統計学　阿部真人

・ポイント１

本当に興味があるのは、小皿のお味噌汁ではなく、鍋のお味噌汁である。

→本当に興味があるのは、標本のデータではなく、母集団である。

ここでの目的は、お鍋のお味噌汁の味を知ることです。いくら小皿のお味噌汁の味を味わっても、お鍋の味を評価しないのであれば意味がありません。

・ポイント２

鍋のお味噌汁を飲み干して調べ尽くすことは困難である。

→母集団の全ての要素を調べ尽くす全数調査は困難である。

お鍋のお味噌汁の味を知ることが目的ですが、お鍋のお味噌汁をすべて飲み干して味を確認することは困難です。

・ポイント３

小皿程度の少ない量のお味噌汁で、鍋のお味噌汁の味を「ほぼ」確認できる。

→小さいサンプルサイズの標本から母集団を推測できる。

味見する量はなぜ小皿程度なのでしょうか。これより少ない、例えば１滴程度であれば、うまく母集団のお味噌汁の味を反映していないことが考えられます（少ないと舌が感覚器として十分に反応しないという問題もありますが、ここでは置いておきます)。

一方、お椀一杯分で味見することに意味はあるでしょうか。お椀一杯の方が、より正確に鍋のお味噌汁の味を反映すると考えられますが、そこまでしなくても、小皿程度の量で、鍋のお味噌汁の味を十分正確に表していると考えられます。

・ポイント４

お玉でお味噌汁をすくうときに、よくかき混ぜること。

→標本を抽出するときには、ランダムサンプリングする必要がある。

・n-1 で割る

一方、第2章で紹介した標本標準偏差 s（または標本分散 s^2）は少し事情が変わってきます。標本標準偏差 s の定義におけるルートの中の分母は n でした。記述統計として、データのばらつき度合いを評価する場合には問題ありません。しかし、これは母集団の標準偏差 σ を過小評価してしまうという問題があります。正しくは、$n-1$ で割った次の式が、母集団の標準偏差 σ の不偏推定量になります。

$$s=\sqrt{s^2}=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2} \qquad \text{(式4.1)}$$

これを、不偏標準偏差（s^2 を不偏分散）と言います。

以下、n で割った場合に、なぜ過小評価になるのかの直感的な理由を説明します。

各値 x_i を標本平均 $\bar{x}$ との差を取って二乗し、ばらつき度合いを測っていましたが、本来、$(x_i-\mu)^2$ として計算したいところを、μ は未知なので、$(x_i-\bar{x})^2$ で代用しています。そして、図4.2.3や図4.2.4で見たように、$\bar{x}$ は μ と一致せず、各値 x_i と μ の位置関係、および各値 x_i と $\bar{x}$ の位置関係を考えると、x_i は μ よりも $\bar{x}$ に近いところにあるのです。そのため、$(x_i-\bar{x})^2$ の和は $(x_i-\mu)^2$ よりも小さめの値になります。そこで、n で割るのではなく、$n-1$ で割ることで、過小評価を補正しているのです。

実際のデータ分析では、さほど意識する必要はありませんが、最初に統計学を学んだ時に、なぜ不偏標準偏差では $n-1$ で割るのかは腑に落ちないところだと思いますので、簡単に説明しました[4)]。

仮説検定では、p 値（p-value）という数値を計算し、仮説を支持するかどうかを判断します。

p 値は科学論文で頻繁に用いられますが、**定義がわかりにくい**ため、きちんと理解せずに使っている方が多く、科学業界における様々な問題を引き起こす原因となっています。

論文などへ記載する際には、「統計的に有意な差は見られなかった(p=0.246)」のように記述します。くり返しますが、**有意な差が見られないことは**、帰無仮説を支持するのではなく、帰無仮説と対立仮説のどちらを支持することもできず、**結論を保留するという判断**であることに注意してください。

図や表では、統計的に有意であることを示すために、＊（アスタリスク）をつけることが一般的です

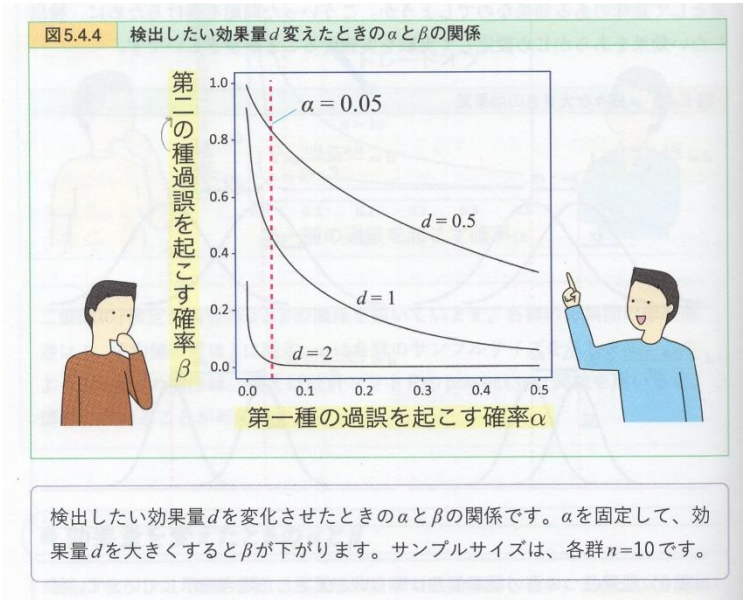
・・・

よく使われる記法は、＊：p＜0.05，＊＊：p＜0.01，＊＊＊：p＜0.001 です。つまり、0.01≦p＜0．05 の場合はアスタリスク 1 つ、0.001≦p＜0.01 の場合はアスタリスク 2 つ、p＜0.001 の場合は、アスタリスク 3 つということです。また、有意ではないことを示すために、N.S.（non－significant）と書くことがあります。

図5.4.1　判断と真実のパターン

判断2パターン	真実2パターン：帰無仮説が正しい	真実2パターン：対立仮説が正しい
帰無仮説を棄却しない	OK	第二種の過誤確率β
帰無仮説を棄却して対立仮説を採択	第一種の過誤確率α	OK

　また、α、β、サンプルサイズ n、効果量 d の 4 つの数値のうち、3 つを決めると**残り 1 つは自動的に決まる値**であるという性質があるので、$\alpha=0.05$，$1-\beta=0.8$ と、検出したい効果量 d をあらかじめ設定することで、仮説検定に必要なサンプルサイズ n を求めることができます。



◆正規性を調べる

　パラメトリック検定では、各群のデータに正規性が必要です。正規性を調べる手法には、視覚的な判断として調べる Q-Q プロットや、仮説検定として調べるシャピロ・ウイルク検定、分布一般の比較に使われるコルモゴロフ・スミノフ検定などがあります。

