

# 地球環境統計解析法<sup>1</sup>

筑波大学陸域環境研究センター

浅沼 順 (asanuma@suii.tsukuba.ac.jp)

平成 19 年 11 月 16 日

<sup>1</sup>第 1.2 版

# 目次

0.1	教科書および参考書	5
第1章	基礎的な統計量	6
1.1	度数分布とヒストグラム	6
1.2	代表値	6
1.2.1	平均 (mean)	7
1.2.2	メディアン (median)	7
1.2.3	最頻値 (モード)	7
1.3	変動あるいは散らばりの尺度	7
1.3.1	レンジ	7
1.3.2	分散・標準偏差	8
1.3.3	変動係数	8
1.4	その他の指標	8
1.4.1	モーメント	8
第2章	確率変数と確率分布	11
2.1	離散確率変数	11
2.1.1	確率分布	11
2.1.2	期待値・分散	11
2.2	連続確率変数	13
2.2.1	確率密度関数	13
2.2.2	期待値・分散	14
2.3	正規分布	14
第3章	母集団と標本、点推定、区間推定	17
3.1	概念	17
3.2	点推定と不偏推定量	17
3.3	区間推定	19
3.3.1	正規分布母集団の平均の区間推定	19
3.3.2	正規分布母集団の分散の区間推定	21
第4章	仮説検定	24
4.1	正規母集団の平均に対する仮説検定	25
4.1.1	母分散 $\sigma$ が既知の時	25
4.1.2	母分散が未知の時	27

<b>第 5 章</b>	<b>2 つの母集団の問題</b>	<b>29</b>
5.1	平均の差	29
5.1.1	母分散が既知の時	29
5.1.2	母分散が未知であるが等しいとき	30
5.1.3	母分散が未知であるが等しくないとき	31
5.2	分散の比	32
<b>第 6 章</b>	<b>相関と回帰分析</b>	<b>35</b>
6.1	2 変数の統計量	35
6.1.1	サンプルの共分散・相関係数	35
6.1.2	母集団の共分散・相関係数	36
6.2	単純線形回帰モデル	36
6.2.1	最小二乗推定法	36
6.2.2	回帰直線の特性と決定係数	37
6.3	様々な線形回帰モデル	38
6.3.1	独立変数と従属変数	38
6.3.2	生態相関	39
6.4	原点を通る線形回帰モデル	40
6.4.1	原点を通る単純回帰モデル	40
6.4.2	原点を通る生態相関	41
6.5	相関係数に関する区間推定・仮説検定	41
6.5.1	区間推定	41
6.5.2	仮説検定	42
6.6	回帰分析に関する区間推定・仮説検定	42
6.6.1	単純回帰直線	42
6.6.2	原点を通る回帰直線に関する検定	44
6.7	演習	45
<b>第 7 章</b>	<b>多変量回帰分析 (Multiple Regression)</b>	<b>48</b>
7.1	線形回帰分析の行列表現	48
7.1.1	最小二乗法の行列表現	48
7.2	多変量線形回帰分析 (Multiple Regression)	50
7.3	偏決定係数	51
7.4	演習	52
<b>第 8 章</b>	<b>時系列データの解析</b>	<b>55</b>
8.1	基本的な概念	55
8.1.1	決定論的と確率論的	55
8.1.2	決定論的データ	55
8.1.3	確率論的データ	57
8.2	確率過程データの解析	59
8.3	定常確率過程	59
8.4	エルゴード性を持つ確率過程	61
8.5	スペクトル密度関数	62

8.5.1	相関関数からのスペクトル . . . . .	62
8.5.2	有限フーリエ変換からのスペクトル . . . . .	63
8.5.3	スペクトルの性質 . . . . .	63
<b>第 9 章</b>	<b>分散分析 (ANOVA, Analasys of Variance)</b>	<b>65</b>
9.1	一因子分散分析 (Single-Factor ANOVA) . . . . .	65
9.1.1	定義 . . . . .	65
9.1.2	前提条件 . . . . .	66
9.1.3	ANOVA 表 . . . . .	67
9.1.4	ANOVA 検定 . . . . .	69

# 目 次

1.1	右から表 1.1 のデータのヒストグラム, 累積度数グラフ、累積相対度数グラフ . . .	7
1.2	歪度が正の変動量 (Tennekes and Lumley, 1972) . . . . .	9
1.3	尖度が大きいの変動量と小さい変動量 (Tennekes and Lumley, 1972) . . . . .	9
2.1	例 2.1 の確率分布、累積確率分布の例 . . . . .	12
2.2	確率密度関数。 $P(a \leq X \leq b)$ は黒く塗った部分の面積に相当する。 . . . .	13
2.3	累積確率密度関数。 $F(x) = P(X \leq b)$ は黒く塗った部分の面積に相当する。 . . .	14
2.4	標準正規分布 . . . . .	15
3.1	母集団, サンプルと統計的推定の概念 . . . . .	17
3.2	$z_{\alpha/2}$ の定義。黒く塗った部分の面積が $1 - \alpha$ になる。 . . . .	19
3.3	$t$ 分布と正規分布 . . . . .	21
3.4	$t_{\alpha/2, \nu}$ の定義。黒く塗った部分の面積が $1 - \alpha$ になる。 . . . .	21
3.5	$\chi^2$ 分布 . . . . .	22
3.6	$\chi^2_{\alpha, \nu}$ の定義。黒く塗った部分の面積が $\alpha$ になる。 . . . .	22
5.1	$F$ 分布の例と $F_{\alpha, \nu_x, \nu_y}$ の定義。黒く塗った部分の面積が $\alpha$ になる。 . . . .	33
6.1	式 6.20 および式 6.33、式 6.36 による線形回帰直線。データはモンゴルにおいて 2 種類の手法によって観測された顕熱フラックスである。 . . . .	39
8.1	三角関数 (左) とそのスペクトル (右) . . . . .	55
8.2	ほぼ周期的なデータの例 $x(t) = \sin 2\pi t + 0.5 \sin 4\pi t + 0.25 \sin 8\pi t$ 。時系列 (左) とスペクトル (右) . . . . .	56
8.3	複雑な周期データの例 $x(t) = \sin 2\pi t + 0.5 \sin 4\pi t + 0.25 \sin 2\sqrt{3}\pi t$ 。時系列 (左) とスペクトル (右) . . . . .	56
8.4	短期的な周期データの例。 . . . .	57
8.5	確率論的データの例。2 秒間の風速成分の瞬間値。3 つの別の地点での同時の観測。 . . . .	57

# はじめに

## 0.1 教科書および参考書

本講義の教科書とする．

- 「統計学入門」： 東京大学教養学部統計学教室 (1991)

以下は日本語の参考書．比較的応用的な内容をわかりやすく説明しており，かつ広範囲にカバーしている

- 「自然科学の統計学」：東京大学教養学部統計学教室 (1992a)

同じく「人文・社会科学の統計学」(東京大学教養学部統計学教室, 1992b) もある．また、時系列解析では以下の本がよい参考書になる。

- 「スペクトル解析」：日野 (1977)

また以下は，英語での参考書である．どちらも日本語のどの専門書よりも広範囲の内容を盛りだくさんにカバーしている．

- “Probability and Statistics for Engineering and the Sciences” by J. L. Devore, 1991, Brooks/Cole Pub. Co. (Devore, 1991)
- “Applied Linear Statistical Models” by J. Neter, W. Wasserman and Michael H. Kutner, 1990, Irwin Inc. (Neter et al., 1990)
- Random Data Analysis and Measurement Procedure: Bendat and Piersol (1971)

# 第1章 基礎的な統計量

資料の性質を表す統計量を以下に挙げる。

## 1.1 度数分布とヒストグラム

**度数分布 (頻度分布)** 一組のデータセットを階級 (class) に分け、それぞれの階級で資料の中の変量がいくつあるか (度数, 頻度, frequency) を表にしたものを度数分布 (頻度分布, frequency distribution) 表と呼ぶ

**相対度数** 資料中のデータ総数に対する相対的な度を相対度数 (relative frequency) と呼ぶ

**累積度数** 度数を下から順に積み上げたときの累積和を累積度数 (cumulative frequency) と呼ぶ

**累積相対度数** 相対度数を下から積み上げたときの累積和を累積相対度数 (cumulative relative frequency) と呼ぶ

表 1.1: 試験得点の度数分布表 (某大学某講義)。(東京大学教養学部統計学教室, 1991, より引用)

階級		階級値	度数	相対度数	累積度数	累積相対度数
以上	未満					
0	10	5	12	0.032	12	0.032
10	20	15	10	0.027	22	0.059
20	30	25	19	0.051	41	0.110
30	40	35	42	0.113	83	0.223
40	50	45	72	0.193	155	0.416
50	60	55	82	0.220	237	0.635
60	70	65	54	0.145	291	0.780
70	80	75	38	0.102	329	0.882
80	90	85	25	0.067	354	0.949
90	100	95	19	0.051	373	1.000

## 1.2 代表値

一組のデータセットが「どのあたりに」位置するかを示す値を代表値と呼ぶ。

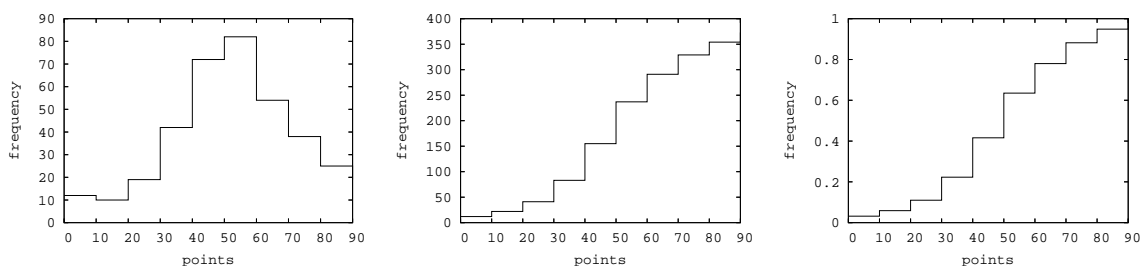


図 1.1: 右から表 1.1 のデータのヒストグラム, 累積度数グラフ, 累積相対度数グラフ

### 1.2.1 平均 (mean)

$n$  個の変数 (variable) を  $x_1, x_2, x_3, x_4, \dots, x_n$  とすると, 代表値として最もよく用いられる統計量が平均 (mean) であり, その中で特に算術平均 (arithmetic mean) である. 算術平均を  $\bar{x}$  で表す.

$$\bar{x} \equiv \frac{1}{n}(x_1 + x_2 + x_3 + x_4 + \dots + x_n) = \frac{1}{n} \sum_{i=1}^n x_i \quad (1.1)$$

算術平均と並んで幾何平均 (geometric mean)

$$x_G \equiv \sqrt[n]{x_1 x_2 \dots x_n} \quad (1.2)$$

や調和平均 (harmonic mean),

$$\frac{1}{x_h} = \frac{1}{n} \sum_{i=1}^n \frac{1}{x_i} \quad (1.3)$$

などがある.

### 1.2.2 メディアン (median)

中央値, 中位数とも呼び,  $\tilde{x}$  で表す. データ  $x_i$  ( $i = 1 \dots n$ ) を小さい方から順に並べた新しいデータを  $x'_i$  ( $i = 1 \dots n$ ) とすると,  $n$  が奇数の場合は  $\tilde{x} = x'_{(n-1)/2+1}$ , 偶数の場合は,  $\tilde{x} = (x'_{n/2} + x'_{n/2+1})/2$  である.

メディアンと同じ考え方のものに, 分位点がある. 小さい方から 100p% のところにある値を 100p パーセンタイル (percentile) または百分位点という. また, 25% 分位点, 50% 分位点, 75% 分位点を第 1, 第 2, 第 3 四分位点ともいう. 第 2 四分位点はメディアンである.

### 1.2.3 最頻値 (モード)

度数分布表において, もっとも頻度が高い値. ヒストグラムのピークを示す値に相当する.

## 1.3 変動あるいは散らばりの尺度

### 1.3.1 レンジ

文字通り, 変動の範囲.  $\max(x_i) - \min(x_i) = x'_n - x'_0$ .

### 1.3.2 分散・標準偏差

変数  $x_i$  の平均との差  $x_i - \bar{x}$  を偏差 (deviation) と呼ぶ。偏差の平均は、ゼロである。

$$\sum_{i=1}^n (x_i - \bar{x}) = \sum_{i=1}^n x_i - \sum_{i=1}^n \bar{x} = n\bar{x} - n\bar{x} = 0 \quad (1.4)$$

偏差に関する有用な情報として、偏差の二乗平均をとりこれを分散と呼び、ここでは  $\sigma^2$  と表す。

$$\sigma^2 \equiv \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (1.5)$$

また、分散の平方根を標準偏差と呼ぶ。

$$\sigma = \sqrt{\sigma^2} \quad (1.6)$$

分散の計算方法は以下の通り

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2 \quad (1.7)$$

標準偏差を用いて、データを以下のように変換することを考える。

$$z_i = \frac{x_i - \bar{x}}{\sigma_x} \quad (1.8)$$

新しくできたデータ系列  $z_i$  の平均と標準偏差は以下の通りです。

$$\bar{z} = \frac{1}{n} \sum_{i=1}^n z_i = \frac{1}{n\sigma_x} \sum_{i=1}^n (x_i - \bar{x}) = 0 \quad (1.9)$$

$$\sigma_z^2 = \frac{1}{n} \sum_{i=1}^n (z_i - \bar{z})^2 = \frac{1}{n} \sum_{i=1}^n z_i^2 = \frac{1}{n\sigma_x^2} \sum_{i=1}^n (x_i - \bar{x})^2 = 1 \quad (1.10)$$

よって、 $z_i$  は平均 0、標準偏差 1 のデータ系列になった。このような  $x$  から  $z$  の変換を標準化 (standardization) あるいは正規化 (normalization) と呼ぶ。

### 1.3.3 変動係数

$$C.V. \equiv \frac{\sigma_x}{\bar{x}} \quad (1.11)$$

を変動係数 (Coefficient of variation) と呼ぶ。

## 1.4 その他の指標

### 1.4.1 モーメント

一般に以下のような量を  $k$  次の平均周りのモーメントと呼ぶ

$$\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^k \quad (1.12)$$

2 次の平均周りのモーメントは分散である。3 次、4 次の平均周りのモーメントを用いて歪度 (skewness)  $S_k$ , 尖度 (flatness)  $F_l$  は以下のように定義される。

$$S_k \equiv \frac{1}{n\sigma^3} \sum_{i=1}^n (x_i - \bar{x})^3 \quad (1.13)$$

$$F_l \equiv \frac{1}{n\sigma^4} \sum_{i=1}^n (x_i - \bar{x})^4 \quad (1.14)$$

また、正規分布の尖度は 3 のため、上記の  $F_l - 3$  を尖度の定義とする場合もある。

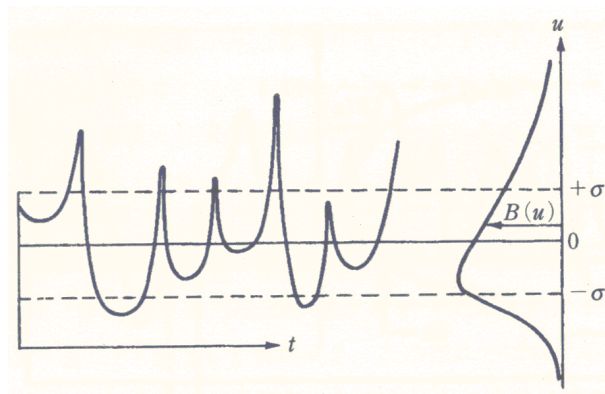


図 1.2: 歪度が正の変動量 (Tennekes and Lumley, 1972)

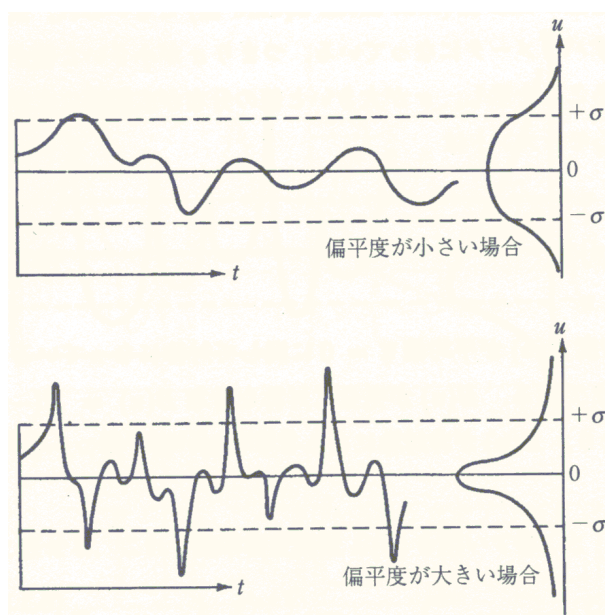


図 1.3: 尖度が大きいの変動量と小さい変動量 (Tennekes and Lumley, 1972)

## 演習

---

### 演習 1.1 基礎統計量の計算

---

サンプル数 80 以上のデータを用いて以下の作業を行うこと。データを持っていない場合は、ホームページ (<http://www.suiri.tsukuba.ac.jp/~asanuma/courses/>) に示すデータを用いてもよい。米国最高裁判所での裁判官の任官期間（単位は年）である (Devore, 1991)。

1. 対象データについて、表 1.1 と同じような度数分布表を作れ
2. 対象データについて、図 1.1 と同じようなヒストグラム、累積度数グラフ、累積相対度数グラフを作れ。
3. 対象データについて、平均値、メディアン、最頻値を求めよ
4. 対象データについて、分散、標準偏差、変動係数を求めよ。
5. 対象データについて、歪度、尖度をもとめよ。

## 第2章 確率変数と確率分布

### 2.1 離散確率変数

確率をもって変化数量を確率変数 (random variable) と呼ぶ．ここでは，確率変数を大文字の  $X$  で表すとする．

例 2.1

セブンイレブンでアルバイトをしている．レジにくる客のうちペットボトルの茶を購入するかどうかを考える．確率変数を

$$X = \begin{cases} 1 & \text{客がペットボトル茶を購入した場合} \\ 0 & \text{客がペットボトル茶を購入しない場合} \end{cases}$$

で定義する． $X$  は 0 または 1 の値を取る離散確率変数である．

例 2.2

春日 4 丁目で，戸別訪問のアンケート調査を行うよう教授に命じられた．一日に 5 件の家の呼び鈴を鳴らし，その中で応じてくれた家でアンケートを取るというものである．ここで考えられる確率変数は，

$$X = \text{1 日のアンケート回収数}$$

であり， $X$  は  $0 \leq X \leq 5$  の範囲を離散的に変化する離散確率変数である．

#### 2.1.1 確率分布

離散確率変数  $X$  がある値  $x$  を取る確率を  $p(x) = P(X = x)$  で表し，確率分布 (probability distribution) と呼ぶ．また，

$$F(x) = P(X \leq x) = \sum_{y: y \leq x} p(y) \quad (2.1)$$

は， $X$  が  $x$  以下の値を取る時  $p(x)$  の総和であり，累積確率分布 (cumulative distribution function) と呼ぶ．

