大学院・医学基礎技術演習・実験基本技術（医学統計学）テキスト

中澤　港（生態情報学　助教授）

2006 年 5 月 10 日

本演習は，「医学情報検索」と合わせて 1 単位とする。実験や調査によって得られる生データからデータファイルを作成し，統計処理ソフトウェア **R**[*1]で解析し，結果を読み，レポートをまとめるという一連の流れを身に付けられるように，コンピュータを使って演習を行う。

問い合わせ先：生態情報学　助教授　中澤　港（e-mail: nminato@med.gunma-u.ac.jp）

1　研究倫理・データ入力・記述統計・図示

研究に先立って必要な研究倫理についての考え方を概説し，研究デザインを説明した後に，基本的なデータ入力の方法と，生データの概要を把握するための記述統計の方法と簡単な図示の方法を示す。

1.1　研究倫理について

研究は，ただやればいいというものではない。その研究をすることが社会に受容されるためには，何らかの社会貢献ができなければならない。社会に余裕があれば，役に立つか立たないかわからない研究でも受容されやすいが，現代では，その意義を十分に説明できないような研究は行えない。とくに医学研究はヒトを対象とするので，対象者への説明責任も大きくなるのが必然であり，対象者が研究に協力することによって蒙る負荷を上回る利益がなければなされるべきではないと考えるのが普通である。

もちろん，対象者のプライヴァシーへの配慮が必要なことはいうまでもない。最近よく EBM（Evidence-based Medicine；根拠に基づく医療）の必要性が言われるけれども，その Evidence としてもっとも強力なものの一つとして，メタアナリシスの結果がある。メタアナリシスとは，多くの研究結果をまとめて分析する方法である。数多くの多様な対象集団について共通してみいだされるメカニズムがあれば，それが強力な Evidence となるのは自明であろう。メタアナリシスのためには，個々の研究データが利用可能な形でデータベースに蓄積されることが理想であるが[*2]，そのときに対象者個人のプライヴァシーが漏れるようなことがあってはならない。

医学研究における倫理を扱った中で，もっとも有名なものは，「ヘルシンキ宣言」[*3]である。ヘルシンキ宣言は，A．序言，B．すべての医学研究の基本原則，C．メディカル・ケアと結びついた医学研究のための追加原則の 3 パートと脚注からなり，A．で宣言自体が適用されるべき範囲を既定し，ジュネーブ宣言が「私の患者の健康を私の第一の関心事とする」ことを医師に義務づけていることと，医の倫理の国際綱領が「医師は患者の身体的及び精神的な状態を弱める影響をもつ可能性のある医療に際しては，患者の利益のためにのみ行動すべきである」と宣言していることに触れた上で，被験者の福利が最優先であることと危険と負担が伴うことを踏まえながらも，医学研究の必要性を訴えている。自国の倫理や法規制上の要請への配慮の必要性にも触れている。B．では被験者がボランティアであること，科学的原則に従うこと，環境への配慮，詳細な研究計画書を作成し事前に倫理審査を受けること，研究計画の公開性の保証，受益者にとっての目的の正当性の保証，インフォームド・コンセントなどの必要性が述べられている。C．では被験者が

[*1] **R** については付録を参照。

[*2] 現実にはなかなかそうなっていない。

[*3] `http://www.med.or.jp/wma/helsinki02_j.html` で，日本医師会による訳が全文読めるが，世界医師会 (WMA) によって発表されたもので（英文は `http://www.wma.net/e/policy/pdf/17c.pdf` でダウンロードできる），正式には，「ヒトを対象とする医学研究の倫理的原則」という。2000 年のエディンバラ改訂からは医師だけでなく，研究者すべてが遵守すべきとされる。2004 年 10 月に東京で行われた WMA 総会で第 30 項への注釈が付加されたのが最新の改訂である。

患者である場合に，研究方法が現在最善とされている予防・診断・治療の方法と比較されねばならないことや，研究終了後に被験者はその成果として得られた最善の方法を受けられねばならないこと，などが定められている。

> 疫学研究は、疾病のり患をはじめ健康に関する事象の頻度や分布を調査し、その要因を明らかにする科学研究である。疾病の成因を探り、疾病の予防法や治療法の有効性を検証し、又は環境や生活習慣と健康とのかかわりを明らかにするために、疫学研究は欠くことができず、医学の発展や国民の健康の保持増進に多大な役割を果たしている。
> 疫学研究では、多数の研究対象者の心身の状態や周囲の環境、生活習慣等について具体的な情報を取り扱う。また、疫学研究は医師以外にも多くの関係者が研究に携わるという特色を有する。
> 疫学研究については、従来から、研究対象者のプライバシーに配慮しながら研究が行われてきたところであるが、近年、研究対象者に説明し同意を得ることが重要と考えられるようになり、さらに、プライバシーの権利に関する意識の向上や、個人情報保護の社会的動向などの中で、疫学研究においてよるべき規範を明らかにすることが求められている。
> そこで、研究対象者の個人の尊厳と人権を守るとともに、研究者等がより円滑に研究を行うことができるよう、ここに倫理指針を定める。
> この指針は、世界医師会によるヘルシンキ宣言や、我が国の個人情報保護に係る論議等を踏まえ、疫学研究の実施に当たり、研究対象者に対して説明し、同意を得ることを原則とする。また、疫学研究に極めて多様な形態があることに配慮して、この指針においては基本的な原則を示すにとどめており、研究者等が研究計画を立案し、その適否について倫理審査委員会が判断するに当たっては、この原則を踏まえつつ、個々の研究計画の内容等に応じて適切に判断することが求められる。
> 疫学研究が、社会の理解と信頼を得て、一層社会に貢献するために、すべての疫学研究の関係者が、この指針に従って研究に携わることが求められている。同時に、健康の保持増進のために必要な疫学研究の実施について、広く一般社会の理解が得られることを期待する。

ヒトを対象とした研究全般について，平成 14 年 6 月 17 日に，文部科学省と厚生労働省が共同で発表した倫理指針が，「疫学研究に関する倫理指針」[*4]である。インフォームド・コンセントと研究の公開性，公益性の保証がポイントである。適用範囲や具体的な運用にいたるまで，かなり詳しく書かれているが，思想的背景となる前文は，枠内の通りである。ヘルシンキ宣言を踏まえていることがわかる。また，臨床研究に関しては，翌平成 15 年 7 月 16 日に厚生労働省より告示され同 30 日から施行された「臨床研究に関する倫理指針」[*5]に従うこととされているが，これもヘルシンキ宣言を踏まえて作成されたものである。

群馬大学でヒトを対象とした研究をしようとする場合は，臨床研究倫理審査専門委員会[*6]か，疫学研究倫理審査委員会[*7]の審査を受け，審査結果が医学倫理委員会に報告され，承認を受けたものについて医学系研究科長によって研究の実施が許可される，というプロセスを踏まねばならない。

1.2 データ入力

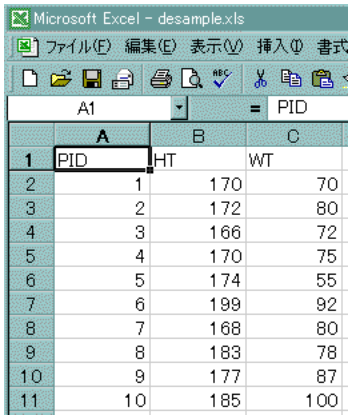
研究によって得られたデータをコンピュータを使って統計的に分析するためには，まず，コンピュータにデータを入力する必要がある。データの規模や利用するソフトウェアによって，どういう入力方法が適当か（正しく入力でき，かつ効率が良いか）は異なってくる。

ごく小さな規模のデータについて単純な分析だけ行う場合，電卓で計算してもよいし，分析する手続きの中で直接数値を入れてしまってもよい。例えば，60 kg, 66 kg, 75 kg という 3 人の平均体重を求めるには，Microsoft Excel では，1 つのセルの中に`=AVERAGE(60,66,75)` とか`=(60+66+75)/3` と打てばいいし，R ならばプロンプトに対して `mean(c(60,66,75))` または `(60+66+75)/3` と打てばいい。

しかし実際にはもっとサイズの大きなデータについて，いろいろな分析を行う場合が多いので，データ入力と分析は別々に行うのが普通である。そのためには，同じ調査を繰り返しするとか，きわめて大きなデータであるとかでなければ，Microsoft Excel のような表計算ソフトで入力するのが手軽であろう。きわめて単純な例として，10 人の対象者についての身長と体重のデータが次の表のように得られているとする。

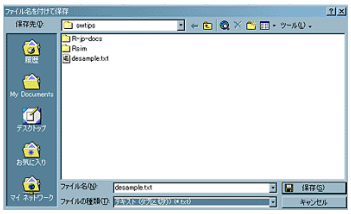
[*4] `http://www.mhlw.go.jp/general/seido/kousei/i-kenkyu/sisin2.html` で全文公開されている。
[*5] `http://www.mhlw.go.jp/topics/2003/07/dl/tp0730-2b.pdf` で，施行に関する細則も付記された形で全文公開されている。
[*6] `http://www.med.gunma-u.ac.jp/buisiness-offices/general-affairs/rinsyokenkyu.html` 参照。
[*7] `http://www.med.gunma-u.ac.jp/buisiness-offices/general-affairs/ekigaku.html` 参照。

対象者 ID	身長 (cm)	体重 (kg)
1	170	70
2	172	80
3	166	72
4	170	75
5	174	55
6	199	92
7	168	80
8	183	78
9	177	87
10	185	100



まずこれを Microsoft Excel などの表計算ソフトに入力する。一番上の行には変数名を入れる。日本語対応 R なら漢字やカタカナ，ひらがなも使えるが，半角英数字（半角ピリオドも使える）にしておくのが無難である。入力が終わったら，一旦，そのソフトの標準の形式で保存しておく。入力完了した状態は，右の画面のようになる。

次に，この表をタブ区切りテキスト形式で保存する。Microsoft Excel の場合，メニューバーの「ファイル (F)」から「名前を付けて保存」を選び，現れるウィンドウの一番下の「ファイルの種類 (T)」のプルダウンメニューから「テキスト（タブ区切り）(*.txt)」を選ぶと，自動的にその上の行のファイル名の拡張子も xls から txt に変わるので，「保存 (S)」ボタンを押せば OK である（下のスクリーンショットを参照）。複数のシートを含むブックの保存をサポートした形式でないとかいう警告が表示されるが無視して「はい」を選んでよい。その直後に Excel を終了しようとすると，何も変更していないのに「保存しますか」と聞く警告ウィンドウが現れるが，既に保存してあるので「いいえ」と答えてよい（「はい」を選んでも同じ内容が上書きされるだけであり問題はない）。この例では，desample.txt ができる。



あとは R で読み込めばいい。この例のように，複数の変数を含む変数名付きのデータを読み込むときは，データフレームという構造に付値するのが普通である。保存済みのデータが `d:\desample.txt` だとすれば，R のプロンプトに対して，

```
dat <- read.delim("d:/desample.txt")
```

と打てば，データが `dat` というデータフレームに付値される[*8]。確認のためにデータを表示させたければ，ただ `dat` と打つ。データ構造を見たければ，`str(dat)` とすればよい。読み込まれた変数に対して分析したいとき，例えばこの例の身長の平均と標準偏差を出したければ，

```
cat("mean=",mean(dat$HT),"sd=",sd(dat$HT),"\n")
```

とする。いちいち `dat$`と打つのが面倒ならば，`attach(dat)` とすることでデフォルトのデータフレームが指定できるので，それ以降のセッション中，`detach(dat)` するまで，`dat$`を入力しなくても良くなる。例えば，このデータで身長と体重の相関係数を出して検定したいときは次のようにできる。

[*8] なお，この程度のデータなら，ファイル保存をしなくても，Microsoft Excel の上で入力した範囲をすべて選んでコピーし（Ctrl キーと C キーを同時に押すか，あるいはマウスで右クリックしてコピーを選ぶ），R のプロンプトに対して，`dat <- read.delim("clipboard")` と打ってもよい。

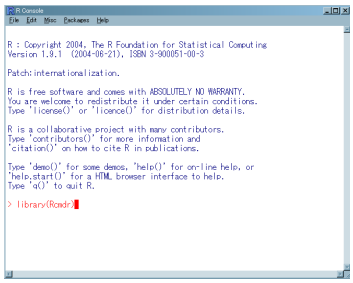
```
attach(dat)
cor.test(HT,WT)
detach(dat)
```

なお，R コマンダー（Rcmdr）を使って，次のようにメニュー形式でデータファイルを読み込むこともできる。ただし，Rcmdr を使うためには，予め，R Console の設定の GUI インターフェースを SDI にしておく必要がある。そのためには，R を起動後，メニューバーの Edit の GUI preferences を開いて，SDI のところをチェックしてから，Save して Finish し，R を再起動すればよい。

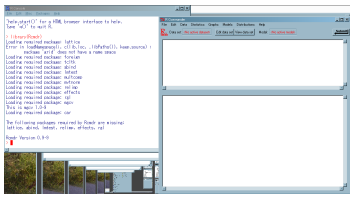
他のライブラリを使えるようにするときも基本的に同じだが，Rcmdr ライブラリを呼び出すには，

```
library(Rcmdr)
```

と入力する。



他のライブラリとは異なり，Rcmdr ライブラリは，呼び出すだけで R コマンダーのウィンドウが起動する。



この状態から，先ほど保存した `d:\desample.txt` を読み込むためには，メニューバーの Data から Import Data の From Text File を開いて，Enter name for data set:の欄に適当な参照名をつけ（変数名として使える文字列なら何でもよいのだが，デフォルトでは Dataset となっている），Field Separator を White space から Tabs に変えて（Tabs の右にある○をクリックすればよい），OK ボタンをクリックすればよい。後は Rcmdr のメニューから選んでいくだけで，いろいろな分析ができる。

なお，データ入力は，入力ミスを防ぐために，2 人以上の人が同じデータを入力し，それを比較するプログラムを実行して誤りをチェックする方法がよいとされる。しかし，現実には 2 人の入力者を確保するのが困難なため，1 人で 2 回入力して 2 人で入力する代わりにするか，あるいは 1 人で入力してプリントアウトした結果を元データと見比べてチェックするといった方法が使われることも多い。

1.3 欠損値の扱い

ここで注意しなければならないのは，欠損値の取扱いである。一般に，統計処理をする対象のデータは，母集団から標本抽出したサンプルについてのものである。サンプルデータを統計解析して，母集団についての情報を得るためには，そのサンプルが正しく母集団を代表していることが何より大切である。質問紙調査の場合でも，実験研究の場合でも，欠損値（質問紙なら無回答，非該当，わからない，等，実験研究なら検出限界以下，サンプル量不足，測定失敗等）をどのように扱うかによって，サンプルの代表性が歪められてしまうことがある。欠損が少なければあまり気にしなくていいが，たとえば，健診の際の食生活質問等で，「甘いものが好きですか」に対して無回答の人は，好きだけれどもそれが健康に悪いと判断されるだろうから答えたくない可能性があり，その人たちを分析から除くと，甘いもの好きの人の割合が，全体よりも少なめに偏った対象の分析になってしまう。なるべく欠損が少なくなるような努力をすべきだけれども，どうしても欠損のままに残ってしまった場合は，結果を解釈する際に注意する。

欠損値のコードは，通常，無回答 (NA) と非該当と不十分な回答が区別できる形でコーディングするが，ソフトウェアの上で欠損値を欠損値として認識させるためのコードは，分析に使うソフトウェアによって異なっているので（欠損