　よく使われるシャピロ・ウイルク検定の詳細な仕組みは、本書のレベルを超えるので述べませんが、この場合、次の前提で仮説検定を実行します。

・帰無仮説「母集団の分布は正規分布である」
・対立仮説「母集団の分布は正規分布ではない」

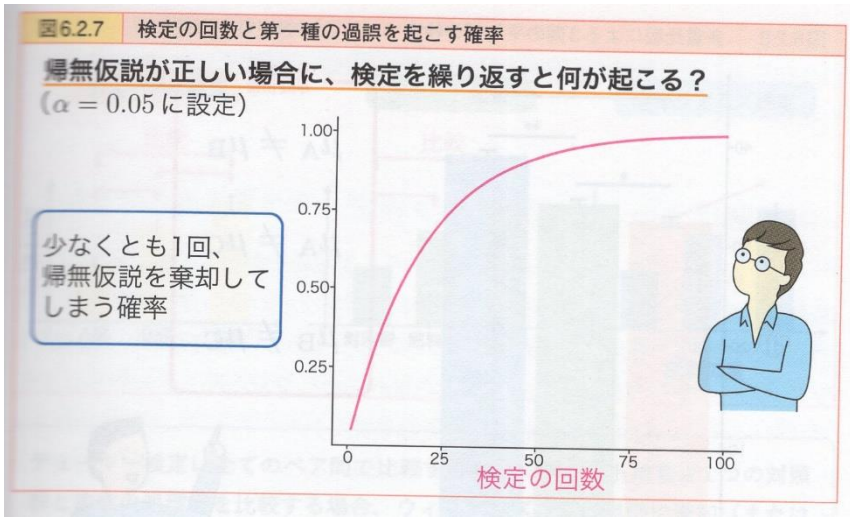
◆等分散性を調べる

　t 検定、それから次に解説する分散分析では、分散が等しい母集団から得られたという条件が必要です。分散が等しいという仮説を調べる検定には、パートレット検定やルビーン検定があります。この場合、この場合、次の前提で仮説検定を実行します。

・帰無仮説「全ての群に対し母集団の分散が等しい」
・対立仮説「少なくとも 1 つの群間に対し母集団の分散が等しくない」

◆多重比較

三群以上ある場合、各ペアの差を調べるために、有意水準α＝0.05 で二標本の t 検定を繰り返し実行してしまうと、第一種の過誤が増加してしまうという問題があります。例えば、三群ある場合には、ペアの数は 3 です。t 検定を 3 回繰り返すと、少なくとも 1 つのペアで第一種の過誤を起こす確率は、1－（どのペアも第一種の過誤を起こさない確率）^{3}＝1－（1-0.95）^{3}＝0.143 となり、分散分析で全体として設定していたα＝0.05 を上回ってしまいます。n 群ある場合には、$_nC_2$ のペアがあるため、群の数が増えるほど、第一種の過誤が起きやすくなってしまいます。つまり、**何度も検定を繰り返すことで、本当は差がないにもかかわらず、差があると言ってしまう誤りが簡単に起きてしまいます**（図 6．2．7）。



仮説検定におけるこの問題は、科学業界においても重大な影響をもたらす可能性があります。これについては、第 9 章で詳しく解説します。

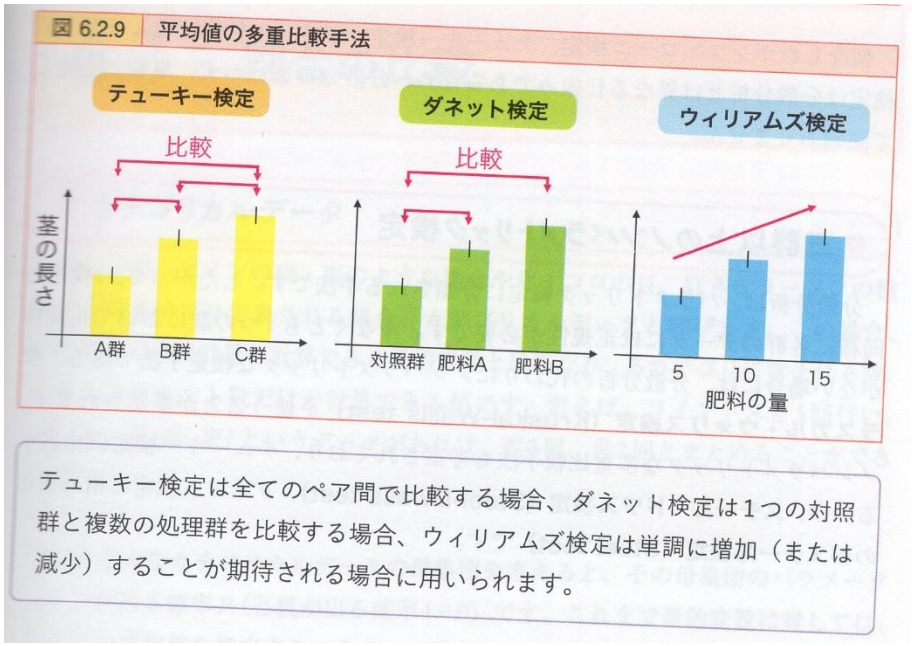
なお、この多重性の問題を回避する 1 つの方法は、多重比較の検定を用いることです。基本的なアイデアは、**検定を繰り返す分だけ有意水準を厳しい値に変更する**というものです。それによって、トータルで、有意水準α＝0.05 を保つようにします。

◆様々な多重比較手法

最も単純な多重比較手法として、**ボンフェローニ法**があります。これは、全体で有意水準αを設定しているのであれば、検定を繰り返す数を k とすると、1 回 1 回の検定では、αを検定の回数で割ったα/k を基準に仮説検定を行う手法です。つまり、各検定では、p＜α/k であれば、対立仮説が採択されるということになります。

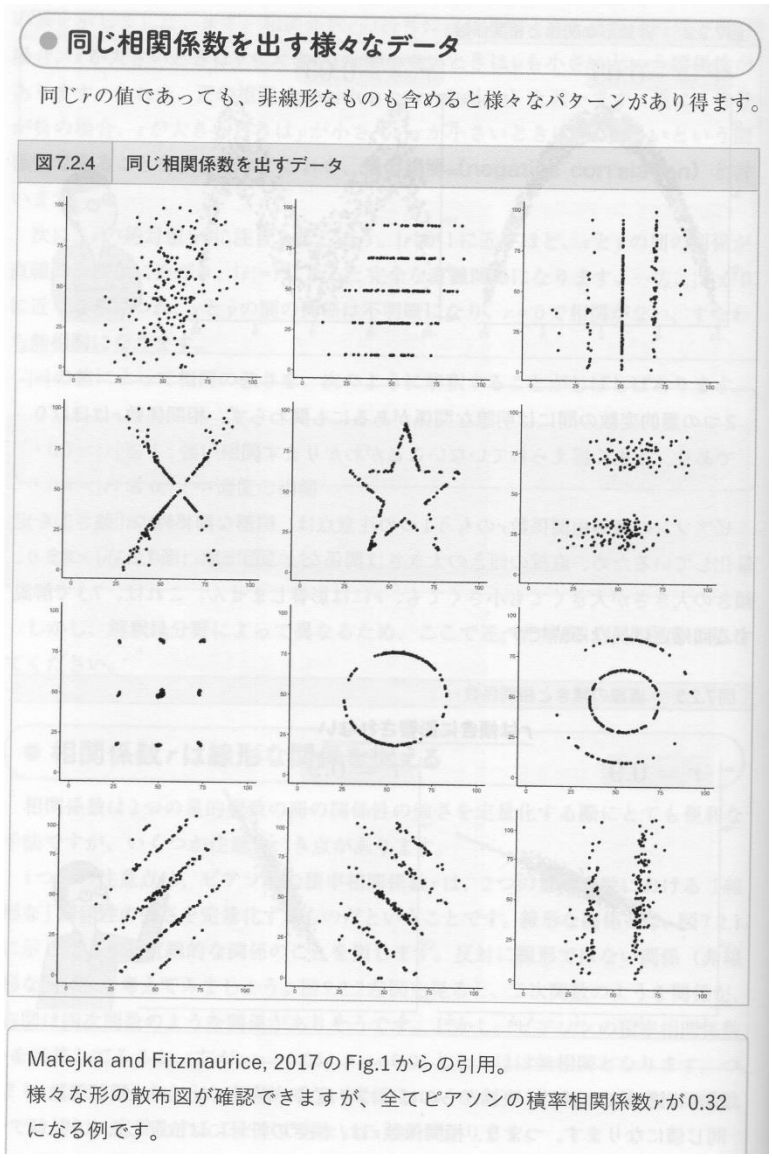
平均値の比較でボンフェローニ法を使うとすれば、t 検定を各ペアで実行し、p＜α/k で差があると判断します。ボンフェローニ法は非常に簡便で、平均値の比較だけでなくあらゆる多重比較において実行可能である利点があります。

