2011 年度前期　統計学で考える　第１４回

地震予知と狼少年 ー ベイズの定理

狼少年は，狼が本当に現れたときには，確かに本当のことを言っている。

地震予知の精度とは

次のような問題を考えます。

> ある地震予知技術では，毎日 1 回地震情報を出します。ある規模以上の地震が起きる日には，前日に 97%の確率で警報を出します。しかし，地震が起きない日にも 5%の確率で誤って警報を出してしまいます。この地域では，警報を出すべき規模の地震が 1 日に起きる確率は 2%であるとします。さて，警報が出たとき，その日に地震が起きる確率はいくらですか。

これを，次の手順で解いてください。以下，地震情報を出した日数を x 日とし，x は十分大きいとします。

1. x 日のうち，地震が起きた日数は何日ですか。
2. 警報が出て，かつ地震が起きた日数は何日ですか。
3. 地震が起きなかった日数は何日ですか。
4. 警報が出て，かつ地震が起きなかった日数は何日ですか。
5. 警報が出た日数は合計何日ですか。
6. 警報が出た日数のうち地震が起きた日数の割合はいくらですか。この値が求める確率です。

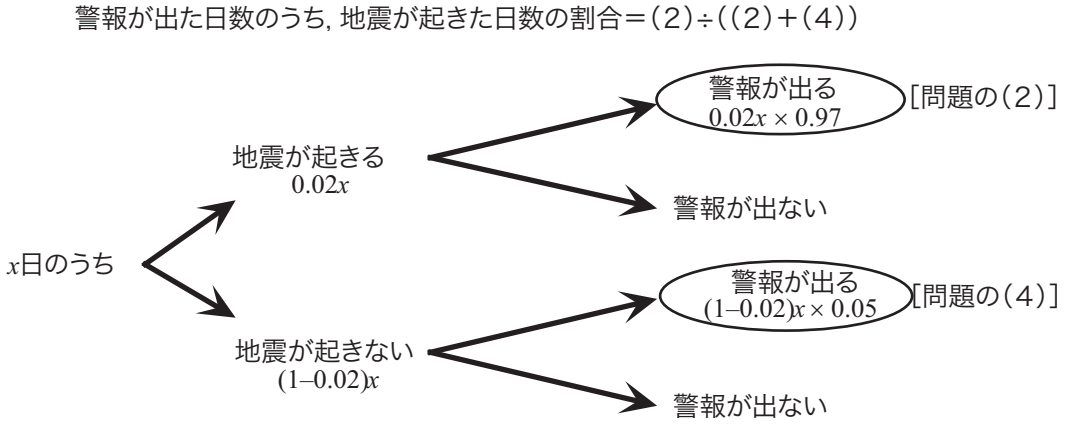

図 1: 例題の考え方

解答はこうなります

1. 「警報を出すべき規模の地震が 1 日に起きる確率は 2%」とありますから，求める日数は $0.02x$ です。

2. 「地震が起きる日には，97%の確率で警報を出す」とあり，地震が起きた日数は $0.02x$ ですから，求める日数は $0.97 \times 0.02x$ です。

3. 「地震が 1 日に起きる確率は 0.02」ですから，地震が起きない確率は $(1 - 0.02)$ です。よって，求める日数は $(1 - 0.02)x$ です。

4. 「地震が起きない日にも 5%の確率で警報を出してしまう」とあり，地震が起きない日数は $(1-0.02)x$ ですから，求める日数は $0.05 \times (1 - 0.02)x$ です。

5. 求める日数は 2. と 4. の合計で，$0.97 \times 0.02x + 0.05 \times (1 - 0.02)x$ です。

6. 求める割合は，5. の日数のうちの 2. の日数の割合で，

$$\frac{0.97 \times 0.02x}{0.97 \times 0.02x + 0.05 \times (1 - 0.02)x} = \frac{0.97 \times 0.02}{0.97 \times 0.02 + 0.05 \times (1 - 0.02)} = 0.284 \tag{1}$$

   となります。図 1 で，この関係を確かめてみてください。

条件付き確率

上の問題では，「地震が起きたときに，警報が出る確率」が 97%，「地震が起きていないときに，警報が出る確率」が 5%となっています。これらの確率は，数学で**条件付き確率**とよばれるものです。その意味を，さいころの各目が出る確率を例にとって，以下で説明します。

さいころで，「3 以下の目が出る確率」を図に表すことを考えます。さいころで，「可能なすべての目」は 1, 2, 3, 4, 5, 6 の 6 通りで，これを集合 Ω で表します。一方，「3 以下の目」は 1, 2, 3 の 3 通りで，これを Ω の内部にある集合 A で表します。

