確率統計の演習問題

目次

確率統計の演習問題です。原本は「統計学演習」村上正康、安田正實；培風館（ISBN 978-4-563-00870-3） のテキストです。この中から選んでいますし、変更や追加もしています。説明不足の箇所は本文テキストを参照。また解答は原本を参照のこと、Web ページでは見られません。問番号にある [ref:x.y] は x 章の問題番号 y や, [ref:x.(z)]x 章の例題 z の意味です。

1 データの整理

調査対象として数値データやアンケートを集計する集団、あるいはさまざまな実験を考える母集団 (popilation) という。集められた数値データや集計値、資料結果、特性を表すものをデータ変量 (data variable) とよぶ。データの特性を表す単位については、その尺度構造を理解しておく。

(I) 項目データ（質的データ、カテゴリーデータ）:

I-1 名義尺度： 単なる区別分類するために割り当てた名称、番号や数値を名義尺度データいう。 例えば、血液型や男女の性別、出身地の都道府県名、個人の学籍番号、電話番号など。

I-2 順序尺度： 順序尺度データとは、割り当てたデータの大小比較はできるが、間隔や比率には意味がないもの、和や差あるいは掛け算、割り算の計算をしても意味をもたない。例. アンケート結果の評価数値、サービスの満足度の値には、大きさを比較するには差が意味をもつが、割合には意味がない。

(II) 量的データ（定量的データ、数値データ）:

II-1 間隔尺度： 数値の和や差は意味をもつが、比率には意味がないもの。たとえば天候での温度変化には、差は意味があるが１０％暑いとか寒いなどとは言わない、接客の満足度の値

II-2 比例尺度： 通常の数値データとして取り扱うもの。データの変数値を和、差、さらに掛け算、割り算に対しても意味がある。

これらのデータ構造の分類について、明確にならない部分も多くあるが、注意すべき点は、得られた統計のデータ値をまとめ縮約するときに、その特性を明示するためには、データの特性をはっきりとつかんで、グラフ表現や統計数値の情報を表すことが大切です。代表的な中心位置を取りまとめるに用いられる 「最頻値（モード）、中央値（メディアン）、平均値」を計算して意味があるかどうかを判断し認識しておかねばならないことです。さらには拡がりの目安としての「標準偏差、分散、四分位数」などを意味があるかどうか認識しておく。

項目データ（カテゴリーデータ）

◎名義尺度
→［○］度数，最頻値
→［×］平均，分散，標準偏差
◎順序尺度　大小比較は可能であるが，間隔や比率には意味がない.
→［○］度数，最頻値，中央値
→［×］平均，分散，標準偏差

[基本の用語]

階級 (class)、階級値 (center value of class)、度数 (frequency)、度数分布 (階級別に分けた度数と対応する個数のペア)、度数分布表 (frequency table、度数分布を表あるいはグラフで表したもの)、相対度数（各々の個数を総和数で割り、比率にした値)、相対度数分布表、ヒストグラム（histogram, グラフではないことに注意）

統計量 (statistics) [データの集まりから、集計、分析のために計算される計算式を意味する。数値として得られたものは統計値という]

各個体のデータ値は数直線上の点で表され、その個体の数（データサイズ、データの大きさ）が n であるとき、データは $\{x_i; i = 1, 2, \cdot, n\}$ と表す。

平均値 （データ値の総和を個数で割った数値、たんに平均と言ってもよいが、計算した結果の数値という意味をもたしている）

$$\overline{x} = \frac{1}{n}\sum_i x_i = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$$

中央値（メディアン）　階級別に小さいものから大きいものへと並べ直して、数値がちょうど中央に順位が属する階級の値)。データ $\{x_i; i = 1, 2, \cdot, n\}$ を大きさに並び替えて、

$$x_{(1)} \leq x_{(2)} \leq x_{(3)} \leq \cdots \leq x_{(n-1)} \leq x_{(n)}$$

とすると、もし n が奇数であれば、$k = \dfrac{n-1}{2}$ とし、この番号に対応する値を $\mathrm{Med}(x) = x_{(k)}$, 偶数であれば, $k = \dfrac{n}{2}$ として、$\mathrm{Med}(x) = x_{(k)}$

最頻値（モード）　階級別に整理した度数の値がもっともおおきい度数の階級値、もし双峰形などであるときには、「モードがない」と考えるか、集団を分けてデータの特徴を整理する方法が考えられる。

四分位数 （中央にくる値で 2 つに分けて（中央値の定義)、それぞれをさらにデータ個数で 2 つ分けて、計 4 つに分割する。これらを小さい順に、第一四分位数 $Q_1(x)$、第二四分位数（中央値のこと）$\mathrm{Med(x)} = \mathrm{Q_2(x)}$、第三四分位数 $\mathrm{Q_3}(x)$ をよぶ）箱型にし、最小値と最大値を直線で伸ばしたものを箱ひげ図（ボックスチャート）という。