値を表すコードの方を変更することも可能），それに合わせておくのも１つの方法である。デフォルトの欠損値記号は，R なら NA，SAS なら.（半角ピリオド）である。Excel ではブランク（何も入力しない）にしておくと欠損値として扱われる，入力段階で欠損値をブランクにしておくと，「入力し忘れたのか欠損値なのかが区別できない」という問題を生じるので，入力段階では決まった記号を入力しておいた方が良い。その上で，もし簡単な分析まで Excel でするなら，すべての入力が完了してから，検索置換機能を使って（Excel なら「編集」の「置換」。「完全に同一なセルだけを検索する」にチェックを入れておく），欠損値記号をブランクに変換すれば用は足りる。

次に問題になるのが，欠損値を含むデータをどう扱うかである。結果を解釈する上で一番紛れのない方法は，「１つでも無回答項目があったケースは分析対象から外す」ということである[*9]（もちろん，非該当は欠損値ではあるが外してはならない）。その場合，統計ソフトに渡す前の段階で，そのケースのデータ全体（Excel 上の１行）を削除してしまうのが簡単である（もちろん，元データは別名で保存しておいて，コピー上で行削除）。質問紙調査の場合，たとえば 100 人を調査対象としてサンプリングして，調査できた人がそのうち 80 人で，無回答項目があった人が 5 人いたとすると，回収率 (recovery rate) は 80%（80/100）となり，有効回収率 (effective recovery rate) が 75%（75/100）となる。調査の信頼性を示す上で，これらの情報を明記することは重要である。目安としては有効回収率が 80%程度は欲しい。

1.4 記述統計

記述統計はデータの特徴を把握する目的で用いる。しかし，あまりにも妙な最大値や最小値，大きすぎる標準偏差などが得られた場合は，入力ミスを疑って，元データに立ち返ってみるべきである。

記述統計量には，大雑把にいって，分布の位置を示す「中心傾向」と分布の広がりを示す「ばらつき」があり，中心傾向としては平均値，中央値，最頻値がよく用いられ，ばらつきとしては分散，標準偏差，四分位範囲，四分位偏差がよく用いられる。Rcmdr からは，メニューバーの Statistics の Summaries から項目を選べばよい。

中心傾向の代表的なものは以下の３つである。

平均値（mean） 分布の位置を示す指標として，もっとも頻繁に用いられる。実験的仮説検証のためにデザインされた式の中でも，頻繁に用いられる。R で平均値を計算する方法は既に記載した通りである。記述的な指標の１つとして，平均値は，いくつかの利点と欠点をもっている。日常生活の中でも平均をとるという操作は普通に行われるから説明不要かもしれないが，数式で書くと以下の通りである。

母集団の平均値 μ（ミューと発音する）は，

$$\mu = \frac{\sum X}{N}$$

である。X はその分布における個々の値であり，N は値の総数である。$\sum$（シグマと発音する）は，一群の値の和を求める記号である。すなわち，$\sum X = X_1 + X_2 + X_3 + ... + X_N$ である。

標本についての平均値を求める式も，母集団についての式と同一である。ただし，数式で使う記号が若干異なっている。標本平均 $\bar{X}$（エックスバーと発音する）は，

$$\bar{X} = \frac{\sum X}{n}$$

である。n は，もちろん標本サイズである[*10]。

ちなみに，重み付き平均は，各々の値にある重みをかけて合計したものを，重みの合計で割った値である。式で書くと，

$$\bar{X} = \frac{n_1(\bar{X}_1) + n_2(\bar{X}_2) + ... + n_n(\bar{X}_n)}{n_1 + n_2 + ... + n_n}$$

中央値（median） 中央値は，全体の半分がその値より小さく，半分がその値より大きい，という意味で，分布の中央である。言い換えると，中央値は，頻度あるいは値の数に基づいて分布を２つに等分割する値である。中央値

[*9] 最初からその方針ならば，１つでも無回答項目があった人のデータは入力しないことに決めておく手もある。通常はそこまで思い切れないので，とりあえず入力は全部することが多い。

[*10] 記号について注記しておくと，集合論では $\bar{X}$ は集合 X の補集合の意味で使われるが，代数では確率変数 X の標本平均が $\bar{X}$ で表されるということである。同じような記号が別の意味で使われるので混乱しないように注意されたい。補集合は X^C という表記がなされる場合も多いようである。標本平均は $\bar{X}$ と表すのが普通である。

を求めるには式は使わない（決まった手続き＝アルゴリズムとして，並べ替え (sorting) は必要)。極端な外れ値の影響を受けにくい（言い換えると，外れ値に対して頑健である)。歪んだ分布に対する最も重要な central tendency の指標が中央値である。R で中央値を計算するには，`median()` という関数を使う。なお，データが偶数個の場合は，普通は中央にもっとも近い２つの値を平均した値を中央値として使うことになっている。

最頻値（Mode） 最頻値はもっとも度数が多い値である。すべての値の出現頻度が等しい場合は，最頻値は存在しない。

平均値は，(1) 分布のすべての値を考慮した値である，(2) 同じ母集団からサンプリングを繰り返した場合に一定の値となる，(3) 多くの統計量や検定で使われている，という特長をもつ。標本調査値から母集団の因果関係を推論したい場合に，もっとも普通に使われる。しかし，(1) 極端な外れ値の影響を受けやすい，(2) 打ち切りのある分布では代表性を失う場合がある[*11]，という欠点があり，外れ値があったり打ち切りがあったりする分布では位置の指標として中央値の方が優れている。最頻値は，標本をとったときの偶然性の影響を受けやすいし，もっとも頻度が高い値以外の情報はまったく使われない。しかし，試験の点で何点の人が多かったかを見たい場合は最頻値が役に立つし，名義尺度については最頻値しか使えない。

ここで上げた３つの他に，幾何平均 (geometric mean) や調和平均 (harmonic mean) も，分布の位置の指標として使われることがある。幾何平均はデータの積の累乗根（対数をとって平均値を出して元に戻したもの)，調和平均はデータの逆数の平均値の逆数であり，どちらもゼロを含むデータには使えない。大きな外れ値の影響を受けにくいという利点があり，幾何平均は，とくにデータの分布が対数正規分布に近い場合によく用いられる。

一方，分布のばらつき（Variability）の指標として代表的なものは，以下の４つである。

四分位範囲 (Inter-Quartile Range; IQR) 四分位範囲について説明する前に，分位数について説明する。値を小さい方から順番に並べ替えて，４つの等しい数の群に分けたときの 1/4, 2/4, 3/4 にあたる値を，四分位数 (quartile) という。1/4 の点が第１四分位，3/4 の点が第３四分位である（つまり全体の 25 %の値が第１四分位より小さく，全体の 75 %の値が第３四分位より小さい)。2/4 の点というのは，ちょうど順番が真中ということだから，第２四分位は中央値に等しい。ちょっと考えればわかるように，ちょうど４等分などできない場合がもちろんあって，上から数えた場合と下から数えた場合で四分位数がずれる可能性があるが，その場合はそれらを平均するのが普通である。また，最小値，最大値に，第１四分位，第３四分位と中央値を加えた５つの値を五数要約値と呼ぶことがある（R では `fivenum()` 関数で五数要約値を求めることができる)。第１四分位，第２四分位，第３四分位は，それぞれ Q1, Q2, Q3 と略記することがある。四分位範囲とは，第３四分位と第１四分位の間隔である。上と下の極端な値を排除して，全体の中央付近の 50%（つまり代表性が高いと考えられる半数）が含まれる範囲を示すことができる。

四分位偏差 (Semi Inter-Quartile Range; SIQR) 四分位範囲を２で割った値を四分位偏差と呼ぶ。もし分布が左右対称型の正規分布であれば，中央値マイナス四分位偏差から中央値プラス四分位偏差までの幅に全データの半分が含まれるという意味で，四分位偏差は重要な指標である。IQR も SIQR も少数の極端な外れ値の影響を受けにくいし，分布が歪んでいても使える指標である。

分散 (variance) データの個々の値と平均値との差を偏差というが，マイナス側の偏差とプラス側の偏差を同等に扱うために，偏差を二乗して，その平均をとると，分散という値になる。分散 V は，

$$V = \frac{\sum (X - \mu)^2}{N}$$

で定義される[*12]。標本数 n で割る代わりに自由度 $n-1$ で割って，不偏分散 (unbiased variance) という値にすると，標本データから母集団の分散を推定するのに使える。即ち，不偏分散 V_{ub} は，

$$V_{ub} = \frac{\sum (X - \bar{X})^2}{n-1}$$

である。

[*11] 氷水で痛みがとれるまでにかかる時間とか，年収とか。無限に観察を続けるわけにはいかないし，年収は下限がゼロで上限はビル・ゲイツのそれのように極端に高い値があるから右すそを長く引いた分布になる。平均年収を出している統計表を見るときは注意が必要である。年収の平均的な水準は中央値で表示されるべきである。

[*12] 実際に計算するときは２乗の平均から平均の２乗を引くとよい。

標準偏差 (standard deviation)　分散の平方根をとったものが標準偏差である。平均値と次元を揃える意味をもつ。不偏分散の平方根をとったものは，不偏標準偏差となる。もし分布が正規分布ならば，Mean±2SD[*13]の範囲にデータの 95%が含まれるという意味で，標準偏差は便利な指標である。

1.5 図示

データの大局的性質を把握するには，図示をするのが便利である。人間の視覚的認識能力は，パターン認識に関してはコンピュータより遥かに優れていると言われているから，それを生かさない手はない。また，入力ミスをチェックする上でも有効である。変数が表す尺度の種類によって，さまざまな図示の方法があるので，それをざっと示すことにする。

離散変数の場合は，以下のものが代表的である。

度数分布図　値ごとの頻度を縦棒として，異なる値ごとに，この縦棒を横に並べた図である。離散変数の名前を X とすれば，R では `barplot(table(X))` で描画される。

積み上げ棒グラフ　値ごとの頻度の縦棒を積み上げた図である。R では

```
fx <- table(X)
barplot(matrix(fx,NROW(fx)),beside=F)
```

で描画される。

帯グラフ　横棒を全体を 100 %として各値の割合にしたがって区切って塗り分けた図である。R では

```
px <- table(X)/NROW(X)
barplot(matrix(pc,NROW(pc)),horiz=T,beside=F)
```

で描画される。

円グラフ（ドーナツグラフ・パイチャート）　円全体を 100 %として，各値の割合にしたがって中心から区切り線を引き，塗り分けた図である。ドーナツグラフでは２つの同心円にして，内側の円内を空白にする。R では `pie()` 関数を用いる。

連続変数の場合は，以下のものが代表的である。

ヒストグラム　変数値を適当に区切って度数分布を求め，分布の様子を見るものである。R では `hist()` 関数を用いる。

正規確率プロット　連続変数が正規分布しているかどうかを見るものである（正規分布に当てはまっていれば点が直線上に並ぶ）。R では `qqnorm()` 関数を用いる。

幹葉表示 (stem and leaf plot)　大体の概数（整数区切りとか 5 の倍数とか 10 の倍数にすることが多い）を縦に並べて幹とし，それぞれの概数に相当する値の細かい部分を葉として横に並べて作成する図。R では `stem()` 関数を用いる。

箱ヒゲ図 (box and whisker plot)　縦軸に変数値をとって，第１四分位を下に，第３四分位を上にした箱を書き，中央値の位置にも線を引いて，さらに第１四分位と第３四分位の差（四分位範囲）を 1.5 倍した線分をヒゲとして第１四分位の下と第３四分位の上に伸ばし，ヒゲの先より外れた値を外れ値として○をプロットした図である。カテゴリによって層別した箱ヒゲ図を横に並べて描くと，大体の分布の様子と外れ値の様子が同時に比較できるので便利である。R では `boxplot()` 関数を用いる。

レーダーチャート　複数の連続変数を中心点から放射状に数直線としてとり，データ点をつないで表される図である。それら複数の変数によって特徴付けられる性質のバランスをみるのに役立つ。１つのケースについて１つのレーダーチャートができるので，他のケースと比較するには，並べて描画するか，重ね描きする。R では `stars()` 関数を用いる。

[*13] 普通このように 2SD と書かれるが，正規分布の 97.5 パーセント点は 1.959964... なので，この 2 は，だいたい 2 くらいという意味である。

散布図 (scatter plot)　2つの連続変数の関係を2次元の平面上の点として示した図である。R では `plot()` 関数を用いる。異なる群ごとに別々のプロットをしたい場合は `plot()` の `pch` オプションで塗り分けたり，`points()` 関数を使って重ね打ちしたりできる。点ごとに異なる情報を示したい場合は `symbols()` 関数を用いることができるし，複数の連続変数間の関係を調べるために，重ね描きしたい場合は `matplot()` 関数と `matpoints()` 関数を，別々のグラフとして並べて同時に示したい場合は `pairs()` 関数を用いることができる。データ点に文字列を付記したい場合は `text()` 関数が使えるし，マウスで選んだデータ点にだけ文字列を付記したい場合は `identify()` 関数が使える。

2　独立2標本の差の検定

医学統計でよく使われるのは，伝統的に仮説検定である。仮説検定は，意味合いからすれば，元のデータに含まれる情報量を，仮説が棄却されるかどうかという2値情報にまで集約してしまうことになる。これは情報量を減らしすぎであって，点推定量と信頼区間を示す方がずっと合理的なのだが，伝統的な好みの問題なので，この演習でも検定を中心に説明する[*14]。

2群の差がないという帰無仮説を検定する（つまり，差がないという帰無仮説の元で，現在得られている値以上に差がある値が偶然得られる確率＝有意確率＝が，偶然ではありえないくらい小さいかどうかを調べる）方法は，以下のようにまとめられる。

1. 量的変数の場合
   (a) 正規分布に近い場合
      i. 2群の間で分散に差がないという帰無仮説で F 検定して仮説が棄却されない場合：t 検定（R では `t.test(x,y,var.equal=T)`）
      ii. 仮説が棄却される場合：Welch の検定（R では `t.test(x,y)`）
   (b) 正規分布とかけ離れている場合：Wilcoxon の順位和検定（R では `wilcox.test(x,y)`）
2. カテゴリ変数の場合：母比率の差の検定（R では `prop.test()`）。

これらの手法は，ほぼすべての初等統計の教科書に載っているが，簡単に説明しておく。

まず，標本調査によって得られた独立した2つの量的変数 X と Y（サンプル数が各々 n_X と n_Y とする）について，平均値に差があるかどうかを検定することを考える。

2.1　母分散が既知で等しい V である場合

$z_0 = |E(X) - E(Y)|/\sqrt{V/n_X + V/n_Y}$ が標準正規分布に従うことを使って検定する[*15]。

2.2　母分散が未知の場合

調査データを分析する場合は母分散が既知であることはほとんどなく，こちらが普通である。手順は以下の通り。

1. F 検定（分散が等しいかどうか）：2つの量的変数 `X` と `Y` の不偏分散 `SX<-var(X)` と `SY<-var(Y)` の大きい方を小さい方で（以下の説明では `SX>SY` だったとする）割った `F0<-SX/SY` が第1自由度 `DFX<-length(X)-1`，第2自由度 `DFY<-length(Y)-1` の F 分布に従うことを使って検定する。有意確率は `1-pf(F0,DFX,DFY)` で得られる。しかし，`F0` を手計算しなくても，`var.test(X,Y)` で等分散かどうかの検定が実行できる[*16]。また，1つ

