

III-2-2. 二項分布の性質

必要な言葉の定義

期待値：ある数値が得られる確率的事象で、得られる確率が最も高い数値。観念的には、あることを無限回繰り返して、そのたびにある値を計算し、その値の無限回分の平均値のこと。

一般的には、ある確率事象が起きたときに得られる値とその事象が起きる確率の積の、起きうる全事象についての総和として計算する。

$$E(f(x)) = \sum_{i=1}^n f(x_i) p_i$$

Where

$f(x_i)$: 事象 i が起こることによって得られる値

$p(i)$: 事象 i がおこる確率

$E(f(x))$: 期待値

式 9

平均値

定義 1. 全データの総和/データの総数

$$\frac{\sum_{i=1}^n x_i}{n}$$

定義 2. $E(x) = \sum_{i=1}^n x_i p_i$ (期待値の表現を用いた新たな定義)

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \sum_{i=1}^n x_i p_i$$

$\bar{x}$: average of x

式 10

二乗和: Sum of square (平均値からの距離の 2 乗の和)、平均値は、対象となっているデータセット (標本集団) の平均値

$$SS = \sum_{i=1}^n (x_i - \bar{x})^2$$

$\bar{x}$: 標本集団の平均値

n : 標本集団のサンプル数

式 11

2 次の積率 (2 次のモーメント)

$$E((x - \mu)^2)$$

x はデータ、 μ は平均

n をデータの総数とすれば標本集団については、2次の積率を SS/n として計算できます。2次の積率は、その標本集団のデータが平均値からどのくらい隔たった値のデータで構成されているかを平均的に表したものです。平均値からの距離の平均でも良さそうですが、 μ は平均値だから、 $E(x - \mu)$ は0になってしまいます。そこで、データのばらつきに指標には $E((x - \mu)^2)$ を使います。

実際に知りたいのは母集団の平均値です。また、その推測値の信頼範囲がどのくらいなのかも考えなくてはなりません。そのような方法はいくつか考えられます。たとえば、最近流行っている MCMC（マルコフチェーン・モンテカルロシミュレーション）では、標本集団からランダムにデータを取りだして、その平均値の変動を記述します。原理的にどのように分布するかではなく、実際のデータをランダム現象と考えて、そこからランダムにデータを取り出して変動を観察したら、結果的にある分布範囲が出てくるという方法を使います。この方法は、コンピュータの計算速度に依存しています。コンピュータがなかった時代には、そんな手間と時間のかかることはできませんから、そういう方法に依存しないで、期待値として母集団の平均値の変動範囲を求める方法を考えたのです。私たちは、漠然と、標本集団の平均値を母集団の平均値に近い値（推測値）だと考えて、標本集団の平均値を求めます。標本集団の平均値を母集団の平均値の推測値だと考えるのは、考えるまでもない当然のような気がします。ここでは、その考え方の妥当性を検討します。そんな面倒なことをするのは、平均値が分布の中心で、母集団のデータの分布の特徴を表す一つの要素だから、これを考えることによって、分布の特徴をどのように記述すれば良いのかを理解するためです。

私たちが知っている確率分布は、以下の二項分布です。

$$W(k) = {}_n C_k p^k q^{(n-k)}$$

$$p + q = 1$$

この式を使えば、 n と p を与えて確率分布の形を描くことが出来ます。この式は一つのピークを持つ（単峰形）であることがわかっています。このピークの位置が分布の特徴の一つです。ピークが一つなので、この式が極値となる k を求めれば、ピークの位置がわかります。

$$\frac{dW(k)}{dk} = 0$$

この式、つまり $C_k p^k q^{(1-k)}$ を k で微分すれば良いのですが、これを微分するには、数学的なテクニックが必要です。このテクニックは少し複雑で長い説明になります。そこで、その説明の前に、素朴で感覚的な方法でピークの位置を求めて、その後に数学的な説明をすることにします。

図6は $B\left(n, \frac{2}{3}\right)$ 、つまり、 $p = \frac{2}{3}$ の時の二項分布の $W(k)$ と n, k の関係を図示したものです。

具体的なイメージとしては、表(A)になる確率が $\frac{2}{3}$ 、裏(B)になる確率が $\frac{1}{3}$ になるコインのトスを n 回繰り返した時に、何回 (k) 表が出るかという確率の問題です。図 6 にみられるように、確率のピークは、試行回数(n) の増加に伴って、右に移動します。

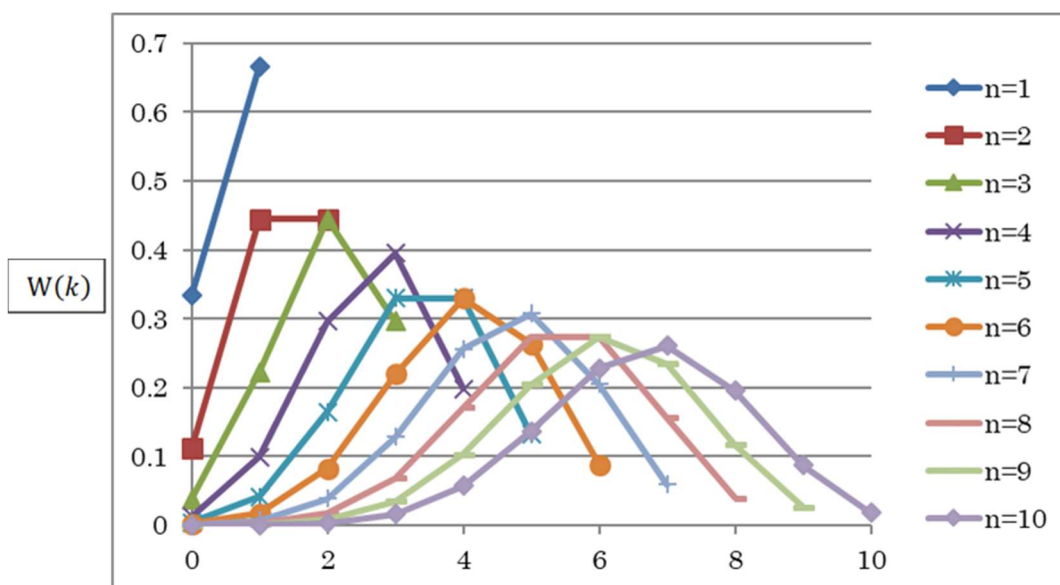


図 6. $W(k) = {}_n C_k p^k q^{(n-k)}, B(n, \frac{2}{3})$

この図を見ると、 $B(9, \frac{2}{3})$ ではピークは $k = 6$ で、 $B(6, \frac{2}{3})$ では $k = 4$ 、 $B(3, \frac{2}{3})$ では $k = 2$ が

ピークです。他のケースではあまりはっきりしませんが、だいたい、 $\frac{k}{n} = p$ となる k つまり

$k = np$ のように見えます。単純に考えて、1回の試行で A になる確率が $\frac{2}{3}$ なのだから、2回それを行えば $k = \frac{2}{3} + \frac{2}{3}$ となり、 n 回試行すれば、最も有りそうな k は np だと考えても良いでしょう。これが「一回の試行で A が起こる確率が p のときに、それを n 回繰り返したと k に、Aが何回出るか」という予測で最も有りそうな回数ということになります。この k のことを期待値と言って、ここでは $\bar{k}$ と表します。

$$\bar{k} = np. \quad (\bar{k} \text{は} k \text{の期待値})$$

感覚的にはこれで良いのですが、きちんと証明する必要があります。すでに述べたようにこの証明は手間がかかるのですが、一応、やります。

上述の感覚的な説明でも用いたように、この場合、期待値は現象が起こる確率とおこる回数の積の総和です。一回も起こらないこともありますから、0 から n 回まで、 $n + 1$ 個の積の和になります。 $\bar{k} = \sum_{k=0}^n W(k) \times k$

$$= \sum_{k=0}^n {}_n C_k p^k q^{(n-k)} \times k$$

$$= \sum_{k=0}^n \frac{kn!}{k!(n-k)!} p^k q^{(n-k)}$$

ここで少しテクニックを使って式を変形します。

$$\begin{aligned} &= \sum_{k=0}^n \frac{kn(n-1)!}{k!((n-1)-(k-1))!} p \times p^{(k-1)} q^{((n-1)-(k-1))} \\ \bar{k} &= \sum_{k=0}^n \frac{n(n-1)!}{(k-1)!((n-1)-(k-1))!} p \times p^{(k-1)} q^{((n-1)-(k-1))} \\ \bar{k} &= np \sum_{k=1}^{n-1} \frac{(n-1)!}{(k-1)!((n-1)-(k-1))!} p^{(k-1)} q^{((n-1)-(k-1))} \end{aligned}$$

