第 9 章 統計的仮説検定

浅川 伸一

2006 年 06 月 22 日

統計的仮説検定とは想定される母集団に関して、ある仮説 H (hypothesis の頭文字) を正しいと判定するか否かを決定するための材料を提供することである。繰り返しになるが、データはある母集団から無作為に抽出されたものとみなすことができることが前提である。

1 具体例による検討

コイントスを 400 回行った結果 215 回表が出た。このコインの表の出る確率は $p = 1/2$ と言って良いか。

コイン投げは無限母集団を想定して良い。そこで仮説 H:$p = 1/2$ を検定することを考える。これは 2 項分布であるから期待値は $np = 400 \times 0.5 = 200$, であり分散は $nqp = 400 \times 0.5 \times 0.5 = 100$, 従って標準偏差は分散の開平 $\sqrt{100} = 10$ である。このことから表が 215 回でた結果を標準得点 (z スコア) に変換して $(215 - 200)/10 = 1.5$ を得る。データ数が $n = 400$ と大きいので正規分布と見なせると考える。そして仮説 H が正しいとしたら、215 回より大きい数、すなわち 400 回コイントスして、215 回以上表がでる確率が小さければ、そのようなことは滅多に起らないと考えて、仮説 H が正しいと見なす。

実際、統計表を引いて調べる、あるいは、表計算ソフトの関数を使うなどして (1-normdist(1.5,0,1,1) あるいは 1-normdist(215,200,10,1))、この確率が約 0.0668 であることを知ることができる。

この場合、100 回に 6 回強 215 回以上表の出る可能性がある訳だが、これを多いと見なすか少いと見なすのか。一般に統計学では 0.0668 では多いと見なさない。従って仮説 H を捨てる (棄却する) ことはできない。仮説 H は採択される。すなわちこのコインの表の出る確率は 1/2 と言える。

2 具体例による検討 (2)

コイントスを 400 回行った結果 220 回表が出た。このコインの表の出る確率は $p = 1/2$ と言って良いか。

上とまったく同じ条件で数字のみ異なる。この場合の z 得点は 2 となり、220 回以上表の出る確率は、統計表、もしくは表計算ソフトの関数により (1-normdist(2,0,1,1) あるいは 1-normdist(220,200,10,1))、0.0228 を得る。上と同様に 100 回繰り返すと 2 回強は 220 回以上表の出ることがあり得る。

ところが、統計学では 0.05(5%) を基準にして、仮説を採択するか、棄却するかを決めることがある。この値のことを 5% の危険率、あるいは有意水準と言ったりする。この場合は 5% の危険率で仮説 H は棄却された、という。

この基準に照らして考えれば、このコインの表の出る確率は 1/2 とは言えない。

この二つの例から分かるとおり、表の出た回数が 215 回と 220 回とでは、異なる結論が導き出される。もちろん確率的事象を扱うので、断言できるものではない。

事実 z 得点にした場合は norminv(0.95,0,1)=1.644853627, あるいは元のデータを扱うことにして norminv(0.95,200,10)=216.4485363 であるから、216 回表が出た場合は仮説 H を採択し、217 回表がでた場合仮説 H を棄却することになる。

換言すれば、統計学における判断、検定の考え方とは明白な基準を用いているというよりも、慣例にしたがって判断を下すことが多いということである。少くとも今までは多くの人たちが、この考え方を受け入れてきた。

なお危険率、あるいは有意水準は 0.05(5%) 以外にも 0.01(1%) や 0.001(0.1%) に設定されたりする。伝統的な統計学のテキストでは、有意水準はデータを集める前にあらかじめ決定しておくべきものとされている。

3 両側検定と片側検定

上の例での仮説 H は $p = 1/2$ ということであった。厳密に考えればこの説の否定は $p \neq 1/2$ であるから $p > 1/2$ であるか $p < 1/2$ のどちらかであるから、仮説 H を否定する場合、期待値 (平均) 回数より多く出るか少く出るかは分からない。従って上下両方に危険率を設定する (危険率を示す区間のことを危険域ともいう)。これを両側検定という。

