２章　確率分布と検定【Skip OK】

2.1 確率密度関数

この章は３章以下の検定についての準備の章です。ある程度知識がある場合は飛ばしてもらっても結構です。

量的データの集計方法としてヒストグラムを描くということを前章で学びましたが、ここではデータの数を十分大きくしていった場合のヒストグラムの形を考えます。図のようにデータ数を多く取って行くと、ヒストグラムはきめが細かくなり、ヒストグラムの上端を繋いだグラフは次第に滑らかになって行き、ある形に近づいて行きます。この形を変数の確率密度関数または単に密度関数と呼びます。但し、縦軸の度数はこの関数が囲む面積が 1 になるような数値に書き換えています。

図 2.1.1 ヒストグラムの極限と密度関数

例えばテストの点数を考える場合、図のような位置に 40 点と 60 点があるとしましょう。確率密度関数のその間にある領域の面積を S とすると、S は 40 点から 60 点の間にいる人の割合になります。

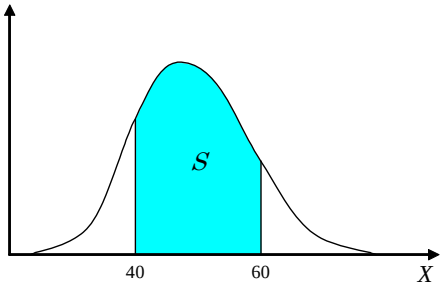


図 2.1.2 面積と確率

もう少し言い換えるとこの集団の中から一人選ぶとすると、40 点から 60 点の間の人を選ぶ確率は S になります。もちろん割合または確率ですので、この面積は 0 以上 1 以下です。またこれを 0 点から 100 点までとしますと、全員その中に入りますので、割

合または確率は 1 になります。確率密度関数が囲む全面積が 1 になるように縦軸を書き換えていると言った理由は全確率が 1 になるように設定するためです。

よく「ある変数の分布は？」という言い方をしますが、これはこの確率密度関数がどんな形かということです。以後代表的な分布について見て行きましょう。

2.2　正規分布と標準正規分布

統計学の基本的な分布は正規分布（normal distribution）と呼ばれる分布です。これはデータの測定誤差などランダムな現象から生じる分布です。また、正規分布以外の分布からでも、ある程度大きな数のデータの平均値は正規分布することが知られています。これは中心極限定理と言って統計学では大変重要な定理です。この節ではこの正規分布についてみて行きましょう。

正規分布は平均と分散で確率密度関数の形が完全に決まる分布です。ある量 X が平均 μ（ミュー）、分散 σ^2（シグマ^2）の正規分布をしていることを $X \sim N(\mu,\sigma^2)$ と表します。正規分布していることを正規分布に従う、とも言います。この正規分布の確率密度関数は以下のようにきれいな富士山形をしています。

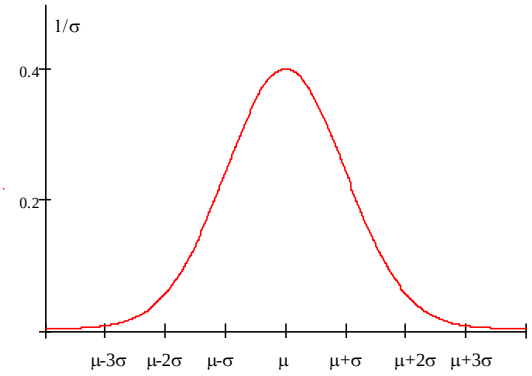


図 2.2.1 $N(\mu,\sigma^2)$ の確率密度関数

正規分布には平均と標準偏差の値がどのように変わってもこの形を維持するという面白い性質があります。すなわち平均と分散の値に関わらず以下の性質があることが示されています。