[*14] もっとも，Rothman とか Greenland といった最先端の疫学者は，仮説検定よりも区間推定，区間推定よりも p 値関数の図示の方が遥かによい統計解析であると断言している。

[*15] 分布がひどく歪んでいる場合には，Mann-Whitney の U 検定（Wilcoxon の順位和検定と数学的に同値）を行う。後述するが，その場合は，代表値としても平均値と標準偏差でなく，中央値と四分位範囲または四分位偏差を表示するのが相応しい。

[*16] 拙著「R による統計解析の基礎」では，「この場合は，R が勝手に入れ替えてくれるので，X の不偏分散の方が Y の不偏分散より大きいかどうか気にしなくてもよい。」と書いていたが，少なくとも現在のバージョン 2.1.1 ではそうではなくて，古川・丹後「医学への統計学」（朝倉書店）で2つの方法の1つとして触れられている，「帰無仮説：SX=SY，対立仮説：SX≠SY」で大小を区別せず F 比を算出して両側検定するのがデフォルトになっているので注意されたい。

の量的変数 X と 1 つの群分け変数 C があって，C の 2 群間で X の分散が等しいかどうか検定するというスタイルでデータを入力してある場合は，var.test(X~C) とすればよい。

2. 分散に差があるか差がないかによって，平均値が等しいかどうかの検定法は異なる（以下に詳述）。分散に差があるときは，その事実をもって別の母集団からとられた標本であると判断し，平均値が等しいかどうかを検定する意味はないとする考え方もあるが，一般には Welch の方法を使うか，ノンパラメトリックな方法を使って検定する。

2.3 分散に差がない場合

母分散 S を S<-(DFX*SX+DFY*SY)/(DFX+DFY) として推定し，

```
t0 <- abs(mean(X)-mean(Y))/sqrt(S/length(X)+S/length(Y))
```

が自由度 DFX+DFY の t 分布に従うことから，帰無仮説「X と Y の平均値には差がない」を検定すると，(1-pt(t0,DFX+DFY))*2 が有意確率となる。

R では，t.test(X,Y,var.equal=T) とする。また，F 検定のところで触れた量的変数と群分け変数という入力の仕方の場合は，t.test(X~C,var.equal=T) とする。ただしこれだと両側検定なので，片側検定したい場合は，t.test(X,Y,var.equal=T,alternative="less") などとする（alternative="less"は対立仮説が X<Y という意味なので，帰無仮説が X>=Y であることを意味する）。

2.4 分散が差がある場合（Welch の方法）

$t_0 = |E(X) - E(Y)|/\sqrt{S_X/n_X + S_Y/n_Y}$ が自由度 ϕ の t 分布に従うことを使って検定する。但し，ϕ は下式による。

$$\phi = \frac{(S_X/n_X + S_Y/n_Y)^2}{\{(S_X/n_X)^2/(n_X-1) + (S_Y/n_Y)^2/(n_Y-1)\}}$$

R では, t.test(X,Y,var.equal=F) だが, var.equal の指定を省略した時は等分散でないと仮定して Welch の検定がなされるので省略して t.test(X,Y) でいい。量的変数 X と群分け変数 C という入力の仕方の場合は，t.test(X~C) とする。

なお，既に平均値と不偏標準偏差が計算されている場合の図示は，エラーバー付きの棒グラフを使うのが常道であるが[*17]，生データを図示する場合は stripchart() 関数を用いる。そのためには，量的変数と群別変数という形にしなくてはいけないので，たとえば，2 つの量的変数 V <- rnorm(100,10,2) と W <- rnorm(60,12,3) があったら，予め

```
X <- c(V,W)
C <- as.factor(c(rep("V",length(V)),rep("W",length(W))))
```

のように変換しておく必要がある。プロットするには次のように入力すればよい。

```
stripchart(X~C,method="jitter",vert=T)
MX <- tapply(X,C,mean); SX <- tapply(X,C,sd); IX <- c(1.1,2.1)
points(IX,MX,pch=18)
arrows(IX,MX-SX,IX,MX+SX,angle=90,code=3)
```

[*17] R では，barplot() 関数で棒グラフを描画してから，arrows() 関数でエラーバーを付ける。

対応のある２標本の平均値の差の検定

各対象について２つずつの値があるときは，それらを独立２標本とみなすよりも，対応のある２標本とみなす方が切れ味がよい。全体の平均に差があるかないかだけをみるのではなく，個人ごとの違いを見るほうが情報量が失われないのは当然である。
対応のある２標本の差の検定は，paired-t 検定と呼ばれ，意味合いとしてはペア間の値の差を計算して値の差の母平均が 0 であるかどうかを調べることになる。R で対応のある変数 `X` と `Y` の paired-t 検定をするには，`t.test(X,Y,paired=T)` で実行できるし，それは `t.test(X-Y,mu=0)` と等価である。

2.5 Wilcoxon の順位和検定

Wilcoxon の順位和検定は，パラメトリックな検定でいえば，t 検定を使うような状況，つまり，独立２標本の分布の位置に差がないかどうかを調べるために用いられる。Mann-Whitney の U 検定と（これら２つほど有名ではないが，Kendall の S 検定とも）数学的に等価である。

データがもつ情報の中で，単調変換に対して頑健なのは順位なので，これを使って検定しようという発想である。以下，Wilcoxon の順位和検定の手順を箇条書きする。

1. 変数 X のデータを $x_1, x_2, ..., x_m$ とし，変数 Y のデータを $y_1, y_2, ..., y_n$ とする。
2. まず，これらをまぜこぜにして小さい方から順に番号をつける[*18]。例えば，$x_8[1], y_2[2], y_{17}[3], ..., x_4[N]$ のようになる（但し $N = m + n$）。
3. ここで問題にしたいのは，それぞれの変数の順位の合計がいくつになるかということである。ただし，順位の総合計は $(N + 1)N/2$ に決まっているので，片方の変数だけ考えれば残りは引き算でわかる。そこで，変数 X だけ考えることにする。
4. X に属する x_i（$i = 1, 2, ..., m$）の順位を R_i と書くと，X の順位の合計は

$$R_X = \sum_{i=1}^{m} R_i$$

   となる。R_X があまり大きすぎたり小さすぎたりすると，X の分布と Y の分布に差がないという帰無仮説 $H0$ が疑わしいと判断されるわけである。では，帰無仮説が成り立つ場合に，R_X はどのくらいの値になるのだろうか？[*19]
5. もし X と Y に差がなければ，X は N 個のサンプルから偶然によって m 個取り出したものであり，Y がその残りである，と考えることができる。順位についてみると，$1, 2, 3, ..., N$ の順位から m 個の数値を取り出すことになる。同順位がなければ，ありうる組み合わせは，${}_NC_m$ 通りある[*20]。
6. $X > Y$ の場合には，${}_NC_m$ 通りのうち，合計順位が R_X と等しいかより大きい場合の数を k とする（$X < Y$ の場合は，合計順位が R_X と等しいかより小さい場合の数を k とする）。
7. $k/{}_NC_m$ が有意水準 α より小さいときに $H0$ を疑う。N が小さいときは有意になりにくいが，N が大きすぎると計算が大変面倒である[*21]。そこで，正規近似を行う（つまり，期待値と分散を求めて，統計量から期待値を引いて分散の平方根で割った値が標準正規分布に近似的に従うという関係を用いて検定する）。

[*18] 同順位がある場合の扱いは後述する。

[*19] 以下説明するように，順位和 R をそのまま検定統計量として用いるのが Wilcoxon の順位和検定であり，R_X, R_Y の代わりに，$U_X = mn + n(n + 1)/2 - R_Y$，$U_Y = mn + m(m + 1)/2 - R_X$ として，U_X と U_Y の小さいほうを U として検定統計量として用いるのが，Mann-Whitney の U 検定である。また，$U_X - U_Y$ を検定統計量とするのが Kendall の S 検定である。有意確率を求めるために参照する表が異なる（つまり帰無仮説の下で検定統計量が従う分布の平均と分散は，これら３つですべて異なる）が，数学的には等価な検定である。R では，Wilcoxon の順位和統計量の分布関数が提供されているので，例えばここで得られた順位和を RS と書くことにすると，`2*(1-pwilcox(RS,m,n))` で両側検定の正確な有意確率が得られる。

[*20] R では `choose(N,m)` によって得られる。

[*21] もっとも，今ではコンピュータにやらせればよい。例えば R であれば，`wilcox.test(X,Y,exact=T)` とすれば，サンプル数の合計が 50 未満で同順位の値がなければ，総当りして正確な確率を計算してくれる。が，つい 15 年くらいまではコンピュータは誰もが使える道具ではなかったし，総当りをするには計算時間がかかりすぎた。今のコンピュータでもサンプルサイズが大きいと，総当りでは計算時間がかかりすぎて実用的でない。

8. 帰無仮説 H_0 のもとでは，期待値は

$$E(R) = \sum_{i=1}^{m} E(R_i) = m(1 + 2 + ... + N)/N = m(N + 1)/2$$

(1 から N までの値を等確率 $1/N$ でとるから)。分散はちょっと面倒で，

$$var(R) = E(R^2) - (E(R)^2)$$

から，

$$E(R^2) = E((\sum_{i=1}^{m} R_i)^2) = \sum_{i=1}^{m} E(R_i^2) + 2\sum_{i<j} E(R_i R_j)$$

となるので[*22]，

$$E(R_i^2) = (1^2 + 2^2 + ... + N^2)/N = (N + 1)(2N + 1)/6$$

と

$$E(R_i R_j) = \frac{1}{N(N-1)}\{(\sum_{k=1}^{N} k)^2 - \sum_{k=1}^{N} k\}$$

$$= \frac{1}{N(N-1)}(\frac{N^2(N+1)^2}{4} - \frac{N(N+1)(2N+1)}{6})$$

$$= \frac{(N+1)(3N+2)}{12}$$

を代入して整理すると，結局，$var(R_X) = m(N + 1)(N - m)/12 = mn(N + 1)/12$ となる。

9. 標準化[*23]して連続修正[*24]し，$z_0 = \{|R_X - E(R_X)| - 1/2\}/\sqrt{var(R_X)}$ を求める。m と n が共に大きければこの値が標準正規分布に従うので，例えば $z_0 > 1.96$ ならば，両側検定で有意水準 5%で有意である。R で有意確率を求めるには，z_0 を `z0` と書けば，`2*(1-pnorm(z0,0,1))` とすればよい。

10. ただし，同順位があった場合は，ステップ 2 の「小さい方から順に番号をつける」ところで困ってしまう。例えば，変数 X が $\{2, 6, 3, 5\}$，変数 Y が $\{4, 7, 3, 1\}$ であるような場合には，X にも Y にも 3 という値が含まれる。こういう場合は，下表のように平均順位を両方に与えることで，とりあえず解決できる。

属する変数	Y	X	X	Y	Y	X	X	Y
値	1	2	3	3	4	5	6	7
順位	1	2	3.5	3.5	5	6	7	8

11. ただし，このやり方では，正規近似をする場合に分散が変わる[*25]。帰無仮説の下で，$E(R_X) = m(N + 1)/2$ はステップ 8 と同じだが，分散が

$$var(R_X) = mn(N + 1)/12 - mn/\{12N(N - 1)\} \cdot \sum_{t=1}^{T}(d_t^3 - d_t)$$

となる。ここで T は同順位が存在する値の総数であり，d_t は t 番目の同順位のところにいくつのデータが重なっているかを示す。上の例では，$T = 1$，$d_1 = 2$ となる。なお，あまりに同順位のものが多い場合は，この程度の補正では追いつかないので，値の大小があるクロス集計表として分析することも考慮すべきである（例えば Cochran-Armitage 検定などが考えられる)。

[*22] 第 1 項が対角成分，第 2 項がそれ以外に相当する。$m = 2$ の場合を考えてやればわかるが，

$$E((\sum_{i=1}^{2} R_i)^2) = E((R_1 + R_2)^2) = E(R_1^2 + R_2^2 + 2R_1R_2) = \sum_{i=1}^{2} E(R_i^2) + 2\sum_{i<j} E(R_i R_j)$$

となる。

[*23] 何度も出てくるが，平均（期待値）を引いて分散の平方根で割る操作である。

[*24] これも何度も出てくるが，連続分布に近づけるために 1/2 を引く操作である。

[*25] 正確な確率を求めることができれば問題ないけれども，同順位がある場合には，R では正確な確率は求められない。

2.6 2 群の母比率の差の検定

たとえば，患者群 n_1 名と対照群 n_2 名の間で，ある特性をもつ者の人数がそれぞれ r_1 名と r_2 名だったとして，その特性の母比率に差がないという帰無仮説を考える。

2 群の母比率 p_1, p_2 が，各々の標本比率 $\hat{p_1} = r_1/n_1, \hat{p_2} = r_2/n_2$ として推定されるとき，それらの差を考える。差 $(\hat{p_1} - \hat{p_2})$ の平均値と分散は，$E(\hat{p_1} - \hat{p_2}) = p_1 - p_2, V(\hat{p_1} - \hat{p_2}) = p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2$ となる。2 つの母比率に差が無いならば，$p_1 = p_2 = p$ とおけるはずなので，$V(\hat{p_1} - \hat{p_2}) = p(1-p)(1/n_1 + 1/n_2)$ となる。この p の推定値として，$\hat{p} = (r_1 + r_2)/(n_1 + n_2)$ を使い，$\hat{q} = 1 - \hat{p}$ とおけば，$n_1 p_1$ と $n_2 p_2$ がともに 5 より大きければ，標準化して正規近似を使い，

$$Z = \frac{\hat{p_1} - \hat{p_2} - E(\hat{p_1} - \hat{p_2})}{\sqrt{V(\hat{p_1} - \hat{p_2})}} = \frac{\hat{p_1} - \hat{p_2}}{\sqrt{\hat{p}\hat{q}(1/n_1 + 1/n_2)}} \sim N(0,1)$$

によって[*26]検定できる。

数値計算をしてみるため，仮に，患者群 100 名と対照群 100 名で，喫煙者がそれぞれ 40 名，20 名だったとする。喫煙率に 2 群間で差がないという帰無仮説を検定するには，

```
p <- (40+20)/(100+100)
q <- 1-p
Z <- (abs(40/100-20/100)-(1/100+1/100)/2)/sqrt(p*q*(1/100+1/100))
2*(1-pnorm(Z))
```

より，有意確率が約 0.0034 となるので，有意水準 5%で帰無仮説は棄却される。つまり，喫煙率に 2 群間で差がないとはいえないことになる。

差の 95%信頼区間を求めるには，サンプルサイズが大きければ正規分布を仮定できるので，原則どおりに差から分散の平方根の 1.96 倍を引いた値を下限，足した値を上限とすればよい。この例では，

```
dif <- 40/100-20/100
vardif <- 40/100*(1-40/100)/100+20/100*(1-20/100)/100
difL <- dif - qnorm(0.975)*sqrt(vardif)
difU <- dif + qnorm(0.975)*sqrt(vardif)
cat("喫煙率の差の点推定値=",dif,"  95%信頼区間= [",difL,",",difU,"]\n")
```

より，[0.076,0.324] となる。しかし，通常は連続性の補正を行うので，下限からはさらに $(1/n_1 + 1/n_2)/2 = (1/100 + 1/100)/2 = 0.01$ を引き，上限には同じ値を加えて，95%信頼区間は [0.066,0.334] となる。

R には，こうした比率の差を検定するための関数 `prop.test()` が用意されており，以下のように簡単に実行することができる。

```
smoker <- c(40,20)
pop <- c(100,100)
prop.test(smoker,pop)
```