このとき，「3 以下の目が出る確率」は，集合 A の要素がおきる確率なので，「事象 A がおきる確率」で，$P(A)$ で表します。$P(A)$ は，「集合 A の要素の数」を $|A|$ で表すと，

$$P(A) = |A|/|\Omega| = 3/6 = 1/2 \tag{2}$$

となります。

さらにもうひとつ，「偶数の目が出る確率」を考えます。同様にして，「偶数の目」は 2, 4, 6 の 3 通りで，これを集合 B で表すと，「偶数の目が出る確率」$P(B)$ は

$$P(B) = |B|/|\Omega| = 3/6 = 1/2 \tag{3}$$

となります。これらを目に見えるように表したのが「ベン図」で，図 2 となります。

では，「3 以下かつ偶数の目が出る」確率を考えましょう。この事象は集合 $A \cap B$ で表されますから，その確率 $P(A \cap B)$ は

$$P(B) = |A \cap B|/|\Omega| = 1/6 \tag{4}$$

となります。

ここで，$|A \cap B|/|B|$ という確率を考えてみましょう。図 3 の太線の部分です。分母が $|\Omega|$ から $|B|$ に変わっていますから，ここでは，「偶数の目」が，ここでの「可能なすべての目」になっています。一方，

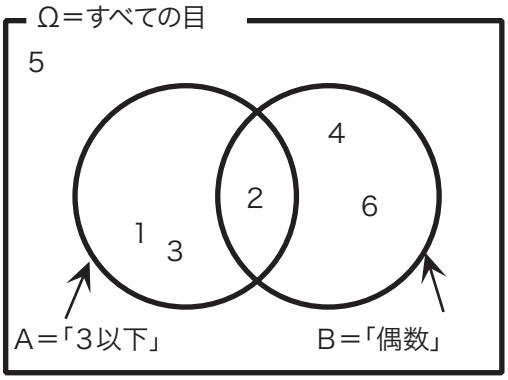

図 2: ２つの事象とベン図

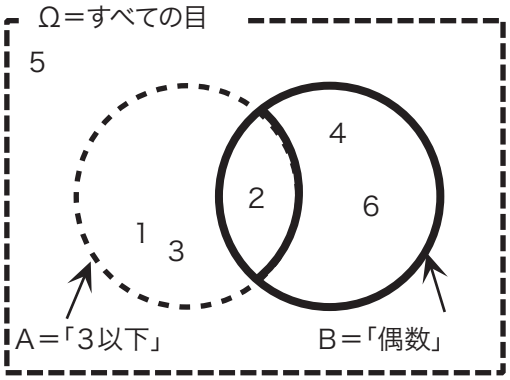

図 3: 条件付き確率

$A \cap B$ は「３以下かつ偶数の目が出る」という事象ですが，今は「偶数の目が出る」という事象の中でしか考えていませんから，この事象は単に「３以下の目が出る」という事象ということができます。したがって，

> $|A \cap B|/|B|$ ＝偶数の目が出るとわかっている時（偶数の目が出るのが確実な時），それが３以下である確率

になります。

これを，**「B を条件とする A の条件付き確率」**といい，$P(A|B)$ で表します。$P(A|B) = |A \cap B|/|B| = 1/3$ ですから，「偶数の目が出た」という情報が得られている時は，そうでないときよりも「３以下の目が出る」確率は小さくなることがわかります。

ところで，

$$P(A|B) = \frac{|A \cap B|}{|B|} = \frac{|A \cap B|/|\Omega|}{|B|/|\Omega|} = \frac{P(A \cap B)}{P(B)} \tag{5}$$

と表され，これを条件付き確率の定義としている本もあります。ただし，この場合，分母分子それぞれの確率は，いずれも同じ $|\Omega|$ を分母とする確率でなければならないことに，注意する必要があります。また，(5) 式から

$$P(A \cap B) = P(A|B)P(B) \tag{6}$$

となります。(6) 式は，簡単に言えば

> 「A と B の両方が起きる確率」＝「B が起きたとしたときに A が起きる確率」
> ×「本当に B が起きる確率」

ということです。$P(A|B)$ と $P(A \cap B)$ の違いも，これでわかると思います。

ベイズの定理

最初の問題では，「地震が起きたときに，警報が出る確率」が 97%，「地震が起きていないときに，警報が出る確率」が 5%となっています。これらの確率は，条件付き確率で表すと，それぞれ

- 「地震が起きること」を条件とする，「警報が出る」条件付き確率 $= 97\%$
- 「地震が起きないこと」を条件とする，「警報が出る」条件付き確率 $= 5\%$

ということになります。そこで，「地震が起きる」ことを事象 A，「警報が出る」ことを事象 B で表します。このとき，「地震が起きない」ことは事象 $\bar{A}$ で表されます。また，上の 2 つの条件付き確率は，$P(B|A) = 0.97, P(B|\bar{A}) = 0.05$ と表されます。また，問題で求める確率は「『警報が出る』ことを条件とする，『地震が起きる』条件付き確率」で，$P(A|B)$ となります。そこで，条件付き確率の考え方を使って，$P(A|B)$ を求めてみましょう。