範囲（レンジ)、最大値、最小値

$$\text{最大値と最小値}：\mathrm{Max(x)} = \max_{\mathrm{i}}\{\mathrm{x_i}\} = \mathrm{x_{(n)}}, \quad \mathrm{Min(x)} = \min_{\mathrm{i}}\{\mathrm{x_i}\} = \mathrm{x_{(1)}}$$

$$\text{範囲}：\mathrm{R}(x) = \mathrm{Max(x)} - \mathrm{Min(x)} = \max_{\mathrm{i}}\{\mathrm{x_i}\} - \min_{\mathrm{i}}\{\mathrm{x_i}\}$$

四分位範囲 (quantile range)

$$\mathrm{Q}_R(x) = \mathrm{Q_3}(x) - \mathrm{Q_1}(x)$$

分散、標本分散 (sample variance)　各データについて、平均値 $\overline{x}$ との偏差をもとめ、2 乗した値 $(x_i - \overline{x})^2; i =$

$1, 2, \cdots, n$ に対して, その平均値を計算したもの。

$$\begin{aligned} s_x^2 &= \frac{1}{n}\sum_i (x_i - \overline{x})^2 = \frac{1}{n}\sum_i x_i^2 - \overline{x}^2 \\ &= (\text{データ値と平均値との 2 乗偏差}) \text{ の平均} = (\text{2 乗値の平均}) - (\text{平均値の 2 乗}) \end{aligned}$$

不偏分散 (unbiansed variance) データ間相互の偏差を 2 乗して総和をとった値、入れ替えても同じであるから、差の 2 乗値を半分にしたものと同じ。

$$\begin{aligned} u_x^2 &= \frac{1}{n-1}\sum_i (x_i - \overline{x})^2 = \frac{1}{n}\sum_i x_i^2 - \overline{x}^2 \\ &= \frac{1}{\binom{n}{2}} \sum_{1 \le i < j \le n} \frac{(x_i - x_j)^2}{2} \\ &= (\text{データの相互偏差 2 乗値の半数を平均したもの}) \end{aligned}$$

標準偏差 (standard deviation, SD) 分散の平方根、$s_x = \sqrt{s_x^2}$ (正の値とする) 分散は 2 乗をするが、標準偏差はその平方根であるから、平均と同じ単位になり、加法の演算「(平均) ±(標準偏差)」は意味を持つが、(平均) + (分散) は意味がない。

二変量データ、多変量データ： 母集団を構成する要素を個体といい、その個体がベクトルとして表現されるときをいう。変数とは言わずに、統計の数値では、変量 (statistical variable) という。

n 個のデータから統計量 T を対応させる： $\quad (x_1, x_2, \cdots, x_n) \quad \to \quad T = T(x_1, x_2, \cdots, x_n)$

個体を表す数値が平面空間の場合が二変量データであり、一般の場合を多変量という。各個体のデータ値が 2 次元のベクトルで、その個体の数 (データサイズ、データの大きさ) が n であるとき、データは $(x_i, y_i); i = 1, 2, \cdot, n$ と表す。

$$\text{変量 } x \text{ の平均：} \quad \overline{x} = \frac{1}{n}\sum_i x_i = \frac{S_x}{n} = \frac{1}{n}(x_1 + x_2 + x_3 + \cdots + x_n)$$

,

$$\text{変量 } y \text{ の平均：} \quad \overline{y} = \frac{1}{n}\sum_i y_i = \frac{S_y}{n} = \frac{1}{n}(y_1 + y_2 + y_3 + \cdots + y_n)$$

x の 2 乗和 (sum of square) とは $S_{xx} = \sum_i x_i^2 = x_1^2 + x_2^2 + x_3^2 + \cdots + x_n^2$. また x, y の積和とは $S_{xy} = \sum_i x_i y_i = x_1 \cdot y_1 + x_2 \cdot y_2 + x_3 \cdot y_3 + \cdots + x_n \cdot y_n$ などと表すこともある。

散布図 (scattering figure) x-y 平面のうえに描かれたデータ $(x_i, y_i); i = 1, 2, \cdots, n$ をプロットしたもの。

共分散 (covatiance)

$$\mathrm{Cov}(x, y) = \frac{1}{n}\sum_i (x_i - \overline{x})(y_i - \overline{y}) = \frac{1}{n} S_{xy} - \overline{x}\,\overline{y}$$

相関係数 (corrlation coefficinet) 後述の第 7 節「相関と回帰」を参照せよ。

問 1.1 (ref:1.(4))**.** ある日、図書館で本を借りた人の年齢を調べたところ、

年齢	*8-12*	*13-20*	*21-60*	*61-64*
人数	9	22	35	14

(1) このデータをヒストグラムで表せ。 (2) この日の集計結果における年齢の平均と標準偏差をもとめよ。

問 1.2 (ref:1.(6))**.** つぎの *8* 個のデータ*:* $6, 8, 8, 10, 21, x, y, z$、ただし *(*x, y, z は未知*)* は、平均が 7 で、モード（最頻値）が 6 であるという。このときメディアン（中央値、中位数）はいくつか？

