小さな数理統計学　佐藤健一（広島大学・原医研）

１．これから、統計学の基礎となる考え方を小さなデータを通して説明します。まず、末尾の乱数表から皆さんの誕生日を見つけて、その日から連続した２０個の０と１のデータを以下の空欄に転記して下さい。このデータは、例えば、ある病院において生まれた２０人の赤ちゃんの性別（女なら１、男なら０）のデータと考えることもできます。２つの事象を２値データとして記録する場合、関心がある事象に１を、ない方に０を対応させて記録することが慣例です。２０個の値を例えば１つの値で代表させる、または、要約できれば便利ですね。

２．それでは、書き込んだ２値データを要約することから始めてみましょう。要約とは、例えば電話で、報告書で、論文で、相手に自分の持っているデータがどういうものかをできるだけ手短に説明することだと思ってください。もちろん、データすべてを読み上げるというのもひとつの手です。電話で全部読み上げるのは可能だとしても、メモを取る方は大変ですね。

３．実は、末尾の表にあるデータはすべて１が出る確率を（秘密にしときますが）ある値に決めて作った乱数です。例えば、１が出る確率を 0.75 とする乱数を作りたければ、ぐるぐる回せるダーツの的をピザ切りの要領で４等分に線を引き、そのうちの３つに色を付けて、その後ぐるぐる回しながらダーツの矢を投げ、色がついた所に刺されば 1、そうでなければ 0 としてデータを作っていきます。末尾の表もそのようにして作ったと思ってもらって構いません。

４．すると、目の前にある２０個のデータは、たまたま得られた一組のデータと考えることができます。この神様がこっそり設定する１が出る確率を、母比率または真の比率とよびます。このようにデータを生成する機構に想定された確率的な仕組みを統計モデルとよびます。さて、目の前の２０個のデータから、データを生成した機構の母比率を当てることはできるでしょうか？たとえ話に戻ると、目の前の２０人の赤ちゃんの女の子の比率から、神様が決めてるであろう女の子が生まれる母比率を当てることができるか、という問題になります。

３．この当てるという行為は、推定とよばれます。真の比率と区別して、目の前のデータから得られた比率を標本比率とよぶことにすると、当てるというのは、標本比率を使って母比率を推定する、と言い換えることができます。あれ？はじめは２値データの要約ということで比率を計算したのですが、もう、目的が変わりましたよ？これが、記述統計と、推測統計の違いです。大げさに言えば、情報処理と統計の違いとも言えます。統計では、データを生成した統計モデルを想定し、目の前のデータからその機構を推測していきます。推測とは、推定と検定を合わせた言葉です。検定の説明は後でします。

４．さて、母比率を推定すると言ったものの、たまたま観測された目の前のデータは、わずか２０個です。もし、もう一度、同じように観測する機会があれば、きっと目の前のデータとは異なる一組のデータとなるでしょう。その度に計算される標本比率はどうでしょうか？変わりますか？変わるでしょうね。でも、それって、どのくらい変わるんでしょう？今日は幸運なことに、あなたと同じデータの生成機構から観測された２０個の一組のデータを、他の方々もお持ちです。みなさんの標本比率がどのようにばらついているか、格調高く言えば、どのように分布しているのか、集計してみましょう。それぞれ、何人いるでしょうか？

比率	0/20	1/20	2/20	3/20	4/20	5/20	6/20	7/20	8/20	9/20
頻度										
10/20	11/20	12/20	13/20	14/20	15/20	16/20	17/20	18/20	19/20	20/20