母比率の推定と，その差があるかどうかの検定[*27]，差の 95%信頼区間を一気に出力してくれる。上で一段階ずつ計

[*26] この Z は離散値しかとれないため，連続分布である正規分布による近似の精度を上げるために，連続性の補正と呼ばれる操作を加え，かつ $p_1 > p_2$ の場合（つまり $Z > 0$ の場合）と $p_1 < p_2$ の場合 (つまり $Z < 0$ の場合）と両方考える必要があり，正規分布の対称性から絶対値をとって $Z > 0$ の場合だけ考え，有意確率を 2 倍する。即ち，

$$Z = \frac{|\hat{p_1} - \hat{p_2}| - (1/n_1 + 1/n_2)/2}{\sqrt{\hat{p}\hat{q}(1/n_1 + 1/n_2)}}$$

として，この Z の値が標準正規分布の 97.5%点（R ならば `qnorm(0.975,0,1)`）より大きければ有意水準 5%で帰無仮説を棄却する。

[*27] 連続性の補正済み，事象が生起しない場合についても考慮してカイ二乗適合度検定をしているのだが，この操作は次回説明する 2 つの変数の

算した結果と一致することを確かめてみよう。

3 群以上の母比率の差の検定

prop.test() 関数は，3 群以上の間でも，「どの群でも事象の生起確率に差がない」という帰無仮説を検定するのに使える。その帰無仮説が棄却されるときに，どの群間で差があるのかをみるには，検定の多重性（後述）が生じるので，平均値の差の場合と同様，第一種の過誤を調整する必要があり，ボンフェローニの方法やホルムの方法を用いることができる。R の関数は pairwise.prop.test() である。なお，3 群以上の間で事象の生起確率に一定の傾向がみられるかどうかを調べたい場合には，コクラン＝アーミテージの検定という手法がある。例えば，漁師 100 人，農民 80 人，事務職 30 人について便の検査をして，日本住血吸虫虫卵陽性者が 60 人，30 人，8 人だったとしたとき，職業的な貝との接触リスクに対して勝手に漁師を 4，農民を 2，事務職を 1 とスコアリングして，陽性割合の増加傾向が，このスコアと同じかどうかを調べることができる。この場合なら，R のコマンドは以下のようになる。

```
total <- c(100,80,30)
epos <- c(60,30,8)
orisk <- c(4,2,1)
prop.trend.test(epos,total,orisk)
```

3 分散分析と多重比較

3 群以上を比較するために，単純に 2 群間の差の検定を繰り返すことは誤りである。なぜなら，n 群から 2 群を抽出するやりかたは ${}_nC_2$ 通りあって，1 回あたりの第 1 種の過誤（本当は差がないのに，誤って差があると判定してしまう確率）を 5%未満にしたとしても，3 群以上の比較全体として「少なくとも 1 組の差のある群がある」というと，全体としての第 1 種の過誤が 5%よりずっと大きくなってしまうからである。

この問題を解消するには，多群間の比較という捉え方をやめて，群分け変数が注目している量の変数に与える効果があるかどうかという捉え方にするのが一つの方法であり，具体的には一元配置分散分析やクラスカル＝ウォリス (Kruskal-Wallis) の検定がこれに当たる[*28]。

そうでなければ，有意水準 5%の 2 群間の検定を繰り返すことによって全体として第 1 種の過誤が大きくなってしまうことが問題なので，第 1 種の過誤を調整することによって全体としての検定の有意水準を 5%に抑える方法もある。このやり方は「多重比較法」と呼ばれる。

独立性のカイ二乗検定と数学的に等価である。

[*28] なお，分散分析は本来，その効果をみるための実験計画をした上で実施するものだから，群ごとのサンプルサイズは揃っているべきだし，効果の有無を効率よく検出するのに適したサンプルサイズが設計されているべきだが，現実には実験計画されていないデータにも適用されている。適切なサンプルサイズは，母集団の均質性，サブグループ数，母集団のパラメータ推定に求めたい正確さ，注目している現象の出現頻度，予算などで変わってくる。詳しくは，永田靖 (2003) サンプルサイズの決め方，朝倉書店を参照されたいが，有意水準 5%，検出力 90%の場合なら，以下の式によって求めるのが基本となる。

- 2 つの集団の平均値の差を調べる場合：予測される標本平均が m1, m2，標本分散が d1, d2 なら，サンプルサイズは

$$\frac{(1.96+1.28)^2(d1+d2)}{(m1-m2)^2}$$

- 2 つの集団の罹患率の差を調べる場合：2 つの集団で予測される罹患率がそれぞれ r1，r2 なら，サンプルサイズは

$$\frac{(1.96+1.28)^2(r1+r2)}{(r1-r2)^2}$$

- 2 つの集団の比率の差を調べる場合：期待される比率を p1，p2 とすると，サンプルサイズは，

$$\frac{\{1.28\sqrt{p1(1-p1)+p2(1-p2)}+1.96\sqrt{(p1+p2)(1-(p1+p2)/2)}\}^2}{(p1-p2)^2}$$

3.1 一元配置分散分析

一元配置分散分析では，データのばらつき（変動）を，群間の違いという意味のはっきりしているばらつき（群間変動）と，各データが群ごとの平均からどれくらいばらついているか（誤差）をすべての群について合計したもの（誤差変動）に分解して，前者が後者よりもどれくらい大きいかを検討することによって，群分け変数がデータの変数に与える効果があるかどうかを調べる。

例えば，南太平洋の 3 つの村 X, Y, Z で健診をやって，成人男性の身長や体重を測ったとしよう。このとき，データは例えば次のようになる（架空のものである）[*29]。

ID 番号	村落 (VG)	身長 (cm)(HEIGHT)
1	X	161.5
2	X	167.0
（中略）		
22	Z	166.0
（中略）		
37	Y	155.5

村落によって身長に差があるかどうかを検定したいならば，HEIGHT という量的変数に対して，VG という群分け変数の効果があるかどうかを一元配置分散分析することになる。R でデータを読み込んでから，`summary(aov(HEIGHT ~ VG))` とすれば（Rcmdr の場合は，メニューバーの Statistics から Means の One-Way ANOVA を選ぶ），例えば次のような結果が得られる。

```
            Df  Sum Sq Mean Sq F value   Pr(>F)
VG           2  422.72  211.36  5.7777 0.006918 **
Residuals   34 1243.80   36.58
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

このような結果の表を分散分析表という。右端の*の数は有意性を示す目安だが，確率そのものに注目してみるほうがよい。Sum Sq のカラムは偏差平方和を意味する。VG の Sum Sq の値 422.72 は，村ごとの平均値から総平均を引いて二乗した値を村ごとの人数で重み付けした和であり，群間変動または級間変動と呼ばれ，VG 間でのばらつきの程度を意味する。Residuals の Sum Sq の値 1243.80 は各個人の身長からその個人が属する村の平均身長を引いて二乗したものの総和であり，誤差変動と呼ばれ，村によらない（それ以外の要因がないとすれば偶然の）ばらつきの程度を意味する。Mean Sq は平均平方和と呼ばれ，偏差平方和を自由度 (Df) で割ったものである。平均平方和は分散なので，VG の Mean Sq の値 211.36 は群間分散または級間分散と呼ばれることがあり，Residuals の Mean Sq の値 36.58 は誤差分散と呼ばれることがある。F value は分散比と呼ばれ，群間分散の誤差分散に対する比である。この場合の分散比は第 1 自由度 2，第 2 自由度 34 の F 分布に従うことがわかっているので，それを使った検定の結果，分散比がこの実現値よりも偶然大きくなる確率（Pr(>F) に得られる）が得られる。この例では 0.006918 なので，VG の効果は 5%水準で有意であり，帰無仮説は棄却される。つまり，身長は村落によって有意に異なることになる。

3.2 クラスカル＝ウォリス (Kruskal-Wallis) の検定

一元配置の分散分析は，各群が等しい母分散をもつ正規分布に従うことを仮定して行っているパラメトリックな分析法なので，データの分布がひどく歪んでいる場合は結果がおかしくなる場合がある[*30]。そこで，多群間の差を調べる

[*29] http://phi.med.gunma-u.ac.jp/grad/sample2.dat として公開してあり，R から read.delim() 関数で読み込み可能な筈である。

[*30] 各群の母分散が等しいかどうかを調べる検定法として，バートレット (Bartlett) の検定と呼ばれる方法がある。R では，量的変数を Y，群分け変数を C とすると，bartlett.test(Y~C) で実行できる。同じ目的のノンパラメトリックな方法として，Fligner-Killeen の検定という方法もあり，fligner.test(Y~C) で実行できる。また，量的変数について，母集団で正規分布しているかどうかを調べる方法としては，既に説明したヒストグラムや正規確率プロットなどのグラフ表示による方法の他に，シャピロ＝ウィルク (Shapiro-Wilk) の検定と呼ばれる方法もある。詳しくは説明しないが，R では shapiro.test(Y) で実行できる。厳密に言えば，これらの検定で等分散性と分布の正規性が確認されない限り，一元配置分散分析の結果を解釈するには注意が必要なのだが，論文や本でもそこまで考慮されずに使われていることが多い。

ためにもノンパラメトリックな方法がある。クラスカル＝ウォリス (Kruskal-Wallis) の検定と呼ばれる方法である。R では，量的変数を `Y`，群分け変数を `C` とすると，`kruskal.test(Y~C)` で実行できる。以下，Kruskal-Wallis の検定の仕組みを箇条書きで説明する。

- 「少なくともどれか 1 組の群間で大小の差がある」という対立仮説に対する「すべての群の間で大小の差がない」という帰無仮説を検定する。
- まず 2 群の比較の場合の順位和検定と同じく，すべてのデータを込みにして小さい方から順に順位をつける（同順位がある場合は平均順位を与える）。
- 次に，各群ごとに順位を足し合わせて，順位和 $R_i (i = 1, 2, ..., k;\ k$ は群の数) を求める。
- 各群のオブザーベーションの数をそれぞれ n_i とし，全オブザーベーション数を N としたとき，各群について統計量 B_i を $B_i = n_i\{R_i/n_i - (N+1)/2\}^2$ として計算し，

$$B = \sum_{i=1}^{k} B_i$$

として B を求め，$H = 12 \cdot B/\{N(N+1)\}$ として H を求める。同順位を含むときは，すべての同順位の値について，その個数に個数の 2 乗から 1 を引いた値を掛けたものを計算し，その総和を A として，

$$H' = \frac{H}{1 - \frac{A}{N(N^2-1)}}$$

により H を補正した値 H' を求める。

- H または H' から表を使って（データ数が少なければ並べかえ検定によって）有意確率を求めるのが普通だが，$k \geq 4$ で各群のオブザーベーション数が最低でも 4 以上か，または $k = 3$ で各群のオブザーベーション数が最低でも 5 以上なら，H や H' が自由度 $k - 1$ のカイ二乗分布に従うものとして検定できる。

3.3 検定の多重性の調整

仮に，上述の南太平洋の島の 3 つの村での健診の例で，一元配置分散分析か Kruskal-Wallis の検定で有意差があったときに，具体的にどの村の間に有意差があるのかを調べるには，単純に考えると，t 検定[*31]や順位和検定[*32]を繰り返せば良さそうである。この方法が使われている本や論文もないわけではない。しかし，3 つの村でこれをやると 3 つから 2 つを取り出す全ての組み合わせについて検定するので，3 回の比較をすることになり，個々の検定について有意水準を 5%にすると，全体としての第 1 種の過誤は明らかに 5%より大きくなる。もし村が 7 つあったら，7 つから 2 つを取り出す組み合わせは 21 通りあるので，1 つくらいは偶然によって有意差が出てしまう比較があっても全然おかしくない。したがって，先に述べた通り，t 検定の繰り返しは第 1 種の過誤が大きくなってしまって不都合である。これに似た方法として無制約 LSD（最小有意差）法や Fisher の制約つき LSD 法（一元配置分散分析を行って有意だった場合にのみ LSD 法を行うという方法）があるが，これらも第 1 種の過誤を適切に調整できない（ただし制約つきの場合は 3 群なら大丈夫）ことがわかっているので，使ってはいけない。現在では，この問題は広く知られているので，t 検定の繰り返しや LSD 法で分析しても論文は accept されない。

多重比較の方法にはいろいろあるが，良く使われているものとして，ボンフェローニ (Bonferroni) の方法，シェフェ (Scheffé) の方法，ダンカン (Duncan) の方法，テューキー (Tukey) の HSD，ダネット (Dunnett) の方法，ウィリアムズ (Williams) の方法がある。しかしこの中で，ダンカンの方法は，新多範囲検定などと呼ばれた時期もあったが，数学的に間違っていることがわかっているので，使ってはいけない。ボンフェローニの方法とシェフェの方法も検出力が悪いので，特別な場合を除いては使わない方が良い。せめてテューキーの HSD を使うべきである。ダネットの方法は対照群が存在する場合に対照群と他の群との比較に使われるので，適用場面が限定されている[*33]。ウィリアムズの方

[*31] R では `t.test(height[vg=="X"],height[vg=="Y"])` など。

[*32] R では `wilcox.test(height[vg=="X"],height[vg=="Y"])` など。

[*33] ただし，対照群が他の群との比較のすべての場合において差があるといいたい場合は，多重比較をするのではなくて，t 検定を繰り返して使うのが正しいので，注意が必要である。もちろんそういう場合は多くはない。

法は対照群があって他の群にも一定の傾向が仮定される場合には最高の検出力を発揮するが，ダネットの方法よりもさらに限られた場合にしか使えない。

上記いくつかの方法が良く使われている原因は，用途が限定されているダネットとウィリアムズを除けば，たんにそれらが歴史的に古く考案され，昔の統計学の教科書にも説明されているからに過ぎない。現在では，かなり広い用途をもち，ノンパラメトリックな分析にも適応可能なホルム (Holm) の方法（ボンフェローニの方法を改良して開発された方法）が第一に考慮されるべきである。その上で，全ての群間の比較をしたい場合はペリ (Peritz) の方法，対照群との比較をしたいならダネットの逐次棄却型検定（これはステップダウン法と呼ばれる方法の１つであり，既に触れたダネットの方法とは別）も考慮すればよい。とはいえ，ソフトウェアによってはこれらの方法をサポートしていない場合もあると思われ，その場合はテューキーの HSD を使うべきである（もちろん場合によっては，ダネットかウィリアムズを使い分けねばならない）[*34]。

多重比較においては，帰無仮説が単純ではない。例えば，4 群間の差を調べるとしよう。一元配置分散分析での帰無仮説は，$\mu_1 = \mu_2 = \mu_3 = \mu_4$ である。これを包括的帰無仮説と呼び，$H_{\{1,2,3,4\}}$ と書くことにする。さて第 1 群から第 4 群までの母平均 $\mu_1 \sim \mu_4$ の間で等号関係が成り立つ場合をすべて書き上げてみると，$H_{\{1,2,3,4\}} : \mu_1 = \mu_2 = \mu_3 = \mu_4$，$H_{\{1,2,3\}} : \mu_1 = \mu_2 = \mu_3$，$H_{\{1,2,4\}} : \mu_1 = \mu_2 = \mu_4$，$H_{\{1,3,4\}} : \mu_1 = \mu_3 = \mu_4$，$H_{\{2,3,4\}} : \mu_2 = \mu_3 = \mu_4$，$H_{\{1,2\},\{3,4\}} : \mu_1 = \mu_2$ かつ $\mu_3 = \mu_4$，$H_{\{1,3\},\{2,4\}} : \mu_1 = \mu_3$ かつ $\mu_2 = \mu_4$，$H_{\{1,4\},\{2,3\}} : \mu_1 = \mu_4$ かつ $\mu_2 = \mu_3$，$H_{\{1,2\}} : \mu_1 = \mu_2$，$H_{\{1,3\}} : \mu_1 = \mu_3$，$H_{\{1,4\}} : \mu_1 = \mu_4$，$H_{\{2,3\}} : \mu_2 = \mu_3$，$H_{\{2,4\}} : \mu_2 = \mu_4$，$H_{\{3,4\}} : \mu_3 = \mu_4$ の 14 通りである。このうち，$H_{\{1,2,3,4\}}$ 以外のものを部分帰無仮説と呼ぶ。すべての２つの群の組み合わせについて差を調べるということは，$\{H_{\{1,2\}}, H_{\{1,3\}}, H_{\{1,4\}}, H_{\{2,3\}}, H_{\{2,4\}}, H_{\{3,4\}}\}$ が，考慮すべき部分帰無仮説の集合となる。一方，例えば第 1 群が対照群であって，他の群のそれぞれが第 1 群と差があるかどうかを調べたい場合は，$\{H_{\{1,2\}}, H_{\{1,3\}}, H_{\{1,4\}}$ が考慮すべき帰無仮説の集合となる。これらの集合をその多重比較における「帰無仮説族」と呼ぶ。

ここで多重比較の目的を「帰無仮説族」というコトバを使って言い換えてみる。個々の帰無仮説で有意水準を 5%にしてしまうと，帰無仮説族に含まれる帰無仮説のどれか１つが誤って棄却されてしまう確率が 5%より大きくなってしまう。それではまずいので，その確率が 5%以下になるようにするために，何らかの調整を必要とするわけで，この調整をする方法が多重比較なのである。つまり，帰無仮説族の有意水準を定める（例えば 5%にする）ことが，多重比較の目的である[*35]。

R では，`pairwise.t.test(HEIGHT,VG,p.adjust.method="bonferroni")` とすれば，ボンフェローニの方法で有意水準を調整した，すべての村落間での身長の差を t 検定した結果を出してくれる[*36]。

また，`pairwise.wilcox.test(HEIGHT,VG,p.adjust.method="bonferroni")` とすれば，ボンフェローニの方法で有意水準を調整した，すべての村落間での身長の差を順位和検定した結果を出してくれる。これらの関数で，`p.adjust.method` を指定しなければホルムの方法になるが，明示したければ，`p.adjust.method="holm"` とすればよい。R でもボンフェローニが可能なのは，一番単純な方法であるという理由と，ホルムの方法に必要な計算がボンフェローニの計算を含むからだと思われる。なお，R を使って分析するのだけれども，データがきれいな正規分布をしていて，かつ古典的な方法の論文しか accept しない雑誌に対してどうしても投稿したい，という場合は，`TukeyHSD(aov(HEIGHT~VG))` などとして，テューキーの HSD を行うことも可能である。また，CRAN (`http://cran.r-project.org/`) から `multcomp` パッケージをインストールすることによって，`simtest(HEIGHT ~ VG, type="Dunnett")` あるいは `simtest(HEIGHT ~ VG, type="Williams")` としてダネットやウィリアムズの方法を使うことも可能である（ただし，この例でこれらの方法を使うことは不適切である）。

Rcmdr の場合なら，Statistics メニューの Means から One-Way ANOVA を選んで実行するときに，Pairwise comparisons of means に左にチェックを入れておけば，自動的に Tukey の HSD で検定の多重性を調整してくれる。

[*34] もっとも，オープンソースで多くのコンピュータで無料で使える R がホルムの方法をデフォルトとしている現実を考えれば，そういう言い訳はもはや通用しない。

[*35] このことからわかるように，差のなさそうな群をわざと入れておいて帰無仮説族を棄却されにくくしたり，事後的に帰無仮説を追加したりすることは，統計を悪用していることになり，やってはいけない。

[*36] ただし，t 検定とは言っても，`pool.sd=F` というオプションをつけない限りは，t_0 を計算するときに全体の誤差分散を使うので，ただの t 検定の繰り返しとは違う。

4 相関と回帰

4.1 相関

相関も回帰も 2 つの変数間の関係を調べるという点は共通している。どちらを調べる場合にも，まずすべきことは散布図を描くことである。先に挙げた南太平洋の 3 つの村の男性のデータについて，身長と体重の関係を調べるためには，たとえば，身長を横軸，体重を縦軸にとって，二次元の散布図を描くことになる。R では，`plot(HEIGHT,WEIGHT,pch=paste(VG))` とすれば，村落によってプロット記号を変えた散布図を描くことができる。

相関と回帰は混同されやすいが，思想はまったく違う。相関は，変数間の関連の強さを表すものである。回帰は，ある変数の値のばらつきが，どの程度他の変数の値のばらつきによって説明されるかを示すものである。回帰の際に，説明される変数を従属変数または目的変数，説明するための変数を独立変数または説明変数と呼ぶ。2 つの変数間の関係を予測に使うためには，回帰を用いる。

関係とか関連とかいっても，その中身は多様である。例えば，$pV = nRT$ のような物理法則は，測定誤差を別にすれば 100%成り立つ関係である。身長と体重の間の関係はそうではないが，無関係ではないことは直感的にも理解できるし，散布図を見ても「身長の高い人は体重も概して重い傾向がある」ことは間違いない。一般に，2 個以上の変量が「かなりの程度の規則正しさをもって，増減をともにする関係」のことを相関関係 (correlation) という。相関には正の相関 (positive correlation) と負の相関 (negative correlation) があり，一方が増えれば他方も増える場合を正の相関，一方が増えると他方は減る場合を負の相関と呼ぶ。例えば，身長と体重の関係は正の相関である。

相関関係があっても，それが見かけ上のものである（それらの変量がともに，別の変量と真の相関関係をもっている）場合がある。具体例としては，血圧と所得の間に正の相関があるという命題は，データをとってみれば，多くの場合に成り立つであろう。しかしこれは，おそらくどちらも年齢や摂取エネルギー量との間に真の相関関係が存在するのであって，それらの影響を制御したら（例えば同年齢で同じような食生活をしている人だけについて見る，という層別化をしたら），血圧と所得の間の正の相関は消えてしまうだろう。この場合，見かけ上の相関があることは，たまたまそのデータで成り立っているだけであって，科学的仮説としての意味に乏しい。

時系列データや地域相関のデータでは，擬似相関 (spurious correalation) が見られる場合もある。例えば，日本の砂糖輸入量と溺死・溺水者数の年次別データをプロットしてみると，負の相関関係があるように見えるのだが，両者の間には真の関係はない。ある年に日本で植えた木の幹の太さと，同じ年にイギリスで生れた少年の身長を 15 年分，毎年 1 回測ったデータをプロットすると，おそらくは正の相関関係があるように見えるのだが，両者の間には関係がないのは明らかである（どちらも年次と真の相関があるとはいえるだろうが)。複数の種類の異なるデータをまとめて見ることで見かけの相関が生じてしまう場合もあるので，注意が必要である。

上で定義したように，相関関係は増減をともにする関係であればいいので，その関係が線形（一次式で表される，散布図で直線として表される）であろうと非線形（二次式以上または階段関数などで表される）であろうと問題ない。しかし，一般には，線形の関係があるという限定的な意味で使われる場合が多い。なぜなら，相関を表すための代表的な指標である相関係数[*37] r が，線形の関係を示すための指標だからである。もっといえば，r が意味をもつためには，厳密に言えば，2 つの変量が二次元正規分布に従っていなければならない。

非線形の相関関係を捉えるには，2 つのアプローチがある。1 つは線形になるように対数変換などの変換をほどこすことで，もう 1 つはノンパラメトリックな相関係数（分布の形によらない，例えば順位の情報だけを使った相関係数）を使うことである。ノンパラメトリックな相関係数にはスピアマン (Spearman) の順位相関係数 ρ や，ケンドール (Kendall) の順位相関係数 τ がある。

ピアソンの積率相関係数とは，X と Y の共分散を X の分散と Y の分散の積の平方根で割った値である。式で書けば，相関係数の推定値 r は，X の平均を $\bar{X}$，Y の平均を $\bar{Y}$ と書けば，

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

[*37] 普通，ただ相関係数といえば，ピアソンの積率相関係数 (Pearson's Product Moment Correlation Coefficient) を指し，通常，r という記号で表す。

となる。母相関係数がゼロかどうかという両側検定のためには，それがゼロであるという帰無仮説の下で，検定統計量

$$t_0 = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

が，自由度 $n-2$ の t 分布に従うことを利用して検定すればよい。

なお R では，

