統計学 メモと追加の問

北海道大学理学部数学科 2012 年 8 月

以下，教科書の引用は，**服部哲弥「統計と確率の基礎」学術図書**の章節.

0　序

講義の前半 $\frac{2}{3}$ では，教科書の主に前半，数理統計学の基礎事項の部分について，各回で指定する教科書の該当章の読みどころや読み方を説明する.

このメモは，各回と教科書の対応関係，および，追加の練習問題である．具体的な講義内容やその他の練習問題は教科書を参照.

1　数理統計学

- 誤差や分布を伴うデータに数学を適用することで，データに基づく意思決定を支援する方法論の研究.

統計学の性質上，確率論の適用は重要.

2　確率論と統計学

確率論：「全測度 1 の測度と測度の定義された空間を定義域とする（可測）関数について数学的に証明できる特徴的な定理」

数理統計学：「確率論の結果を現実の場面（データ）においてどう理解するか」

- この講義の立場：数学としての確率論は確率論の講義に任せて，いくつかの確率論の結果を認めた上で，それをどのようにデータの理解に用いるかを講義

3　共通の基礎

数理統計学：抽象的結果の具体的案理解のしかたについてのインターフェース
インターフェース：対象分野やデータの性質で異なる事情の結果として多様

- インターフェースを場面に応じて分類し始めるときりがない.
- 実際の適用に必要な個別の知見はそれぞれの専門に任せて，個別の事情によらない共通の基礎（**確率論から直接続く部分，しかも，古典的な部分**）を講義.

1　確率測度と確率変数，平均と分散（教科書第 1,2 章）

1　確率空間

(i) **全体集合**Ω,
(ii) Ω の部分集合達（全てあるいは一部）集めた**集合族**$\mathcal{F}$,
(iii) **集合関数**（確率測度）$\mathrm{P}:\ \mathcal{F} \to \mathbb{R}$

の三つ組 $(\Omega, \mathcal{F}, \mathrm{P})$ であって公理を満たすもの.

公理：$\mathcal{F}$ は σ 加法族（可算和と補集合について閉じていること），P は σ 加法性（どの 2 つも共通部分のない，自然数添え字の集合列について，和集合の確率が各集合の確率の級数の和に等しいこと）を満たす非負値集合関数で $\mathrm{P}[\,\Omega\,] = 1$ を満たす.

確率変数 $X:\ \Omega \to \mathbb{R}$：大雑把には Ω 上の（普通の意味の）関数のこと．正確には，可測関数のこと.

可測関数：全ての実数 a に対して $\{X \geqq a\} := \{\omega \in \Omega \mid X(\omega) \geqq a\}$ の確率が測れる関数（つまり $(\forall a)\ \{X \geqq a\} \in \mathcal{F}$）

注：

- 集合論からは，Ω と，Ω の部分集合の集合 $\mathcal{F} \subset 2^{\Omega}$ までを考えるので，前者を集合，後者を集合族，と呼び分けることもある.
- Ω の部分集合に値を対応させる P は，集合論的には $\mathcal{F}$ 上の関数と呼んで差し支えないが，集合と集合族しか使わないことに対応して，後者を集合関数と呼び分けることもある.

確率を測度として定義する数学的価値（測度論の良さ）

(i) 古典的に知られた具体例（後述）を統一的に測度という一つの概念で扱う.
(ii) （大数の法則や中心極限定理を始め多くの極限定理が知られるが,）測度や期待値を取ること（確率測度に関する積分）と列の極限を取ることの順序が自由に交換できる（極限集合が定義域に自動的に入るなど）.
(iii) 長さ・面積・体積などの素朴な「大きさ」に基づく確率（$\mathbb{R}^n$ 上の確率）にとどまらず，関数の集合上の確率（ブラウン運動）など，抽象的な集合の大きさを考えられる.（たとえば独立確率変数列の大数の強法則は，可算個の独立確率変数の列を一つの Ω 上で考えているので，暗黙に '$\{0,1\}^{\infty}$' などの「無限次元」を考えている.）

2　講義で踏み込まないこと

数学から応用までの「細い理屈の線」を引くために，以下の数学の講義としては不満足な講義スタイルを取る.

- 確率の定義・公理が何を意味するか（基礎性質），どのような対象がそれを満たすか（具体例）はルベーグ積分（測度論）や確率論の講義に任せる.
- 数理統計学の話をする上で必要になる基礎性質はその時々に必要な範囲で紹介する.
- 以下の対象は確率空間であることを認めて先に進む.（途中でも適宜増やすかも知れない.）
  - 初等確率論：Ω は有限集合，$\mathcal{F} = 2^{\Omega}$，各要素 $i \in \Omega$ に対して $\{i\} \in \mathcal{F}$（根元事象）の確率 $p_i = \mathrm{P}[\,\{i\}\,]$ が与えられている．ただし，$0 \leqq p_i \leqq 1, \sum_{i \in \Omega} p_i = 1$.（確率測度の σ 加法性は有限加法性と同値.）

- 非負整数（整数など要素が自然数で添え字付けられる（可算無限集合）ならば同じ）上の確率測度：$\Omega = \mathbb{Z}_+$ 以外は初等確率論と同条件．ただし，$\sum_{i\in\Omega} p_i = 1$ の和の記号は級数の和の意味．そして，確率測度の σ 加法性も級数の和の意味で成り立つ．($p_i \geqq 0$ で全部加えても 1 なので，確率計算に現れる級数は全て収束.)
- $\mathbb{R}$ 上（$\mathbb{R}^n$ 上でも）の密度関数 ρ を持つ確率測度:$\mathrm{P}[\,A\,] = \int_A \rho(x)\,dx$．ただし，$\rho:\ \mathbb{R} \to \mathbb{R}$ は非負値で $\int_{\mathbb{R}} \rho(x), dx = 1$ を満たす区分的に連続な関数（区分的に連続でなくても可測関数ならば良い）で，積分はルベーグ積分の意味，P の定義域は，ルベーグ可測集合.

- 統計学のデータは有限個なので,（極限に関する数学的な裏付けがあることを認めるならば）現実の計算で測度論に戻る必要性は少ない．**（という言い訳のもとで，講義を先に進める.）**

3　離散分布の平均と分散

全体集合，確率測度（離散分布→根元事象），非負値性と $\sum_{k=0}^{n} Q(\{k\}) = 1$

平均と分散（標準偏差）の定義（教科書第 1 章 §2）　$\sum_{k=0}^{n} k\,Q(\{k\}) = \mu$

例題：　有限集合 $\Omega = \{0, 1, \cdots, n\}$ 上の平均 μ の離散分布 Q について，
$\sum_{k=0}^{n} (k-\mu)^2\,Q(\{k\}) = \sum_{k=0}^{n} k^2\,Q(\{k\}) - \mu^2$ を証明せよ．　◇

例：2 項分布（教科書第 1 章 §1）$B_{n,p}$：ベルヌーイ試行の表の和 N_n の分布として
2 項定理→平均 Np, 分散 $Np(1-p)$

4　可算集合上の分布の例

- **可算集合**＝要素が $\mathbb{N}$ で番号付けられる（ような大きさの）集合（カントール）
  - $\mathbb{N}$, $\mathbb{Z}_+$, $\mathbb{Z}$, $\mathbb{Q}$
  - $\Omega = \{\omega_i \mid i \in \mathbb{Z}_+\}$
  - **根元事象**の確率 $p_i = \mathrm{P}[\,\{\omega_i\}\,] \geqq 0$ を
  - $\mathrm{P}[\,\Omega\,] = \sum_{i=0}^{\infty} p_i = 1$ を満たすように与えれば確率空間（σ 加法性 OK）
- 非可算集合 $\mathbb{R}$（対角線論法）
  - 根元事象で定義する方法は不可（測度論，拡張定理）
  - 数列の集合，無限コイン投げ（ランダムウォーク），関数の集合→こういう全体集合を扱うために測度論に基づく近現代確率論が成立した（コルモゴロフ）