一方、検出力が低い傾向があるため、本当は差があるときに差があると主張しにくくなるという欠点があります。通常、分散分析を行ったあとは、ボンフェローニ法よりも優れた（検出力が改善された）手法である、テューキーの検定（Tukey's test）等の他の手法が使われることが多いです。



テューキー検定は全てのペア間で比較する場合、ダネット検定は1つの対照群と複数の処理群を比較する場合、ウィリアムズ検定は単調に増加（または減少）することが期待される場合に用いられます。

r=0.32



Matejka and Fitzmaurice, 2017 の Fig.1 からの引用。
様々な形の散布図が確認できますが、全てピアソンの積率相関係数rが0.32になる例です。

ノンパラメトリックな相関係数
　スピアマンの順位相関係数

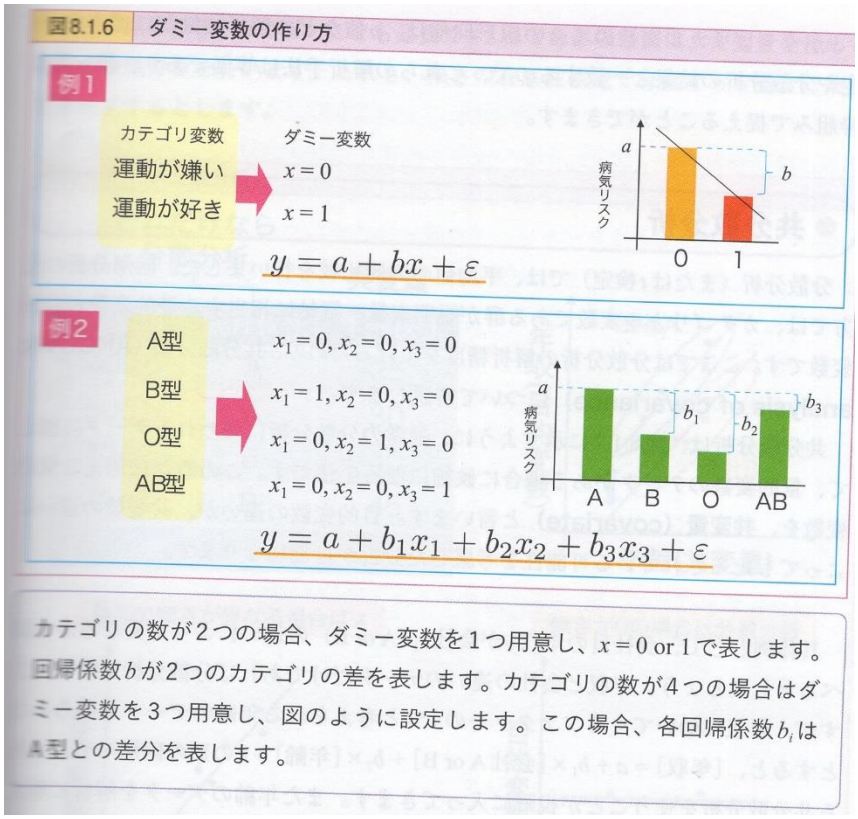
決定係数

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - f(x_i)\right)^2}{\sum_{j=1}^{n}\left(y_j - \bar{y}\right)^2}$$

右辺第二項は、説明されずに残っている残差の割合を表します。これを 1 から引き算することで、決定係数 R2 はデータによって説明された割合を示します。

　説明変数が 1 つの一次関数の線形回帰で最小二乗法を用いた場合、決定係数 R2 は x と y の間のピアソンの積率相関係数 r を二乗した値に一致します。そのため、相関係数とデータの散らばり具合の関係を感覚として知っておくと、決定係数 R2 からどの程度当てはまっているかが大まかにわかります。しかし、決定係数 R2 は**説明変数の数が増える**（例えば y=a+b1x1+b2x2）**と大きくなる**という性質があるため、意味のない説明変数を導入することで見かけ上、回帰モデルの説明力が上がって見えてしまうことがあります。そのため、説明変数の数 k こ応じて調整された、**調整済み決定係数** R'2（A 句舶 R－Squared）を用いるのが一般的です。

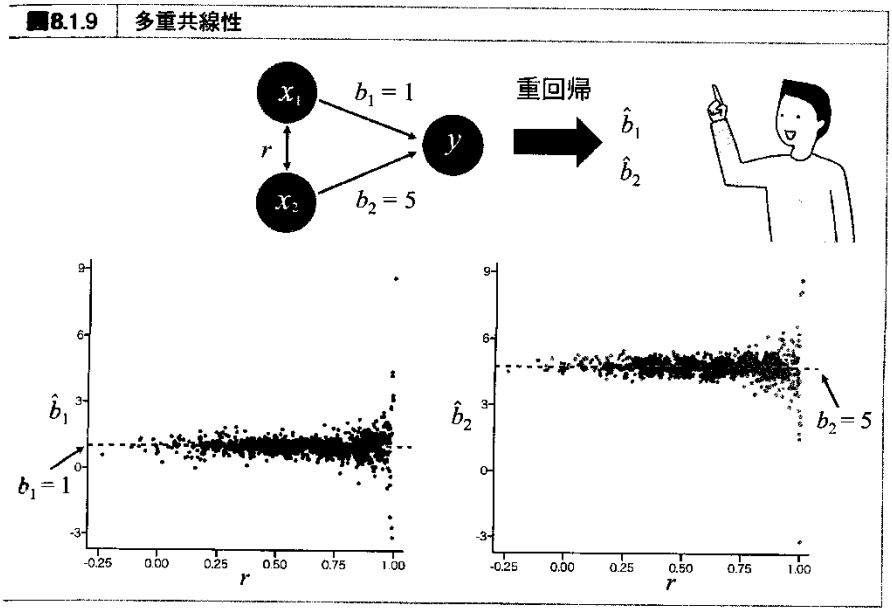
◆ダミー変数の作り方



次元の呪い
回帰モデルの説明変数の数（次元）が多くなると、推定に必要なデータ数が爆発的に増えてしまう。

◆多重共線性
　説明変数が複数個ある重回帰において、説明変数同士が強く相関している場合、多重共線性（multicollinearity）があると言います。そして多重共線性がある場合、回帰係数の推定の誤差が大きくなる問題が起こる可能性があります。つまり、推定値の信頼性が落ちるということです。
　y＝x1+5x2+ ε という重回帰モデルの例を見てみましょう（図 8.1.9）。この場合、