```
r <- cov(X,Y)/sqrt(var(X)*var(Y))
n <- NROW(X)
t0 <- r*sqrt(n-2)/sqrt(1-r^2)
```

として，`2*(1-pt(t0,n-2))` で有意確率が得られるが，`cor.test()` 関数(下記)を使う方が簡単である[*38]。`cor.test()` 関数を使った場合は，信頼区間も計算される。なお，信頼区間は，サンプルサイズがある程度大きければ（通常は 20 以上），正規近似を使って計算できる。すなわち，

$$a = \frac{1}{2}\ln\frac{1+r}{1-r} - \frac{1}{\sqrt{n-3}}Z(\alpha/2), \quad b = \frac{1}{2}\ln\frac{1+r}{1-r} + \frac{1}{\sqrt{n-3}}Z(\alpha/2)$$

と書くことにすると[*39]，母相関係数の $100\times(1-\alpha)\%$信頼区間の下限は $(\exp(2a)-1)/(\exp(2a)+1)$，上限は $(\exp(2b)-1)/(\exp(2b)+1)$ である[*40]。

順位相関係数は，非線形の相関関係を捉えたい場合以外にも，分布が歪んでいたり，外れ値がある場合に使うと有効である。スピアマンの順位相関係数 ρ は[*41]，値を順位で置き換えた（同順位には平均順位を与えた）ピアソンの積率相関係数になる。X_i の順位を R_i，Y_i の順位を Q_i とかけば，

$$\rho = 1 - \frac{6}{n(n^2-1)}\sum_{i=1}^{n}(R_i - Q_i)^2$$

となる。スピアマンの順位相関係数がゼロかどうかという両側検定は，サンプル数が 10 以上ならばピアソンの場合と同様に，$T = \frac{\rho\sqrt{n-2}}{\sqrt{1-\rho^2}}$ が自由度 $n-2$ の t 分布に従うことを利用して行うことができる。

ケンドールの順位相関係数 τ は，

$$\tau = \frac{(A-B)}{n(n-1)/2}$$

によって得られる。ここで A は順位の大小関係が一致する組の数，B は不一致数である。

いずれにせよ，R では `cor.test(X, Y, method="pearson")` とすればピアソンの相関係数が得られる[*42]。同時に，`alternative` を指定しないときは，「相関係数がゼロである」を帰無仮説として両側検定した有意確率と 95%信頼区間が表示される。なお，例えば `cor.test(X, Y, alternative="g")` とすれば，ピアソンの相関係数が計算され，対立仮説を「正の相関がある」とした片側検定の結果が得られる。なお，ケンドールに関しては並べ換えによる正確な確率も求めることができ，その場合は `exact=T` というオプションを指定する。

4.2 2 つのカテゴリ変数間の関係

2 つの量的変数間の関係を調べるには相関をみればよいわけだが，カテゴリ変数間の関係を調べるにはどうしたらよいだろうか？　もちろん，カテゴリ変数についても関連の強さをみる指標はあって，ファイ係数（記号は ρ を用いるのが普通）と呼ばれる指標は，要因の有無，発症の有無を 1,0 で表した場合のピアソンの積率相関係数と同じ計算式で得られる。θ_1, θ_2 を発症者中の要因あり割合，非発症者中の要因あり割合として，$\rho = \sqrt{(\pi_1-\pi_2)(\theta_1-\theta_2)}$ である。また，疫学研究では，人数あるいは人年の比を取ることによって，要因があった群が，要因がなかった群に比べて，どれ

[*38] また，`cov(X,Y)/sqrt(var(X)*var(Y))` と同値な関数として `cor(X,Y)` がある。

[*39] $Z(\alpha/2)$ は，標準正規分布の $100\times(1-\alpha/2)$ パーセント点，つまり R では（α を `alpha` と書くことにして），`qnorm(1-alpha/2,0,1)` である。例えば $\alpha = 0.05$ なら，`qnorm(0.975,0,1)` である。

[*40] なお，ln は自然対数，exp は指数関数を表す。

[*41] ピアソンの相関係数の母相関係数を ρ と書き，スピアマンの順位相関係数を r_s と書く流儀もある。

[*42] `method="spearman"`ならスピアマンの順位相関係数，`method="kendall"`ならケンドールの順位相関係数が得られる。

くらい発症しやすいかを調べることが多い（オッズ比やリスク比やハザード比を求め，その 95%信頼区間が 1 を含まないかどうかで，要因の有無が発症の有無に有意に影響しているかどうかを判定することが慣例的に行われる)。

しかし，2 つのカテゴリ変数の関係を考えるとき，一般に，もっともよく行われるのは，それらが独立であるという帰無仮説を立てて検定することである。

カテゴリ変数のもつ統計的な情報は，カテゴリごとの度数だけである。そこで，2 つのカテゴリ変数の間に関係について検討したいときには，まず**それらの組み合わせの度数**を調べた表を作成する（R では `table()` という関数が使える)。これをクロス集計表と呼ぶ。とくに，2 つのカテゴリ変数が，ともに 2 値変数のとき，そのクロス集計は 2 × 2 クロス集計表（2 × 2 分割表）と呼ばれ，その統計的性質が良く調べられている。

4.3 独立性のカイ二乗検定

独立性の検定としては，2 つのカテゴリ変数の間に**関連がない**と仮定した場合に推定される期待度数を求めて，それに観測度数が適合するかを検定するカイ二乗検定が最も有名である（だから，実はカイ二乗適合度検定と同じ原理である)。もちろん，ある種の関連が仮定できれば，その仮定の元に推定される期待度数と観測度数との適合を調べてもいいが，一般に，2 つのカテゴリ変数の間にどれくらいの関連がありそうかという仮定はできないことが多い。そこで，関連がない場合の期待度数を推定し，それが観測値に適合しなければ関連がないとはいえない，と推論するのである。

	A	$\bar{A}$
B	a 人	b 人
$\bar{B}$	c 人	d 人

2 つのカテゴリ変数 A と B が，それぞれ「あり」「なし」の 2 つのカテゴリ値しかとらないとき，これら 2 つのカテゴリ変数の組み合わせは「A も B もあり（$A \cap B$)」「A なし B あり（$\bar{A} \cap B$)」「A あり B なし（$A \cap \bar{B}$)」「A も B もなし（$\bar{A} \cap \bar{B}$)」の 4 通りしかない。それぞれの度数を数えあげた結果が，上記の表として得られたときに，母集団の確率構造が，

	A	$\bar{A}$
B	π_{11}	π_{12}
$\bar{B}$	π_{21}	π_{22}

であるとわかっていれば，$N = a + b + c + d$ として，期待される度数は，

	A	$\bar{A}$
B	$N\pi_{11}$	$N\pi_{12}$
$\bar{B}$	$N\pi_{21}$	$N\pi_{22}$

であるから，

$$\chi^2 = \frac{(a - N\pi_{11})^2}{N\pi_{11}} + \frac{(b - N\pi_{12})^2}{N\pi_{12}} + \frac{(c - N\pi_{21})^2}{N\pi_{21}} + \frac{(d - N\pi_{22})^2}{N\pi_{22}}$$

として, 自由度 3 のカイ二乗検定をすればよいことになる。しかし, 普通, π は未知である。そこで, $Pr(\bar{A}) = 1 - Pr(A)$ かつ，この 2 つのカテゴリ変数が独立ならば $Pr(A \cap B) = Pr(A)Pr(B)$ と考えれば良いことを使って，$Pr(A)$ と $Pr(B)$ を母数として推定する[*43]。$Pr(A)$ の点推定量は，B を無視して A の割合と考えれば $(a + c)/N$ であることは自明である。同様に，$Pr(B)$ の点推定量は，$(a + b)/N$ となる。したがって，$\pi_{11} = Pr(A \cap B) = Pr(A)Pr(B) = (a + c)(a + b)/(N^2)$ となる。

同様に考えれば，母集団の各セルの確率は下式で得られる。

$$\pi_{12} = (b + d)(a + b)/(N^2)$$

$$\pi_{21} = (a + c)(c + d)/(N^2)$$

$$\pi_{22} = (b + d)(c + d)/(N^2)$$

[*43] $Pr(X)$ はカテゴリ X の出現確率を示す記号である。また，2 つの母数をデータから推定するので，得られるカイ二乗統計量が従う分布の自由度は 3 より 2 少なくなり，自由度 1 のカイ二乗分布となる。

これらの値を使えば，

$$\chi^2 = \frac{\{a-(a+c)(a+b)/N\}^2}{\{(a+c)(a+b)/N\}} + \frac{\{b-(b+d)(a+b)/N\}^2}{\{(b+d)(a+b)/N\}} + \frac{\{c-(a+c)(c+d)/N\}^2}{\{(a+c)(c+d)/N\}} + \frac{\{d-(b+d)(c+d)/N\}^2}{\{(b+d)(c+d)/N\}}$$

$$= \frac{(ad-bc)^2\{(b+d)(c+d)+(a+c)(c+d)+(b+d)(a+b)+(a+c)(a+b)\}}{(a+c)(b+d)(a+b)(c+d)N}$$

分子の中括弧の中は N^2 なので，結局，

$$\chi^2 = \frac{N(ad-bc)^2}{(a+c)(b+d)(a+b)(c+d)}$$

ただし通常は，イェーツの連続性の補正を行う。カイ二乗分布は連続分布なので，各度数に 0.5 を足したり引いたりしてやると，より近似が良くなるという発想である。この場合，

$${\chi_c}^2 = \frac{N(|ad-bc|-N/2)^2}{(a+c)(b+d)(a+b)(c+d)}$$

が自由度 1 のカイ二乗分布に従うと考えて検定する。ただし，$|ad-bc|$ が $N/2$ より小さいときは補正の意味がないので，$\chi^2 = 0$ とする。

実際の検定は R を使えば，クロス集計表が既に得られているとき，例えば a=12, b=8, c=9, d=10 などとわかっていれば，`x <- matrix(c(12,9,8,10),nr=2)` として表を与え（あまり 2 × 2 の場合は意味がないが，必要なら `mosaicplot(x)` として図示してから），`chisq.test(x)` とするだけでいい[*44]。各度数が未知で，各個人についてカテゴリ変数 A と B の生の値が名義尺度として得られているときは，`table(A,B)` とすればクロス集計表が作成できる。そこで，`chisq.test(table(A,B))` とすれば，独立性のカイ二乗検定ができる[*45]。

R では，`chisq.test()` 関数の中で，`simulate.p.value=TRUE` というオプションを使えば，シミュレーションによってそのカイ二乗値より大きなカイ二乗値が得られる確率を計算させることもできる。この方がたんなるカイ二乗検定よりも正確な p 値が得られるが，遅いコンピュータだと計算時間がかかる欠点がある。

例題

肺ガンの患者 100 人に対して，1 人ずつ性・年齢が同じ健康な人を対照として 100 人選び（この操作をペアマッチサンプリングという），それぞれについて過去の喫煙の有無を尋ねた結果，患者群では過去に喫煙を経験した人が 80 人，対照群では過去に喫煙を経験した人が 55 人だった。肺ガンと喫煙は無関係といえるか？　独立性のカイ二乗検定をせよ。

帰無仮説は，肺ガンと喫煙が無関係（独立）ということである。クロス集計表を作ってみると，

	肺ガン患者群	健康な対照群	合計
過去の喫煙経験あり	80	55	135
過去の喫煙経験なし	20	45	65
合計	100	100	200

となる。この表は，`matrix(c(80,20,55,45),nr=2)` で得られる。肺ガンと喫煙が無関係だという帰無仮説の下で期待される各カテゴリの人数は，

	肺ガンあり	肺ガンなし
喫煙あり	$135 \times 100/200 = 67.5$	$135 \times 100/200 = 67.5$
喫煙なし	$65 \times 100/200 = 32.5$	$65 \times 100/200 = 32.5$

となる。従って，連続性の補正を行なったカイ二乗統計量は，

$$\chi_c^2 = (80-68)^2/67.5 + (55-67)^2/67.5 + (20-32)^2/32.5 + (45-33)^2/32.5 = 13.128...$$

[*44] 連続性の補正を行わないときは `chisq.test(x,correct=F)` とするが，通常その必要はない。

[*45] 実は `chisq.test(A,B)` でもカイ二乗検定は可能だが，表を与える形にしておく方がよい。

となり，自由度１のカイ二乗分布で検定すると 1-pchisq(13.128,1) より有意確率は 0.00029... となり，有意水準5%で帰無仮説は棄却される。つまり，肺ガンの有無と過去の喫煙の有無は独立とはいえない。R では