#### 2.1.2 期待値・分散

離散確率変数  $X$  が集合  $D$  の値をとり，確率分布が  $p(x)$  で与えられるとき， $X$  の期待値 (Expected value) あるいは平均値 (mean value) は  $E(X)$  あるいは  $\mu_X$  で表され、

$$E(X) = \mu_X = \sum_{x \in D} xp(x) \quad (2.2)$$

で与えられる。 $E(\cdot)$  は演算子と考えられ、以下の性質を持つ。

$$E(aX + b) = aE(X) + b \quad \text{あるいは} \quad \mu_{aX+b} = a\mu_X + b \quad (2.3)$$

離散確率変数  $X$  が集合  $D$  の値をとり、確率分布が  $p(x)$ 、期待値が  $\mu$  で与えられるとき、 $x$  の分散  $V(X)$  あるいは  $\sigma_X^2$  で表され、

$$V(X) = \sigma_X^2 = \sum_{x \in D} (x - \mu)^2 p(x) = E[(X - \mu)^2] \quad (2.4)$$

で与えられる。 $X$  の標準偏差は、 $\sigma_X = \sqrt{\sigma_X^2}$  である。上式は以下のように書き換えられる。

$$V(X) = \sum_{x \in D} x^2 p(x) - \mu^2 = E(X^2) - \{E(X)\}^2 \quad (2.5)$$

また、 $V(\cdot)$  は以下の性質を持つ

$$V(aX + b) = a^2 V(X) \quad (2.6)$$

### 例 2.1 の場合

平均的に 10 人に 2 人の割合で、ペットボトル茶を購入するならば、

$$p(0) = P(X = 0) = 0.8$$

$$p(1) = P(X = 1) = 0.2$$

$$p(x) = 0 \quad \text{for } x \neq 0 \text{ or } 1$$

あるいは、

$$p(x) = P(X = x) =$$

と表される。グラフに表現すると、図 2.1 の通りとなる。また、期待値・分散はそれぞれ、

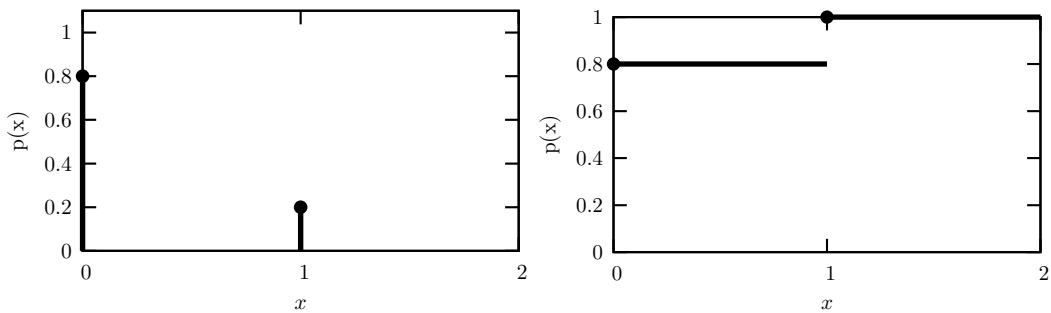


図 2.1: 例 2.1 の確率分布、累積確率分布の例

$$E(X) = 0 \cdot p(0) + 1 \cdot p(1) = 0.2 \quad (2.7)$$

$$V(X) = (0 - 0.2)^2 p(0) + (1 - 0.2)^2 p(1) = 0.04 \cdot 0.8 + 0.64 \cdot 0.2 = 0.16 \quad (2.8)$$

のようになる。

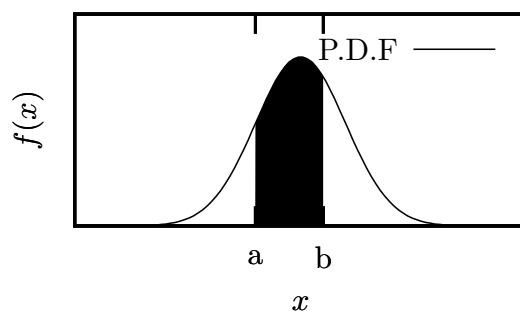


図 2.2: 確率密度関数。  $P(a \leq X \leq b)$  は黒く塗った部分の面積に相当する。

## 2.2 連続確率変数

確率変数が連続的に変化する場合、連続確率変数と呼ぶ。

### 例 2.3

某講師の部屋の壁には日本地図が張っており、地図にダーツを投げて当たった場所を次の調査地域にするのが某講師の研究スタイルである。ダーツの当たった地点の標高は、日本全国からランダムに抽出された地点と考えられる。

$X$  = 日本全国からランダムに抽出された地点の標高

は、連続的に変化するので連続確率変数であり、 $0 \leq X \leq 3778$  である。

### 2.2.1 確率密度関数

連続確率変数  $X$  の確率分布あるいは確率密度関数 (probability density, probability density function, p.d.f) を  $f(x)$  で表すと、 $f(x)$  は以下の式を満たす。

$$P(a \leq X \leq b) = \int_a^b f(x)dx \quad (2.9)$$

当然、 $f(x)$  は以下の式を満たす

$$f(x) \geq 0 \quad \text{すべての } x \text{ について} \quad (2.10)$$

$$\int_{-\infty}^{\infty} f(x)dx = 1 \quad (2.11)$$

連続確率変数  $X$  の累積確率密度関数 (cumulative density function, c.d.f.)  $F(x)$  は、以下のように定義される。

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(y)dy \quad (2.12)$$

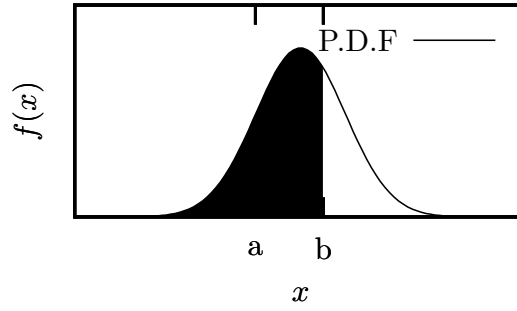


図 2.3: 累積確率密度関数。  $F(x) = P(X \leq b)$  は黒く塗った部分の面積に相当する。

### 2.2.2 期待値・分散

確率密度関数  $f(x)$  を持つ連続確率変数  $X$  の期待値あるいは平均値は

$$E(X) = \mu_X = \int_{-\infty}^{\infty} x f(x) dx \quad (2.13)$$

で表される。また、分散は、

$$V(X) = \sigma_X^2 = \int_{-\infty}^{\infty} (x - \mu_X)^2 f(x) dx = E[(X - \mu_X)^2] \quad (2.14)$$

である。 $X$  の標準偏差は、 $\sigma_X = \sqrt{\sigma_X^2}$  である。離散確率変数の時と同様に以下の式が成り立つ。

$$V(X) = E(X^2) - \{E(X)\}^2 \quad (2.15)$$

また、同様に 3 次、4 次のモーメントも定義され、歪度  $S_X$ 、尖度  $F_X$  は以下のように定義される。

$$S_X = \frac{1}{\sigma^3} \int_{-\infty}^{\infty} (x - \mu_X)^3 f(x) dx \quad (2.16)$$

$$F_X = \frac{1}{\sigma^4} \int_{-\infty}^{\infty} (x - \mu_X)^4 f(x) dx \quad (2.17)$$

## 2.3 正規分布

平均  $\mu$ 、標準偏差  $\sigma$  をもつ連続確率変数  $X$  の確率密度関数が以下の式で表されるときこれを、 $X$  は正規分布 (normal distribution)  $N(\mu, \sigma^2)$  とよぶ。

$$f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x - \mu)^2}{2\sigma^2}\right\} \quad -\infty < x < \infty \quad (2.18)$$

このとき、 $X \sim N(\mu, \sigma^2)$  と表される。特に  $\mu = 0, \sigma = 1$  のとき、 $N(0, 1)$  を標準正規分布 (standard normal distribution) と呼ぶ。また標準正規分布に従う確率変数を特に、 $Z$  で表す。 $Z$  の確率密度関数は、以下のように表される。

$$f(z; 0, 1) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{z^2}{2}\right\} \quad (2.19)$$

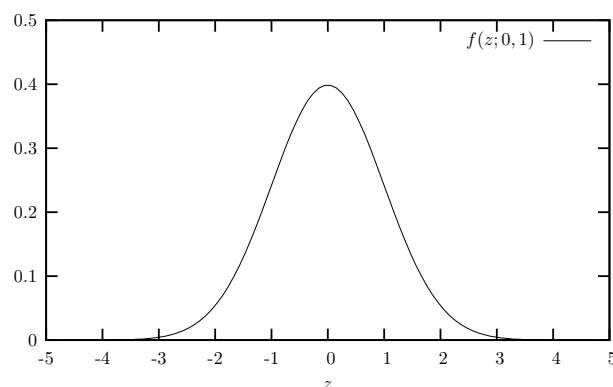


図 2.4: 標準正規分布

$X$  が正規分布に従うとき、 $Z = (X - \mu)/\sigma$  は標準正規分布に従う。すなわち

$$X \sim N(\mu, \sigma^2) \Rightarrow Z = \frac{(X - \mu)}{\sigma} \sim N(0, 1) \quad (2.20)$$

である。 $N(0, 1)$  の累積確率密度関数を  $\Phi(z)$  で表す。すなわち、

$$\Phi(z) = P(Z \leq z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{y^2}{2}\right\} dy \quad (2.21)$$

である。

正規分布以外にも、確率密度関数の例としては、 $t$  分布 (student-t)、 $\chi^2$  (カイ二乗) 分布などがある。

## 演習

---

### 演習 2.1

---

1. 例 2.2 において、各戸がアンケートに応じてくれる確率を  $\lambda$  として、確率分布、累積確率分布、期待値、分散を計算せよ。
2.  $\lambda = 0.25, 0.5, 0.75$  について、確率分布、累積確率分布を図示せよ。
3. 一日に訪問する戸数を増やすとどうなるか。

---

### 演習 2.2

---

1. 式 2.3 を式 2.2 から導け。
2. 式 2.5 を式 2.2、2.3、2.4 から導け。

---

### 演習 2.3 ランダムウォーク

---

最初に  $x = 0$  の位置にいるとして、サイコロを振り、奇数の目が出た時に右へ一歩、偶数の目が出たときに左へ一歩移動するとする。一歩の移動量は  $\Delta x$  で一定であるとする。サイコロを  $n$  回振ったあとの位置を確率変数  $X$  としたときに、その密度関数  $p(x) = P(X = x)$  を求めよ。

$n = 10, 100, 1000$  の時の  $p(x)$  をグラフにプロットし,  $n$  が大きくなると,  $p(x)$  が正規分布に近づくことを確認せよ.

---

## 演習 2.4

---

正規分布について, 以下の設問に答えよ. 必要に応じて  $\Phi(Z)$  の数表<sup>1</sup> を使用せよ.

1. 正規分布において,  $\mu - 3\sigma \leq X \leq \mu + 3\sigma$ ,  $\mu - 2\sigma \leq X \leq \mu + 2\sigma$ ,  $\mu - \sigma \leq X \leq \mu + \sigma$  となる確率を求めよ.
2. 任意に選ばれたある種のダイオードの破壊電圧は正規分布し,  $\mu = 40, \sigma = 1.5$  ボルトである. あるダイオードの破壊電圧が 39 から 42 ボルトの間にある確率は, どれくらいか? (Devore, 1991)
3. ある機械が 1 時間に使用する蒸留水の量は正規分布し,  $\mu = 64$  リットル,  $\sigma = 0.78$  リットルである. 蒸留水の供給が足りなくなるのが 0.5% のみであるような, 供給量  $C$  を求めよ. (Devore, 1991, を改編).

---

<sup>1</sup><http://www.suiri.tsukuba.ac.jp/~asanuma/courses/resources/norm.html>

## 第3章 母集団と標本、点推定、区間推定

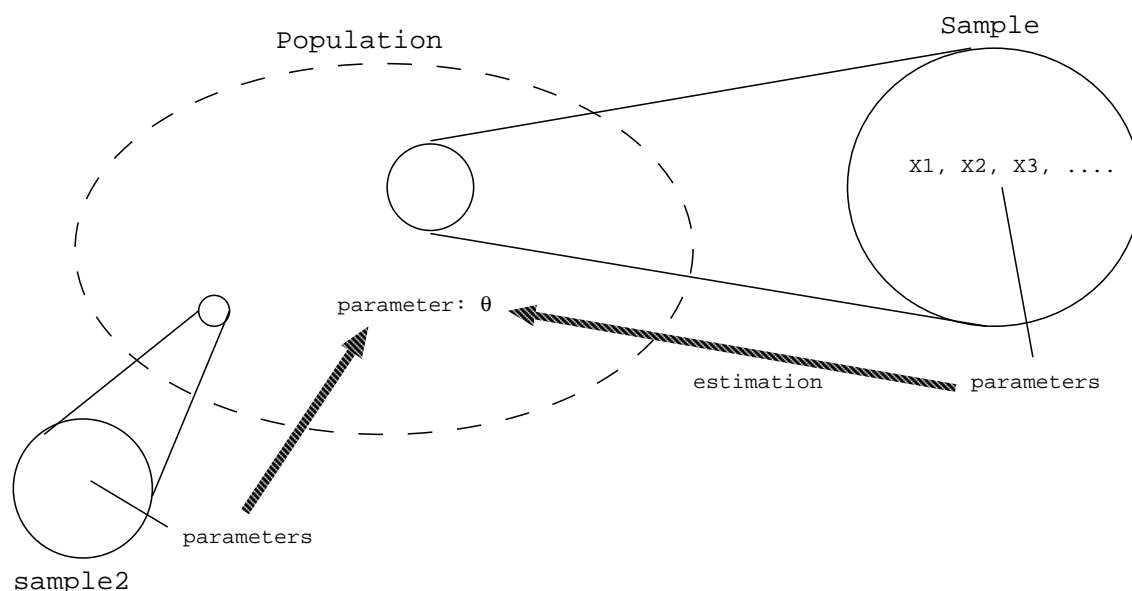


図 3.1: 母集団，サンプルと統計的推定の概念

### 3.1 概念

統計的推定とは，対象となる母集団 (population) から実験・観測・調査 (サンプリング) などを通じてサンプル (標本、sample) を抽出し，サンプルの性質から母集団の性質を推定 (estimate) する作業である．地球環境科学の多くのケースでは，母集団の数も特定できない場合が多く，母集団全てのサンプルは不可能である．サンプル統計値から母集団の性質を推定することが多い．実験・観測・調査は，複数回行うことも可能であり，当然ながら多くの回数実施すれば，推定精度は向上すると考えられる (ことが多い) ．

### 3.2 点推定と不偏推定量

母集団から  $n$  個のサンプル  $X_1, X_2, X_3, \dots, X_n$  を抽出する．このサンプルから母集団の性質を表すパラメータ  $\theta$  を推定することを点推定 (point estimate) と呼ぶ．このとき， $\theta$  の推定量 (estimator)

を  $\hat{\theta}$  で表す．例えば，母集団平均  $\mu = E(X)$  の推定量  $\hat{\mu}$  の推定量の例として，

$$\hat{\mu} = \bar{X} = \frac{\sum X_i}{n} \quad (3.1)$$

$$\hat{\mu} = \tilde{X} \quad (3.2)$$

$$\hat{\mu} = \frac{\max(X_i) + \min(X_i)}{2} \quad (3.3)$$

などが考えられる．推定量の中で最も確からしいものとして，不偏推定量 (unbiased estimator) は以下のような性質を持つ推定量である．

$$E(\hat{\theta}) = \theta \quad (3.4)$$

ここで  $E(\cdot)$  は期待値であり，ここでは複数のサンプリングの平均を意味する．また， $E(\hat{\theta}) - \theta$  をバイアス (bias) と呼ぶ．不偏推定量は複数存在することもある．

サンプル平均は，母集団平均の不偏推定量である．

$$E(\bar{X}) = \mu \quad (3.5)$$

また，母集団の確率分布が対象であるときは，メディアン  $\tilde{X}$  も母集団平均の不偏推定量である．  
サンプル分散  $S^2$

$$\hat{\sigma}^2 = S^2 = \frac{1}{n-1} \sum (X_i - \bar{X})^2 \quad (3.6)$$

は，母集団分散  $\sigma^2$  の不偏推定量である．

証明.

$$\begin{aligned} E(S^2) &= \frac{1}{n-1} E \left( \sum (X_i - \bar{X})^2 \right) \\ &= \frac{1}{n-1} E \left[ \sum X_i^2 - \frac{1}{n} \left( \sum X_i \right)^2 \right] \\ &= \frac{1}{n-1} \left\{ \sum E(X_i^2) - \frac{1}{n} E \left[ \left( \sum X_i \right)^2 \right] \right\} \end{aligned}$$

2.5 を使用すると、 $E(X^2) = V(X) + \{E(X)\}^2 = \sigma^2 + \mu^2$  であるので、

$$\begin{aligned} E(S^2) &= \frac{1}{n-1} \left( \sum (\sigma^2 + \mu^2) - \frac{1}{n} \left[ V \left( \sum X_i \right) + \left\{ E \left( \sum X_i \right) \right\}^2 \right] \right) \\ &= \frac{1}{n-1} \left[ n\sigma^2 + n\mu^2 - \frac{1}{n} \{n\sigma^2 + (n\mu)^2\} \right] \\ &= \frac{1}{n-1} [n\sigma^2 - \sigma^2] = \sigma^2 \end{aligned}$$

□

また、 $n-1$  の代わりに  $n$  を用いた

$$S_p^2 = \frac{1}{n} \sum (X_i - \bar{X})^2 \quad (3.7)$$

は、

$$E(S_p^2) = \frac{n-1}{n} \sigma^2 \quad (3.8)$$

となり、母分散を過小評価する。特に  $n$  が小さいときに要注意である。

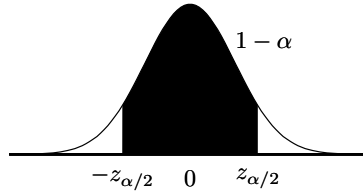


図 3.2:  $z_{\alpha/2}$  の定義。黒く塗った部分の面積が  $1 - \alpha$  になる。

### 3.3 区間推定

点推定は母集団パラメータの値を推定するが、区間推定 (Interval estimation) は、母集団パラメータの確率的な範囲を推定する。すなわち、母集団のパラメータ  $\theta$  が

$$P(L \leq \theta \leq U) = 1 - \alpha \quad (3.9)$$

となるような、 $L, U$  を求めることである。 $L \leq \theta \leq U$  を  $100(1 - \alpha)$  信頼区間 (Confidence interval)、 $100(1 - \alpha)$  を信頼度 (Confidence level) と呼ぶ。 $1 - \alpha$  には、0.99, 0.95 すなわち、99%, 95% などを用いる。

#### 3.3.1 正規分布母集団の平均の区間推定

—— 中心極限定理より ——

平均  $\mu$ 、分散  $\sigma^2$  の任意の確率分布を持つ母集団から、無作為に抽出されたサンプル  $X_1, X_2, \dots, X_n$  の平均値  $\bar{X}$  は、サンプル数  $n$  が十分大きいとき、期待値  $\mu$ 、標準偏差  $\sigma/\sqrt{n}$  を持つ正規分布  $N(\mu, \sigma^2/n)$  にほぼ従う。

分散が既知の時

中心極限定理から、正規分布  $N(\mu, \sigma^2)$  に従う母集団から抽出したサンプルの平均  $\bar{X}$  は、正規分布  $N(\mu, \sigma^2/n)$  に従う。よって、

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \quad (3.10)$$

は、標準正規分布に従う。標準正規分布では、例えば、

$$P(-1.96 < Z < 1.96) = 0.95 \quad (3.11)$$

なので、

$$P(-1.96 < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < 1.96) = 0.95 \quad (3.12)$$

