統計学 2020 Lecture 3: 2 項分布と統計的推測

北門　利英（東京海洋大学海洋生物資源学科）

2020 年 5 月 27 日

Attention:

- 授業 HP に授業に関する情報をアップデートしていきますので参照ください．(URL: https://toshihidekitakado.github.io/STAT2020/index.html)
- 今回の授業に関してわからないことがあれば，メールかリアルタイム接続時に遠慮なく質問してください（5 月 27 日は 12:45-13:45 とします）.

Point:

- 2 項分布とは何か説明できるようになること
- 2 項分布の期待値と分散の式を自身で導けるようにすること
- 2 項分布の確率分布を利用して様々な計算ができること
- 推定量と推定値の違いを理解する事
- 2 項分布の推定量の期待値，分散，標準偏差の計算ができること

1　2 項分布とは

1.1　2 項分布の定義

いま，成功か失敗の試行を N 回行うとする．各回の試行は独立で，成功の確率は試行を通して同一とする．ここでの成功/失敗は，例えば母集団から個体を選んだときの雄/雌，成熟/未成熟，ウイルス感染/非感染と読み替えても同様である．さて，第 i 回目 $(i = 1, 2, \dots, N)$ の試行の結果を

$$X_i = \begin{cases} 0 & (\text{第 } i \text{ 回目の試行が失敗}) \\ 1 & (\text{第 } i \text{ 回目の試行が成功}) \end{cases}$$

とする．確率変数 $X_1, \dots, X_N n$ は独立同一に確率分布

$$\begin{aligned} P(X_i = x_i) &= \begin{cases} 1 - p & \text{if } x_i = 0 \\ p & \text{if } x_i = 1 \end{cases} \\ &= p^{x_i}(1-p)^{1-x_i} \quad (x_i = 0, 1) \end{aligned}$$

にしたがう．このような 0 か 1 の値をとる試行をベルヌーイ試行 (Bernoulli trial) という．

この試行を N 回繰り返したときに成功する回数もまた確率的に変動する．これを $Y=\sum_{i=1}^{N}X_i$ で表すことができる．確率変数 Y の確率分布を考えよう．最初は簡略化のために，$Nn=4$ の場合を考えると，$Y=\sum_{i=1}^{4}X_i$ と元の確率ベクトル $X=(X_1,X_2,X_3,X_4)$ との間に以下の対応関係があることがわかる．

$$
\begin{aligned}
&Y=0 \Longleftrightarrow X=(0,0,0,0)\\
&Y=1 \Longleftrightarrow X=(1,0,0,0) \text{ or } (0,1,0,0) \text{ or } (0,0,1,0) \text{ or } (0,0,0,1)\\
&Y=2 \Longleftrightarrow X=(1,1,0,0) \text{ or } (1,0,1,0) \text{ or } (1,0,0,1) \text{ or } (0,1,1,0) \text{ or } (0,1,0,1) \text{ or } (0,0,1,1)\\
&Y=3 \Longleftrightarrow X=(1,1,1,0) \text{ or } (1,1,0,1) \text{ or } (1,0,1,1) \text{ or } (0,1,1,1)\\
&Y=4 \Longleftrightarrow X=(1,1,1,1)
\end{aligned}
$$

例えば $Y=1$ に対する確率は

$$
\begin{aligned}
P(Y=1) &= P(X=(1,0,0,0)\text{or}(0,1,0,0)\text{or}(0,0,1,0)\text{or}(0,0,0,1))\\
&= P(X=(1,0,0,0))+P(X=(0,1,0,0))\\
&\quad +P(X=(0,0,1,0))+P(X=(0,0,0,1))\\
&= 4p(1-p)^3
\end{aligned}
$$

となる．このように，確率 $p(1-p)^3$ に $Y=1$ に対応する $X=(X_1,X_2,X_3,X_4)$ の組み合わせ数が掛け合わされることで成功数 $Y=1$ の確率が求まる．他の Y の値についても同様である．これらを一般化すると以下のようになる．

離散型分布 1 [2 項分布] 確率変数 Y が確率関数

$$
P(Y=y)=\binom{N}{y}p^y(1-p)^{N-y}, \quad y=0,1,\dots,N \quad (0\le p\le 1) \tag{1}
$$

