

「確率とその応用」ノート、その壱

逆瀬川浩孝

このノートは、テキストの「確率とその応用」(サイエンス社)を読むための補助教材というつもりで書きました。テキストはページ数の制約があり、無味乾燥な記述だけで済ませているところが多々あり、苦勞させられるかもしれません。その行間を埋めるように、追加の例題や、問題練習用の練習問題、演習問題を補足しました。「ちゃんと勉強すれば誰にでも分かる確率とその応用」を目指しています。ちゃんと勉強しても分からないところがあったら、遠慮無く質問してください。ホームページにもサポート資料を随時掲載するつもりですので、そちらも参考にしてください。

<http://www.f.waseda.jp/sakas/probability/>

1 確率ア・ラ・カルト

1.1 世論調査

確率があるおかげでこんなことができる、という一つの例として世論調査を取り上げます。大新聞社は毎月のように内閣の支持率を調べて発表しています。誰に聞いても「そんな調査、されたことない」と答えるでしょう。それもそのはず、調査結果の掲載されている紙面を注意深く読むと、調査対象は1000人(あるいは3000人)と書いてあります。1000人では調査された経験が0でも不思議ではありませんが、ほとんどの人が支持不支持を表明しないのに、これが全体の意見です、と言っても良いというのは不思議です。

このような調査の仕方は**標本調査**と呼ばれています。もし、「ちゃんと(正確に)」調べようと思えば、一億人くらいの有権者全員の意見を聞いて回り、支持する人の数を数えなければいけません。誰がいつどのようにして意見を集めることができるでしょうか? 1000人が調査に当たるとして、一人あたり10万人を調査しなければいけません。一人1分で終わるとしても10万分、1日8時間労働でも200日かかります。そんなことしなくても、全員に調査依頼を送って、回答をインターネットで返してもらえばそんなに時間はかからないはず、と思う人もいるかもしれないけれど、依頼状をちゃんと読んで素直に反応してくれることは期待できません。

あるいはまた、別の視点で考えてみましょう。標本調査結果を報じる記事の中に回収率という言葉を見ることができます。調査対象として決めた人たちに調査した中で、意見を聞いた人の割合が回収率です。良くて6割くらいの数値が書かれていますが、短時間に補足できる人の割合はそんなものです。5年に一度、国の統計を取るために、全国民を対象とした国勢調査が行われます。膨大なお金をかけて人海戦術(って分かりますか?)で調査したとしても数%は回収できないらしい。というわけで、1億人全員の意見を集めることなど、およそ不可能です。

このようにして、費用、時間、実行可能性を考えると、全員の中から何らかの意味で全体を代表するような一部を取りだして調査し、それを全体の意見の代わりとする、という方法を取らざるを得ません。そのような調査でも、やり方を工夫すると、ある程度全体の意見を反映した結果が得られる、ということが、確率論を使って保証されています。その理屈を考えてみましょう。

1.1.1 無作為抽出(ランダムサンプリング)

全体の一部を調べて、それが全体の意見を代表していると言えるためには、その調べた全体(標本といいます)が、全体のミニチュア、縮小模型、になっている必要があります。内閣支持率

調査の場合、全体は日本人すべてですから、そのミニチュアを作るのはそう簡単ではないでしょう。しかし、今の場合、日本人全体のミニチュアを作るというように考えないで、日本人全体の「意見の」ミニチュアを作ると考えれば良いのです。つまり、抽出された標本に含まれる支持者の割合が、全体の支持率とほぼ等しい、と言えるような標本を抽出すれば良いのです。

全体の支持率が分からないのに、それに近い標本を集めるなんて無理、という反論が予想されますが、くじ引き、あるいは抽選、の考え方を使って、目的の標本抽出を実現することができます。こういう実験を考えてみてください。壺の中に、多くの赤い玉と白い玉を入れ、よくかき混ぜてから 100 個を取り出すことにします。その中に 30 個の赤い玉が含まれていたとすると、壺の中の赤い玉と白い玉の個数に関して、何か言えませんか。あるいは、一見いびつなさいころを 100 回振ったら 1 の目が 30 回も出たとして、このさいころの 1 の目が出る確率について何か言えませんか。1000 回振って 300 回だとしたら、どうですか。

赤い玉と白い玉の個数は 3:7、1 の目が出る確率は 0.3 という「推定」に対して、多くの人は納得するに違いありません。それは、よくかき混ぜるという動作、あるいはさいころを転がすということで、実験をする人の意思とは無関係に実験結果が出現すると考えられるからです。逆に、ちゃんとしたさいころを 100 回振って 1 の目が 30 回出るか、と考えると、そんなことはほとんどあり得ないので、30 回 1 の目を出すさいころは歪んでいると考えるのが合理的です。

練習問題 1.1 正しいサイコロを 100 回振って 1 の目が 30 回以上出る確率はいくつですか。

標本調査にもこのような抽選のメカニズムを組み込むことができれば、調査する人の思惑が結果に影響を与えない、公平な結果が得られそうです。このような仕組みを取り入れた標本の抽出の仕方は**無作為抽出（ランダムサンプリング）**と呼ばれています。