一方、あらかじめ $p < 1/2$ であることが知られていれば $p > 1/2$ の可能性を考慮しなくても良いので、この場合は片側検定と呼ばれる。

norminv(0.95,0,1)=1.644853627, norminv(0.975,0,1)=1.959963985 より、標準化得点に変換したの値であれば片側検定で有意水準 5% の値は 1.645, 両側検定では 1.96 であることがわかる。両側検定では、分布の裾野の左右両端部分の面積を合わせて全体の 5% となるような領域に設定されることに注意。

上の例では、norminv(0.95,200,10)=216.4485363, norminv(0.975,200,10)=219.5996398 なので、片側検定で危険率 5% のときは 217 回以上、両側検定では 220 回以上となる。

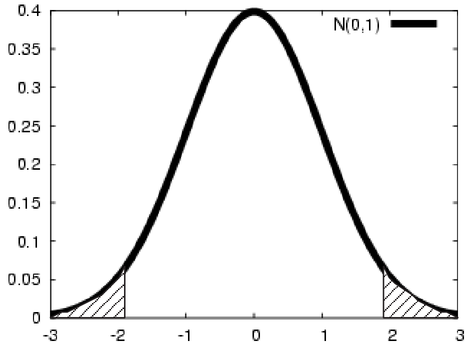


図 1: 両側検定

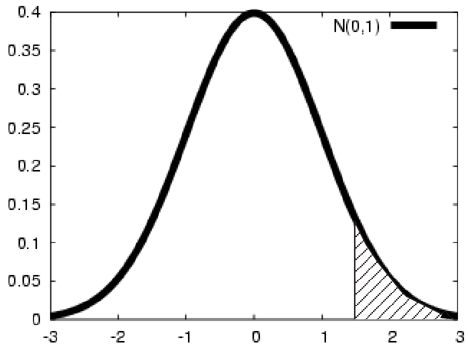


図 2: 片側検定

上の例のような微妙な数値の場合、危険率 (信頼水準) の設定、両側検定か片側検定か、どちらを用いるかによって解釈が異なってしまうことがある。ここでも伝統的な統計学のテキストでは、データを集める **以前に** 危険率や片側検定か両側検定かを設定すべきであると主張している。データを見てから態度を変えてはいけないとされる。400 回中 219 回表が出たことが明らかになった時点で、帰無仮説を $p = 1/2$ から $p < 1/2$ に変更してはいけない。結論が異なる。

4 帰無仮説 null hypothesis

統計的推論は「検定の結果 5% で有意であった。」などと記述される。上の例によれば、これはデータから計算されたある量は、危険率の範囲に入ったことを意味している。

統計的な推論方法とは次のとおりである。ある言明 A を主張したいとき

に、故意に否定の言明 not A を検討すべき仮説として考える。そして否定の言明 not A が起る可能性が滅多にない、すなわち危険域に入った場合には、その仮説 (否定の言明 not A) は採択されず、従って、言明 A が正しいものとみなすことにしておく。そうしてから、データを集つめ、集計した結果に基づいて検定する。その結果、ある有意水準でしか起こり得ないデータを得る。否定の言明 not A が真であるとした仮説、(危険率以下の水準でしか) 起こり得ないことが起きたのだから、否定の言明 not A は支持されず、従って、言明 A が正しいであろう、とすることのである。

このような推論方法は、意図的に自身の主張したことの逆を仮説として考え、この仮説を棄却することによって自説の正しさを主張することに他ならず、棄却されるべき仮説のことを帰無仮説 null hypothesis という。帰無仮説を H_0 などと表記する。帰無仮説の否定のことを対立仮説といい H_1 などと表記する。

例えば、自分の関心のある変数が、ある行動に影響を与えることを主張したい場合、「この変数によって、その行動は変化しない」ことを帰無仮説とする。帰無仮説が棄却されることで、自説が正しいことを主張する。

