# ９章 検定の基礎

## 9.1 検定とは

これまではデータの集計方法を中心に話を進めてきましたが、これからしばらくは統計学の中で大きな位置を占める検定と呼ばれる分野の話をします。一般的な説明をする前に、検定とはどんなことをするのか、意味は後にして、一度行ってみようと思います。次の例を見て下さい。

**例**

超能力を持つという人にコインの表裏を当てる実験をしてもらい、100 回の施行で70%の正解率を得た。この人には本当に超能力があると考えられるか？有意水準を 5%として判定せよ。20 回の試行ではどうか。

この問題はテレビなどでよく見る超能力実験を簡単にモデル化したものです。ここで、100 回中 70%の正解率というのは、結構当っているのでしょうか？何を基準に当っている、当っていないを判断するのでしょうか。通常超能力のない人の場合、表裏を当てる確率は 0.5 ですから、100 回中 50 回当てることは難しいことではありません。では、70 回を偶然当てるのは難しいのでしょうか。これを判断するのは、偶然に 70%以上の正解率を得る確率です。この確率が極端に小さい場合、偶然でこんなことは起こり得ないと判断して、何らかの超能力があるのではないか、と解釈します。また、確率があまり小さくない場合は、たぶん偶然だろうから、超能力があるとは言えないと解釈します。ではどの位の確率がこの判断を分けるのでしょうか。通常の場合、この確率は 0.05 とされています。超能力があると判断した場合、5%は判断を間違える（事象は偶然に起こった）かも知れないという基準です。人命がかかるような重要な判断には、この基準を 0.01 に下げます。判断が間違っている可能性を低く押さえるためです。では、なぜ 0.05 や 0.01 なのでしょうか。これには特に絶対的な理由はありません。統計学の世界で広く一般に認められている共通の判断基準なのです。この確率のことを有意水準といいます。問題では有意水準を 5%として、とありますから、この基準の確率は 0.05 となります。有意水準は超能力があると判断して間違える危険性の確率でもありますので、危険率とも呼ばれます。

さて、実際に偶然起こったものとして、100 回の試行で 70%以上の正解率を得る確率を近似的に求めてみます。

まず、以下の量 $\chi^2$ を計算します。

$$\chi^2 = \frac{(\text{当った数} - \text{当る予想数})^2}{\text{当る予想数}} + \frac{(\text{外れた数} - \text{外れる予想数})^2}{\text{外れる予想数}}$$

ここに、偶然に当る確率は 0.5 ですから、当る予想数は$100 \times 0.5 = 50$回、外れる確率も 0.5 ですから、外れる予想数も 50 回です。この量は当った数が予想値を超えて多くなるほど大きくなっていく量です。実際に$\chi^2$の値を計算してみます。

$$\chi^2 = \frac{(70-50)^2}{50} + \frac{(30-50)^2}{50} = 2 \times \frac{400}{50} = 16$$

さて、今計算してもらったこの$\chi^2$という量ですが、これは、前の章で学んだ自由度 1 の$\chi^2$分布に従うことが知られています。そのため、Excel を用いて簡単に$\chi^2 \geq 16$となる確率$p$を求めることができます。

$$p = chidist(16,1) = 6.33425\text{E}-05 \cong 0.00006 < 0.05$$

これは、一般的な教科書の書き方だと、以下のようになります。

$$\chi_1^2(p) = 16, \quad p \cong 0.00006 < 0.05$$

左の式は自由度 1 で上側確率$p$の$\chi^2$値 $\chi_1^2(p)$ の値が 16 に等しいとき、というように見ます。我々は、Excel を中心に考えていますので、主に前者の書き方で示すことにします。

パソコンが利用できず、確率の値が簡単に求められない場合、予め統計表などから自由度 1 で上側確率 0.05 の$\chi^2$値を与えておいて、計算した$\chi^2$の値がこの値に比べて大きいか小さいかで、判断する場合もあります。この例の場合では、以下のようになります。

$$\chi^2 > \chi_1^2(0.05) = 3.841$$

確率$p$の値を有意水準と比較した結果、この場合は、超能力があるといえると判断します。これは言葉を変えると、これが偶然とすると 70%的中の実現確率は 0.05 に比べて小さいといえる、となります。