例：幾何分布 $p_i = q^i(1-q)$（始めて表が出るまでに裏の出た回数の分布）（教科書第 2 章章末問題）

例：パスカル分布 $p_i = {}_{a+i-1}\mathrm{C}_i q^a (1-q)^i$ （a 回表が出るまでに裏の出た回数の分布．$a = 1$ のとき幾何分布）（教科書第 1 章章末問題）

$\sum_{i=0}^{\infty} {}_{a+i-1}\mathrm{C}_i (1-q)^i = q^{-a}$ $(a = 1, 2, \ldots,\ 0 < q < 1)$

5　積分が定義する確率測度（教科書第 2 章 §1）

$\Omega = \mathbb{R}$ （$\mathbb{R}^n$，$\mathbb{R}$ の区間等も可）

密度関数 ($\rho:\ \mathbb{R} \to \mathbb{R}_+;\ \int_{\mathbb{R}} \rho(x), dx = 1$) があれば，$\mathrm{P}[\ A\ ] = \int_A \rho(x)\, dx$ は $\Omega = \mathbb{R}$ 上の確率測度を定義する（普通の積分）．P の定義域 $\mathcal{F}$ は区間，可算個の区間等を含む.

例：標準正規分布 $N(0,1)$: $\rho(x) = \dfrac{1}{\sqrt{2\pi}} e^{-x^2/2}$

正規分布 $N(m,v)$: $\rho(x) = \dfrac{1}{\sqrt{2\pi v}} e^{-(x-m)^2/(2v)}$

n 次元正規分布： $\Omega = \mathbb{R}^n$, $\rho(\vec{x}) = \dfrac{1}{\sqrt{2\pi v}^n} e^{-(\vec{x}-\vec{m})^2/(2v)}$

指数分布： $\Omega = \mathbb{R}_+$，$\rho(x) = w\, e^{-w\, x}$

平均，分散，標準偏差

例題：　コーシー分布 ，すなわち密度関数が $\rho(x) = \dfrac{1}{\pi}\dfrac{1}{1+x^2}$ で与えられる実数上の連続分布，は，確率分布ではあるけれども，期待値が無いことを示せ. ◇

6　確率変数（教科書第 2 章 §2）

確率変数（可測関数）：　$X:\ \Omega \to \mathbb{R}$ であって，$\mathrm{P}[\ X \geqq a\ ]$ が全ての a について定義されているもの ($\{X \geqq a\} = \{\omega \in \Omega \mid X(\omega) \geqq a\} \in \mathcal{F}$)

X の分布：$Q([a,\infty)) = \mathrm{P}[\ X \geqq a\ ]$ が定義する，$\mathbb{R}$ 上の確率測度

$Q(A) = \mathrm{P}[\ X \in A\ ] = \mathrm{P} \circ X^{-1}(A)$

（離散値確率変数のときは $Q(\{i\}) = \mathrm{P}[\ X = i\ ]$）

$f:\ \mathbb{R} \to \mathbb{R}$ がたとえば区分的に連続ならば $f(X) := f \circ X$ も確率変数.

期待値：　$\mathrm{E}[\ X\ ] = \int_\Omega X(\omega)\, \mathrm{P}[\ d\omega\ ] = \int_{\mathbb{R}} x\, Q(dx)$

Q が密度 ρ を持つとき $Q(dx) = \rho(x)dx$，Q が離散分布（X の値域が $\{x_0, x_1, \ldots\}$）のとき $\int_{\mathbb{R}} x\, Q(dx) = \sum_{i=0}^{\infty} x_i Q(\{x_i\})$

期待値の線形生

分散：　$\mathrm{V}[\ X\ ] = \mathrm{E}[\ (X - \mathrm{E}[\ X\ ])^2\ ]$

例題：　確率変数 X の分散 $\mathrm{V}[\ X\ ] = \mathrm{E}[\ (X - \mathrm{E}[\ X\ ])^2\ ]$ と定数 a に対して $\mathrm{V}[\ a\, X\ ] = a^2 \mathrm{V}[\ X\ ]$ を証明せよ. ◇

離散値確率変数の場合：　$\mathrm{E}[\ X\ ] = \sum_{\omega \in \Omega} X(\omega)\, \mathrm{P}[\ \{\omega\}\ ]$, $\mathrm{V}[\ X\ ] = \mathrm{E}[\ (X - \mathrm{E}[\ X\ ])^2\ ]$

例題：　X が $\mathrm{P}[\ X = k\ ] = \dfrac{a^k}{k!} e^{-a}$, $k = 0, 1, 2, \cdots$, （ポワッソン分布）に従うとき，期待値 $\mathrm{E}[\ X\ ]$ と分散 $\mathrm{V}[\ X\ ]$ を求めよ．a は正の定数とする. ◇

7　並列並びと一列並び（教科書第 2 章 §3）

例題：　かつて 1990 年頃 JR 東日本Y駅のみどりの窓口では，いったん開始した購入客の「一列並び」をまもなくやめた．問い合わせに対する回答によると，やめた理由は

(i) 平均時間は変わらない，
(ii) 列が長く見えてY駅は混んでいると思われてしまう，
(iii) 誘導人員が確保できない，

ということであった．このうち第 1 点については講義および教科書で説明するように，適切な理由とは言えない．実際，Y駅では 2000 年頃から一列並びを再開して，現在では定着している．第 1 点の誤解は指定券購入客にも多いであろう．この観点から，1990 年代に戻ったつもりになって，客の理解を得られない，と一列並びを渋る Y 駅の広報担当に代わり，一列並びについての客の理解を得るための広告説明文を提案せよ．なお，客の大多数は確率論の専門家ではないし，この講義を聞いてもいない．　◇

窓口 M，客 N，客 i の窓口処理時間 S_i，
$S_i : \Omega \to \mathbb{R}$ は平均 $\bar{t}$ 分散 σ^2 の**独立同分布確率変数**
独立確率変数の分散の加法性（次回の先取り）
分散の計算

バラツキの計測・考察が重要なとき：　待ち時間の余裕 $t_0 > \bar{t}$ のときに，間に合う確率
（確率の評価には具体的な分布が必要なので，今回の講義や教科書第 2 章では分散で直感的に判断するにとどめる）

問 1 ．　$0 < p < 1$ を定数とする．表の出る確率が p の硬貨を 1 回投げて表または裏にお金を賭ける賭博を行う．簡単のため，賭ける金額は一人当たり等しく 1 銭（架空のお金の単位）とし，手数料等は無く，賭けた金は全て集めて，勝った参加者に平等に分けるものとする．以下の小問に答えよ．

(i) $0 < q < 1$ とする．参加者の N 人中 $N\,q$ 人が表に賭け，残り $N\,(1-q)$ 人が裏に賭けたとき，表に賭けた人は平均（期待値で）いくら儲かるか？計算も示して答えよ．（まず 1 銭差し出すことを忘れないように．）
(ii) 参加者全員が合理的に考え，硬貨の表が出る確率が p であることを知っていて，皆の賭ける様子を見ながら賭を変えることができて，皆が納得してから硬貨が投げられるとすると，表に賭ける人の割合 q はいくらになるはずか？理由も示して答えよ．
(iii) $p > 0.5$ とする．参加者全員が公平な硬貨（表が出る確率が 0.5）と信じていて，前小問と同様の合理的行動をするとき，表に賭けた人の期待値を求め，正になることを確認せよ．

◇

2　データと独立確率変数と極限定理（教科書第 3,4 章）

1　データとは

立論・計算の基礎となる、既知のあるいは認容された事実・数値。資料。与件 [広辞苑]

目的のために範囲と方法を明確にして収集した（ここでは）数値（数値の組，ラベル付きの数値などを含む）の集まり.