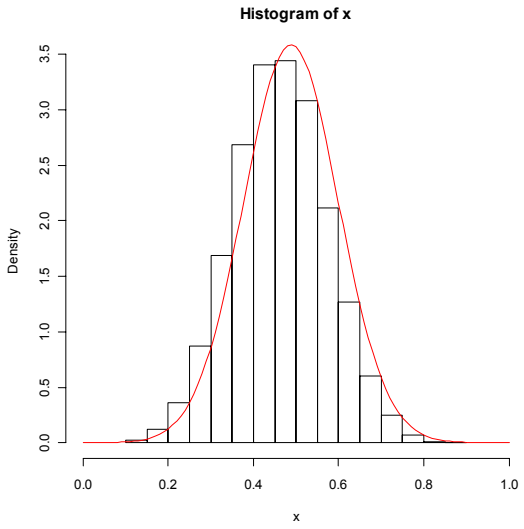
5．標本比率の分布は、比率の標本分布ともよばれます。今度は、この標本分布を要約してみましょうか。はじめの２値データと違って、色々な値をとるので（と言っても、２１通りですが)、この標本分布を説明するのは難しそうですね。はじめに思いつくのは、やはり平均でしょう。クラスの平均点、ボーナスの平均額、などなど日常的に分布の要約値として使われていますね。平均というのは、言うまでもなく、データの値をすべて足して、その個数で割るのですが、これって、２値データなら・・・。とりあえず、集計表から標本比率の平均を算出してみましょう。みなさんの標本比率は、だいたい、どのくらいだったのでしょうか。

標本分布の平均　＝

6．比率の標本分布の要約値として平均を考えましたが、平均はデータの真ん中を示す値ではありますが、分布のばらつき、もしくは横の広がり具合を示すものではありません。そこで登場するのが、分散もしくは、その平方根である標準偏差という統計量です（ちなみに、９の平方根は３、３の２乗が９ですね)。特に、標本分布の標準偏差は標準誤差とよばれ、あらゆる統計解析の結果に記述されています。標準偏差は平均と対をなすもので、平均からの（平均的な）離れ具合を示します。

標本分布の標準偏差（標準誤差）＝

7．平均と標準偏差が分かれば、データの真ん中とそこからのばらつき具合が分かります。というか、分かると言われています。みなさん、平均と標準偏差を聞いただけで分布の形が分かりますか？実は、それだけでは分からないのです。分かるという人もいるでしょうが、その人達はもう一つ、重要な事を知っています。それが、正規分布です。正規分布は分布の形の一つで、釣鐘状の、もしくは、なだらかな山のような分布をしています。そして、平均が山の頂上のＸ座標を、標準偏差が山の斜面のなだらかさを示します。みなさんから集計した標本比率の分布を、算出した平均と標準偏差を持つ正規分布と重ねて比較してみます。どうですか？平均と標準偏差がわかれば、正規分布を使って分布の形が近似できそうですね。実は、比率の標本分布はデータの数が増えると、本当に正規分布になっていくことが数学的に証明できます。正規分布の形というのは式で書こうとすると、とんでもなく複雑なのですが、いくつかの有用なことが知られています。ここでは、標本分布を



近似する形で正規分布を紹介しましたが、もちろん、一般的なデータの分布を正規分布で近似することができます。近似と聞くと難しそうですが、必要なのは平均と標準偏差だけです。電話で、データの平均と標準偏差を言えば、聞き手はその平均と標準偏差を持つ正規分布を思い浮かべて、なるほど、こういう分布をするデータなのかとイメージが持てます。

8．正規分布の形は、平均の近くで一番多く観測されていて、平均から遠ざかるほど左右対称にその頻度がなだらかに減少していきます。平均からどのくらい離れると、どのくらい減っていくのか、または、平均の周りがどのくらい観測されやすいか、気になるところです。ズバリ、平均から左右に２倍（正確には、１．９６倍）の標準偏差、略して、平均±２σ（シグマと読む）の範囲に９５％の確率で観測されることが分かります。忘れてはいけません、残り５％は、もっと裾側で観測される可能性があります。そして、３標準偏差の範囲で、およそ９９％です。したがって、大まかに言えば、正規分布の山の形は、２σの範囲にその大体が、３σまでの範囲にほぼすべてが、すっぽり入ります。

9．さて、話を比率の標本分布の話に戻しましょう。わずか２０個のデータから母比率を推定しようとしていたのですが、みなさんがそれぞれ算出した比率は色々な値を取ってはいるものの、全体としてどの辺にどの程度集中しているか分かってきました。ここで、数理統計学の出番です。ちょっと難しいので、聞き流して下さい・・・。実は、比率の標本分布の中心は母比率と一致しており、「母比率±２標準誤差の範囲に、標本比率が観測される確率は９５％」であり、言い換えると「標本比率±２標準誤差の範囲に母比率がある確率は９５％」となります。標本比率は、すでに手持ちでのデータで算出できています。標準誤差は？あと、標準誤差、すなわち比率の標本分布の標準偏差が分かれば、手持ちでのデータだけでも母比率の居場所の範囲が判るのですが、実際、標準誤差も手持ちのデータから推定可能です。というわけで、手持ちの２０個のデータから、母比率は次の範囲に９５％の確率で存在することが分かります。これを、９５％信頼区間とよびます。みなさんも手持ちのデータから母比率の９５％信頼区間を作って下さい。