無作為抽出を取り入れた調査は次のようにすれば実現できます。抽出対象の日本国民全員に通し番号を付け、最初の 10 万人の中からランダムに一人選びます（さいころを 7 回振って、出た目の数マイナス 1 を並べたものを 6 進法 7 桁の数だと思えば良い。2,3,5,5,1,6,2 だとしたら、1244051 を 6 進法の数とみなし、それを 10 進数に変換すると 68287 となります）。そこから 10 万番おきに取り出せば 1000 人くらいの標本が得られます。

実際には、通し番号を付けるのは難しいので、ランダムデジタルダイアリング法という方法を使って調査しています。ランダムに電話番号（固定の）を生成し、その電話番号を持つ人に対して電話調査をする、という方法です。世帯を抽出するので、大家族の中の一員と単身世帯では当たる確率が違います。また、携帯しか持っていない人は調査対象に含まれていないなど、完全な無作為抽出ではありません。

1.1.2 無作為抽出法の結果の評価

無作為抽出による標本調査は、全体のミニチュアを作るという意味で、多くの人を納得させるものですが、さらに、確率論を使うことによって、理論的にも正しい方法であることが裏付けられ、その結果がどれくらい（未知の）真の値と食い違う可能性があるのかを数量的に評価できます。

実際、今まで知っている確率論の知識から、次のような計算ができます。もし、壺の中に 10 個の赤い玉と 10 個の白い玉があったとすると、その中からランダムに 6 個を取り出したとき、その中に赤い玉が 3 個含まれる確率は 2 項係数

$${}_nC_k = \binom{n}{k} = \frac{n!}{k!(n-k)!} \quad (1.1)$$

を使って $\binom{10}{3}\binom{10}{3}/\binom{20}{6}$ と計算されます。玉の数がもっと多くなっても同じことです。一般に、赤い玉が n 個、白い玉が m 個ある時、その中から k 個の玉を選ぶ場合、その中に i 個の赤い玉が含まれる確率は

$$\frac{\binom{n}{i}\binom{m}{k-i}}{\binom{n+m}{k}} \quad (1.2)$$

と計算できます。

この考え方を使うと、いろいろな計算ができます。一人一人の有権者の意見を玉の色に置き換えます。支持ならば赤い玉、不支持ならば白い玉を持ってもらいます。有権者が 1 億人として、1 億個の玉をよくかき混ぜてその中から 1000 個の玉を選び、その中に含まれる赤い玉の比率を数える、というのが標本調査のやり方になっていることが分かるでしょう。赤い玉、白い玉の個数 n, m は分かりません。調査で分かるのは $k = 1000$ と i です。知りたいのは調査結果 i/k と真の値 $n/(n+m)$ がどれくらい近いのか、ということです。

でたらしめに取り出すわけですから、下手すると、全員支持すると答えるかもしれない、その場合は支持率 100%(!)、いくらなんでも、こんな調査は誰も信用しませんね。そういう可能性はどれくらいあるか、計算することができます。たとえば、 $n = m = 50000000, k = 1000$ として $i = 0$ の場合の確率を計算すると、 10^{-302} くらいになります。そんなことが起きることはありえません。 $i < 450$ となる確率を計算すると、約 0.0008、これもあり得ない話です。ということは、もし支持率が 5 割だったとき、1000 人に聞けばほぼ確実に 45% から 55% の間の数字が得られるということになります。

この議論を逆に使うと、もし調査対象者の 4 割しか支持していない、という調査結果が出たとすれば、全体の支持率が 5 割ということはありませんか。今までの計算は壺の中の赤い玉白い玉の個数が分かっているものとして計算しましたが、そういう計算をたくさんしておく、壺の中の玉の数が分からなかったとしても、取り出した玉の数の比率から壺の中の赤い玉の割合がある程度の精度で推定可能になるのです。これが無作為抽出を取り入れる最大の効果です。

参考 組み合わせ確率の計算

$\binom{n}{k}$ を計算する Excel の関数は COMBIN(n, k) ですが、 $n = 50000000$ を計算させようとするとき「#NUM!」と表示されてしまいます。計算できるのはせいぜい $n = 10000$ 程度のように。 n, m が非常に大きく、それに対して、 k, i はほとんどゼロと見なしても構わないくらい小さい、という場合は、次のような近似計算に切り替えると、 $n = 50000000$ の場合にもどうにか計算することができるようになります。

$$\begin{aligned} \frac{\binom{n}{i}\binom{m}{k-i}}{\binom{n+m}{k}} &= \binom{k}{i} \frac{n(n-1)\cdots(n-i+1) \times m(m-1)\cdots(m-k+i+1)}{(n+m)(n+m-1)\cdots(n+m-k+1)} \\ &\approx \binom{k}{i} \frac{n^i m^{k-i}}{(n+m)^k} = \binom{k}{i} \left(\frac{n}{n+m}\right)^i \left(\frac{m}{n+m}\right)^{k-i} \end{aligned}$$

