それぞれ $\mu = 140\text{cm}$ 、 $\sigma = 5\text{cm}$ 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 $H_0; \mu = 140$) 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 $\bar{X} = 137\text{cm}$ 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。

$$z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} = \frac{137 - 140}{5 / \sqrt{25}} = -3 \quad \text{付表4から } z = 3.0 \text{ の面積は } 0.0013 \text{ となる。}$$

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 $H_0; \mu = 157.5$ 、対立仮説 $H_1; \mu \neq 157.5$

(2)検定統計量を求める;この場合、15人の身長の平均 $\bar{X} = 162.0\text{cm}$ 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。

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

自由度 $\nu = n - 1 = 15 - 1 = 14$ 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、 $t = \text{差の平均値} / \text{差の標準誤差}$ 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、式立てをしてみてください。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; $x_1, x_2, x_3, \dots, x_n$ 夏の血圧値; $y_1, y_2, y_3, \dots, y_n$

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

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。

$$t = \frac{\bar{z} - \mu_z}{\frac{s_z}{\sqrt{n}}} \quad \text{ここで、} t \text{ は自由度}(n-1)\text{の}t\text{分布である。}$$

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 $\mu_x \neq \mu_y$ ($\mu_z \neq 0$) 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

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

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（マンローホイトニー法）、コロモゴロフスミルノフの検定、

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。

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リック法を用いる。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) (過去問)、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 (過去問) 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。

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。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。

$$y = bx + a \quad (\text{この式の場合、} y \text{ が従属変数、} x \text{ が独立変数になる})$$

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。

多変量解析 (過去問)

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 時点をとっていても、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 変量のもつ情報のある程度表現しようとしたもの。

因子分析は、かす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。典型的なもの、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、その距離の算出方法は無数。代表的なものは、重みつきクークリッド距離、ミンコフスキーの距離、マハラノビスの汎距離など。医学、看護学の分野では、人の性格特性の分類から、疾病の分類、さらには DNA の分類に至るまで様々な分野で用いられている。

疫学と統計学

A 疫学の考え方

いわゆる疫病の原因を探るために発達した医学・健康科学の一分野に疫学（epidemiology）という領域がある。近年は、慢性非感染疾患（糖尿病、心筋梗塞、がんなど）で死亡する人が多いため、疫学の取り扱う領域も慢性非感染性疾患が主流となっている。

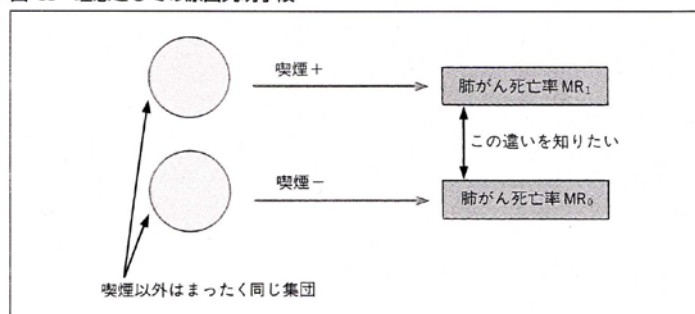
また、近年になって医療は科学的根拠に基づいて（evidence-based medicine、EBM）行われなくてはならないという考え方が広まってきている。この EBM において科学的根拠を与えるための手法としても、疫学が用いられている。

このように非常に重要な位置を占める疫学も統計学と密接に関連している。ここでは、いわゆる「統計学」の枠を多少こえるが、看護学生によって将来必要となってくるとされる疫学について概説する。

① 疫学とは

疫学とは、人間集団を対象にして疾病あるいは健康現象の原因をさぐる科学の一分野である。「人間集団を対象にする」ということには利点と欠点が存在する。実際には不可能であっても、どのような方法が理想的かをまず考えてみよう。喫煙によって肺がんが増加するかどうか調べる場合に、図 53 のように全く同じ集団