すなわち、

$$P(\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}}) = 0.95 \quad (3.13)$$

表 3.1: 主な信頼度に対する  $z_{\alpha/2}$  の値

信頼度 $100(1 - \alpha)$	$\alpha$	$\alpha/2$	$z_{\alpha/2}$
90%	0.10	0.050	1.645
95%	0.05	0.025	1.960
99%	0.01	0.005	2.576

となり、 $\mu$  の 95% 信頼区間は

$$\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}} \quad (3.14)$$

であるということができる。

一般的に、 $z_{\alpha/2}$  を図 3.2 のように定義するならば、信頼区間は以下の通りとなる。

正規分布母集団の平均の区間推定（分散が既知のとき）

分散  $\sigma$  が既知であり、また、観測値が  $X_1 = x_1, X_2 = x_2, \dots$  のように得られたとする。このとき、 $\bar{x}$  を式 3.14 の  $\bar{X}$  に代入すると正規分布母集団の平均  $\mu$  の  $100(1 - \alpha)\%$  の信頼区間は、

$$\left( \bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right) \quad (3.15)$$

である。

また、代表的な信頼度について、 $z_{\alpha/2}$  の値を表 3.1 に挙げた。

分散が未知の時

母集団の分散  $\sigma^2$  が未知の時、これをサンプル分散  $S^2 = \sum (X_i - \bar{X})^2 / (n - 1)$  で代用する。このとき、サンプル数  $n$  が十分に大きいとき、

$$Z = \frac{\bar{X} - \mu}{S/\sqrt{n}} \quad (3.16)$$

は、近似的に標準正規分布に従う。一方、 $n$  があまり大きくないときは、同じ値は  $t$  分布に従う。

$t$  分布

平均値  $\mu$  をもつ正規分布母集団から得られたサンプル数  $n$  のサンプルの平均値が  $\bar{x}$  のとき、

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \quad (3.17)$$

は自由度  $n - 1$  の  $t$  分布 (Student  $t$ ) に従う。

$t$  分布は、一つのパラメータ（自由度）に依存する確率分布である。平均 0 で左右対称であり、標準正規分布よりも広がり大きい。自由度  $\nu$  の  $t$  分布曲線  $t_\nu$  は、 $\nu$  が大きいほど標準正規分布に近くなる。正規分布と同様に  $t_{\alpha, \nu}$  を図 3.4 に示すように定義すると、

$$P(-t_{\alpha/2, \nu} \leq T \leq t_{\alpha/2, \nu}) = 1 - \alpha \quad (3.18)$$

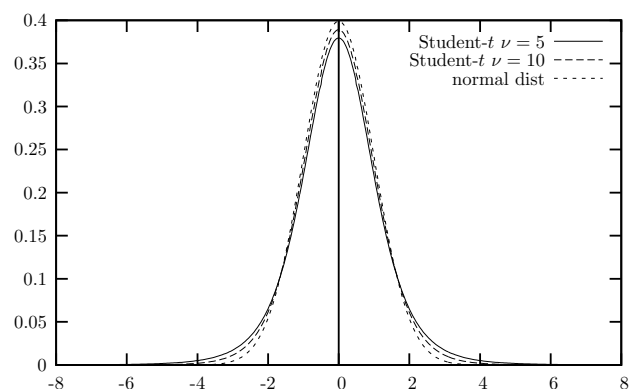


図 3.3:  $t$  分布と正規分布

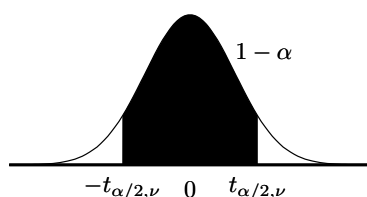


図 3.4:  $t_{\alpha/2, \nu}$  の定義。黒く塗った部分の面積が  $1 - \alpha$  になる。

である。この  $t$  分布を用いると以下のような区間推定ができる。

正規分布母集団の平均の区間推定（分散が未知のとき）――

正規分布から得たサンプルの平均が  $\bar{x}$ 、サンプル分散が  $s^2$  の時、母集団の平均  $\mu$  の  $100(1 - \alpha)\%$  の信頼区間は、

$$\left( \bar{x} - t_{\alpha/2, n-1} \frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2, n-1} \frac{s}{\sqrt{n}} \right) \quad (3.19)$$

### 3.3.2 正規分布母集団の分散の区間推定

$\chi^2$  分布――

平均値  $\mu$ 、分散  $\sigma^2$  を持つ正規母集団  $N(\mu, \sigma^2)$  から得られたサンプルの分散を  $s^2$  とすると、

$$\frac{(n-1)S^2}{\sigma^2} = \frac{\sum (X_i - \bar{X})^2}{\sigma^2} \quad (3.20)$$

は、自由度  $n - 1$  の  $\chi^2$  分布に従う。

$\chi^2$  は一つのパラメータ（自由度）によって決定される確率分布であり、左右非対称である。 $x > 0$  についてのみ確率密度関数が正であり、 $n$  が大きくなると全体的に右にシフトし、かつ対象に近くなる。 $\chi_{\alpha, \nu}^2$  をこれまでと同様に定義すると、

$$P \left( \chi_{1-\alpha/2, \nu}^2 \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi_{\alpha/2, \nu}^2 \right) = 1 - \alpha \quad (3.21)$$

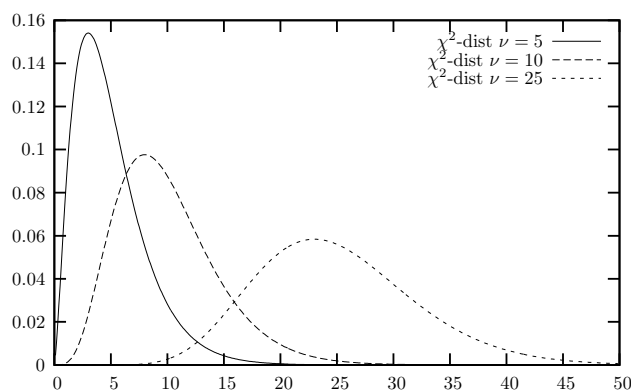


図 3.5:  $\chi^2$  分布

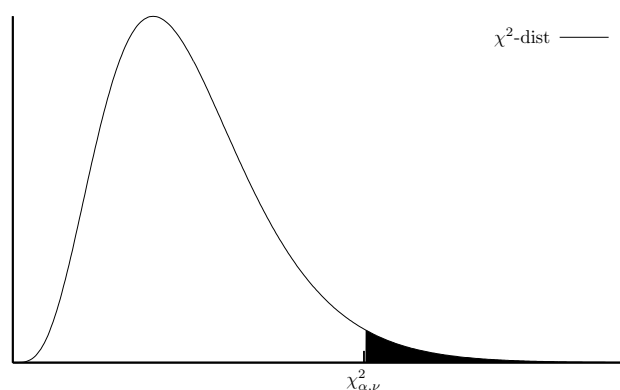


図 3.6:  $\chi^2_{\alpha, \nu}$  の定義。黒く塗った部分の面積が  $\alpha$  になる。

である。これから以下のような区間推定が可能である。

正規分布母集団の分散の区間推定

平均値  $\mu$ 、分散  $\sigma^2$  を持つ正規母集団  $N(\mu, \sigma^2)$  から得られたサンプルの分散を  $s^2$  とすると、母分散  $\sigma^2$  の  $100(1 - \alpha)\%$  の信頼区間は、

$$\left( \frac{(n-1)s^2}{\chi^2_{\alpha/2, n-1}}, \frac{(n-1)s^2}{\chi^2_{1-\alpha/2, n-1}} \right) \quad (3.22)$$

である。

## 演習

### 演習 3.1 母集団統計量とサンプル統計量

宿題 1.1 で用いたデータを母集団と見て、以下の統計シミュレーションを行え。

1. 母集団から、 $n=10$  個のサンプルを任意に抜き出し、新しいデータセットを作る
2. 新しいサンプルデータセットについて、平均と分散を計算する。

3. 上記 1,2 を 10 回繰り返し,  $n=10$  のサンプルから得られた平均, 分散と元の母集団の平均分散が, 一致するかどうか, 比較せよ.

---

### 演習 3.2

---

人間工学の技術者が、労働者が操作する機械の入力装置の高さを、生産性をもっとも大きくなるように設定しようとしている。 $n = 31$  人の被験者に対してそれぞれの最適な入力装置の高さを計測したところ、平均値は  $\bar{X} = 80$  cm であった。最適な入力装置の高さは、 $\sigma = 2.0$  cm で分布していると仮定して、すべての労働者に対する最適な高さの平均値の 95%信頼区間を求めよ。99.5%信頼区間は、どうなるか (Devore, 1991)。

---

### 演習 3.3

---

以下の値を求めよ。必要に応じて  $t$  分布表<sup>1</sup> を用いよ。

1.  $t_{0.1,15}$
2.  $t_{0.05,15}$
3.  $t_{0.05,25}$
4.  $t_{0.05,40}$
5.  $t_{0.005,40}$

---

### 演習 3.4

---

ある地点でのサンプリングから土壌水分（体積含水率）が 10.4, 8.1, 9.5, 8.9, 10.7 % と計測された。この地点の平均土壌水分量を 95%の信頼度での信頼区間を求めよ。また、99.5%の信頼区間を求めよ。

---

<sup>1</sup><http://www.suiri.tsukuba.ac.jp/~asanuma/courses/resources/tdist.html>

## 第4章 仮説検定

仮説検定 (hypothesis testing) の目的は、母集団について仮定された命題を、標本に基づいて検証することである (東京大学教養学部統計学教室, 1991) . 仮説から予測される結果と実際の観測結果の差が、偶然による (統計的に有意でない) ものであるか、意味がある (統計的に有意 (significant) である) ものであるかを、確率の基準で評価する . ここでたてられた仮説を統計的仮説 (statistical hypothesis), または仮説と呼ぶ . したがって、仮説検定とは、統計的仮説の有意性の検定である .

### 例 7.1

XX 商店街でやっている YY 大福引き大会で、一等賞を取った人の話を聞いたことが無い . よって、実は一等賞はないに違いない .

どの程度の確率までを有意と考えるかを有意水準 (significant level) と呼び、通常  $\alpha$  で表す .  $\alpha = 0.1$  とは、この確率で起こることは統計的に意味があるが、この確率以下の事象は統計的に有意でない偶然の事象、と考えることである .

仮説検定は、以下の順序を経て行う .

1. 検定の対象となる統計的仮説を立てる . これを帰無仮説 (null hypothesis) と呼び、これを  $H_0$  と表す .
2. 帰無仮説と対立する対立仮説 (alternative hypothesis) をたてる . これを  $H_a$  と表す .
3. 検定に用いる統計量 (検定統計量, test statistic) を選ぶ .
4. 検定に必要な有意水準を決定する .
5. 統計検定量に対して、帰無仮説を棄却すべき範囲 (棄却域, rejection region) を求める . 棄却域とは反対に、帰無仮説を採択すべき範囲を採択域 (acceptance region) と呼び .
6. サンプルの情報から統計検定量を計算する .
7.  $H_0$  が棄却されるべきかを決定する .

### 例 7.2

AさんとBさんがじゃんけんを10回して、Aさんが6回勝った。よって、AさんがBさんよりもじゃんけんが強い。

この場合、AさんがBさんに勝つ確率を $\lambda$ と置くと、帰無仮説は

$$H_0 : \lambda = 0.5$$

であり、対立仮説は

$$H_a : \lambda > 0.5$$

である。

## 4.1 正規母集団の平均に対する仮説検定

### 4.1.1 母分散 $\sigma$ が既知の時

母集団から得られたサンプル $X_1, X_2, \dots, X_n$ を用いて、もとの母集団の平均 $\mu$ がある値( $\mu_0$ )であるかどうかを検定するとしよう。

この場合、帰無仮説は、

$$H_0 : \mu = \mu_0 \quad (4.1)$$

であり、対立仮説は、

$$H_a : \mu \neq \mu_0 \quad (4.2)$$

である。前章の中心極限定理を使用すれば、 $\bar{X}$ は平均 $E(\bar{X}) = \mu$ 、分散 $V(\bar{X}) = \sigma^2/n$ の正規分布に従う。すなわち、

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad (4.3)$$

あるいは、

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1) \quad (4.4)$$

である。もしも $H_0$ が正しければ( $\mu = \mu_0$ ならば)、

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1) \quad (4.5)$$

となる。 $Z$ は観測値平均 $\bar{X}$ と、仮定されている母集団平均値 $\mu_0$ の差が標準偏差で無次元化されたものであると考えられる。 $Z$ の値が大きければ、観測値と仮定されている値との間に大きな差があることになる。すなわち、ある値よりも $Z$ が大きければ仮説が棄却されることになる。 $\alpha$ を有意水準とすれば棄却域は、以下の通りである。

$$z \geq z_{\alpha/2} \quad \text{あるいは} \quad z \leq -z_{\alpha/2} \quad (4.6)$$

これを両側検定 (two-tailed test) と呼ぶ。

一方、 $\mu$  がある数よりも大きいかどうかを調べるときは、同じように帰無仮説を

$$H_0 : \mu = \mu_0 \quad (4.7)$$

とし、対立仮説を

$$H_a : \mu > \mu_0 \quad (4.8)$$

とする。棄却域は、

$$z \geq z_\alpha \quad (4.9)$$

となる。これを右片側検定 (upper-tailed test) と呼ぶ。

同様に  $\mu$  がある値  $\mu_0$  よりも小さいかどうかを調べるときは、

$$H_a : \mu < \mu_0 \quad (4.10)$$

とする。棄却域は、

$$z \leq -z_\alpha \quad (4.11)$$

となる。これ左片側検定 (lower-tailed test) と呼ぶ。

このように、対象となる変数がある値と同じかどうかを検定するには両側検定を行うが、大小の判定には片側検定 (one-tailed test) を用いる。

以上をまとめると以下の通りである。

母集団平均に関する検定 ( $\sigma$  が既知の時)

帰無仮説	$H_0 : \mu = \mu_0$
検定統計量	$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$
対立仮説	棄却域
$H_a : \mu > \mu_0$	$z \geq z_\alpha$
$H_a : \mu < \mu_0$	$z \leq -z_\alpha$
$H_a : \mu \neq \mu_0$	$z \geq z_{\alpha/2}$ あるいは $z \leq -z_{\alpha/2}$

このような検定を  $Z$  検定 ( $Z$ -test) と呼ぶ。分散が既知の時およびサンプルが多いとき ( $n \gtrsim 30$ ) に用いることができる。

### 例 7.3

空調システムの作動状況を調べるために、設定温度を 25 度とし、7 日間にわたって室内温度を測定したところ、次のような結果を得た。

24.2, 25.3, 26.2, 25.7, 24.4, 25.1, 25.6

母集団の標準偏差が  $\sigma = 0.7$  と既知の時、このシステムが正しく働いているかを有意水準 5% で検定しよう。

これは、 $\mu = 25.0$  という仮定が正しいかを検定することになる。

帰無仮説  $H_0 : \mu = 25$

$$\text{検定統計量 } z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = 0.81$$

$z_{\alpha/2} = 1.96$  であるので、

棄却域  $z \geq 1.96$  あるいは  $z \leq -1.96$

となり、 $H_0$  は棄却されず、採択された。

### 4.1.2 母分散が未知の時

母分散が未知の時、あるいは  $n$  が小さいときには、 $t$  分布に基づく  $t$  検定 ( $t$ -test) を行うことになる。帰無仮説  $H_0 : \mu = \mu_0$  と対立仮説  $H_a : \mu \neq \mu_0$  に関する検定を考える。

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \quad (4.12)$$

は自由度  $n - 1$  の  $t$  分布に従うので、検定統計量  $t = (\bar{X} - \mu_0)/(S/\sqrt{n})$  を用いると、棄却領域が  $t \geq t_{\alpha/2}$  あるいは  $t \leq -t_{\alpha/2}$  となる。他の対立仮説の時も含めてまとめると以下の通りとなる。

母集団平均に関する検定 ( $\sigma$  が既知の時)

帰無仮説  $H_0 : \mu = \mu_0$

$$\text{検定統計量 } t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

対立仮説 棄却域

$$H_a : \mu > \mu_0 \quad t \geq t_{\alpha, n-1}$$

$$H_a : \mu < \mu_0 \quad t \leq -t_{\alpha, n-1}$$

$$H_a : \mu \neq \mu_0 \quad t \geq t_{\alpha/2, n-1} \quad \text{あるいは} \quad t \leq -t_{\alpha/2, n-1}$$

## 演習

---

### 演習 4.1 母集団の平均に関する $Z$ 検定

---

ある自動車会社は、新車購入後の最初の点検を走行距離 3000km で行うように推薦している。実際に点検に持ち込まれた車の 50 台を調べたところ、平均 3208km であった。標準偏差が 273km であるとする。このとき、点検に持ち込まれる車の平均走行距離は、3000km であるといえるか、有意水準 0.01 で検定せよ (Devore, 1991, より改編)。

---

### 演習 4.2 母集団の平均に関する $Z$ 検定

---

ある建設現場で、ある種のブロックの強度が  $3200 \text{ kg/m}^2$  未満で無ければ、このブロックを利用することにした。無作為抽出された 36 個のサンプルが強度試験に掛けられ、その平均、標準偏差は  $3109 \text{ kg/m}^2$ ,  $156 \text{ kg/m}^2$  であった。

1. 必要な仮説を述べ、有意水準 0.05 で検定せよ。
2. もしも、ブロックの強度が  $3200 \text{ kg/m}^2$  を越える時にこのブロックを使用するとした場合はどうか。(Devore, 1991, より改編)。

---

### 演習 4.3 母集団の平均に関する $t$ 検定

---

ある自動車メーカーが新しく開発したコンパクトタイプ車の燃費の向上を試験するため、一般のドライバーにこの車を運転してもらったところ、結果は、27.2, 29.3, 31.5, 28.7, 30.2, 29.6 マイル/ガロンであった。この自動車メーカーは、この車の燃費が最低 30 マイル/ガロンであることを宣伝したいと考えているが、この実験結果はこれと矛盾しないかどうか、有意水準 0.05 で検定せよ。 $\mu$  を真の燃費とすると、帰無仮説は  $H_0: \mu = 30$ 、対立仮説は  $H_a: \mu < 30$  である (Devore, 1991, より改編)。

## 第5章 2つの母集団の問題

これまで1つの母集団から得られた標本についての統計的推定、仮説検定を考えてきた。この章では、2つの異なる母集団から得られた2つの標本について考える。これは例えば、1つの実験材料に対して、2つの異なる処理を行った場合などを想定している。

サンプル  $X_1, \dots, X_m$  が正規分布  $N(\mu_x, \sigma_x^2)$  に従う母集団からのサンプルで、サンプル  $Y_1, \dots, Y_n$  が正規分布  $N(\mu_y, \sigma_y^2)$  からのサンプルであり、両者がお互いに独立である場合を考える。

### 5.1 平均の差

このとき、

$$E(\bar{X} - \bar{Y}) = E(\bar{X}) - E(\bar{Y}) = \mu_x - \mu_y \quad (5.1)$$

$$V(\bar{X} - \bar{Y}) = V(\bar{X}) + V(\bar{Y}) = \frac{\sigma_x^2}{m} + \frac{\sigma_y^2}{n} \quad (5.2)$$