ここで、 $n-k \approx n, m-k \approx m$ であることを使っています。これにより、 $n = m = 50000000, k = 1000$ として $i = 0$ の場合の確率は「=COMBIN(1000,0)*0.5^1000」によって（近似的に）計算することができるようになり、 9.3326×10^{-302} が得られます。

1.2 ランダム回答法

世論調査は、回答者全員が本音を回答することが前提です。もし、質問項目が他の人に知られたくない内容を含むようなものがあつたとしたら、普通の調査では正確な回答が期待できません。そのような場合にとられる方法として、回答を封印して回収するとか、回答を匿名にして郵送させるとかの手段を取りますが、ホントウに秘密が守られるのか、という不安が残ります。対面調査（調査する人とされる人が顔を付き合わせる）で、回答者の秘密が守られるような都合の良い調査のやり方はないのでしょうか。

世論調査のようなアンケートは、**一人一人の答えを知るのが目的ではなく、集団としてどのような意見が多いかを調べるものです**。一人一人の回答は調査した人には分からないけれど、全体の意見（割合）は大ざっぱに掴むことができる、というような調査法があれば、それで充分です。でもそんな都合の良い話があるのでしょうか。その無茶な要求に応える一つの例がランダム回答法です。

例を使って説明しましょう。あまり切実な問題を例に出すと、注意がそちらに集中してしまい、肝心なことが見えなくなってしまう恐れがありますので、ここでは

あなたは、喫煙の習慣がありますか？

という質問を例として取り上げることにしましょう（「喫煙は人を殺す」というメッセージをたばこのケースに印刷しないと売ってはいけない、という国もあるというのに、日本ではせいぜい「妊娠中の喫煙は、胎児の発育障害や早産の原因の一つとなります...」「喫煙は、あなたにとって肺気腫を悪化させる危険性を高めます」など、ソフト路線です。有害は分かっているのに。<http://www.mhlw.go.jp/topics/tobacco/qa/>）。このクラスの中の喫煙習慣のある人のパーセンテージを調べようというのが目的です。分かったからどう、ということでもありませんが。

さて、この目的のためには、クラスの1人1人に上の質問に答えてもらう必要がありますね。その調査で「はい」と答えた人の数を全体の人数で割ればこのクラスの喫煙率が計算できます。しかし、その調査結果は信用できるのでしょうか。たばこを吸っているかいけないかということくらいならば（答えたくない人もいるでしょうが）正直に答えてくれるかもしれませんが、もっと微妙な問題の場合（たとえば「ドラッグの経験がありますか」「カンニングの経験は？」）には正直に答えてくれる人ばかりとは限りません。

そこで、ちょっと工夫します。もしも

質問1： あなたは、喫煙の習慣がありますか？

質問2： あなたは、喫煙の習慣がありませんね？

という二つの質問のどちらかをでたらめに選んで、その質問の**答えだけ**を（どちらの質問を選んだかはもちろん言わずに）答えてください、というように聞かれたらどうでしょう。

質問の答えは「はい」か「いいえ」です。調査した人が受け取るのはその二つの答えのどちらかだけです。「はい」と答えた人は

- 質問1に対する答えの「はい（喫煙の習慣があります）」なのか
- 質問2に対する答えの「はい（喫煙の習慣はありません）」なのか

分かりません。つまり、答えを聞いただけでは、答える人が喫煙の習慣があるかどうかを知るこ

とはできません。プライバシーが守られているわけですから、直接質問した場合よりは正直に答えようとする人の割合が増えることで期待できます。

「二つの質問をでたらめに選ぶ」と書きましたが、全くでたらめに選んだとすれば、当然のことながら調査によって得られるものではありません。そこで、ちょっとゆがんだでたらめさを決めて、それに従って選んでもらうことにします。たとえばさいころを振って出た目が1か2ならば質問1、さもなければ質問2に答える、というようにしてもらうことにします。以上の説明から、実際の調査の仕方をまとめてみたのが次の質問です。

私に見えないようにしてさいころを振ってください。
もし出た目が1か2だったら、次の質問に「はい」か「いいえ」で答えて下さい
「あなたは、喫煙の習慣がありますか？」
出た目がそれ以外だったら、次の質問に「はい」か「いいえ」で答えて下さい
「あなたは、喫煙の習慣がありませんね？」

このようなアンケート法を**ランダム回答法**といいます。

自分がこの調査に答えなければいけないとしたとき、「はい」と答えても「いいえ」と答えても、自分が喫煙の習慣があるかないかを調査員に知られることはない、ということが「ホントウに」納得できますか。

1.3 ランダム回答法の理屈

ランダム回答法では、回答が「はい」だといっても、その人が喫煙の習慣があるのかわかりません。両方のケースが混ざっているのです。こんなあいまいな調査をして、その数字から何か分かることがあるのでしょうか。