```
X <- matrix(c(80,20,55,45),nr=2)
chisq.test(X)
```

と入力すれば，下枠内の結果が得られる。

```
        Pearson's Chi-squared test with Yates' continuity correction

data:  X
X-squared = 13.1282, df = 1, p-value = 0.0002909
```

この検定は，肺ガン群と対照群の間で，過去の喫煙者の割合に差があるかどうかを検定することと数学的に同値である。下枠内を実行すれば，まったく同じ検定結果が得られる。

```
smoker <- c(80,55)
pop <- c(100,100)
prop.test(smoker,pop)
```

ただし，カイ二乗検定はあくまで正規近似なので，ある程度各カテゴリの組み合わせごとの期待頻度が大きくないと近似が悪くなってしまう。一般に，期待度数が５以下の組み合わせが検討すべき組み合わせ数の 20%以上あるときは（例えば２×２クロス集計表なら，１つでも期待度数５以下の組み合わせがあれば）カイ二乗検定は適当でないとされる。

4.4 フィッシャーの直接確率（正確な確率）

期待度数が低い組み合わせがあるときには，カテゴリを併合して変数を作り直す方法もあるが，もっといい方法が考案されている。

ここで調べたいのは組み合わせの数なので，周辺度数を固定して（各々の変数については母比率が決まっていると仮定して）すべての組み合わせを考え，それらが起こる確率（超幾何分布に従う）を１つずつ計算し，得られている集計表が得られる確率よりも低い確率になるような表が得られる確率をすべて足し合わせてしまえば，２つのカテゴリ変数の間に関連がないという帰無仮説の下でそういう表が偶然得られる確率がどれほど低いのかを，直接計算することができる。こうして計算される確率を，フィッシャーの直接確率，あるいは，フィッシャーの正確な確率（検定）という。これなら，近似ではないので，期待度数が低い組み合わせがあっても問題ない。

もう少し丁寧に言うと，サイズ N の有限母集団があって，そのうち変数 A の値が 1 である個体数が m_1，1 でない個体数が m_2 あるときに，変数 B の値が 1 である個体数が n_1 個（1 でない個体数が $n_2 = N - n_1$ 個）あるという状況を考え，この n_1 個のうち変数 A の値が 1 である個体数がちょうど a である確率を求めることになる。これは，m_1 個から a 個を取り出す組み合わせの数と m_2 個から $n_1 - a$ 個を取り出す組み合わせの数を掛けて，N 個から n_1 個を取り出す組み合わせの数で割った値になる。これと同じ周辺度数をもつ２×２分割表のうち，確率がこれと同じかこれよりも小さい表の確率をすべて足し合わせたものが，「変数 A と変数 B が独立」という帰無仮説が成り立つ確率になる[*46]。

フィッシャーの正確な確率は，R では，fisher.test() 関数で実行できる。この方がカイ二乗検定よりも正確であ

[*46] 有限母集団からの非復元抽出になるので，平均 $E(a)$ と分散 $V(a)$ は，

$$E(a) = n_1 m_1 / N$$

$$V(a) = \{(N - n_1)/(N - 1)\} n_1 (m_1/N)(m_2/N) = (m_1 m_2 n_1 n_2)/\{N^2(N - 1)\}$$

となる。実際には組み合わせ計算が多いので，手計算で実行することはまずありえず，統計ソフトにやらせることになる。また，個々の２×２分割表の確率は離散値をとるので，同じ確率の表がありうる場合に，それを足し算に含めるのかどうかは難しい点である。これを乱数によって決める「ランダム検定」という手法もあるが，あまり一般的ではない。

る。クロス集計表を使って２つのカテゴリ変数間の独立性の検定をするときは，コンピュータが使えるならば，サンプルサイズがよほど大きくない限り，常にカイ二乗検定ではなく，フィッシャーの正確な確率を求めるべきである。

4.5 回帰

実験によって，あるサンプルの濃度を求めるやり方の１つに，検量線の利用がある。検量線とは，予め濃度がわかっている標準物質を測ったときの吸光度のばらつきが，その濃度によってほぼ完全に（通常 98%以上）説明されるときに（そういう場合は，散布図を描くと，点々がだいたい直線上に乗るように見える），その関係を利用して，サンプルを測ったときの吸光度からサンプルの濃度を逆算するための回帰直線である（曲線の場合もあるが，通常は何らかの変換をほどこし，線形回帰にして利用する）。検量線の計算には，(A) 試薬ブランクでゼロ点調整をした場合の原点を通る回帰直線を用いる場合と，(B) 純水でゼロ点調整をした場合の切片のある回帰直線を用いる場合がある。いずれも，量がわかっているもの（この場合は濃度）を x，誤差を含んでいる可能性がある測定値（この場合は吸光度）を y として $y = bx + a$ という形の回帰式の係数 a と b を最小二乗法で推定し，サンプルを測定した値 y から $x = (y - a)/b$ によってサンプルの濃度 x を求める。回帰直線の適合度の目安としては，学生実習でも相関係数の 2 乗が 0.98 以上あることが望ましい。また，データ点の最小，最大より外で直線関係が成立する保証はない。従って，サンプル測定値が標準物質の測定値の最小より低いか，最大より高いときは，限界を超えていることになってしまう[*47]。

測定点 (x_1, y_1), (x_2, y_2), ..., (x_n, y_n) が得られたときに，検量線 $y = bx + a$ を推定するには，図に示した線分の二乗和が最小になるように a と b を設定すればよい，というのが最小二乗法の考え方である。つまり，

$$f(a,b) = \sum_{i=1}^{n} \{y_i - (bx_i + a)\}^2$$

$$= b^2 \sum_{i=1}^{n} x_i^2 - 2b \sum_{i=1}^{n} x_i y_i + 2ab \sum_{i=1}^{n} x_i - 2a \sum_{i=1}^{n} y_i + na^2 + \sum_{i=1}^{n} y_i^2$$

が最小になるような a と b を推定すればよい。通常，a と b で偏微分した値がそれぞれ 0 となることを利用して計算すると簡単である。つまり，

$$\frac{\partial f(a,b)}{\partial a} = 2na + 2(b \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} y_i) = 0$$

$$i.e. \quad na = \sum_{i=1}^{n} y_i - b \sum_{i=1}^{n} x_i$$

$$i.e. \quad a = (y \text{ の平均}) - (x \text{ の平均}) * b$$

$$\frac{\partial f(a,b)}{\partial b} = 2b \sum_{i=1}^{n} x_i^2 + 2(a \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i y_i) = 0$$

$$i.e. \quad b \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i - a \sum_{i=1}^{n} x_i$$

を連立方程式として a と b について解けばよい。これを解くと，

$$b = \frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

が得られる[*48]。b の値を上の式に代入すれば a も得られる。検量線に限らず，一般の回帰直線でも，計算方法は原則として同じである。名称の説明をしておくと，一般に，$y = bx + a$ という回帰直線について，b を回帰係数 (regression coefficient)，a を切片 (intercept) と呼ぶ。

データから得た回帰直線は，$pV = nRT$ のような物理法則と違って，完璧にデータに乗ることはない。そこで，回帰直線の当てはまりのよさを評価する必要が出てくる。a と b が決まったとして，$z_i = a + bx_i$ とおいたとき，$e_i = y_i - z_i$

[*47] 余談だが，このような場合はサンプルを希釈するか濃縮して測定するのが普通である。

[*48] 分母分子を n^2 で割れば，b は $x_i y_i$ の平均から x_i の平均と y_i の平均の積を引いて，x_i の二乗の平均から x_i の平均の二乗を引いた値で割った形になる。

を残差 (residual) と呼ぶ。残差は，y_i のばらつきのうち，回帰直線では説明できなかった残りに該当する。つまり，残差が大きいほど，回帰直線の当てはまりは悪いと考えられる。残差にはプラスもマイナスもあるので二乗和をとり，

$$Q = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - z_i)^2$$

$$= \sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2/n - \frac{(n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i)^2}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}/n$$

として，この Q が回帰直線の当てはまりの悪さを示す尺度となる。Q を「残差平方和」と呼び，それを n で割った Q/n を残差分散という。この残差分散（$var(e)$ と書くことにする）と Y の分散 $var(Y)$ とピアソンの相関係数 r の間には，$var(e) = var(Y)(1 - r^2)$ という関係が常に成り立つので，$r^2 = 1 - var(e)/var(Y)$ となる。このことから r^2 が 1 に近いほど回帰直線の当てはまりがよいことになる。その意味で，r^2 を「決定係数」と呼ぶ。また，決定係数は，Y のばらつきがどの程度 X のばらつきによって説明されるかを意味するので，X の「寄与率」と呼ぶこともある。

回帰直線は最小二乗法でもっとも残差平方和が小さくなるように選ぶわけだが，データの配置によっては，何通りもの回帰直線の残差平方和が大差ないという状況がありうる。例えば，独立変数と従属変数（として選んだ変数）が実はまったく無関係であった場合は，データの重心を通るどのような傾きの線を引いても残差平方和はほとんど同じになってしまう。その意味で，回帰直線のパラメータ（回帰係数 b と切片 a）の推定値の安定性を評価することが大事である。そのためには，t 値というものが使われている。いま，Y と X の関係が $Y = a_0 + b_0 X + e$ というモデルで表されるとして，誤差項 e が平均 0，分散 σ^2 の正規分布に従うものとすれば，回帰係数の推定値 a も，平均 a_0，分散 $\sigma^2/n)(1 + M^2/V)$（ただし M と V は x の平均と分散）の正規分布に従い，残差平方和 Q を誤差分散 σ^2 で割った Q/σ^2 が自由度 $(n-2)$ のカイ二乗分布に従うことから，

$$t_0(a_0) = \frac{\sqrt{n(n-2)}(a - a_0)}{\sqrt{(1 + M^2/V)Q}}$$

が自由度 $(n-2)$ の t 分布に従うことになる。しかしこの値は a_0 がわからないと計算できない。a_0 が 0 に近ければこの式で $a_0 = 0$ と置いた値（つまり $t_0(0)$。これを切片に関する t 値と呼ぶ）を観測データから計算した値が $t_0(a_0)$ とほぼ一致し，自由度 $(n-2)$ の t 分布に従うはずなので，その絶対値は 95%の確率で t 分布の 97.5%点（サンプルサイズが大きければ約 2 である）よりも小さくなる。つまり，データから計算された t 値がそれより大きければ，切片は 0 でない可能性が高いことになる。t 分布の分布関数を使えば，「切片が 0 である」という帰無仮説に対する有意確率が計算できることになる。回帰係数についても同様に，

$$t_0(b) = \frac{\sqrt{n(n-2)V}b}{\sqrt{Q}}$$

が自由度 $(n-2)$ の t 分布に従うことを利用して，「回帰係数が 0」であるという帰無仮説に対する有意確率が計算できる。

以上の説明からすると，身長と体重のように，どちらも誤差を含んでいる可能性がある測定値である場合には，一方を独立変数，他方を従属変数とすることは，本当は妥当でないかもしれない。一般には，身長によって体重が決まってくるというように方向性が仮定できれば，身長を独立変数と見なしてもよいことになっているが，回帰分析をしてしまうと，独立変数に測定誤差がある可能性が排除されてしまうことには注意しておくべきである。つまり，測定誤差が大きい可能性がある変数を独立変数とした回帰分析は，できれば避けたほうが良い（が，そうもいかないのが実情である）。また，最小二乗推定の説明から自明なように，独立変数と従属変数を入れ替えた回帰直線は一致しない。従って，どちらを従属変数とみなし，どちらを独立変数とみなすか，ということは，因果関係の方向性に基づいて（先行研究や biological なメカニズムを参照して）きちんと決めるべきである。

回帰を使って予測をするとき，外挿には注意が必要である。前述の通り，検量線は，原則として外挿してはいけない。実際に測った濃度より濃かったり薄かったりするサンプルに対して，同じ関係が成り立つという保証はどこにもないからである（吸光度を y とする場合は，濃度が高くなると分子の重なりが増えるので飽和 (saturate) してしまい，吸光度の相対的な上がり方が小さくなっていき，直線から外れていく）。しかし，外挿による予測は，実際にはかなり行われている。例えば世界人口の将来予測とか，河川工学における基本高水計算式とか，感染症の発症数の将来予測は，回帰

の外挿による場合が多い。このやり方が妥当性をもつためには，その回帰関係が（1）かなり説明力が大きく，（2）因果関係がある程度認められ，（3）それぞれの変数の分布が端の切れた分布でない（truncated distribution でない）という条件を満たす必要がある。そうでない場合は，その予測結果が正しい保証はどこにもない。

R では，今回説明したような線形回帰を行うための関数は `lm()` である。例えば，独立変数を `X`，従属変数を `Y` として，`lm(Y~X)` のように用いれば，回帰直線の推定値が得られる。決定係数や回帰係数と切片の検定結果は，`summary(lm(Y~X))` とすれば出力される。Rcmdr の場合は，データを読み込んだ後，メニューバーの Statistics（統計処理）の Fit Models（モデルの当てはめ）の Linear Regression（線型回帰）を選び，Response Variable（応答変数＝従属変数と同義）として `Y` を，Explanatory Variable(s)（説明変数 [群] ＝独立変数 [群] と同義）として `X` を選んで OK ボタンをクリックすればよい。

4.6 共分散分析

複数のグループがあって，どのグループに属するサンプルについても，同じ独立変数と従属変数が調べられているときに，独立変数と従属変数の関係がグループによって異なるかどうか調べたい場合がある。共分散分析は，このような場合に用いることができる分析手法である。

典型的には，共分散分析は，$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2 + \varepsilon$ というモデルになる。2 値変数 X_1 によって示される 2 群間で，量的変数 Y の平均値に差があるかどうかを比べるのだが，Y が量的変数 X_2 と相関がある場合に（このとき X_2 を共変量と呼ぶ），X_2 と Y の回帰直線の傾き (slope) が X_1 の示す 2 群間で差がないときに，X_2 による影響を調整した Y の修正平均（adjusted mean; 調整平均ともいう）に，X_1 の 2 群間で差があるかどうかを検定する。

R では，X_1 を示す変数名を `C`（注：`C` は `factor` である必要がある），X_2 を示す変数名を `X` とし，Y を示す変数名を `Y` とすると，

