・母平均と母分散

下記の表のような確率をもつ離散型の確率分布の母平均と母分散を求めよ。

X	-4	1	3	11
確率	$\frac{1}{10}$	$\frac{9}{20}$	$\frac{1}{5}$	$\frac{1}{4}$

なお、母分散は $E[(X-\mu)^2] = E[X^2] - \mu^2$ を用いて計算しても良い。

・母平均 μ

$$\begin{aligned}\mu = E[X] &= (-4) \times \frac{2}{20} + 1 \times \frac{9}{20} + 3 \times \frac{4}{20} + 11 \times \frac{5}{20} \\ &= \frac{(-4) \times 2 + 1 \times 9 + 3 \times 4 + 11 \times 5}{20} \\ &= \frac{-8+9+12+55}{20} = \frac{68}{20} = \frac{17}{5} = 3.40\end{aligned}$$

・母分散 σ^2《定義通り》

$$\begin{aligned}\sigma^2 &= E[(X-\mu)^2] \\ &= \left(-4 - \frac{17}{5}\right)^2 \times \frac{2}{20} + \left(1 - \frac{17}{5}\right)^2 \times \frac{9}{20} + \left(3 - \frac{17}{5}\right)^2 \times \frac{4}{20} + \left(11 - \frac{17}{5}\right)^2 \times \frac{5}{20} \\ &= \left(-\frac{37}{5}\right)^2 \times \frac{2}{20} + \left(-\frac{12}{5}\right)^2 \times \frac{9}{20} + \left(-\frac{2}{5}\right)^2 \times \frac{4}{20} + \left(\frac{38}{5}\right)^2 \times \frac{5}{20} \\ &= \frac{1369 \times 2 + 144 \times 9 + 4 \times 4 + 1444 \times 5}{500} = \frac{11270}{500} = \frac{1127}{50} = 22.54\ (= 22.5)\end{aligned}$$

・母分散 σ^2《$E[X^2]$ を使う》

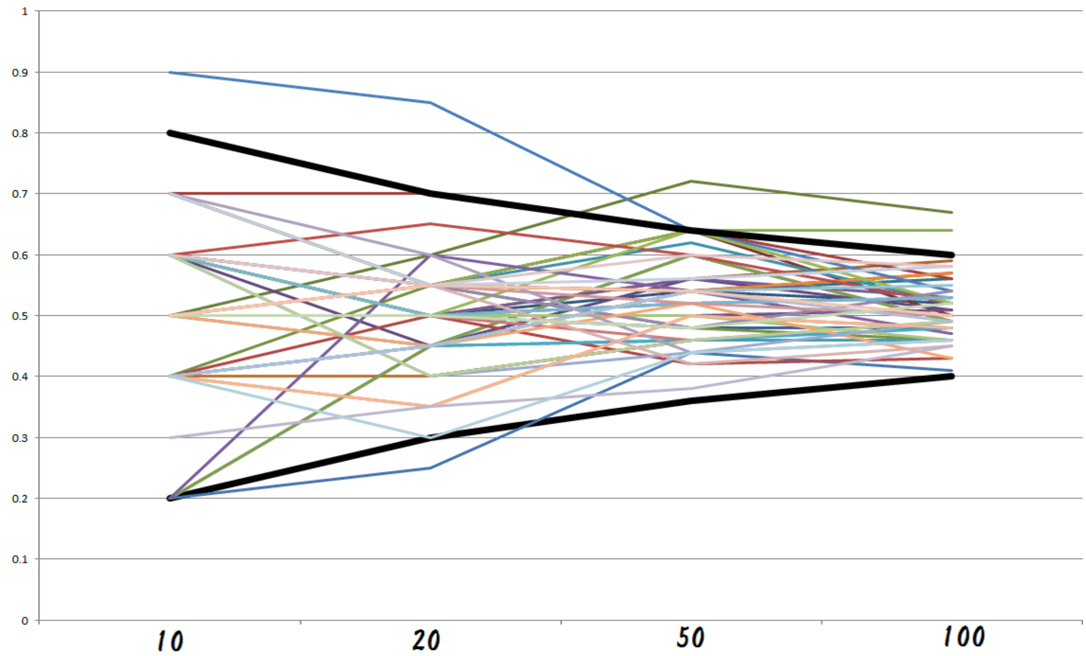
$$\begin{aligned}E[X^2] &= (-4)^2 \times \frac{2}{20} + 1^2 \times \frac{9}{20} + 3^2 \times \frac{4}{20} + 11^2 \times \frac{5}{20} \\ &= \frac{16 \times 2 + 1 \times 9 + 9 \times 4 + 121 \times 5}{20} = \frac{682}{20} = \frac{341}{10}\end{aligned}$$