この場合も、赤い玉白い玉を使って標本調査の理屈を計算したのと同じように、あらかじめ分かっている数字を使って計算してみます。もしクラス120人の中で、喫煙率（これを r と書きましょう）が40%($r = 0.4$)と分かっているとすれば、「はい」という答えのおおよその数を見積もることができます。喫煙率40%ですから、喫煙の習慣がある人は120人中48($= 120r$)人くらいいます。「はい」と答えるのはそのうちの1か2の目が出た人です。1か2が出る確率は3分の1と言って良いでしょう。したがって、48人の3分の1、16人くらいが「はい」と答えるはずです。

一方喫煙習慣がない人が「はい」と答えるためには1か2以外の目が出る必要があります。その確率はだいたい3分の2と言って良いでしょう。したがって72人の3分の2、48人くらいが「はい」と答えると考えて良いでしょう。

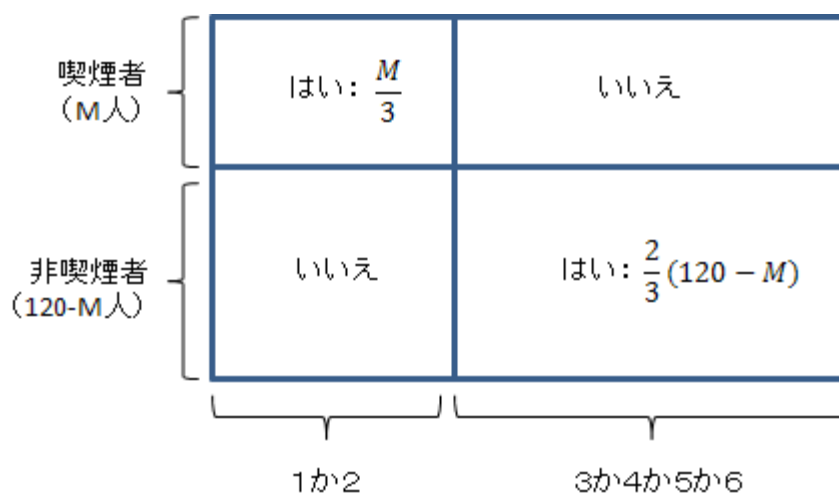
さいころの目	喫煙習慣あり (40%)	喫煙習慣なし (60%)
1 または 2	16 名	24 名
3,4,5,6	32 名	48 名

結局、「はい」と答える人はおおよそ $16 + 48 = 64$ 人いることになります。つまり全体の $53.3\%(= 64/120)$ が「はい」と答えることになります。同様に、もし喫煙率が30%のクラスで調査した場合は $12 + 56 = 68$ 人が「はい」と答え、20%のクラスで調査した場合は $8 + 64 = 72$ 人が「はい」と答えるということも計算できます（確かめてください）。表にまとめると、以下のようになります。

喫煙率	「はい」回答の数
40%	64
30%	68
20%	72
?	76

ここまで来れば、喫煙率を知らずに、「はい」回答の数を聞いただけで、おおよその喫煙率が分かりそうですね。「はい」回答が 76 ならば喫煙率は 10%? 「はい」回答が 60 ならば喫煙率は 50%?

そのとおりです。「はい」の数が分かると、おおよその喫煙率が計算できるのです。次のような図で考えてみましょう。



クラスの喫煙者の数を M とすると、喫煙者で「はい」と答える人はそのうちの 3 分の 1 です。非喫煙者 $120 - M$ 人の中で「はい」と答える人はそのうちの 3 分の 2 です。

実際の調査で 68 人が「はい」と答えたとすれば、

$$\frac{1}{3}M + \frac{2}{3}(120 - M) = 68$$

という式が成り立っています。未知数は M 一つだけですから、この方程式を解くことで

$$M = 36 \Leftrightarrow r = \frac{M}{N} = \frac{36}{120} = 0.3(30\%)$$

のように、喫煙率を計算することができました。

この計算過程を一般化してみましょう。調査対象が N 人で、未知の喫煙率 r が知りたいものです。喫煙者数は $M = rN$ です。肯定的な質問を選ぶ確率を p として、調査結果として「はい」回答が m だったとしましょう。そうすると、次の方程式が得られます。

$$\begin{aligned} pM + (1-p)(N-M) &= m \\ \Leftrightarrow r &= \frac{M}{N} = \frac{\frac{m}{N} - (1-p)}{p - (1-p)} = \frac{1}{2p-1} \left(\frac{m}{N} - (1-p) \right) \end{aligned}$$

最後に得られた右辺の式を、喫煙の習慣がある人の比率 r の**推定値**と言います。結局、質問を選ぶ確率 p と、その調査結果から得られた「はい」の回答数 m 、調査の参加者数 N が分かると、(直接質問しなくても) 喫煙率が推定できたことになります。