図 53 理想としての原因究明手段



が2つあると考える。この一方のすべてに喫煙させ（曝露群：曝露とはここでは有害物質や病原菌などにさらされることを指している）、もう一方には喫煙させない（非曝露群）で数十年間集団を観察する。そして曝露群の肺がん死亡率 MR_1 と非曝露群の肺がん死亡率 MR_0 とを比較することにより、喫煙が肺がんの死亡率を増加させるかどうかを決定することができる。しかしながら、まったく同じ集団というのは空想科学小説でなければありえない。また、人間集団であれば、健康を害すると思われる原因に関して、一方を曝露群にして、他方を非曝露群にするなどということは倫理的に大きな問題である。この問題を避けるために、実験的方法と疫学的方法では異なるアプローチをとる（実験的な疫学的方法もあるし、細胞を用いた実験などもあるが、ここでは取り上げない）。

実験的方法

実験的方法では、倫理的な問題を避けて実験動物を用いる。また、曝露群と非曝露群との振り分けを無作為に行うこと（random assignment）によって統計的に2群を等質にする。これらの作業によって、理想に近い状態で曝露の影響を評価することができる。実験的方法の問題は、①動物と人間とは異なること、②一般に実験に用いられる動物の数が少ないこと、などである。①の例としては、人間ではがんがおこらないのにマウスでは発生する場合がある。それでも動物実験を行うのは、共通の生理的メカニズムを持つ部分も多いため、共通の要因で共通の影響がみられる、と考えてのことである。②の例として、20匹を10匹ずつの曝露群と非曝露群に分け、曝露群から4匹、非曝露群から1匹に疾病の発生が

確認されたとする。このときフィッシャーの直接確率は 0.3 となり、一般に用いられる有意水準の 0.05 よりもずっと大きい。このように、曝露によって発症が 4 倍になっても有意な結果が得られない。倍の実験動物（40 匹）を用いて、やっと p 値が 0.06 程度と、一般によく用いられる有意水準に近くなる。

疫学的方法

疫学的方法では、多くの場合、倫理的な問題から曝露群、非曝露群を研究者が決定することはできない。そのため無作為割り付けができない。喫煙の例でいえば、ある集団で喫煙している人を集めて曝露群とし、喫煙していない人を非曝露群とする。このことは、両群がほぼ等質であるという前提がおけないことを意味する。大きな欠点である。一方、疫学的方法では人間集団そのものを対象としているために、動物実験のような人間とのギャップという問題は存在しない。また、多くの場合、対象とする集団は数百人から数十万人の規模に及び、小さな相違でも検出可能なことが多い。これらは大きな利点であり、このために化学物質の発がん性を格付けする国際機関においても疫学的研究は実験研究以上に重要であると考えられている。

ここ数十年の近代疫学の歴史は、無作為割り付けの不可能な人間集団において、いかに妥当な結論を導くかに関する理論発達の世界史であった。以下にその初歩に関する概要を述べる。

② 原因論（ロスマン Rothman のモデル）

たとえば喫煙が、肺がんの原因であるかどうかを考えると、「喫煙者でも肺がんにならない人が多い。だから喫煙は肺がんの原因ではない。」とか「喫煙しなくても肺がんになる人がいる。だから喫煙は肺がんの原因ではない。」とかいわれることがある。これは正しいだろうか。

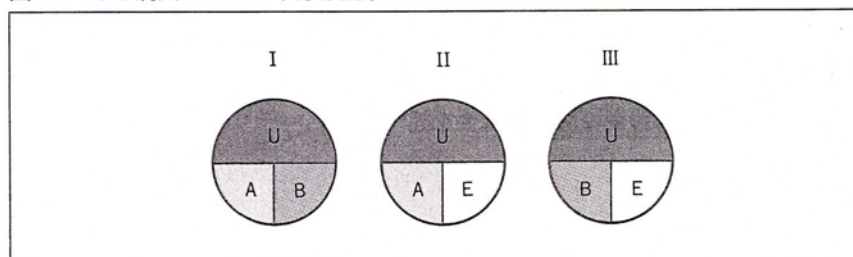
日本脳炎は、コガタアカイエカが媒介して日本脳炎ウイルスによっておこる。しかし、ウイルスがコガタアカイエカによって体内に注入された人のうち、発病するのは 1% にも満たないといわれている。喫煙者でも肺がんにならない人がいるから喫煙を肺がんの原因でないとすれば、日本脳炎もウイルスによっておこると言えなくなってしまう。一方、最近問題となっているスギ花粉症では、アレルギー性鼻炎症状をおこす。アレルギー性鼻炎を考えたとき、カモガヤもブタクサでも症状のおこることがある。つまり、スギ花粉の存在しないところでもアレルギー性鼻炎は存在する。しかし、スギ花粉はアレルギー性鼻炎の原因でないと考える者はいない。