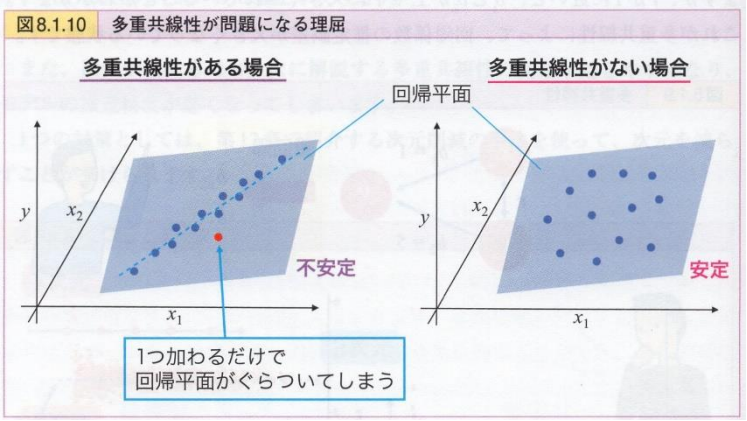
母集団のパラメータは b1＝1、b2＝5 です。x1 と x2 の相関係数 r を色々と変えて、乱数を発生させて仮想のデータを作り、重回帰分析して推定した b1 ハットと b2 ハットの値を観察してみます。r の広い範囲で、b1 ハットと b2 ハットは正解である b1＝1、b2＝5 付近に分布していますが、r が 1 に近いと、b1 ハットと b2 ハットが上や下に大きくぶれていることがわかります。これが多重共線性によって、回帰係数の推定誤差が大きくなっている状態です。



様々なrに対し、サンプルサイズn=20で、$y = x_1 + 5x_2 + \varepsilon$からデータを発生させ、パラメータを推定しました。グラフ中の青い点線が、パラメータの正解の値です。rが1に近いところで、大きく上や下にぶれており、推定の精度が急に悪くなることがわかります。

◆多重共線性の直感的理解

なぜ、このような問題が起こるかを直感的に理解するために、図8.1.10を見てみましょう。2つの変数x_1, x_2の間に相関があると、2つの変数は直線的な関係にあるので、y, x_1, x_2の空間の中では、データが直線的な領域に分布することになります（図8.1.10左）。



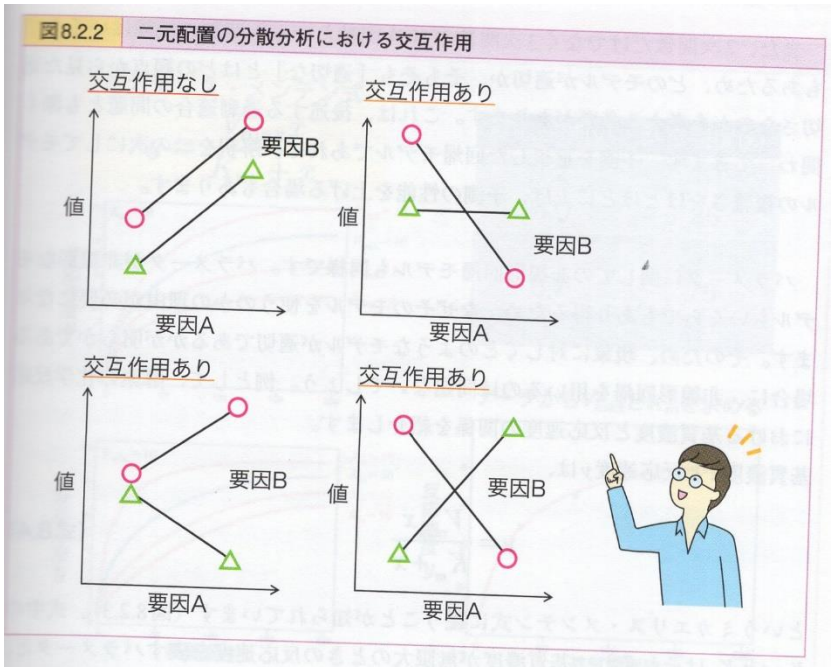
◆交互作用

これまで、線形回帰のモデルは説明変数の数をk個として、

$$y = a + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k + \varepsilon$$

で表されるモデルを扱ってきました。このモデルは、ある説明変数x_iは他の説明変数とは独立に、1増えるごとにb_1だけ目的変数が変化することを仮定しています。しかし、現実のデータにおいては、x_iが1増えたときのyの増え方が、他の説明変数に影響される場合も考えられます。このような説明変数間同士の相乗効果を**交互作用**と言い、線形回帰モデルの中に掛け算$cx_i x_j$として導入することができます。

交互作用がなければ、２つの傾きは平行になる。



一般化線形モデル

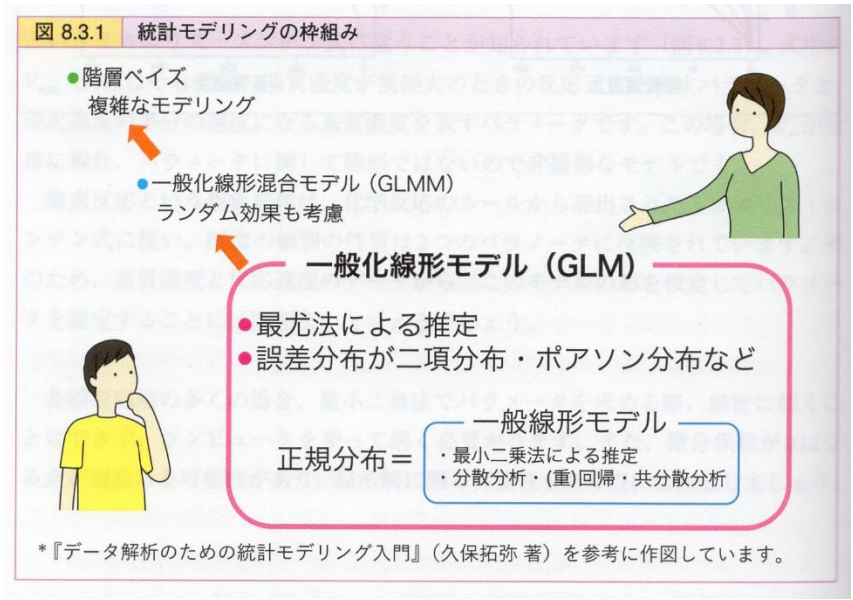
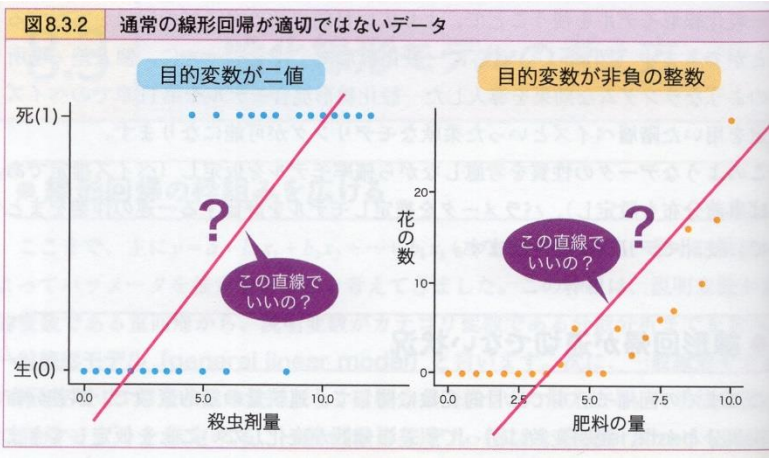