であるので、 $\bar{X} - \bar{Y}$  は、母平均の差の不偏推定値である。

#### 5.1.1 母分散が既知の時

それぞれの母分散  $\sigma_x, \sigma_y$  が既知の時、

$$\bar{X} \sim N(\mu_x, \sigma_x^2/m) \quad (5.3)$$

$$\bar{Y} \sim N(\mu_y, \sigma_y^2/n) \quad (5.4)$$

である。よって、

$$\bar{X} - \bar{Y} \sim N\left(\mu_x - \mu_y, \frac{\sigma_x^2}{m} + \frac{\sigma_y^2}{n}\right) \quad (5.5)$$

である。

以下が成り立つ。

正規母集団の標本平均の差の標本分布 (分散が既知の時)

2つの正規母集団  $N(\mu_x, \sigma_x^2), N(\mu_y, \sigma_y^2)$  からの独立な標本平均  $\bar{X}, \bar{Y}$  の差  $\bar{X} - \bar{Y}$  に関して、

$$Z = \frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{m} + \frac{\sigma_y^2}{n}}} \quad (5.6)$$

は、標準正規分布  $N(0, 1)$  に従う。

よって、区間推定は以下の通り。

正規母集団の母平均の差の区間推定 (分散が既知の時)

2つの正規母集団  $N(\mu_x, \sigma_x^2), N(\mu_y, \sigma_y^2)$  の母平均の差  $\mu_x - \mu_y$  の  $100(1 - \alpha)\%$  区間推定は、

$$\left( \bar{x} - \bar{y} - z_{\alpha/2} \sqrt{\frac{\sigma_x^2}{m} + \frac{\sigma_y^2}{n}}, \bar{x} - \bar{y} + z_{\alpha/2} \sqrt{\frac{\sigma_x^2}{m} + \frac{\sigma_y^2}{n}} \right) \quad (5.7)$$

である。

また、この場合の仮説検定は以下の通りとなる。

正規母集団の母平均の差に関する仮説検定 (分散が既知の時)

$$\text{帰無仮説 } H_0 : \mu_x - \mu_y = \Delta_0$$

$$\text{推定統計量 } z = \frac{\bar{x} - \bar{y} - \Delta_0}{\sqrt{\frac{\sigma_x^2}{m} + \frac{\sigma_y^2}{n}}}$$

$$\text{対立仮説 } \text{棄却域}$$

$$H_a : \mu_x - \mu_y > \Delta_0 \quad z \geq z_{\alpha}$$

$$H_a : \mu_x - \mu_y < \Delta_0 \quad z \leq -z_{\alpha}$$

$$H_a : \mu_x - \mu_y \neq \Delta_0 \quad z \geq z_{\alpha/2} \quad \text{あるいは} \quad z \leq -z_{\alpha/2}$$

### 5.1.2 母分散が未知であるが等しいとき

二つの正規分布母集団の母分散が等しいとき ( $\sigma_x = \sigma_y = \sigma$ )、

$$V(\bar{X} - \bar{Y}) = \sigma^2 \left( \frac{1}{m} + \frac{1}{n} \right) \quad (5.8)$$

となる。

この二つの母集団それぞれからの標本を用いた母分散の推定は、二つの標本をあわせた合併推定値 (pooled estimator) で行う。すなわち、

$$s_p^2 = \frac{m-1}{m+n-2} s_x^2 + \frac{n-1}{m+n-2} s_y^2 \quad (5.9)$$

は、 $\sigma$  の点推定値である。よって、標準化して、

$$Z = \frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{\sigma \sqrt{\frac{1}{m} + \frac{1}{n}}} \quad (5.10)$$

は、 $N(0, 1)$  に従う。よって、 $\sigma$  を  $s_p$  で入れ替えた量は  $t$  分布に従う。

正規母集団の標本平均の差の標本分布 (分散が等しいとき)

2つの正規母集団  $N(\mu_x, \sigma^2), N(\mu_y, \sigma^2)$  からの独立な標本平均  $\bar{X}, \bar{Y}$  の差  $\bar{X} - \bar{Y}$  に関して、

$$T = \frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{s_p \sqrt{\frac{1}{m} + \frac{1}{n}}} \quad (5.11)$$

は、自由度  $m + n - 2$  の  $t$  分布に従う。

よって、区間推定は以下の通り。

正規母集団の母平均の差の区間推定 (分散が等しいとき)

2つの正規母集団  $N(\mu_x, \sigma^2), N(\mu_y, \sigma^2)$  の母平均の差  $\mu_x - \mu_y$  の  $100(1 - \alpha)\%$  区間推定は、

$$\left( \bar{x} - \bar{y} - t_{\alpha/2, m+n-2} \cdot s_p \sqrt{\frac{1}{m} + \frac{1}{n}}, \bar{x} - \bar{y} + t_{\alpha/2, m+n-2} \cdot s_p \sqrt{\frac{1}{m} + \frac{1}{n}} \right) \quad (5.12)$$

である。

また、この場合の仮説検定は以下の通りとなる。

正規母集団の母平均の差に関する仮説検定 (分散が等しい時)

帰無仮説	$H_0 : \mu_x - \mu_y = \Delta_0$
推定統計量	$t = \frac{\bar{x} - \bar{y} - \Delta_0}{\sqrt{\frac{1}{m} + \frac{1}{n}}}$
対立仮説	棄却域
$H_a : \mu_x - \mu_y > \Delta_0$	$t \geq t_{\alpha, m+n-2}$
$H_a : \mu_x - \mu_y < \Delta_0$	$t \leq -t_{\alpha, m+n-2}$
$H_a : \mu_x - \mu_y \neq \Delta_0$	$t \geq t_{\alpha/2, m+n-2}$ あるいは $t \leq -t_{\alpha/2, m+n-2}$

### 5.1.3 母分散が未知であるが等しくないとき

それぞれの母分散  $\sigma_x, \sigma_y$  が未知でかつ等しく無いときは、上記の例のような標本分布の特定、区間推定、仮説検定は厳密には不可能である。

近似的に以下の方法が可能である。

正規母集団の標本平均の差の標本分布 (分散が未知の時)

2つの正規母集団  $N(\mu_x, \sigma_x^2), N(\mu_y, \sigma_y^2)$  からの独立な標本平均  $\bar{X}, \bar{Y}$  の差  $\bar{X} - \bar{Y}$  に関して、

$$T' = \frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{\sqrt{\frac{s_x^2}{m} + \frac{s_y^2}{n}}} \quad (5.13)$$

は、自由度

$$\nu = \frac{\left(\frac{s_x^2}{m} + \frac{s_y^2}{n}\right)^2}{\frac{(s_x^2/m)^2}{m-1} + \frac{(s_y^2/n)^2}{n-1}} \quad (5.14)$$

の  $t$  分布に近似的に従う。ただし、 $\nu$  が整数でないときはもっとも近い整数を用いる。

また、この場合の仮説検定は以下の通りとなる。

正規母集団の母平均の差に関する仮説検定 (分散が未知の時)

帰無仮説  $H_0: \mu_x - \mu_y = \Delta_0$

推定統計量  $t' = \frac{\bar{x} - \bar{y} - \Delta_0}{\sqrt{\frac{s_x^2}{m} + \frac{s_y^2}{n}}}$

対立仮説 近似的な棄却域

$H_a: \mu_x - \mu_y > \Delta_0 \quad t' \geq t_{\alpha, \nu}$

$H_a: \mu_x - \mu_y < \Delta_0 \quad t' \leq -t_{\alpha, \nu}$

$H_a: \mu_x - \mu_y \neq \Delta_0 \quad t' \geq t_{\alpha/2, \nu} \quad \text{あるいは} \quad t' \leq -t_{\alpha/2, \nu}$

## 5.2 分散の比

第3章で、正規母集団  $N(\mu, \sigma^2)$  から得られたサンプルの分散  $S^2$  は、 $(n-1)S^2/\sigma^2$  は自由度  $n-1$  の  $\chi^2$  分布に従うことを紹介した。ここでは、二つの正規母集団の分散の比が統計分布と、それを利用した統計検定を紹介する。

$X_1$  が自由度  $\nu_1$  の  $\chi^2$  分布に従い、 $X_2$  が自由度  $\nu_2$  の  $\chi^2$  分布に従うとすると、

$$F = \frac{X_1/\nu_1}{X_2/\nu_2} \quad (5.15)$$

は、自由度  $(\nu_1, \nu_2)$  の  $F$  分布に従う。これと式 3.20 を用いると、

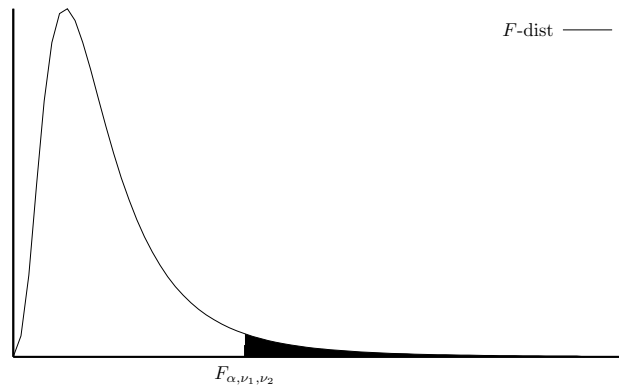


図 5.1:  $F$  分布の例と  $F_{\alpha, \nu_x, \nu_y}$  の定義。黒く塗った部分の面積が  $\alpha$  になる。

正規母集団からの標本分散の比

2つの正規母集団  $N(\mu_x, \sigma_x^2), N(\mu_y, \sigma_y^2)$  それぞれからの独立な標本 (それぞれ  $m, n$  個) の分散を  $s_x^2, s_y^2$  とすると、

$$F = \frac{s_x^2 / \sigma_x^2}{s_y^2 / \sigma_y^2} \quad (5.16)$$

は、自由度  $(m-1, n-1)$  の  $F$  分布に従う。

また、母分散の比に関する仮説検定は以下の通りとなる。

正規母集団の母分散に関する仮説検定

$$\text{帰無仮説 } H_0 : \sigma_x^2 = \sigma_y^2$$

$$\text{推定統計量 } f = \frac{s_x^2}{s_y^2}$$

対立仮説 近似的な棄却域

$$H_a : \sigma_x > \sigma_y \quad f \geq F_{\alpha, m-1, n-1}$$

$$H_a : \sigma_x < \sigma_y \quad f \leq F_{1-\alpha, m-1, n-1}$$

$$H_a : \sigma_x \neq \sigma_y \quad f \geq F_{\alpha/2, m-1, n-1} \quad \text{あるいは} \quad f \leq F_{1-\alpha/2, m-1, n-1}$$

## 演習

### 演習 5.1

2種類のワイヤーロードに対する強度試験を行った結果、以下の通りとなった。

種類	サンプル数	サンプル平均	サンプル標準偏差
AISI 1064	$m = 129$	$\bar{x} = 107.6$	$s_1 = 1.3$
AISI 1078	$n = 129$	$\bar{y} = 123.6$	$s_2 = 2.0$

AISI 1064 および AISI 1078 の真の平均値を  $\mu_1, \mu_2$  とするとき,  $\mu_1 - \mu_2$  の 95%信頼区間を求めよ (Devore, 1991) .

---

### 演習 5.2

---

ある種の殺虫剤が, いくつかの鳥を絶滅に追いやっていることが指摘されている. この殺虫剤が散布された領域で, 20 匹のキジを捕獲し ( $m = 20$ ), 体内の酵素 A を計測したところ平均  $\bar{x} = 1.55$ , 標準偏差  $s_1 = 0.39$  であった. 殺虫剤の散布領域外で捕獲したキジで同様の計測を行ったところ, 9 匹で ( $n = 9$ ) 平均  $\bar{y} = 1.60$ , 標準偏差  $s_2 = 0.40$  であった. この酵素 A の体内量の平均値の殺虫剤の散布領域と散布領域外での平均の差 ( $\mu_1 - \mu_2$ ) の 95%信頼区間を求めよ (Devore, 1991) .

## 第6章 相関と回帰分析

2 変数  $x$  と  $y$  , あるいは 3 変数  $x, y, z$  など多次元データを扱う場合がある . 多次元の変数は , 例えば 2 次元ならば  $x$  と  $y$  が独立でなく , 相互に関係を持っている場合があり , 1 次元の統計とはまた異なる側面を持つ . ここでは , 2 次元のデータを主に取り上げ , 両者の関係を統計的に表す方法を議論する .

### 6.1 2 変数の統計量

#### 6.1.1 サンプルの共分散・相関係数

2 つの確率変数  $X$  と  $Y$  に関するペアのサンプル  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  がある時 , このサンプルの  $x$  および  $y$  それぞれの平均 , 分散は以下の通り .

$$\begin{aligned}\bar{x} &= \frac{1}{n} \sum_{i=1}^n x_i, & S_x^2 &= \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \\ \bar{y} &= \frac{1}{n} \sum_{i=1}^n y_i, & S_y^2 &= \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2\end{aligned}$$

$x, y$  の共分散 (covariance) は以下のように定義される .

$$S_{xy} \equiv \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \quad (6.1)$$

また , 相関係数は以下のように定義される .

$$r_{xy} = \frac{S_{xy}}{S_x S_y} = \frac{S_{xy}}{\sqrt{\frac{1}{n} \sum (x_i - \bar{x})^2} \sqrt{\frac{1}{n} \sum (y_i - \bar{y})^2}} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}} \quad (6.2)$$

相関係数は、以下の関係を満たす。

$$-1 \leq r \leq 1 \quad (6.3)$$

また、

$$z_i = \frac{x_i - \bar{x}}{S_x} \quad w_i = \frac{y_i - \bar{y}}{S_y} \quad (6.4)$$

とおくと、式 6.2 は

$$r_{xy} = \frac{1}{n} \sum z_i w_i \quad (6.5)$$

とおくことができる。すなわち相関係数は、無次元化した 2 変量同士の共分散である。

### 6.1.2 母集団の共分散・相関係数

離散型の2つの確率変数  $X$  と  $Y$  の同時確率分布 (Joint probability distribution) を

$$P(X = x, Y = y) = f(x, y) \quad (6.6)$$

と記す。 $f(x, y)$  は、以下の式を満たす。

$$f(x, y) \geq 0 \quad \sum_x \sum_y f(x, y) = 1 \quad (6.7)$$

$X, Y$  が連続型の場合は、同時確率密度関数  $f_{XY}(x, y)$  は以下の式を満たす。

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1 \quad (6.8)$$

$$P(a \leq X \leq b, c \leq Y \leq d) = \int_a^b \int_c^d f(x, y) dx dy \quad (6.9)$$

同時確率密度分布あるいは関数を用いて、共分散は以下のように定義される。

$$Cov(X, Y) = \sum_x \sum_y (x - \mu_x)(y - \mu_y) f_{XY}(x, y) \quad (X, Y) \text{ が離散型} \quad (6.10a)$$

$$Cov(X, Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_x)(y - \mu_y) f_{XY}(x, y) dx dy \quad (X, Y) \text{ が連続型} \quad (6.10b)$$

また、以下のようにも表される

$$Cov(X, Y) = E\{(X - \mu_x)(Y - \mu_y)\} = E(XY) - \mu_X \mu_Y \quad (6.11)$$

母集団の相関係数は以下のように定義される。

$$\rho_{X,Y} = \frac{Cov(X, Y)}{\sigma_X \sigma_Y} \quad (6.12)$$

サンプル相関係数は、母集団の相関係数の不偏な点推定量である。

$$\hat{\rho}_{X,Y} = r_{xy} \quad (6.13)$$

## 6.2 単純線形回帰モデル

### 6.2.1 最小二乗推定法

$n$  個のペアのサンプル  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  があるとする。 $x$  が独立変数 (independent variable) で、 $y$  がその従属変数 (dependent variable) であり、 $x$  から  $y$  が以下の式で推定できるとする。

$$Y = \beta_0 + \beta_1 X + \epsilon \quad (6.14)$$

この直線関係の  $y$  切片  $\beta_0$ 、傾き  $\beta_1$  の上記のサンプルからの推定値をそれぞれ  $b_0, b_1$  とすると、推定される回帰直線は、

$$y = b_0 + b_1 x \quad (6.15)$$

で表される。 $x_i$  による  $y$  の推定値は

$$\hat{y}_i = b_0 + b_1 x_i \quad (6.16)$$

であり、推定誤差は

$$\epsilon_i = y_i - \hat{y}_i = y_i - b_0 - b_1 x_i \quad (6.17)$$

である。最小二乗法による推定値  $b_0, b_1$  は、

$$e = \sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n (y_i - b_0 - b_1 x_i)^2 \quad (6.18)$$

を最小にする。よって、

$$\frac{\partial e}{\partial b_0} = 0 \quad \frac{\partial e}{\partial b_1} = 0 \quad (6.19)$$

が必要。これを解くと以下の関係を得る。

単純線形回帰直線

$$b_1 = \hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{S_{xy}}{S_x^2} \quad (6.20a)$$

$$b_0 = \hat{\beta}_0 = \frac{\sum y_i - \hat{\beta}_1 \sum x_i}{n} = \bar{y} - \hat{\beta}_1 \bar{x} \quad (6.20b)$$

### 6.2.2 回帰直線の特性と決定係数

係数推定 6.20 を用いると回帰直線は、以下のようにあらわされる。

$$y - \bar{y} = \frac{S_{xy}}{S_x^2} (x - \bar{x}) \quad (6.21)$$

この回帰直線は、以下の性質を持つ。

- 点  $(\bar{x}, \bar{y})$  を必ず通る。
- 傾きは、 $S_{xy}/S_x^2$ 。

切片と傾きの推定値を用いて誤差平方和 (Square sum of errors,  $SSE$ ) が以下のように定義される。

$$SSE = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \quad (6.22)$$

$SSE$  は以下の式のいずれかより計算される。

$$\begin{aligned} SSE &= \sum y_i^2 - \hat{\beta}_0 \sum y_i - \hat{\beta}_1 \sum x_i y_i \\ SSE &= \sum (y_i - \bar{y})^2 - \frac{\{\sum (x_i - \bar{x})(y_i - \bar{y})\}^2}{\sum (x_i - \bar{x})^2} \\ SSE &= \left\{ \sum y_i^2 - \frac{1}{n} \left( \sum y_i \right)^2 \right\} - \frac{\left( \sum x_i y_i - \frac{1}{n} \sum x_i \sum y_i \right)^2}{\sum x_i^2 - \frac{1}{n} \left( \sum x_i \right)^2} \end{aligned}$$

$SSE$  を用いて、平均二乗誤差 (error mean square, mean squared error,  $MSE$ ) が定義される。

$$MSE = \frac{SSE}{n-2} \quad (6.23)$$

$MSE$  は、回帰直線 6.15 がどの程度、実際のデータにあっているか (goodness of fit) の指標の一つであり、回帰直線周りの  $y$  の分散  $\sigma^2$  の不偏推定量である。すなわち、

$$E(MSE) = \sigma^2 \quad (6.24)$$

$y$  の元々の平均周りの変動の二乗和を総平方和 (total sum of squares) と呼び、 $SST$  で表す。

$$SST = \sum (y_i - \bar{y})^2 \quad (6.25)$$

である。 $SST$  と  $SSE$  を用いて、決定係数 (Coefficient of Determination)  $R^2$  を以下のように定義する。

決定係数

$$R^2 = 1 - \frac{SSE}{SST} \quad (6.26)$$

$SSE$  は回帰分析の後で残された分散であるから、 $R^2$  は  $y_i$  の分散のうち回帰分析によって説明された分散の割合と解釈することができる。

