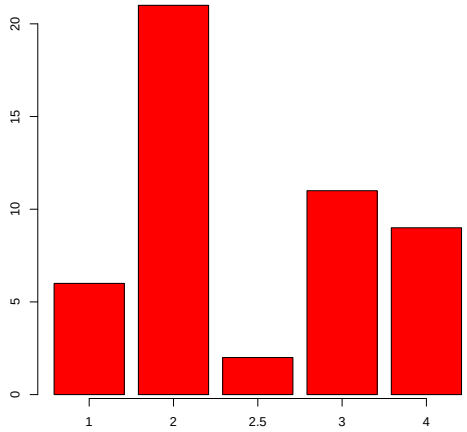
統計学　第２回　「統計的な考え方の基礎＝確率と確率分布」

前回のアンケートの集計結果

右の図は項目Ａの集計結果を示す度数分布図である。横軸は数学のできる程度（１：数式はまったく駄目，２：四則演算の変形程度はＯＫ，３：対数などもＯＫ，４：微積分もＯＫ），縦軸はそれに該当する回答数を示している。L1Q.txt という名前のタブ区切りのテキストファイルに回答を入力し，R を使って，

```
> x< −read.delim("L1Q.txt")
> barplot(table(x$QA))
```

として作成した（table は値ごとの頻度を集計する R の関数）。



約９割の人が四則演算の変形程度はＯＫ（というとき，暗黙のうちに回答１と回答２以上で再カテゴリ化をし，割合を求めていることに注意）なので，若干の式を示しながら講義を進める。

項目Ｂ（この講義に期待すること）としては，全回答者 49 人のうち，統計的なものの考え方を知りたい人が 33 人，応用事例を知りたい人が 16 人，統計解析手法を身につけたい人が 44 人だったので，解析手法の系統的な説明[1] に重点をおきながら，その背景となる考え方も適宜説明することにする。

今回の講義テーマ

前回，統計学とは「不確実性を考慮した論理的推論である」と述べた。不確実性とは，0%でも 100%でもないファジーな確率をもっているということである。

しかし，ここで簡単に「確率」と言ってしまったが，さて，改めて確率とは何かと訊かれたら，答えに詰まってしまうのではないだろうか。

そこで，今回は，あらゆる統計的な考え方の基礎となる「確率」というものを徹底的に考えてみることにする。高校数学の復習になってしまうかもしれないが，頭を整理しておくという意味で，役に立つのではないだろうか。なお，本日のテーマについて，もっと詳しく知りたい人には，伏見正則 (1987)「理工学者が書いた数学の本　確率と確率過程」（講談社，本体 2000 円，ISBN4-06-186837-3）をお薦めする。

確率的な現象を統計的事象と呼ぶ

▼ どういう現象が確率的か？

- サイコロを振ったときの目：振ってみるまでは１から６のどれが出るかはわからない。どの目がでる可能性も等しいから。
- 天気予報：「明日の天気予報は晴れ」といっても「必ず晴れる」とは限らない。「曇ったり雨が降ったりする可能性も少しはあるが，晴れる可能性が高い」ことを意味する。
- 喫煙と肺がんの関係：「タバコを吸うと肺がんになる」という命題は，「タバコを吸った人と吸わなかった人を比べて，肺がんになった人の割合が吸った人の方で高い」という関係を示す。タバコを吸っても肺がんにならない人もいるし，吸わなくても肺がんになる人もいる。

▼ こういう「不確かさ」に潜む法則性（長期間繰り返し観察したり，大集団で観察すると見られる）を考える学問を確率論と呼ぶ。大雑把に言えば，この種の法則性をもつ現象を，「統計的事象」と呼び，

[1] 第３回は尺度の説明とデータの図示のいろいろ，第４回は代表値のいろいろ，第５回から何回かはカテゴリ変数の分析（母比率に関する検定と推定，クロス集計表の解析でカイ二乗検定，フィッシャーの正確な確率，オッズ比，リスク比など），その後量的変数の分析（平均値の差の検定，分散分析，相関係数のいろいろ，回帰分析など），最後に余裕があれば多変量解析（共分散分析，ロジスティック回帰分析，重回帰分析，主成分分析など）を説明する。