理屈は分かって、何とも不思議な話ですね。みんな答える質問をさいころに従って選んで答えたので、その個人のホントウの答えは分からないのに、全体の比率は分かる、というのは、そう簡単に納得できる話ではないでしょう。でも、どこにもトリックを使った形跡は見あたりませんし、何度読み返しても、インチキは見つからないはずです。しかし、方法自体は間違いではなくても、「何となくうさんくさい」「納得が行かない」というのが大方の感想でしょう。何がその「うさんくさ」さの原因か、よく考えてください。

練習問題 1.2 ランダム回答法で「はい」の回答数が1増えたとき、最終結果の % にどれくらい影響が出るかを計算しなさい。

練習問題 1.3 分母が0、すなわち p が0.5の場合に r は推定できません、なぜでしょう。直感的な説明を考えなさい。

1.4 コンピュータを用いた確率実験

ランダム回答法は肯定的質問と否定的質問をランダムに選んで回答してもらうことにより、回答者の協力を得やすくし、プライバシーに関わる事項の集団的特徴を調査する手法でした。平均的にはこうするはずだ、ということを根拠として推定値を導きましたが、そんな正確に分かるはずがない、どこか変だ、と考える人も多いことでしょう。「どこか変だ」で止めないで調査法の論理をチェックして行くと、その原因が見えてくるはずです。

1.4.1 平均値とばらつき、シミュレーション実験

「はい」の数を数えることが基本でしたが、「平均的に」求めた数が調査でそのまま得られるという保証はありません。コインを 10 回投げれば平均的に 5 回は表が出るはず、というのは間違いではありませんが、実際に 10 回投げた時に必ず 5 回表が出るというわけではありません。実際には表が 6 回以上出る確率は $\frac{3}{8}$ 以上あります（正確な数字を計算してみてください）。

そこで、「平均的な」数を使って求めた推定値と、実際に調査した場合の推定値の差について調べてみることにしましょう。一人一人にさいころを渡して、質問をサイコロの目にしたがって選択し答えてもらう、ということを繰り返せばよいのですが、いくら協力的な人でも、そのうちいやになってしまうでしょう。そこで、この実験をコンピュータに代行させることを考えます。コンピュータにさいころを振らせ、そのときの回答を集めて推定値を計算するというのを繰り返します。コンピュータのさいころ振りは乱数を使って実現することができます（C 言語のプログラミングを思い出して下さい）。このように乱数を使ってランダム事象の分析をするやり方を**モンテカルロ法**、あるいは**モンテカルロシミュレーション**といいます。詳しくは教科書の 1 章 1.10 と 6 章、C テキストの 12 章を見てください。

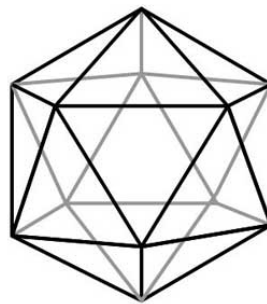


図 1 正二十面体

1.4.2 乱数

確率の理論や、確率に関する問題で分からないことがあったら、乱数を使って実験をすることで理解を深める、あるいはヒントを得ることができる。今のうちに慣れておくと良いでしょう。乱数について復習しておきましょう。乱数とは「デタラメのようにみえる数列」のことです。お手本はさいころ数列です。普通のさいころは 6 面体ですが、正二十面体の 20 個の面に 0 から 9 までの数を 2 回ずつ書いたものをさいころにすると、十進数の乱数列が得られます。その乱数列をよく観察すると、

1. それまでに出了目の記録を見ても、次にどの目が出るか誰も予測できない
2. さいころ投げの記録を集計していくと、0 から 9 までの数が同じような頻度になる

という特徴があることが分かります。逆に、このような性質を持つ数列のことを乱数ということにしましょう。

もし、数式を使ってそのような特徴を持つ数列が生成できるならば、それを乱数といってもよい、ということから**擬似乱数列**が考えられました。Excel の `rand()`、C 言語の `rand()` はそのような数列を生成します。どれくらいランダムか、実験してみましょう。

1.4.3 Excel の実験、rand 関数

セル A4 に「=RAND()」と入力してください。セル A4 には「0.742593」のような数が表示されるはずです。次にセル A4 をセル A5:A23 にコピーペーストして下さい。20 個の表示されている数を見て、何か規則性が感じられますか。次に「再計算キー（ファンクションキーの F9）」を押してみてください。セル A4:A23 には別の 20 個の数が表示されるはずですが、今度はどうですか。続けて何度か F9 キーを押して、セル A4:A23 がどのように変化するか、しばらく様子を見てください。

1.4.4 Excel の実験、コイン投げ

コインを 20 回投げて、おもての出る相対度数を調べる：セル B4 に「=if(A4<0.5,"おもて","うら")」と入力してください。セル B4 には「おもて」か「うら」のどちらかが表示されるはずです。次にセル B4 をセル B5:B23 にコピーペーストして下さい。これが実際に 20 回コイン投げをした結果、といわれても納得してしまうでしょう。実際は乱数を使って「計算」しただけなのに。次いで、セル B24 に「=COUNTIF(B4:B23,"おもて")/20」と入力して下さい。COUNTIF 関数は条件を満たすデータの個数を数えるときに使います。これで、セル B24 に「おもて」の相対度数が計算されます。こうしてから「再計算キー（ファンクションキーの F9）」を押してみてください。セル B4:B24 には別のコイン投げの結果が表示されるはずですが、前回とどう変わっていますか。