単純な線形回帰モデル 6.15 に最小二乗法による係数推定 6.20 を用いた場合、決定係数は相関係数の二乗に等しい。これは、重回帰、あるいは非線形回帰のときには当てはまらないので注意。

$$R^2 = r^2 \quad \text{ただし、単純線形回帰モデルの場合} \quad (6.27)$$

## 6.3 様々な線形回帰モデル

### 6.3.1 独立変数と従属変数

線形回帰直線の推定値 6.20 は、式 6.18 で定義される  $e$  が最小になる推定値である。これは、 $X$  から  $Y$  を推定するにあたって、その推定誤差がもっとも確率論的に最小になることを意味する。

反対に  $Y$  から  $X$  を推定するときの推定誤差を最小にするような、回帰直線の推定も可能である。回帰直線を

$$x = b'_0 + b'_1 y \quad \text{あるいは} \quad y = \frac{1}{b'_1} x - \frac{b'_0}{b'_1} \quad (6.28)$$

とする。式 6.20 の導出と同様に

$$e' = \sum (x_i - \hat{x}_i)^2 = \sum (x_i - b'_0 - b'_1 y_i)^2 \quad (6.29)$$

を最小にする回帰係数を求めると

$$b'_1 = \frac{S_{xy}}{S_y^2} \quad (6.30)$$

$$b'_0 = \bar{x} - b'_1 \bar{y} \quad (6.31)$$

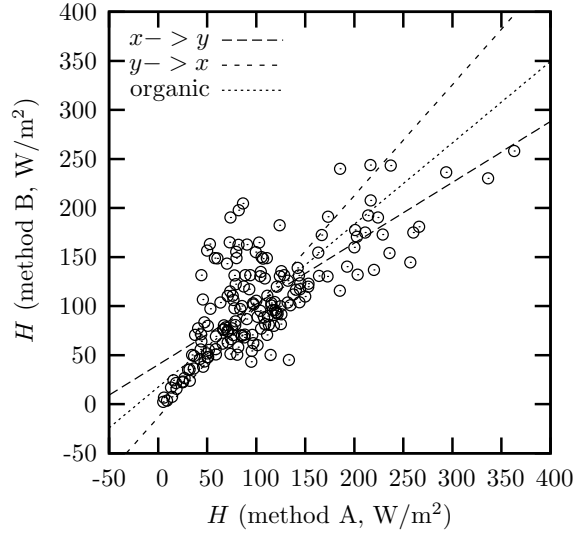


図 6.1: 式 6.20 および式 6.33、式 6.36 による線形回帰直線。データはモンゴルにおいて 2 種類の手法によって観測された顕熱フラックスである。

が得られる。これによって回帰直線は、

$$y - \bar{y} = \frac{S_y^2}{S_{xy}}(x - \bar{x}) \quad (6.32)$$

となる。別の方法による  $\beta_0, \beta_1$  の推定値

$$\hat{\beta}_1 = \frac{\sum (y_i - \bar{y})^2}{\sum (x_i - \bar{x})(y_i - \bar{y})} = \frac{S_y^2}{S_{xy}} \quad (6.33a)$$

$$\hat{\beta}_0 = \frac{\sum y_i - \hat{\beta}_1 \sum x_i}{n} = \bar{y} - \hat{\beta}_1 \bar{x} \quad (6.33b)$$

が導き出されたことになる。

### 6.3.2 生態相関

応用事例によっては、 $X$  と  $Y$  の従属関係が明確でないことがある。そのような場合は、 $X$  と  $Y$  を同等の変数として取り扱うような回帰直線が必要になる場合がある。そのような例として、生態相関 (organic correlation Kermack and Haldane, 1950; Kruskal, 1953; Kritsky and Menkel, 1968; Helsel and Hirsch, 1993) がある。生態相関では、回帰直線

$$y = b_0 + b_1 x \quad (6.34)$$

が式 6.18, 6.29 の代わりに以下の式を最小にする。

$$e'' = \sum (x_i - \hat{x}_i)(y_i - \hat{y}_i) = \sum \left(x_i - \frac{b_0}{b_1} - \frac{1}{b_1} y_i\right)(y_i - b_0 - b_1 x) \quad (6.35)$$

以前と同様に、回帰係数が求められる。

$$b_1 = \hat{\beta}_1 = \text{sgn}(S_{xy}) \sqrt{\frac{\sum (y_i - \bar{y})^2}{\sum (x_i - \bar{x})^2}} = \text{sgn}(S_{xy}) \sqrt{\frac{S_y^2}{S_x^2}} \quad (6.36a)$$

$$b_0 = \hat{\beta}_0 = \frac{\sum y_i - \hat{\beta}_1 \sum x_i}{n} = \bar{y} - \hat{\beta}_1 \bar{x} \quad (6.36b)$$

ここで、 $\text{sgn}(x)$  は

$$\text{sgn}(x) = \quad (6.37)$$

で  $x$  の符号を表す。

図 6.1 は、式 6.20 および式 6.33、式 6.36 の 3 種類の方法で推定された回帰直線の例である。式 6.36 によって推定された回帰直線は、式 6.20、式 6.33 による回帰直線の真ん中になることがわかる。

## 6.4 原点を通る線形回帰モデル

### 6.4.1 原点を通る単純回帰モデル

応用例によっては  $X, Y$  が必ず、 $(0, 0)$  をとらなければならない場合がある。この場合は、式 6.15 の代わりに、以下のモデルを観測されたサンプルに当てはめる。

$$Y = \beta_1 X + \epsilon \quad (6.38)$$

$\beta_1$  の推定値を  $\hat{\beta}_1 = b_1$  とすれば、これは、原点を強制的に通る直線

$$y = b_1 x \quad (6.39)$$

をデータに対してフィットさせることと同一である。式 6.18 と同様に、

$$\sum \epsilon_i^2 = \sum (y_i - b_1 x_i)^2 \quad (6.40)$$

を最小化させるような回帰係数は以下の通りである。

原点を通る単純線形回帰直線

$$b_1 = \hat{\beta}_1 = \frac{\sum x_i y_i}{\sum x_i^2} \quad (6.41)$$

式 6.41 を用いた回帰直線に関する  $SSE, MSE$  は、以下のように求められる。

$$SSE = \sum \epsilon_i^2 = \sum (y_i - b_1 x_i)^2 \quad (6.42a)$$

$$MSE = \frac{SSE}{n - 1} \quad (6.42b)$$

### 6.4.2 原点を通る生態相関

生態相関を用いて原点を通る線形回帰直線を求めることができる。その場合は、  
 原点を通る生態相関

$$b_1 = \hat{\beta}_1 = \text{sgn}(S_{xy}) \sqrt{\frac{\sum y_i^2}{\sum x_i^2}} \quad (6.43)$$

となる。

## 6.5 相関係数に関する区間推定・仮説検定

### 6.5.1 区間推定

2つの確率変数  $X, Y$  からのペアのサンプル  $(X_1, Y_1), \dots, (X_n, Y_n)$  を考える。 $X, Y$  が母平均  $\mu_X, \mu_Y$ 、母分散  $\sigma_X^2, \sigma_Y^2$  の、母共分散が  $\sigma_{XY}$  の2次元正規母集団  $N((\mu_X, \mu_Y), (\sigma_X^2, \sigma_Y^2, \sigma_{XY}))$  に従うとする。母共分散  $\sigma_{XY}$  は、

$$\sigma_{XY} = E[(X - \mu_1)(Y - \mu_2)] \quad (6.44)$$

であり、これを用いると母集団の相関係数は、以下のように定義される。

$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \quad (6.45)$$

サンプル共分散

$$s_{XY} = \frac{1}{n-1} \sum (X_i - \bar{X})(Y_i - \bar{Y}) \quad (6.46)$$

とサンプル分散  $s_X^2, s_Y^2$  を用いると、サンプル相関係数は  $\rho$  の点推定値であり、

$$\hat{\rho} = R = \frac{s_{XY}}{s_X s_Y} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2} \quad (6.47)$$

で定義される。 $R$  の標本分布は近似的にフィッシャー変換を用いて以下のように求められる。

サンプル相関係数の標本分布

$(X_1, Y_1), \dots, (X_n, Y_n)$  が2次元正規分布からのサンプルであるとき

$$V = \frac{1}{2} \ln \left( \frac{1+R}{1-R} \right) \quad (6.48)$$

は平均、分散が、

$$\mu_V = \frac{1}{2} \ln \left( \frac{1+\rho}{1-\rho} \right), \quad \sigma_v = \frac{1}{n-3} \quad (6.49)$$

の正規分布  $N(\mu_V, \sigma_v^2)$  に従う。

これを用いると、区間推定は以下のようにしてできる。

### 母集団の相関係数の区間推定

サンプルから求められた相関係数を  $r$  とし、

$$v = \frac{1}{2} \ln \left( \frac{1+r}{1-r} \right), \quad c_1 = v - \frac{z_{\alpha/2}}{\sqrt{n-3}}, \quad c_2 = v + \frac{z_{\alpha/2}}{\sqrt{n-3}} \quad (6.50)$$

とすれば、 $\rho$  の  $100(1-\alpha)\%$  信頼区間は、

$$\left( \frac{e^{2c_1} - 1}{e^{2c_1} + 1}, \frac{e^{2c_2} - 1}{e^{2c_2} + 1} \right) \quad (6.51)$$

である。

## 6.5.2 仮説検定

相関係数に関する検定は、式 6.48 が正規分布に従うことを利用する。

相関係数に関する検定 ( $\sigma$  が既知の時)

帰無仮説  $H_0 : \rho = \rho_0$

$$\text{検定統計量 } Z = \frac{V - \frac{1}{2} \ln[(1+\rho_0)/(1-\rho_0)]}{1/\sqrt{n-3}}$$

対立仮説 棄却域

$$H_a : \rho > \rho_0 \quad z \geq z_\alpha$$

$$H_a : \rho < \rho_0 \quad z \leq -z_\alpha$$

$$H_a : \rho \neq \rho_0 \quad z \geq z_{\alpha/2} \quad \text{あるいは} \quad z \leq -z_{\alpha/2}$$

## 6.6 回帰分析に関する区間推定・仮説検定

### 6.6.1 単純回帰直線

最小二乗法による単純回帰直線  $y = \beta_0 + \beta_1 x$  の  $\beta_0$  と  $\beta_1$  の推定値 (式 6.20)

単純線形回帰直線

$$b_1 = \hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$b_0 = \hat{\beta}_0 = \frac{\sum y_i - \hat{\beta}_1 \sum x_i}{n} = \bar{y} - \hat{\beta}_1 \bar{x}$$

について、区間推定と検定を行う。

## 単純回帰直線の傾き $\beta_1$

$b_1$  の標本分布は平均、分散を

$$E(b_1) = \beta_1 \quad (6.53)$$

$$V(b_1) = \frac{\sigma^2}{\sum (x_i - \bar{x})^2} \quad (6.54)$$

とする正規分布に従う。式 6.24 によれば、 $MSE$  は  $y$  の回帰直線周りの分散  $\sigma^2$  の不偏推定値である。よって、 $V(b_1)$  の推定値は、

$$s_{b_1}^2 = \frac{MSE}{\sum (x_i - \bar{x})^2} \quad (6.55)$$

は  $b_1$  の分散の不偏点推定値である。よって、

$b_1$  の標本分布

$$\frac{b_1 - \beta_1}{s_{b_1}} \text{ は自由度 } n - 2 \text{ の } t \text{ 分布に従う。} \quad (6.56)$$

である。これを用いると  $\beta_1$  の区間推定は以下の通り。

$\beta_1$  の区間推定

$\beta_1$  の  $100(1 - \alpha)\%$  の信頼区間は、

$$(b_1 - t_{1-\alpha/2, n-2} \cdot s_{b_1}, b_1 + t_{1-\alpha/2, n-2} \cdot s_{b_1}) \quad (6.57)$$

である。

また、 $\beta_1$  に関する仮説検定は以下の通りとなる。

最小二乗法による回帰直線の傾き  $\beta_1$  に関する仮説検定

帰無仮説  $H_0 : \beta_1 = \beta_{10}$

検定統計量  $T = \frac{b_1 - \beta_{10}}{s_{b_1}}$

対立仮説 棄却域

$H_a : \beta_1 > \beta_{10} \quad t \geq t_{\alpha, n-2}$

$H_a : \beta_1 < \beta_{10} \quad t \leq -t_{\alpha, n-2}$

$H_a : \beta_1 \neq \beta_{10} \quad t \geq t_{\alpha/2, n-2} \text{ あるいは } t \leq -t_{\alpha/2, n-2}$

## 単純回帰直線の切片 $\beta_0$

$b_0$  の標本分布は平均、分散を

$$E(b_0) = \beta_0 \quad (6.58)$$

$$V(b_0) = \sigma^2 \frac{\sum x_i^2}{n \sum (x_i - \bar{x})^2} = \sigma^2 \left[ \frac{1}{n} + \frac{\bar{x}^2}{n \sum (x_i - \bar{x})^2} \right] \quad (6.59)$$

よって  $V(b_0)$  の推定値は、

$$s_{b_0}^2 = MSE \frac{\sum x_i^2}{n \sum (x_i - \bar{x})^2} = MSE \left[ \frac{1}{n} + \frac{\bar{x}^2}{n \sum (x_i - \bar{x})^2} \right] \quad (6.60)$$

よって、

$b_0$  の標本分布

$$\frac{b_0 - \beta_0}{s_{b_0}} \text{ は自由度 } n - 2 \text{ の } t \text{ 分布に従う。} \quad (6.61)$$

である。これを用いると  $\beta_0$  の区間推定は以下の通り。

$\beta_0$  の区間推定

$\beta_0$  の  $100(1 - \alpha)\%$  の信頼区間は、

$$(b_0 - t_{1-\alpha/2, n-2} \cdot s_{b_0}, b_0 + t_{1-\alpha/2, n-2} \cdot s_{b_0}) \quad (6.62)$$

である。

また、 $\beta_0$  に関する仮説検定は以下の通りとなる。

最小二乗法による回帰直線の切片  $\beta_0$  に関する仮説検定

帰無仮説	$H_0 : \beta_0 = \beta_{00}$
検定統計量	$T = \frac{b_0 - \beta_{00}}{s_{b_0}}$
対立仮説	棄却域
$H_a : \beta_0 > \beta_{00}$	$t \geq t_{\alpha, n-2}$
$H_a : \beta_0 < \beta_{00}$	$t \leq -t_{\alpha, n-2}$
$H_a : \beta_1 \neq \beta_{00}$	$t \geq t_{\alpha/2, n-2}$ あるいは $t \leq -t_{\alpha/2, n-2}$

## 6.6.2 原点を通る回帰直線に関する検定

原点を通る回帰直線  $y = \beta_1 x$  の  $\beta_1$  の推定値 (式 6.41 )

原点を通る単純線形回帰直線

$$b_1 = \hat{\beta}_1 = \frac{\sum x_i y_i}{\sum x_i^2} \quad (6.63)$$

に関する区間推定は以下の通りとなる。

$b_1$  は平均は

$$E(b_1) = \beta_1 \quad (6.64)$$

で、その分散の推定値は

$$s_{b_1}^2 = \frac{MSE}{\sum X_i^2} \quad (6.65)$$

である。よって、 $\beta_1$  の区間推定は

原点を通る回帰直線の傾き  $\beta_1$  の区間推定

$\beta_0$  の  $100(1 - \alpha)\%$  の信頼区間は、

$$(b_1 - t_{1-\alpha/2, n-2} \cdot s_{b_0}, b_1 + t_{1-\alpha/2, n-2} \cdot s_{b_0}) \quad (6.66)$$

となる。

## 6.7 演習

### 演習 6.1

ある草原における  $x$ : 6 月から 8 月までの降水量 (mm) と  $y$ : 多年生植物の生産量 (kg/ha) の 10 年間の各年の観測値は以下の通りである (Devore, 1991, p.495)。

$x$ : 22.05 35.04 30.48 11.89 27.28 9.63 17.63 22.20 17.27 19.63

$y$ : 291 629 823 307 660 263 375 366 563 558

このデータを用いて、

1.  $x$  と  $y$  をそれぞれ横軸、縦軸にして、上記のデータをプロットせよ。
2.  $x$  と  $y$  のそれぞれの平均、分散、 $x$  と  $y$  の共分散、そして相関係数を求めよ。

### 演習 6.2

ある交差点の右折レーンでの 1 回の信号待ちの時の自動車の数 ( $X$ ) とバスの数 ( $Y$ ) の同時確率密度分布  $f(x, y)$  は以下の通りである。 (Devore, 1991, p.198)

		$y$		
		0	1	2
$x$	$f(x, y)$	0	1	2
	0	0.025	0.015	0.010
	1	0.050	0.030	0.020
	2	0.125	0.075	0.050
	3	0.150	0.090	0.060
	4	0.100	0.060	0.040
5	0.050	0.030	0.020	

1. 信号待ちが自動車が 1 台でかつ、バスが 1 台である確率はいくつか。
2. 信号待ちが自動車が 1 台以下、バスが 1 台以下である確率はいくつか。
3. 信号待ちが自動車が 1 台である確率はいくつか。バスが 1 台である確率はいくつか。
4. 右折レーンは、5 台分しかないとする。バス 1 台は自動車 3 台に相当する。右折レーンがいっぱいになる確率はいくつか。
5.  $X$ 、 $Y$  のそれぞれの平均、分散を求めよ。

6.  $X, Y$  の共分散を求めよ。また相関係数を求めよ。
7.  $X$  と  $Y$  は独立か、従属か

---

### 演習 6.3

---

宿題 6.1 の例で、回帰直線の傾きと切片を求めよ。

---

### 演習 6.4

---

図 6.1 にならい、各自の持つデータの中から同じような例を挙げ、そのデータに対して式 6.20 および式 6.33、式 6.36 の 3 種類の回帰直線を求め、グラフにデータとともにプロットせよ。適切なデータが無い場合には、図 6.1 のデータ<sup>1</sup> を用いよ。

---

### 演習 6.5

---

式 6.42 を展開せよ

---

### 演習 6.6 原点を通る回帰直線 1

---

宿題 6.1 のデータを用いて、式 6.20 および式 6.41 による回帰直線を求め、グラフにデータとともにプロットせよ。またそれぞれについて  $SSE$ ,  $MSE$  および決定係数  $R^2$  を求めよ。

---

### 演習 6.7 原点を通る回帰直線 2

---

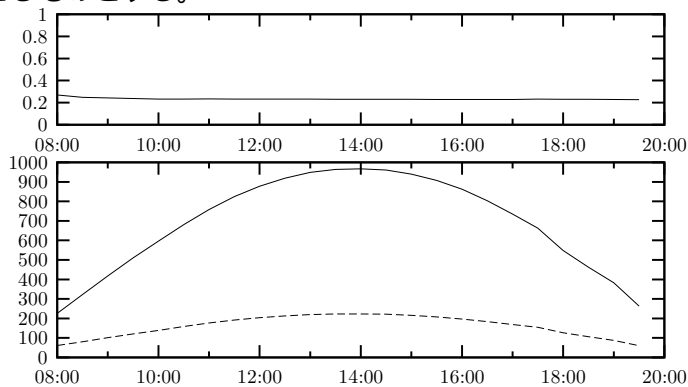
宿題 6.4 のデータを用いて、および式 6.41 および式 6.43 による回帰直線を求め、データとともにプロットせよ。