$P(\mu-\sigma \le X \le \mu+\sigma) = 0.683$　　外側　32%

$P(\mu-2\sigma \le X \le \mu+2\sigma) = 0.954$　　外側　5%

$P(\mu-3\sigma \le X \le \mu+3\sigma) = 0.997$　　外側　0.3%

ここに $P(a \le X \le b)$ は変数 X が a から b の値を取る確率を表します。例えば、ある試

験の結果が、平均 65 点、標準偏差 8 点の正規分布をする場合、一番上の式を利用すると、57（=65-8）点から 73（=65+8）点までの間に約 68%の人がいることになります。この値は非常に重要で、特に外側確率の 32%, 5%, 0.3%という概数は統計を学ぶ際にはぜひ覚えておくべきでしょう。

正規分布の中で特に重要なものがあります。それは平均 0、分散 1 の正規分布です。この分布の確率密度関数は以下の図のような形をしており、計算機が発達する前から細かい範囲で確率を与える表が作られていました。

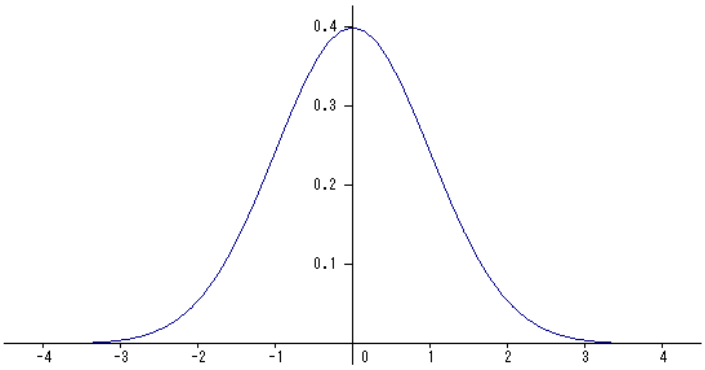


図 2.2.2 標準正規分布の確率密度関数

しかし、実際の分布が都合よく標準正規分布になるということはまずないのに、なぜ詳しく調べられたのでしょうか。それは正規分布に以下の性質があるからです。

$$X \sim N(\mu, \sigma^2) \quad \text{ならば、} \quad X' = \frac{X-\mu}{\sigma} \sim N(0,1)$$

これだけでは何のことか分からないと思いますので、前に使った試験の点数の分布 $N(65, 8^2)$ を例に上の式の意味を考えてみます。例えばこの試験で 70 点以上の人の確率（割合） $P(X \geq 70)$ を計算しようとしても、当然一般の正規分布では細かな確率は与えられていませんのでそのままでは不可能です。そこで上の式を利用して、変数 X を X' に変換します。この場合 $\mu = 65$ 、 $\sigma = 8$ ですから、変換の式は以下のようになります。

$$X' = \frac{X-65}{8} \sim N(0,1)$$

この式を使って $X \geq 70$ を X' の関係に直すと

$$X' = \frac{X - 65}{8} \geq \frac{70 - 65}{8} = 0.625$$

$X' \geq 0.625$ となります。この X' は標準正規分布に従うので、細かい確率も計算できます。すなわち結果は以下のようになります。

$$P(X \geq 70) = P(X' \geq 0.625) = 0.2660$$

このようにして標準正規分布の確率を詳しく求めるということは、一般の正規分布の確率を詳しく求めると同じであることが分かりました。

一般の分布の場合、上の変換は単に平均を 0 にして分散を 1 にする変換で、正規分布のときだけ変換後も正規分布であり続けるところが大きな特徴です。このようにして変数 X を変数 X' に変換することを標準化と呼び、後で学ぶ多変量解析ではよく利用されます。

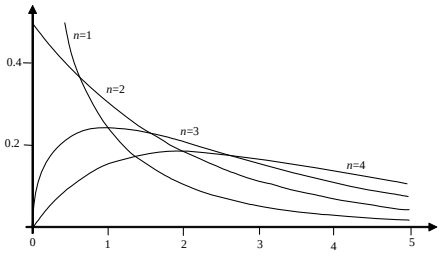
正規分布の性質は他にもいろいろありますが、興味のある人は「基礎からの統計学」を参照して下さい。