図8.3.2　通常の線形回帰が適切ではないデータ



● 尤度と最尤法

最小二乗法では、回帰モデルとデータの間の差（残差）の二乗を計算し、最小化することでパラメータを求めました。最小二乗法における、回帰モデルのデータへの当てはまりの良さは、回帰モデルとデータの距離（残差）に基づいています。しかし、前述したような二値変数である目的変数のデータや、非負の整数である目的変数といったデータでは、ほとんど同じ値を取る領域（例えば、図8.3.2左の説明変数が小さいまたは大きい領域）や、値が大きくばらつく領域があったりします（例えば、図8.3.2右の説明変数が大きい領域）。

そのため、距離をもとにモデルとデータの適合度合いを測るよりも、データが確率的に生成されたことを考え、確率的な起こりやすさをもとにデータへの当てはまりの良さを評価すると良さそうです。

確率分布はパラメータを決めると形状が決まり、どの値がどの程度起きやすいか（確率）を表しました。手元にあるデータをx $(x_1, x_2, \ldots, x_n)$、確率分布のパラメータをθと表記すると、データxがあるパラメータθの確率分布からどの程度起きやすいかは、次のように書くことができます。

$$P(x|\theta) = P(x_1|\theta) \times P(x_2|\theta) \times \cdots \times P(x_n|\theta)$$

$P(x|\theta)$ を、データxに対してパラメータθの関数として見たものを、**尤度（ゆうど, likelihood）** と言います。すなわち、

$$L(\theta|x) = P(x|\theta)$$

です。手元のデータxに対しθを変えることで、尤度が変化するイメージです。尤度が大きいことは、そのθで手元のデータが起きやすいことを表しています。そこで、θを様々変えて尤度を最大化するθを見つけ、推定値とすれば、手元のデータへ最も当てはまるパラメーターθを得ることができます。

この手法を、**最尤法（maximum likelihood method）または、最尤推定（maximum likelihood estimation）** と言います。通常、尤度の対数を取って対数尤度$\log L(\theta|x)$ として計算することが多いです。対数を取っただけなので、尤度を最大化するθと対数尤度を最大化するθは同じ値になります。

一般化線形モデルは、目的変数の誤差の確率分布を指定し、尤度にもとづいてパラメータを推定する回帰と言えます。ただし、線形回帰で通常仮定される正規分布の誤差を持つモデルに対し、最小二乗法を用いてパラメータを推定した結果と、最尤法を用いて推定した結果は一致します。その意味で、一般化線形モデルはこれまでの正規分布の誤差を仮定した最小二乗法を包含し、さらに正規分布以外の確率分布を用いることができるという点で自然な一般化になっています。

◆ロジスティック関数

図8.3.4の例は、様々な濃度で殺虫剤を虫に与えて、生死を調べた実験データです。この場合、目的変数yは二値のカテゴリ変数（生:0, 死:1とします）、または上限のあるカウント数（$0, 1, \ldots, N$）であるため、ロジスティック回帰を用いてみましょう。殺虫剤濃度x（$x_1, x_2, \ldots, x_n$）と虫の各個体の生死データy（$y_1, y_2, \ldots, y_n$）に対し、回帰モデルは殺虫剤濃度x_iから個体の死亡率p_i（$0 \leqq p_i \leqq 1$）を関係づけて、そのp_iを持つ二項分布$B(N, p_i)$から各データy_iが発生したと考えて、

$$y_i \sim B(N, p_i) \qquad \text{(式8.5)}$$

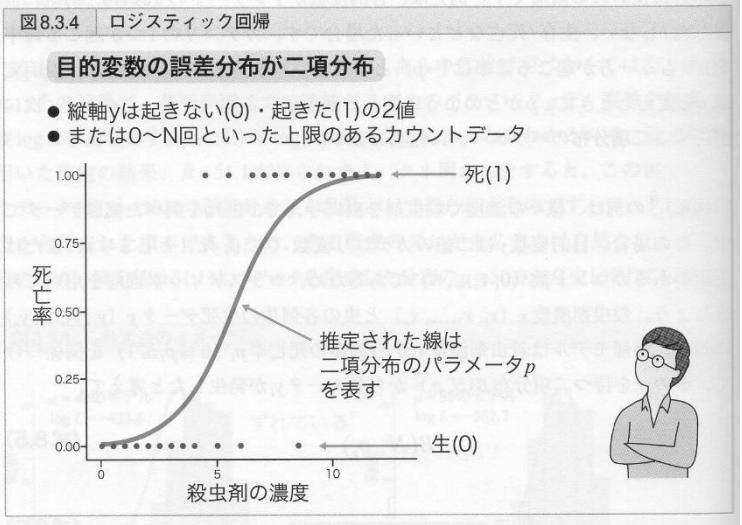
$$p_i = \frac{1}{1+\exp\left(-(a+bx_i)\right)} \qquad \text{(式8.6)}$$

となります（〜は、左の確率変数が右の分布に従うことを表し、$\exp(x)$は自然対数eのx乗を表します）。

式8.6を見ると、$a+bx$という最も単純な一次式がexpの中に含まれることがわかります。ロジスティック回帰を用いた分析では、データからパラメータa, bを推定し、起こる（この場合、死亡する）確率pのモデルを構築します。これまでの回帰と同様に、パラメータbが0であれば、説明変数xが変わっても目的変数が変わらないことになります。

図8.3.4の例に対し、最尤推定でパラメータを推定すると、$\hat{a} = -4.82$, $\hat{b} = 0.84$が得られます。図中の線が式8.6に推定されたパラメータを代入した回帰モデルで、殺虫剤濃度を上げると死亡率が増加していくことがわかります。

図8.3.4　ロジスティック回帰



式8.6を式変形すると次のようになり、

$$a+bx=\log\frac{p}{1-p} \qquad \text{(式8.7)}$$

この右辺を**ロジット関数（logit function）**と言います。

一般化線形モデルの枠組みでは、左辺の$a+bx$の部分を**線形予測子**と言い、線形予測子と目的変数の確率分布のパラメータ（二項分布ではp）をつなぐ関数を、**リンク関数**と言います。すなわち、ロジスティック回帰のリンク関数はロジット関数ということになります。リンク関数を変更することで、目的変数の誤差の確率分布（誤差構造）を指定し、柔軟なモデリングが可能になります。この線形予測子とリンク関数が、一般化線形モデルの主要部分です。