---

### 演習 6.8 原点を通る回帰直線: 比の平均の計算

---

下の図は、2003/5/23 のモンゴル草原における日射量  $S_d$  と反射量  $S_u$  の観測値<sup>2</sup> である。上のパネルは  $(\text{反射量}) \div (\text{日射量})$  でアルベドと呼ぶ。今これは時間によって変化しない一定値をとるものとする。



この一日の平均アルベドを以下の 3 つの方法で求め、比較せよ。

1.  $S_d/S_u$  の平均値として。
2.  $S_d$  の総和と  $S_u$  の総和の比として。
3.  $S_d$  を  $x$  軸、 $S_u$  を  $y$  軸にプロットして、その原点を通る回帰直線の傾きとして。

---

<sup>1</sup><http://www.suiri.tsukuba.ac.jp/~asanuma/courses/currnet/geostats/>

<sup>2</sup><http://www.suiri.tsukuba.ac.jp/~asanuma/courses/currnet/geostats/>

---

**演習 6.9      相関係数に関する区間推定**

---

水中の窒素除去方法の研究によれば、流入水中の窒素濃度  $x(\text{mg/l})$  と、ある方法による窒素の除去率  $y$  に関する 20 日間の統計量は以下の通り。

$$\begin{aligned}\sum x_i &= 285.9, & \sum x_i^2 &= 4409.55, \\ \sum y_i &= 690.30, & \sum y_i^2 &= 29040.29, & \sum x_i y_i &= 10869.71\end{aligned}$$

1. サンプル相関係数を求めよ。
2. 母相関係数の 95%信頼区間を求めよ。

---

**演習 6.10      相関係数に関する仮説検定**

---

宿題 6.9 において、流入水中の窒素濃度と窒素除去率の相関係数が少なくとも 0.5 以上であることを検定せよ。5%の有意水準で検定せよ (Devore, 1991, より改編)。

---

**演習 6.11      回帰直線の傾きに関する仮説検定**

---

宿題 7.2 の例を用いて、単純回帰直線の傾き  $\beta_1$  の 95%信頼区間を求めよ。また、製作個数と総労働時間の間に線形関係があるかどうか ( $\beta_1 = 0$  であるかどうか) を、5%の有意水準で検定せよ (Neter et al., 1990)。

---

**演習 6.12      回帰直線の切片に関する区間推定**

---

宿題 7.2 の例を用いて、単純回帰直線の傾き  $\beta_0$  の 95%信頼区間を求めよ (Neter et al., 1990)。

## 第7章 多変量回帰分析 (Multiple Regression)

### 7.1 線形回帰分析の行列表現

#### 7.1.1 最小二乗法の行列表現

観測値  $(x_i, y_i)$ 、回帰係数  $b_0, b_1$ 、誤差  $\epsilon_i$  を用いて、以下の行列を定義する。

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \quad \mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \quad (7.1)$$

このとき

$$\boldsymbol{\epsilon} = \mathbf{Y} - \mathbf{X}\mathbf{b} \quad (7.2)$$

である。よって、式 6.18 は、以下のように行列式で表される。

$$e = \sum \epsilon_i^2 = \boldsymbol{\epsilon}^T \boldsymbol{\epsilon} = (\mathbf{Y} - \mathbf{X}\mathbf{b})^T (\mathbf{Y} - \mathbf{X}\mathbf{b}) \quad (7.3)$$

ここで、 $^T$  は転置行列である。上式を展開すると、

$$e = \mathbf{Y}^T \mathbf{Y} - 2\mathbf{b}^T \mathbf{X}^T \mathbf{Y} + \mathbf{b}^T \mathbf{X}^T \mathbf{X} \mathbf{b}$$

となる。最小二乗の条件 (式 6.19) は、

$$\frac{\partial}{\partial \mathbf{b}} e = \begin{bmatrix} \frac{\partial e}{\partial b_0} \\ \frac{\partial e}{\partial b_1} \end{bmatrix} = 0 \quad (7.4)$$

と表されるので、

$$\frac{\partial}{\partial \mathbf{b}} e = -2\mathbf{X}^T \mathbf{Y} + 2\mathbf{X}^T \mathbf{X} \mathbf{b} = 0$$

すなわち、

$$\mathbf{X}^T \mathbf{X} \mathbf{b} = \mathbf{X}^T \mathbf{Y} \quad (7.5)$$

ここで、

$$\mathbf{X}^T \mathbf{X} = \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} \quad \mathbf{X}^T \mathbf{Y} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix}$$

である。よって、式 6.20 の行列表現

$$\mathbf{b} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} \quad (7.6)$$

を得る。

式 7.6 を用いると、 $\mathbf{Y}$  の推定値は、

$$\hat{\mathbf{Y}} = \mathbf{X}\mathbf{b} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} \quad (7.7)$$

あるいは、

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \quad (7.8)$$

とすると

$$\hat{\mathbf{Y}} = \mathbf{H}\mathbf{Y} \quad (7.9)$$

と表される。

$SST$ ,  $SSE$  は、以下の通りとなる。まず  $SST$  は式 6.25 より、

$$SST = \sum y_i^2 - n\bar{y}^2 = \sum y_i^2 - \frac{(\sum y_i)^2}{n} \quad (7.10)$$

よって、

$$\begin{aligned} SST &= \mathbf{Y}^T \mathbf{Y} - \frac{1}{n} \mathbf{Y}^T \mathbf{J} \mathbf{Y} \\ &= \mathbf{Y}^T \left[ \mathbf{I} - \frac{1}{n} \mathbf{J} \right] \mathbf{Y} \end{aligned} \quad (7.11)$$

ここで、

$$\mathbf{J} = \begin{bmatrix} 1 & \cdots & 1 \\ 1 & \cdots & 1 \\ \cdot & & \cdot \\ \cdot & & \cdot \\ 1 & \cdots & 1 \end{bmatrix} \quad (7.12)$$

である。また  $SSE$  は、

$$\begin{aligned} SSE &= \sum \epsilon_i^2 = \boldsymbol{\epsilon}^T \boldsymbol{\epsilon} = (\mathbf{Y} - \mathbf{X}\mathbf{b})^T (\mathbf{Y} - \mathbf{X}\mathbf{b}) \\ &= \mathbf{Y}^T \mathbf{Y} - \mathbf{b}^T \mathbf{X}^T \mathbf{Y} \\ &= \mathbf{Y}^T \left[ \mathbf{H} - \frac{1}{n} \mathbf{J} \right] \mathbf{Y} \end{aligned} \quad (7.13)$$

となる。

## 7.2 多変量線形回帰分析 (Multiple Regression)

目的とする  $Y$  が一つ以上の変数に依存するときは多くある．このとき  $Y$  が依存する変数が互いに独立であるとは限らない．そこでこれらの変数を「説明変数」(carrier) と呼ぶ．

説明変数が 2 つの時、これを  $X_1, X_2$  とすると、 $Y$  との間に

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon \quad (7.14)$$

の関係を仮定することができる．同様に、説明変数が  $p-1$  個ある時、

$$Y = \beta_0 + \sum_{i=1}^{p-1} \beta_i X_i + \epsilon \quad (7.15)$$

と仮定することになる．これは  $p-1$  次元における平面をデータにフィッティングさせることに他ならない． $\beta_i$  の推定値を  $b_i$ 、 $n$  組のデータを  $(x_{11}, x_{12}, \dots, x_{1,p-1}, y_1), \dots, (x_{n1}, x_{n2}, \dots, x_{n,p-1}, y_n)$  をとすると、

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1,p-1} \\ 1 & x_{21} & \cdots & x_{2,p-1} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{n,p-1} \end{bmatrix} \quad \mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_{p-1} \end{bmatrix} \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \quad (7.16)$$

で回帰式は

多変量回帰直線

$$\mathbf{Y} = \mathbf{X}\mathbf{b} + \boldsymbol{\epsilon} \quad (7.17)$$

である。

最小二乗による  $\mathbf{b}$  の推定値は、単純線形回帰の場合と同じく、

多変量回帰直線

$$\mathbf{b} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} \quad (7.18)$$

である。

$\mathbf{Y}$  の推定値は

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \quad (7.19)$$

とすると

$$\hat{\mathbf{Y}} = \mathbf{X}\mathbf{b} = \mathbf{H}\mathbf{Y} \quad (7.20)$$

であるので、

$$\boldsymbol{\epsilon} = \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I} - \mathbf{H})\mathbf{Y} \quad (7.21)$$

よって  $SST$ ,  $SSE$ ,  $MSE$  は以下の通り

$$SST = \mathbf{Y}^T \left[ \mathbf{I} - \frac{1}{n} \mathbf{J} \right] \mathbf{Y} \quad (7.22)$$

$$SSE = \boldsymbol{\epsilon}^T \boldsymbol{\epsilon} = \mathbf{Y}^T [\mathbf{I} - \mathbf{H}] \mathbf{Y} \quad (7.23)$$

$$MSE = \frac{SSE}{n - p} \quad (7.24)$$

よって、重決定係数 (coefficient of multiple determination) は、

$$R^2 = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST} \quad (7.25)$$

となる。重決定係数は、説明変数のセット  $(X_1, X_2, \dots, X_{p-1})$  を使用した時に、 $Y$  の変動がどの程度減少するかを示している。

一般に次元の数が増えるに従って、決定係数は 1 に近づく。よって、次元数を考慮に入れた以下の修正決定係数 (adjusted coefficient of determination) を用いることがある。

$$R_a^2 = 1 - \frac{\frac{SSE}{n-p}}{\frac{SST}{n-1}} = 1 - \left( \frac{n-1}{n-p} \right) \frac{SSE}{SST} \quad (7.26)$$

### 7.3 偏決定係数

地球環境科学の分野においては、目的変数  $Y$  が依存する説明変数  $X_i$  の候補のうち、どれが真に  $Y$  を支配するか、不明であることが多くある。このような場合、説明変数の候補からいく通りかの変数の組を用いた多変量解析が可能である。

たとえば、以下のように説明変数が 2 つのモデルを例にとる。

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon \quad (7.27)$$

このとき、 $X_1$  のみを用いて回帰分析をした場合、すなわち  $X_1$  のみをモデルに用いた時の  $SSE$  を  $SSE(X_1)$  と書くことにする。また  $SSE(X_1, X_2)$  は、 $X_1$  と  $X_2$  をモデルに用いた時の  $SSE$  である。よって、 $SSE(X_1) - SSE(X_1, X_2)$  は、 $X_1$  がすでにモデルに使用されていて、 $X_2$  をモデルに加えたときに  $Y$  の変動の減少である。

よって、

$$r_{Y2,1}^2 \equiv \frac{SSE(X_1) - SSE(X_1, X_2)}{SSE(X_1)} = 1 - \frac{SSE(X_1, X_2)}{SSE(X_1)} \quad (7.28)$$

は、すでに  $X_1$  があったときに、 $X_2$  をモデルに加えるとどの程度の割合で、 $Y$  の変動が減少するかを示し、偏決定係数 (coefficient of partial determination) と呼ぶ。同じように

$$r_{Y1,2}^2 \equiv 1 - \frac{SSE(X_1, X_2)}{SSE(X_2)} \quad (7.29)$$

は、すでに  $X_2$  がモデルに入っているときに、 $X_1$  を加えることによって、どの程度モデル性能が向上するか、を示している。

一般的に、

$$r_{Y3,12}^2 \equiv 1 - \frac{SSE(X_1, X_2, X_3)}{SSE(X_1, X_2)} \quad (7.30)$$

$$r_{Y2,13}^2 \equiv 1 - \frac{SSE(X_1, X_2, X_3)}{SSE(X_1, X_3)} \quad (7.31)$$

$$r_{Y1,23}^2 \equiv 1 - \frac{SSE(X_1, X_2, X_3)}{SSE(X_2, X_3)} \quad (7.32)$$

$$r_{Y4,123}^2 \equiv 1 - \frac{SSE(X_1, X_2, X_3, X_4)}{SSE(X_1, X_2, X_3)} \quad (7.33)$$

のように記す。

また、最適なモデルを探索するアルゴリズムとして、ステップワイズ回帰 (stepwise regression) (例えば Neter et al., 1990, p.453) などがある。このようなアルゴリズムは、たいてい、代表的な統計計算ソフトに組み込まれている。

## 7.4 演習

---

### 演習 7.1 回帰計算の行列表現 1

---

宿題 6.1 のデータを用いて、式 6.20 による回帰直線を行列表現を用いて計算せよ。

---

### 演習 7.2 回帰計算の行列表現 2

---

ある工場では、毎月、需要に応じて製品を製作する。過去 10 ヶ月間の各月の製品の製作個数と、それに要した総労働時間 (人・時間) は以下の通りである。(Neter et al., 1990, p.40)

月番号	製造製品数	総労働時間 (人・時間)
$i$	$x_i$	$y_i$
1	30	73
2	20	50
3	60	128
4	80	170
5	40	87
6	50	108
7	60	135
8	30	69
9	70	148
10	60	132

行列表現を用いて、回帰直線の切片と傾きを求めよ。

---

### 演習 7.3

---

ある化粧品会社があるスキนครリームを 15 の地域で独占的に販売している。各地域の販売個数を、人口と平均年収から予測したいと考えている。各地域のデータは以下の通り (Neter et al., 1990, p.249)。

地域	販売個数	人口	平均年収
	$y_i$	$x_{i1}$	$x_{i2}$
1	162	274	2,450
2	120	180	3,254
3	223	375	3,802
4	131	205	2,838
5	67	86	2,347
6	169	265	3,782
7	81	98	3,008
8	192	330	2,450
9	116	195	2,137
10	55	53	2,560
11	252	430	4,020
12	232	372	4,427
13	144	236	2,660
14	103	157	2,088
15	212	370	2,605

$Y$  を  $X_1, X_2$  で以下のように表すとき,

$$y = b_0 + b_1x_1 + b_2x_2$$

$\beta_0, \beta_1, \beta_2$  を求めよ．また，決定係数を求めよ．

---

#### 演習 7.4

---

下の表は，25 から 34 歳の健康な女性の被験者から得た，体脂肪量と，その説明変数として上腕三頭筋での皮下脂肪量，太股周囲，中腕周囲のデータである (Neter et al., 1990, p.271) ．

番号	皮下脂肪量	太股周囲	中腕周囲	体脂肪量
	$x_{i1}$	$x_{i2}$	$x_{i3}$	$y_i$
1	19.5	43.1	29.1	11.9
2	24.7	49.8	28.2	22.8
3	30.7	51.9	37.0	18.7
4	29.8	54.3	31.1	20.1
5	19.1	42.2	30.9	12.9
6	25.6	53.9	23.7	21.7
7	31.4	58.5	27.6	27.1
8	27.9	52.1	30.6	25.4
9	22.1	49.9	23.2	21.3
10	25.5	53.5	24.8	19.3
11	31.1	56.6	30.0	25.4
12	30.4	56.7	28.3	27.2
13	18.7	46.5	23.0	11.7
14	19.7	44.2	28.6	17.8
15	14.6	42.7	21.3	12.8
16	29.5	54.4	30.1	23.9
17	27.7	55.3	25.7	22.6
18	30.2	58.6	24.6	25.4
19	22.7	48.2	27.1	14.8
20	25.2	51.0	27.5	21.1

以下の 4 通りの回帰直線とそれぞれでの  $SSE, R^2$  を求めよ。

$$y = b_0 + b_1x_1$$

$$y = b_0 + b_2x_2$$

$$y = b_0 + b_1x_1 + b_2x_2$$

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3$$

また、 $r_{Y3,12}, r_{Y1,2}, r_{Y2,1}$  をそれぞれ計算せよ。

## 第8章 時系列データの解析

### 8.1 基本的な概念

#### 8.1.1 決定論的と確率論的

時系列データには、決定論的 (deterministic) データと確率論的 (random) データの二通りがある。決定論的データは、数学的に明確な関係式で表現できる。しかしながら、物理現象であっても完全に数学的な表現で表せないような場合が多い。例えば海の波浪の高さ、ノイズ発生器の出力などは予測できない現象の代表である。

決定論的データと確率論的データは、以下のように分類される。

- |            |   |   |
|------------|---|---|
| 1. 決定論的データ | $\left\{ \begin{array}{l} \text{周期データ} \\ \text{非周期データ} \end{array} \right\}$ | $\left\{ \begin{array}{l} \text{三角関数} \\ \text{複雑な周期データ} \\ \text{ほぼ周期的なデータ} \\ \text{短期的な非周期データ} \end{array} \right\}$ |
| 2. 確率論的データ | $\left\{ \begin{array}{l} \text{定常データ} \\ \text{非定常データ} \end{array} \right\}$ | $\left\{ \begin{array}{l} \text{エルゴード的定常データ} \\ \text{非エルゴード的定常データ} \end{array} \right\}$                               |

以下、その例を挙げる。

#### 8.1.2 決定論的データ

##### 三角関数

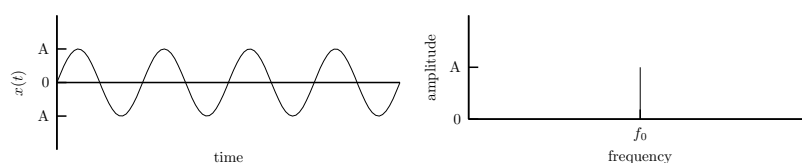


図 8.1: 三角関数 (左) とそのスペクトル (右)

##### 三角関数

$$x(t) = A \sin(2\pi f_0 t + \theta) \quad (8.1)$$

は代表的な周期データである。ここで、

$A$  : 振幅

$f_0$  : 周波数。単位時間あたりの回数

$\theta$  : 時間 0 のときの初期位相角。

である。周期  $T_p$  は

$$T_p = \frac{1}{f_0} \quad (8.2)$$

である。

### 複雑な周期データ

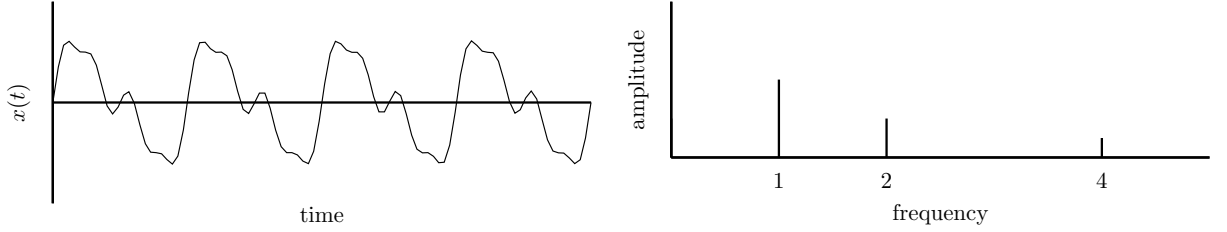


図 8.2: ほぼ周期的なデータの例  $x(t) = \sin 2\pi t + 0.5 \sin 4\pi t + 0.25 \sin 8\pi t$ 。時系列（左）とスペクトル（右）

一般に周期  $T_p$  をもつ周期 (periodic) データは以下の条件を満たす。

$$x(t) = x(t + nT_p) \quad (8.3)$$

例外を除いて、一般の周期データは三角関数の和として、フーリエ級数で表される。

$$x(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} (a_n \cos 2\pi n f_1 t + b_n \sin 2\pi n f_1 t) \quad (8.4)$$