```
summary(glm(Y~C+X))
```

とすれば，`X` の影響を調整した上で，`C` 間で `Y` の修正平均（調整平均）が等しいという帰無仮説についての検定結果が得られる（結果出力で `C2` と表示される行の右端に出ているのがその有意確率である）。ただし，この検定をする前に，2 本の回帰直線がともに有意にデータに適合していて，かつ 2 本の回帰直線の間で傾き (slope) が等しいかどうかを検定して，傾きが等しいことを確かめておかないと，修正平均の比較には意味がない。そこで，まず例えば，`summary(lm(Y[C==1]~X[C==1]); summary(lm(Y[C==2]~X[C==2])` として 2 つの回帰直線それぞれの適合を確かめ，`summary(glm(Y~C+X+C*X))` として傾きが等しいかどうかを確かめなければならない。傾きが有意に違っていることは，`C` と `X` の交互作用項が有意に `Y` に効いていることと同値なので，`Coefficients` の `C2:X` と書かれている行の右端を見れば，「傾きが等しい」を帰無仮説とした場合の有意確率が得られる。そもそも回帰直線の適合が悪ければその独立変数は共変量として考慮する必要がないし，傾きが違っていれば群分け変数と独立変数の交互作用が従属変数に関して有意に影響しているということなので，2 群を層別して別々に解釈する方が良い。

参考までに数式でも説明しておく。いま，`C` で群分けされる 2 つの母集団における，(X, Y) の間の母回帰直線を，$y = \alpha_1 + \beta_1 x$，$y = \alpha_2 + \beta_2 x$ とすれば，次の 2 つの仮説が考えられる。まず傾きに差があるかどうか？　を考える。つまり，$H_0 : \beta_1 = \beta_2$，$H_1 : \beta_1 \neq \beta_2$ である。次に，もし傾きが等しかったら，y 切片も等しいかどうかを考える。つまり，$\beta_1 = \beta_2$ のもとで，$H_0' : \alpha_1 = \alpha_2$，$H_1' : \alpha_1 \neq \alpha_2$ を検定する。各群について，X と Y の平均と変動と共変動を出しておけば[*49]，仮説 H_1 のもとでの残差平方和

$$d_1 = SS_{Y1} - (SS_{XY1})^2/SS_{X1} + SS_{Y2} - (SS_{XY2})^2/SS_{X2}$$

と仮説 H_0 のもとでの残差平方和

$$d_2 = SS_{Y1} + SS_{Y2} - (SS_{XY1} + SS_{XY2})^2/(SS_{X1} + SS_{X2})$$

を計算して $F = (d_2 - d_1)/(d_1/(N-4))$ が H_0 のもとで第 1 自由度 1，第 2 自由度 $N-4$ の F 分布に従うことを使って傾きが等しいかどうかの検定ができる。H_0 が棄却されたときは，$\beta_1 = SS_{XY1}/SS_{X1}$，$\beta_2 = SS_{XY2}/SS_{X2}$ として

[*49] サンプルサイズ $N1$ の第 1 群に属する x_i, y_i について，$E_{X1} = \sum x_i/N1$，$SS_{X1} = \sum (x_i - E_{X1})^2$，$E_{Y1} = \sum y_i/N1$，$SS_{Y1} = \sum (y_i - E_{Y1})^2$，$E_{XY1} = \sum x_i y_i/N1$，$SS_{XY1} = \sum (x_i y_i - E_{XY1})^2$。第 2 群も同様。

別々に傾きを推定し，y 切片 α もそれぞれの式に各群の平均値を入れて計算できる。H_0 が採択されたときは，共通の傾き β を，$\beta = (SS_{XY1} + SS_{XY2})/(SS_{X1} + SS_{X2})$ として推定する。この場合はさらに y 切片が等しいという帰無仮説 H_0' のもとで全部のデータを使った残差平方和 $d_3 = SS_Y - (SS_{XY})^2/SS_X$ を計算して，$F = (d_3 - d_2)/(d_2/(N-3))$ が第 1 自由度 1，第 2 自由度 $N-3$ の F 分布に従うことを使って検定できる。H_0' が棄却された場合は各群の平均を共通の傾きに代入すれば各群の切片が求められるし，採択されたら，要するに 2 群間に差がないということになるので，2 群を一緒にして普通の単回帰分析をしていいことになる。

5 生存時間解析

5.1 生存時間解析とは

実験においては，化学物質などへの 1 回の曝露の影響を時間を追ってみていくことが良く行われる。時間ごとに何らかの量の変化を追うほかに，エンドポイントを死亡とした場合，死ぬまでの時間を分析することで毒性の強さを評価することができる。このような期間データを扱う方法としては，一般に生存時間解析 (Survival Analysis または Event History Analysis) と呼ばれるものがある。なかでもよく知られているものが Kaplan-Meier の積・極限推定量である（現在では一般に，カプラン・マイヤ推定量と呼ばれている）。カプラン・マイヤ推定量は，イベントが起こった各時点での，イベントが起こる可能性がある人口（リスク集合）あたりのイベント発生数を 1 から引いたものを掛け合わせて得られる，ノンパラメトリックな最尤推定量である。カプラン・マイヤ推定量は非常によく使われるので，具体例で説明しておく（後述）。複数の期間データ列があったときに，それらの差を検定したい場合は，ログランク検定や一般化ウィルコクソン検定が使われる。細かいことをいえば，ログランク検定でも Mantel-Haenzel 流のログランク検定と Peto and Peto 流のログランク検定があったり，一般化ウィルコクソン検定でも Gehan-Breslow 流と Peto-Prentice 流があったりして，非常に面倒な話になってくるので，ここでは Mantel-Haenzel 流のログランク検定のみ[*50]説明する（後述）。

それらのノンパラメトリックな方法とは別に，イベントが起こるまでの時間が何らかのパラメトリックな分布に当てはまるかどうかを調べる方法もある。当てはめる分布としては指数分布やワイブル分布がある。イベントが起こるまでの期間に何らかの別の要因が与える効果を調べたいときはコックス回帰（それらが基準となる個体のハザードに対して $\exp(\sum \beta_i z_i)$ という比例定数の形で掛かるとする比例ハザード性を仮定する方法）と，パラメトリックなモデルに対数線形モデルの独立変数項として入れてしまう加速モデルがある。R では生存時間解析をするための関数は `survival` パッケージで提供されており，`library(survival)` とすれば使えるようになる。カプラン・マイヤ法は `survfit()` 関数，コックス回帰は `coxph()` 関数，加速モデルは `survreg()` 関数で実行できる。なお，生存時間解析について，より詳しく知りたい方は，大橋，浜田 (1995) などを参照されたい。

5.2 カプラン・マイヤ法

では，簡単な例を使って，カプラン・マイヤ推定量を説明しよう。表 1 は，ソロモン諸島のある村で，既婚女性全員に，自分の誕生日，第 1 子誕生日，第 2 子誕生日，……，末子誕生日（まだ出産を完了していない年齢の女性も含めて，ともかくそれまでに産んだ子どもの誕生日を全部）聞き取った結果である。間隔データを使わなければ，このデータから出生力について何かいうためには，出産を完了した女性についての平均出産数（平均完結パリティという）くらいしか指標がないが，間隔データを使えば，時間当たりの出生力を考えることができるので，出産を完了していない女性のデータも使うことができる。

この種のデータには，以下の利点と欠点がある。

- 母親に対して，全ての子どもの出生年月日を聞き取ることは，統計がしっかりしていない社会でも比較的信頼性の高い方法である。
- 人口規模が小さくても使える上，過去の推計もできるという利点がある。
- 古くなるほど誤差が大きくなるバイアスや，他に影響を受ける要因が多いのは欠点。

[*50] Peto and Peto 流のものは，コンピュータが使えない場合に手計算するには有用だが，それ以上の意味はない。

MO_ID	MO_BD	C1_BD	C2_BD	C3_BD	C4_BD	C5_BD	C6_BD	C7_BD	C8_BD	C9_BD	C10_BD	C11_BD
20102	390000	0	640600	680000	711014	760000						
60202	250000	480415	560921	630000								
50102	400000	550000	590000	630000	660810	681011	710319	741018	760611	0		
30602	450000	580000	601004	630000	650000	670000	670000	720000	740000	750000	780714	
10502	400000	600716	630000	650807	670000	690609						
10102	400000	651103	681225	0	720200	0	790517	0	820000	840503	860527	890302
30102	490000	680000	700000	720000	750000	770000	820927					
10202	490000	680000	720826	760000	830000							
40302	580000	700000	780606	820906	901012	910606						
40102	570000	710114	730000	750000	770000	810621	840101	870802	920813			
20502	580000	720906	740704	761106	800407	811126	860516	910406				
50302	520000	730000	780000	800000	830000	870000	0					
10402	441101	730324	760723	770801	880119							
60302	460000	740000	770000	790000	800000	820000						
70202	550000	740000	780000	800000	840000	870000	890000	920000	941100			
70302	600000	750000	780000	800000	820000	850500	860000	880000	920000	940000		
20302	610000	760709	771020	790309	811002	850415	890803					
30702	600000	810500	820000	830000	840000	850000	900924	930430	950604			
30502	301205	820921	840803	881228								
60402	330000	830212	850216	900916	950921							
10802	650521	840623	861009	890727	920329	940416						
50402	670000	861114	880430	900130	910000	930325	950108					
20602	651114	870904	881111	900519	911104							
60102	570000	880000	950905									
10902	670000	900000	910000	930319								
30202	710000	900408	920210	940305								
40202	640000	901007	931109									
60204	680000	910000	920000									
50202	640000	911001	921020									
20202	711014	920801	931127									
10302	720826	920823	940308									
11002	700917	930303	950513									
10702	670304	930701										
80102	720229	940125										
11102	670809	940406										
30302	720000	940611										
50303	730000	950300										
10602	740700	950317										
20504	740704	950905										
60303	740000	951024										
70102	0	420000	450000	470000	520000	531225	550000	630000	670000			

図 1: ソロモン諸島のある村の女性全員の再生産史

- 結婚から第１子誕生までの期間や，第１子と第２子の出生間隔がよく使われるが，上にあげたソロモン諸島の社会では，結婚記念日はあまり正確に記憶されていなかったために，第１子と第２子の出産間隔を使うことにした。第１子と第２子の出産間隔には，第２子の在胎期間が含まれるために，その期間のハザードは原理的にゼロであることに注意する必要がある[*51]。

まず，カプラン・マイヤ推定量についての一般論を示す。イベントが起こる可能性がある状態になってから，イベントが起こった時点を $t_1, t_2, ...$ とし，t_1 時点でのイベント発生数を d_1，t_2 時点でのイベント発生数を d_2，以下同様であるとする。また，時点 $t_1, t_2, ...$ の直前でのリスク集合の大きさを $n_1, n_2, ...$ で示す。リスク集合の大きさとは，その直前でまだイベントが起きていない（この例では第１子出産後で第２子出産前の）個体数である。観察途中で死亡や転居などによって打ち切りが生じるために，リスク集合の大きさはイベント発生によってだけではなく，打ち切りによっても減少する。従って n_i は，時点 t_i より前にイベント発生または打ち切りを起こした個体数を n_1 から除いた残りの数となる。なお，イベント発生と打ち切りが同時点で起きている場合は，打ち切りをイベント発生直後に起きたと見なして処理するのが慣例である。このとき，カプラン・マイヤ推定量 $\hat{S}(t)$ は，

$$\hat{S}(t) = (1 - d_1/n_1)(1 - d_2/n_2)... = \prod_{i<t}(1 - d_i/n_i)$$

として得られる。その標準誤差はグリーンウッドの公式により，説明は省略するが，

$$var(\hat{S}) = \hat{S}^2 \times \sum_{i<t} \frac{d_i}{n_i(n_i - d_i)}$$

で得られる。なお，カプラン・マイヤ推定量を計算するときは，階段状のプロットを同時に行うのが普通である。

R では，`library(survival)` としてパッケージを呼び出し，`dat <- Surv(`**生存時間，打ち切りフラグ**`)` 関数で生存時間データを作り（打ち切りフラグは 1 でイベント発生，0 が打ち切り。ただし区間打ち切りの場合は 2 とか 3 も使う），`res <- survfit(dat)` でカプラン・マイヤ法によるメディアン生存時間が得られ，`plot(res)` とすれば階段関数が描かれる。イベント発生時点ごとの値を見るには，`summary(res)` とすればよい[*52]。

[*51] 例えば，在胎期間の推定値として 9 ヶ月を引いた値をデータにしたり，または在胎期間を切片として含んだハザード関数を推定することも考慮するべきである。

[*52] 参考までに書いておくと，生データがイベント発生の日付を示している場合，間隔を計算するには `difftime()` 関数や `ISOdate()` 関数を使うと便利である。例えば，

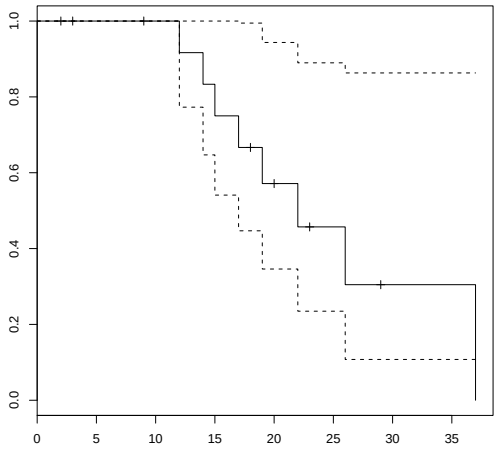


図 2: ソロモン諸島女性の第 1 出産間隔についてのカプラン・マイヤプロット

例えば，区間打ち切り（イベント発生までの時間がある幅をもってしかわからないデータ）を無視して，上で示したソロモン諸島のデータのうち，第 1 子出生が 1986 年以降のものの出産間隔データを R で分析すると[*53]，右側打ち切りを考慮した出産間隔のメディアンが 22 ヶ月であることがわかる（プロットを図 2 に示す）。

5.3 ログランク検定

次に，ログランク検定を簡単な例で説明する。

8 匹のラットを 4 匹ずつ 2 群に分け，第 1 群には毒物 A を投与し，第 2 群には毒物 B を投与して，生存期間を追跡したときに，第 1 群のラットが 4,6,8,9 日目に死亡し，第 2 群のラットが 5,7,12,14 日目に死亡したとする。この場合，観察期間内にすべてのラットが死亡し，正確な生存時間がわかっているので，観察打ち切りがないデータとなっていて計算しやすい。

ログランク検定の思想は，大雑把に言えば，死亡イベントが起こったすべての時点で，群と生存／死亡個体数の 2 × 2 クロス集計表を作り，それをコクラン＝マンテル＝ヘンツェル流のやり方で併合するということである。上記の例では，死亡イベントが起こった時点 1〜8 において各群の期待死亡数を計算し，各群の実際の死亡数との差をとって，それに時点の重みを掛けたものを，各時点における各群のスコアとして，群ごとのスコアの合計を求める。2 群しかない

