慶應義塾大学ビジネス・スクール

統計学　ノート（1）

1　はじめに

統計学はもともと社会における現象の法則性に対する人間のさまざまな実践的関心から起こった。現象の法則性を知るために，すべてを詳しく調べ，規則性から法則を見つけてもよいし，また，一部を観察して，そこから論理に基づいた推論を行い，全体の法則性を発見してもよい。

全体の法則性を知るために，すべてを調べなければならないという考え方は全数調査と呼ばれ，実際には国勢調査などで用いられている。

一方，現在の統計学は「一部」から「全体」を知ろうとする，科学的推論のための方法論の理論的体系となっている。この方法論は，確率論に基づく数学的根拠とともに，広い応用範囲と妥当性をもっている。

ここで，統計の目的は得られたデータからデータを集めたもとの集団についての情報を引き出すことである。このとき，得られたデータ(観測値)の集まりを標本(sample)といい，もととなる集まりを母集団(population)という。

この標本と母集団を区別して考え，標本から母集団への橋渡しとして確率論が用いられるようになって，近代統計学の基礎が確立した。そして，これに基づいて，統計的推測の論理が築かれた。母集団に対する推定(estimation)の理論と仮説検定(hypothesis testing)である。

2　標本

2.1　データの収集，分類とその記述

母集団に関する妥当な結論を得るためには，どのようにして母集団から標本を抽出すればよいかということが問題である。その標本抽出の方法は，母集団を構成するどの個体も標本に選ばれる確率が同じになるような抽出法であり，無作為抽出という。

データは量的データ，質的データに分けられる。数値などで定量的に表されるデータを量的データとよぶ。また，量的データは連続型変数と離散型変数に分けられる。長さ，重さ，温度，時間などのある区間内の任意の値をとりうる変数は連続型変数であり，事故の件数，人数，金額などは離散型変数の例である。数値として観測することができず，ある

カテゴリーに属していることや，ある状態にあることだけがわかるデータを質的データとよぶ。性別，学歴，天気などがその例であり，これらはダミー変数とよばれる 0 または 1 をとる変数によって数量化が可能となる。

標本データの分類は普通連続型変数に対して行われる。観測値をいくつかの等間隔の階級に分け，それぞれの階級に分類された観測値がいくつあるか度数を数えて，表にしたものが度数分布表である。ここで，階級の上限値と下限値を階級境界値という。また，階級を代表する値を階級値とよび，普通階級の中心値を階級値とする。

標本データはその分布の特徴を直観的に理解するために，連続型変数に対してヒストグラムと呼ばれるグラフが描かれる。離散型変数に対しても近似的に連続型と同様にヒストグラムを描くこともある。

2.2 分布の特性値

ヒストグラムによって直観的な理解は得られるが，より正確な分布に関する情報は数量的に得られる。分布の特徴を表す特性値としてヒストグラムの中心の位置を与える特性値と，散らばりの程度を与える特性値を考える。これらはそれぞれ位置の測度，変動の測度とよばれる。

2.2.1 平均

変数 X についての n 個の標本値がある母集団からとられたとする。これらの値を $X_1, X_2, \ldots, X_n$ で表す。このとき，X の平均（$\overline{X}$ で表す）は

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} \tag{1}$$

で与えられる。

次に分類されたデータの場合を考える。度数分布表に k 個の階級があり，その階級値を $x_1, x_2, \ldots, x_k$，それぞれの度数を $f_1, f_2, \ldots, f_k$ とする。このとき，平均を $\bar{x}$ で表すとすると，次の式で与えられる。

$$\overline{X} = \frac{x_1 f_1 + x_2 f_2 + \cdots x_k f_k}{n} \tag{2}$$

また，Σ記号を用いると，(1)，(2) はそれぞれ次のように書ける。

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{3}$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} x_i f_i \tag{4}$$

2.2.2 標準偏差

観測値が平均の周りにどの程度変動しているかを測る尺度を考える。最も単純な尺度は範囲であり，観測値の最大値から最小値を引いた差で表される。しかし，この尺度は次節に述べる望ましくない性質をもつので，次の尺度を考える。

位置の測度として平均が用いられるので，平均からの偏差を用いるのが適当であろう。散らばりを測るためには偏差の絶対値を考えればよいが，絶対値は扱いにくいので偏差の2乗をとりその平均を考える。したがって，