さて、試行回数 20 回ではどうでしょうか。

$$\chi^2 = \frac{(14-10)^2}{10} + \frac{(6-10)^2}{10} = 2 \times \frac{16}{10} = 3.2$$

$$p = chidist(3.2,1) = 0.073638 \cong 0.074 > 0.05$$

（$\chi_1^2(p) = 3.2, \quad p \cong 0.074 > 0.05$）

これより、超能力があるとはいえないと結論されます。

## 9.2 何を検定するか

検定はどんな状況で実施するのでしょうか。この節では母集団、標本、検定の関係について見ていこうと思います。

最初に１つの母集団を考えます。この母集団の平均や分散などは未知であるとします。最初に考えるのはこの母集団のある統計量が我々が想定する値と差があるかどうかという検定です。この判断のために我々は母集団からランダムサンプリングによって標本を取り出し、そこで求めた統計量と想定された統計量とを比較します。例えば模擬テストの全国平均が分っているものとして、これとある地域の生徒の平均点を比較する問題を考えます。指定値としてはその全国平均の値を利用し、標本としてその地域の生徒の集団を取り、標本から得た平均と全国平均とを比較します。これによってこの地域の生徒の成績が全国平均と比べて差があるかどうか判定します。また、ある標本に対するアンケート調査の質問で賛成が 60%であったとします。この標本を抽出した母集団で過半数が賛成していると言えるかという問題もこの分類に含まれます。この場合は標本の比率と想定比率 50%とを比べて過半数かどうか判定します。この教科書ではこのような検定について「指定値との比較」として手法を説明します。

もう 1 つの状況は、2 つの母集団があり、どちらも情報が未知の場合です。2 つの母集団の情報を得るために、それぞれから標本を取ります。この標本の情報をもとに 2 つの母集団間の性質に差があるかどうか調べます。例えば、ある 2 つの地域の住民を母集団として標本を取り出し、平均や分散について比較を行う場合や、アンケートによる支持率比較などが考えられます。この本ではこの問題について、「標本間の比較」として様々な手法を紹介します。以上の関係を直感的に描いたものが図 9-1 と図 9-2 です。

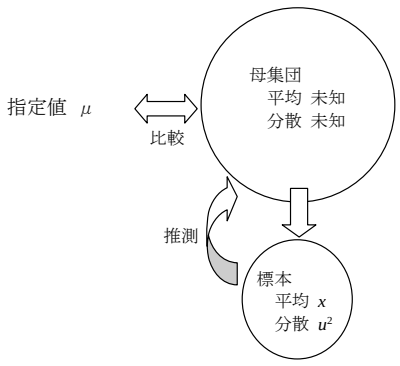


図 9-1 指定値との比較

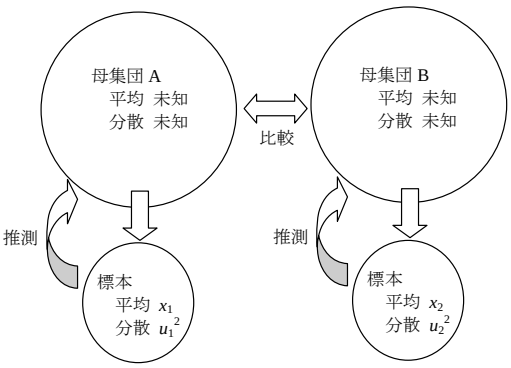


図 9-2 標本間の比較

これらの図は検定をできるだけ統一的に理解するために描いたものです。統計量とし

ては、平均と分散の他に中央値や相関係数なども考えます。

図 9-2 で２つの母集団を別々に描いていますが､検定結果によっては２つの母集団が実は同一であるということになる場合があります。そのため２つの標本は同一の母集団から抽出されたものか検定するという考え方で、１つの母集団から２つの標本を取り出して比較するような図が多く見られます。また図 9-1 でも、母集団の平均が分かっていて、標本は本当にこの母集団から抽出されたものかという見方がむしろ一般的です。ただこれらの考え方は初心者には馴染みにくいものと考え、思いきってもう少し直感的な表現法にすることにしました。