統計的事象の確かさの度合いを示すのに便利なモノサシが「確率」である。そこで,「確率」をきちんと定義してみることにする。その前に，いくつかの準備が必要である。

準備その１：「標本空間」

▼ 統計的事象を捉えるには,「どんなことが起こりうるか」という範囲を定めることが必要である。

▼ 現象は一般に多面的で，様々な観察方法がある。以下３点によって統計的現象を捉えた，記号化された結果の集合のことを「標本空間」と呼ぶ。

- ＊ 観察を行う面を特定する
- ＊ 起こりうる結果の範囲を規定する
- ＊ その範囲内の各結果に記号を対応させる

▼ 個々の結果の起こりうる可能性を示す数値（これを「確率」と呼ぶ）を考える。一般には「どの結果も同程度に起こる」と考える。各結果に対応付けられた確率は０から１までの数値であり，各確率の値の総和は１にならねばならない。

▼ サイコロの目では，標本空間は｛1，2，3，4，5，6｝

準備その２：「事象」

▼ 問題は，個々の結果の可能性よりも，いくつかの結果が複合された集合（これを「事象」と呼ぶ）の起こる可能性がどのくらいか，ということである。つまり,「事象」＝「標本空間の部分集合」である。

▼ サイコロの例では,「目が偶数（丁）」とか「目が５以上」とか「目が１」とかいうことが事象。

▼ ある事象の確率は，その事象に含まれる各結果の生起確率の和である。従って，各結果の生起確率が等しい場合は，その事象に含まれる結果の場合の数をすべての場合の数で割ると，その事象の確率になる。サイコロの例では,「目が５以上」という事象の確率は，2/6=0.333... である。

準備その３：余事象・和事象・積事象・排反事象

▼ 起こりうるすべての結果の集合を「全事象」という。つまり，全事象は標本空間に等しい。

▼ 決して起こらない事象を「空事象」といい，空集合 ϕ で表す。

▼ 事象 E に対して，E が起こらないという事象を E の「余事象」という。E の余事象を $\bar{E}$ と書く。サイコロの例では,「目が偶数」という事象の余事象は「目が奇数」である。$Pr(E)+Pr(\bar{E})=1$ が常に成立する。

▼ 事象 E と F の少なくとも一方が起こるという事象を，E と F の「和事象」といい，$E \cup F$ で表す。

▼ 事象 E と F の両方が起こるという事象を，E と F の「積事象」といい，$E \cap F$ で表す。

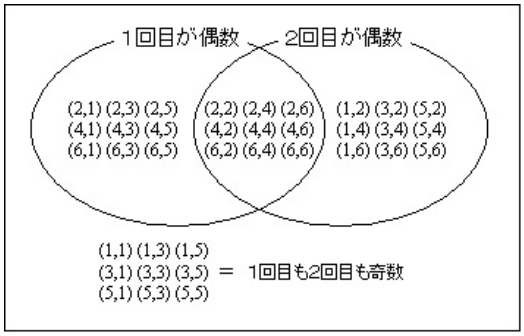
▼ 事象 E が起これば F は決して起こらないとき，E と F は「排反事象」であるという。E と F が排反事象なら， $E \cap F = \phi$ である。

準備その４：相互排反性と加法定理

▼ サイコロで考えると，１回振ったとき「偶数の目が出る」という事象 E が起こる確率（これを $Pr(E)$ という記号で書くことにする）は，2,4,6 の場合の数３を，1,2,3,4,5,6 の場合の数６で割った値なので $Pr(E)=0.5$。

▼ ２回振って「少なくとも１回は偶数の目」の確率は?