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \tag{5}$$

で与えられる。分類されたデータの場合は(4)の平均を用いて次のように与えられる。

$$\frac{1}{n}\sum_{i=1}^{k}(x_i - \bar{x})^2 f_i \tag{6}$$

しかし，統計的推論の問題を扱うときにはこの式の分母のnを$n-1$に修正した式を使う。修正式を用いる理由はあとで述べる。修正式をs^2で表し，分散という。分散は分類されていないデータに対して

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \tag{7}$$

で，また分類されたデータに対しては

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{k}(x_i - \bar{x})^2 f_i \tag{8}$$

で定義される。

また，分散の単位は観測値の単位の2乗になっている。分布を表す量は元の観測値と同じ単位をもつことが望ましい。平均はこれを満たしているが，分散は満たしていない。そこで，分散の正の平方根をとることによってこれを満たすことができる。分散の正の平方根を標準偏差という。したがって，標準偏差は分類されていないデータに対して

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2} \tag{9}$$

で，分類されたデータに対しては

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{k}(x_i - \bar{x})^2 f_i} \tag{10}$$

で定義される。

実際にデータから計算する場合には(9)，(10)を変形した次の式を用いるのが便利である。

$$s = \sqrt{\frac{\sum X_i^2 - \frac{1}{n}\left(\sum X_i\right)^2}{n-1}}$$

$$s = \sqrt{\frac{\sum x_i^2 f_i - \frac{1}{n}\left(\sum x_i f_i\right)^2}{n-1}}$$

2.3 その他の測度

位置の測度として平均を挙げたが，その他に最頻値，中央値が用いられることも多い。最大の度数をもつ観測値が1つあるとき，その観測値を最頻値（モード）と定義する。また，観測値の組に対して観測値を大きさの順に並べたとき，ちょうど中央にくる観測値が1つあれば，その観測値を中央値（メディアン）と定義する。

変動の測度として前に述べた範囲がある。範囲は計算が容易で理解しやすいのでよく使われている。しかし，標本の大きさが増えると範囲の値も増えてしまう傾向がある。また，データの中に極端に大きい（小さい）異常値が存在する場合には，これに大きく左右されてしまう。そこで，中央値を定義するときに用いた方法と同様な方法で，ヒストグラムの面積を4等分する値を見つけることができる。これらの値を四分位数とよぶ。最小の四分位数を第1四分位数，中央値を第2四分位数，最大の四分位数を第3四分位数という。簡単な変動の測度として，第3四分位数と第1四分位数の差が四分位範囲とよばれて用いられる。

3 確率

統計学においては問題に対する解は確率的に表現される。この節では確率の基本的事項について述べる。

3.1 標本空間

統計学や確率論では起こりうる事柄を事象と呼ぶが，これは集合を用いて説明する必要がある。繰り返しが可能な実験において可能な結果のそれぞれを点で表現することが便利である。この点を標本点，その全体の集合を標本空間とよぶ。

定義1　実験の可能な結果を表す点（標本点）全体の集合をその集合の標本空間という。

1つの標本点からなり分解できない事象を単一事象，複数の標本点からなる事象を複合事象とよぶ。単一事象は普通 e に添字をつけて，$e_1, e_2, \ldots$ のように表す。また複合事象は $A, B, C, \ldots$ のように大文字で表す。

3.2 事象の確率

繰り返し実験の各標本点に対して確率を割り当てるとき，多数の繰り返し実験から得られた期待相対度数が用いられる。このとき，相対度数の合計は常に1であるから，各確率は0と1の間の数で，かつその合計は1でなければならない。

複合事象は単一事象の集まりであるので，複合事象の確率の定義は次のように与えられる。

定義2　複合事象 A が起こる確率は A を構成している単一事象の確率の和である。

偶然を伴うゲームの実験では，可能な結果のすべてが同じ相対度数で起こると期待される場合が多い。そのような実験では，標本空間のすべての標本点に等しい確率が割り当てられる。標本点の総数を n 個とすれば，$1/n$ の確率が割り当てられる。単一事象の確率がすべて等しいという，単純な場合には複合事象の確率の計算は容易である。複合事象 A が $n(A)$ 個の単一事象からなるとすれば，確率 $P\{A\}$ は公式

$$P\{A\} = \frac{n(A)}{n} \qquad (11)$$

で与えられる。

3.3 加法定理

いま，2 つの事象 A_1, A_2 を考える。ここで，A_1 と A_2 の両方がともに起こる確率を知りたいとする。この結合事象を $(A_1 \text{ and } A_2)$ で，確率を $P\{A_1 \text{ and } A_2\}$ で表す。また，A_1 と A_2 の少なくとも 1 つが起こる確率を知りたいとする。このような事象を $(A_1 \text{ or } A_2)$ で，確率を $P\{A_1 \text{ or } A_2\}$ で表す。