問 1.3. あるデータを集計したところ、度数分布表はつぎの結果 (i) となった。

(1) この (i) の表から、その標本平均と標本分散、標準偏差を求めよ。

(2) また (ii) の表における場合の計算を (i) の結果から計算せよ。

(i)

値	*110*	*120*	*130*	*140*	*150*	計
度数	2	6	3	7	2	20

(ii)

値	*10*	*20*	*30*	*40*	*50*	計
度数	2	6	3	7	2	20

問 1.4. 2 つのグループ *A,B* のうち、A は大きさ 4、平均 5*,* 分散 2 であり、B は大きさ 6*,* 平均 7*,* 分散 3 であった。この *2* つを合体させたとき、平均と分散をもとめよ。

問 1.5. 3 個のデータ $\{x_1, x_2, x_3\}$ について 、平均値を $\overline{x}$ とするとき

$$\frac{1}{3}\left[\frac{(x_1-x_2)^2}{2}+\frac{(x_1-x_3)^2}{2}+\frac{(x_2-x_3)^2}{2}\right]=\frac{x_1^2+x_2^2+x_3^2}{3}-\overline{x}^2$$

を示せ。またデータの個数が $n \geq 3$ の場合でも成り立つことを確かめよ。

2 確率と確率分布

確率の基本事項

事象の記号: 全事象 Ω, 空事象 ϕ;

(i) 和事象 (union): A or B, $A \cup B$, (ii) 積事象 (intersection); A and B, $A \cap B = AB$, (iii) 補事象、余事象 (complement) ; negation, not A ; A^c, $\overline{A}$

加法性:

$$\begin{aligned} P(A\cup B) &= P(A)+P(B)-P(AB)=P(A\overline{B})+P(\overline{A}B)+P(AB) \\ P(A\cup B\cup C) &= P(A)+P(B)+P(C)-P(AB)-P(BC)-P(CA)+P(ABC) \\ &= P(A\,B\,C)+P(A\,B\,\overline{C})+P(A\,\overline{B}\,C)+P(\overline{A}\,B\,C) \\ &\quad +P(A\,\overline{B}\,\overline{C})+P(\overline{A}\,B\,\overline{C})+P(\overline{A}\,\overline{B}\,C) \end{aligned}$$

問 2.1. 2 つの事象 A, B の確率が $P(A\cup B)=\frac{7}{8}$, $P(A\cap B)=\frac{1}{4}$, $P(\overline{A})=\frac{5}{8}$, と与えられている。つぎの

確率をもとめよ。 (i) $P(A)$ (ii) $P(\overline{B})$ (iii) $P(A \cap B)$ (iv) $P(A \cup \overline{B})$

問 2.2. 事象 A , B は互いに排反 *(*同時に起ることはない*)* 事象で $P(A) = 0.2$, $P(B) = 0.8$ で、さらにある事象 C の条件つき確率が $P(C \,|\, A) = 0.4$ と $P(C \,|\, B) = 0.5$ で与えられている。$P(A \,|\, C)$ と $P(B \,|\, C)$ をもとめよ。答えは分数のままでもよい。

問 2.3. 全事象 $S = \{1, 2, 3, 4\}$ とし、それぞれ 4 個の点に対して、同じ確率 $\frac{1}{4}$ ずつをもつとする。3 つの事象 $A = \{1, 2\}$, $B = \{1, 3\}$, $C = \{1, 4\}$ において、(i) $P(AB) = P(A)P(B)$, $P(AC) = P(A)P(C)$, $P(BC) = P(B)P(C)$ を示せ。 (ii) しかし、3 つが同時に起こる積事象の確率 $A \cap B \cap C = ABC$ はそれぞれの確率の積にはならない。すなわち

$$P(A \cap B \cap C) \neq P(A)P(B)P(C)$$

期待値、平均と分散

(i) $\mu_X = E[X] = \sum_k k\, p_X(k)$
(ii) $\sigma_X^2 = V[X] = E\left[(X - \mu_X)^2\right] = \sum_k (k - \mu_X)^2\, p_X(k) = \sum_k k^2\, p_X(k) - (\mu_X)^2$
(iii) 確率変数の関数 $h(X)$ に関する期待値は

$$E[h(X)] = \sum_k h(k)\, p_X(k)$$

(iv) 2 変数の場合 $f_{X,Y}(i, j) = P(\{X = i\} \cap \{Y = j\})$ とするとき

$$E[h(X, Y)] = \sum_i \sum_j h(i, j)\, f_{X,Y}(i, j)$$

(v) 2 つの和 $X + Y$ の期待値は $h(X, Y) = X + Y$ で、独立であっても、独立でなくても、

$$E[X + Y] = E[X] + E[Y]$$

であり、一般に線形式では $E[aX_1 + bX_2 + \cdots + cX_n] = aE[X_1] + bE[X_2] + \cdots + cE[X_n]$ となる。
(vi) 積の期待値では、もし独立であれば、$f_{X,Y}(i, j) = P(X = i) \cdot P(Y = j) = f_X(i) f_Y(i)$ より、

$$E[XY] = \sum_i \sum_j (i \times j)\, f_{X,Y}(i, j) = E(X) \times E(Y)$$

を得るが、期待値が積になっていても、もとの確率変数は独立であるとは限らない。

問 2.4. 確率変数 X の平均（期待値） が $E(X) = m$, 分散が $V(X) = s^2$ であるという。つぎの確率変数 (a) $Y = 3X + 2$, (b) $Z = X - m$, (c) $W = \dfrac{X + b}{a}$ （ここで a, b は定数*)* について、それぞれの平均と分散をもとめよ。