- ＊ 0.5+0.5=1.0 ではないのは自明。
- ＊ 『１回目に「偶数の目が出る」事象 E_1 と２回目に「偶数の目が出る」事象 E_2 とは排反ではない』ことに注意。
- ＊ 集合のベン図（右の図）から考えると，$Pr(E_1 \cup E_2) = Pr(E_1) + Pr(E_2) - Pr(E_1 \cap E_2)$ であることが直感的にわかる。この式を「加法法則」と呼ぶ。
- ＊ 「２回とも奇数」の余事象なので，$1 - Pr(\bar{E_1} \cap \bar{E_2}) = 1 - 9/36 = 0.75$ と考えてもよい。
- ＊ なお，事象 E と事象 F が排反なら，$Pr(E_1 \cap E_2) = 0$ なので，$Pr(E \cup F) = Pr(E) + Pr(F)$ という「加法定理」が成立する。



準備その５：事象の独立性と乗法定理

- ▼ 事象 E が起こったときに事象 F が起こる確率を,「E が起こったときの F の条件付き確率」といい, $Pr(F|E)$ と書く。
- ▼ 「E が起こった」うちの「E も F も起こった」場合なので，$Pr(F|E) = Pr(F \cap E)/Pr(E)$ である。
- ▼ E と F が互いに無関係（＝独立）なら， $Pr(F|E) = Pr(F)$。逆にいえば，$Pr(F) = Pr(F|E)$ のときに事象 E と事象 F は互いに独立であるという。独立でないとき「従属である」という。
- ▼ 上記２つの式から，E と F が独立なら， $Pr(F \cap E) = Pr(F) \times Pr(E)$ という”乗法定理”が成立する。

確率を定義するための４種類のアプローチ

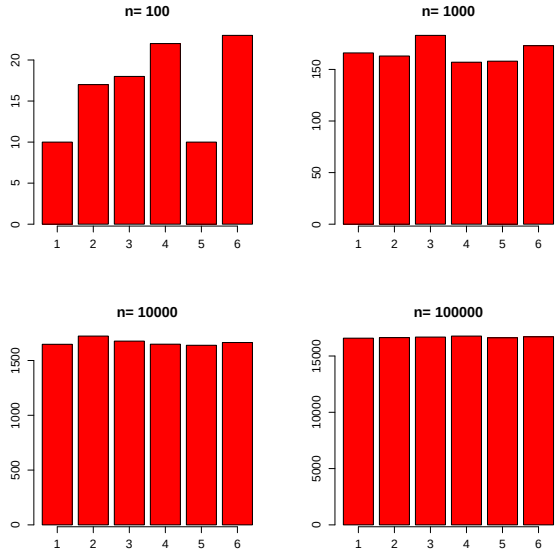
- ▼ 操作的アプローチ（統計的定義）: 数多く試したときの相対度数の極限。例えば，事象 E が起こる確率 $Pr(E)$ は，N 回試したときに N_1 回事象 E が起こるとして，N_1/N という相対度数が，N を無限大にしたときに漸近する値である。
- ▼ 対称的確率：サイコロの場合，６通りの目の出る確率はどれも等しくなければならず，その和は１でなくてはならないので，例えば１の目が出る確率は $1/6$ となる。限定的かつ循環論法。
- ▼ 公理的客観確率:標本空間の各要素を e_i として，$Pr(e_i) >= 0$ かつ $Pr(e_1)+Pr(e_2)+...+Pr(e_N) = 1$ かつ事象 E が起こる確率 $Pr(E) = \sum Pr(e_i)$ を公理とする[2)]。
- ▼ 主観確率：観念的にも二度と繰り返すことのできない事象についての「見込み」を扱う。決定理論において重要。

大数の法則（操作的接近の根拠）

- ▼ R のプログラムでは（runif は一様乱数を発生させる関数）,

```
> a <- c(100,1000,10000,100000)
> op <- par(mfrow=c(2,2))
> for (i in 1:4) {
> y <- as.integer(runif(a[i],1,7));
> s <- paste("n=",as.integer(a[i]))
> barplot(table(y),main=s)}
> par(op)
```