$\sum_{k=1}^{n-1} \frac{(n-1)!}{(k-1)!((n-1)-(k-1))!} p^{(k-1)} q^{((n-1)-(k-1))}$ は $n = n-1$ の時の確率の総和だから

$$\sum_{k=1}^{n-1} \frac{(n-1)!}{(k-1)!((n-1)-(k-1))!} p^{(k-1)} q^{((n-1)-(k-1))} = 1$$

したがって

$$\bar{k} = np$$

証明終わり

これで、期待値が $W(np)$ であることはわかりますが、 $(np - W(np))$ がピークであることを証明するには、 $k = np$ で以下の微分式が成立つことを証明しなくてはなりません。

$$\frac{dW(k)}{dk} = 0$$

$$W(k) = {}_n C_k p^k q^{(n-k)}$$

$$p + q = 1$$

$W(k) = {}_n C_k p^k q^{(n-k)}$ を k についての合成関数に変形すれば、合成関数の微分として計算できます。この方法は入れ子型の合成関数の微分法を知っていれば簡単に解けます。まず、この方法で微分します。

入れ子型の合成関数の微分の公式は以下の通りです。

$$\frac{df(g(x))}{dx} = \frac{df(g(x))}{dg(x)} \frac{dg(x)}{dx}$$

二項分布の公式は以下の通りです。

$$W(k) = {}_n C_k p^k q^{(n-k)}$$

$$= \frac{n!}{k!(n-k)!} p^k q^{(n-k)}$$

$$p + q = 1$$

ここで、このままでは入れ子型の合成関数の形にならないので、両辺の対数をとります。

$$\log W(x) = \log(n!) - \log(x!) - \log(n-x)! + k \log(p) + (n-x) \log(q)$$

その後、もう一度対数から、普通の形に戻すと

$$W(x) = e^{(\log(n!) - \log(x!) - \log(n-x)! + x \log(p) + (n-x) \log(q))}$$

$$g(x) = \log(n!) - \log(x!) - \log(n-x)! + x \log(p) + (n-x) \log(q)$$

$$W(g(x)) = e^{g(x)}$$

のように、入れ子型の合成関数の形になります。

$$\frac{dW(g(x))}{dg(x)} = e^{g(x)}$$

$$\frac{dg(x)}{dx} = -\log x + \log(n-x) + \log(p) - \log(q)$$

$$\begin{aligned} \frac{dW(g(x))}{dx} &= \frac{dW(g(x))}{dg(x)} \frac{dg(x)}{dx} = e^{g(x)} (-\log x + \log(n-x) + \log(p) - \log(q)) \\ &= e^{g(x)} \left(\log \frac{(n-x)p}{x(1-p)} \right) \end{aligned}$$

極値を与える、 x を求めます。

$$\frac{dW(g(x))}{dx} = e^{g(x)} \left(\log \frac{(n-x)p}{x(1-p)} \right) = 0$$

$$e^{g(x)} > 0 \quad (\text{not } 0)$$

$$\frac{(n-x)p}{x(1-p)} = 1$$

$$(n-x)p = x(1-p)$$

$$np - xp = x - xp$$

$$x = np$$

証明終わり

これで、ピークの値が、期待値であり、平均値だということがわかります。入れ子型の合成関数の微分の公式を知っていることが前提になりますが、この証明が最も普通の証明です。テクニックとしては、対数をとってそれを元に戻すというやり方を思いつく必要があります。少し工夫が必要ですが、入れ子型の合成関数の微分の公式を使わない素朴な方法もあります。そちらをやってみます。そんな必要はないという人は読み飛ばしてください。まず、二項分布から出発して、両辺の対数をとるところまでは同じです。

$$W(k) = {}_n C_k p^k q^{(n-k)}$$

$$= \frac{n!}{k!(n-k)!} p^k q^{(n-k)}$$

$$p + q = 1$$

$$\log W(k) = \log(n!) - \log(k!) - \log(n-k)! + k \log(p) + (n-k) \log(q)$$

k は不連続な変数の記号として使われることが多いので、連続関数として扱おうという意味で、 $k = x$ と書き換えます。

$$\log W(x) = \log(n!) - \log(x!) - \log(n-x)! + k \log(p) + (n-x) \log(q)$$

式 12

ここからが違って、 $\log W(x)$ が最大の時に $W(x)$ も最大になるという関係を使います。もう一つ、

$$\lim_{x \rightarrow \infty} \log x! = \lim_{x \rightarrow \infty} \int_1^x \log t \, dt$$

という式も使います。この式の意味は、 x が十分大きい時に、 $\int_1^x \log t \, dt \cong \log x!$ だということですが、この等式は比較的単純で、感覚的にそうだと受け入れる人もいるでしょうし、よく意味が解らないという人もいます。この方法の証明は、大変に複雑で手間がかかります。ただ、その証明法は微分や積分の証明によく使われる技術で、これを知ると、微分に詳しくない読者（たとえば高校で微分について詳しく学んでいない読者）も微分積分が感覚的にわかって、微分について苦手意識を解消できるかもしれません。すでに指摘したように、そんな証明は退屈で嫌いだという読者はこの部分を読み飛ばしてください。

証明するのは

$$\lim_{x \rightarrow \infty} \int_1^x \log t \, dt = \lim_{x \rightarrow \infty} \log x!$$

ですが、これを変形して、

$$\lim_{x \rightarrow \infty} \frac{\int_1^x \log_e t \, dt}{\log_e x!} = 1$$

を証明のゴールとします。この等式の中に

$$\int_1^x \log t \, dt$$

があります。この式の意味は図 7 の $\log t$ の曲線と x 軸に挟まれた部分の面積です。次のように表せます。

$$S_t = \int_1^x \log t \, dt$$

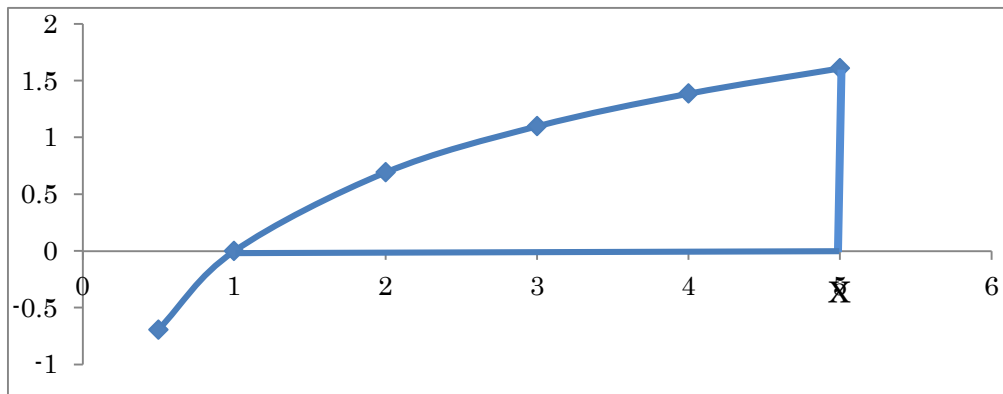


図 7. $\int_1^X \log t dt$ が示す面積

この図にいくつかの矩形を重ね合わせたものが、図 8-1 です。さらに、図 8-2、図 8-3 はそれぞれの矩形を取り出してみました。図 8-2 は変動の上限で図 8-3 は変動の下限です。