標準偏差＝｛標本比率＊（１－標本比率)｝の平方根

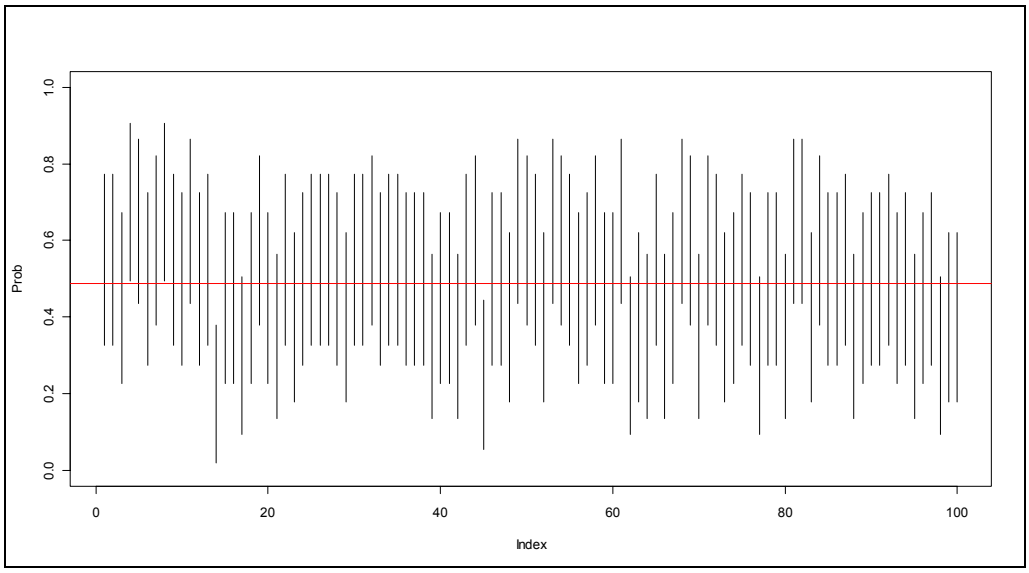
標準誤差＝　標準偏差÷（２０の平方根）

母比率の９５％信頼区間　＝標本比率±１．９６＊標準誤差

1 の数	1	2	3	4	5	6	7	8	9	10
標本比率	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
標準誤差	0.05	0.07	0.08	0.09	0.10	0.10	0.11	0.11	0.11	0.11
信頼区間下限	0.00	0.00	0.00	0.02	0.06	0.10	0.14	0.19	0.23	0.28
信頼区間上限	0.15	0.23	0.31	0.38	0.44	0.50	0.56	0.61	0.67	0.72
1 の数	11	12	13	14	15	16	17	18	19	20
標本比率	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	1.00
標準誤差	0.11	0.11	0.11	0.10	0.10	0.09	0.08	0.07	0.05	0.00
信頼区間下限	0.33	0.39	0.44	0.50	0.56	0.62	0.69	0.77	0.85	1.00
信頼区間上限	0.77	0.81	0.86	0.90	0.94	0.98	1.00	1.00	1.00	1.00

【標本比率、標準誤差および９５％信頼区間】

１０．　いつの間にか、母比率の９５％信頼区間まで構成してしまいました。信頼区間の構成は、区間推定ともよばれています。それに対して、標本比率で母比率をピンポイントで推定することを点推定とよびます。はじめに、推定と検定をあわせて統計的推測とよぶ、という話をしましたが、最後に簡単に検定の話をします。母比率の９５％信頼区間を構成したのですが、それは言い換えると、その区間の外に母比率がある可能性が５％あるということです。このように、ある値から外側の値をとる確率を外側確率とよぶことがあります。それでは、確かめてみましょう。みなさんにお配りした２値データの母比率は実は・・・なのですが、自分で作った９５％信頼区間に入ってなかった人はどのくらいますか？１００人いれば平均的に５人程度は、母比率を含まない信頼区間を作っているはずです。それが、９５％信頼区間の意味です。