とすれば，右図のように，試行回数を増やすと，サイコロの特定の目が出る割合が，ある一定値に近づくことがわかる。「１の目が出る」事象 E_1 が起こる確率 $Pr(E_1) = p$ とおけば，N 回サイコロを振って N_1 回 1 の目が出たとして，任意の正の小さな数 ε に対して，$\lim_{N\to\infty} Pr(|N_1/N - p| < \varepsilon) = 1$ ということで，これをベルヌーイ (Bernoulli) の大数の法則という。



確率変数と期待値（スロットマシンの例による説明）

- ▼ スロットマシンでは，ごくたまに，投入金額の何十倍ものコインが出てくることがある。
- ▼ マシン利用者全員に返ってくる賞金の合計を利用回数で割った値が，１回に期待される賞金額で，これを賭け金で割った値を「賞金還元率」と呼ぶ。すべての賭け事で胴元が儲かるようになっているのは，賞金還元率が 100 %未満だからである。宝くじでは 40%，競馬では 75%と言われる。
- ▼ 一般に，賞金額が $x_1, x_2, x_3, ...$ で，その賞金が得られる確率が $p_1, p_2, p_3, ...$ のように設定されたスロットマシンの期待賞金額 M は，$M = x_1p_1 + x_2p_2 + x_3p_3 + ...$ で与えられる。

[2)] 要素は互いに排反であるということ。当然である。もっと厳密に書くには，確率空間というものを定義しなくてはならず，面倒なのでここでは省略する。詳しく知りたい場合は，伏見による前掲書を参照されたい。

▼ このスロットマシンのようなものを確率変数といい，期待賞金を期待値と呼ぶ（厳密には後述）。

分散（スロットマシンの例による説明）

▼ 期待賞金が同じでも，値動きの幅が小さいと一喜一憂の程度が小さく，逆に幅が大きいと滅多に当たらないが当たったときの喜びは大きくなる。つまり，ギャンブル性は，値動きの幅と，チャンスの大きさに依存している。

▼ 各賞金がどれくらい期待賞金から隔たりがあり，それを獲得できる可能性がどれくらいあるのかを見積もれば，ギャンブル性が表せる。

▼ マシンのギャンブル性を V とおけば，$V = \sum$（期待値からの隔たり）$\times$（可能性）という値が定義できる。この V を「分散」と呼ぶ[3]。なお，各賞金額 x と期待値 M の隔たりは，差を二乗した値 $D = (x - M)^2$ で表す。

確率変数を数式で書く。確率分布とは？

▼ 一般に，とりうる値の集合 $\mathbf{x} = (x_1, x_2, x_3, ...)$ と，それぞれの値が実現する確率 $\mathbf{p} = (p_1, p_2, p_3, ...)$ が与えられていて，事象として $\mathbf{x}$ のうちどれか１つの値のみ実現するとき，$(\mathbf{x}, \mathbf{p})$ という１セットを,「確率変数」と呼んで，X で表す。

▼ 期待値は $E(X) = \mu = \sum x_i p_i$

▼ 分散は $V(X) = \sigma^2 = \sum (x_i - \mu)^2 p_i$

▼ 分散の平方根 σ を標準偏差と呼ぶ。

▼ 横軸に $\mathbf{x}$ の各々の値を示す位置をとり，その各々に $\mathbf{p}$ の各々の可能性を示す高さの棒を立ててみれば，これが確率変数の「確率分布」ということになる。

ベルヌーイ試行と２項分布

▼ １回の実験で事象 S か事象 F のどちらかが起こり，しかもそれらが起こる可能性が，$Pr(S) = p, Pr(F) = 1 - p = q$ で何回実験しても変わらないとき，これを「ベルヌーイ試行」という。

▼ ベルヌーイ試行では，事象 F は事象 S の余事象になっている。

▼ 例えば，不透明な袋に黒い玉と白い玉が 500 個ずつ入っていて，そこから中を見ないで１つの玉を取り出して色を記録して（事象 S は「玉の色が黒」，事象 F は「玉の色が白」）袋に戻す実験はベルヌーイ試行である（注：袋に戻さないと１回実験するごとに事象の生起確率が変わっていくのでベルヌーイ試行にならない）。