2 つの事象 A_1 と A_2 が一方が起これば他方は決して起こらないという性質をもつとき，これらの事象は互いに排反であるという。このとき，$n(A_1)$ を複合事象 A_1 を構成する標本点の数，$n(A_2)$ を複合事象 A_2 を構成する標本点の数とすると，A_1 または A_2 が起こることに相当する標本点の数はこれら 2 つの点の和に等しい。したがって，n を標本空間の点の総数とすると，公式(11)から

$$P\{A_1 \text{ or } A_2\} = \frac{n(A_1) + n(A_2)}{n}$$

$$= \frac{n(A_1)}{n} + \frac{n(A_2)}{n}$$

$$= P\{A_1\} + P\{A_2\}$$

加法定理　A_1 と A_2 が互いに排反ならば，

$$P\{A_1 \text{ or } A_2\} = P\{A_1\} + P\{A_2\}$$

3.4　乗法定理

この節では個々の事象の確率を用いて，$P\{A_1 \text{ and } A_2\}$ に対する公式を導く。そのために，条件付き確率という概念を導入する。

いま，A_1 が起こったことが分かっているとき，A_2 が起こる確率を知りたいとする。このとき，A_1 が起こることが条件となっているので，問題の標本空間は A_1 を構成する単一事象の集まりに縮小される。これらの中で A_2 にも含まれる点，A_1 と A_2 の重複部分にある点が A_1 と A_2 の両方が起こることに対応する点である。$n(A_1)$ を A_1 内の点の数，$n(A_1 \text{ and } A_2)$ を A_1 と A_2 の両方に含まれる点の数とすると，標本空間を A_1 内の点の集まりに制限したときの A_2 の起こる確率は，公式 (11) から $n(A_1 \text{ and } A_2)/n(A_1)$ で与えられる。この条件付き確率を記号 $P\{A_2 \mid A_1\}$ で表すと，

$$P\{A_2 \mid A_1\} = \frac{n(A_1 \text{ and } A_2)}{n(A_1)}. \tag{12}$$

一般論で，条件付き確率を考えてみると，

$$P\{A_1\} = \frac{n(A_1)}{n},$$

$$P\{A_1 \text{ and } A_2\} = \frac{n(A_1 \text{ and } A_2)}{n}.$$

この第 2 式を第 1 式で割ると，

$$\frac{P\{A_1 \text{ and } A_2\}}{P\{A_1\}} = \frac{n(A_1 \text{ and } A_2)}{n(A_1)}$$

この結果と公式 (12) から次の公式が得られる。

$$P\{A_2 \mid A_1\} = \frac{P\{A_1 \text{ and } A_2\}}{P\{A_1\}}. \tag{13}$$

この式で A_1 を与えたときの A_2 の条件付き確率が定義される。これを積の形に変形したものが確率の乗法定理であり，次のように表される。

乗法定理

$$P\{A_1\}P\{A_2 \mid A_1\} = P\{A_1 \text{ and } A_2\} \tag{14}$$

この式は A_1 と A_2 を入れ替えてもかまわない。なぜなら，この 2 つの事象の番号は便宜的なものであり，2 つの事象の起こり方の時間的順序を意味していないからである。慣例により $P\{A_2 \mid A_1\}$ を A_1 が起こったことが分かっているときの A_2 の起こる確率といっている。しかし，時間的順序の関係をもつ事象の対を考えることも多い。

3.5 独立な事象の乗法定理

2つの事象A_1, A_2において，A_2の起こる確率がA_1の起こるかどうかに関係しないとき，A_2はA_1と独立であるといい，次のように表す。

$$P\{A_2 \mid A_1\} = P\{A_2\}$$

このとき，乗法定理は次のようになる。

$$P\{A_1 \text{ and } A_2\} = P\{A_1\}P\{A_2\}.$$

事象$(A_1 \text{ and } A_2)$は事象$(A_2 \text{ and } A_1)$と同じであるから，(14)でA_1とA_2を入れ替えることができ，

$$P\{A_1 \text{ and } A_2\} = P\{A_2\}P\{A_1 \mid A_2\}$$

が得られる。上の2つの式の右辺を比べると，$P\{A_1 \mid A_2\} = P\{A_1\}$となり，以上より$A_2$が$A_1$と独立なら，$A_1$は$A_2$と独立であることが示された。独立性は相互的なものであるから，A_1とA_2が独立であるという言い方が正しい。