これらこのことから明らかなことは、一般的にある疾病の原因は数種類あり、また、ある 1 つの要因だけで疾病を発症することはないということである。この多要因の原因論を疫学で使えるようにモデル化したのがロスマン Rothman である。

図 54 の U、A、B、E はそれぞれの原因となる要素を示し、I、II、III がそれぞれ十分原因をあらわす。十分原因 sufficient cause とは、幾つかの要因から成り立ち、そのすべてがそろった時点で疾病が発生するものを言う。図 54

の十分原因 I では、U と B をすでに持っている人に A が加わると疾病がおこると考える。たとえば、疾病が肺がん、A が「1 日 20 本の紙巻きタバコ喫煙 20 年間」という場合を考える。U と B をもっている人は、A によって十分原因 I が完成し、肺がんを発症してしまう。

図 54 ある病気の 3 つの十分原因 (Rothman & Greenland, 1998)



図の例は単純化してあり、肺がんの原因がわずか3つの要因で構成されているなどということはない。しかし、いくらその要因の数が多かろうと、考え方は同様である。つまり、たとえば、ある集団で起こった肺がんのうち、Ⅰで起こった人の割合が20%、Ⅱで起こった人の割合が50%、Ⅲで起こった人の割合が30%とすると、これら肺がんのうち、Aすなわち、「1日20本の紙巻きタバコ喫煙20年間」、という条件をはずす。つまり禁煙させると十分原因ⅠとⅡが成立しなくなり、70%が防げたであろうことが分かる。このように、取り除くことで疾病が予防できる（十分原因が完成しなくなる）要因を、一般に原因とよぼうというのである。

疫学は、疾病を予防するために発達してきたので、この考えは有用である。

B 疫学の方法論

疫学とは、前述したように人間集団における疾病発生に関する1つの科学である。人間集団全てを調査することは困難であるから、疫学の指標や手法は統計学のそれを援用することが多い。ここでは、従来の統計学の教科書では古くなってしまった部分や記述されていない部分のうち、特に重要と思われる基礎的事項について解説する。

① 割合・危険度・率

これらについてはエランドト・ジョンソンとジョンソン（Elandt-Johnson and Johnson：1986年）が明快な定義を与えているので、まずそれを紹介し、あわせてそれらの意義について述べる。

比 (ratio)： 広くは、ある数を別の数で割った結果であるが、科学の領域では、通常2つの互いに排他的（mutually exclusive）な部分を比較するために用いられる。たとえば、ある集団の性比は次の式であらわされる。性比 = (男性の数) / (女性の数)

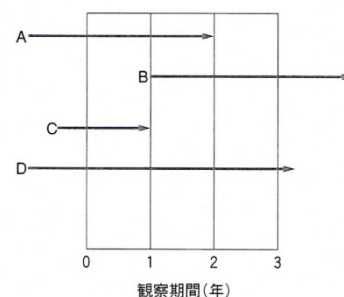
割合 (proportion)： 割合は広義の比の一種で、分子が分母の一部分のものである。一般には次の式であらわされる。 $p = a / (a + b)$ 百分比は、この割合を100倍したものである。衛生統計でよく用いられる有病率（prevalence）は、ある集団のうちで、ある疾病を持っている者の割合である。割合は0以上1以下で、単位は無名数である。

危険度 (リスク、risk)： 死亡危険度は、死亡の危険度を割合のかたちであらわしたものである。たとえば、ある集団10,000人のうち、3年間で10人の肺がん患者が死亡したとすれば、この集団における3年間の肺がんによる死亡リスクは0.001となる。

率 (rate)： 率とは、たとえば自動車の速度のように、ある量（時間）の単位あたり、もう1つの量（走行距離）がどの程度変化したかを示すものである。

死亡率 (mortality rate)： 自動車の速度は絶対的な率であるが、死亡率は集団の大きさを補正した相対的な率であり、人の死にゆく速さの尺度である。具体的には、図55で説明するが、式であらわすと次のようになる。死亡率 = 死亡者数 / 人時間