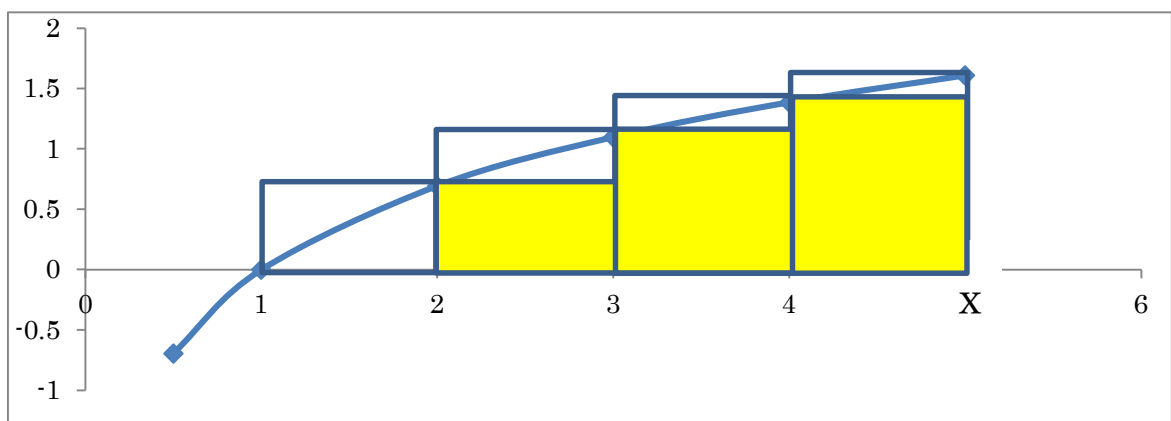


図 8-1. 区分法による面積計算の図

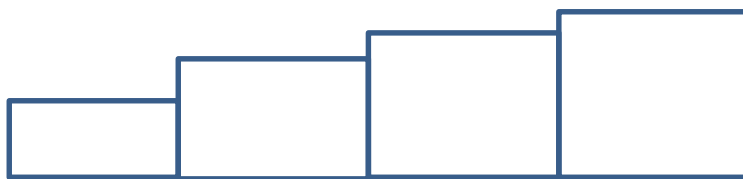


図 8-2. 変動の上限を表す矩形の面積

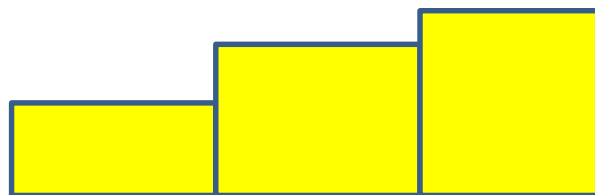


図 8-3. 変動の下限を表す矩形の面積

それぞれの矩形の幅は1ですから、例えば図 8-2 の一番左の矩形の面積は $\log 2$ 、その右側に
ある矩形の面積は、それぞれ $\log 3$, $\log 4$ and $\log 5$ で、その総和は

$$S_l = \log 2 + \log 3 + \log 4 + \log 5 = \log 5!$$

となり、 $t = x$ とすると

$$S_l = \log x!$$

同様に、図 8-3 の黄色い矩形の問責の総和は

$$S_s = \log(x-1)!$$

図の 8-1 に示したように、これらの面積の大きさの関係は次のようになっています

$$S_s \leq S_t \leq S_l$$

$$\log(x-1)! \leq \int_1^x \log t dt \leq \log x!$$

ここで、 $x > 1$ だから、

$$\log x! > 0$$

す。これが正の値なので、これで等式の各項を割っても、不等号の向きは変わりません。

$$\frac{\log(x-1)!}{\log x!} \leq \frac{\int_1^x \log t dt}{\log x!} \leq \frac{\log x!}{\log x!}$$

一番右の式が

$$\frac{\log x!}{\log x!} = 1$$

であることは自明です。一番左の式は

$$\begin{aligned} & \frac{\log(x-1)!}{\log x!} \\ &= \frac{\log(x-1) + \log(x-2) + \cdots + \log 1}{\log x!} \\ &= \frac{\log x + \log(x-1) + \log(x-2) + \cdots + \log 1 - \log x}{\log x!} \\ &= \frac{\log x! - \log x}{\log x!} \\ &= 1 - \frac{\log x}{\log x!} \end{aligned}$$

となりますが、著者には $\lim_{x \rightarrow \infty} \frac{\log x}{\log x!} = 0$ は自明ですが、一般的には証明が必要でしょう。

証明

$$\lim_{x \rightarrow \infty} \frac{\log x}{\log x!} = 0$$

まず、以下の式を証明します。

$$\log k! > \log k + \log(k-1) + \cdots + \log \left\lfloor \frac{k}{2} \right\rfloor > \left(\frac{k}{2} - 1 \right) \log \left\lfloor \frac{k}{2} \right\rfloor$$

この式で $\left\lfloor \frac{k}{2} \right\rfloor$ は $\frac{k}{2}$ を超えない最大の整数の意味で、 $k=5$ の時 $\left\lfloor \frac{k}{2} \right\rfloor = 2$ 、 $k=4$ の時 $\left\lfloor \frac{k}{2} \right\rfloor = 2$ です。

$\log k! > \log k + \log(k-1) + \cdots + \log \left\lfloor \frac{k}{2} \right\rfloor$ の意味は、この不等式の左右はともに正ですから、部分は全体よりも小さいという意味です。

$$\log k + \log(k-1) + \cdots + \log \left\lfloor \frac{k}{2} \right\rfloor$$

が下の式の一部です。

$$\log k! = \log k + \log(k-1) + \cdots + \log 1$$

したがって

$$\log k! > \log k + \log(k-1) + \cdots + \log \left\lfloor \frac{k}{2} \right\rfloor$$

次に

$\log k + \log(k-1) + \cdots + \log \left\lfloor \frac{k}{2} \right\rfloor > \left(\frac{k}{2} - 1 \right) \log \left\lfloor \frac{k}{2} \right\rfloor$ を証明します。この式は、9図の黄色い矩形

形の面積の和と青い矩形の面積の和の比較です

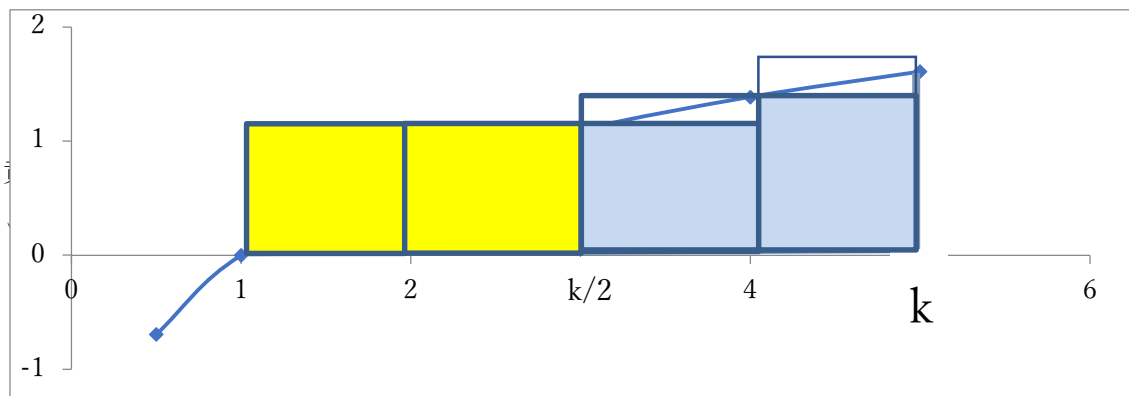


図 9. 矩形の面積の比較

この図から、黄色の矩形の面積の和よりも青い矩形の面積の和が大きいことは明らかです。したがって

$$\log k + \log(k-1) + \cdots + \log \left\lfloor \frac{k}{2} \right\rfloor > \left(\frac{k}{2} - 1 \right) \log \left\lfloor \frac{k}{2} \right\rfloor$$

であり、

$$\log k! > \log k + \log(k-1) + \cdots + \log \left\lfloor \frac{k}{2} \right\rfloor > \left(\frac{k}{2} - 1 \right) \log \left\lfloor \frac{k}{2} \right\rfloor$$

の関係が証明されます。したがって、

$$\log k! > \left(\frac{k}{2} - 1\right) \log \left[\frac{k}{2}\right]$$

k を x と書いて

$$\log x! > \left(\frac{x}{2} - 1\right) \log \left[\frac{x}{2}\right]$$

この関係は $x \rightarrow \infty$ のように x を無限大に拡大しても変わらないので

$$\lim_{x \rightarrow \infty} \log x! > \lim_{x \rightarrow \infty} \left(\frac{x}{2} - 1\right) \log \left[\frac{x}{2}\right]$$

不等式の左辺も右辺も 1 より大きいので

$$\lim_{x \rightarrow \infty} \frac{1}{\log x!} < \lim_{x \rightarrow \infty} \frac{1}{\left(\frac{x}{2} - 1\right) \log \left[\frac{x}{2}\right]} = 0$$