をもつとき，Y は 2 項分布 (binomial distribution) にしたがうという．この 2 項分布は $Bin(N,p)$ で表すこととする．

```r
par(mfrow = c(2, 2), mar = c(2, 4, 4, 3))
ylim=c(0, 0.22); PP <- "Probability"
y <- seq(0, 20, 1)
plot(y, dbinom(y, 20, 0.2), "h", ylim=ylim, ylab=PP, main="(a) N=20, p=0.2")
plot(y, dbinom(y, 20, 0.5), "h", ylim=ylim, ylab=PP, main="(b) N=20, p=0.5")
y <- seq(0, 50, 1);
plot(y, dbinom(y, 50, 0.2), "h", ylim=ylim, ylab=PP, main="(c) N=50, p=0.2")
plot(y, dbinom(y, 50, 0.5), "h", ylim=ylim, ylab=PP, main="(d) N=50, p=0.5")
```

1.2 2 項分布の性質

Y が 2 項分布 $Bin(N,p)$ にしたがうとき，Y の期待値と分散はそれぞれ $E[Y]=Np, V[Y]=Np(1-p)$ となる．

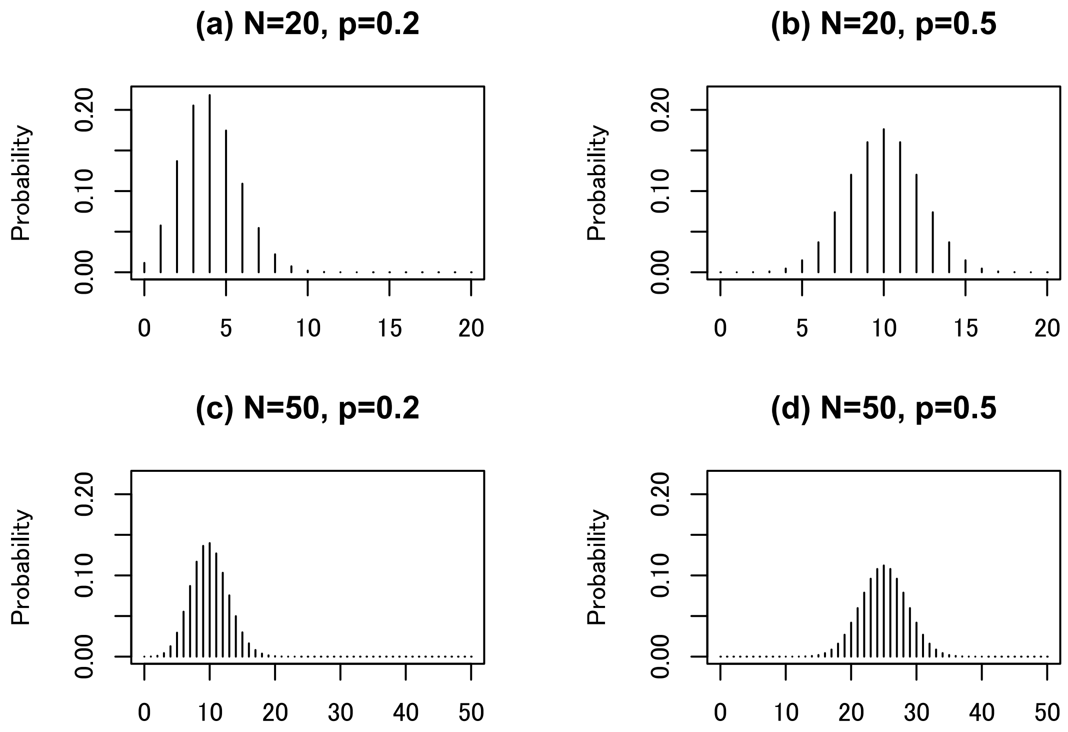

図1　2 項分布の確率分布の例

1.2.1　期待値

期待値は以下のように求められる．

$$\begin{aligned} E[Y] &= \sum_{y=0}^{N} y\binom{N}{y}p^y(1-p)^{N-y} \\ &= Np\textstyle\sum_{y=1}^{N}\frac{(N-1)!}{(y-1)!(N-y)!}p^{y-1}(1-p)^{N-y} \\ &= Np\displaystyle\sum_{y=0}^{N-1}\frac{(N-1)!}{y!(N-1-y)!}p^y(1-p)^{N-1-y} \\ &= Np. \end{aligned}$$