## 9.3 検定の手順と過誤（誤り）

さて、この節では検定の方法とその問題点について見てみようと思います。例として、ある集団の身長の平均と全国平均との比較を考えます。これは図 9.1 の場合に相当し、指定値として全国平均を、母集団としてその集団の対象者を考えます。今、身長の全国平均を $\mu$ としますが、母集団について平均は未知で、仮に $\mu_1$ としておきます。母集団から標本を取り出し、その統計量を調べたところ、データ数 $n$，標本平均 $\bar{x}$，不偏分散 $u^2$ の結果を得たものとします。

検定を行うにはまず指定値と母集団の平均が等しい、即ち $\mu = \mu_1$ を仮定します。この仮定を帰無(きむ)仮説といい、記号 $H_0$ で表わします。それに対して我々の疑問は、$\mu \neq \mu_1$ ではないかということです。この疑問は帰無仮説と対立することから、対立仮説といい、記号 $H_1$ で表わします。

帰無仮説 $H_0$： $\mu = \mu_1$

対立仮説 $H_1$： $\mu \neq \mu_1$

次にやるべきことは帰無仮説を仮定して、確率変数から分布のよく知られた何らかの統計量を求めることです。この統計量については昔から統計学者によってどのようなものが良いのか議論されてきました。ここでは、データの正規性を仮定して、自由度 $n-1$ の t 分布に従うことが知られている統計量 $t$ を用いてみます。

$$t = \frac{\sqrt{n}(\bar{x} - \mu)}{u} \sim t_{n-1} \text{ 分布}$$

この量は母集団より取り出された標本データから計算され､その実現値 $t$ から 8.3 節で述べたようにして外側確率 $p$ を計算します｡この関係を図 9-3 に表わしておきます。

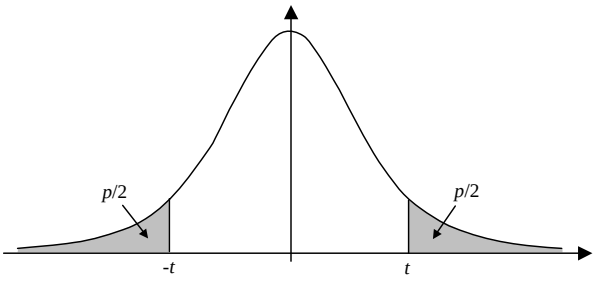


図 9-3　t 値と確率

この確率 $p$ が有意水準（危険率）$\alpha$ よりも小さければ帰無仮説 $H_0$ を棄却し、対立仮説 $H_1$ を採択します。逆に有意水準$\alpha$ よりも大きければ決め手に欠けるとして帰無仮説 $H_0$ をそのまま採択します。特に帰無仮説を棄却する領域という意味で、両側確率が$\alpha$ となる点の外側を棄却域と呼びます。

ところで、これら判断には間違った判定を下す可能性が 2 つあります。1 つは対立仮説 $H_1$ を採択したけれども実は $\mu = \mu_1$ であった場合、もう 1 つは帰無仮説 $H_0$ を採択したけれども $\mu \neq \mu_1$ であった場合です。前者の間違いを第 1 種の過誤といい、これが有意水準に相当します。後者の間違いを第 2 種の過誤といい、複数の検定統計量のうちどの統計量が優れているかの判断基準になります。

この教科書での言葉の使い方として、少し注意しておきたいことがあります。例えば、平均値の検定を行う際に、例として問題を提示する場合、「平均に差があるといえるか、有意水準 5%で判定せよ。」と書いています。本来の主旨は「平均が等しいと仮定して、有意水準 5%で検定せよ。」となるべきでしょうが、少し回りくどく数学の苦手な人には直感的でないと考え、敢えて「判定」という言葉を用いました。統計の専門家の方には批判を受けるかも知れませんが、ご容赦願います。

## 9.4 両側検定と片側検定

t 分布や、標準正規分布を用いて検定を行う場合、対立仮説の取り方によって検定の基準が 2 通りあります。前節で行った平均値の検定の場合、帰無仮説を $\mu = \mu_1$、対立仮説を $\mu \neq \mu_1$ としましたが、これには $\mu < \mu_1$ の場合もあれば、$\mu > \mu_1$ の場合もあります。有意水準を$\alpha$ とした場合、我々は図 9-4 のように t 分布の両端の確率の合計が$\alpha$ であると考えています。これを両側検定といいます。