【１００人が作成した９５％信頼区間と母比率】

１１．　理解が深まったところで、次の（帰無）仮説が正しいかどうか統計的に判断することに利用してみましょう。

仮説「母比率は０．２である」

実は、難しいことを考えなくても、すでにみなさんは答えをお持ちです。みなさんが作った９５％信頼区間に０．２が含まれていなければ、５％の過ちを犯す可能性はあるものの、仮説は棄却されます。なぜなら、みなさんの９５％信頼区間に母比率が含まれる確率は９５％なのですから。このように、ある仮説を立てて、統計的にその仮説の正しさを判断することを統計的検定、もしくは、単に検定とよびます。このときの誤って仮説を棄却する確率は有意水準とよばれます。また、０．２がぎりぎり入らないように信頼区間伸び縮みさせたときの外側確率はＰ値とよばれます。９５％信頼区間の外側に０．２がある人は、Ｐ値を５％よりも小さくすることができますね。

１２．　最後に何をやったか、おさらいです。２値データの要約値として比率を計算しましたが、目の前のデータは、たまたま観測された一組のデータに過ぎないという立場から、統計学の力を借りて、その母比率の存在するであろう信頼区間を構成しました。そして、さらに検定問題にも触れました。実は統計学の存在意義は、この検定にあると言っても過言ではありません。もし、統計を使わなければ、上の仮説をどう扱えばよいでしょう？自分の標本比率が

ちょうど０．２のときだけ、仮説は正しいのでしょうか？それとも、０．３くらいなら、まあまあ許されますか？その基準は誰が決めますか？このような無益な議論を回避するために、統計的検定は『データから客観的な判断を行う唯一の方法』として科学分野で広く利用されています。

誕生日	1月	2月	3月	4月	5月	6月	7月	8月	9月	10月	11月	12月
1日	1	1	1	1	1	0	1	1	1	1	1	0
2日	1	1	1	0	1	1	1	0	0	0	0	0
3日	1	1	1	1	0	1	0	0	1	0	0	1
4日	0	1	1	1	0	1	1	1	1	1	0	1
5日	0	1	0	0	0	0	1	1	0	0	0	1
6日	0	1	1	0	1	1	0	1	0	0	1	0
7日	0	0	1	1	0	0	0	0	1	1	0	0
8日	0	1	1	0	0	1	0	0	1	1	0	0
9日	0	1	0	1	1	0	0	0	0	1	0	0
10日	1	0	1	1	0	1	1	0	1	0	0	1
11日	0	1	0	0	0	0	0	0	0	1	0	1
12日	0	0	0	0	1	1	1	1	0	0	0	1
13日	1	1	0	1	0	1	0	0	1	1	1	0
14日	0	0	0	0	0	1	0	0	0	1	0	0
15日	0	1	0	1	1	1	1	0	1	0	0	1
16日	1	0	1	0	0	0	1	1	0	1	0	1
17日	1	0	0	0	1	0	1	1	0	1	1	1
18日	1	0	0	1	1	0	0	0	0	1	0	0
19日	0	1	1	0	0	1	0	1	1	1	0	0
20日	0	0	0	0	0	1	0	1	0	1	1	0
21日	1	0	1	0	1	0	0	0	1	0	1	0
22日	0	1	0	0	1	1	1	0	1	1	0	0
23日	1	0	0	1	0	1	0	1	0	1	0	1
24日	0	1	1	0	1	1	1	0	0	0	1	0
25日	1	0	0	1	0	1	1	0	0	1	0	1
26日	1	0	1	1	0	1	1	0	1	1	1	1
27日	0	1	0	0	0	0	1	1	0	1	1	0
28日	0	1	0	0	0	1	1	0	0	0	0	1
29日	0		1	0	0	0	0	1	1	1	0	1
30日	0		0	0	0	1	1	0	1	1	1	0
31日	0		0		0		1	0		1		0