1.2.2　分散

分散の公式は，

$$\begin{aligned} V[Y] &= E[(Y-E[Y])^2] \\ &= E[Y^2-2YE[Y]+E[Y]^2] \\ &= E[Y^2]-2E[Y]\cdot E[Y]+E[Y]^2 \\ &= E[Y(Y-1)+Y]-E[Y]^2 \\ &= E[Y(Y-1)]+E[Y]-E[Y]^2 \end{aligned}$$

とかける．そして $E[Y]$ の場合とほぼ同様な計算から

$$\begin{aligned} E[Y(Y-1)] &= \sum_{y=0}^{N} y(y-1)\binom{N}{y} p^y(1-p)^{N-y} \\ &= N(N-1)p^2 \textstyle\sum_{y=2}^{N} \frac{(N-2)!}{(y-2)!(N-y)!} p^{y-2}(1-p)^{N-y} \\ &= N(N-1)p^2 \sum_{y=0}^{N-2} \frac{(N-2)!}{y!(N-1-y)!} p^y(1-p)^{N-2-y} \\ &= N(N-1)p^2. \end{aligned}$$

であることが分かり，したがって分散は

$$V[Y] = E[Y(Y-1)] + E[Y] - E[Y]^2 = N(N-1)p^2 + Np - (Np)^2 = Np(1-p)$$

となる．

1.2.3　2 項分布の再生性

定理 1　[2 項分布の再生性]　確率変数 Y_i $(i = 1, 2, \ldots, m)$ が独立同一に $Bin(n_i, p)$ にしたがうとき，$T = \sum_{i=1}^{m} Y_i$ は 2 項分布 $Bin(N, p)$ にしたがう．ただし，$N = \sum_{i=1}^{m} n_i$ である．

1.3　2 項分布の例

例 1　ある充分に大きな母集団における雌の割合を p とする．この母集団から大きさ N のサンプルを抜き出したときの雌の数 Y は 2 項分布 $Bin(N, p)$ にしたがう．

例 2　ある充分に大きな母集団において，ある特定のハプロタイプをもった個体の割合を p とする．この母集団から大きさ N のサンプルを抜き出したとき，このハプロタイプをもつ個体数 Y は 2 項分布 $Bin(N, p)$ にしたがう．

このほか，2 項分布に関連した幾つかの例を見ていこう（今回はこういう例もある，という程度で構いません）．

例 3　[枠取り調査法 (quadrat method)] 面積 A の漁場に n 個体の貝がランダムに分布しているとする．そこに，面積 a の区画を設ける．その区画内の貝の総数を X とすると，この X の確率分布は 2 項分布 $Bin(n, p)$ となる．ただし，$p = a/A$ である．

次に，区画内の個体数 X はダイバーによってある一定の割合で見逃されるとしよう．この見逃し率を $1 - \theta$

とするとき（したがって発見率は θ），区画内の発見個体数 Y の周辺分布は，以下の全確率の公式で求まる．

$$\begin{aligned}
P(Y=y) &= \sum_{x=0}^{n} P(X=x)\cdot P(Y=y|X=x) \\
&= \sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x} \cdot \binom{x}{y} \theta^y (1-\theta)^{x-y} \\
&= \frac{n!}{y!(n-y)!}(\theta p)^y \sum_{x=y}^{n} \frac{(n-y)!}{(n-x)!(x-y)!}\{(1-\theta)p\}^{x-y}(1-p)^{n-x} \\
&= \frac{n!}{y!(n-y)!}(\theta p)^y \sum_{z=0}^{n-y} \frac{(n-y)!}{(n-y-z)!z!}\{(1-\theta)p\}^{z}(1-p)^{n-y-z} \\
&= \frac{n!}{y!(n-y)!}(\theta p)^y (1-\theta p)^{n-y}.
\end{aligned}$$

したがって，Y の周辺分布は 2 項分布 $Bin(n,\theta p)$ となる．

例 4　[死亡過程のモデル] ある個体群では生まれてから 1 年間の各個体の死亡率が独立同一に ϕ_1 $(0<\phi_1<1)$ であることが知られているとする．ある年に n 個体生まれたとき，出生 1 年後の死亡数 Y_1 は 2 項分布 $Bin(n,\phi_1)$ にしたがう．また，2 年目の 1 年間の死亡率も独立同一が ϕ_2 $(0<\phi_2<1)$ であるとき，1 年目の死亡数 $Y_1=y_1$ を与えた時の 2 年目の死亡数 Y_2 の条件付き確率分布は $Bin(n-y_1,\phi_2)$ となる．