### ほぼ周期的なデータ

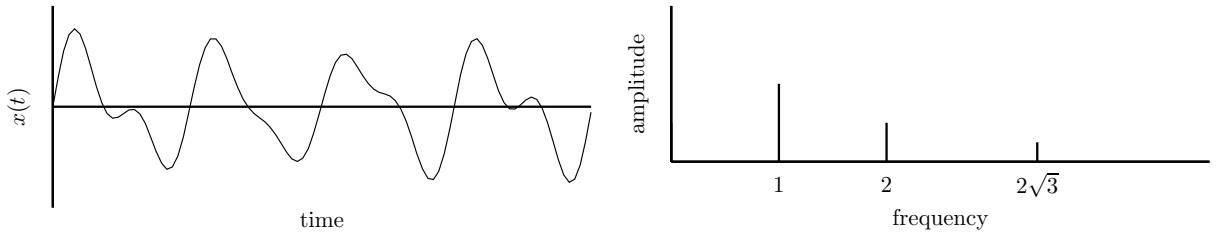


図 8.3: 複雑な周期データの例  $x(t) = \sin 2\pi t + 0.5 \sin 4\pi t + 0.25 \sin 2\sqrt{3}\pi t$ 。時系列（左）とスペクトル（右）

周期データが三角関数の和として表される一方で、三角関数の和がすべて周期データとは限らない。例えば、

$$x(t) = A_1 \sin(2t + \theta_1) + A_2 \sin(3t + \theta_2) + A_3 \sin(7t + \theta_3) \quad (8.5)$$

のように、各三角関数の周波数同士の比がすべて有理数であるとき、周期データとなる。この場合は、周期が  $T_p = 1$  である。これに対し、

$$x(t) = A_1 \sin(2t + \theta_1) + A_2 \sin(3t + \theta_2) + A_3 \sin(\sqrt{50}t + \theta_3) \quad (8.6)$$

は式 8.3 を満たさない。よって、一般的にほぼ周期的なデータは、

$$x(t) = \sum_{n=1}^{\infty} (a_n \sin 2\pi f_n t + b_n \sin 2\pi f_n t) \quad (8.7)$$

と表され、 $f_n/f_m$  が有理数にならない場合である。

このような時系列を生む物理現象の例として、2 つ以上の無関係の周期的な現象が混在する場合である。

### 短期的な非周期データ

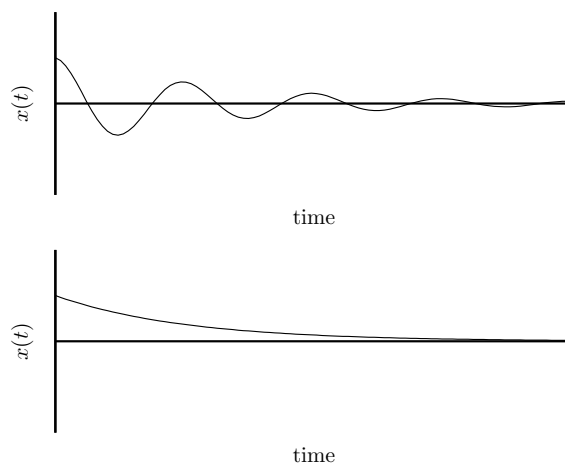


図 8.4: 短期的な周期データの例。

ほぼ周期的なデータ以外の非周期データはすべてここに分類される。単純な例を図 8.4 に示す。

### 8.1.3 確率論的データ

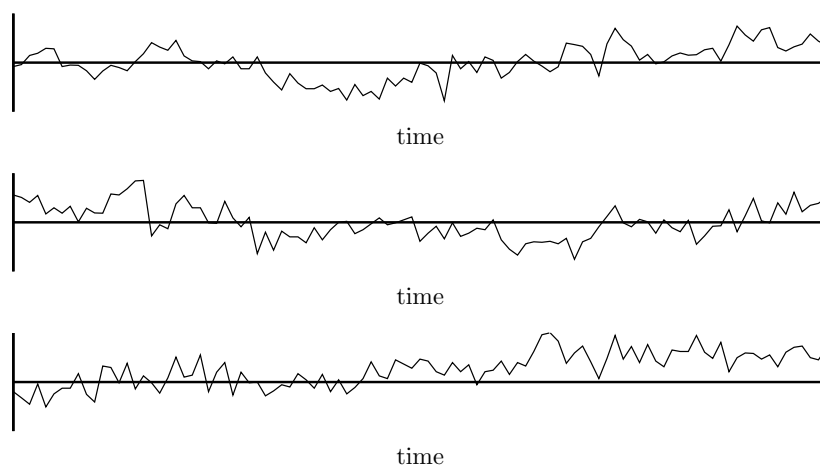


図 8.5: 確率論的データの例。2 秒間の風速成分の瞬間値。3 つの別の地点での同時の観測。

確率論的データは同じ現象を観測しても、観測ごとに観測結果は異なるのが特徴である。図 8.5 は、確率論的データの例であり、風速成分の瞬間値を 2 秒間、3 つの別の地点で同時に観測した観測値である。このように、一つ一つの観測時系列は、単に無限に存在し得る観測時系列の一つにかすぎない。このように一つの観測時系列を標本時系列、あるいは標本レコード (sample record) とよび、すべての標本レコードの集まりを生成するプロセスを確率過程 (random process、あるいは stochastic process) と呼ぶ。よって、一つの標本レコードは、確率過程の一つの具現化されたもの、と考えられる。

### 確率過程の定常性

確率過程の標本レコードの集合 (アンサンブル、ensemble と呼ぶ) がある時、このアンサンブルを  $\{x(t)\}$  であらわす。このとき平均は、

$$\mu_x(t_1) = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n x_k(t_1) \quad (8.8a)$$

また、自己相関関数 (auto-correlation function) は、以下のように定義される。

$$R_{xx}(t_1, t_1 + \tau) = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n x_k(t_1) x_k(t_1 + \tau) \quad (8.8b)$$

このような平均方法をアンサンブル平均 (ensemble average) と呼ぶ。

$\mu_x(t_1)$  と  $R_{xx}(t_1, t_1 + \tau)$  が時間  $t_1$  によらないとき、 $x(t)$  は広い意味で定常 (stationary) であると呼ぶ。広い意味で定常な確率過程では、平均は一定であり、自己相関関数は時間によらず  $\tau$  の関数である。すなわち、

$$\mu_x(t_1) = \mu_x \quad (8.9a)$$

$$R_{xx}(t_1, t_1 + \tau) = R_{xx}(\tau) \quad (8.9b)$$

である。またすべての高次モーメントが時間に  $t_1$  によらないとき、確率過程は厳密な意味で定常であると呼ぶ。実用的には、広い意味での定常性は厳密な意味での定常性と同一である。

### 確率過程のエルゴード性

平均や自己相関関数などの確率過程の特性値は、時間平均によっても計算可能である。 $k$  番目の標本レコードの平均、自己相関係数は以下のように表される。

$$\mu_x(k) = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T x_k(t) dt \quad (8.10a)$$

$$R_{xx}(\tau, k) = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T x_k(t) x_k(t + \tau) dt \quad (8.10b)$$

定常でかつ、 $\mu_x(k)$ ,  $R_{xx}(\tau, k)$  が  $k$  (標本レコード) によって変わらないとき、確率過程はエルゴード性を持つ (ergodic) と呼ぶ。エルゴード性を持つ確率過程においては、時間平均による平均および自己相関関数とアンサンブル平均による平均および自己相関関数は等しい。すなわち、

$$\mu_x(k) = \mu_x \quad (8.11a)$$

$$R_{xx}(\tau, k) = R_{xx}(\tau) \quad (8.11b)$$

である。また、定常なデータのみがエルゴード性を持つ。また、実用的には、定常性を持つ確率過程はエルゴード性を持つ。エルゴード性を持つ確率過程は、アンサンブル平均を用いずに時間平均によってその特性値を計算できる。

## 8.2 確率過程データの解析

一つの標本レコードの統計的特徴を表す統計量として、以下があげられる。

- 平均と分散
- 確率密度関数
- 自己相関関数
- スペクトル密度関数

## 8.3 定常確率過程

確率過程  $\{x_k(t)\}$ ,  $-\infty < t < \infty$  (あるいは時系列) は確率密度関数によって特徴づけられる関数のアンサンブルである。それぞれの  $x_k(t)$  はこのアンサンブルの具現化されたものであり、標本関数、あるいはサンプル関数 (sample function) と呼ぶ。実用的には、 $x_k(t)$  は一つの観測時系列と考えられ、 $k$  は観測番号を意味する。また、標本関数のアンサンブルが定常確率過程を作り上げる。一つの  $x_k(t)$  のみで全体の確率過程  $\{x_k(t)\}$  を類推するのは、不適切である場合もあるが、エルゴード性が成立する場合は  $x_k(t)$  の時間平均から  $\{x_k(t)\}$  の特性に関する統計量を求めることができる。

平均

今、二つの確率過程  $\{x_k(t)\}, \{y_k(t)\}$  を考え、そのサンプル関数を  $x_k(t), y_k(t)$  とする。時間  $t$  における  $k$  に関するアンサンブル平均は、以下のように表される。

$$\mu_x(t) = E[x_k(t)] \quad (8.12a)$$

$$\mu_y(t) = E[y_k(t)] \quad (8.12b)$$

一般に異なる時間におけるこの平均値は、異なる値を取る。

$$\mu_x(t_1) \neq \mu_x(t_2) \quad \text{if } t_1 \neq t_2 \quad (8.13a)$$

$$\mu_y(t_1) \neq \mu_y(t_2) \quad \text{if } t_1 \neq t_2 \quad (8.13b)$$

共分散関数

時刻  $t_1 = t$  と  $t_2 = t + \tau$  共分散関数 (covariance functions) は以下のように定義される。

$$C_{xx}(t, t + \tau) = E[(x_k(t) - \mu_x(t))(x_k(t + \tau) - \mu_x(t + \tau))] \quad (8.14a)$$

$$C_{xy}(t, t + \tau) = E[(x_k(t) - \mu_x(t))(y_k(t + \tau) - \mu_y(t + \tau))] \quad (8.14b)$$

$$C_{yx}(t, t + \tau) = E[(y_k(t) - \mu_y(t))(y_k(t + \tau) - \mu_y(t + \tau))] \quad (8.14c)$$

一般にこれらは、 $t_1, t_2$  の異なる組み合わせによって異なる値を取る。  $\tau = 0$  の場合は、以下の通りとなる。

$$C_{xx}(t, t) = E[(x_k(t) - \mu_x(t))^2] = \sigma_x^2(t) \quad (8.15a)$$

$$C_{xy}(t, t) = E[(x_k(t) - \mu_x(t))(y_k(t) - \mu_y(t))] = C_{xy}(t) \quad (8.15b)$$

$$C_{yy}(t, t) = E[(y_k(t) - \mu_y(t))^2] = \sigma_y^2(t) \quad (8.15c)$$

よって、 $C_{xx}(t, t), C_{yy}(t, t)$  は時刻  $t$  における分散を、また  $C_{xy}(t, t)$  は時刻  $t$  における  $\{x_k(t)\}, \{y_k(t)\}$  の間の共分散を表す。定常な確率過程に関しては、共分散関数は、時間依存せず、 $\tau$  のみの関数になる。すなわち、 $C_{xx}(t, t + \tau) = C_{xx}(\tau), C_{xy}(t, t + \tau) = C_{xy}(\tau), C_{yy}(t, t + \tau) = C_{yy}(\tau)$  である。

### 自己相関関数・相互相関関数

定常な確率過程に関して、自己相関関数、相互相関関数 (cross correlation function) を以下のように定義される。

$$R_{xx}(\tau) = E[x_k(t)x_k(t + \tau)] \quad (8.16a)$$

$$R_{xy}(\tau) = E[x_k(t)y_k(t + \tau)] \quad (8.16b)$$

$$R_{yy}(\tau) = E[y_k(t)y_k(t + \tau)] \quad (8.16c)$$

ここで、 $R_{xx}, R_{yy}$  を自己相関関数、 $R_{xy}$  を相互相関関数と呼ぶ。もしも  $\mu_x = \mu_y = 0$  ならば、式 8.15 の  $C$  は  $R$  と等しくなる。一般には、

$$C_{xx} = R_{xx}(\tau) - \mu_x^2 \quad (8.17a)$$

$$C_{xy} = R_{xy}(\tau) - \mu_x\mu_y \quad (8.17b)$$

$$C_{yy} = R_{yy}(\tau) - \mu_y^2 \quad (8.17c)$$

である。

$\{x(t)\}$  が定常な確率過程であるならば、以下が成り立つ。

$$R_{xx}(-\tau) = R_{xx}(\tau) \quad (8.18a)$$

$$R_{yy}(-\tau) = R_{yy}(\tau) \quad (8.18b)$$

すなわち、定常な確率過程の自己相関関数は偶関数である。また、 $R_{xy}$  については以下が成り立つ。

$$R_{xy}(-\tau) = R_{yx}(\tau) \quad (8.18c)$$

また、 $R_{xx}(0)$  で除した値、

$$\rho_{xx} = \frac{R_{xx}(\tau)}{R_{xx}(0)}, \quad \rho_{yy} = \frac{R_{yy}(\tau)}{R_{yy}(0)}, \quad \rho_{xy} = \frac{R_{xy}(\tau)}{R_{xy}(0)} \quad (8.19)$$

を自己相関係数 (auto-correlation coefficient), 相互相関係数 (cross-correlation coefficient) と呼ぶ。

例：自己相関関数の例

$x(t)$	$R_{xx}(t)$
一定値 ( $x(t) = c$ )	$R_{xx}(\tau) = c^2$
正弦関数 ( $x(t) = A \sin[2\pi f_0 t + \theta]$ )	$R_{xx}(\tau) = \frac{A^2}{2} \cos 2\pi f_0 \tau$
白色雑音	$R_{xx}(\tau) = a\delta(\tau)$

一般に、自己相関関数は確率過程の「メモリーの持続特性」を表し、相互相関は一方に対する他方の「遅れ」を表す。

## 8.4 エルゴード性を持つ確率過程

エルゴード性を持つ確率過程では、アンサンブル平均が時間平均に等しく、統計量を時間平均から計算することができる。

平均

定常な確率過程  $x_k(t), y_k(t)$  の時間平均は、以下のように定義される。

$$\mu_x(k) = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T x_k(t) dt \quad (8.20a)$$

$$\mu_y(k) = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T y_k(t) dt \quad (8.20b)$$

分散関数・共分散関数

また、時間平均による分散関数は以下のように定義される。

$$\begin{aligned} C_{xx}(\tau, k) &= \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T [x_k(t) - \mu_x(k)][x_k(t + \tau) - \mu_x(k)] dt \\ &= R_{xx}(\tau, k) - \mu_x^2(k) \end{aligned} \quad (8.21a)$$

$$\begin{aligned} C_{yy}(\tau, k) &= \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T [y_k(t) - \mu_y(k)][y_k(t + \tau) - \mu_y(k)] dt \\ &= R_{yy}(\tau, k) - \mu_y^2(k) \end{aligned} \quad (8.21b)$$

また、共分散関数は以下のように定義される。

$$\begin{aligned} C_{xy}(\tau, k) &= \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T [x_k(t) - \mu_x(k)][y_k(t + \tau) - \mu_y(k)] dt \\ &= R_{xy}(\tau, k) - \mu_x(k)\mu_y(k) \end{aligned} \quad (8.21c)$$

## 自己相関関数・相互相関関数

また、ここで自己相関、相互相関関数は、以下のように定義される。

$$R_{xx}(\tau) = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T [x_k(t)x_k(t+\tau)]dt \quad (8.22a)$$

$$R_{xy}(\tau) = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T [x_k(t)y_k(t+\tau)]dt \quad (8.22b)$$

$$R_{yy}(\tau) = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T [y_k(t)y_k(t+\tau)]dt \quad (8.22c)$$

エルゴード性を持つ定常確率過程では、以下のようにアンサンブル平均と時間平均が等しい。

$$\mu_x(k) = \mu_x \quad \mu_y(k) = \mu_y \quad (8.23a)$$

$$C_{xx}(\tau, k) = C_{xx}(\tau) \quad C_{xy}(\tau, k) = C_{xy}(\tau) \quad C_{yy}(\tau, k) = C_{yy}(\tau) \quad (8.23b)$$

## 8.5 スペクトル密度関数

スペクトル密度関数 (spectral density function)、あるいはスペクトル (spectra) は、元の信号にどの周波数の成分が存在しているかを示す。以下の3つの方法で定義・計算することができる

- 相関関数より。
- 有限フーリエ変換より
- フィルター・四角化・平均化操作より

ここでは、最初の2つについて説明する。

### 8.5.1 相関関数からのスペクトル

確率過程  $\{x_k(t)\}, \{y_k(t)\}$  は、平均0の確率過程であるとする。このとき、スペクトル・コスペクトルは自己相関関数・相互相関関数のフーリエ変換として定義される。

$$S_{xx}(f) = \int_{-\infty}^{\infty} R_{xx}(\tau) e^{-i2\pi f\tau} d\tau \quad (8.24a)$$

$$S_{yy}(f) = \int_{-\infty}^{\infty} R_{yy}(\tau) e^{-i2\pi f\tau} d\tau \quad (8.24b)$$

$$S_{yx}(f) = \int_{-\infty}^{\infty} R_{xy}(\tau) e^{-i2\pi f\tau} d\tau \quad (8.24c)$$

$S_{xx}, S_{yy}$  をスペクトル密度関数 (スペクトル、(auto)spectral density function)、 $S_{xy}$  をクロススペクトル密度関数 (クロススペクトル、cross-spectral density function) と呼ぶ。 $x(t), y(t)$  が実数関数ならば、以下のように簡略化される。

$$S_{xx}(f) = \int_{-\infty}^{\infty} R_{xx}(\tau) \cos 2\pi f\tau d\tau = 2 \int_0^{\infty} R_{xx}(\tau) \cos 2\pi f\tau d\tau \quad (8.25a)$$

$$S_{yy}(f) = \int_{-\infty}^{\infty} R_{yy}(\tau) \cos 2\pi f\tau d\tau = 2 \int_0^{\infty} R_{yy}(\tau) \cos 2\pi f\tau d\tau \quad (8.25b)$$

$$S_{xy}(f) = \int_{-\infty}^{\infty} R_{xy}(\tau) \cos 2\pi f\tau d\tau \quad (8.25c)$$

$R_{xx}, R_{yy}$  は偶関数であるので、

$$S_{xx}(-f) = S_{xx}(f) \quad S_{yy}(-f) = S_{yy}(f) \quad S_{xy}(-f) = S_{yx}(f) \quad (8.26)$$

よって、スペクトルは偶関数である。式 8.24 の逆フーリエ変換は以下の通りである。

$$R_{xx}(\tau) = \int_{-\infty}^{\infty} S_{xx}(f) e^{i2\pi f\tau} df \quad (8.27a)$$

$$R_{yy}(\tau) = \int_{-\infty}^{\infty} S_{yy}(f) e^{i2\pi f\tau} df \quad (8.27b)$$

$$R_{yx}(\tau) = \int_{-\infty}^{\infty} S_{xy}(f) e^{i2\pi f\tau} df \quad (8.27c)$$

これも  $x(t), y(t)$  が実数であれば、上記と同様に簡略化され、

$$R_{xx}(\tau) = \int_{-\infty}^{\infty} S_{xx}(f) \cos 2\pi f\tau df = 2 \int_0^{\infty} S_{xx}(f) \cos 2\pi f\tau df \quad (8.28a)$$

$$R_{yy}(\tau) = \int_{-\infty}^{\infty} S_{yy}(f) \cos 2\pi f\tau df = 2 \int_0^{\infty} S_{yy}(f) \cos 2\pi f\tau df \quad (8.28b)$$

$$R_{yx}(\tau) = \int_{-\infty}^{\infty} S_{xy}(f) \cos 2\pi f\tau df \quad (8.28c)$$