図8.3.6　ロジスティック関数

ロジスティック関数

$$p=\frac{1}{1+\exp(-(a+bx))}$$

- pは0以上1以下
- aは横に平行移動のパラメータ
- bは変化の度合いを決めるパラメータ

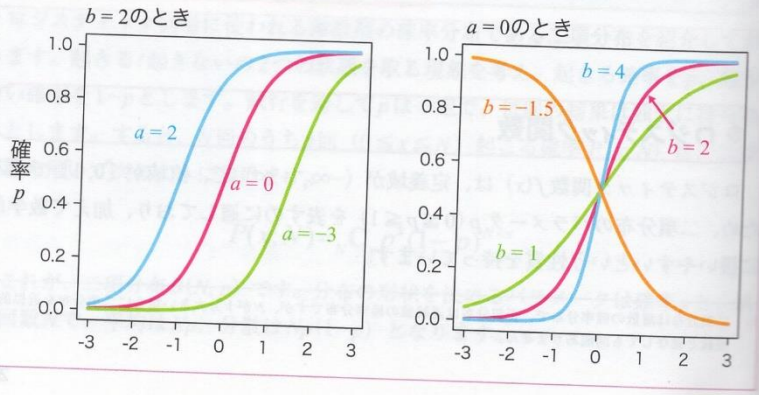


図8.3.7　オッズ比

オッズ(odds)

確率pで起きる出来事に対し$\frac{p}{1-p}$をオッズという

例　$p=0.5$でオッズ1
　　$p<0.5$でオッズ<1
　　$p>0.5$でオッズ>1

オッズ比(odds ratio, OR)

確率p, qに対し

$$\mathrm{OR}=\frac{\left(\frac{p}{1-p}\right)}{\left(\frac{q}{1-q}\right)}$$

pとqがどれくらい違うかを定量化

例　$p=q$ なら OR$=1$
　　$p>q$ なら OR>1
　　$p<q$ なら OR<1

具体例

風邪薬を飲む場合、3日以内に治る確率 $p=0.8$
風邪薬を飲まない場合、3日以内に治る確率 $p=0.4$

$$\mathrm{OR}=\frac{\frac{0.8}{0.2}}{\frac{0.4}{0.6}}=6$$

医療統計で特によく使われる(95%信頼区間CIも記述することが多い)

◆ポアソン回帰

負の値をとらない整数値、特にカウント数が目的変数である場合に候補となる一般化線形モデルが、**ポアソン回帰（Poisson regression）**です。例えば、目的変数が1つの木に咲く花の数の場合、負の値を取ることはありませんし、整数値で表されるため、ポアソン回帰が候補となります。ポアソン回帰は、目的変数（の誤差項）が次に説明する**ポアソン分布**という確率分布に従う回帰で、回帰式は、ポアソン分布の平均を表すパラメータλを表します（図8.2.8）。

ポアソン回帰における説明変数x_iと目的変数y_iは、パラメータλ_iを通して次のような関係になります。

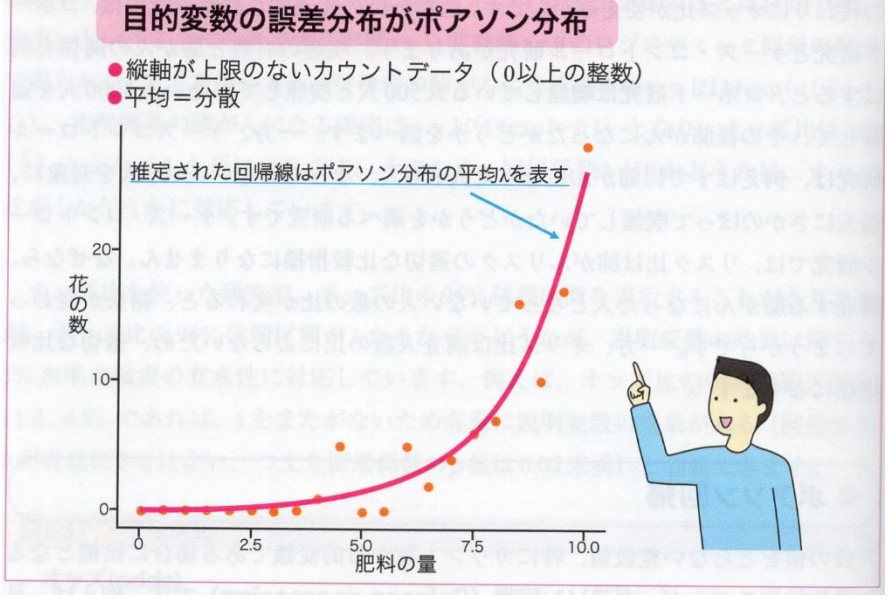
$$y_i \sim Poisson(\lambda_i) \quad \text{(式8.8)}$$

$$\lambda_i = \exp(a + bx_i) \quad \text{(式8.9)}$$

この場合のリンク関数は、式8.9の両辺の対数を取ることで、$\log(\lambda)$であることがわかります。

図8.3.8の例では、最尤推定の結果、$\hat{a} = -2.68$, $\hat{b} = 0.594$が得られ、与えた肥料の量に対して花の数が変化していることがわかります。

図 8.3.8　ポアソン回帰



◆Wald 検定

通常の線形回帰モデルと同様に、一般化線形モデルにおいても推定した回帰係数に関する仮説検定を、帰無仮説：回帰係数＝0、対立仮説：回帰係数≠0として実行することができます。推定値のばらつきを表す標準誤差を用いて、最尤推定で得られた推定値/標準誤差をWald統計量と言います。最尤推定量が正規分布に従うと仮定すれば、Wald統計量を用いて信頼区間やp値を得ることができます。この検定手法を、**Wald検定**と言います。ただし、サンプルサイズnが大きくない場合には、正規分布が従うという仮定は怪しくなる傾向があり、次に紹介する尤度比検定の方が信頼できるとされています。

驚くべきことに、仮説検定の使い方次第で p 値が 0.05 を下回るように操作することができてしまいます。自分にとって都合の良いように p 値を操作することを **p-hacking** と言い、これが非常に大きな問題になっています。

その後、ネイマンとピアソンは対立仮説を設定し、第一種の過誤、第二種の過誤を考える現代の仮説検定の潮流を作りました。ネイマン・ピアソンの枠組みでは、p 値は有意水準 α 未満か以上かにだけ着目し、仮説の棄却／採択の結論を下します。そのため、p 値が 0.01 であろうが 0.001 であろうが、$p<0.05$ という点で同ジであり、ともに統計的に有意という結果になります。ただし、ネイマン・ピアソン流の検定では、あらかじめ検出したい効果の大きさを決め、設定した α と β から、必要なサンプルサイズを決定しておく必要があります。**サンプルサイズが大きいと、ほんの小さな差であっても帰無仮説が棄却されてしまう**からです。

p 値の問題であった帰無仮説と対立仮説の非対性は、**ベイズファクター**では解消され、2 つの仮説を対等に比較することができ、帰無仮説を支持することも可能です。

◆HARKing

例3の状況を考えてみます。薬を開発することを想像してください。開発した薬Aに効果があるかを調べるために実験を行い、その結果、$p \geqq 0.05$ となり、統計的に有意な効果が見つからなかったとします。次に、新たに薬Bを開発し実験を行い、これも同様に $p \geqq 0.05$ でした。次は薬Cを試して…ということを繰り返します。