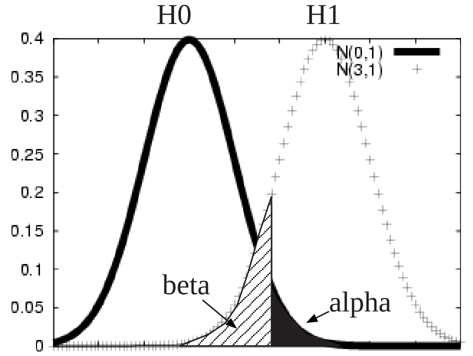
直接言明 A を正しいとみなして、データから言明 A が正しい可能性は 99% である、と言った説明はなされてこなかった。

5　二種類の誤り

危険率が設定された場合、帰無仮説が誤って棄却されてしまう危険性は、危険率だけ存在する。これを第一種の誤り type I error という。

ところがもう一つ誤りが存在する。それは対立仮説が正しいにもかかわらず、帰無仮説を採択してしまう誤りである。こちらは第二種の誤り type II error と呼ばれる。

第一種の誤りを α 第二種の誤りを β と表記する文献もある。



図には α と β の関係が示されている。「良い」統計的検定とは、第一種の誤りである危険率 α が小さいだけでなく、第二種の誤り β も小さいことが望ましい。$1-\beta$ のことを検定力 power と言ったりする。

ところが、第一種の誤り α はデータを集計する人間が決めることができるが、第二種の誤り β を明確に定めることは困難である。そもそも β が不明なことも多い。

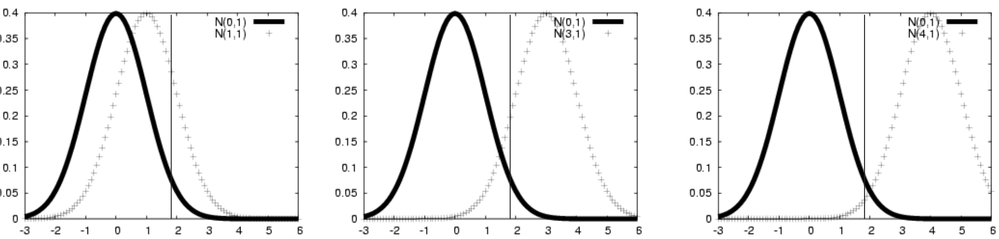


図 3: α 一定の条件のもとでの β の変化

演習：上の図で第一種の誤り α の領域と第二種の誤り β の領域を示せ。

ただ、8 章の標本平均と標本平均の分散についての議論から、データ数が多くなればなるだけ検定力は上がる。

6 平均値の差への応用

二つのグループ A 群と B 群、実験条件 A 条件と B 条件、などからデータを得て、標本平均を考えたときに、これらの両者が同じと言えるかどうかを検定することを考える。

帰無仮説としては、H_0:両群に差がない、をとる。

条件 A の標本平均を $\bar{x}_A$, 条件 B の標本平均を $\bar{x}_B$ とする。条件 A の標本平均の分散は $V(\bar{x}_A)=\sigma_A^2/n_A$ とし、条件 B の標本平均の分散は $V(\bar{x}_B)=\sigma_B^2/n_B$ とする。ここで、n_A は条件 A のデータ数、n_B は条件 B のデータ数である。ここで帰無仮説に従えば両標本平均の分散は等しい $\sigma_A^2=\sigma_B^2$ はずなので、これを σ^2 とする。

条件 A の標本平均の分散と、条件 B の標本平均の分散との差の分散は、各分散を用いて $V(\bar{x}_A-\bar{x}_B)=\sigma^2/n_A+\sigma^2/n_B$ と表せる (8 章を参照)。

従って、条件間の標本平均の差は

$$z=\frac{\bar{x}_A-\bar{x}_B}{\sqrt{\frac{\sigma^2}{n_A}+\frac{\sigma^2}{n_B}}} \tag{1}$$

となる。データ数が多ければこの式は限りなく $N(0,1)^2$ に近づく。従って、この量を元に適当な有意水準で検定が可能となる。