2.3 標準正規分布から導かれる分布

この後検定について話を始めますが、そのとき出てくる分布の名前は χ^2 分布、F 分布、t 分布の 3 つです。下にそれぞれの確率密度関数の形や定義式などを書いてありますが、その部分はほとんど知らなくても結構です。ただこれらの分布が標準正規分布を基礎にしていることとそれぞれの分布に自由度と呼ばれる整数のパラメータがあって、その値によって確率密度関数の形が変わるということは記憶しておいて下さい。各分布の自由度の数は、χ^2 分布で 1 つ、F 分布で 2 つ、t 分布では 1 つです。またこれらの分布は自由度を下に付けて以下の略号で表されます。

χ^2 分布 ： χ_n^2 ,　　F 分布： F_{n_1, n_2} ,　　t 分布： t_n

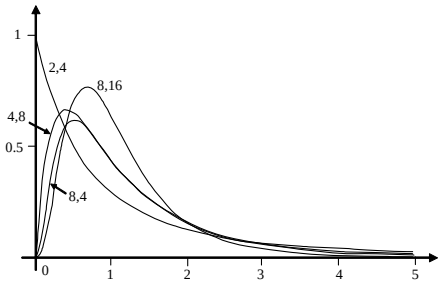
$\chi 2$ 分布



$X_i \sim N(0,1)$ 分布で独立なとき、

$\chi^2 = \sum_{i=1}^{n} X_i^2 \sim \chi_n^2$ 分布（自由度 n の χ^2 分布）

F 分布

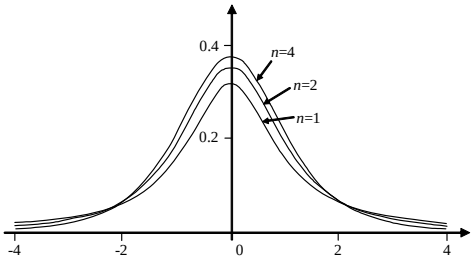


$\chi_1^2 \sim \chi_{n_1}^2$ 分布，$\chi_2^2 \sim \chi_{n_2}^2$ 分布で独立なとき、

$$F = \frac{\chi_1^2 / n_1}{\chi_2^2 / n_2} \sim F_{n_1, n_2} \text{ 分布}$$

（自由度 n_1, n_2 の F 分布）

t 分布



$X \sim N(0,1)$ 分布，$\chi^2 \sim \chi_n^2$ 分布で独立なとき、

$$t = \frac{X}{\sqrt{\chi^2 / n}} \sim t_n \text{ 分布（自由度 } n \text{ の t 分布）}$$

注） $t^2 \sim F_{1,n}$ 分布

注） $n \to \infty$ で $N(0,1)$ 分布

図 2.3.1 確率密度関数

以下の問題は分布の統計値と確率の関係を与えるものですが、Excel の関数を使った演習は「基礎からの統計学」で勉強してもらうとして、良く使うものは College Analysis でも計算できますので簡単に説明しておきます。ただ、現実にアンケートの分析などでこの機能を使うことはまずありません。