なお、人時間（person-time）とは、集団の構成員の延べ生存時間という。通常は人年（person-years）が用いられる。死亡率の次元は「時間⁻¹」である。これから明らかなように、死亡率が負になることはないが、1よりも大きくなることもある。よく人口10万人対死亡率というが、これは「死亡数 / (10万人)」のことである。したがって、たとえば、100



万人の集団を約 0.1 年観察して人年が 10 万のときに、このうち 20 万人が死亡すれば、人口 10 万人対死亡率は 20 万ということになる。

図 55 簡略化した職業疫学の死亡率の計算例：

これは、ある工場の例である。横軸が観察期間で、ここでは 3 年である。矢印の先端は死亡を示す。A 氏；観察を開始する前からこの工場で働いており、観察開始後、2 年で死亡した。A 氏の人年は 2 である。B 氏；観察開始 1 年後にこの工場に採用された。観察期間中は 2 人年の貢献をし、観察終了後の時点では生存していた。C 氏；観察開始前から勤務していたが、1 人年の貢献ののち観察期間中に死亡した。D 氏；観察開始前から勤務しており、3 人年の貢献をして観察終了時点でなお生存していた。この 4 人が研究対象であるとする、分子としての死亡数は 2 人、人年は、 $2+2+1+3=8$ となる。よって死亡率は、次のようになる。 $2 \text{ (人)} \div 8 \text{ (人年)} = 0.25/\text{年}$

割合と率との相違：1990（平成 2）年の日本における悪性新生物による死亡者数は 217,413 人で、同年の日本の人口は 122,721,397 人（これは厳密には年央人口であるが、便宜上年初人口と考えよう）であるから、1990 年の 1 年間の悪性新生物による死亡危険度は次のようになる。 $217,413/122,721,397 = 0.001772$

死亡による人口の減少は日本人総人口の 1%未満、出生による増加も 1%前後であるから、すべての人が 1 年間生きているとしても誤差は小さい。したがって、近似値として人年は $122,721,397 \times 1 = 122,721,397$ で、結局、死亡率 $= 0.001772/\text{年}$ となり、単位は異なるものの、値としては死亡危険度と同じになる。

では、なぜわざわざ死亡率を求めるのであろうか。理由の 1 つは図 55 で述べた。もう 1 つ極端な例であるが、2001（平成 13 年）年生まれの日本人集団もソマリアの集団も、200 年間の死亡危険度は 1.0 であるが、子どもたちが飢餓や感染症などによって多数死亡していくソマリアの死亡率よりも、日本の死亡率の方が著しく低く、死亡率においては日本の方がはるかに望ましい。つまり危険度は、一定の期間を区切ったなら、その期間内であればいつ死亡しても同じとみなすが、死亡率は各人の生存期間に応じて決定されるのである。疫学は、一般に生存期間を延長することを目的としているので、率の方がよい指標といえる。

一方、奇形の発生に関する調査では、出生時という一時点での奇形の存在の有無が問題であり、時間とは関係ない。このような場合には危険度を用いて解析を行う。

なお、上に述べた死亡率と危険度とを合わせて、広義の危険度と呼ぶこともあり、やや古い教科書では死亡割合を死亡率と書いてあることもある。しかし、今後、疫学的研究を行う際には両者を混同することは許されない。

② 曝露の指標

疫学は、疾病をおこしやすくする要因、あるいは逆に疾病に罹患しにくくする要因を選び出そうとするものであるから、その要因の関与の度合いをあらわす指標が必要になる。

人間は環境要因への曝露 exposure を受け、その用量（体内での作用量 dose）に応じて発病したり、死亡したりする。そこで、化学物質などであれば、たとえば工場での濃度が曝露濃度として測定され、そこで働く工員の血液検査などによって用量が明らかにできることがある。しかし、用量は量的に測定可能なものばかりではなく、本来量的であるが測定の不可能なものや、測定しにくいものなどもある。量的に個人レベルでの用量が測定できるのが理想的であるが、それができない場合には曝露量を、またそれでもできなければ曝露量の相対的な測定を可能にするような方法を探ることになる。曝露を受ける職