数理統計学を適用するためにはデータが独立確率変数列のサンプルとなるように集めるべきである.（高校教科書の無作為抽出．実験では制御不能な誤差.）

データの大きさ (data size)：集めた数値が $x_1, x_2, \ldots, x_n$ のときの n

データは（無作為抽出ないしは制御できない誤差によって），たまたまその値が実現したと考える＝我々が住む世界（その値が実現した世界）は大きな（神のみぞ知る仮想的な「パラレルワールド」）Ω のサンプル $w \in \Omega$ における，確率変数列 $X_1, X_2, \ldots, X_n$ たちの値：$x_i = X_i(\omega)$, $i = 1, 2, \ldots, n$

講義と教科書では，X_i をデータ，x_i や ω をサンプル，X_i の分布を母分布，と呼ぶことにする．ω の集合 Ω が母集団だが，曖昧に母分布を母集団とも言うかもしれない．

講義と教科書ではデータを独立確率変数列のサンプルとしたときに何が言えるかを扱う．テーマは（有限サイズの）データから最大限言えることは何か？

(i) 数学的にはデータを独立確率変数列と捉える限り明確（極限定理が成り立つ列の最初の有限項の分布を見る，というだけ）.

(ii) 結論を現実の人がどう理解するかは難しい．時代でも分野でも表現・説明の仕方が変わる．**講義では仮説検定に基づく古典的な推測統計学の立場で説明．**

(iii) データが独立確率変数列として扱えるための実験（観測・測定）のしかたは非常に難しい．分野ごとに力点も変わるし，そもそも数学にならないだろう．講義では扱わない．独立確率変数列としてのデータが与えられたとして始める.

2　独立確率変数列としてのデータ（教科書第 3 章 §2）.

以下を認める（必要に応じて測度論や確率論で学ぶことを期待）：

- X_i, $i = 1, 2, \ldots, n$, が独立とは結合分布が分布の積で書けること
- $\mathrm{P}[\, X_i \leqq a_i,\ i = 1, 2, \ldots, n \,] = \mathrm{E}[\, \prod_{i=1}^{n} \mathbf{1}_{X_i \leqq a_i} \,]$ によって，X_i の関数の期待値が積で書けることと同値

分散の加法性

密度を持つ場合：結合分布の密度が周辺分布の密度の積で書けることと同値

独立でない 2 つの確率変数：**共分散，相関係数**

相関係数の絶対値は 1 以下である（± 1 に等しいとき，確率 1 で比例関係）

例題：　母平均が m で母分散 v の母集団から無作為抽出で取り出した大きさ n のデータ（独立同分布確率変数列）$X_1, \cdots, X_n$ の標本平均 $\bar{X}_n = \dfrac{1}{n}(X_1 + \cdots + X_n)$ の期待値 $\mathrm{E}[\, \bar{X}_n \,]$ と分散 $\mathrm{V}[\, \bar{X}_n \,]$ を求めよ（m, v, n だけを用いて表せ）. ◇

例題：　X_1, X_2 は独立な確率変数で，それぞれ区間 $[0, 1]$ 上の一様分布に従うとする．このとき，確率 $\mathrm{P}[\, X_1 + X_2^2 < 1 \,]$ を計算せよ．ここで $[0, 1]$ 上の一様分布とは密度関数 ρ が $\rho(x) = 1$, $0 \leqq x \leqq 1$, で与えられる $\Omega = [0, 1]$ 上の連続分布をいう. ◇

3　極限定理（教科書第 4 章 §2，教科書第 1 章 §3）.

この部分は教科書に準拠して既知の確率論の基礎事項を講義する.
以下の確率論の定理を認める（教科書第 4 章 §2）：

(i) LLN
(ii) 正規分布 $N(m, v)$
(iii) 弱収束（極限が密度を持つ $\mathbb{R}$ 上の確率測度の場合は区間の確率の収束）
（極限が $\mathbb{Z}$ 上の確率測度などの離散分布ならば，根元事象の確率の収束）
(iv) CLT

補足と参考事項：

- 4 次モーメントがある独立同分布確率変数列の LLN の証明と，CLT の証明の骨子に出てくる特性関数の説明の概要は教科書参照（教科書第 4 章 § 補足）. ただし，うかつにも誤植が残っているので注意（誤植の訂正は気づき次第（教わり次第）`http://web.econ.keio.ac.jp/staff/hattori/gakjutu.htm` にアップしています.）.
- 独立同分布確率変数列の LLN は,
  - 「（ドリフトのない）ランダムウォークの位置は歩数に比べて高次の微小量」という，変数列に依存する精密な意味（期待値が存在すれば成り立つ）と,
  - 「揺らぎの小さいものの多数の平均の揺らぎはたいへん小さい」という，変数列について一様な意味（たとえば 4 次モーメント有限ならば成り立つ）がある.
- 例：2 項分布の CLT と局所 CLT（教科書第 1 章 §3），スターリングの公式と e の定義

問 2 .　離散値確率変数 X と Y が独立なとき，すなわち，$\mathrm{P}[\, X = a,\ Y = b \,] = \mathrm{P}[\, X = a \,] \times \mathrm{P}[\, Y = b \,]$ が全ての取り得る値の組 (a, b) に対して成り立つとき，$\mathrm{E}[\, X^2 Y^3 \,] = \mathrm{E}[\, X^2 \,]\, \mathrm{E}[\, Y^3 \,]$ を示せ. ◇

3　統計的検定・推定の基本原理（教科書第 4 章 §1，第 5,6 章）

- 母分布の族，すなわち 2 項分布や正規分布やポワッソン分布のようにパラメータ（**母数**）を持つ族を与えて，データから母数について最大限言えることを言うための原理.
- 母数は，測定やサンプルの誤差の母数と背後の法則の定数や母分布のパラメータ両方を意味しうる.

1　統計の可能性と限界

- 少ないデータで最大限言う
- どんなサンプルも出る可能性があるから，そこから母集団について何も言えない，という立場もある
- 統計的推測（推測統計学の立場でもベイズ統計学の立場でもその他どんな立場でも）そこに登場する「確率」は存在しない過去または未来（または存在しない法則）の中での定量化．しかし，確率論という，整合的な数学としての一意的な結論（値）が得られるのだから,「少ないデータで最大限言う」以上はその数値を提出すべき.
- あくまで意志決定支援情報．意志を間違えないように表現（インターフェース）するためにどうすればよいかという，非数学的な部分について，数々の立場の間の対立や流行があるかもしれない.

2　点推定（教科書第 4 章 §1 ）

母数のもっともらしい値を統計量（データの関数）で与えること.

- **母平均** (E[X]) の推定・検定にはデータの**算術平均**，**母分散** (V[X]) の推定・検定にはデータの**不偏分散**，を用いるのが素直（教科書第 4 章 § 補足）.
- 自然な統計量を探すセンスが無いときは尤度のような一般論（教科書第 10 章）.

地震間隔の長期予測の例：表の数値を BPT(brownian passage time) 分布に当てはめる＝分布の族のパラメータ 2 つを決める．将来についての確率予測は条件付き確率.

3　検定（教科書第 5 章）

母数の値のもっともらしさの度合いを意志決定支援の形で定式化すること.

正規分布 (§2，§ 補足)：　密度，対称性，平均と分散，Z 変換 $Z = \dfrac{X - m}{\sqrt{v}}$，正規分布に従う独立な確率変数の和（特性関数のほうがやさしい），CLT，数表

検定の原理 (§3)：　帰無仮説，危険率・有意水準（信頼係数・信頼水準），危険域・棄却域，棄却・採択，第 1 種の過誤（第 2 種の過誤）

例題：　統計的検定の基礎的な手続きに関する次の文章の空欄 (1)–(5) を埋めよ.

以下の手続きを統計的検定とよぶ.

仮説 H:「母集団の分布は P である」を立てる（この，最初に立てる仮説を [(1)　　] とよぶ）．小さな正数である，と自分が判断する実数 α （[(2)　　] などとよぶ）を固定し，$\alpha = P(A)$ をみたす事象 A（[(3)　　] などとよぶ）を定める.