メニュー［分析－基本統計－分布と確率］を選択すると以下の画面が表示されます。

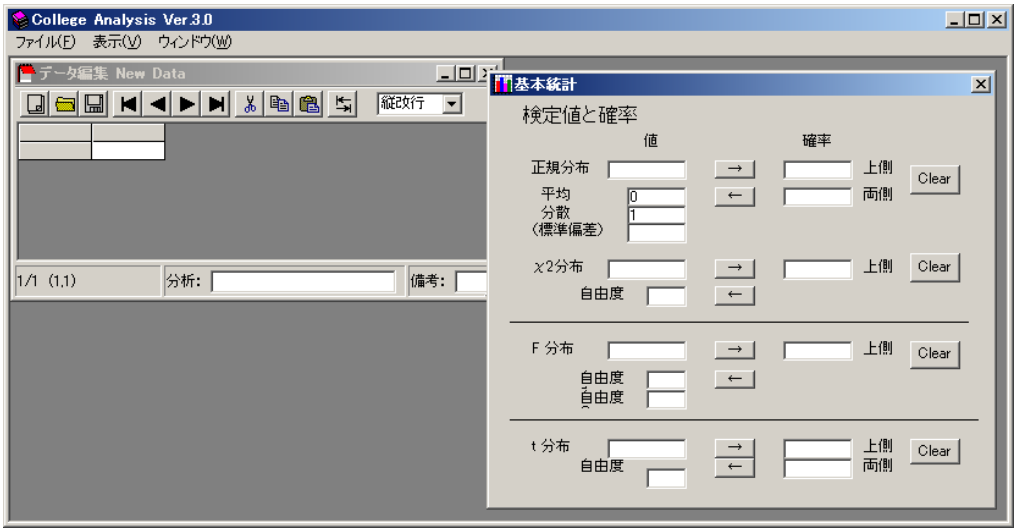


図 2.1.4 分布と確率画面

例えば試験成績の分布で $N(65, 8^2)$ の場合、$x \geq 74$ の確率を求めるには、正規分布の平均と分散（または標準偏差）の値を、分布名の横の値のところに 170 を書き込み、「→」ボタンをクリックすると以下のように上側確率と両側確率が表示されます。逆に上側確率か両側確率かを書き込み、「←」ボタンをクリックすると平均値より大きい側の検定値が表示されます。

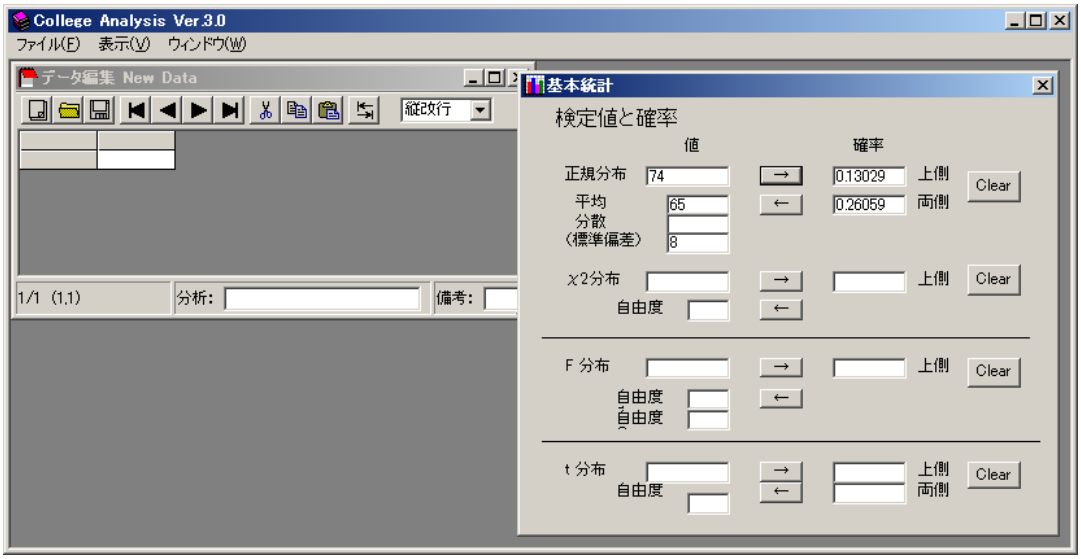


図 2.3.2 正規分布と検定確率

同様に χ^2 分布についても、検定値と自由度を書き込んで「→」ボタンをクリックすると上側確率が、逆に上側確率を書き込んで「←」ボタンをクリックすると検定値が表示されます。F 分布や t 分布についても同じです。