場での勤務日数で職業上の曝露量を代用するというのは、その例である。

表 27 に、職業疫学分野における曝露情報の種類を示した。

表 27 職業疫学分野における曝露情報の種類

データの種類	関与量への近似度
1. 個人的な量的測定値	最良
2. 部署・業務別の量的測定値	↑
3. ランク付けされた業務	
4. ある産業に勤務していた期間	↓
5. ある産業に勤務したことの有無	最悪

③ 関連の指標

疫学研究の方法で、実験との関連から最も理解しやすいのは、集団を無作為に 2 つに分け、一方は曝露群、もう一方非曝露群として、その 2 者を追跡してその死亡（疾病）状況を比べるというものであろう。医学の使命は人間の健康に寄与することであるから、倫理的には許されないことではあるが、方法論としてはわかりやすいので、この前提に沿って説明を進める。はじめにデータレイアウト表 28 に示した。他のタイプの疫学研究としては、疾病をもつ者ともたない者を選び、曝露の割合を調査する患者対照研究などがある。

表 28-a 率型のデータレイアウト

	曝露	非曝露	合計
罹患者	a	b	M
人年	N ₁	N ₀	T

表 28-b リスク型のデータレイアウト

	曝露	非曝露	合計
疾病あり	a	b	M ₁
疾病なし	c	d	M ₀
合計	N ₁	N ₀	T

1. 比

率比（レイト比）；曝露群と非曝露群との相対的な比較に用いられる。死亡率・罹患率などの率の比をとるのが率比 **rate ratio (RR)**、危険度の割合をとるのが危険度比 **risk ratio** である。表 28 のレイアウトによると、両者ともに次のように表現される。

$$RR = (a/N_1) / (b/N_0) \dots\dots\dots (6-1)$$

区間推定値の近似式としては、基準正規分布をする z 値から、例えば 95%信頼区間を与える 1.96 を用いると、率比は、

$$\exp \{ \ln (RR) \pm 1.96 \sqrt{ (1/a + 1/b) } \},$$

リスク比は、 $\exp \{ \ln (RR) \pm 1.96 \sqrt{ (c/aN_1 + d/bN_0) } \}$ で表される。ただし、 $\exp$ は自然対数の底、 $\ln$ は自然対数をあらわす。曝露が疾病の発生に影響を与えない場合、 $RR=1$ であるから、上記の 95%信頼区間が 1 を含まなければ、曝露が疾病と関係ない、すなわち $RR=1$ という帰無仮説は棄却できる。この場合は、5%水準で曝露と疾病とは関係があると考え。標本が小さい場合には直接確率を求めるが、これについては本書の範囲を超える。

オッズ比 odds ratio；オッズとは、あたる確率とはずれる確率との比である。例えば曝露群の罹患のオッズは、次の式であらわされる。

$$(a/N_1) / (c/N_1) = a/c \dots\dots\dots (6-2)$$

曝露群の死亡（罹患）確率のオッズと、非曝露群の死亡（罹患）確率のオッズとの比が、死亡（罹患）のオッズ比（OR）であり、表 28-b のレイアウトによると次のようにあらわされる。

$$OR = \{(a/N_1) / (c/N_1)\} / \{(b/N_0) / (d/N_0)\} = (ad) / (bc) \dots\dots\dots (6-3)$$

区間推定値としては、z 値から自分の望む信頼区間を与える値 z_0 を用いて、

$$\exp \{ \ln (OR) \pm z_0 \sqrt{ (1/a + 1/b + 1/c + 1/d) } \} \dots\dots\dots (6-4)$$

で与えられる。

オッズ比に計算と 95%信頼区間

以下のような仮のデータを考えてみる。

	食塩摂取		
	1 日 15g 以上 (+)	1 日 15g 未満 (-)	計
脳血管疾患ケース	84	36	120
対照	66	54	120
計	150	90	240

オッズ比（OR）は、 $OR = (ad) / (bc) = (84 \times 54) / (36 \times 66) = 1.91$

95%信頼区間（CI）は $CI = \exp \{ \ln (OR) \pm 1.96 \sqrt{ (1/a + 1/b + 1/c + 1/d) } \}$

$\ln (OR) = \log_e 1.91 = 0.647$

$S = 1.96 \sqrt{ (1/a + 1/b + 1/c + 1/d) } = 1.96 \sqrt{ (1/84 + 1/36 + 1/66 + 1/54) } = 0.533$