一方

$$\log x > 0 \text{ and } \frac{1}{\log x!} > 0$$

$$\frac{\log x}{\log x!} > 0$$

だから、これらを組み合わせると

$$0 \leq \lim_{x \rightarrow \infty} \frac{\log x}{\log x!} \leq \lim_{x \rightarrow \infty} \frac{\log x}{\left(\frac{x}{2} - 1\right) \log \frac{x}{2}} = \lim_{x \rightarrow \infty} \frac{\log x}{\left(\frac{x}{2} - 1\right) (\log x - \log 2)} = \lim_{x \rightarrow \infty} \frac{1}{\left(\frac{x}{2} - 1\right) \left(1 - \frac{2}{\log x}\right)} = 0$$

$$0 \leq \lim_{x \rightarrow \infty} \frac{\log x}{\log x!} \leq 0$$

$$\lim_{x \rightarrow \infty} \frac{\log x}{\log x!} = 0$$

となるので、

証明終わり

$$\lim_{x \rightarrow \infty} \frac{\log x}{\log x!} = 0$$

が証明されます。こういう方法を挟み撃ち法と言います。これによって

$$\lim_{x \rightarrow \infty} \left(1 - \frac{\log x}{\log x!}\right) = 1$$

となりますが

$$1 - \frac{\log x}{\log x!} = \frac{\log x! - \log x}{\log x!} = \frac{\log \frac{x!}{x}}{\log x!} = \frac{\log(x-1)!}{\log x!}$$

なので

$$\lim_{x \rightarrow \infty} \frac{\log(x-1)!}{\log x!} = 1$$

$$1 = \lim_{x \rightarrow \infty} \frac{\log(x-1)!}{\log x!} \leq \lim_{x \rightarrow \infty} \frac{\int_1^x \log t dt}{\log x!} \leq \lim_{x \rightarrow \infty} \frac{\log x!}{\log x!} = 1$$

ここも挟み撃ち法を使って

$$\lim_{x \rightarrow \infty} \frac{\int_1^x \log t dt}{\log x!} = 1$$

$$\lim_{x \rightarrow \infty} \int_1^x \log t dt = \lim_{x \rightarrow \infty} \log x!$$

となるので、すべての証明が終了して

$$\log x! \doteq \int_1^x \log t dt$$

となります。そこで式 1 2 に戻って

$$\log W(x) = \log(n!) - \log(x!) - \log(n-x)! + x \log(p) + (n-x) \log(q)$$

$$\doteq \log(n!) - \int_1^x \log t dt - \int_1^{n-x} \log t dt + x \log p + (n-x) \log q$$

$x \rightarrow \infty$ ならば、上のような変形が出来て、これを微分します。

$$\log W(x) = \log(n!) - \int_1^x \log t dt - \int_1^{n-x} \log t dt + x \log p + (n-x) \log q$$

右辺の第一項は x を含まない定数項だから微分すればなくなって、第二項と第三項の微分ですが

$$\int \log t dt = t \log t - t + C$$

$$\therefore \int_1^x \log t dt = x \log x - x - 1 \log 1 + 1 = x \log x - x + 1$$

$$\therefore \frac{d(\int_1^x \log t dt)}{dx} = 1 \times \log x + x \frac{1}{x} - 1 = \log x$$

ですから、右辺の各項の微分はつぎのようになります。

$$\frac{d(\log n!)}{dx} = 0, \frac{d(x \log p)}{dx} = \log p, \frac{d((n-x) \log q)}{dx} = \log q = \log(1-p)$$

ですから微分全体は

$$\begin{aligned} \{\log W(x)\}' &= -\log x + \log(n-x) + \log p - \log(1-p) \\ &= \log \frac{(n-x)p}{x(1-p)} \end{aligned}$$

となります。

$$\log \frac{(n-x)p}{x(1-p)} = 0$$

ならば $\log 1 = 0$ だから

$$\frac{(n-x)p}{x(1-p)} = 1,$$

$$\frac{(n-x)p}{x(1-p)} = 1$$

$$(n-x)p = x(1-p)$$

$$np - xp = x - xp$$

$$x = np$$

$(\log W(x))'$ は単調減少関数だから、 $(\log W(x))' = 0$ すなわち、 $x = np$ の時に最大値になります。

これによって、期待値＝平均値がピークの値と一致することがわかります。

次に分布の特徴を表しているのは、単峰形の尖り方と幅です。これは、全体を代表するピークを与える x と個々のデータの距離の平均で表せば良いでしょう。しかし、単純に平均値からの距離を合計すると次の式のように、その合計は 0 になってしまいます。

$$M(x - \bar{x}) = E(x - \bar{x}) = \frac{\sum_{i=1}^n (x - \bar{x})}{n} = 0$$

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

そこで、次の式のように 2 次の積率を計算します。つまり、平均値周りの分散を計算します。

$$M((x - \bar{x})^2) = \frac{\sum_{i=1}^n (x - \bar{x})^2}{n}$$

x : 個々のデータの値

$\bar{x}$: 平均値

n : サンプルサイズ (データの数)

ここで、議論しようとするのは、標本集団の平均値周りの 2 次の積率を母集団の 2 次の積率の推定値として使えるかという議論です。この議論は、数学的な論理によって説明することもできますが、もっと、大切なことは感覚的に納得することです。そこで、実際に計算してみることにします。つまり、 n, p の値を与えて、 $B(n, p)$ について、 $p(k)$ を計算し、そこから、二次の積率の期待値を計算してみることにします。

具体的な作業

二項分布するモデル $B(n, p)$ について、そのモデル通りに理想的に標本 (データ) が得られるものとして、それらのデータから平均や 2 次の積率 (分散) を求めて、モデルが理論的に与える平均値や分散と一致するかどうかを考えます。1 試行の中で n を変化させるという

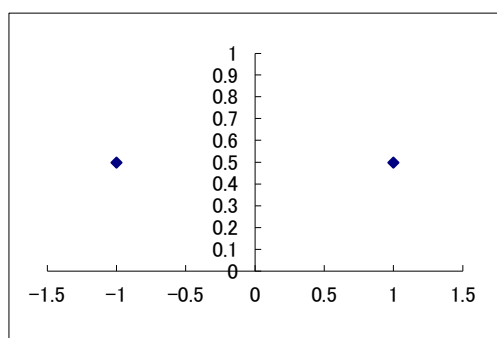
ことは、 n 回の繰り返しを 1 セットとし、そのセットごとに平均や SS を求めて、それを無限回繰り返すというイメージです。この値を用いて、考えられる方法で、2 次の積率（母集団のばらつきの指標）を計算してみます。これを無限回、繰り返した時の期待値を計算して、それが既知の母集団の積率と一致するかどうかを検討してみます。

最も簡単なモデルで試してみます。

1/2 の確率で -1、1/2 の確率で 1 となる事象

例として、コインを投げて表なら 1 円もらえて裏ならば 1 円払うことを考えます。この場合平均値は 0 であり 2 次の積率も 1 であることは自明でしょう。

単純に考えれば、標本集団の積率 (SS/n) が母集団の積率であると考えて良さそうですが、本当にそうでしょうか、試してみましょう。ついでに $SS/(n-1)$ についても計算してみます。



繰り返し 1 の二項分布の例

1 回の繰り返しの場合

$n=1$

実際のデータ	確率	平均値	SS	SS/n	SS/(n-1)
-1	1/2	-1	0	0	-
1	1/2	1	0	0	-
期待値			0	0	

2 回の繰り返しの場合

$n=2$

実際のデータ	確率	平均値	SS	SS/n	SS/(n-1)
-1 -1	1/4	-1	0	0	0
-1 1	1/2*	0	2**	1	2
1 1	1/4	1	0	0	0
期待値			1	1/2***	1****

* : ${}_2C_1(1/2)*(1/2)$

** : $(-1-0)^2 + (1-0)^2 = 2$

$$*** : 0 \times 1/4 + 1 \times 1/2 + 0 \times 1/4 = 1/2$$

$$**** : 0 \times 1/4 + 2 \times 1/2 + 0 \times 1/4$$

以下、実際に計算してみてイメージをつかみましょう

3回の繰り返しの場合

n=3

実際のデーター	確率	平均値	SS	SS/n	SS/(n-1)
- 1 - 1 - 1	1/8	- 1	0	0	0
- 1 - 1 1	3/8	- 1/3	24/9*	8/9**	12/9***
- 1 1 1	3/8	1/3	24/9	8/9	12/9
1 1 1	1/8	1	0	0	0
期待値			2	2/3****	1*****

$$* : (-1 - (-1/3))^2 + (-1 - (-1/3))^2 + (1 - (-1/3))^2 = 24/9$$

$$** : (24/9)/3$$

$$*** : (24/9)/2$$

$$**** : 0 \times 1/8 + (8/9) \times (3/8) + (8/9) \times (3/8) + 0 \times (1/8)$$

$$***** : 0 \times 1/8 + (12/9) \times (3/8) + (12/9) \times (3/8) + 0 \times (1/8)$$