ここで上側確率と両側確率については正規分布を例に取ると以下のようになります。

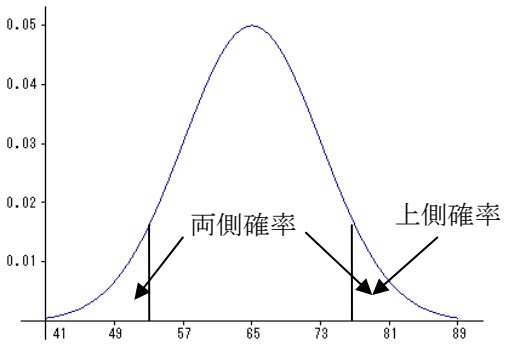


図 2.3.3 上側確率と両側確率

問題１　以下の値を求めよ。

１）$N(0,1)$ 分布，x 値 1.5 のときの上側確率 $p/2$　［　　］

２）$N(0,1)$ 分布，x 値 1.5 のときの両側確率 p　［　　］

３）$N(170,64)$ 分布，x 値 180 のときの上側確率 $p/2$　［　　］

４）χ^2_5 分布，χ^2 値 10 のときの上側確率 p　［　　］

５）χ^2_{10} 分布，上側確率 0.05 のときの χ^2 値　［　　］

６）$F_{8,4}$ 分布，F 値 10 のときの上側確率 p　［　　］

７）$F_{10,5}$ 分布，上側確率 0.05 のときの F 値　［　　］

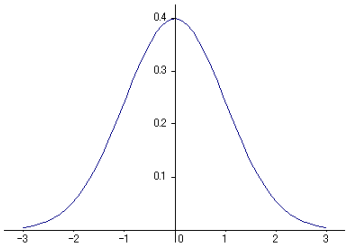
８）t_{10} 分布，t 値 2 のときの上側確率 $p/2$　［　　］

９）t_{10} 分布，t 値 2 のときの両側確率 p　［　　］

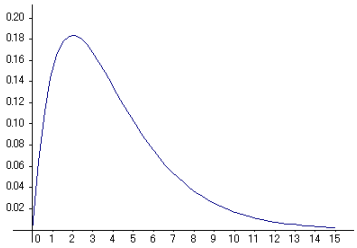
10）t_{10} 分布，両側確率 0.05 のときの t 値　［　　］

問題２　以下のグラフを描け。

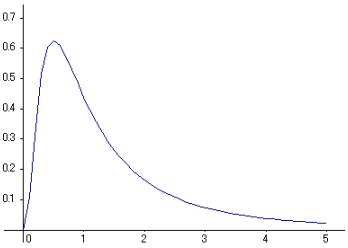
１）$N(0,1)$ 分布



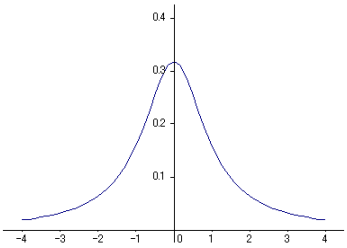
２）自由度 4 の χ^2 分布



３）自由度 8,4 の F 分布



４）自由度 1 の t 分布



2.4 検定の基礎

統計では調査したい対象を母集団と言います。しかし、例えば広島県の成人などとした場合、調査の費用は莫大かかりますので、実際は広島県の人の中から適当な方法で何人か選んで調べ、それを元に広島県の人全体を推測します。この実際に調査する対象を標本と言い、母集団の中から無作為抽出（ランダムサンプリング）によって選びます。図 2.4.1 はその概念図です。