信頼区間の下限値を CI_1 、上限値を CI_2 とすると、

$\ln (OR) - S = 0.647 - 0.533 = 0.114$ 、 $\ln (OR) + S = 0.647 + 0.533 = 1.18$ から

$CI_1 = e^{0.114} = 1.12$ 、 $CI_2 = e^{1.18} = 3.25$

ゆえに、オッズ比は 1.91、95%信頼区間は「1.12～3.25」となる。これは 1.00 よりも大きいので、1 日に 15g 以上の食塩摂取は脳血管疾患の危険因子である可能性が大きい。

2. 差

数学的な取り扱いの関係で、比ほどは論文などに出てこないが、曝露群と非曝露群との率あるいは危険度の差（RD）も用いられることがある。表 28 のレイアウトによると、次の式であらわされる。

$$RD = a/N_1 - b/N_0 \dots\dots\dots (6-5)$$

これは曝露群と非曝露群との絶対的な比較であり、特に衛生行政などの分野では重要である。たとえば、疾患 A の死亡率比が 10 で、物質 m_A への非曝露群の死亡率が 10^{-6} （年⁻¹）、曝露群の死亡率が 10×10^{-6} （年⁻¹）とする。一方、罹患 B の死亡率比が 1.2 で、物質 m_B への非曝露群の死亡率が 100×10^{-6} （年⁻¹）、曝露群の死亡率が 120×10^{-6} （年⁻¹）とする。死亡率比でみれば、圧倒的に A のほうが危険であるが、物質 m_A への曝露を防ぐことによって可能となる死亡率の減少は、 9×10^{-6} （年⁻¹）であるのに対し、物質 m_B への曝露を防げば、 20×10^{-6} （年⁻¹）の死亡率減少が見込める。これは、単純な比較で、実際には曝露人口の割合、無害な物質による代替の可能性、罹患者にかかる費用など、行政上で考慮すべきことは多いが、それにしても、このような場合に比よりは差のほうが有用であることにはかわりはない。

寄与危険割合（attributable proportion）；ある因子がある疾病の原因であると考えられる場合、寄与危険割合という考え方も、率差と同様に有用である。これは、曝露群が曝露の除去によって非曝露群の罹患率になる場合に、防げる罹患者の割合として定義される。曝露群の寄与危険割合（ AP_E ）は、次の式

であらわされる。

$AP_E = (RR - 1) / RR$ 曝露群・非曝露群を含む集団での寄与危険割合 (AP_T) は、その集団における曝露者の割合を P であらわすと次のようになる。

$$AP_T = (RR - 1) / (RR + 1/P - 1) = AP_E \cdot P$$

相対危険度と寄与危険度；relative risk と attributable risk で

相対危険度 = (曝露群の発生率) / (非曝露群の発生率)

寄与危険度 = 曝露群の発生率 - 非曝露群の発生率

寄与危険度百分率 = (曝露群の発生率 - 非曝露群の発生率) / 曝露群の発生率 = 1 - 非曝露群の発生率 /

曝露群の発生率 = 1 - 1 / 相対危険度

相対危険度とは、問題の要因を持っている人が、もっていない人を 1 とした場合、何倍その病気になりやすいかを表したものの。

寄与危険度とは、集団として問題の要因をもった場合、どれほどその病気がその集団内で増えるかを表すもの。寄与危険度は集団としての危険度、相対危険度は個人としての危険度を表している。

寄与危険度百分率 (attributable risk percent) は曝露群の発生全体のうち何%が問題の要因によって発生したかを表すもの。

たとえば、肉食を好む 10000 人と好まない人 10000 人を 20 年間追跡し、大腸がんの発生をみた場合、

	大腸がん (+)	大腸がん (-)	計
肉食 (+)	80	9920	10000
肉食 (-)	20	9980	10000

相対危険度は $(80 \div 10000) \div (20 \div 10000) = 4$ となる。つまり、肉食者は非肉食者に比べ、4 倍大腸がんにかかりやすい、という意味。

寄与危険度は、 $80 - 20 = 60$ (対 10000) で、10000 人当たりの大腸がん発生 80 人のうち、60 人は肉食による、という意味。これは、20 人は肉食をしなくても発生するので、肉食の責任分は 80 人中 60 人である、という意味でもある。寄与危険度百分率は $60 \div 80 \times 100 = 75\%$ となる。