実際の実験・調査・観測などによって標本 x を得たとする．$x \in A$ ならば,『仮説 H が正しいのに A が生じるとはとうてい思えない,』と判断して，仮説 H を [(4)　　] し,「母集団の分布は P で

ない」と結論する．$x \notin A$ ならば仮説 H を [(5)　　] し,「母集団の分布は P でないとはいえない（さらなる研究が必要だ）」と結論する．

$\alpha = P(A)$ は H が正しいにもかかわらず [(4)　　] する確率，すなわち，[(6)　　] が起きる確率を表す．

これに対して，[(7)　　] の確率，すなわち，H が誤りにもかかわらず [(5)　　] する確率は上記の手続きでは求められない．これを求めるためにはたとえば対立仮説 H' を立てて，いずれかが正しい状況で上記手順を 2 度行う 2 仮説検定などの手続きを要する．　◇

4　区間推定（教科書第 6 章）

検定で採択される母数の範囲のこと．

- **母平均** ($\mathrm{E}[\,X\,]$) の推定・検定にはデータの**算術平均**を用いるのが素直→データサイズが大きければ CLT によって母分布の形によらずに正規分布が使える**（大標本理論）**
- 母平均が未知の時，不偏分散は独立確率変数の和ではないので CLT は自明ではない．たとえば母分布が正規分布の場合は先に進める（教科書第 7,8 章）．

例題：　ある集団中の無作為抽出標本 1 万個体中 100 個体にある病原菌の感染があった．この集団への病原菌の感染率 p を信頼水準 95% で区間推定せよ.（標本の感染率は本当は 2 項分布に従うが，これを正規分布で近似して計算せよ．ここで，確率変数 W_n の期待値と分散を $\mathrm{E}[\,W_n\,]$ と $\mathrm{V}[\,W_n\,]$ とするとき，W_n の分布を正規分布で近似するとは，$\frac{1}{\sqrt{\mathrm{V}[\,W_n\,]}}(W_n - \mathrm{E}[\,W_n\,])$ の分布が標準正規分布 $N(0,1)$ であるとすることを言う.）　◇

例題：　視聴率 x% （百分率で表示したもの）を x が $0 < x < 100$ のどの値であっても信頼水準 95% で小数第 1 位まで正確に得る，つまり，x の信頼区間の幅が 0.1 以内になるようにする，ために必要な調査対象世帯数を求めよ．視聴率は 2 項分布だが，正規分布で近似して計算せよ．　◇

問 3．　某テレビ番組によれば，さいころは目の数によって材料を削り取る量が異なるため目の出る確率が正確に $p = 1/6$ ずつではない，という．これを統計学的に検証するために，さいころを $n = 10000$ 回投げて 6 の出た回数 m を調べ，比 $X = m/n$ をデータとして帰無仮説　H: $p = 1/6$ を危険率（有意水準）1% で検定することにした．以下の問に答えよ．

(i) 帰無仮説 H の下で 6 の目が 1 万回中出る回数の期待値を求めよ．答は四捨五入して整数で．

(ii) H が棄却されるのは m が平均（期待値）からいくつ以上ずれたときか?「約 2100 以上ずれたとき」など有効数字 2 桁程度で答えよ．なお，H の下で X が従う分布は平均 p，分散 $v_n = \dfrac{p(1-p)}{n}$ の正規分布 $N(p, v_n)$ とし，必要ならば標準正規分布の表と $\sqrt{5} = 2.236$ を用いよ．

(iii) このサイコロを 2 回投げて 2 回とも 6 が出たとき，H を危険率 5% で検定せよ．結論はこの危険率の下でこのサイコロについてどういう判断を下すことになるか？　◇

4　正規母集団の統計学（教科書第 7,8 章）

母集団が正規分布（正規母集団）のときは，古くに詳しく調べられた.

正規分布が特に重視された（おそらく）理由：

- 精密実験の誤差→制御しきれない細かい微少な原因の集まり→ CLT から正規分布で近似できると期待される.
- 母分布に関係なく，データの算術平均（重み付きでも可）は CLT によって母平均に収束.

正規母集団 $N(m,v)$ から取った大きさ n のデータの算術平均は（CLT と無関係に）正規分布 $N(m,v/\sqrt{n}$ に従う（教科書第 5 章 §2）→正規母集団でその分散 v が既知ならば，データの算術平均から母平均を検定・推定することについて厳密な統計的推測の手順が取れる

1　χ^2 分布（教科書第 7 章 §1,2）

χ^2 分布：定義と密度の具体形（教科書第 7 章 §1）

母集団が $N(m,v)$ のとき $\dfrac{n-1}{v}V_n$ は χ^2_{n-1} に従う→正規母集団ならばデータの不偏分散から母分散を検定・推定することについて厳密な統計的推測の手順が取れる

尤度比検定（教科書第 10 章 §4,5）もデータが大きいとき χ^2 検定に帰着する（漸近理論）

例題：　母平均と母分散の両方とも未知の正規母集団から大きさ 51 の標本を無作為抽出し不偏分散 $V(\omega)$ を計算した．帰無仮説 H_0: 母分散が v，を有意水準 5% で検定するときの棄却域を v で表せ．なお，自由度 50 のカイ平方分布 χ^2_{50} について $\chi^2_{50}((32.36,\infty)) = 0.975$（32.36 を超える値が起きる確率が 0.975 という意味）および $\chi^2_{50}((71.42,\infty)) = 0.025$ であることを用いてよい．　◇

2　t 分布（教科書第 7 章 §3,4，第 8 章 §3）

t 分布：定義と密度の具体形（教科書第 7 章 §3）

母集団が $N(m,v)$ のとき $\sqrt{\dfrac{n}{V_n}}(\bar{X}_n - m)$ は T_{n-1} に従う→正規母集団ならばデータの算術平均と不偏分散から**（母分散についての情報無しに）**母平均を検定・推定することについて厳密な統計的推測の手順が取れる

- χ^2 分布と t 分布によって，正規母集団のパラメータ m,v の検定・推定が完成した.
- 1 つの正規母集団の推測統計学が完成したので，歴史は 2 つ以上の正規母集団の比較の理論に進んだ（t 分布と F 分布）.
  - 母分散の比の検定は母平均の異同に関係なく F 分布で可能.（教科書第 8 章 §2,4）
  - 等分散の 2 つの母集団の母平均の差の検定は t 分布で可能.（教科書第 8 章 §3）
    正規分布に従う 2 つの独立確率変数の差も正規分布に従う（教科書第 5 章 §2）ことから 1 つの母集団と同様に t 分布．→正規母集団ならばデータの算術平均と不偏分散から母平均を検定・推定することについて厳密な統計的推測の手順が取れる
    - 分散も異なると平均の差だけで決まる統計量がなく近似分布（Welch 分布）止まり.
    - 等分散の場合の母平均差の検定は，複数の差の要因が加法的な場合に母集団の個数についての一般化が F 分布によって完成.（分散分析：教科書第 8 章 § 補足）

    回帰分析（教科書第 9 章）も（データの誤差がわかっていれば）分散分析が可能.