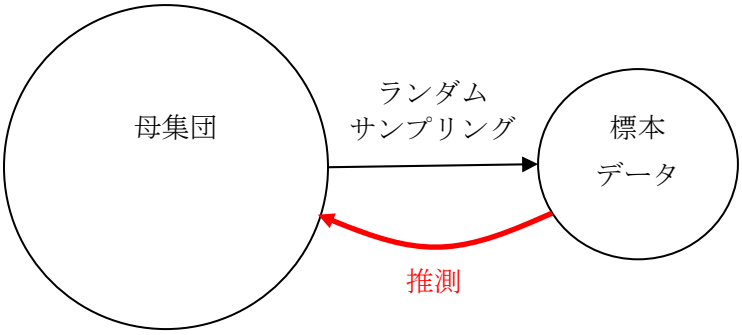


図 2.4.1 母集団と標本

この推測の方法を学ぶのが推定と検定です。考え方は推定が簡単ですが、適用範囲が狭く、実用では検定と呼ばれる方法がよく使われます。ここでは検定の考え方を簡単な例を使って説明します。

例

超能力を持つという人にコインの裏表を当てる実験をしてもらい、100 回の試行で 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。

有意水準（危険率）：超能力があると判定して間違う確率

この問題では 70% の正解率が確かに超能力によって起こったものか、偶然に起こったものかを判定します。

答えを得るには適合度検定または χ^2 検定という検定手法を利用します。多少説明不足のところがありますが、直感を優先させるためにご容赦下さい。

まず我々はこの 70%の正解率は全くの偶然であるという仮説を考えます。この仮説を統計用語で帰無仮説と呼びます。さて、この帰無仮説の下では、裏表が当たる確率も外れる確率も 0.5 です。そうすると試行回数が 100 回ですから、偶然当たる回数の予測値は 50 回で、外れる回数の予測値も 50 回になります。これらの回数を使って、以下の式 χ^2 を考えます。

$$\chi^2 = \frac{(\text{当たった回数} - \text{当たる予測値})^2}{\text{当たる予測値}} + \frac{(\text{外れた回数} - \text{外れる予測値})^2}{\text{外れる予測値}}$$

このように定義した χ^2 は、統計学者により自由度 1 の χ^2 分布に従うことが示され

ています。実際にこの値を計算してみましょう。

$$\chi^2 = \frac{(70-50)^2}{50} + \frac{(30-50)^2}{50} = 2 \times \frac{400}{50} = 16 \quad (\sim \chi_1^2)$$

この値の 16 はどのような数値でしょうか。以下に自由度 1 の χ^2 分布の確率密度関数のグラフを描いてみます。

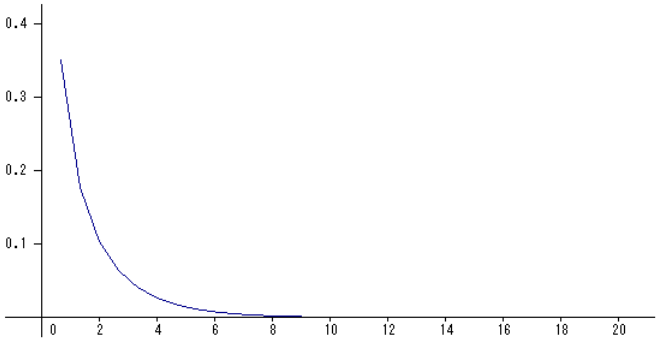


図 2.4.2 自由度 1 の χ^2 分布

この横軸の 16 のところがこの数値で、χ^2 の値が 16 以上の領域（グラフと横軸の間の面積）が、偶然に 70 回以上当たる確率です。図 2.4.2 では横軸に張り付いてこの領域は見えませんが実際にその面積を求めると $p = 0.00006$ （誤差は大きいでしょうが）になります。例えばこの値は Excel で =chidist(16, 1) として求めることができます。

この確率から考えて、70 回当たることは珍しいことでしょうか？この確率は 10 万回に 6 回のことですから、ほとんどの人が珍しいと答えるでしょう。ではこの珍しいことが偶然この場合に起こったと考えるべきでしょうか。やはりそうは思えません。ではどこに問題があるのでしょうか。