問 2.5. 離散型確率変数 (X, Y) の結合分布 : $f(x, y) = P(X = x, Y = y)$ がつぎで与えられている。

$$f(1,1)=1/8,\ f(1,2)=1/4,\ f(2,1)=1/8, f(2,3)=3/8,\ f(2,4)=1/8,$$

その他の値では 0。このとき、(1) 和 $X+Y$ の平均をもとめよ。(2) 積 XY の平均をもとめよ。(3) この 2 つの確率変数独立となるかどうかを調べよ。

問 2.6 (ref:2.(9))**.** 箱の中に 4 個の赤球 と 3 個 の青球がある。 この箱から, 「非復元抽出」で青球が出るまで球のとり出しを続ける。X を青球が出るまでの球のとり出し回数とする。

(1) X の確率分布を求めよ。 (2) X の平均 $E(X)$ と 分散 $V(X)$ を求めよ。

3 2 項分布と正規分布

正規分布の確率を計算するには、原本テキストの末尾部分にある「統計分布表」か、「表計算ソフトの命令」をつかえば簡単に求められる。

密度関数、確率分布

- 一様分布（離散型）: $X\sim \mathrm{Unif}\{1,2,\cdots,N\},\ \ p_X(k)=\dfrac{1}{N}, k=1,2,\cdots,N$
- 一様分布（連続型）: $X\sim \mathrm{Unif}[0,1],\ \ f_X(x)=1, x\in[0,1]$
- 2 項分布: $X\sim \mathrm{Binom}(n,p),\ \ p_X(k)=\binom{n}{k}p^k(1-p)^{n-k}, k=0,1,2,\cdots,n$
- 標準正規分布: $Z\sim \mathrm{N}(0,1),\ \ \varphi(x)=\dfrac{1}{\sqrt{2\pi}}e^{-x^2/2}, \Phi(x)=P(-\infty<Z\le x)=\displaystyle\int_{-\infty}^{x}\varphi(t)\,dt$
- 一般の正規分布: $X\sim \mathrm{N}(\mu,\sigma^2),\ f_X(x)=\dfrac{1}{\sqrt{2\pi\sigma^2}}e^{-(x-\mu)^2/(2\sigma^2)}, -\infty<x<\infty$

問 3.1. *2* 項分布 $X\sim Binom(n,p)$ において、 $p=\dfrac{4}{5}$, $P(X=4)=5$, $P(X=5)$ のとき、平均 $E(X)$ をもとめよ。

問 3.2. ある集団には左利きの人が 10% いるという。この中から n を選んだとき、そのうちに少なくとも一人以上左利きの人が含まれる確率を 0.95 以上にするには、n の値はいくつにしなければならないか。

問 3.3. *2* 項分布 $X\sim Binom(8,\frac{1}{2})$ において、 つぎをもとめよ。

(1) 平均 $m=E(X)$ および分散 $s^2=V(X)=E(X-m)^2$ (2) 確率 $P(|X-m|\ge 2s)$ の値

問 3.4 (ref:5.(1))**.** $Z\sim N(0,1)$ のとき、正規分布表（*WEB* 配布の資料あるいは演習書の *p.203, 204*）または表計算ソフト (=norm.s.dist(引数), =norm.s.inv(引数) など) から、つぎの確率を計算せよ。

(1) $P(\,Z<1\,)$ (2) $P(\,Z<1.24\,)$ (3) $P(\,-1\le Z<1.24\,)$

(4) $P(\,2Z+3<4\,)$ (5) $P(\,|\,Z-1\,|<0.24\,)$

問 3.5 (ref:5.(9))**.** コインを 400 回投げるとき、表の出た回数が 180 回から 210 回の間となる確率を正規近似をもちいて計算せよ。

4 標本分布、中心極限定理

正規分布の一次結合と中心極限定理

- 正規分布の和はまた正規分布にしたがう。
  もし $X\sim N(\mu_X,\sigma_X^2), Y\sim N(\mu_Y,\sigma_Y^2)$ ならば、その和 $Z=aX+bY\sim N(\mu_Z,\sigma_Z^2)$ であり、