### 8.5.2 有限フーリエ変換からのスペクトル

$x_k(t), y_k(t)$  の有限フーリエ変換を

$$\tilde{x}_k(f) = \int_0^T x_k(t) e^{-i2\pi ft} dt \quad (8.29)$$

$$\tilde{y}_k(f) = \int_0^T y_k(t) e^{-i2\pi ft} dt \quad (8.30)$$

と表すと、スペクトルは以下のように定義される。

$$S_{xy} = \lim_{T \rightarrow \infty} \frac{1}{T} \tilde{x}_k^*(f) \tilde{y}_k(f) \quad (8.31)$$

ただし、\* は複素共役を表す。式 8.24 と式 8.31 が同等であることは、数学的に証明できる。

### 8.5.3 スペクトルの性質

式 8.27 において、 $\tau = 0$  とすれば左辺は分散・共分散となる。よって、

$$\sigma_x^2 = \int_{-\infty}^{\infty} S_{xx}(f) df \quad (8.32)$$

$$\sigma_y^2 = \int_{-\infty}^{\infty} S_{yy}(f) df \quad (8.33)$$

$$C_{xy} = \int_{-\infty}^{\infty} S_{xy}(f) df \quad (8.34)$$

スペクトル・クロススペクトルは、分散・共分散の周波数空間の分解であり、各周波数成分の分散・共分散への寄与を表す。

## 演習

---

### 演習 8.1

---

10 個の長さ 200 のデータ列がある<sup>1</sup>。それぞれは、標本レコードであり、ほぼ同時に観測された鉛直風速の時間データである。これをアンサンプルとみて、以下の計算を行え。

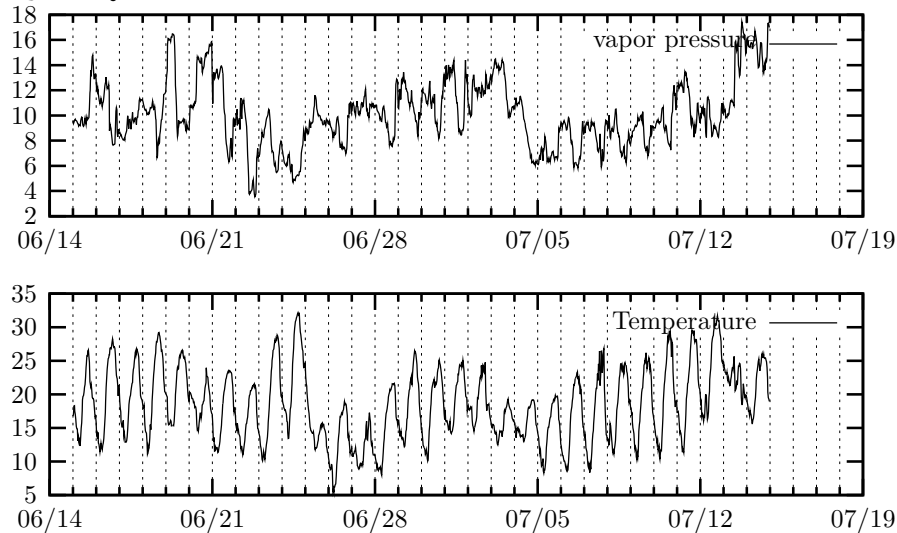
1. 平均をアンサンプル平均として計算し、標本レコードとともにグラフにプロットせよ。
2. 自己相関関数  $R_{xx}(t_1, t_1 + \tau)$  をアンサンプル平均 (ただし  $\tau = 20$  まで) を求めよ
3. 最初の標本レコードについて自己相関関数  $R_{xx}(\tau, 1)$  を時間平均として求めよ。また、上のアンサンプル平均とともにプロットし、比較せよ。

---

### 演習 8.2

---

モンゴル高原における 2003/6/15 から 1 ヶ月間の 30 分平均気温 ( $x(t)$ , degC)、湿度 ( $y(t)$ , hPa) を与える<sup>2</sup>。



1. 自己相関関数、相互相関関数をもとめ、グラフにせよ。
2. スペクトル  $S_{xx}(f)$ ,  $S_{yy}(f)$  をもとめよ。

---

<sup>1</sup><http://www.suiri.tsukuba.ac.jp/~asanuma/courses/currnet/geostats/>

<sup>2</sup><http://www.suiri.tsukuba.ac.jp/~asanuma/courses/currnet/geostats/>

## 第9章 分散分析 (ANOVA, Analasys of Variance)

第5章では、2つの異なる母集団のそれぞれから得られた標本についての統計的な推論、検定について議論した。ここでは、2つ以上の標本があるときを考える。これは、複数の母集団がありそれぞれから標本を取得するときや、同じ実験標本に対して異なる処理 (treatment) をした結果などが考えられる。具体的には、以下のような例がある。

例 9.1

5種類のガソリンがあり、種類によって燃費 (km/ℓ) にどのような違いが出てくるかを実験的に調べたい。

例 9.2

化学工場で製品中のある成分の含有量を多くするため、反応温度を、 $50^{\circ}\text{C}$ ,  $55^{\circ}\text{C}$ ,  $60^{\circ}\text{C}$ ,  $65^{\circ}\text{C}$  の4段階に変えて、製品 1kg 中の含有量を測定した。

このように、標本毎の違いを作り出す異なる実験処理などを因子 (factor)、また因子に与える条件を水準 (level) と呼ぶ。例 9.1 の例では、因子はガソリンの種類であり因子は1つで5種類なので水準は5つ、すなわち単因子5水準である。例 9.2 は単因子4水準である。

上記の例 9.1 の場合は因子は定性的であるのに対し、例 9.2 では定量的な因子となっている。因子が定量的である場合は、第6章で取り扱う回帰分析を適用することができる。

分散分析の目的は、複数ある因子水準のそれぞれの因子水準平均  $\mu_i (i = 1, \dots, I)$  が、母集団ごとの違い、すなわち因子水準によって生じる違いがあるかどうかを判別することである。帰無仮説  $H_0$  として  $\mu_i$  がすべて等しいこととし、 $F$  検定によって  $H_0$  が棄却されるかどうかを判別し、 $H_0$  が棄却された場合は、さらにどの  $\mu_i$  同士に優位な違いがあるかを探し出すことになる。ここでは一因子分散分析を例にとって解説する。

### 9.1 一因子分散分析 (Single-Factor ANOVA)

#### 9.1.1 定義

まず、 $I$  を水準の数とする。例 9.3 では、 $I = 4$  である。ここで、 $\mu_i$  を各母集団の平均値でとすると、帰無仮説は、 $\mu_i$  がすべて等しい、すなわち

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_I$$

である。対立仮説はすなわち、

$$H_a : \text{少なくとも2つの } \mu_i \text{ が異なる}$$

である。例 9.3 に示すように  $X_{i,j}$  を  $i$  番目の水準からの  $j$  番目の観測値の確率変数、 $x_{i,j}$  を  $X_{i,j}$  の観測値を表すものである。 $X_{i,j}$  を簡略して  $X_{ij}$  と書くこともある。また、 $n_i$  を各水準毎の観測数（サンプル数）とし、観測は全部で  $n_T = \sum_{i=1}^I n_i$  個となる。

### 例 9.3

ある会社で朝食用のシリアルの新パッケージの市場調査を行った。売り上げが等しい、10 件の小売店を選択し、4 つのパッケージの 1 つをランダムに割り当てた。2 つのパッケージにそれぞれ 3 つの小売店を、2 つのパッケージにそれぞれ 2 つの小売店を割り当てた。パッケージ以外の、値段や商品内容は同じである。調査期間中のそれぞれのパッケージの売り上げは以下の通りである。

パッケージ番号	売り上げ		計	平均	小売り店数
$i$	$x_{ij}$			$\bar{x}_{i\cdot}$	$n_i$
1	12	18	30	15	2
2	14	12 13	39	13	3
3	19	17 21	57	19	3
4	24	30	54	27	2

$i$  番目の水準のサンプルの平均は

$$\bar{X}_{i\cdot} = \frac{\sum_{j=1}^{n_i} X_{ij}}{n_i} \quad (i = 1, \dots, I) \quad (9.1)$$

で表される。 $\bar{X}_{i\cdot}$  の  $\cdot$  は二つめの添え字について、加算を行ったことを表す。また、すべての観測値についての平均は

$$\bar{X}_{..} = \frac{\sum_{i=1}^I \sum_{j=1}^{n_i} X_{ij}}{n_T} \quad (9.2)$$

で表され、一般平均（grand mean）と呼ぶ。また、

$$S_i^2 = \frac{\sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{i\cdot})^2}{n_i - 1} \quad (i = 1, \dots, I) \quad (9.3)$$

は  $i$  番目のサンプルの中での分散である。

## 9.1.2 前提条件

一因子分散分析における前提条件は、5.1.2 節における合併統計値による  $t$  検定を行ったときの拡張であり、 $I$  個の母集団が  $\sigma^2$  を分散とする正規分布となることを前提とする。

— 因子分散分析における前提条件 —

$I$  個の水準のそれぞれに対応する母集団が正規分布に従い、その分散が  $\sigma^2$  に等しい。すなわち  $X_{ij}$  は正規分布に従い、

$$E(X_{ij}) = \mu_i, \quad V(X_{ij}) = \sigma^2$$

である。

この前提条件は、 $X_{ij}$  を以下のように考えることと等しい。

$$X_{ij} = \mu_i + \epsilon_{ij} \quad (i = 1, \dots, I; j = 1, \dots, n_i) \quad (9.4)$$

ここで、 $\epsilon_i$  は、平均が 0、分散が  $\sigma^2$  の正規分布 ( $N(0, \sigma^2)$ ) に従う。また、 $\mu_., \tau_i$  を

$$\mu_ = \frac{\sum_{i=1}^I n_i \mu_i}{n_T} \quad (9.5)$$

$$\tau_i = \mu_i - \mu_ . \quad (9.6)$$

と定義する。 $\tau_i$  は、第  $i$  水準の効果 (effect) と呼ばれる。式 9.4 は以下のように表される。

$$X_{ij} = \mu_ . + \tau_i + \epsilon_{ij} \quad (9.7)$$

これは、 $X_{ij}$  が、すべての観測値に共通の平均 ( $\mu_ .$ ) と第  $i$  水準の効果 ( $\tau_i$ )、ランダムな誤差 ( $\epsilon_{ij}$ ) の和として表されることになる。

### 9.1.3 ANOVA 表

一般平均からの変動値

$$X_{ij} - \bar{X}_{..} \quad (9.8)$$

は、水準毎の平均  $\bar{X}_{i.}$  を用いて以下のように表される。

$$X_{ij} - \bar{X}_{..} = (\bar{X}_{i.} - \bar{X}_{..}) + (X_{ij} - \bar{X}_{i.}) \quad (9.9)$$

式 9.9 で右辺第 1 項は全体の平均からの各水準平均の推定値の変動、右辺第 2 項は各水準平均の推定値からの変動である。両辺二乗して、すべての観測値についての和を取ると、以下の式を得る。

$$\sum_i \sum_j (X_{ij} - \bar{X}_{..})^2 = \sum_i n_i (\bar{X}_{i.} - \bar{X}_{..})^2 + \sum_i \sum_j (X_{ij} - \bar{X}_{i.})^2 \quad (9.10)$$

式 9.10 の左辺は、一般平均からの総変動量の二乗和であり、総平方和 (Total sum of squares) と呼ばれ、 $SST$  と表される (式 6.25 参照)。

$$SST = \sum_i \sum_j (X_{ij} - \bar{X}_{..})^2 \quad (9.11)$$

表 9.1: 一因子 ANOVA の ANOVA 表

要因	平方和	自由度	二乗平均	二乗平均の期待値	F 値
水準間	$SSTr$	$I - 1$	$MSTr$	$\sigma^2 + \frac{1}{I-1} \sum n_i (\mu_i - \mu.)^2$	$MSTr/MSE$
誤差	$SSE$	$n_T - I$	$MSE$	$\sigma^2$	
計	$SST$	$n_T - 1$			

式 9.10 の右辺第 1 項、第 2 項はそれぞれ、級間平方和 (Treatment sum of squares)、誤差平方和 (Error sum of squares) と呼ばれ、 $SSTr$ 、 $SSE$  とあらわされる。

$$SSTr = \sum_i n_i (\bar{X}_{i.} - \bar{X}_{..})^2 \quad (9.12)$$

$$SSE = \sum_i \sum_j (X_{ij} - \bar{X}_{i.})^2 \quad (9.13)$$

すなわち、式 9.10 は以下のように書き直される。

$$SST = SSTr + SSE \quad (9.14)$$

式 9.14 は、総平方和  $SST$  が、各水準内のばらつき ( $SSE$ ) と各水準平均推定量の間のばらつき ( $SSTr$ ) に分けられる。 $SST$ ,  $SSE$ ,  $SSTr$  は実用的には以下の式から計算できる。

$$SST = \sum_i \sum_j X_{ij}^2 - n_T \bar{X}_{..}^2 \quad (9.15)$$

$$SSTr = \sum_i n_i \bar{X}_{i.}^2 - n_T \bar{X}_{..}^2 \quad (9.16)$$

$$SSE = \sum_i \sum_j X_{ij}^2 - \sum_i n_i \bar{X}_{i.}^2 \quad (9.17)$$

また、 $SST$ ,  $SSTr$ ,  $SSE$  それぞれの自由度は、 $SST$  は平均からの偏差を取った時点で自由度が 1 経るので自由度  $n_T - 1$  となる。また、 $SSTr$ ,  $SSE$  はそれぞれ、自由度  $I - 1$ ,  $n_T - I$  である。よって、二乗平均は以下のように定義できる。

$$MSTr = \frac{SSTr}{I - 1} \quad (9.18)$$

$$MSE = \frac{SSE}{n_t - I} \quad (9.19)$$

それぞれ、級間二乗平均 (Treatment mean square)、誤差二乗平均 (Error mean square) である。以上をまとめて記したものが ANOVA 表である。例 9.3 のパッケージの市場調査のデータについての実例を表 9.2 に示す。

表 9.2: 例 9.3 の ANOVA 表

要因	平方和	自由度	二乗平均	F 値
水準間	258	3	86	11.2
誤差	46	6	7.67	
計	304	9		

### 9.1.4 ANOVA 検定

$MSTr, MSE$  のそれぞれの期待値は以下ようになる。

$$E(MSE) = \sigma^2 \quad (9.20)$$

$$E(MSTr) = \sigma^2 + \frac{\sum n_i(\mu_i - \mu.)^2}{I - 1} \quad (9.21)$$

よって、 $MSE$  は無条件で、 $X_{ij}$  の誤差項  $\epsilon_{ij}$  (式 9.4) の分散  $\sigma^2$  の普遍推定量である。その一方で、 $MSTr$  はすべての  $\mu_i$  が等しければ、すなわち  $H_0$  が棄却されなければ、 $MSTr$  は  $\sigma^2$  の普遍推定量となる。しかしながら、2 つ以上の  $\mu_i$  が等しく無ければ、 $MSTr$  は  $\sigma^2$  を過大評価する。この性質を用いて、 $\mu_i$  がすべて等しいかどうかの統計検定を構築することができる。すなわち、 $MSTr$  が  $MSE$  と同じ程度の値ならば、因子水準平均  $\mu_i$  はすべて等しく、 $MSTr$  が  $MSE$  よりも十分に大きければ、因子水準平均  $\mu_i$  にばらつきがあると考えられる。

前述したように、帰無仮説は、

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_I$$

で、対立仮説は

$$H_a : \text{少なくとも 2 つの } \mu_i \text{ が異なる}$$

である。検定統計量は

$$F = \frac{MSTr}{MSE} \quad (9.22)$$

である。帰無仮説  $H_0$  が真ならば、 $F$  は  $F$  分布  $F(I - 1, n_T - I)$  に従う。よって、ANOVA 検定は以下の通りとなる。

ANOVA 検定

帰無仮説	$H_0 : \mu_1 = \mu_2 = \dots = \mu_I$
対立仮説	$H_a : \text{少なくとも 2 つの } \mu_i \text{ が異なる}$
検定統計量	$f = \frac{MSTr}{MSE}$
棄却域	$F_{1-\alpha, I-1, n_T-I} < f$

例 9.3 の場合は、 $f = 11.2 > F_{0.95, 3, 6} = 4.76$  であるので、 $H_0$  は棄却された。

## 演習

### 演習 9.1

4 つの異なる水泳指導方法の効果を明らかにするために、初級の水泳の講義で実験を行った。4 つの指導方法は、a) 口頭指導、b) 口頭とビデオ指導、c) ビデオ指導、d) 何もなし、である。24 人の学生をランダムに 4 つのグループに分け、講義の最後に 8m の水泳の時間を計測した結果を以下に示す (Devore, 1991, Ex 10.1)。

指導法		時間 (秒)							標本平均	標本分散
	i	$x_{ij}$							$\bar{x}_{i\cdot}$	$S_i^2$
指導法 a	1	18.7	21.1	17.9	19.5	22.1	18.3		19.60	2.78
指導法 b	2	19.9	17.6	18.2	20.0	16.9	17.5		18.35	1.71
指導法 c	3	18.6	20.3	21.7	19.7	20.9	20.8		20.33	1.16
指導法 d	4	19.1	18.9	18.4	18.8	17.7	20.5		18.90	0.86

このデータを用いて、以下を計算せよ

1. このデータを用いて ANOVA 表を作成せよ
2.  $f$  値を計算し、有意水準 0.05 で検定せよ

## 関連図書

- Bendat, J.S. and Piersol, A.G. (1971): Random Data Analysis and Measurement Procedure. Wiley Inter-Science. 594 pp.
- Devore, J.L. (1991): Probability and Statistics for Engineering and the Sciences. Brooks/Cole Pub. Co., 3rd edition. 716 pp.
- Helsel, D. and Hirsch, R. (1993): Statistical Methods in Water Resources. Elsevier Sci Pub. Co.
- Kermack, K.A. and Haldane, J.B.S. (1950): Organic correlation and allometry. *Biometrika*, **37**(1/2), pp. 30–41
- Kritsky, S.N. and Menkel, J.F. (1968): Some statistical methods in the analysis of hydrologic data. *Soviet Hydrology, Selected Papers*, **1**, pp. 80–98
- Kruskal, W.H. (1953): On the uniqueness of the line of organic correlation. *Biometrics*, **9**(1), pp. 47–58
- Neter, J., Wasserman, W., and Kutner, M.H. (1990): Applied Linear Statistical Models. Irwin Inc., 3rd edition
- Tennekes, H. and Lumley, J.L. (1972): A First Course in Turbulence. The MIT Press, Cambridge, MA. 300 pp.
- 日野幹雄 (1977): スペクトル解析. 朝倉書店
- 東京大学教養学部統計学教室 (編) (1991): 統計学入門. 基礎統計学 I. 東京大学出版会
- 東京大学教養学部統計学教室 (編) (1992a): 自然科学の統計学. 基礎統計学 III. 東京大学出版会
- 東京大学教養学部統計学教室 (編) (1992b): 人文・社会科学の統計学. 基礎統計学 II. 東京大学出版会