例えば、7番目の薬Gで $p < 0.05$ となり、統計的に有意な効果があるという結果が得られたとしましょう。ここで、「薬Gに効果があると仮説を立て、実験を行った。その結果、$p < 0.05$ で仮説が支持された」という論文を書いてしまいがちですが、これは *p*-hacking になります。特に、*p*-hacking に関連する概念である **HARKing** の例です。HARKing とは、Hypothesis After the Results are Known の略で、データを得て結果を見てから仮説を作ることです。具体的には、多くの実験を繰り返す、またはデータの様々な変数をこねくり回して意味のありそうな結果だけをピックアップし、最初から立てていた仮説として報告することです（図9.3.2）[6]。HARKing も *p*-hacking 同様に、再現性の低下につながります。

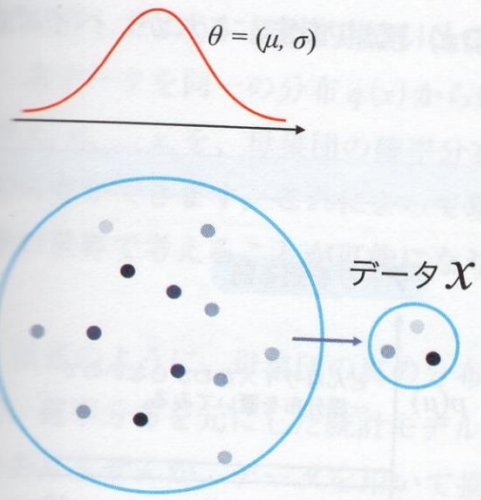
HARKing が再現性の低下につながる理由は、意味のないゴミデータをたくさん集めると、意味のありそうに見える結果がたまたま得られるからです。例えば、全く効果がない薬を 20 個（薬 A，B，…，T）用意して実験し、それぞれ有意水準 5 α=0.05 で検定を実施すると、20 個のうち平均的に 1 個が有意になります。あるいは、**20 個のうち少なくとも 1 つが統計的に有意と判定される確率は、64%**にもなります。これは検定を繰り返し実施することにおける、**多重性の問題と同じ**です。

ピックアップされず捨てられてしまった結果が論文に載っていなくても、論文の読者は捨てられたことに気づくことができません。そのため、HARXing していかを見抜くことが難しいという問題があります。

因果推論の根本問題

したしないの両方を観測できなにから、因果関係を調べることは原理的に不可能。

図11.1.1 頻度主義とベイズの不確実性の違い



頻度主義

θ は真の値が存在し、固定された値として考える $p(x \mid \theta)$

不確実性は、標本の抽出のみに由来

最尤法は $p(x \mid \theta)$ を θ の関数と見て尤度 $L(\theta \mid x)$ を最大化する θ^* を見つける

点推定

ベイズ

θ についての不確実性を確率分布として考える

データを知る前　**事前分布** $p(\theta)$

データを知った後　**事後分布** $p(\theta \mid x)$

最尤法

第8章で解説した最尤法を、おさらいしておきましょう。あるパラメータθに対し、データ $x_1, x_2, \ldots, x_n$ の起こりやすさ[4]は統計モデルを用いて、次のように表されます。

$$p(x_1, x_2, \ldots, x_n \mid \theta) \qquad \text{(式11.1)}$$

この式に対し、データを固定してパラメータθの関数と見なし、

$$L(\theta \mid x_1, x_2, \ldots, x_n) = p(x_1, x_2, \ldots, x_n \mid \theta) \qquad \text{(式11.2)}$$

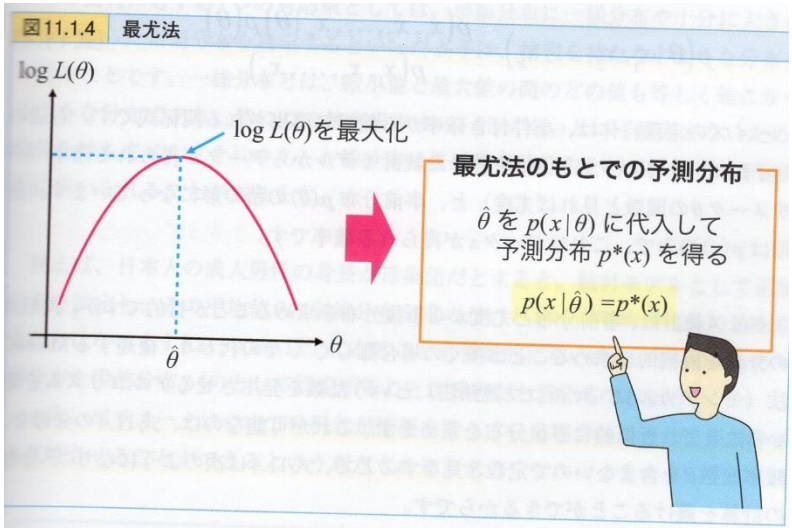
としたLを、尤度または尤度関数と呼びました。

尤度が大きいことは、指定したパラメータθでデータが起こりやすいことを意味するので、尤度関数 $L(\theta)$ を最大化するθを持つモデルが、手元にあるデータを生成する可能性がもっとも高いだろうと考えたわけです。このアイディアに基づき、パラメータθを決定するのが最尤法または最尤推定です。数式として書くと、次のようになります。

$$\hat{\theta} = \arg\max L(\theta \mid x_1, x_2, \ldots, x_n) \qquad \text{(式11.3)}$$

argmax $L(\theta)$ は、関数 $L(\theta)$ を最大化するθを表します。尤度を最大化するθの値を、$\hat{\theta}$ と表し、これが最尤法によって得られたパラメータで、最尤推定値と呼ばれます。最尤法、そして頻度主義統計では、パラメータの推定は「点」として1つの値が得られるのが特徴です。

この推定したパラメータ $\hat{\theta}$ を統計モデルのパラメータθに代入した $p(x|\hat{\theta})$ を、最尤推定で得られる予測分布 $p^*(x)$ と言います。



式 11.4 の分母を直接的に求めることは多くの場合難しい。

ベイズ統計学の考え方

ベイズ統計では、統計モデルのパラメータθを確率変数として扱い、その確率分布を考えます。そしてベイズ統計における推定は、データを知る前に持っているパラメータθに関する情報がデータを知ることで更新され、どういったθの値が起きやすいかを知ることができるようになるイメージです。

具体的な方法について解説していきましょう。ベイズ統計の準備として、統計モデル $p(x|\theta)$ に加え、分析者がデータを知る前の段階におけるθの確率分布、すなわち**事前分布 $p(\theta)$ (prior distribution)** を用意する必要があります。それらをもとに、データを知ったあとのパラメータθの確率分布である、**事後分布 $p(\theta|x)$ (posterior distribution)** を得ることが、ベイズ統計における推定です。

データ x、統計モデル $p(x|\theta)$、事前分布 $p(\theta)$ から事後分布 $p(\theta|x)$ を得るために、**ベイズの定理**を用います。

$$p\left(\theta \mid x_1, x_2, \ldots, x_n\right) = \frac{p\left(x_1, x_2, \ldots, x_n \mid \theta\right) p\left(\theta\right)}{p\left(x_1, x_2, \ldots, x_n\right)} \qquad \text{(式 11.4)}$$