▼ ベルヌーイ試行を n 回行って，S がちょうど k 回起こる確率は，$Pr(X = k) =_n C_k p^k q^{n-k}$。

▼ ${}_nC_k$ は言うまでもなく n 個のものから k 個を取り出す組み合わせの数である。２項係数と呼ばれる。このような確率変数 X は,「２項分布に従う」といい，$X \sim B(n, p)$ と表す。$E(X) = np, V(X) = npq$ である。

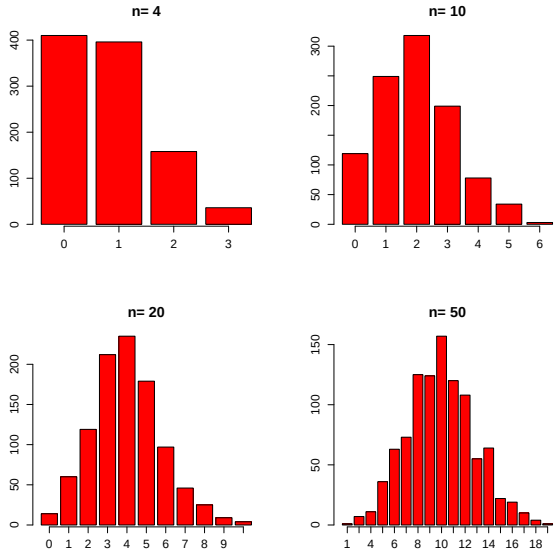
２項分布のシミュレーション

▼ 正二十面体（各面には１から 20 までの数字が割り振られている）サイコロを n 回 $(n = 4, 10, 20, 50)$ 投げたときの，１から４までの目が出る回数を１試行と考えれば，これはベルヌーイ試行である。１回投げたときに１から４までの目が出る確率は 0.2 であるとして（＝母比率を 0.2 とする），試行 1000 セットの度数分布を図に示す。R のプログラムは下記の通り。

---

[3] 分散とはギャンブル性である，というこの説明は，鈴木義一郎「情報量基準による統計解析入門」（講談社サイエンティフィク）から引いてきたものだが，うまい解釈だと思う。

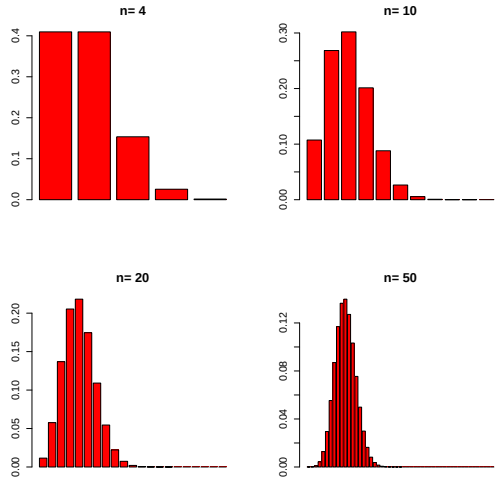
```
> times <- function(n) {
> hit <- 0
> dice <- as.integer(runif(n,1,21))
> for (j in 1:n) {
> if (dice[j]<5) {hit <- hit+1}}
> return(hit)}
>
> a <- c(4,10,20,50)
> op <- par(mfrow=c(2,2))
> for (i in 1:4) { nx<-a[i]; y<-c(1:1000)
> for (k in 1:1000) { y[k]<-times(nx) }
> barplot(table(y),main=paste("n=",nx)}}
> par(op)
```



2項分布の理論分布

▼ この例で，各nについての理論的な確率分布は，$Pr(X = k) =_n C_k 0.2^k 0.8^{n-k}$ より右図のようになる。R のプログラムは下記の通り。

```
> a <- c(4,10,20,50)
> op <- par(mfrow=c(2,2))
> for (i in 1:4) { n <- a[i]
> k <- 0; chk <- c(1:n+1)
> while (k <= n) {
> chk[k+1] <- choose(n,k)*(0.2^k)*(0.8^(n-k))
> k <- k+1 }
> barplot(chk,col='red',main=paste("n=",n))}
> par(op)
```