$P(A|B)$ を表す (5) 式の A と B を入れ替えると

$$P(B|A) = \frac{P(B \cap A)}{P(A)} \tag{7}$$

となりますから，

$$P(B \cap A) = P(A)P(B|A) \tag{8}$$

となります。

一方，$P(A|B)$ を表す (5) 式の分母は，

$$P(B) = P(B \cap A) + P(B \cap \bar{A}) \tag{9}$$

となります。これは，事象 A（地震が起きる）と事象 $\bar{A}$（地震が起きない）が，「同時にはおこらず，しかもどちらかが必ず起きる」という関係にあるからです。これを**排反**といいます。

(9) 式の右辺を，上と同様に条件付き確率で表すと，

$$P(B) = P(A)P(B|A) + P(\bar{A})P(B|\bar{A}) \tag{10}$$

となります。

以上をあわせると，

$$P(A|B) = \frac{P(A)P(B|A)}{P(A)P(B|A) + P(\bar{A})P(B|\bar{A})} \tag{11}$$

となります。この関係を**ベイズの定理**といい，今回最初に示した計算はこの関係を計算したことになっています。

ここで，$P(A)$ は「地震が起きる確率」で，最初の問題は 2%となっています。この確率は，地震警報の精度がうんぬんという段階では，本当はわからないはずで，実際には「目分量で見積った」確率です。実は，これは前回の講義の「意思決定」のところで出てきたのと同じ**事前確率**です。これに対して，今回求めた「警報が出るという条件のもとでの，地震が起きる条件付き確率」$P(A|B)$ を**事後確率**といいます。また，ここで述べているような事前確率・事後確率を導入した統計学を**ベイズ統計学**とよんでいます。

ところで，上の計算で最初の問題の答えを計算すると，「警報を出した日に本当に地震が起きる」確率は 0.284 となります。ということは，この地震警報はほとんど役に立たないことを意味しています。地震が起きるときには 97%の確率で警報を出すのに，どうしてこういうことになるのでしょうか？ そのわけは，「地震が起きる確率」すなわち事前確率が 2%と小さい，という点です。「地震が起きる日」はほとんどないわけですから，「地震が起きる日に警報を出す能力が高い」ことよりも，「地震が起きない日に警報を出さない能力が高い」ことのほうが，警報全体の信用度を上げるのに大きく影響します。

このことから思い出されるのは，「狼少年」の話です。この場合も，狼はそう頻繁には現れませんから，いくら本当に狼が現れたときに少年が「狼が来た」と叫んでも，ふだん狼が現れていないときに頻繁に叫んでいれば「叫んだときに狼が現れる確率が小さい」すなわち「叫びは信用できない」ということになるわけです。

最近，ベイズの定理を応用した新しいソフトウェアが普及してきています。それは「迷惑メールフィルタ」です。このソフトウェアでは，あらかじめ「迷惑メールとわかっているメール」と「迷惑メールでないとわかっているメール」を用意しておきます。そして，迷惑メールに入っていると思われる単語（たとえば “viagra”）について，「迷惑メールであるとわかっているとき，そのメールがこの単語を含む条件付き確率」「迷惑メールでないとわかっているとき，そのメールがこの単語を含む条件付き確率」を計算し，ベイズの定理を使って「この単語を含むメールが，迷惑メールである確率」を求めます。この確率が大きければ，そのメールは迷惑メールであると判断します。

このとき，「迷惑メールがやってくる確率」すなわち事前確率は，それまでに受け取ったメールのうちの迷惑メールの割合と考えることができます。したがって，このソフトウェアを使って，メールのふりわけを行えば行うほど，事前確率や上の条件付き確率は正確になり，ソフトウェアの能力はあがってゆきます。

「独立」の概念

ここまでの講義で，「独立」という言葉が何度か出てきました。ここまで，何となく「できごとが互いに無関係」という感じで説明してきましたが，その正確な意味は条件付き確率によって定義されます。

条件付き確率を説明した，今日のさいころの例で，事象 A が「3以下の目」ではなく「2以下の目」だったらどうでしょう。このときは，「2以下の目が出る確率」$P(A) = 1/3$ です。一方，$P(A \cap B) = 1/6$ や $P(B) = 1/2$ は変わりませんから，$P(A|B) = |A \cap B|/|B| = 1/3$ もかわりません。

したがって，このときは $P(A|B) = P(A)$ となります。このときは，「事象 A が起きる確率」と「事象 B が起きるとわかっているときに，事象 A が起きる確率」が同じですから，事象 B が起きるかどうかには関係がないことを意味しています。このとき，事象 A と事象 B は**独立**であるといいます。

事象 A と事象 B が独立のとき，(5) 式から

$$P(A \cap B) = P(A)P(B) \tag{12}$$

となります。**事象 A と事象 B が独立のときこうなるのであって，いつもこうなるのではない**ことに注意してください。