3　F 分布（教科書第 8 章 §1,2, 補足）

F **分布：**定義と密度の具体形（教科書第 8 章 §1）

母集団が $N(m,v)$ **と** $N(m',v')$ **のとき** $\dfrac{v'}{v}\dfrac{V_m}{V'_n}$ **は** F_{n-1}^{m-1} **に従う**→正規母集団ならばデータの不偏分散から母平均の比を検定・推定することについて厳密な統計的推測の手順が取れる

例題：　以下の表は，標準正規分布に従う独立確率変数列の標本平均と不偏分散およびそれらの比の分布の数表の一部分である．標準正規分布 $N(0,1)$ と t 分布 T_n については $P((-\infty,-a)\cup(a,\infty))=0.05$ となる a，χ^2 分布 χ_n^2 と F 分布 F_n^m については $P((a,\infty))=0.05$ となる a，をそれぞれ与える部分を教科書の表から抜き出した．但し χ^2 分布については都合上教科書と行と列を入れ替えた．空欄の (1) から (5) に入るべき数字を解答用紙に (1) 0.00 ⋯ (5) 9.99 などのように書け．　なお同じ番号の空欄には同じ数字が入る．　◇

T_n

$n\setminus\alpha$	0.05
4	2.776
5	2.571
6	2.447
7	2.365
10	2.228
∞	(4)

$N(0,1)$

$P((a,\infty))$	a
0.025	1.96

F_n^m

$n\setminus m$	1	2	3	4
4	(1)	6.94	6.59	6.39
5	(2)	5.79	5.41	5.19
6	5.99	5.14	4.76	4.53
7	5.59	4.74	4.35	4.12
10	4.96	4.10	3.71	3.48
∞	(5)	(3)	2.60	2.37

χ_n^2

$\alpha\setminus n$	1	2	3	4
0.05	(5)	5.99	7.81	9.49

問 4 ．　(i)　10 個パックの卵の重さを量ったところ，グラム単位で 61, 62, 64, 64, 68，58, 63, 64, 66, 67，であった．鶏が産む卵の重さは正規分布 $N(\mu,v)$ にしたがってばらつくと仮定して，母分散 v を信頼水準 **95** % で両側区間推定せよ．ただし，上記 10 個の数字は，平均が $\bar{X}_{10}(w)=63.7$ で，$\bar{X}_{10}(w)$ との差の 2 乗の和が $9V_{10}(w)=\sum_{i=1}^{10}(X_i(w)-\bar{X}_{10}(w))^2=78.1$ であることを用いて良い．また自由度 n の χ^2 分布に従う確率変数 χ_n^2 について $\alpha=\mathrm{P}[\chi_n^2>c]$ となる c の数値は下表を用いてよい．(表の中の空欄 (ア)(イ) はあとの問で用いる．)

$n\setminus\alpha$	0.99	0.975	0.95	0.05	0.025	0.01
9	2.09	2.70	3.33	16.92	19.02	21.67
10	2.56	3.25	3.94	18.31	20.48	23.21
5000	4770	(ア)	4837	(イ)	5198	5236

(ii)　$X_i,\ i=1,\cdots,n,$ が独立で標準正規分布 $N(0,1)$ に従う確率変数のとき，$\chi_n^2=X_1^2+\cdots+X_n^2$ は自由度 n の χ^2 分布に従う．$\mathrm{E}[\,X_i^2\,]=1$, $\mathrm{V}[\,X_i^4\,]=2$ を用いると，n が大きいとき，中心極限定理から $\dfrac{1}{\sqrt{2n}}(\chi_n^2-n)$ の分布は $N(0,1)$ に近い．このことを用いて，問 2 の χ^2 分布の表の $n=5000$ の行の空欄 (ア)(イ) の近似値をそれぞれ求めよ．答えだけでなく，計算の経過も示すこと．$N(0,1)$ に従う確率変数 Z について，$\alpha=\mathrm{P}[Z>c]$ となる c の数値は次の表を用いてよい．　◇

α	0.05	0.025	0.01
c	1.6448	1.9600	2.326

5　回帰分析（教科書第 9 章，改訂版）

1　最小 2 乗法．

データを信号とノイズにわけて信号を定量的に説明する法則を導くことを目指すとき，もっとも単純な定量的法則は定数である．たとえば，真空中の光の速さ 3×10^8[m/s] は実験結果から統計的処理を経て得た定数である．本書でも最初に取り上げた母平均はばらつくデータ標本の背後に共通する定数である．

実験条件を制御できるときやデータが数値の対で与えられる場合を考える．たとえば実験条件が数値 x で与えられ，実験結果が数値 y で与えられるとき，大きさ n のデータの標本は，$i=1,2,\dots,n$ に対して $x=x_i$ と $y=y_i$ の組 (x_i,y_i) として記録される．2 変数 x,y のあいだの関数関係をデータから統計的に推測するとき，定数の次に単純な法則は直線関係（一次式）$y=a+bx$ である．y の x の上への**回帰式**（回帰方程式）ともいう．回帰式によって従属変数 y を独立変数 x で説明することを「**被説明変数**y を**説明変数**x に回帰する」ともいう．法則を定める定数（**母集団回帰係数**）$a=a^*$ と $b=b^*$ を統計的に推測することが**回帰分析**である．後戻りを意味する回帰 (regression) は奇異な用語だが，身長の高い親からも低い親からも平均に近い子供の生まれることが多く，世代と共に平均に戻る（回帰する），と命名者ガルトンは考えたらしい．

真の法則 $y=f(x)$ が説明変数 x の一次式でなくても，パラメータ（母集団回帰係数）について一次式であれば，たとえば x の二次式 $y=b\,x^2$ に回帰したい場合は，$x'=x^2$ で変数変換して，標本を $(x'_i,y_i)=(x_i^2,y_i)$ たちとして $y=b\,x'$ に回帰すればよい．さらに，係数が複雑な形で法則に入る場合も，f がなめらかで説明変数の変化が大きくなければ，数学的に言い換えると，f がテイラーの定理をみたし，かつ，定義域が定点 x_0 を中心とする剰余項が小さい範囲ならば，一次式 $y=f'(x_0)x+f(x_0)-x_0f'(x_0)$ で近似できるので，定数項と 1 次のの係数を改めて求めるべきパラメータとして回帰すること（近似法則）も考えられる．本章は一次式に絞って回帰分析を紹介する．

最小 2 乗法による母集団回帰係数の推定量 $a=\hat{a}(y_1,\dots,y_n)$ と $b=\hat{b}(y_1,\dots,y_n)$ を**標本回帰係数**（回帰係数）とよび，$\chi^2(a,b)=\sum_{i=1}^{n}(y_i-(a+b\,x_i))^2$ を最小にする a, b として定義する．標本と法則の偏差の 2 乗の和を最小にするようにパラメータを決める（点推定する）ことから**最小 2 乗法**とよばれる．

式を短くするために標本平均を $\bar{x}_n=\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}x_i$, $\bar{y}_n=\bar{y}_n(y_1,\dots,y_n)=\frac{1}{n}\sum_{i=1}^{n}y_i$, とおく．また，$x_1,\dots,x_n$ は固定して，関数の変数として書くのを略す．このとき，

定理． 標本回帰係数は

$$C=\sum_{i=1}^{n}(x_i-\bar{x}_n)(y_i-\bar{y}_n),\quad D=\sum_{i=1}^{n}(x_i-\bar{x}_n)^2,$$

とおくとき，$D\neq0$ ならば，

$$\begin{aligned}\hat{a}&=\hat{a}(y_1,\dots,y_n)=\bar{y}_n-\frac{C}{D}\,\bar{x}_n\,,\\ \hat{b}&=\hat{b}(y_1,\dots,y_n)=\frac{C}{D}\,,\end{aligned}$$

である．◇

証明．$\dfrac{\partial\chi^2}{\partial a}=0$ を解くと $\bar{y}_n=\hat{a}+\hat{b}\bar{x}_n$ を得る．これに x_i をかけて i について足したものを $\dfrac{\partial\chi^2}{\partial b}=0$ から引いて少し整理すると，主張を得る．

少なくとも一つの x_i の値を変えれば $D>0$ なので，x を変えたときの y の変化を知るための標本ならば定理の仮定 $D\neq0$ はみたされるはずである．

標本回帰係数を係数とする直線 $y = \hat{a} + \hat{b}\,x$ を**標本回帰直線**（回帰直線）といい，説明変数 x を与えたときの回帰直線の値を**予測値**（回帰値）という．たとえば，標本回帰係数を求めるために用いた値については