ベイズの定理自体は、条件付き確率から簡単に導出される関係式です。右辺の分子には、あるパラメータθを持った統計モデルからデータが得られる確率（パラメータθの関数と見れば尤度）と、事前分布 $p(\theta)$ の積の形になっています。分母は $p(x)$ のみで、これはデータ x が得られる確率です。

ベイズ統計は、事前分布と尤度から事後分布を求めることが目的ですが、式11.4の分母を直接的に求めることは多くの場合難しく[5]、その代わり、後述するMCMC法（モンテカルロ・マルコフ連鎖法）という乱数を発生させるアルゴリズムを用いることで、近似的に事後分布を求めます。これが可能なのは、式11.4の分母を、確率変数θを含まないので定数と見なすことで、式11.4は次のようになり[6]、分母の計算を避けることができるからです。

（事後分布）∝（尤度）×（事前分布）

事後分布の形状は、尤度と事前分布の積に比例し、確率を直接計算できなくても、発生させた乱数のヒストグラムの形状から事後分布の近似的な分布（経験分布）を得ることができます。

頻度主義統計の多くの仮説検定手法では、計算を簡単に実行することができます。それに比べると、ベイズ統計はやや複雑な手順を踏む必要があります。

◆ランダムウォーク

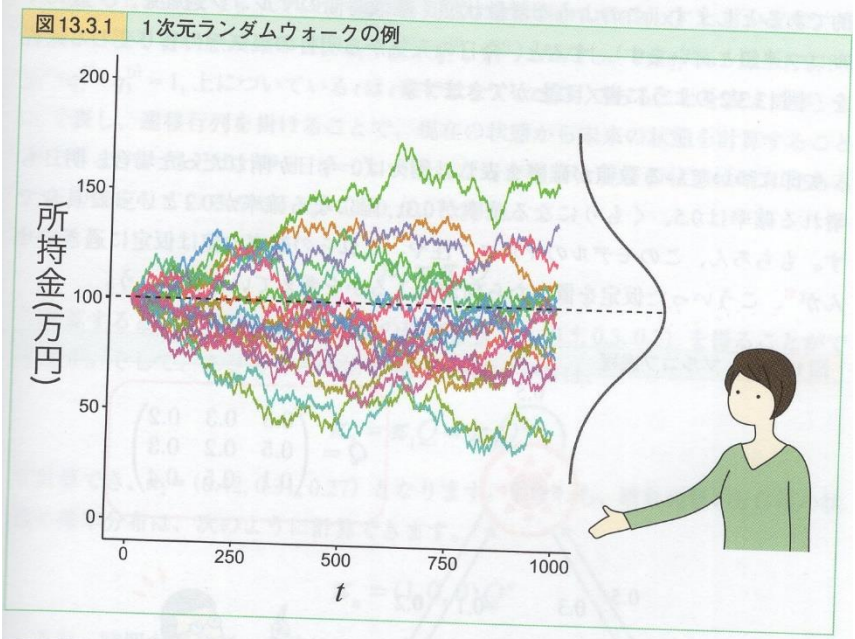
確率的モデルの最もシンプルな例として、ランダムウォーク（random walk）を紹介します。

あるギャンブラーが、100 万円の資金を元手にギャンブルを始めたとしましょう。各時刻 t において、確率 1/2 で 1 万円を得て、確率 1/2 で 1 万円を失う、勝ち負けが均等のギャンブルを考えます。単純化のために借金も可能、つまり持ち金がマイナスになることもあるとしましょう。

このとき、元手の 100 万円は、時間とともにどのように変化するでしょうか。これを表現したのが、1 次元ランダムウォークというモデルです。図 13.3.1 に、確率的なシミュレーションを実行した例を示しました。確率的なモデルなので毎回結果が異なりますが、面白いことに、ある法則性が現れてきます。

勝敗の確率は均等なので、お金の増減の期待値は 0 です。しかし、時間を経るにつれ、徐々に大きい値や小さい値を取ることが増えてきます。実は、時刻が t 分経つと、手元のお金の分布は平均 100 万円、標準偏差 $\sigma\sqrt{t}$（ここでの σ は、1 と −1 が均等に出る確率分布の標準偏差＝1）の正規分布で近似できます。これは、＋1＋（−1）＋（＋1）＋…という確率変数の和が正規分布に従う中心極限定理から説明できます。そして、この正規分布の標準偏差が $\sigma\sqrt{t}$ であるため、各人の手元の**お金は、時間とともに 100 万円から離れていく傾向がある**ことがわかります。

図 13.3.1　1 次元ランダムウォークの例

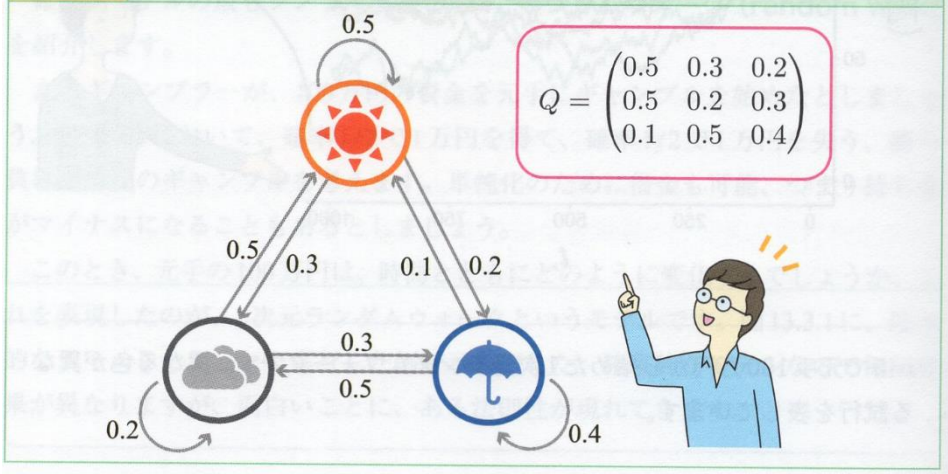


$t=0$ で元手 100 万円から始めた 1 次元ランダムウォークです。異なる色が異なる試行を表しています。

「過去に依らず、現在によって未来が決まる性質」を、マルコフ性と言います。

ランダムウォークでは、ギャンブルの勝敗はどのような過去があったかに関係なく決まる（と仮定している）ため、マルコフ性を持つと言えます。

図13.3.2　マルコフ過程



$$Q = \begin{pmatrix} 0.5 & 0.3 & 0.2 \\ 0.5 & 0.2 & 0.3 \\ 0.1 & 0.5 & 0.4 \end{pmatrix}$$

丸が各状態を表し、状態間の数値が遷移確率を表します。

このマルコフ過程を解析するためには、各状態に晴れ＝1, くもり＝2, 雨＝3として番号をつけ、状態iから状態jへの遷移確率であるp_{ij}を並べた行列

$$Q = \begin{pmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{pmatrix} \quad \text{(式13.12)}$$

を考えると便利です。これを、**遷移行列（transition matrix）**と言います。時刻tにおける各状態の確率を、$\boldsymbol{\pi} = (a^{(t)}$