4回の繰り返しの場合

n=4

実際のデーター	確率	平均値	SS	SS/n	SS/(n-1)
- 1 - 1 - 1 - 1	1/16	- 1	0	0	0
- 1 - 1 - 1 1	4/16	- 1/2	48/16	12/16	16/16
- 1 - 1 1 1	6/16	0	4	1	4/3
- 1 1 1 1	4/16	1/2	48/16	12/16	16/16
1 1 1 1	1/16	1	0	0	0
期待値			3	3/4	1

わかったこと、

この表を見るとSS/n（標本を集団の平均まわりの積率）は母集団の2次の積率と一致しません。母集団の2次の積率をあらわしているのは、SS/(n-1)です。意外なことが発見されました。

何でこんなことになるのか

標本集団ではそれぞれの試行（繰り返しの 1 回分）の平均値と個々のデーターの差を求めています。この平均値を求めるときには、差を求めたデーターそのものも用いられています。したがって、その平均値は、その分だけ、母集団の平均値に近づいているのです。

2 回繰り返しのときは $1/2$ 、3 回繰り返しのときは $1/3$ 、4 回のときは $1/4$ 近づいていることに注目しましょう。

2 項分布を考えると、本当は n が大きくなると、分布範囲が広がって、平均値や分散も大きくなります。ここに示した例では、回数そのものではなくて、平均化しているので -1 から 1 の間にある、いわば、初めから分散 1 となるような形で例示しているので多少説明としてはごまかしているところがあります。後半の説明で、2 項分布の数式から母集団の平均値分散を計算すると、正確に理解できます。

この場合の $n-1$ のように、母集団の 2 次の積率（標本集団の分散）を求めるために、SS を割る数、いわば SS の σ^2 に対する比率のことを**自由度**といいます。自由度をどのようにするかは、場合によって異なります。いつでも標本数 -1 と覚えてはいけません。

繰り返しの数が違うと繰り返しごとに求められる平均値の分布はどのように変化するでしょうか。

前に行った確立 $1/2$ の例から標本集団の平均値とその平均値が出る確率の図を作ってみます。横軸が求められる平均値の値、縦軸がその値となる確率です。

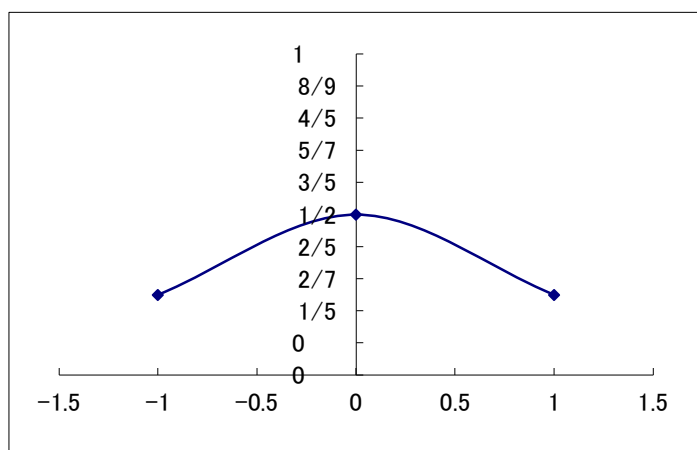


図 10. $n=2$ の場合

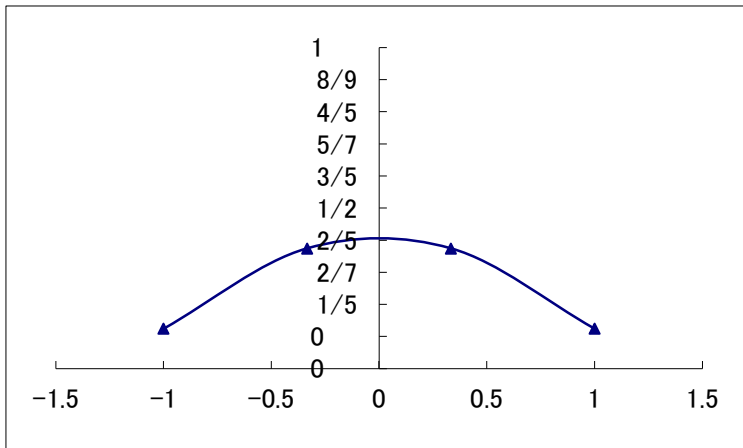


図 11. $n=3$ の場合

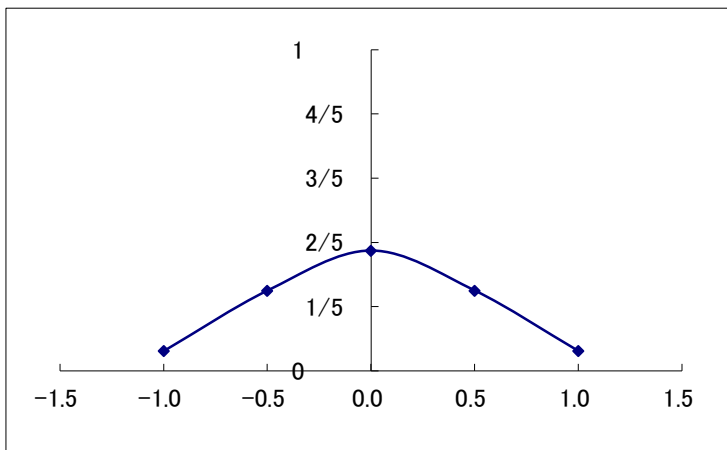


図 12. $n=4$ の場合

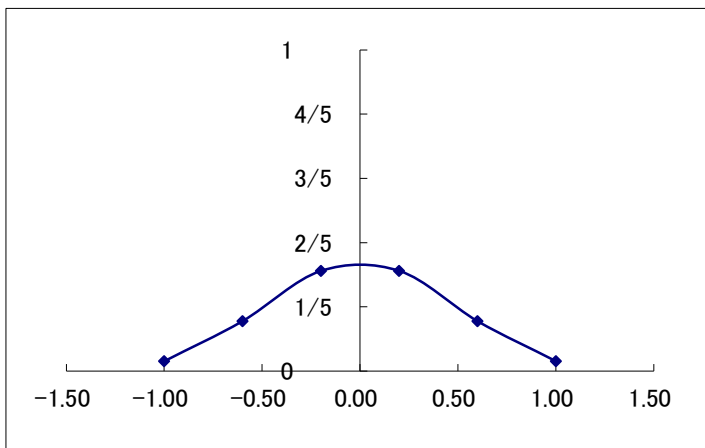


図 13. $n=5$ の場合

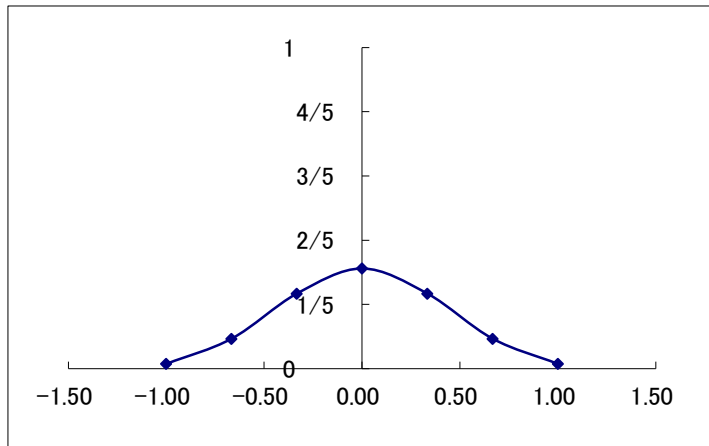


図 14 $n=6$ の場合

n が大きくなると両側の値が小さくなり、尖った形になり、正規分布に近づきます。 n が無限大の時の二項分布が正規分布です。

ここまでは極めて単純なケースを用いて考察を行ってきました。一般的な確率事象はもう少し複雑です。そのような場合にも今まで考察してきた結果があてはまるか確認をしてみましょう。実際計算してみると感覚が身につきます。