我々は最初、70% の正解率は偶然によるという帰無仮説を考えました。そう考えるから $p = 0.00006$ という結果が出てきましたので、どうやらこの帰無仮説に問題がありそうです。実はこの帰無仮説が正しくなく、実際に何らかの超能力があると考えればこのような問題は生じません。そこでこの帰無仮説を捨てて、この人には超能力がある、すなわち当てる確率は 0.5 ではないと考えるべきでしょう。後者の仮説は対立仮説と呼ばれます。これは統計用語で言うと、帰無仮説を棄却し、対立仮説を採択するとなります。この例題の答えは「何らかの超能力がある」になりました。

さてこの理論にはもうひとつ問題が残っています。それは珍しい、珍しくないを決める際に人の感覚を用いているところです。統計学者はこれを排除するために、ある確率を考えました。1 つは最もよく利用される 0.05 という数値です。これを用いて、

計算した結果が $p < 0.05$ のとき珍しいと判断して帰無仮説を棄却し、$p \geq 0.05$ の場合は偶然の可能性が高いとしてそのまま帰無仮説を採択する方法が採られます。この 0.05 という数値に根拠はありませんが、これは超能力があると言ったときの間違える確率でもありますから、まあ 100 回に 5 回くらいは容認しようといったところです。珍しいかどうかを判断する基準は他に 0.01 がよく使われ、もっと厳密に判断したいという場合に使われます。これらの 0.05 や 0.01 の確率は有意水準と呼ばれます。この例題の場合はもちろん $p < 0.05$ ですから、5% の有意水準で超能力があるといえるとなります。

試行回数が 20 回の場合、以下のような結論になります。

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

$p = 0.07364 > 0.05$ より、超能力があるといえない。

これらの結論からデータ数が多いほど差が見えてくるということが分かります。この検定方法は適合度検定または χ^2 分布の性質を利用することから、単純に χ^2 検定と呼びます。

2.5 検定の形式

検定には大きく分けて２つの比較方法があります。１つは以下の図のように 1 つの母集団から 1 つの標本を取り出し、それを使って母集団の平均などの統計量をある指定値と比較する検定です。

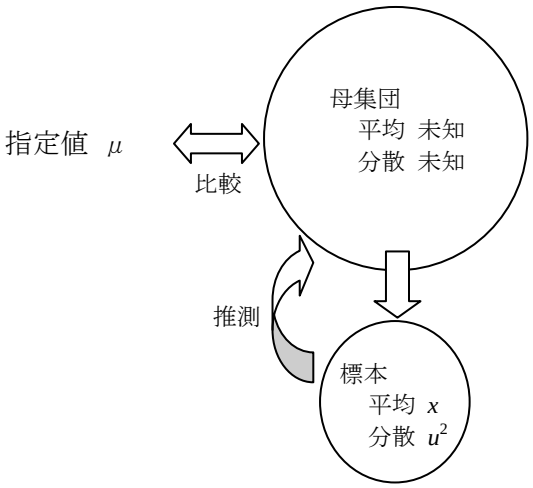


図 2.5.1 母集団の統計量と指定値との比較

先に例で述べた場合もこの群に入り、母集団は無限個の成功と失敗の事例の集合でその中から 100 個標本をとり、比較するものは母集団の成功比率と指定比率 0.5 です。