練習問題 1.4 コイン投げの回数を増やして (A,B 列の 25 行目以下にも同じ数式をコピーペーストする)、おもての出る相対度数の数値がどのように変わるか、調べて下さい。

1.4.5 Excel の実験、世論調査の実験

内閣支持率が 30% の集団から 1000 人を対象にして支持率調査を行った結果を Excel でシミュレーションする：新しいシートを表示させて、セル B3 に「0.3」、セル A7 に「=rand()」、セル B7 に「=if(a7<\$B\$3,"支持","不支持")」と入力し、セル A7:B7 をセル A8:B1006 にコピーペーストして下さい。B 列の文字列は抽出された標本の回答を表していると考えられます。セル C3 にその調査結果として「=COUNTIF(B7:B1006,"支持")/1000」と入力して下さい。再計算キーを押すとセル C3 の表示がどう変わりますか？ また、セル B3 の数値を変えると、調査結果はどう変わりますか？ B3 の数値を変えても変わらない特徴はありませんか？

1.4.6 ランダム回答法の実験

世論調査の実験で「支持」の代わりに「肯定（的質問選択）」と表示させれば、その調査対象がどちらの質問を選んだのかが分かります（ランダムに決まります）。その調査対象が喫煙習慣があれば肯定的質問に答えるので「はい」、さもなければ「いいえ」と答えるので、そのように表示させる数式を使えばランダム回答法がコンピュータで実験できます。

これからの実験では、次のようなシートを作るのが目標です。

	A	B	C	D	E
1	ランダム回答法実験				
2	前提条件	調査参加者数		100	
3	前提条件	喫煙率		0.3	
4	前提条件	肯定質問選択確率		0.3	
5	計算結果	肯定質問選択数		33	0.33
6	計算結果	"はい"回答数		57	
7	計算結果	喫煙率の推定値		0.325	
8					
9	氏名	喫煙	乱数	質問	回答
10	ダル	1	0.066113	肯定	はい
11	イチロー	0	0.468253	否定	はい
12	青木	1	0.152737	肯定	はい
13	4	1	0.542504	否定	いいえ
14	5	1	0.330635	否定	いいえ
15	6	1	0.692385	否定	いいえ
16	7	1	0.294625	肯定	はい

図2 ランダム回答法コンピュータ実験のシート見本

10行目から下は調査結果の個人記録を表にしています。「氏名」「喫煙（喫煙するか否か）」「乱数（質問選択のための乱数）」「質問（どっちの（肯定、否定）質問を選んだか）」「回答（最終回答（はい、いいえ））」の5項目です。セルD2:D7は調査対象が100人だった場合の集計結果を表示しています。

シートの作り方

- セルA1:B7、セルA9:E9にシート見本のような文字列を入力、セルD2に「100」、セルD4に「0.3」と入力する。セルA10:A109に調査協力者の名前を適当に100人分（通し番号でも構いません）、セルB10:B39に1（喫煙習慣有り）、B40:B109に0を入力。
- セルD3にはセルB10:B109の平均値（つまり、100人の喫煙率）を計算する式を入力してください（平均値関数は「=AVERAGE(..)」です）。
- セルC10には「=RAND()」と入力してください
- セルD10には（上の例では）「ダル」がどっちの質問を選んだかを「肯定」「否定」の文字列で表示します。やり方は上の実験で説明済み。この式は後でコピーペーストしますので、セルD4を「絶対参照（\$D\$4）」する必要があります。

5. セル E10 はちょっと複雑です。喫煙者（セル B10 が 1）が肯定的質問を選ぶ（セル D10 が「肯定」）か、非喫煙者（セル B10 が 0）が否定的質問を選ぶ（セル D10 が「否定」）とき回答は「はい」で、それ以外は「いいえ」です。Excel の関数では
「A かつ B」は「=AND(A,B)」(=AND(B10=1,D10="肯定"))、
「A または B」は「=OR(A,B)」(=OR(AND(B10=1,...))
と表します。（最終形は、IF(OR(AND(...),AND(...)),"はい","いいえ")のような形）
6. セル C10:E10 をセル C11:E109 にコピーペーストします。これで 100 人分の調査が完成です。
7. (集計) セル D5 は肯定的質問を選ぶ人の数 (countif)、セル D6 は「はい」回答の総数を計算する式 (countif) を入力して下さい。そして、最後にセル D7 に喫煙比率の推定式を入力して、完成です。ただし、乱数の偏りによって、推定結果がマイナスになったり、1 を超えることがあるかもしれませんから、その場合はそれぞれ、0,1 で置き換えるものとします (max,min)。