$$\mu_Z=a\mu_X+b\mu_Z,\quad \sigma_Z^2=a^2\sigma_X^2+b^2\sigma_Y^2$$

  (i) 同一の分布 $X,Y\sim N(\mu,\sigma^2)$ であれば、$Z=X+Y$ に対して、$\mu_Z=2\mu$, $\sigma_Z^2=2\sigma^2$ で、差 $Z=X-Y$ では $\mu_W=\mu-\mu=0$, $\sigma_W^2=\sigma^2+\sigma^2=2\sigma^2$
  (ii) 同一分布の n 個では $V=X_1+X_2+\cdots+X_n$ に対し、 $\mu_V=n\mu$, $\sigma_V^2=n\sigma^2$
- 標準化した確率変数とその極限の分布
  標本平均 (算術平均の形) $\overline{X}=\dfrac{1}{n}\sum_i X_i$ に対し、$\mu_{\overline{X}}=\dfrac{1}{n}(n\mu)=\mu$, $\sigma_{\overline{X}}^2=\left(\dfrac{1}{n}\right)^2\left(n\sigma^2\right)=\dfrac{\sigma^2}{n}$

$$\overline{X}\sim N\left(\mu,\frac{\sigma^2}{n}\right),\qquad Z=\frac{\overline{X}-\mu}{\sqrt{\sigma^2/n}}\sim N(0,1)$$

問 4.1 (ref:6.(2))**.** $1,2,3$ の目がそれぞれ *2* 個づつ記されたサイコロがある。これを投げて、結果とその確率を示し、出る目の平均 μ と分散 σ^2 を求めよ。

a. このサイコロを 3 回投げるとき、出る目の (i) 平均値（出た目 3 個の算術平均)、(ii) 中央値（メディアン）(3 個を大きさの順に並べた中央の値)、の標本分布を導け。
b. それぞれの標本分布について, 平均はいづれも μ に等しいなることを示せ。また分散を計算して、大小を比較せよ。

問 4.2. 2 つの正規分布にしたがう $X\sim N(5,9)$, $Y\sim N(7,16)$ は独立であるとき、
(1) $X-Y$ の分布は何か、 (2) $3X+Y$ の分布は何か (3) $\overline{X}_{25}-\overline{Y}_{16}$ の分布は何か
ここで $\overline{X}_{25},\overline{Y}_{16}$ はそれぞれの分布からの大きさ $25,16$ の標本平均とする。

問 4.3. 2 つの独立な正規分布の一次結合のつくる分布を調べる。X,Y はそれぞれ標準正規分布 $N(0,1)$ にしたがうとき、
(1) $2X+Y$ の平均 (2) $2X+Y$ の分散 (3) 確率 $P(2X+Y>3)$ (4) 確率 $P(1<2X+Y<3)$

問 4.4. *5* 個の正規分布にしたがう $X_i \sim N(2,1)$, $i = 1, 2, \cdot, 5$ に対して、平均 $Y = \dfrac{1}{5}(X_1 + \cdots + X_5)$ をつくる。

(1) Y の平均と分散をもとめよ。 (2) $P(2 < Y < 4)$ をもとめよ。

問 4.5. 平均 μ, 分散 σ^2 の正規母集団から大きさ $n = 4$ の標本平均を $\overline{X}$ とする。

(1) 確率 $P(|\overline{X} - \mu| < \sigma)$ の値をもとめよ。

(2) 確率 $P(|\overline{X} - \mu| > 2\sigma)$ の値をもとめよ。

定理 1 (中心極限定理)**.** 一般に同じ平均 μ, 分散 σ^2 をもつ独立な確率変数に対して、(正規分布でなくてもよい) 標準化した確率変数は

$$W_n = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \frac{\sum_i X_i - n\mu}{\sqrt{n}\sigma} \to N(0,1)$$

ここでの収束のおおまかな定義は、n が十分に大きければ、W_n の分布 $P(W_n \le x)$ と $N(0,1)$ の分布 $\Phi(x) = P(Z \le x) = \displaystyle\int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz$ が近似的に等しい値となる。

2 乗和の分布がカイ二乗分布、正規分布と t-分布

- 正規分布を 2 乗した分布はカイ二乗分布。
  ガンマ関数 $\Gamma(1/2) = \sqrt{\pi}$ であって、この分布が自由度 1 のカイ 2 乗分布である。
- 独立な正規分布の和はやはり正規分布の族 (パラメータは異なる) であり、カイ二乗分布でも和がカイ二乗分布の族をなす。
- t 分布は母集団が正規分布 (正規母集団) から抽出した n 個の標本平均 $\overline{X_n}$ に関するもの。すなわち中心極限定理 (定理 1) では $W_n = \dfrac{\overline{X_n} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$ であったことに対し、もし分母のある標準偏差 σ が σ^2 が未知であれば、その代替として不偏分散 $u_n^2 = \sum(X_i - \overline{X})^2/(n-1) = \dfrac{1}{\binom{n}{2}} \sum_{i<j} \dfrac{(X_i - X_j)^2}{2}$ とした $T_n = \dfrac{\overline{X_n} - \mu}{u_n/\sqrt{n}} \sim t(n)$ が自由度 n の t-分布にしたがう。