▼ ただし，R には様々な確率分布についての関数があり，`choose(n,k)*(0.2^k)*(0.8^(n-k))` は `dbinom(k,n,0.2)` と同値である。このように，確率変数が取りうる各値に対して，その値をとる確率を与える関数を確率密度関数という。値が小さいほうからそれを全部足した値を与える関数（つまり，その確率変数の標本空間の下限から各値までの確率密度関数の定積分）を分布関数（あるいは確率母関数，累積確率密度関数）と呼ぶ。

正規分布

▼ n が非常に大きい場合は，2項分布 $B(n,p)$ の確率 $Pr(X = np + d)$ という値が，

$$\frac{1}{\sqrt{2\pi npq}} \exp\left(-\frac{d^2}{2npq}\right)$$

で近似できる。

▼ 一般にこの極限（n を無限大にした場合）である，

$$Pr(X = x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

という形をもつ確率分布を正規分布と呼び，$N(\mu, \sigma^2)$ と書く。

▼ $z = (x - \mu)/\sigma$ と置けば，

$$Pr(Z = z) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-z^2}{2}\right)$$

となる。これを標準正規分布と呼び，$N(0, 1)$ と書く。

▼ 統計学でよく使われる確率分布であるカイ二乗分布とか t 分布とか F 分布は，正規分布から導かれる。

参考：よく使われる確率密度関数，分布関数，分位点関数についてRでの表現の一覧

分布の種類	確率密度関数 (probability density function)	分布関数＝確率母関数＝累積確率密度関数 (distribution function = probability generating function cumulative probability density function)	分位点関数 (quartile function)
カイ二乗分布	dchisq(カイ二乗値, 自由度)	pchisq(カイ二乗値, 自由度)	qchisq(%, 自由度)
２項分布	dbinom(生起回数, 試行回数, 母比率)	pbinom(生起回数, 試行回数, 母比率)	qbinom(%, 試行回数, 母比率)
ポアソン分布	dpois(生起回数, 期待値)	ppois(生起回数, 期待値)	qpois(%, 期待値)
正規分布 (1)	dnorm(Ｚスコア，平均値，標準偏差)	pnorm(Ｚスコア，平均値，標準偏差)	qnorm(%, 平均値，標準偏差)
対数正規分布 (2)	dlnorm(Ｚスコア, 対数平均値，対数標準偏差)	plnorm(Ｚスコア，対数平均値 ，対数標準偏差)	qlnorm(%, 対数平均値 ，対数標準偏差)
一様分布 (3)	dunif(値，最小値，最大値)	punif(値，最小値，最大値)	qunif(%, 最小値，最大値)
t 分布	dt(t 値，自由度)	pt(t 値，自由度)	qt(%, 自由度)
F 分布	df(F 値，第１自由度，第２自由度)	pf(F 値，第１自由度，第２自由度)	qf(%, 第１自由度，第２自由度)

(1) 平均値と標準偏差は省略可能。省略時は標準正規分布 (平均 0, 標準偏差 1) になる。

(2) 対数平均値と対数標準偏差は省略可能。省略時は対数平均 0, 対数標準偏差 1 になる。なお，対数平均とは自然対数をとった値の平均，対数標準偏差とは自然対数をとった値の標準偏差をいう。dlnorm(1) は dnorm(0) と等しい。

(3) 閉区間である。省略時は 0 と 1 になる。

（注）これらの分布関数に従う乱数を生成する関数もある。例えば，0 から 1 までの一様乱数を 1000 個生成する関数は runif(1000,0,1) である。試行回数 100 回，母比率 0.2 の２項分布に従う乱数を 1000 個発生させるには，rbinom(1000,100,0.2) とすれば良い。

練習問題

▼ ８頭で出走する競馬のレースがあり,「どの馬が勝つチャンスも等しい」と仮定した場合，ある特定の馬が勝つと予想して当たる確率は 1/8 となるが，２回のレースの少なくともどちらか一方に当たる確率はいくらか？