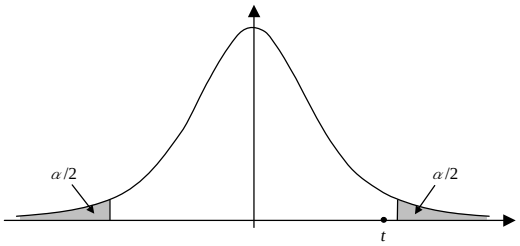


図 9-4 両側検定

これに対して、$\mu < \mu_1$が殆ど確実な場合、対立仮説を$\mu < \mu_1$とすることがあります。この場合、有意水準を$\alpha$とすると、図 9-5 のように t 分布の片側で確率$\alpha$と考えれば十分であると思われます。これを片側検定と言います。

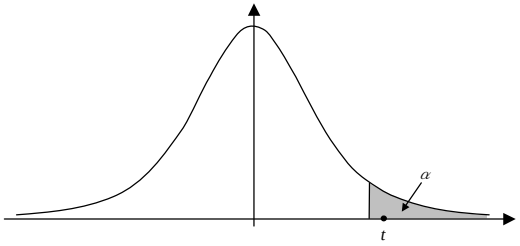


図 9-5 片側検定

両側検定と片側検定とでは検定の厳しさに差があります。図 9-4 と図 9-5 に検定に用いる統計量 $t$ の例を 1 つ描き加えてあります。これは同じ位置に描いているのですが、両側検定では棄却域の内側で、差があるとはいえないと判定され、片側検定では棄却域に入って、差があると判定されます。このことから、一般に両側検定は片側検定に比べて厳しい（差が出にくい）検定であるといえます。通常 t 分布や標準正規分布を用いる検定では両側検定を使うのが慣習になっています。また、$\chi^2$ 分布や F 分布を用いる検定では、統計量の定義域が正の値しか取らないことから、通常片側検定を用いることが慣習になっています。この教科書では、これらの慣習に従って検定を行うことにします。

## 9.5 検定選択ツリー

具体的な検定手法の説明に入る前に、全体を概観しておきましょう。検定手法はデータが質的か量的かによって大きく分かれ、さらに指定値との比較か、群間の比較かによって分かれます。また群間の比較の場合、データ間に対応があるかないかによって異なった検定手法を用います。ここでは質的データと量的データに分けて、検定手法をまとめておきましょう。

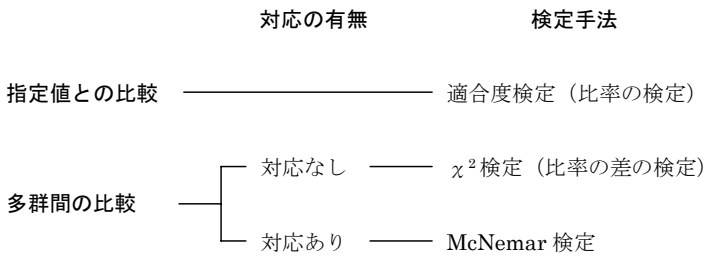


図 9-6 質的データの検定手法

図 9-6 は質的データの検定手法についてまとめたものです。まず、指定値（指定比率）との比較であるか群間の比較であるかについて検定手法が分かれます。指定値との比較の場合、適合度検定と呼ばれる検定手法を利用します。特に事象が「起きた・起きない」や、「はい・いいえ」などのように 2 つである場合、比率の差の検定と呼ばれる分かり易い方法も使えます。

群間の比較の場合、データの対応の有無によって検定手法が変わってきます。例えば、あるキャンペーンを行なった前後のアンケート調査のような場合、同じ対象について 2 度データを取りますから、2 つの標本間でデータに 1 対 1 の関係があります。これを対応と呼びます。また、ある要因の有無によって 2 つの標本を集める場合、精度を上げるために、性別や年齢ごとに対となるデータを選ぶようなことをします。これも対応がある例です。これに対して、別々に集めた男女間で意見の比率の差を見るような場合は、対応のない場合です。対応のない場合には 2 つの標本のデータ数が違っても問題ありません。

対応のない場合、比率の差の検定には $\chi^2$ 検定を用います。この検定は統計量として $\chi^2$ 分布に従うものを利用するために、この名前が付いています。上で述べた適合度検定も $\chi^2$ 分布に従う統計量を用いるので、$\chi^2$ 検定と呼ぶことがありますが、我々は両者を区別するために適合度検定の呼び名を使うことにします。適合度検定や群間の比較の $\chi^2$ 検定は、2 群以上の比較も可能です。