```
dob
1964-8-21
```

という形のテキストファイル L12-1.dat で，日付が与えられているとしよう。これを x <- read.delim("L12-1.dat") として読み込み，2003 年 6 月 11 日までの間隔を計算したければ，difftime(ISOdate(2003,6,11),x$dob) とすれば，その間の経過日数が DateTimeClasses のオブジェクトとして得られる。日数から年に変換したければ，例えば as.integer(difftime(ISOdate(2003,6,11),x$dob))/365.24 とすればいいし，さらに 12 を掛ければ月単位になる。

[*53] プログラムは下記の通り。

```
library(survival)
time <- c(17,14,22,37,12,15,19,26,29,23,20,18,9,9,3,2)
event <- c(1,1,1,1,1,1,1,1,0,0,0,0,0,0,0,0)
dat <- Surv(time,event)
res <- survfit(dat)
print(res)
summary(res)
plot(res)
```

ので，各時点において群１と群２のスコアの絶対値は同じで符号が反対になる。２群の生存時間に差がないという帰無仮説を検定するためには，群１のスコアの２乗を分散で割った値をカイ二乗統計量とし，帰無仮説の下でこれが自由度１のカイ二乗分布に従うことを使って検定する。なお，重みについては，ログランク検定ではすべて１である。一般化ウィルコクソン検定では，重みを，２群を合わせたリスク集合の大きさとする（そうした場合，もし打ち切りがなければ，検定結果は，ウィルコクソンの順位和検定の結果と一致する)。つまり，ログランク検定でも一般化ウィルコクソン検定でも，実は期間の情報はまったく使われず，死亡順位の情報だけが使われているのである。

記号で書けば次の通りである。第 i 時点の第 j 群の期待死亡数 e_{ij} は，時点 i における死亡数の合計を d_i，時点 i における j 群のリスク集合の大きさを n_{ij}，時点 i における全体のリスク集合の大きさを n_i とすると，

$$e_{ij} = d_i \cdot n_{ij}/n_i$$

と表される。上の例では，$e_{11} = 1 \cdot n_{11}/n_1 = 4/8 = 0.5$ となる。時点 i における第 j 群の死亡数を d_{ij}，時点の重みを w_i と表せば，時点 i における群 j のスコア u_{ij} は，

$$u_{ij} = w_i(d_{ij} - e_{ij})$$

となり，ログランク検定の場合（以下，重みは省略してログランク検定の場合のみ示す）の群 1 の合計スコアは

$$u_1 = \sum_i d_{i1} - e_{i1}$$

となる。上の例では，

$$u_1 = (1 - 4/8) + (0 - 3/7) + (1 - 3/6) + (0 - 2/5) + (1 - 2/4) + (1 - 1/3) + (0 - 0/2) + (0 - 0/1)$$

である。これを計算すると約 1.338 となる。分散は，分散共分散行列の対角成分を考えればいいので，

$$V = V_{jj} = \sum_i \frac{(n_i - n_{ij})n_{ij}d_i(n_i - d_i)}{{n_i}^2(n_i - 1)}$$

となる。この例の数値を当てはめると，

$$V = \frac{(8-4)\times 4}{8^2} + \frac{(7-3)\times 3}{7^2} + \frac{(6-3)\times 3}{6^2} + \frac{(5-2)\times 2}{5^2} + \frac{(4-2)\times 2}{4^2} + \frac{(3-1)\times 1}{3^2}$$

となり，計算すると，約 1.568 となる。したがって，$\chi^2 = 1.338^2/1.568 = 1.14$ となり，この値は自由度１のカイ二乗分布の 95%点である 3.84 よりずっと小さいので，有意水準 5%で帰無仮説は棄却されない。つまりこれだけのデータでは，差があるとはいえないことになる（もちろん，サンプルサイズを大きくすれば違う結果になる可能性もある)。

R では，`Surv(time,event)` と `group`（注：ここで time は生存時間，event は 1 がイベント観察，0 が観察打ち切りを示すフラグ，group がグループを示す）を，`survdiff()` 関数に与えることによってログランク検定が実行できる。打ち切りレコードがない場合は，event は省略できる。なお，生存時間解析の関数は `survival` パッケージに入っているので，まず `library(survival)` とすることは必須である。

上で説明した例では，

```
library(survival)
time <- c(4,6,8,9,5,7,12,14)
event <- c(1,1,1,1,1,1,1,1)
group <- c(1,1,1,1,2,2,2,2)
dat <- Surv(time,event)
# カプラン・マイヤ法で各グループの生存関数を計算
res <- survfit(dat~group)
print(res)
summary(res)
# ログランク検定
res2 <- survdiff(dat~group)
print(res2)
```

とすると解析結果が得られる。一番下を見ると，$\chi^2 = 1.2$，自由度 1，$p = 0.268$ となっているので，有意水準 5%で，2 群には差がないことがわかる。

6 文献

- 大橋靖雄，浜田知久馬 (1995) 生存時間解析: SAS による生物統計. 東京大学出版会.
- 古川俊之 [監修]，丹後俊郎 [著] (1983) 医学への統計学. 朝倉書店.
- 中澤　港 (2003) R による統計解析の基礎. ピアソン・エデュケーション.
- 永田　靖 (2003) サンプルサイズの決め方. 朝倉書店.

7 付録

なぜ R で演習するのか？

R の最大の利点は，オープンソースなフリーソフトで，かつ拡張性が高い点である（世界中の研究者が GIS を含む空間統計解析やゲノム解析などに至るまでさまざまな追加ライブラリを公開しているし，自分で新しい拡張関数を付け加えることもできる）。しかも，SPSS などに比べるとプログラムコードが洗練されているため，ソフトウェアのサイズが小さく，動作が軽快である。完全に無料で利用できるため，卒業後も覚えた技術が無駄にならない（SAS や SPSS のような高価なソフトウェアで実習を受けた場合，卒業しても使えるような環境に就職できる人はほんの一握りなので，多くの学生はその技術を活用することができなくなってしまう）。また，国際協力などの場面でもライセンスを気にすることなく共用することができる。英文のみならず，仏文，西文などのマニュアルも公開されている。Windows だけでなく，Macintosh でも Linux でも FreeBSD でも動作するので，さまざまな環境で同じ統計解析を行なうことができる。同種のソフトウェアとして CDC が提供している EPIINFO があるが，利用できる統計解析手法の種類は R の方がずっと多いし，EPIINFO は Windows 専用である。R には SPSS でさえ実装されていないような新しい分析手法が多く含まれている。結果の信頼性も高く，最近では多くの学術論文が統計解析に R を使っている[*54]。

統計ソフトとしての R の特筆すべき利点は，実行したすべてのプロセスを，テキストファイルとして記録し，保存しておけるので，後になって，どういう分析をしたかをチェックすることができることである。しかも，保存しておいたファイルは（例えば test.R というファイル名だとすると），source("test.R") とすると再実行できる。どんなに複雑な作業をしても，それを何度でも簡単に再現できるということである。逆に考えれば，適当なテキストエディタでプログラムとして R のコマンドを書き連ねておいたものを読み込ませれば，複雑な分析手続きでも 1 回の操作で終わらせることができる。美しい図を作るのも実に簡単で，しかもその図を pdf とか postscript とか png とか jpeg とか Windows 拡張メタファイル (emf) の形式で保存でき，他のソフトに容易に取り込める。例えば emf 形式で保存すれば，Microsoft PowerPoint や OpenOffice.org の Draw などの中で，ベクトルグラフィックスとして再編集できる。

以前は，多くの日本人にとって最大の難点は，日本語が使えないことだったが[*55]，中間英治さんが日本語も扱えるようにするパッチを開発して公開されたのでこの問題は解決した。バージョン 2 からは本体が国際化対応したので，日本の R ユーザ有志の手によってメッセージまで日本語化されたものも使えるようになった。2006 年 5 月 9 日現在，最新版は 2.3.0 であり，会津大学や筑波大学など国内のミラーサーバから入手することができる。

以上の条件を考え合わせ，本実習では **R** を用いることにした。欧米の主な大学では R を使って演習を行なう学部や教室が増えてきているし，国内でも既にかなり多くの大学が採用している。演習室のコンピュータはすべてインストール済みだし，生態情報学セミナー室にも，インストール済みの Windows マシンを複数設置してあるので利用可能である。

[*54] 例えば, 2004 年 2 月には Nature にも R を使って分析された論文が掲載されている。Morris RJ, Lewis OT, Godfray HCJ: Experimental evidence for apparent competition in a tropical forest food web. Nature 428: 310-313, 2004.

[*55] データとしては入っていても大丈夫だったが変数名に使えなかったしグラフ内で使えなかった。

R のインストール情報

R-2.3.0 をインストールするには，以下のサイトの情報が参考になるだろう。

```
http://www.okada.jp.org/RWiki/index.php?R%B7%C7%BC%A8%C8%C4#content_1_1
```

筑波大学ミラーからダウンロードしたファイルをインストールするだけだと文字化けするので，Rconsole と Rdevga を編集する必要があることに注意されたい。

Macintosh 環境では本学社会情報学部の青木繁伸教授のサイト（下記）を参照されたい。

```
http://aoki2.si.gunma-u.ac.jp/R/begin.html
```

Linux，FreeBSD 環境を使っている方なら，自力でコンパイルすることにそれほど大きな困難はないはずである。Debian Linux ではコンパイル済みパッケージが公開されている。

R を使うための文献リスト

- 舟尾暢男『The R Tips データ解析環境 R の基本技・グラフィックス』九天社：CDROM 付き。リファレンスマニュアルのように使える本。
- Maindonald and Braun "Data Analysis and Graphics Using R", Cambridge Univ. Press：名著
- 間瀬，神保，鎌倉，金藤『工学のためのデータサイエンス入門 フリーな統計環境 R を用いたデータ解析』数理工学社：非常にバランスがいい本。

8 課題

以下 2 つの課題に，各々 A4 の用紙に 2 ページ以内で回答せよ。提出期限は 6 月 2 日までとする。生態情報学・中澤助教授室前に設置する箱に提出されたい。

8.1 課題１

`http://phi.med.gunma-u.ac.jp/grad/sample1.dat` は，web サイトで Excel 形式のファイルとして公開されている「厚生統計要覧」[*56]から作ったタブ区切りテキスト形式のデータである。`PREF` は都道府県名，`REGION` は東日本か西日本か，`MEH14` は平成 14 年の国民医療費（単位は億円），`PPMEH14` は平成 14 年の一人当たり国民医療費（単位は千円），`HBEDH14` は平成 14 年の病院病床数，`HNH16` は平成 16 年の病院数，`PPHNH16` は平成 16 年の一人当たり病院数，`PPNDH14` は平成 14 年の一人当たり医師数を示す変数である。

このデータから，東日本と西日本で一人当たり国民医療費に差があるかを検討せよ。さらに，もし差があるとしたら，一人当たり病院数で調整してもその差が残るかも検討せよ。

8.2 課題２

R のライブラリ `MASS` には, 組み込みデータとして, 有名な Gehan の白血病治験データが含まれており, `data(gehan)` とすると利用できるようになる。42 人の白血病患者を 2 群に分け，6-MP 治療群と対照群で生存時間を比較し，6-MP の治療効果を調べた治験データである。含まれている変数は，`pair` がペアを示すラベル，`time` が生存時間，`cens` が打ち切りフラグ，`treat` が 6-MP 治療群か対照群かを示すカテゴリ変数である。このデータを生存時間解析し解釈せよ。

[*56] `http://wwwdbtk.mhlw.go.jp/toukei/youran/index-kousei.html`