まとめると、

相対危険度は発生率の比、個人としての危険度

寄与危険度は発生率の差、集団としての危険度

寄与危険度百分率は問題の要因が責任を分担する割合

人口寄与危険度割合；曝露要因を取り除けば曝露群と非曝露群を合わせた集団全体の疾患をどのくらい減らせるかを示すもので、公衆衛生対策の優先順位の判定にも用いる。

人口寄与危険度割合 = (集団全体の罹患率 - 非曝露群の罹患率) / (集団全体の罹患率)

研究スタートの時点では、

コホート研究は要因判明済み、患者未発生

患者・対照研究は要因未判明、患者発生済み

オッズ比は同じ同士を集めてたすき掛け

精度が高いとデータは固まる (低いと散らばる)

正確度が高いとデータの平均は真に近い (低いと偏る)

死亡率の分母は人口、致命率の分母は患者

罹患率は、集団の中で一定の期間に発生した患者の数。一定期間内に罹患する確率で（「発生率」と一般には同義で用いられる）、疾病の量を表す指標。

有病率は、ある時点で存在する患者数を、人口に対する比率で表したもので、疾病の量を表す指標。

累積罹患率は、最初に定めた観測対象集団全員を、一定期間追跡して、その罹患率を調べるもので、通常、人口 10 万対あるいは 100 万対の比率で示される。

累積罹患率＝（観察期間中の新たな罹患人数）／（最初に観察対象に規定された人口集団の数）

人年法（person-year method）；移動の激しい地域住民や職場集団など、対象者の観察期間が異なる集団の発生頻度を測定する場合、1 人 1 年間の観察を 1 単位（人年）として分母とする方法を人年法という（2 人 1 年間は 2 人年、1 人 2 年間も 2 人年）。

罹患率＝（観察期間中の新たな罹患人数）／（1 人 1 人の観察期間の総和である人年）

④ 偶然・偏り（バイアス）・交絡

偶然の変動は実験でもおこるが、当然、疫学研究においても起こりうる。これには純粹に偶然の変動もあるが、測定が不可能な因子によるばらつきも含まれている。同じ研究方法でも、異なる地域や人種などを対象にした研究はその意味でむだではなく、重要である。これらの研究の総合的な判断から、次に述べる偏り（バイアス bias）や交絡 confounding に関するヒントを得ることがあるからである。

この章の冒頭で述べたように、疫学は実際に人間に起こった疾病・死亡現象を扱うという、動物実験にはない長所をもつ。しかしながら、逆に人間を扱うがために、倫理上、有害物質の投与群と非投与群とに振り分けて観察することは不可能である。このように実験不可能な集団の観察における大きな問題が、偏りと交絡であり、疫学研究の優劣はこれらをいかに制御するかにかかっている。これは疫学にかぎらず、人間集団を対象とする科学の領域では非常に重要なことである。

① 偏り（バイアス）

ひとことでいえば、偏りとは、観測から得られる推定値と真の値との間のずれである。これには、偶然の誤差によるずれは含まれない。偶然の誤差であれば、同様の研究を繰り返し行うことによって、その平均が真の値に一致するからである。偏りは、対象の選択における偏り（selection bias）と、対象から得られる情報の偏り（information bias）の 2 つに分けられる。対象選択における偏りの例としては、次のようなものがある。

循環器疾患を減少させるために、運動によるプログラムを作り、このプログラムへの参加をある保健所管内で住民に呼びかけた。応募した人のその後 10 年間の循環器疾患死亡率は、その管内全体の循環器疾患死亡率よりも低かった。これは、もちろん運動による効果のためかもしれないが、運動するプログラムに参加したいと思う人は、運動ができるほど健康であったはずで、運動のプログラムに参加しなくても、心疾患をもつ人を含む一般集団よりも循環器疾患死亡率は低かったはずである。このことを考慮に入れないと、運動による循環器疾患死亡率の減少を過大評価することになってしまう。