母集団の確率分布がゆがんでいる場合

－ 1 となる確率が 1 となる確率の 2 倍ある場合を考えます。

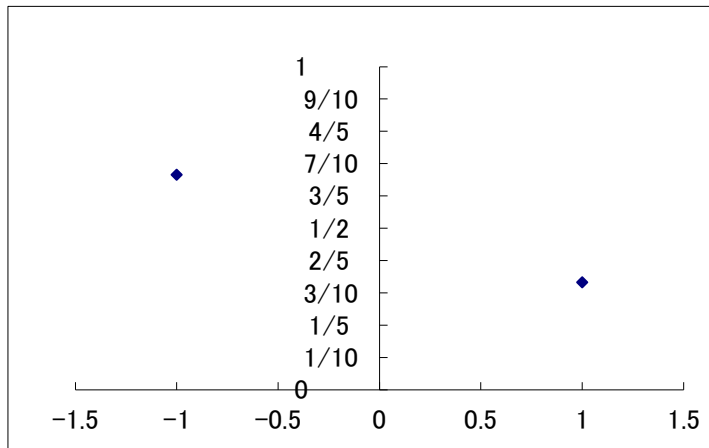


図 15. $n=1$ $p=1/3$ の二項分布

平均値は $-1/3$ 2 次の積率は $8/9$

1 回の繰り返しの場合

$n=1$

実際のデータ	確率	平均値	SS	SS/n	SS/(n-1)
-1	$2/3$	$-1/3$	$2/3$	$2/3$	$2/3$
1	$1/3$	$-1/3$	$2/3$	$2/3$	$2/3$

- 1	2/3	- 1	0	0	-
1	1/3	1	0	0	-
期待値			0	0	

2 回の繰り返しの場合

n= 2

実際のデーター	確率	平均値	SS	SS/n	SS/(n-1)
- 1 - 1	4/9	- 1	0	0	0
- 1 1	4/9*	0	2**	1	2
1 1	1/9	1	0	0	0
期待値			8/9	4/9***	8/9****

$$* : {}_2C_1(2/3) * (1/3)$$

$$** : (-1 - 0)^2 + (1 - 0)^2 = 2$$

$$*** : 0 \times 4/9 + 1 \times 4/9 + 0 \times 1/9 = 4/9$$

$$**** : 0 \times 4/9 + 2 \times 4/9 + 0 \times 1/9$$

3 回の繰り返しの場合

n= 3

実際のデーター	確率	平均値	SS	SS/n	SS/(n-1)
- 1 - 1 - 1	8/27	- 1	0	0	0
- 1 - 1 1	12/27	-1/3	24/9*	8/9**	12/9***
- 1 1 1	6/27	1/3	24/9	8/9	12/9
1 1 1	1/27	1	0	0	0
期待値			2	48/81****	8/9*****

$$* : (-1 - (-1/3))^2 + (-1 - (-1/3))^2 + (1 - (-1/3))^2 = 24/9$$

$$** : (24/9)/3$$

$$*** : (24/9)/2$$

$$**** : 0 \times 8/27 + (8/9) \times (12/27) + (8/9) \times (6/27) + 0 \times (1/27)$$

$$***** : 0 \times 8/27 + (12/9) \times (12/27) + (12/9) \times (6/27) + 0 \times (1/27)$$

4 回の繰り返しの場合

n= 4

実際のデーター	確率	平均値	SS	SS/n	SS/(n-1)
- 1 - 1 - 1 - 1	16/81	- 1	0	0	0
- 1 - 1 - 1 1	32/81	-1/2	48/16	12/16	16/16

-1	-1	1	1	24/81	0	4	1	4/3
-1	1	1	1	8/81	1/2	48/16	12/16	16/16
1	1	1	1	1/81	1	0	0	0
期待値					3	2/3	8/9	

$B(6, \frac{1}{3})$ について図を作ってみます。横軸が求められる平均値の値、縦軸がその値となる確率です。

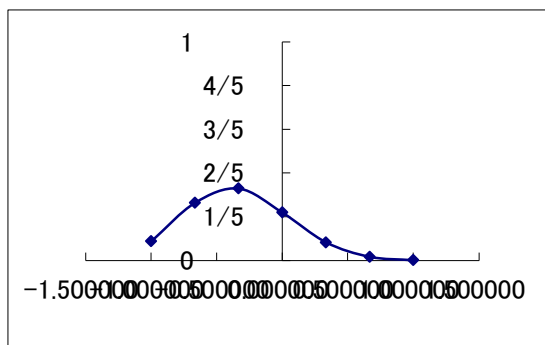


図 16. $n=6$ の時

わかったこと、

$SS/(n-1)$ が、母集団の2次の積率(母集団の原点まわりのバラツキの指標)表わしています。また、 n が大きくなると正規分布に近づきます。

推定された母集団の平均値の確からしさを考える

n が大きくなると、母集団の平均値の推定値の確率分布が尖ってくる。正確にいうと、予測値が母集団の平均値の周りの狭い範囲に集まってきます(これを中心極限定理といいます。無限大回繰り返せば幅が0となり、予測値は母集団の平均値そのものになるはずです)。だから、 n が大きいほど予測値の確からしさは増します。このことは、よりたくさんのデータから母集団の推定値を推測したほうが確かだという経験則にも一致します。これをある確率での信頼範囲という見方をすると、その範囲は母集団の真の平均の周りに在って、その範囲の広がり n の増加に伴って狭くなるということです。あるいは、サンプルサイズの増加に伴って、サンプル集団の平均値が母集団の平均値に近づくと考えても良いでしょう。

そこで、今までやって来たことと同様に、母集団の分布がわかっている事例を考えて、それについて n 回の繰り返しで求めた平均値の推定値と母集団の平均値の差を求め、その2次の積率が n の増加によってどのように変化するかを考えます。

我々が知っているのは2項分布の確率分布です。二項分布では平均値(μ)が np となるということもすでに学習しました。ある事象によって得られる値とその事象が起こる確率の総和ですから、二項分布の二次の積率の期待値は以下の式で表せます。

$$\begin{aligned}
E((k - \mu)^2) &= \sum_{k=0}^n (k - np)^2 W_{(k)} \\
&= \sum_{k=0}^n (k^2 - 2knp + n^2 p^2) W_{(k)} \\
&= \sum_{k=0}^n k^2 W_{(k)} - \sum_{k=0}^n 2knp W_{(k)} + \sum_{k=0}^n n^2 p^2 W_{(k)} \\
&= \sum_{k=0}^n k^2 W_{(k)} - 2np \sum_{k=0}^n k W_{(k)} + n^2 p^2 \sum_{k=0}^n W_{(k)}
\end{aligned}$$

右辺の第一項、 $\sum_{k=0}^n k^2 W_{(k)}$ は k^2 の期待値です。

$$\sum_{k=0}^n k^2 W_{(k)} = E(k^2)$$

第二項の $\sum_{k=0}^n k W_{(k)}$ は二項分布 $B(n, p)$ の母集団の平均値の推定値 (μ) ですから

$$\mu = np$$

$$\sum_{k=0}^n W_{(k)} = 1$$

∴ 確率の総和は 1

$$\begin{aligned}
E((k - \mu)^2) &= \sum_{k=0}^n k^2 W_{(k)} - 2n^2 p^2 + n^2 p^2 \\
&= E(k^2) - n^2 p^2
\end{aligned}$$

$$E\{k(k-1)\} = E(k^2 - k) = E(k^2) - E(k)$$

$$\begin{aligned}
E\{k(k-1)\} &= \sum_{k=0}^n k(k-1) \cdot {}_n C_k p^k q^{n-k} \\
&= \sum_{k=0}^n \frac{k(k-1)n!}{(n-k)!k!} \cdot {}_n C_k p^k q^{n-k}
\end{aligned}$$

$$\begin{aligned}
&\sum_{k=0}^n \frac{n(n-1)k(k-1)(n-2)!}{(n-2-(k-2))!k!} \cdot {}_n C_k p^k q^{n-k} \\
&n(n-1) \sum_{k=0}^n \frac{k(k-1)(n-2)!}{(n-2-(k-2))!k(k-1)(k-2)!} \cdot {}_n C_k p^k q^{n-k} \\
&n(n-1) \sum_{k=0}^n \frac{(n-2)!}{((n-2)-(k-2))!(k-2)!} \cdot {}_n C_k p^2 p^{k-2} q^{(n-2)-(k-2)} \\
&n(n-1)p^2 \sum_{k=0}^n \frac{(n-2)!}{((n-2)-(k-2))!(k-2)!} \cdot {}_n C_k p^{k-2} q^{(n-2)-(k-2)}
\end{aligned}$$