ところで，このとき Y_1 と Y_2 の同時確率分布はどのような分布となるであろうか．

$$\begin{aligned}
P(Y_1=y_1,Y_2=y_2) &= P(Y_1=y_1)P(Y_2=y_2|Y_1=y_1) \\
&= \binom{n}{y_1}\phi_1^{y_1}(1-\phi_1)^{n-y_1}\binom{n-y_1}{y_2}\phi_2^{y_2}(1-\phi_2)^{n-y_1-y_2} \\
&= \frac{n!}{y_1!y_2!(n-y_1-y_2)!}\phi_1^{y_1}\{(1-\phi_1)\phi_2\}^{y_2}\{(1-\phi_1)(1-\phi_2)\}^{n-y_1-y_2}
\end{aligned}$$

と書ける．3 年目以降も同様にモデリングすることが可能である．なおこの同時確率分布は，後に述べる多項分布の一例であることがわかる．

その他，参考までに...

例 5　[成熟率のモデル & 網目選択性試験] 寿命が A 歳である魚種について，年齢 a $(a=0,1,\ldots,A)$ の魚の成熟率を p_a とする．母集団から年齢 $0,1,\ldots,A$ の魚を $n_0,n_1,\ldots,n_A$ 個体サンプリングしたとき，その中に含まれる成熟個体の数 $Y_0,Y_1,\ldots,Y_n$ は独立で，それぞれ 2 項分布 $Bin(n_a,p_a)$ にしたがう．このモデルは合計 $A+1$ 個のパラメータで表現されている．もし，成熟率 p_a が年齢 a のロジスティック関数

$$p_a = p(a) = \frac{e^{\alpha+\beta a}}{1+e^{\alpha+\beta a}}$$

で表せるならば，このモデルは 2 個のパラメータ α,β のみで表現することが可能となる．

同様のモデルは，カバーネット網目選択性試験でもよく利用される．この試験では，網目選択性を調べたいトロール網のコッドエンドにカバーネットを装着する．ただし，カバーネットの目合の大きさは，魚が逃げないくらい十分小さくする．体長階級 l の魚に対する網目選択率を

$$s(l) = \frac{e^{\alpha+\beta l}}{1+e^{\alpha+\beta l}}$$

とする．すなわち，$s(l)$ は網目に遭遇した魚がコッドエンドに保持される確率である．コッドエンドとカバーネットで漁獲された体長階級 l ($l = 1, 2, \ldots, L$) の魚の合計を n_l 個とし，カバーネットで漁獲された個体数を Y_l とすると，Y_l は 2 項分布 $Bin(n_l, s(l))$ に従う．

2　2 項分布における統計的推測

前節では，2 項分布の定義と性質について述べた．2 項分布にしたがう観測値はパラメータ N, p の値に依存することが分かった．すなわち，例えば $N = 10$ のとき，p の値が大きければ確率変数の実現値は大きな値が出やすく，逆に p の値が小さければ小さな値が出やすい．逆に，p の値が未知の場合には，観測値を基に p に関する情報を引き出すことができる．

一般的に，未知の値や構造等に対して，観測値から情報を引き出し何らかの決定をすることを統計的推測という．統計的推測には，点推定，区間推定，仮説検定などいくつかのタイプがあり，これらは後の講義で詳しく解説するが，ここでは 2 項分布を題材にした簡単な例で大まかな概略をつかんでほしい．

2.1　問題の設定

例題

あなたはリサーチ会社に勤務しており，上司から「東京都日曜午後 6 時４０分におけるサザエさん視聴率調査の企画案を作成すること，ただし視聴率の推定値の標準偏差を 0.01 以下にすること」と依頼された．そこであなたは，全世帯から何世帯かをランダムにサンプリングし，アンケート調査を行うこととしたが，上司の依頼に応えるためには何世帯のサンプリングを提案すればよいであろうか？

少々設定に無理があるが，細かい点は別として，「サザエさんを見ている」あるいは「サザエさんを見ていない」の 2 つの属性があり，サザエさんの視聴率が知りたい対象なので，これはまさしく 2 項分布の典型的な例となる．