次に情報の偏りの例を 1 つあげよう。肺がんの遺伝的要因を調べるのに、ある病院の肺がん患者 100 人の家族と、椎間板ヘルニア患者 400 人の家族とに、家族歴に差があるかどうかを質問した。肺がん患者の家族は肺がんに関して敏感になっており、親戚の肺がんについてもよく知るようになるが、椎間板ヘルニアの患者の家族ではそれほどでもない。そのため家族歴を聴取すると、実際の家族歴の有無が過大評価されてしまうことになる。以上の例から、対象集団の選定と情報の収集のしかたには十分な注意が必要であることが理解できる。

② 交絡

交絡とは、複数の因子の影響の混合をさす。以下に、単純化した例で説明してみる。いま、ある爆発性の化学物質 A を製造する会社の工員 10,000 名と、その会社の事務員 10,000 名を対象に、A が肺がんの原因物質であるかどうかを調査することになり、真の肺がん罹患率は表 29 のようであったとする。喫煙者群でも非喫煙者群でも A への曝露の有無による罹患率比は 2.0 となり、A は肺がんの原因の 1 つと考えられる。同様に、A への曝露のある集団でもない集団でも、喫煙の有無による罹患率比は 10.0 である。A は爆発性のため勤務時間中の喫煙は禁止されており、工員の喫煙者は少なくても 1,000 人であったが、事務員にはそのような規則はなく、喫煙は自由であったため、喫煙者は 7,000 人、非喫煙者は 3,000 人であった。この場合、1 年間の調査による肺がんの罹患者数は次のようになる。ただし、罹患率が低く、期間が短いので、年初の観察人員×観察期間で人年を近似させ、それに罹患率をかけて罹患者数としてある。

表 29 喫煙・A への曝露別肺がん罹患率 (年⁻¹)

	非喫煙	喫煙
A への曝露なし (事務員)	0.0001	0.001
A への曝露あり (工員)	0.0002	0.002

工員の喫煙者は罹患率が 0.002、人年が $1,000 \times 1 = 1,000$ 、罹患者数は、 $0.002 \times 1,000 = 2.0$

工員の非喫煙者からの罹患者は同様に、 $0.0002 \times 9,000 = 1.8$ 合計で 3.8 人となり、人年が 10,000 で近似できるので、罹患率は次のようになる。 $3.8/10,000 = 0.00038$ 、一方、事務員の喫煙者では、罹患者数が $0.001 \times 7,000 = 7.0$ 、非喫煙者では、 $0.0001 \times 3,000 = 0.3$ 合計 7.3 人となり、罹患率は次のようになる。 $7.3/10,000 = 0.00073$

喫煙を無視した工員と事務員との罹患率比は、 $0.00038/0.00073 = 0.52$ となり、A への曝露が肺がんに対して予防的に働くかに見える。しかし、この見かけ上の結果は、喫煙の影響を A への曝露の影響と混合したために起こったものである。

交絡因子は、疾患の発生、曝露因子の両方に関係していなくてはならない。上の例では、交絡因子としての喫煙の有無で肺がん罹患率が異なるし、喫煙者の割合が物質 A への曝露群と非曝露群とで異なっている。

喫煙者と非喫煙者とを別々にみると、A の影響が正確に把握された。これを層別というが、標本さえ大きければ非常に優れた交絡因子の制御法である。そのほか集団の限定（たとえば非喫煙者だけを対象とする方法）や標準化、多変量解析などが用いられる。

集団の限定は、人的・経済的に層別が難しい時に用いられるが、基本的には層別のほうが良い。標準化は、強力な交絡因子の 1 つである年齢（交絡因子の代表的な指標である）の制御に用いられることが多い。一方、多変量解析は、制御する因子が多すぎる場合に用いられるもので、マンテル－ヘンセル Mantel－Haenszel 法、多変量ロジスティック回帰モデル multiple logistic regression model、コックス比例ハザードモデル Cox proportional hazards model を用いた解析などがある。

なお、当然のことながら、未知の交絡因子は制御が不可能である。しかし、未知の交絡因子についても予想が可能なことがある。上述の物質 A の例でいうと、喫煙が肺がんの原因ということが未知であっても、他のいくつかの工場での同様の調査でも、工場ごとに、工員と事務員の喫煙者の割合が異なれば結果が一定にならない。ということが、物質 A 以外の肺がんの要因を疑わせるのである。