A_1とA_2が独立ならば，

$$P\{A_1 \text{ and } A_2\} = P\{A_1\}P\{A_2\}. \qquad (15)$$

3.6 ベイズの定理

条件付き確率を応用して，得られた結果から原因を推定する問題を考える。ここで，2段階実験が与えられたとする。第1段階はk個の可能な結果のどれか1つが起こっていなければならない，ということである。これらの可能な結果を$e_1, e_2, \ldots e_k$で表す。次に第2段階ではm個の可能な結果のどれか1つが起こっていなければならず，それらを$o_1, o_2, \ldots o_m$で表す。第1段階の可能な結果$e_1, e_2, \ldots e_k$のそれぞれに対する確率が与えられているとして，それらを$P\{e_1\}, P\{e_2\}, \ldots P\{e_k\}$で表す。第1段階の事象$e_i$が起こったことが分かったとき，第2段階の事象$o_j$が起こる確率，つまり条件付き確率$P\{o_j \mid e_i\}$もすべての$e_i, o_j$の組に対して与えられているとする。ここで問題は第2段階の事象o_jが起こったことが分かったとき，第1段階の事象e_iが起こっていたという確率を求めることである。この条件付き確率は$P\{e_i \mid o_j\}$で表される。

このとき，公式(13)は次のようになる。

$$P\{e_i \mid o_j\} = \frac{P\{e_i \text{ and } o_j\}}{p\{o_j\}} \qquad (16)$$

また，公式(14)は

$$P\{e_i \text{ and } o_j\} = P\{e_i\}P\{o_j \mid e_i\} \qquad (17)$$

(17)の右辺の2つの確率は既知であるから，(16)の分子は得られる。また，(16)の分母$P\{o_j\}$は第1段階との関連で起こる，互いに排反なすべての場合を考えることによって得られる。それは第1段階でe_1が起こり第2段階でo_jが起こる，第1段階でe_2が起こり第2段階でo_jが起こる，…，第1段階でe_kが起こり第2段階でo_jが起こる場合のどれかで

ある。これらの事象は互に排反だから，加法定理より

$$P\{o_j\} = P\{e_1 \text{ and } o_j\} + P\{e_2 \text{ and } o_j\} + \cdots + P\{e_k \text{ and } o_j\}.$$

これは，

$$P\{o_j\} = P\{e_1\}P\{o_j \mid e_1\} + P\{e_2\}P\{o_j \mid e_2\} + \cdots + P\{e_k\}P\{o_j \mid e_k\}$$

これをさらにまとめると，

$$P\{o_j\} = \sum_{i=1}^{k} P\{e_i\}P\{o_j \mid e_i\}$$

これと (17) から次の式が得られる。これをベイズの定理という。

ベイズの定理

$$P\{e_i \mid o_j\} = \frac{P\{e_i\}P\{o_j \mid e_i\}}{\sum_{i=1}^{k} P\{e_i\}P\{o_j \mid e_i\}} \tag{18}$$

3.7 順列・組合せ

実験が多数の段階からなり，それぞれの段階が多数の可能な結果が考えられる場合，標本点を数え上げるのに公式を用いると便利である。

まず，n 個の相異なるものから k 個を選んで並べることを考える。この並べ方（順列）の総数を ${}_n\mathrm{P}_r$ とすると，

$${}_n\mathrm{P}_r = n(n-1)(n-2)\cdots(n-r+1) \tag{19}$$

で与えられる。${}_n\mathrm{P}_r$ を n 個のものから r 個のものをとる順列の数とよぶ。

次に，n 個の相異なるものから r 個を選ぶことを考える。この選び方の組合せの総数は次の公式で与えられる。

$$\binom{n}{r} = \frac{n(n-1)(n-2)\cdots(n-r+1)}{r(r-1)(r-2)\cdots 1} \tag{20}$$

$\binom{n}{r}$ は ${}_n\mathrm{C}_r$ という記号で表されることもある。

ここで，順列，組合せの数を次の記号で n の階乗を用いて表すと便利である。

$$n! = n(n-1)(n-2)\cdots 3\cdot 2\cdot 1$$

また，$0! \equiv 1$ と定義する。階乗の記号を用いると，n 個のものから r 個とる順列の数は

$${}_n\mathrm{P}_r = \frac{n!}{(n-r)!} \tag{21}$$

また n 個のものからr個とる組合せの数は

$${}_n\mathrm{C}_r = \frac{n!}{r!(n-r)!} \tag{22}$$

で表される。

4 問題

1. 11 人の子供の体重を示した次のデータにおいて，データを分類しないで，(a) 平均，(b) 標準偏差 を求めよ:39，52，40，45，46，55，48，40，43，47，44。