$$\sigma^2 = E[X^2] - \mu^2 = \frac{341}{10} - \left(\frac{17}{5}\right)^2 = \frac{1705 - 578}{50} = \frac{1127}{50}$$

母分散はどちらの方法で計算しても $\frac{1127}{50} = 22.54\ (= 22.5)$ となる。

・実際にコイン投げを１００回行い、１０回・２０回・５０回・１００回までの結果から表の出る確率をそれぞれ推定せよ。解答用紙には１００回分の結果と１０回・２０回・５０回・１００回のときの推定値を書くこと。

期限内に提出された 49 人分の前回の演習問題の結果をグラフにまとめると



のようになる。太線は９５％の確率で出現する値（信頼区間）である。値にすると

	10	20	50	100
区間	$[0.200, 0.800]$	$[0.300, 0.700]$	$[0.360, 0.640]$	$[0.400, 0.600]$

のように回数が増えると範囲が狭くなっていることがわかる。ちなみに 1,000 回投げたときは $[0.469, 0.531]$ であり、10,000 回投げたときは $[0.4902, 0.5098]$ となり、だんだん理論値である 0.500 に近づいている。

今回の結果では 1 人区間外になっているが、９５％の確率で入る区間であるため、残り５％の確率で区間外になる。つまり、２０人もいれば 1 〜 2 人程度区間外が出てもまったくおかしなことではない。

資料置場

https://www.gen.kanagawa-it.ac.jp/takeda/class/

中心極限定理と二項分布の正規近似

・中心極限定理

$\{X_n\}$ は独立で同一分布に従う確率変数列とし、$\mu = E[X_1],\ \sigma^2 = V[X_1]$ は共に存在るものとする。Z は標準正規分布 $N(0, 1^2)$ に従う確率変数とし、$S_n = X_1 + \cdots + X_n$ とおくと、確率変数 $\frac{\frac{S_n}{n} - \mu}{\frac{\sigma}{\sqrt{n}}}$ は Z に法則収束する。

つまり、平均と分散を持つ母集団から得られる標本平均の分布はデータの数 n が十分大きければ正規分布で近似できることを示している。正規分布に近づく早さは母集団の分布によって異なるが、二項分布は特に早く正規分布に収束する。

例． 二項分布の場合

表が出る確率 p のコイン投げを n 回投げたとき、表の出る回数 X は二項分布に従う。二項分布の母平均は $\mu = np$，母分散は $\sigma^2 = np(1-p)$ であるから、次のような変換によって、標準正規分布で近似することができる。

$$Z = \frac{X - \mu}{\sigma} = \frac{X - np}{\sqrt{np(1-p)}}$$

例題． $p = \frac{1}{4}$ のコイン投げを 1,000 回行ったとき、$[220, 270]$ の間に入る確率を求めよ。

二項分布の母平均と母分散は $\mu = 1000 \times \frac{1}{4} = 250$,　$\sigma^2 = 1000 \times \frac{1}{4} \times \frac{3}{4} = 187.5$ なので、$[220, 270]$ の間に入る確率は

$$\begin{aligned} P(220 \leqq X \leqq 270) &= P\left(\frac{220 - 250}{\sqrt{187.5}} \leqq Z \leqq \frac{270 - 250}{\sqrt{187.5}}\right) \\ &= P\left(-2.19 \leqq Z \leqq 1.46\right) \\ &= 0.48574 + 0.42786 = 0.91360 = 0.914 \end{aligned}$$

この確率を二項分布を使ってまともに計算するとかなり近い値を得る。

$$\sum_{k=220}^{270} {}_{1000}C_k \left(\frac{1}{4}\right)^k \left(1 - \frac{1}{4}\right)^{1000-k} = 0.9197\cdots$$