問 4.6. $Z \sim N(0,1)$ ならば $Z^2 \sim \chi^2(1)$ (自由度 *1* のカイ二乗分布) を示せ。

［ヒント］ $Z \sim N(0,1)$ に対し、Z^2 の分布関数は $F(u) = P(Z^2 \le u) = P(-\sqrt{u} \le Z \le \sqrt{u}) = \displaystyle\int_{-\sqrt{u}}^{\sqrt{u}} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\, dz$ となる。ここで変数変換 $z = \sqrt{y}$ とすると、$dz = \dfrac{dy}{2\sqrt{y}}$ で対称性から、$F(u) = \displaystyle\int_0^u \frac{1}{\sqrt{\pi}} y^{-1/2} e^{-y} dy$, 微分すると密度関数が $f(u) = \dfrac{1}{\sqrt{\pi}} y^{-1/2} e^{-y}, u \ge 0$ を得る。

定理 2.（正規分布を標準化した *2* 乗値の和）一般の n に対しては、自由度（*degree of freedom, d.f.* と略することもある）n のカイ *2* 乗分布 $\chi^2(n)$ となる：

$$Z_i \sim N(0,1), i = 1, 2, \cdots, n \quad \Rightarrow \quad \sum_{i=1}^{n} X_i^2 \sim f(x) = \frac{1}{\Gamma(n)} x^{n-1} e^{-x}, x \geq 0$$

すなわち確率変数 $Z_i, i = 1, 2, \cdots, n \sim N(0,1)$ で独立であれば

$$Z_1^2 + Z_2^2 + \cdots + Z_n^2 \sim \chi^2(n)$$

ここで $\chi^2(n)$ は自由度 n のカイ *2* 乗分布を意味する。もし $X_i \sim N(\mu_i, \sigma_i^2), i = 1, 2, \cdots, n$ で独立であれば、

$$W_n = \sum_i \frac{(X_i - \mu_i)^2}{\sigma_i^2} \sim \chi^2(n)$$

となる。

5 母数の推定、信頼区間

母集団分布を表す未知母数 (parameter)θ を、抽出された標本データ (date) からつくる統計量 (statistics)（標本の関数）で、未知とした値を推定 (estimate) する。

点推定

未知母数を推定するために望ましい性質：

- 不偏推定量 (unbiansedness): 推定統計量 T の期待値 (平均) が推定したい値 θ に等しい、ズレていない
- 一致推定量 (consistency):標本の大きさを増やすとズレる確率が小さくなる
- 有効推定量 (efficency):分散を比較して、小さいものを「より有効な」推定量といい、不偏性をもつなかで分散が最も小さいものを有効推定量という

推定量を見つけ出す方法のひとつに最尤法があり、これで定めたものが最尤推定量。母集団分布 $f(x;\theta)$ とし、標本値を $x_1, x_2, \cdots, x_n$ とおく。

- 尤度関数 (Likelyhood function);

$$L(\theta) = f(x_1;\theta) f(x_2;\theta) \cdots f(x_n;\theta)$$

- 最尤推定量 (Likelyhood estimator):

$$\text{尤度を最大にするもの} : \hat{\theta} \Leftrightarrow L(\hat{\theta}) = \max_{\theta} L(\theta)$$

区間推定とは、２つの推定量をつくり、その範囲のなかに含まれる確率という形で推定をおこなう。確率 α から定めた区間を信頼区間といい、確率のことをとくに信頼係数とよび、端点を信頼限界という。問題の設定から、90%, 95% ,99% の確率値が経験的によく用いられる。以下では代表的な例を示す。より詳しい場合に

はテキストを参照しなさい。

定理 3. (i) 大標本の場合で分散が未知のとき、「平均 μ に関する信頼区間」（正規分布で計算）
(ii) 「2 つの平均の差 $\mu_1 - \mu_2$ に関する信頼区間」（正規分布で計算）
(iii) 分散 σ^2 に関する信頼区間（χ^2（カイ 2 乗）分布で計算）
(iv) 2 項事象における比率 p の信頼区間（大標本の場合として正規分布で計算）

問 5.1. 正規母集団 $N(\mu, \sigma^2)$ からの n 個の標本を $X_1, X_2, \cdots, X_n$ とおく。標本平均 $\overline{X}_n$ から母平均 μ の推定をおこなう。

(1) 分散が既知のばあいに、信頼係数 $100(1-\alpha)\%$ の信頼区間の記述せよ。
(2) 分散が未知の場合で、大標本では分散の部分にはどう修正したらよいか？

問 5.2. ある学校の生徒の集団から、40 人を無作為に選んでテレビの視聴時間を調べた。そのデータの平均は 18.2 時間、標準偏差は 5.4 時間であった。このときこの学校の生徒集団に対してテレビ視聴時間の分布を正規分布として、その平均の μ を推定する。

(1) 95% 信頼区間　(2) 98% 信頼区間
をもとめよ。

問 5.3. 大標本のばあいの平均の差に対する信頼区間を調べる。

A 社の電球 80 個の寿命時間を調べたところ、平均 1070 時間、分散 472 で、一方 B 社の電球 60 個の寿命時間を調べたところ、平均 1042 時間、分散 366 であった。2 つの集団分布の平均差 $\mu_A - \mu_B$ を推定する。2 つの集団の平均差 $\mu_A - \mu_B$ の公式から