2 つ目の方法は以下の図のように 2 つの母集団間の同じ統計量同士の比較です。

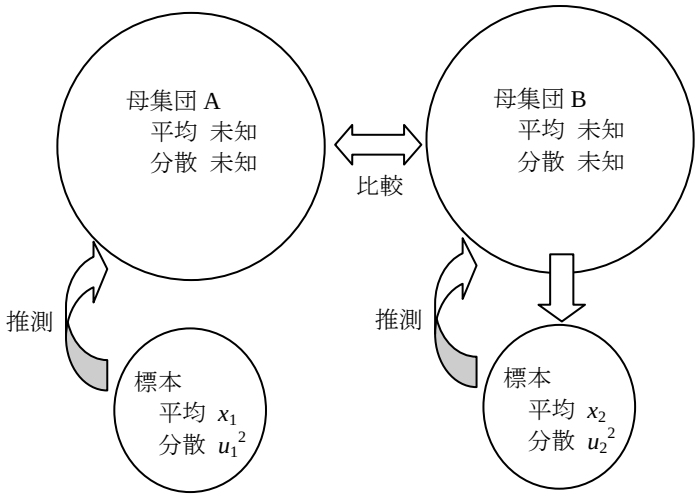


図 2.5.2 母集団間の統計量の比較

これには例えば広島県と岡山県の共通模試の平均点の比較や男女別にみたアンケートの賛成比率の比較など、現実の調査で知りたい多くの内容が含まれています。またこの比較ではあるダイエット食品の使用前の体重と使用後の体重の比較のように、1 人の人が両方の母集団に含まれているような場合もあります。

2.6 検定選択ツリー

この節では今後我々が学ぶ検定手法についてまとめておきます。右端の x x 検定と名前が付いているところが選択すべき検定手法で、そこに到達するまでに、例えば量的なデータでは、対応の有無、正規性の有無、等分散性の有無で分けられて行きます。この流れが理解できるようになればこの本の目的はほぼ達成されたことになります。

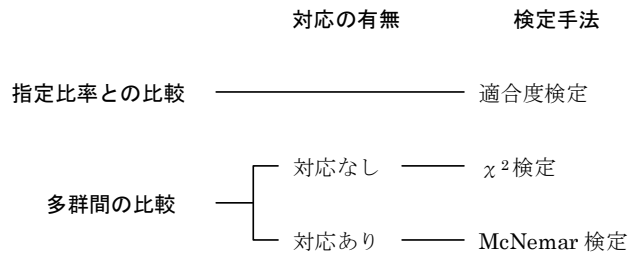


図 2.6.1 質的データの検定手法

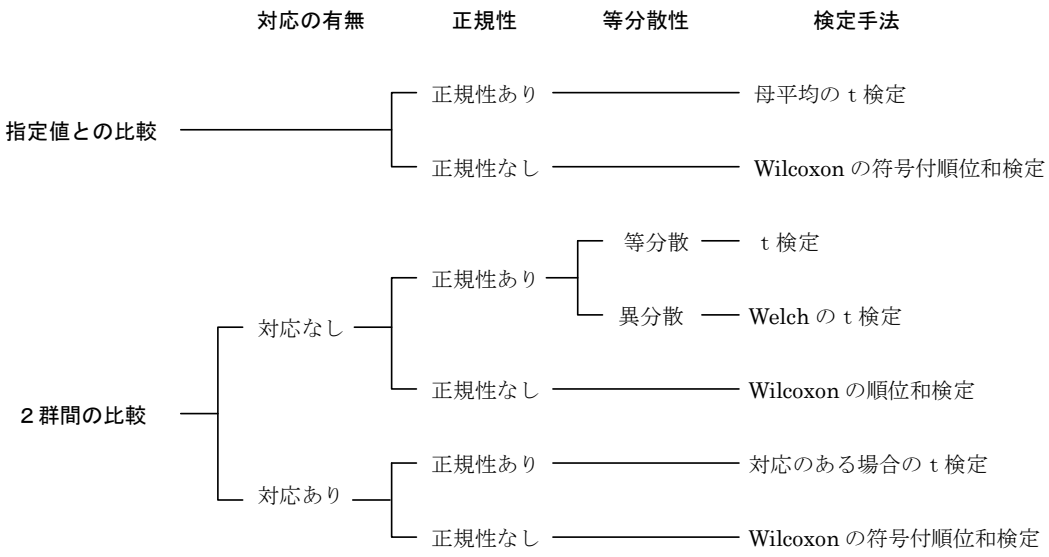


図 2.6.2 量的データの検定手法

以後、これらの検定を詳細に見て行きます。