ここで $\sum_{k=0}^n \frac{(n-2)!}{((n-2)-(k-2))!(k-2)!} \cdot {}_n C_k p^{k-2} q^{(n-2)-(k-2)}$ は $B(n-2, p)$ の確率の総和ですから、

$$\sum_{k=0}^n \frac{(n-2)!}{((n-2)-(k-2))!(k-2)!} \cdot {}_n C_k p^{k-2} q^{(n-2)-(k-2)} = 1$$

です。したがって

$$E\{k(k-1)\} = n(n-1)p^2$$

一方、 $E\{k(k-1)\} = E(k^2) - E(k)$ ですから

$$E(k^2) = E\{k(k-1)\} + E(k)$$

$$E(k^2) = n(n-1)p^2 + np = n^2p^2 - np^2 + np$$

したがって

$$\begin{aligned} E((k-\mu)^2) &= \\ &= E(k^2) - n^2p^2 \\ &= n^2p^2 - np^2 + np - n^2p^2 \\ &= np(1-p) \\ &= npq \\ &\because p+q=1 \end{aligned}$$

この公式で母集団の平均値周りの二次の積率（平均値の推定値の分散）を求めることが出来ます。

補足（以下の公式はここでの説明と直接関係はありませんが、平均値の簡便な計算法としてよく使われるものなので覚えておきましょう。）

$$\begin{aligned} E((k-\mu)^2) &= \\ &= E(k^2) - n^2p^2 \\ &= E(k^2) - \mu^2 \\ &= E(k^2) - E(k)^2 \\ E((k-\mu)^2) &= V_k \\ V_k &= E(k^2) - E(k)^2 \\ V_k &: \text{variance of } k \end{aligned}$$

式 13

元に戻って、母集団の平均値周りの2次の積率を求める式ができたので、これを使って簡単な事例について、2次の積率を計算します。

今まで用いた考察のために用いてきた 1/2 確立で起こる事象のモデルをそのまま用いることにします。下には 1/2 確率3回の繰り返しの場合についての場合を示しました。この表の、下線引いた値（推定される平均値）と、母集団の平均値0の間の差を求め、その2乗にそれが起こる確率を乗じて、その総和を求め、それを母集団の平均値の分散として、その値と繰り返しの回数nの関係を考察します。

3回の繰り返しの場合

n=3

実際のデータ	確率	平均値	SS	SS/n	SS/(n-1)
0	1/8	0	0	0	0
1	3/8	1	3	1	1.5
2	3/8	2	3	1	1.5
3	1/8	3	3	1	1.5

- 1 - 1 - 1	1/8	<u>- 1</u>	0	0	0
- 1 - 1 1	3/8	<u>- 1/3</u>	24/9*	8/9**	12/9***
- 1 1 1	3/8	<u>1/3</u>	24/9	8/9	12/9
1 1 1	1/8	<u>1</u>	0	0	0
期待値			2	2/3*****	1*****

この場合求める値の計算は以下のとおり

$$(-1)^2 \times (1/8) + (-1/3)^2 \times (3/8) + (1/3)^2 \times (3/8) + 1^2 \times (1/8) = 1/3$$

$$\begin{aligned} & (-1)^2 \cdot \frac{1}{8} + \left(-\frac{1}{3}\right)^2 \cdot \frac{3}{8} + \left(\frac{1}{3}\right)^2 \cdot \frac{3}{8} + (1)^2 \cdot \frac{1}{8} \\ &= \frac{1}{8} + \frac{1}{24} + \frac{1}{24} + \frac{1}{8} = \frac{3+1+1+3}{24} = \frac{1}{3} \end{aligned}$$

$B\left(n, \frac{1}{2}\right) (\sigma^2=1)$ について、 $n=4$ までの計算結果を示します。

n=1	1	$\sigma^2/1$
n=2	1/2	$\sigma^2/2$
n=3	1/3	$\sigma^2/3$
n=4	1/4	$\sigma^2/4$

$B\left(n, \frac{1}{3}\right)$ 、 $(\sigma^2=8/9)$ については

n=1	8/9	$\sigma^2/1$
n=2	4/9	$\sigma^2/2$
n=3	8/27	$\sigma^2/3$
n=4	8/36	$\sigma^2/4$

となります。実際にやってみるともっとよくわかります。

推測

母集団の真の平均値と推定された平均値の真の平均値周りの2次の積率 $E((M - \mu)^2)$

は、 $\frac{\sigma^2}{n}$ で求められそうです。

$$E((M - \mu)^2) = \frac{\sigma^2}{n}$$

M : 標本集団の平均値

μ : 母集団の平均値

式 14

III-2-2-2. 代数的な証明

母集団の2次積率の推定値である σ^2 を個々のデータに基づく期待値として計算することを

考えます。1セットの中の試行数を n として、これを m セット繰り返すことを考えます。
簡便化のために母集団の平均を $\mu = 0$ として、無限回のセットの繰り返しを考えます。

$$\begin{aligned}\sigma^2 &= \lim_{m \rightarrow \infty} \frac{1}{m} \sum_{j=1}^m \frac{1}{n} \sum_{i=1}^n (M_j + e_{ij})^2 \\ &= \lim_{m \rightarrow \infty} \frac{1}{m} \sum_j M_j^2 + 2 \lim_{m \rightarrow \infty} \frac{1}{m} \frac{1}{n} \sum_j \left(M_j \sum_{i=1}^n e_{ij} \right) + \lim_{m \rightarrow \infty} \frac{1}{m} \frac{1}{n} \sum_{j=1}^m \sum_{i=1}^n e_{ij}^2\end{aligned}$$

n : 1セットの中の試行回数

m : セットの数

M : 1セットの中の平均

e_{ij} : 平均値からの偏差

$\mu = 0$ なので右辺の第一項 $\lim_{m \rightarrow \infty} \frac{1}{m} \sum_j M_j^2 = E((M - \mu)^2)$,

e_{ij} は平均値からの偏差で $\sum_{i=1}^n e_{ij} = 0$ だから第二項 $\lim_{m \rightarrow \infty} \frac{1}{m} \frac{1}{n} \sum_j \left(M_j \sum_{i=1}^n e_{ij} \right) = 0$

$\sum_{i=1}^n e_{ij}^2 = SS$ だから、第三項 $\lim_{m \rightarrow \infty} \frac{1}{m} \frac{1}{n} \sum_{j=1}^m \sum_{i=1}^n e_{ij}^2$ は $\frac{SS}{n}$ の推定値で一方、母集団の分散は

$$\frac{SS}{n-1} = \sigma^2$$

だから

$$\lim_{m \rightarrow \infty} \frac{1}{m} \frac{1}{n} \sum_{j=1}^m \sum_{i=1}^n e_{ij}^2 = \sigma^2 - \frac{1}{n} \sigma^2$$

これらを総合すると

$$\sigma^2 = \lim_{m \rightarrow \infty} \frac{1}{m} \sum_{j=1}^m \frac{1}{n} \sum_{i=1}^n (M_j + e_{ij})^2 = E(M^2) + \sigma^2 - \frac{1}{n} \sigma^2$$

となり、

$$E(M^2) = \frac{1}{n} \sigma^2$$

証明終わり

これは2次の積率なので、平均値の推定値の分布範囲はルートを開いて以下の式になります。これを標準誤差(standard error)と言います。

$$\sqrt{E(M^2)} = \frac{\sigma}{\sqrt{n}}$$

式 15

二項分布の特徴の要約

1. 分布の中心と分布の範囲で集団のデータ分布の特徴を表す。分布の形は用いる確率モデルによって決まる。

2. 二項分布 $B(n, p)$ の形は n と p で決まる。

二項分布 $B(n, p)$ では、平均と分散は次の式で与えられる

$$\mu = np$$

$$\sigma^2 = np(1 - p)$$

3. 母集団の分布中心は平均値であり、母集団の平均値を標本集団の平均値から推測する。

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

$$\mu = \bar{x}$$

$\bar{x}$: 標本集団の平均値

n : 標本集団のデータ数 (サンプルサイズ)