再計算キーを押しながらセル C7 がどのような範囲で動くのか、よく観察してください。セル E5 に「=D5/D2」という数式を入力しておけば、肯定質問の選択比率が表示されるので、それとセル D4 の差がどれくらいあるかということにも注意が必要です。

練習問題 1.5 再計算キーを押しながら、喫煙率の推定値をいくつかメモしなさい。これによって何が分かりますか。

練習問題 1.6 セル E5 とセル D4 の違いを観察して、分かったことを説明しなさい。

練習問題 1.7 肯定的質問の選択確率を 0.05 としたとき、推定値のばらつきはどうなりますか。肯定的質問の選択確率を 0.45 としたとき、推定値のばらつきはどうなりますか。肯定的質問の選択確率と推定値のばらつきについて分かったことを説明しなさい。

練習問題 1.8 肯定的質問の選択確率とプライバシーの保護の関係を説明しなさい。

問題 1.1 以上の結果を踏まえて、次の問題に対して、理工学部学生にふさわしい報告書を書きなさい。

1. ランダム回答法で、肯定的質問の選択確率と喫煙率の推定精度の関係を説明しなさい。
2. ランダム回答法による喫煙率の推定精度と、普通の調査法による喫煙率の推定精度について比較検討しなさい（「喫煙」が「ドラッグ」だったらどうですか）。
3. ランダム回答法という調査法についての評価、感想を書いてください（調査の精度とプライバシー保護の関係など）。

注意

「最初から推定値が分かっている実験は意味があるのか？」こういう疑問を持つ人も少なからずいるようです。喫煙率を推定することが目的ならば確かにその通り、分かっている数値を推定して何になる？ 「乱数実験」の目的を理解しないとそのような否定的評価が出てきます。目的は

**1 回しか調査できないときに、調査結果から推定値を計算して
これが真の値ですよ、と主張することはどれくらい意味のあることなのか**

を知ることです。つまり、推定値と真の値とのずれがどれくらいあるかということが知りたいのです。真の値を隠して、実験結果だけを人に見せた場合、その人が真の値についてどれくらい推測可能か？ ということを調べたいのです。

それがたとえば 5% 以内ということが分かれば推定値を真の値と考えて行動してもあまり間違いはありませんが、30% 位の誤差は覚悟してください、と言われると、調査自体の信頼性が疑問視され、その推定値だけでは次の行動は起こせない、ということになります。この 5% とか 30% という数字が N や p 、あるいは r と関連させてどのようになるのか、それを知るのがこの乱数実験の目的なのです。

エクセルヒント：グラフの描き方

Excel で散布図を描く方法

0.05, 0.1, 0.3, 0.45, 0.55, ... のように不等間隔の値に対して計算した結果を表示する場合は、数学の関数を書くときと同じで「折れ線グラフ」ではなく「散布図」を使って横軸の目盛りを単位長さの倍数にしないといけません。「挿入／グラフ／散布図」の中から「折れ線」オプションを選択してください。データがない場合は空白にしておけば、線は描かれませんが、

おわり

(参考) ランダム回答法の改良 (好奇心が旺盛な人へ)

こんな粗っぽい方法でも何にも分からないよりはましですが、その推定精度を良くする方法についても研究されています。その一つを説明します。これも Excel で確かめることができますから、どれくらい改良されるのか、実験してみてください。

ある確率を決めて、ランダムに質問を選んでもらうことは同じなのですが、今度は「肯定的質問」と「否定的質問」を選んでもらう代わりに、「肯定的質問」と「回答の確率が分かっている質問」をランダムにを選んでもらう、という方法です。「回答の確率が分かっている質問」というのは、たとえば「コインを振っておもてが出たら”はい”、さもなければ”いいえ”と答えてください」というようなものです。つまり、

私に見えないようにしてさいころとコインを振ってください。

もし出た目が 1 か 2 だったら、次の質問に「はい」か「いいえ」で答えて下さい

「あなたは、喫煙の習慣がありますか？」

もし出た目がそれ以外だったら、次の質問に「はい」か「いいえ」で答えて下さい

「コインは表ですか？」

普通の質問に回答する人はコイン投げをする必要はないのですが、全員にコイン投げをしてもらう、というのがみそです（もし該当する人だけがコインを投げたらどうなりますか？）。

調査協力をお願いしておきながら、肝心の質問に答えずにコイン投げの結果だけを答えさせるとは何事か、と怒られそうな方法です。また、本当の質問に回答する人が減ってしまうわけですから、ますます推定値の精度が悪くなりそうな感じがします。驚いたことに、このようなやり方で「改良」になっているのですね。

質問の選択確率を p 、無関連質問の「はい」回答の確率を q 、喫煙者数を M 、調査協力者数を N 、調査における「はい」回答の数を m とすると、以下の式が成り立ちます。

$$pM + (1 - p)Nq = m$$

従って M/N の推定値は以下の式で表すことができます。

$$M = \frac{1}{p}(m - (1 - p)Nq) \Leftrightarrow \frac{M}{N} = \frac{1}{p}\left(\frac{m}{N} - (1 - p)q\right)$$