この問題を数学的に表現してみると，サンプリングする世帯数を N, 東京全体におけるサザエさん視聴率を p, そして N 世帯中においてサザエさんを見ている世帯数を表す確率変数を Y とおくと，Y は以下の確率分布にしたがう．ここでは，区によって視聴率が異なるなどの細かい点は考えず，東京都全体の様子を捉えることとする．

$$Y \sim Bin(N, p)$$

いま p が知りたいパラメータ値であるから，その推定の方法として次の方式を考える．

$$\hat{p}(Y) = \frac{Y}{N} \tag{2}$$

この式の意味は，N 世帯のサンプリングをするので N は固定されているのに対し，Y は確率変数でありどのような値が観測されるか分からないが，その値に関わらず p の推定のための計算方法を予め規定している．このような式のことを**推定量 (estimator)** という．推定量を敢えて $\hat{p}(Y)$ のように Y の関数として表記しているのには理由がある．$\hat{p}(Y)$ は確率変数 Y の関数なので，$\hat{p}(Y)$ 自身も確率変数であり，したがってその期待値

や分散などの特性値（すなわち推定量の性質）を調べることができるのである．詳しくは次のセクションで述べる．

これに対し，例えば 100 世帯サンプリングするとした場合には $N = 100$ であり，実際に $Y = 15$ などのように具体的な値が観察された後に，

$$\hat{p} = \hat{p}(15) = \frac{15}{100} = 0.15$$

のように具体的な数値の入った値のことを**推定値 (estimate)** という．これはすでに値が決まっているので，確率変数ではない点に注意する．

2.2 推定量の性質

ところで，式 (2) の推定量 $\hat{p}(Y)$ を何のと断りもなく定義したが，これは直感的にも受け入れやすい推定量の形であろう．しかし，いつもこのように直感的な推定方式が決まるわけではなく，実際には何らかの原理原則も存在する．Lecture5 ではそれについてもう少し概念的に説明する予定であるが，ここではその助走として，推定量に対する 3 つの値に注目する．

2.2.1 推定量の期待値

推定量の期待値 (expectation) とは，Lecture5「統計基礎理論 (3) 推定と誤差」で説明する推定量の**不偏性 (unbiasedness)** という概念と密接に関係する．ここで重要なのは，真の p の値は唯 1 つであるが，$\hat{p}(Y) = \frac{Y}{N}$ の値は確率的に変動する事である．例えば $N = 100$ とし，真の p が 0.2 であるときに，$Y = 20$ を観測すれば誤差なしで推定することができるが，$Y = 18$ や $Y = 25$ という値も確率的には取り得るため，推定値はそれに応じてばらつく．したがって，偶然に真値を言い当てられる場合もあるが，むしろずれることの方が多い．ただ，このずれ方に癖があるといけない．この場合の癖とは偏りのことであるが，まずは偏りがないこと，すなわち不偏性であるかどうかを $\hat{p}(Y) = \frac{Y}{N}$ について確認してみよう．

推定量 $\hat{p}(Y) = \frac{Y}{N}$ の期待値は

$$E[\hat{p}(Y)] = E\left[\frac{Y}{N}\right]$$

と形式的に書けるが，ここで Lecture2 のセクション 2.3 で述べた期待値に関する次の性質を思い出してほしい．

$$E[a + bX] = a + bE[X]$$

この式で $a = 0$ とおくと「確率変数の定数倍 aX の期待値は，もとの確率変数の期待値の定数倍」になることが分かるから，それと 2 項分布の期待値の性質 $E[Y] = Np$ を利用すると，以下の式を得る．

$$E[\hat{p}(Y)] = E\left[\frac{Y}{N}\right] = \frac{1}{N}\ E[Y] = \frac{1}{N}\ Np = p$$

また，この式はどんな $p \in (0, 1)$ に対しても成り立つから，p の真値が何であろうとも $\hat{p}(Y) = \frac{Y}{N}$ の期待値は p であり，偏りなく（癖なく）平均的には真値を言い当てられる推定方式であると言える．（ただし，あくまで期待値による平均的な値であるので，運が悪いと $Y = 10$ などの値を観測し推定値が大きく真値からずれてしまうこともある．では少々運が悪くともある程度推定の性能をよくするにはどうすればよいだろうか？直感的には N を多くすればよいことが分かるだろう．それを次に分散と標準偏差を通して述べる．