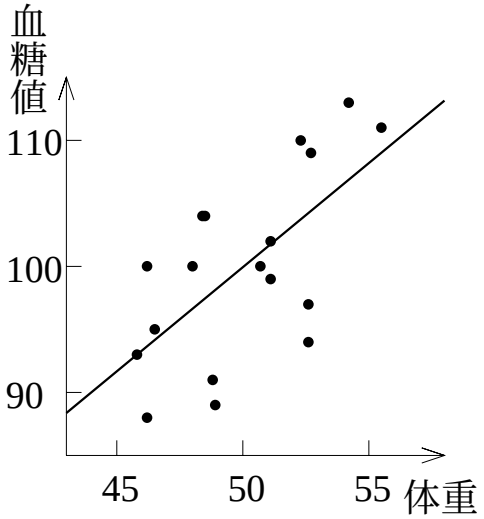
$$\hat{y}_i = \hat{a} + \hat{b}\,x_i$$

などとなり，

$$\frac{1}{n}\sum_{i=1}^{n} \hat{y}_i = \bar{y}_n$$

となる．例題として，著者の過去の健康診断結果から体重と血糖値の対と回帰直線を表とグラフ

i	体重 x_i	血糖値 y_i	予測値 $\hat{y}_i$
1	55.5	111	109
2	51.1	102	102
3	52.7	109	104
4	52.6	94	104
5	48.8	91	98
6	48.9	89	98
7	51.1	99	102
8	52.6	97	104
9	54.2	113	107
10	52.3	110	104
11	50.7	100	101
12	46.5	95	94
13	46.2	88	94
14	46.2	100	94
15	45.8	93	93
16	48.0	100	97
17	48.4	104	97
18	48.5	104	97
平均	50.0	100	
偏差	2.9	8	



健康診断の体重と血糖値の結果（黒丸）．直線は回帰直線．数値の単位は標準のものではない．

にした．ただし，プライバシー保護の観点から，日付は省き，測定値は定数倍して架空の単位系に取り直した．（決定係数の値やその検定結果は単位系のとり方によらない．）表の値を当てはめると，$C = 243$, $D = 147$, $\hat{a} = 1.65$, $\hat{b} = 17.3$ を得る．

体重や血糖値は生活習慣病の危険の手軽な目安として健康診断で重視される．再検査の基準値は医学の進歩か医療を取り巻く社会的状況の影響か時代に連れて変化するが，たとえば血糖値は，21 世紀初めのある時期の日本では，成人男子で 70–110[mg/dl] の範囲から外れると，再検査の呼び出しを受ける．家庭に血糖値の測定器が無くても体重計があれば，体重に目配りすることで，再検査呼び出しの可能性を減らせる．過去の健康診断の結果は忘れるというのも一つの生き方だが，数理統計学の立場は手持ちのデータから最大限言えることは何かを追求する．

以上は一個人についての回帰分析だが，標準の値は多数の人たちの結果に基づいて得られる．たとえば体重 y（キログラム）の身長 x（メートル）の上への回帰曲線としていわゆる標準体重 $y = 22x^2$ がある．一個人ごとに BMI 指数 (body mass index) y/x^2 を計算すると，ほとんどの人は 22 からずれる．回帰式のまわりにデータが分布することは個人差としてこの指数を利用する際に重要になる．BMI 指数の場合は 25 以上を肥満として，生活習慣病のリスクの目安とするようである．歴史的には一次式の回帰式 $y = 90(x - 1)$（ブローカの桂変法）などが用いられた．回帰式の選び方は，とくに得られた法則を**外挿**するときに注意を要する．外挿とは法則を決めるのに用いた変数 x の値の範囲の外側の x の値について，法則にもとづいて予測することをいう．x が時刻を表すとき，外挿は文字通り過去のデータに基づく未来の予測である．本章では一次式のみを

考えたが，関数（法則）の選択まで考えるときモデリングともいう．回帰分析をモデリングに応用する際は外挿の結果（予測）が関数の選び方で大きく変わることは注意を要する．

2　最小 2 乗法の根拠．

本書前半で紹介した考え方に基づいて，最小 2 乗法の根拠を簡単に説明する．説明変数 x_i は誤差のない数値で，被説明変数 $y_i = Y_i(\omega)$ は x_i から回帰式で説明できる値（信号）と統計的なばらつき（ノイズ）$Z_i(\omega)$ の和

$$Y_i = a^* + b^* x_i + Z_i, \quad i = 1, \ldots, n,$$

とする．本書前半と同様に，標本は無作為抽出，すなわち，Z_i は独立同分布確率変数列とし，

$$\mathrm{E}[\, Z_i \,] = 0, \quad i = 1, \ldots, n,$$

および，分散が存在する（有限な）こと $\mathrm{E}[\, Z_i^2 \,] < \infty$ を仮定する．仮定は**偏り**（バイアス，bias）がないことを意味する．偏りがあるとは，たとえば，高温下で延びたまま戻らなくなった質の悪いものさしで計ると数値が小さめに出る状況である．ものさしなら質のよいものに買いかえるが高価な機械はそうそう買いかえられないから，通常は較正 (calibration)，すなわち正しい値を得るように機械を調整するか，機械が較正に対応していなければ得られた数値を偏りのない数値に換算すること，を行なって偏りが起きないよう準備する．

定義から

$$\begin{aligned}
\frac{1}{n}\chi^2(a,b) &= \frac{1}{n}\sum_{i=1}^{n}(Y_i(\omega) - (a + b\, x_i))^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}((a^* - a) + (b^* - b)\, x_i^*)^2 \\
&\quad + \frac{2}{n}\sum_{i=1}^{n}((a^* - a) + (b^* - b)\, x_i^*)\, Z_i(\omega) + \frac{1}{n}\sum_{i=1}^{n} Z_i(\omega)^2.
\end{aligned}$$

右辺第 2 項は，たとえば $\{x_i\}$ が有界ならば，大数の強法則によって $n \to \infty$ で 0 に（確率 1 の ω で）収束し，第 3 項も（$X_i = Z_i^2$ についての大数の強法則によって）定数 $\mathrm{E}[\, Z_1^2 \,]$ に収束する．したがって，データの大きさ n が十分大きければ，$\chi^2(a,b)$ の最小値を与える $(\hat{a}, \hat{b})$ は正しい法則 (a^*, b^*) に近い．これが最小 2 乗法による推定の根拠である．

最小 2 乗法は回帰式からの偏差の 2 乗の和を最小にするようパラメータを決める．ここで見たように，その根拠には誤差 Z_i が同分布であることを用いるので，誤差が説明変数 x によって異なっていて，その分散の大きさが見積もれるときは，偏差の 2 乗を誤差の分散で割った量の和を最小にするほうが，より有効である．たとえばポワッソン分布（放射線の発生や事故発生件数など）のように平均と分散が等しい場合は $\chi^2(a,b) = \sum_{i=1}^{n} \frac{1}{y_i}(y_i - (a + bx_i))^2$ を最小にする．

3　決定係数と相関係数．

回帰残差，すなわち標本 y_i と予測値の差，の 2 乗和

$$S_e(y_1, \ldots, y_n) = \sum_{i=1}^{n}(y_i - \hat{y}_i(y_1, \ldots, y_n))^2$$

を**残差変動**，予測値と標本平均の差の 2 乗和

$$S_r(y_1, \ldots, y_n) = \sum_{i=1}^{n}(\hat{y}_i(y_1, \ldots, y_n) - \bar{y}_n(y_1, \ldots, y_n))^2$$

を**回帰変動**，とよぶ．回帰変動と残差変動の和が**全変動**，すなわち，標本と標本平均の差の 2 乗和，になること

$$S_{\rm tot} := \sum_{i=1}^{n}(y_i - \bar{y}_n(y_1,\ldots,y_n))^2 = S_r + S_e$$

が具体的に計算するとわかる．

標本 y_i のばらつきを，説明変数 x の変化で説明できる部分とできない部分の和に表して，前者を回帰式と呼ぶと考える．すると，回帰式の当てはまりの良さの目安として，回帰変動と全変動の比 $R^2 = \dfrac{S_r}{S_{\rm tot}}$ が自然である．R^2 を**決定係数**あるいは寄与率とよぶ．