この「改良」法では質問の選択確率についての制約はなくなります。むしろ、選択した質問がどちらなのかを隠すためには p は 0.5 に近い方が望ましくなります。

練習問題 1.9 この新しい方法について、Excel を使って推定値の精度を計算し、オリジナルの方法と比較しなさい。何が分かりましたか。

1.5 乱数を使った実験の例

Excel の `rand()` 関数を使うと、確率論のいろいろな問題が実験できます。頭で考えても分からない場合、適切な乱数実験に置き換えて、大雑把な結果を求め、そこから正解を類推する、というような使い方は確率論の理解に役立つでしょう。次のような練習をやりながら、自分でも問題を見つけて実験してみてください。

練習問題 1.10 4 問の 3 択問題をまぐれで全問正解する確率を、Excel を使って実験的に計算しなさい。

ヒント：二つのセルに「`=int(rand()*3)`」と入力して、両者が等しければ正解、さもなければ不正解と判定する。

練習問題 1.11 3 枚のカードがあります。1 枚目は表裏とも赤色、2 枚目は表裏とも黒色、3 枚目は表が赤色、裏が黒色になっています。さて、でたらめに 1 枚選んで机の上に置いたところ赤色でした。そのカードの裏が黒色である確率はいくつでしょう、という問題を Excel を使って計算してください。

ヒント：1 枚をランダムに選び、表が「赤」の場合だけ、裏の色を確かめる、という式を入力すればよい。ランダムに選ぶには、「`=if(a4<1/3,"赤赤",if(a4<2/3,"黒黒",if(a4<5/6,"赤黒","黒黒")))`」とすればよい。最初の文字が「赤」であることを確かめるには「`=if(left(b4,1)="赤",...)`」とすればよい。

練習問題 1.12 ゆがみのないさいころを 100 回振った時に出る目を表示する Excel のシートを作りなさい。

ヒント：コイン投げは条件文を使って、`RAND()` が 0.5 未満かそうでないかによって、「おもて」と「うら」を表示していましたが、さいころの場合は、`RANDBETWEEN(1,6)` という関数を使うと簡単です。

練習問題 1.13 さいころ振り実験で出る目の度数分布を作図しなさい。実験回数をふやして、度数分布が安定していく様子を観察しなさい。

ヒント：`FREQUENCY(.)` という関数を調べなさい。

練習問題 1.14 1000 回くらいのコイン投げ実験をして、各回毎に表の相対度数を計算し、それを折れ線グラフで表示することにより、回数が増えると折れ線グラフの変動が小さくなっていく様子を観察しなさい。

ヒント：縦軸の表示範囲を $[0, 1]$ に固定する（軸の「書式設定」で指定する）と分かりやすい。

練習問題 1.15 40 人のクラスに、同じ誕生日（年は問わず月日のみ一致）を持つペアがいる確率はどれくらいあるか、Excel で確かめて下さい。

ヒント：RANDBETWEEN(1,365) とすれば 1 から 365 までのデタラメな整数が生成される。それを 40 個生成し、countif で自分と同じ誕生日を持つ人を数え、その最大値が 2 以上ならば同じ誕生日を持つペアが 1 組以上いたということになる。20 通りの計算結果のうち、同じ誕生日のペアがいるのはどれくらいの比率か数えれば良い。

参考図書

伏見正則「乱数」東京大学出版会、1989.

逆瀬川浩孝「ちゃんと勉強すれば誰にでも分かる C 言語プログラミング入門」、2012.

1.6 参考 微積分の知識

まだ先の話ですが、4章以降ではいろいろな微分積分が出てきます。すべての公式を覚えておく必要はありませんが、公式を適用して計算できるスキルを身につけておいてください。

指数関数の微分

$$\frac{d}{dx}e^x = e^x$$

マクローリン展開

$$f(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(0)}{k!} x^k, \quad \text{例: } e^x = 1 + x + \frac{x^2}{2} + \frac{x^3}{3!} + \frac{x^4}{4!} + \cdots$$

有理関数の微分

$$\frac{d}{dx} \frac{f(x)}{g(x)} = \frac{f'(x)g(x) - g'(x)f(x)}{(g(x))^2}$$

その例:

$$\frac{d^2}{dx^2} \frac{a}{a-x}, \quad \frac{d}{dx} \frac{px}{1-(1-p)x}$$

合成関数の微分

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

その例: ($\exp(x)$ は e^x と同じ)

$$\frac{d}{dx} \exp\left(-\frac{(x-m)^2}{2}\right)$$

部分積分

$$\int_a^b f(x)g(x)dx = [F(x)g(x)]_a^b - \int_a^b F(x)g'(x)dx, \quad \text{ただし、} F(x) = \int f(x)dx$$

その例:

$$\int_0^{\infty} x e^{-ax} dx, \quad \int_0^{\infty} x e^{-ax^2} dx, \quad \int_0^1 x^3(1-x)^2 dx$$

多重積分

$$\int_0^1 \left(\int_0^x (x+y)dy \right) dx$$