2.2.2 推定量の分散

Lecture 2 のセクション 2.4 で分散 (variance) という概念について説明した．分散とは「確率変数の期待値の周りでの散らばりの尺度」あるいはもう少しフォーマルにいうと「確率変数と期待値の差の 2 乗の期待値」であった．前述の通り，推定量 $\hat{p}(Y) = \frac{Y}{N}$ も確率変数であり，その期待値は p であった．そこで，この推定量の期待値 p の周りでの散らばり具合，すなわち分散を評価してみよう．

推定量 $\hat{p}(Y) = \frac{Y}{N}$ の分散は

$$V[\hat{p}(Y)] = V\left[\frac{Y}{N}\right]$$

と書け，一般に分散の性質

$$V[a + bX] = b^2 E[X]$$

が成り立つから，以下のように求められる．

$$V[\hat{p}(Y)] = V\left[\frac{Y}{N}\right] = \frac{1}{N^2}\ E[Y] = \frac{1}{N^2}\ Np(1-p) = \frac{p(1-p)}{N}$$

この式からわかるように，サンプル数 N を大きくすれば分散がどんどん小さくなっていく．すなわち，どのような N の値でも期待値は p であるが，その周りでの散らばり方が N の値によって異なる．

2.2.3 推定量の標準偏差

さて上司の依頼は推定量 $\hat{p}(Y) = \frac{Y}{N}$ の分散に関してではなく標準偏差 (standard deviation) であった．確率変数の標準偏差とは，分散の平方根を意味し，次のように求められる

$$SD[\hat{p}(Y)] = \sqrt{V[\hat{p}(Y)]} = \sqrt{\frac{p(1-p)}{N}}$$

このことから上司の依頼を数学的に表現すれば

$$\sqrt{\frac{p(1-p)}{N}} \leq 0.01$$

のようになる．ただ，この式には 2 つの変数が入っているため，このままでは未だ答えを導き出せない．我々が求めたいのは N の値であり，何らかの値を p に関して想定する必要がある．

i) ワーストケースを考える：N を固定した時に，左辺の値を一番大きくするのは $p = 0.5$ の時である（2 次関数の性質）．$p = 0.5$ を仮定するとき，以下のように $N = 2500$ 世帯以上のサンプル数が必要となる．

$$\sqrt{\frac{0.5(1-0.5)}{N}} \leq 0.01 \Longleftrightarrow N \geq 2500$$

ii) 視聴率 50% のようなお化け番組はもはや存在しない．サザエさんの視聴率を推定しているわけだから正確な値は調査前には分からないが，せいぜい 20% を想定したらどれくらいの世帯をサンプリングすればよいか．10% ではどうか．計算してみて下さい．

2.3 推定量の性質まとめ

ここでは 2 項分布のパラメータ p に関する推定量とその性質について学んだ．大切な性質は，

Point:

- 真の p の値は唯 1 つであるが，推定量値は確率的に変動する
- 推定量 $\hat{p}(Y) = Y/N$ は $Y \sim Bin(N, p)$ の仮定の下では直感的に妥当な推定方式である
- 実際，$\hat{p}(Y) = Y/N$ の期待値は p の真値に関わらず p であり，少なくとも偏りのない推定方式である
- また，$\hat{p}(Y) = Y/N$ の分散はサンプル数 N に反比例し，N を増やすことによって真値 p の周りでの散らばりを小さくさせることができる
- 分散や標準偏差の評価式を通して，必要とされる推定精度に応じてサンプル数 N の値を事前に決めることも可能である

2.4 R による推定量の性能評価

N の値と p の値を変えて，パラメータ p の推定性能を R によるシミュレーションで簡単に見てみよう（分かりやすいように古典的な書き方のコードにしています）．後の講義でシミュレーションについては再度説明します．

```
Nit <- 10^4
Nvec <- c(10,100,1000)
pvec <- c(0.1, 0.3, 0.5)
pest <- array(NA,c(Nit,9))
col.list <- c("blue","orange","green")

par(mfrow=c(3,3), mar=c(3,3,4,2))
for(i in 1:3){
for(j in 1:3){
   tmp <- rbinom(Nit,Nvec[i],pvec[j])/Nvec[i]
   hist(tmp, xlab="Estimate", xlim=c(0,1), col=col.list[j],
        main=paste("N=",Nvec[i]," p=",pvec[j]))
}}
```