標本平均を $\bar{x}_n$ と $\bar{y}_n$ と書くとき，標本相関係数

$$\tilde{R} = \frac{\displaystyle\sum_{i=1}^{n}(x_i - \bar{x}_n)(y_i - \bar{y}_n)}{\sqrt{\displaystyle\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 \sum_{i=1}^{n}(y_i - \bar{y}_n)^2}}$$

は決定係数の平方根 R に等しい: $R = \tilde{R}$．2 つの確率変数が独立 $(r = 0)$ か本質的に同じか $(r = \pm 1)$ どちらの状況に近いかを相関係数が判定する目安になりうる．相関係数は母分布の（神のみぞ知る）値，標本相関係数はその推定量である．先ほどの例では，$S_t = 973$, $S_r = 402$, $S_e = 571$, $R^2 = 0.41$, $\tilde{R} = R = 0.64$, となる．

4　回帰係数の検定．

5　重回帰分析．

説明変数を多変数にして 2 以上の整数 p に対して，$y = b_0 + b_1 x_1 + \cdots b_p x_p$ なる回帰方程式に大きさ n の標本 $(x_{1,i},\ldots,x_{p,i},y_i)$, $i = 1,\ldots,n$, を当てはめる分析を重回帰分析という．最小 2 乗法は 1 変数の場合と同様に,

$$\chi^2 = \sum_{i=1}^{n}(y_i - (b_0 + b_1 x_{1,i} + \cdots + b_p x_{p,i}))^2$$

を最小にする $b_j = \hat{b}_j(y_1,\ldots,y_n)$, $j = 1,2,\ldots,p$, を推定量（回帰係数）とする．具体形は

$$\hat{b}_j = \hat{b}_j(y_1,\ldots,y_n) = \sum_{k=1}^{p}(D^{-1})_{j,k}\, C_k\,,\ j = 1,\ldots,p,$$
$$\hat{b}_0 = \hat{b}_0(y_1,\ldots,y_n) = \bar{y}_n - \sum_{j=1}^{p}\hat{b}_j \bar{x}_{j,n}$$

で与えられる．ここで $\bar{y}_n = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} y_i$，$\bar{x}_{j,n} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} x_{j,i}$，$C_k = \displaystyle\sum_{i=1}^{n}(y_i - \bar{y}_n)(x_{k,i} - \bar{x}_{k,n})$，とおき，$(D^{-1})_{j,k}$ は $D_{j,k} = \displaystyle\sum_{i=1}^{n}(x_{j,i} - \bar{x}_{j,n})(x_{k,i} - \bar{x}_{k,n})$ で (j,k) 成分が与えられる $p \times p$ 行列 D の逆行列の (j,k) 成分を表す．$p = 1$ の場合と同様に，$x_{j,i}$ たちは D^{-1} が存在するように選ぶとする．

$p = 1$ の場合と同様に，回帰値 $\hat{y}_i = \hat{b}_0 + \displaystyle\sum_{j=1}^{p}\hat{b}_j x_{j,i}$ は y_i のばらつきのうち x_{ji} たちによって説明できる部分，回帰残差 $y_i - \hat{y}_i$ は制御できない擾乱 Z_i たちによるばらつきと解釈する．対応して，

回帰変動と残差変動

$$S_r(y_1,\dots,y_n)=\sum_{i=1}^{n}(\hat{y}_i(y_1,\dots,y_n)-\bar{y}_n(y_1,\dots,y_n))^2$$
$$S_e(y_1,\dots,y_n)=\sum_{i=1}^{n}(y_i-\hat{y}_i(y_1,\dots,y_n))^2$$

を定義すると $p=1$ の場合と同様に両者の和は全変動になる：

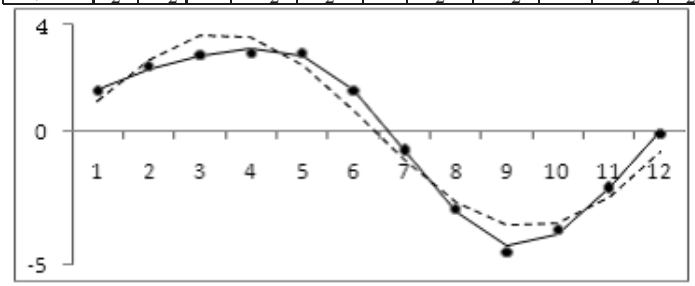
$$S_{\mathrm{tot}}(y_1,\dots,y_n)=\sum_{i=1}^{n}(y_i-\bar{y}_n(y_1,\dots,y_n))^2=S_r+S_e.$$

S_r と S_{tot} の比を決定係数（寄与率）とよぶのも $p=1$ と同様である．

回帰係数 $\hat{b}_j$ の検定や区間推定は $p=1$ の場合と同様に可能である．

章の冒頭で，説明変数は元の変数の関数でも差し支えないことを注意した．重回帰分析との関連では，元の変数のいくつかの関数の和で書かれた法則を考えて，各項の比例係数をデータから求めることになる．図は，気象庁のウェブページにある気象統計情報の一つで，西暦 2000 年前後

月 i	1	2	3	4	5	6	7	8	9	10	11	12	重回帰
濃度 y_i	1.5	2.4	2.8	2.9	2.9	1.5	-0.7	-2.9	-4.5	-3.7	-2.1	-0.1	係数 b_j
$x_{1,i}$	$\frac{1}{2}$	$\frac{\sqrt{3}}{2}$	1	$\frac{\sqrt{3}}{2}$	$\frac{1}{2}$	0	$-\frac{1}{2}$	$-\frac{\sqrt{3}}{2}$	-1	$-\frac{\sqrt{3}}{2}$	$-\frac{1}{2}$	0	3.53
$x_{2,i}$	$\frac{\sqrt{3}}{2}$	$\frac{1}{2}$	0	$-\frac{1}{2}$	$-\frac{\sqrt{3}}{2}$	-1	$-\frac{\sqrt{3}}{2}$	$-\frac{1}{2}$	0	$\frac{1}{2}$	$\frac{\sqrt{3}}{2}$	1	-0.78
$x_{3,i}$	$\frac{\sqrt{3}}{2}$	$\frac{\sqrt{3}}{2}$	0	$-\frac{\sqrt{3}}{2}$	$-\frac{\sqrt{3}}{2}$	0	$\frac{\sqrt{3}}{2}$	$\frac{\sqrt{3}}{2}$	0	$-\frac{\sqrt{3}}{2}$	$-\frac{\sqrt{3}}{2}$	0	0.04
$x_{4,i}$	$\frac{1}{2}$	$-\frac{1}{2}$	-1	$-\frac{1}{2}$	$\frac{1}{2}$	1	$\frac{1}{2}$	$-\frac{1}{2}$	-1	$-\frac{1}{2}$	$\frac{1}{2}$	1	0.76



月平均二酸化炭素濃度（黒丸）．単位は百万分率 ppm．数値は南鳥島の 1995 年から 2005 年の平均値から，この期間の平均濃度上昇 (年 1.9) と平均濃度 (369.8) を引いた値．点線は $x_{1,i}=\sin\frac{2\pi i}{12}$ と $x_{2,i}=\cos\frac{2\pi i}{12}$ による回帰 ($p=2$)，実線は $x_{3,i}=\sin\frac{4\pi i}{12}$ と $x_{4,i}=\cos\frac{4\pi i}{12}$ も加えた回帰 ($p=4$)．