(1) 信頼係数 95% のとき、　(2) 信頼係数 90% のとき、
をもとめよ。

6 検定

真の母数の値が未知であるとき、抽出した標本データから情報で、その値を推定する方法に対して、経験的な知識から、真であろうと考えた母集団分布の状況を真偽を判断すること。しかしデータを活かしていても、間違う可能性が大いにあり得る。これらの間違いを避けるために、その過誤から生じる影響の重大さからを 2 種類: (I) 予想をから立てた仮説が真であったにもかかわらず、これを正しいと判断しない（棄却する）場合 (false positive) と (II) 仮説は正しくなかったが、これを正しいものとして受け入れて（採択）しまうこと (false negative) がある。これらの誤りを少なく、正しければそれを採択し、間違っていれば棄却することの確率が大きくすることが検定である。本文テキストの 115 頁から 118 頁、その一覧表を参照しなさい。

キーワード：帰無仮説、対立仮説、仮説の採択あるいは棄却、第 1 種の過誤の確率、第 2 種の過誤の確率、有意水準、p 値

仮説検定には、正規母集団に関する仮説を調べることが多いが、一方さまざまな状況に関してカイ二乗分布を統計量として検定するもの、これをカイ二乗検定をよぶ。

定理 4. ここではつぎのものをカイ二乗検定として述べる。詳しくは本文テキストや *Web* を参照せよ。

- 適合度検定　カテゴリーデータによって、分類したクロス集計（項目によって分別した多次元の度数分布表）を仮説にもとづいて得られる理論値と、観察された実測値データとを比較して、理論値から得られる分布との適合性を判断するもの。たとえば、ある集団の血液型の分布に対して、理論値により得られる数値と、実測して得たクロス集計（度数集計）との差異をカイ二乗統計量の計算によって仮説を検定することで、結論の判断をおこなう。
- 独立性の検定　確率の概念の重要なものの一つに、事象の独立 $P(A\,B) = P(A)P(B)$ があるが、この命題を実測値として得られて分類をクロス集計したもの（分割表）を、仮説として立てた「命題によって分類した値（期待値とよぶ）について独立であるかどうか」を調べるために、カイ二乗分布の統計量によって判断しようとする。観察値（実測値）と期待値（理論値）との差異をカイ二乗分布の値の大きさによって検定する。

問 6.1. ある箱のなかにある球の個数は、仮説 A：赤球 3 個と黒球 7 個、または 仮説 B：赤球 6 個と黒球 4 個のいづれかである。このとき、$\left\{ \begin{array}{l} \text{帰無仮説 } H_0 : \text{仮説 } A \text{ が正しい} \\ \text{対立仮説 } H_1 : \text{仮説 } B \text{ が正しい} \end{array} \right\}$ としたとき、標本を 2 個選んだ結果から判断する。非復元抽出をおこない、その結果が少なくとも *1* 個が黒であれば H_0 を採択し、それ以外は H_0 を棄却する。このような検定であれば、第 1 種の過誤、第 2 種の過誤の確率はいくつかになるか？

問 6.2. 正規母集団 $N(\mu, \sigma^2)$ の平均値 μ に対する検定をつぎの (1) ～ (4) の場合に決定せよ。ただし n, n_1, n_2 は各集団における標本サイズ、$\overline{X}, \overline{X_1}, \overline{X_2}$ は各変数に対する標本平均、σ^2 は母分散、s, s_1, s_2 は標本標準偏差、α は有意水準をあらわす。

(1) $n = 100$, $\overline{X} = 38$, $\sigma = 25$, $\alpha = 0.05$ のとき、帰無仮説 $H_0 : \mu = 40$, 対立仮説 $H_1 : \mu \neq 40$
(2) $n = 16$, $\overline{X} = 240$, $\sigma = 20$, $\alpha = 0.05$ のとき、帰無仮説 $H_0 : \mu = 250$, 対立仮説 $H_1 : \mu < 250$
(3) $n = 14$, $\overline{X} = 51.52$, $s = 2.13$, $\alpha = 0.01$ のとき、帰無仮説 $H_0 : \mu = 50$, 対立仮説 $H_1 : \mu > 50$
(4) $n_1 = 50$, $n_2 = 40$, $\overline{X_1} = 22.31, \overline{X_2} = 21.54$, $s_1 = 3.8, s_2 = 3.2$, $\alpha = 0.05$ のとき、帰無仮説 $H_0 : \mu_1 = \mu_2$, 対立仮説 $H_1 : \mu_1 \neq \mu_2$

問 6.3. つぎの母集団比率の検定においては大標本と考え、正規近似を当てはめて検定せよ。

(1) $n = 120$, $\overline{X} = 90$, $\alpha = 0.05$ のとき、帰無仮説 $H_0 : p = 0.7$, 対立仮説 $H_1 : p \neq 0.7$
(2) 2 集団の比較：$n_1 = 420$, $\overline{X_1} = 118$, $n_2 = 530$, $\overline{X_2} = 121$ について、 $\alpha = 0.05$ のとき、帰無仮説 $H_0 : p_1 = p_2$, 対立仮説 $H_1 : p_1 > p_2$