μ : 母集団の平均値

4. データ分布の範囲は平均値周りの二次の積率で表す。

$M(x - \mu)^2$: 平均値周りの二次の積率

5. $M(x - \mu)^2$ 平均値周りの二次の積率、平均値からの差の二乗の期待値

$$M(x - \mu)^2 = E(x - \mu)^2$$

SS: 平方和

$$SS = \sum_{i=1}^n (x_i - \bar{x})^2$$

$$s^2 = \frac{SS}{n}$$

$\bar{x}$: 標本集団の平均値

n : 標本集団の標本数 (サンプルサイズ)

s^2 : 標本集団の分散

s : 標本集団の標準偏差

母集団の分散を標本集団から求めるときには以下の式を使う。

$$M(x - \mu)^2 = E(x - \mu)^2$$

$$= \frac{SS}{n - 1}$$

$$\sigma^2 = \frac{SS}{n - 1}$$

σ^2 : 歪みのない母集団の分散

σ : 母集団の分散

$$s^2 = \frac{SS}{n}$$

s^2 : 標本集団の分散

s : 標本集団の標準偏差

6. 母集団の平均値周りの二次の積率 $M(\bar{\mu} - \mu)^2$ は推測された母集団の平均値の信頼範囲

を表している。これは、標準誤差の二乗である。

$$M(\bar{\mu} - \mu)^2 = E(x - \mu)^2 = \frac{\sigma^2}{n}$$

$$\frac{\sigma^2}{n} : \text{標準誤差の二乗}$$

III-2-2-3. 二項分布の応用

二項分布の特性について詳しく見てきましたが、この知識を何にどのように使えば良いのかが分からないとせっかくの知識が役立ちません。二項分布の考え方を発展させたものが正規分布です。また、二項分布のような比率データの違いの検定にはカイ二乗検定というのがあります。これもよく使う便利なデータの検定法です。ここでは、そのような発展形ではなくて、2 項分布のままでどのように使うかを考えます。

二つの例を挙げます。

例 1. ある小学校のクラスから 5 人の生徒を標本として選びます。そのクラスの生徒の男女の性比は 1 対 1 だとします。つまり、一人の生徒が男子である確率は $p = \frac{1}{2}$ です。そのような条件で、5 人の内で何人の生徒が男子であるかという確率を計算します。

$$p(5) = {}_5C_5 \times \left(\frac{1}{2}\right)^5 = 1 \times \frac{1}{32} = \frac{1}{32} = 0.03125$$

$$p(4) = {}_5C_4 \times \left(\frac{1}{2}\right)^5 = \frac{5 \cdot 4 \cdot 3 \cdot 2}{4 \cdot 3 \cdot 2 \cdot 1} \times \frac{1}{32} = 5 \times \frac{1}{32} = \frac{5}{32} = 0.15625$$

$$p(3) = {}_5C_3 \times \left(\frac{1}{2}\right)^5 = \frac{5 \cdot 4 \cdot 3}{3 \cdot 2 \cdot 1} \times \frac{1}{32} = 10 \times \frac{1}{32} = \frac{10}{32} = 0.3125$$

$$p(2) = {}_5C_2 \times \left(\frac{1}{2}\right)^5 = \frac{5 \cdot 4}{2 \cdot 1} \times \frac{1}{32} = 10 \times \frac{1}{32} = \frac{10}{32} = 0.3125$$

$$p(1) = {}_5C_1 \times \left(\frac{1}{2}\right)^5 = \frac{5}{1} \times \frac{1}{32} = 5 \times \frac{1}{32} = \frac{5}{32} = 0.15625$$

$$p(0) = {}_5C_0 \times \left(\frac{1}{2}\right)^5 = 1 \times \frac{1}{32} = \frac{1}{32} = 0.03125$$

選ばれた 5 人の生徒が全員男女どちらか性別だという確率は 0.03125 で、もし性比が本当に 1 対 1 ならば、そういうことが起こるのは 3.125%, しかないことがわかります。そういうことが起きた場合、私たちは、おそらくそのクラスの男女の性比は 1 対 1 ではないか、5 人の生徒が無作為に選ばれたのではないだろうと考えます。そういうことがたまたま偶然に起こる可能性は、3.125% しかないからです。一人だけ女子が選ばれていたとします。そういうことが起こる確率は 0.15625 でランダムに選ばれていたとしてもそういうことが起こることは 15% 以上あるし、全く選ばれない可能性も含むと、1 人以下しか選ばれない可能性は、19% ぐらいはあるので、性比が 1 対 1 ではない。あるいはランダムに選ばれたので

はないと結論するのをためらうでしょう。できるだけ誤った判断をしたくないと考える人は、確率の限界を 1% とか 0.1%のように、極めて少ない確率を判断の限界にするでしょう。つまり、間違えたくない、できるだけ誤りが少ないという安全側に判断する場合は、少ない確率限界を採用します。

第二の例

α と β の二つの箱があり、箱の中に赤いボールと白いボールが入っていますが、私たちは箱の中を見ることが出来ません。A の箱の中には、赤いボールと白いボールが 1:2 で入っています。B の箱には、赤いボールと白いボールが 1:1 の割合で入っています。私たちは 2 つの箱のどちらが α でどちらが β か知りません。この状態で箱からボールを一つずつ取りだして、ボールの色を確かめます。これを 3 回繰り返した結果は、次の通りでした。

箱	1	2	3
A	Red	White	White
B	Red	Red	Red

この状態で、私たちは A の箱が α で、B の箱が β だと判断できます、何故ならば、B の箱が α である可能性は $\frac{1}{27}$ に過ぎないからです。この判断に至る過程では次のような計算をしています。

- (1) A の箱が α であったとき、赤いボールが一回、白いボールが 2 回取り出される可能性は次の通りです。

$$p(1) = {}_3C_1 \left(\frac{1}{3}\right) \left(\frac{2}{3}\right)^2 = 3 \times \frac{2^2}{3^3} = \frac{4}{9}$$

- (2) A の箱が β であったとき、赤いボールが 1 回、白いボールが 2 回取り出される可能性は次の通りです。

$$p(1) = {}_3C_1 \left(\frac{1}{2}\right) \left(\frac{1}{2}\right)^2 = 3 \times \frac{1}{2^3} = \frac{3}{8}$$

- (3) B の箱が α であった時、3 回つづけて赤が出る可能性は次の通りです。

$$p(3) = {}_3C_3 \left(\frac{1}{3}\right)^3 = 1 \times \frac{1}{3^3} = \frac{1}{27}$$

- (4) B の箱が β であった時、3 回続けて赤が出る可能性は次の通りです。

$$p(3) = {}_3C_3 \left(\frac{1}{2}\right)^3 = \frac{1}{2^3} = \frac{1}{8}$$

危険率 4%で考えると

A= α の可能性は否定されません。 (1)

A= β の可能性も否定されません。 (2)

B= α の可能性は否定されます。 (3)

B= β の可能性は否定されません。 (4)

明瞭に可能性が否定されたのは $B=\alpha$ だから、 $B \neq \alpha$ で、 $B=\beta$ 、 $A=\alpha$ となります。ここで、大切なことは、可能性が否定されなかったということは、可能性が肯定されたということではないということです。つまり、(2)で帰無仮説 $A=\beta$ が否定されなかったということは、 $A = \beta$ が肯定されたということではないということです。

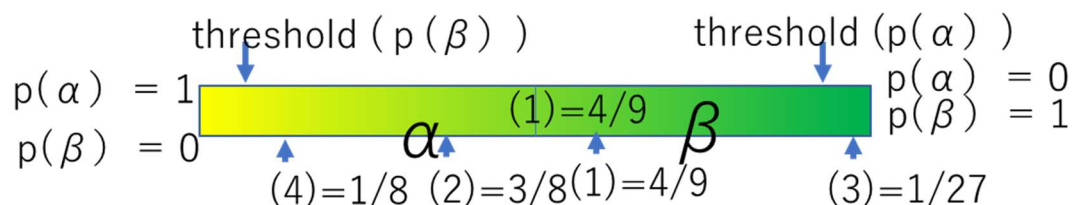


図 17. 確率分布と判定

図 17 は、確率の分布と、判定の結果の関係を示しています。しかし、多くの読者の中には、この判定の閾値が、誤判定の危険性やリスクを冒すことで得られるメリット、データが置かれている状況などを考慮して恣意的に決められるということから、次のような疑問を持つでしょう。この場合、判定の閾値を 0.04 にしたから $B=\alpha$ の可能性だけを否定できたが、閾値を 0.125 以上に設定すると、 $B=\beta$ も否定されてしまうので、どうすればよいのかということです。著者が何故閾値を 0.04 に設定したかですが、たとえば、3つのボールを取り出すという試行を4回くり返します。そのうち1回でも、 β の箱から赤いボールが3つ出てくるという可能性は、 $\frac{1}{8} \times 4 = \frac{1}{2}$ つまり半分以上あることになります。ところが、 α の箱から赤いボールが三つ出てくるのが4回の試行のうち一回でも起こることは、 $0.04 \times 4 = 0.16$ ですから、何回繰り返しても頑健な結果になると考えて、0.04 という閾値を設定したのです。