の鳥島での i 月（$i=1,2,\dots,n,\ n=12$) の平均二酸化炭素 (CO_2) 濃度 y_i(ppm) の値である．但し，紙面の節約のため，気象庁のデータを加工した．まず，数値の月別の変化に注目するために，平均濃度を引いて表の数値の平均を 0 にとり，また，この時期の CO_2 が年とともに高くなる傾向が知られているので，平均上昇率に相当する値を差し引いた．濃度上昇のトレンドを差し引いたので表のデータは季節変化に対応する 12ヶ月の周期関数と考えられる．そこで，月の数 i について 12 を周期とする周期関数で回帰するのが自然である．基本的な周期関数として三角関数を選び，$q=1,2,3,4,5$ に対して，$x_{2q-1,i}=\sin\dfrac{2q\pi i}{12}$, $x_{2q,i}=\cos\dfrac{2q\pi i}{12}$，および前者で $q=6$ とした $x_{11,i}$ の 11 個の説明変数をとる．これに b_0 に対応する $x_{0,i}=1$ を加えた 12 個の $x_{k,i}$ たちの線形結合で，任意の標本，すなわち任意の $\{y_i\}_{i=1}^{12}$ を（連立一次方程式を解くことで）表せる．特に，平均を 0 にとったので $b_0=0$ である．誤差無しで表せるのは本書の立場からは「説明のしすぎ」なので，説明変数を減らして $p=2(q=1)$ と $p=4(q=1,2)$ の重回帰分析を試みる．

結果は図のように，基本周期 $q=1$ のみでは残差が大きく，$q=2$ まで入れると変化をよく説明するように見える．CO_2 の濃度は主に陸と海の分布や植生の緯度に対する分布と地球の太陽に対する傾きの関係で決まると想像すると，三角関数の 2 乗は半角公式によって周期が半分の三角関

数と定数で表せるので，地球の傾きの線形な直接の影響だけでなく，2 乗という非線形な効果が大きいことを回帰の結果は示唆する．

ところで，ここで選んだ三角関数の系 $\{x_{k,\cdot}\}_{k=0}^{p}$ は完全性，すなわち（p を増やせば）全ての周期変化を表せる性質，を持つだけでなく，直交性という著しい性質がある．実際 $D_{j,k}$ に上記の $x_{k,i}$ と i についての平均 $\bar{x}_{j,n}=0$ $(n=12)$ を代入すると,（三角関数の性質または表を参照することで,）$D_{j,k}=\sum_{i=1}^{n} x_{j,i}x_{k,i}=6\delta_{j,k}$ を満たす．すなわち，D は対角行列である（直交性）．特に逆行列が $(D^{-1})_{j,k}=\frac{1}{6}\delta_{j,k}$ となるので，平均 $\bar{y}_n=0$ と合わせると，$\hat{b}_j=\frac{1}{6}C_j=\frac{1}{6}\sum_{i=1}^{n} y_i x_{j,i},\ j=1,\ldots,p,$ を得る．一般の重回帰分析では p 個の説明変数の組み合わせを選んでから D^{-1} を計算しなければならないが，説明変数の間に直交性があれば回帰係数 $\hat{b}_j$ は対応する説明変数 $x_{j,i}$ とデータ y_i だけで決まり，他の説明変数の取り方にも個数 p にもよらない．特に，当てはまり具合を見ながら，どこまで説明変数を増やすかを決めることが容易である．

補足：定理の証明のあらすじ．

Z_i たちが独立同分布で正規分布 $N(0,v)$ に従うから，
$\hat{b}(Y_1,\ldots,Y_n)-b^*$ が $N(0,\frac{v}{D})$ に従うこと，$\frac{1}{v}S_e(Y_1,\ldots,Y_n)$ が自由度 $n-2$ の χ^2 分布に従うこと，そして両者が独立であることをいえば t 分布の定義から $\hat{b}$ についての証明が終わる．

以下，n 成分列ベクトルを $\vec{p}$，その第 i 成分を p_i のように書く．特に，全成分が $n^{-1/2}$ のベクトルを $\vec{p}^{(1)}$，第 i 成分が $(x_i-\bar{x}_n)\,D^{-1/2}$ のベクトルを $\vec{p}^{(2)}$ とおく．ベクトル $\vec{p}$ と $\vec{q}$ の内積を $\vec{p}\cdot\vec{q}$ と書くと，$\vec{p}^{(1)}\cdot\vec{p}^{(1)}=\vec{p}^{(2)}\cdot\vec{p}^{(2)}=1$，$\vec{p}^{(1)}\cdot\vec{p}^{(2)}=0$，すなわち，$\vec{p}^{(1)}$ と $\vec{p}^{(2)}$ は直交する単位ベクトルである．シュミットの直交化法により，第 1,2 列がそれぞれ $\vec{p}^{(1)}$ と $\vec{p}^{(2)}$ の実直交行列 O がある．O の転置行列を O^T，n 次単位行列を I_n と書くと，$O^T O=O\,O^T=I_n$ である．O の第 i 列からなるベクトルを $\vec{p}^{(i)}$ とおく．$O^T O=I_n$ から $\vec{p}^{(i)}\cdot\vec{p}^{(j)}=\delta_{i,j}$ である．$\vec{Z}$ を第 i 成分が Z_i の列ベクトルとすると，$(O^T\vec{Z})_i=\vec{p}^{(i)}\cdot\vec{Z}$ と書ける．

$\hat{b}$ について，少し整理すると

$$\hat{b}-b^*=\hat{b}(Y_1,\ldots,Y_n)-b^*=\frac{1}{D}\sum_{i=1}^{n}(x_i-\bar{x}_n)\,Z_i=\frac{1}{\sqrt{D}}\vec{p}^{(2)}\cdot\vec{Z}$$

となる．これは正規分布 $N(0,v)$ に従う独立確率変数 Z_i たちの一次式だから正規分布に従う．その期待値は 0，分散は

$$\mathrm{V}[\,\hat{b}(Y_1,\ldots,Y_n)\,]=\frac{v}{D^2}\sum_{i=1}^{n}(x_i-\bar{x}_n)^2=\frac{v}{D}\,.$$

S_e も $\hat{b}$ と同様に計算すると

$$S_e=S_e(Y_1,\ldots,Y_n)=\vec{Z}\cdot\vec{Z}-(\vec{p}^{(1)}\cdot\vec{Z})^2-(\vec{p}^{(2)}\cdot\vec{Z})^2$$

と書ける．$O\,O^T=I_n$ と $(O^T\vec{Z})_i=\vec{p}^{(i)}\cdot\vec{Z}$ から，$\vec{Z}\cdot\vec{Z}=\sum_{i=1}^{n}(\vec{p}^{(i)}\cdot\vec{Z})^2$ なので，$S_e=\sum_{i=3}^{n}(\vec{p}^{(i)}\cdot\vec{Z})^2$ である．$i=1,2$ の項が消えることに注意．

$\vec{p}^{(i)}\cdot\vec{p}^{(j)}=\delta_{i,j}$ なので，$\vec{p}^{(k)}\cdot\vec{Z}$, $k=1,2,\ldots,n$, は独立同分布確率変数列で，その分布は $N(0,v)$ である．確率変数たちが独立ならばその関数たちも独立なので，$\hat{b}-b^*$ と S_e は独立で，また，$\frac{1}{v}S_e$ は標準正規分布に従う $n-2$ 個の独立確率変数の 2 乗の和となるから，自由度 $n-2$ の χ^2 分布に従う．

$\hat{a}=\hat{a}(Y_1,\ldots,Y_n)$ についても $\hat{b}$ と同様の計算によって

$$\hat{a}-a^*=\sum_{i=1}^{n}\Big(\frac{1}{n}-\frac{\bar{x}_n}{D}(x_i-\bar{x}_n)\Big)Z_i=\Big(\frac{\vec{p}^{(1)}}{\sqrt{n}}-\frac{\bar{x}_n\vec{p}^{(2)}}{\sqrt{D}}\Big)\cdot\vec{Z}$$

を得るので，平均は a^*，分散は

$$\mathrm{V}[\,\hat{a}(x_1,\ldots,x_n,Y_1,\ldots,Y_n)\,]=v\sum_{i=1}^{n}\left(\frac{1}{n}-\frac{\bar{x}_n}{D}(x_i-\bar{x}_n)\right)^2=\frac{v}{nD}\sum_{i=1}^{n}x_i^2$$

となることに注意すると，同様に証明できる．