問 6.4. あるクラス別の科目の成績評価 {*A,B,C,D* } が つぎの結果であった。クラス間に成績の優劣が認められるかどうか、有意水準 *1%* で検定せよ。

	クラス 1	クラス 2	クラス 3
評価 A	8	12	6
評価 B	25	32	10
評価 C	10	9	14
評価 D	7	5	12
計	50	58	42

7 相関と回帰

相関係数、回帰直線

- 相関係数 (corelation) とは、2 変量のデータから得られる平面にプロットした（散布図）に対して、その変化の傾向を直線的な集中度を示す尺度をいう。それぞれの変量を平均と分散で標準化 : $\dfrac{x_i-\overline{x}}{s_x}$, $\dfrac{y_i-\overline{y}}{s_y}$ とし、 このデータについての共分散 : $Cov\left(\dfrac{x_i-\overline{x}}{s_x},\dfrac{y_i-\overline{y}}{s_y}\right)=\dfrac{1}{n}\sum_i\left(\dfrac{x_i-\overline{x}}{s_x}\cdot\dfrac{y_i-\overline{y}}{s_y}\right)$ が相関係数である。もし分散を $s_x^2=\dfrac{1}{n}\sum(x_i-\overline{x})^2=s_{xx}, s_y^2=\dfrac{1}{n}\sum(y_i-\overline{y})^2=s_{yy}$ と表し、共分散を $s_{xy}=$ と表してみると、

$$r=\frac{s_{xy}}{s_xs_y}=\frac{n(\sum x_iy_i)-(\sum x_i)(\sum y_i)}{\sqrt{n\sum x_i^2-(\sum x_i)^2}\sqrt{n\sum y_i^2-(\sum y_i)^2}}$$

  末尾の項では、各 x,y のデータの和、2 乗和に加えて、積和から、相関係数を求める式である。
- 回帰直線 (regression line) とは、「直線の当てはめ」であって、直線の一次式を $y=a+bx$ として、観測データ値 $(x_i,y_i), i=1,2,\cdots,n$ にもっとも当てはまる基準として、最小二乗値として $\sum_i(y_i-a-bx_i)^2$ を選んで定めた係数 a,b によって示される直線を変量 Y の変量 X への回帰直線という。標本データから x,y の平均、その共分散から y 切片は $a=\overline{y}-b\overline{x}$, 傾き b は $b=\dfrac{s_{xy}}{s_x^2}$ であり、

$$y-\overline{y}=\frac{s_{xy}}{s_x^2}(x-\overline{x})=r\frac{s_y}{s_x}(x-\overline{x})$$

直線は、各データの平均値 $(\overline{x},\overline{y})$ を通り、相関係数 r が係数の正負を定める（標準偏差は常に正の値だから)、相関係数が正の値であれば、右上がりの直線であり、負の値では逆となる。ただし、相関係数は、データ増減の傾向を表すが、基本は「直線への集中度」を示すもの。

問 7.1. つぎを示せ。(i) $(s_{xy})^2\le s_x^2s_y^2$, (ii) $-s_xs_y\le s_{xy}\le s_xs_y$, (iii) 相関係数は $-1\le r\le 1$

問 7.2. つぎの場合について、2 つの変量間の相関係数を求めよ。また散布図を描き、位置関係を確認せよ。

(a) 5 個のデータ (X,Y):

X :	10	20	30	40	50
Y :	2	4	6	8	10

(b) 8 個のデータ (Z,W):

Z :	1	1	1	0	0	−1	−1	−1
W :	1	0	−1	1	−1	1	0	−1

(c) 7 個のデータ (U,V):

U :	−3	−2	−1	0	1	2	3
V :	8	3	0	−1	0	3	8

問 7.3. (i) X, Y の共分散を $s_{x,y}$ とするとき、この変量を $Z = aX + b$, $W = cY + d$ と変換したならば、Z, W の共分散 $s_{z,w}$ ともとの共分散 $s_{x,y}$ のあいだに成り立つ関係式を示せ。(ii) 2 変量 X, Y を一次変換して変量 $Z = aX + b$, $W = cY + d$ と定めた相関係数でも同じ値となることを示せ。

問 7.4. 10 人の生徒について, 身長 (X) とひじの長さ (Y) を測定したら、次のデータを得た。単位は c m）

X :	152	145	164	153	132	148	149	138	142	150
Y :	22	21	25	23	20	21	22	21	22	24

このとき、(i) 2 変量の間の相関係数を求めよ。(ii)　変量 Y の変量 X への回帰直線をもとめよ。(iii) この回帰直線から、身長が 150 c m の生徒に対するひじの長さはどのくらいか？

問 7.5. n 個の標本が 2 つ の変量 x, y に ついてそれぞれ順位がつけられているとき, これら 2 組の順位 $(x_1, x_2, \cdots, x_n)$ と $(y_1, y_2, \cdots, y_n)$ の間の相関係数（スピアマンの順位相関係数） r_s は

$$r_s = 1 - \frac{6\sum_{i=1}^{n}(x_i - y_i)^2}{n(n^2-1)}$$

となることを示せ。