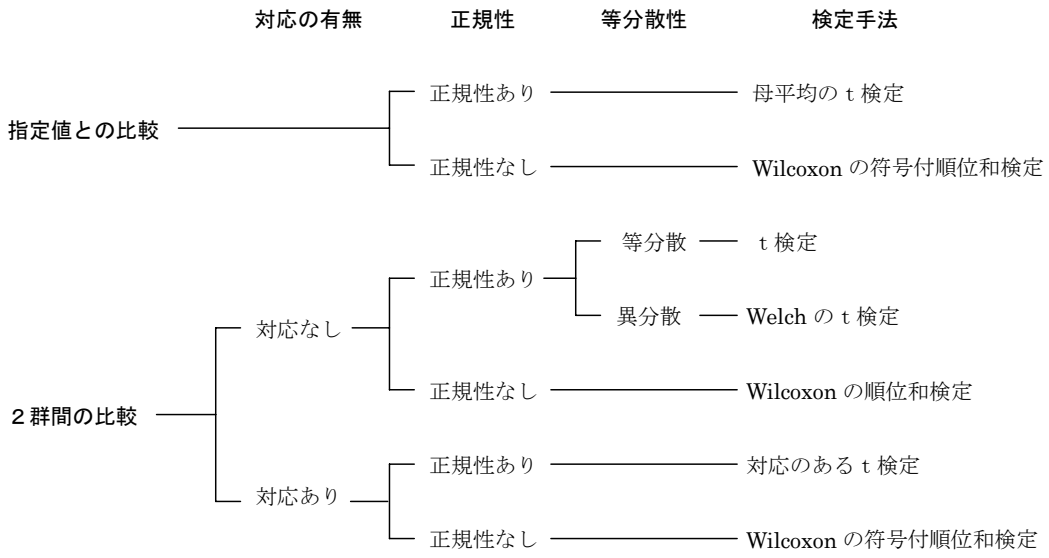


図 9-7 量的データの検定手法

図 9-7 は量的データについての検定手法のまとめです。検定は質的なものと同様に、指定値との比較か、群間の比較かに分かれます。質的データの場合、$\chi^2$ 検定は多群間でも利用できましたが、ここでは話を 2 群間に限ります。多群間の比較は分散分析と呼ばれる手法等を利用しますが、これは別の本に譲ります。

2 群間の比較の場合、対応の有無で検定手法が変わってくることは、質的データの場合と同じです。また、量的データの場合、データに正規性があるかどうかで検定方法が著しく異なってきます。正規性がある場合の検定方法は、パラメトリック検定と呼ばれ、前章で説明した正規分布から導かれる分布を利用します。それに対して、正規性のない場合は、ノンパラメトリック検定と呼ばれ、データの大きさの順位などを用いた処理を行います。通常、初心者用の教科書ではこのノンパラメトリック検定の話を省略してある場合が多いのですが、正規性の有無を確かめずに間違った検定を行ってしまう危険性もありますので、警鐘のために付け加えておくことにしました。

さて、図 9-7 に示した正規性の有無や、正規性がある場合の等分散性の確認方法ですが、これらにももちろん検定手法があります。特に等分散性については F 検定と呼ばれる検定手法で容易に確認することができます。ところが、正規性については、データ数が多い場合はヒストグラムで視覚的に確認できますが、データ数が少ない場合、統計ソフトが無いと少々厄介です。しかし、この検定のためには、昔から正規確率紙と呼ばれるグラフ用紙が売られています。とは言っても正規確率紙は普通の文房具店

にはありませんので、これに相当する処理を Excel にさせてやることにします。

この教科書では、これらの検定手法の他に、相関係数の検定や回帰分析の検定についても学びます。また、検定と裏腹にある母集団の推定と呼ばれる処理についても学びます。以下の章は、これらの検定や推定のマニュアルのような雰囲気を持っています。ただ、少しだけ検定に利用される式が導かれる経緯などもお話しようと思います。

複雑な式を覚える必要はありません。分からないときにはいつでもこの教科書を辞書のように参照して下さい。