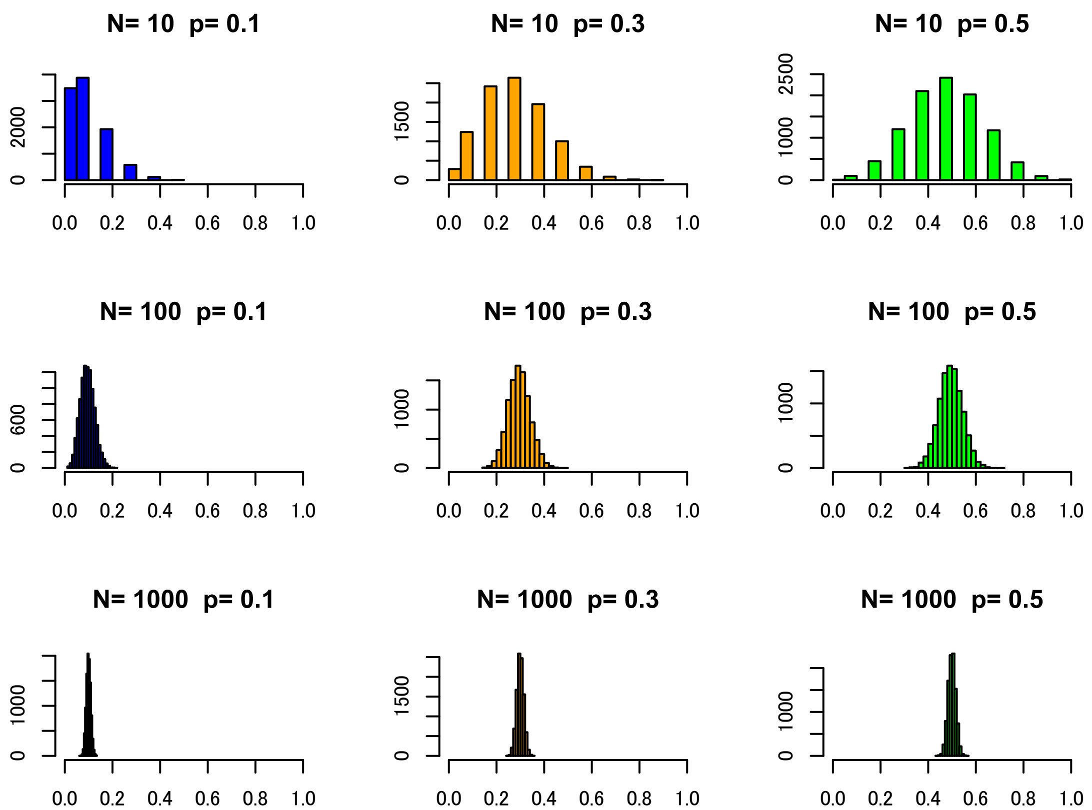

図2　2 項分布のパラメータの推定量の確率分布シミュレーション

3 提出課題

Lecture 3 と次回の Lecture4 の課題と併せて一つの課題とします．必ずまとめて 1 つのファイルで提出してください．提出期限は 6 月 9 日 23:59 です．

Lecture03-HW1

雌雄の比率が 1:1 である生物群集から N 個体のサンプルを取得するとし，雄の個体数を表す確率変数を Y とするとき，確率 $P(Y=y)$ $(y=0,1,...,N)$ の値を求めなさい．ただし，N は学籍番号の末尾の数字とします．(0 の人は N=10 としてください)

Lecture03-HW2

マウスに薬を投与して 80% 以上の割合で効果があった際に次の実験ステージに進めるとする．50% の確率でしか効かない薬をマウス 10 個体に投与した際に，そのような薬が次のステップに進んでしまう確率はどの程度か求めなさい．

Lecture03-HW3

ある養殖場で生産される魚は 1% の確率で奇形であるとする．この養殖場の魚を 100 個体選んだとき，奇形である個体が 3 個体以下である確率を求めよ．

Lecture03-HW4:

ある生物集団におけるメスの成熟率を推定したい．推定量の標準偏差を 0.05 以下とするためには何個体をサンプリングする必要があるか？