2. 20 人の学生の評定平均を示した次のデータにおいて，データを分類しないで，(a) 平均，(b) 標準偏差 を求めよ:2.4，1.2，1.4，2.4，1.1，1.8，1.9，1.4，1.8，3.2，2.4，2.2，2.4，1.8，3.6，1.8，1.2，2.4，2.0，3.4。

3. 次の度数分布表は普通の種類の潅木 60 本について，フィート単位で測った幹の直径に関するものである。

x	1	2	3	4	5	6	7	8	9	10	11	12
f	1	7	12	16	10	4	5	2	1	1	0	1

このデータに対して，(a) $\bar{x}$，(b) s を求めよ。

4. r 個のクラスにテストを行う。第 i クラスの人数を n_i，平均を $\bar{x}_i$ とする $(i = 1, \cdots, r)$。$n_1 + \cdots + n_r = N$，全平均を $\bar{x}$ とするとき $\bar{x} = \frac{1}{N}\sum_{i=1}^{r} n_i \bar{x}_i$ を示せ。

5. 2 個のサイコロを転がす実験で次の事象の確率を求めよ。(a) 出た目の和が 9 にならない，(b) 1，2，3 のどの目も出ない，(c) 両方とも 3 以上の目が出る，(d) 同じ目が出ない，(e) どちらか一方だけが 1 の目を出す。

6. ある標本空間が互いに排反な 3 つの事象 A_1, A_2, A_3 に分けられているとする。$P\{A_1\} = 3/6, P\{A_2\} = 2/6, P\{A_3\} = 1/6$ として，次の事象の確率を求めよ。
   (a) $P\{A_1 \text{ or } A_3\}$
   (b) $P\{A_1 \text{ or } A_2 \text{ or } A_3\}$
   (c) $P\{A_1$ でない$\}$
   (d) $P\{A_1 \text{ or } A_2$ でない$\}$

7. 白球 2 個と黒球 3 個と緑球 5 個を含む壷から 2 個の球を取り出すとき，(a) 両方とも緑球が得られる確率，(b) 両方とも同じ色の球が得られる確率，を求めよ。

8. 同じ製品が A, B, C3 種の機械でつくられる。A は全体の 50%，B, C はそれぞれ 30%，20% つくる。不良品が出る率はそれぞれ 3%，4%，5%である。全製品中から 1 個取ったら不良品であった。それが各機械でつくられた確率はいくらか。

9. $A_1, A_2, \ldots, A_n$ が独立で，$P\{A_i\} = p_i$ とするとき，これらのどれもが起こらないという事象の確率を $p_1, p_2, \ldots, p_n$ で表せ。

10. 1/5，女子生徒は 1/25 であるとする。そのとき，次の事象の確率を求めよ。(a) 無作為に選んだ 1 人の生徒が男子の理科専攻生である，(b) 無作為に選んだ 1 人の生徒が理科専攻生である，(c) 無作為に選んだ 1 人の理科専攻生が男子である。

11. 52 枚 1 組のトランプ札から 4 枚の札を抜くとき，(a) 4 枚ともスペードである確率を求めよ，(b) 4 枚とも同じ種類の札である確率を求めよ，(c) 4 枚の中にスペードが含まれない確率を求めよ。

12. 40 個の良品と 10 個の不良品のヒューズが入った箱があって，それから 8 個を取り出すとき，8 個とも良品である確率を求めよ。ただし，組合せの公式を用いて計算すること。

13. 次の 2 つの賭けを比較せよ。
    (a) サイコロを 4 回投げるとき，6 の目が少なくとも 1 回出るのに賭けるか，1 回も出ないほうに賭けるか。
    (b) サイコロを 2 個同時に 24 回投げるとき，(6, 6) の目が少なくとも 1 回出るのに賭けるか，1 回も出ないほうに賭けるか。

なお，$4 \times (1/6) = 24 \times (1/36)$ に注意する。

参考文献

[1] 初等統計学 原書第 4 版，P. G. ホーエル，培風館 (1981)。

[2] 統計学入門，東京大学教養学部統計学教室編，東京大学出版会 (1991)。

慶應義塾大学ビジネス・スクール

Contents Works Inc.