しかしこの計算の組み合わせの数は莫大（${}_{1000}C_{250}$ で 200 桁近くになる）でまともに計算はできない。このような場合、標準正規分布を用いた近似計算は非常に役立つ。

・不偏推定量と最尤推定量

推定には必ず誤差が生じるため、推定の良し悪しを判断する基準が必要になる。一般的に用いられる基準としては下記の 2 つがよく用いられている。

・不偏推定量（期待値による基準）

推定量 $h(X)$ の期待される値として、期待値が次のように定義されている。

$$E\Big[h(X)\Big] = \begin{cases} \displaystyle\sum h(x_i) \times \Pr\{X = x_i\} & \text{(離散型の場合)} \\ \displaystyle\int_{-\infty}^{\infty} h(x) \times f(x)\,dx & \text{(離散型の場合)} \end{cases}$$

推定量の期待値が実際の値と一致している場合、不偏推定量（意訳：偏りの無い推定量）と呼ぶ。実際に標本平均 $\bar{x}$ の期待値を計算すると、

$$\begin{aligned} E[\bar{X}] &= E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E\left[X_i\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} \mu = \mu \end{aligned}$$

となるので、標本平均は母平均の不偏推定量である。さらに標本分散も母分散の不偏推定量となっている。

$$E[S^2] = E\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \mu)^2\right] = \sigma^2$$

・最尤推定量（尤度関数による基準）

母集団のパラメータが θ で確率密度関数が $f(x|\theta)$ とする。このとき、n 個の標本 $x_1, x_2, \cdots, x_n$ が得られた場合の尤もらしさを表す値として、尤度関数が次のように定義されている。

$$L(\theta|x_1, \cdots, x_n) = f(x_1|\theta) \times f(x_2|\theta) \times \cdots \times f(x_n|\theta)$$

離散型の場合は確率密度関数 $f(x|\theta)$ のかわりに確率 $p(x|\theta)$ で計算するが、いずれにしても標本 x_i が現れた出現確率の積とみなすことができる（厳密にいえば連続型の場合、確率密度の積）。よって、この尤度関数の値が大きいほど分布の当てはまりが良いことを示している。そこで n 個の標本が得られた場合の尤度関数の値を最大にするような、パラメータ θ の推定量を最尤推定量と呼ぶ。正規分布を仮定した場合、標本平均 $\bar{x}$ と n で割った標本分散が最尤推定量となっている。

ただし、最尤推定量と不偏推定量が常に同時に成立しているわけではないことに注意が必要である。実際正規分布を仮定した場合、母平均 μ の推定量 $\bar{x}$ は不偏推定量かつ最尤推定量であるが、母分散 σ^2 の不偏推定量は $n-1$ で割った標本分散 s^2 であり、最尤推定量は n で割った標本分散である。実際、分散の最尤推定量の期待値は実際の値より少し小さい値となる。

また、不偏推定量になる推定量はいくつでも作ることができるが、推定量の分散が小さいほど推定の精度が良いことになるので、一般的には分散が最小になる推定量を選ぶ。

2019年度 確率統計　小テスト解答用紙　2019.10.29

・大標本分布（二項分布の正規近似）

表の出る確率 $p = \frac{1}{6}$，試行回数 $n = 1620$ の二項分布に従う確率変数が $[255, 294]$ に入る確率 $P(255 \leqq X \leqq 294)$ を正規分布で近似して求めよ。ただし、標準正規分布の確率を計算する場合、変換後の両端の値は小数第３位を四捨五入した値を用いること。

（裏面にもう１問あります）

2019年度神奈川工科大学 確率統計 演習問題	学科	学年	組	学　籍　番　号	氏　　　名	

提出先：K3-3309号室前　17番のボックス　提出期限：１０月３０日（水）１７時頃まで

・最尤推定量

指数分布 $f(x|\lambda) = \lambda e^{-\lambda x}$ の尤度関数および対数尤度関数を求め、パラメータ λ の最尤推定量を求めよ。

ヒント：尤度関数より、対数尤度関数を λ で微分した方が簡単です。