**統計学のススメ**



## １．はじめに

人の感覚は十人十色で、感じ方はそれぞれ異なります。例えば、あるやせ薬を服用して1kg痩せたとします。この結果から、痩せる効果があると思う人もいれば、偶然に過ぎないと思う人もいます。もちろん個体差もあるので判断するのは容易ではありません。そこで、生物統計学という専門的な処理によって結果を導き出します。

これから説明するのは、ちょっとした統計処理をするのに是非知っておきたい生物統計学です。特に統計学の分厚い本に圧倒されてなかなか専門書を開けないという人にオススメの内容です。また、数字を扱うことが多い社会人にとっても知っておいて損はないと思います。私自身、教職関係者ということでそのような事例もいくつか取り上げますが参考にしてみて下さい。

## ２．基礎知識

はじめに統計処理を行うためにあたり最低限知っておかなければならない、あるいは知っておくとよいキーワードを紹介します。

### １）母集団と標本

例えば、テレビの視聴率を例にとります。ご存じの通り視聴率とはテレビなどの放送番組の見たり聞いたりされる割合を意味します。しかし、実際には日本全国の全世帯のテレビの受信状況を調査しているわけではありません。詳しいことは私もわかりませんが、調査対象となる世帯が各地域にあり、その受信状況から視聴率を割り出しているとか･･･。ということならば、視聴率とは絶対的な数字では無く、推測ということになります。間違いのない視聴率を調査したいのならば理論的に全世帯のテレビの受信状況を調べれば良いのでしょうが、それは無理です。ゆえに日本全体の世帯からいくつか抽出して使用状況を割り出し、日本全国の視聴率を推測しているのです。

このように、あることを調べようとしても数が多過ぎて調べきれないときは、いくつかデータを抽出して全体を推測します。このとき、膨大なデータを「母集団」、そこから抽出されたデータを「標本」と呼びます。視聴率の例でいうと、全世帯のテレビが母集団、実際に調査されているテレビが標本ということになります。

細菌を扱った実験では地球に存在する細菌に実験を施せるものでもなく、また、薬の効果を調べるために全国の人に試してみることもできませんので、ほとんどの場合、実験データは膨大な母集団から抽出された標本を統計処理して結果を「推測」していることになります。

### ２）平均

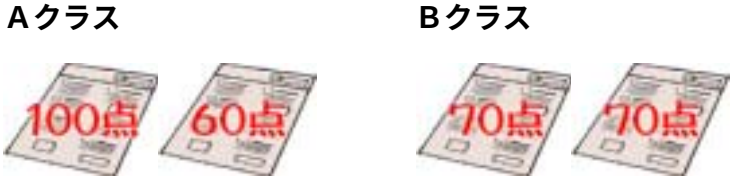
私たちが普段使っている平均とは「算術平均値」のことです。データを全部足して標本数で割るというもので、私たちが最も一般的に使っているものです。

> 例）10, 12, 14, 16の算術平均値
> (10＋12＋15＋16)／4＝13

しかし、使う機会はほとんどありませんが、「調和平均」や「幾何平均」などという平均値の求め方も存在します。もちろん計算の仕方が異なるので、算術平均とは微妙に異なる結果を出したりします。分野（標本の数値が著しく異なる場合や経済的, 時間的計算など）によって相応しい計算の仕方を選択しますが、私たちが平均といえば算術平均のことを指しています。

### ３）分散と標準偏差

例えばAとBという2つのクラスがあったとします。Aのクラスは平均点が80点で、Bのクラスは平均点が70点だったとします。




みなさんはどちらのクラスが優秀だと思いますか？単純に比較してしまえばAのクラスと言うでしょう。しかし、例えばAのクラスのうち半分が100点で残り半分が60点という構成から平均して80点となり、一方のBの

クラスは全員が70点だったとすると、一長一短があり、どちらのクラスが優秀かというのは甲乙をつけがたいですよね。

このようにデータというのは単に平均値だけでは判断できません。そこで全体の様子を知る手がかりとして「分散」や「標準偏差」などを用います。簡単に言うと、分散も標準偏差も平均値を中心にどれくらいばらついた位置に標本が分布しているかを示し、数字が小さいほどばらつきが無いことを意味します。ちなみに異なるグループ間のバラツキを比較するときは「変異係数」を利用しますが、この変異係数は標準偏差をもとに算出します。

分散は、それぞれのデータから標本の平均値を引き二乗したものを合計すれば求められます。また、分散の平方根を標準偏差といいます。(標準偏差$^2$＝分散)

例) 5, 7, 15の分散
平均値＝(5＋7＋15)／3＝**9**
分散＝{(5－**9**)$^2$＋(7－**9**)$^2$＋(15－**9**)$^2$}／3
18.67
標準偏差　4.32 (分散の平方根)

上記のBクラスを例にとります。

全員が70点だったという場合は、各個人の70点から平均点である70点を引いて2乗してそれらを全部合計するのですが、70点から平均の70点を引いて二乗しても0にしかなりませんので、分散も、さらにその平方根である標準偏差も0になります。Bクラスのように全員が同じ70点である場合はバラツキもありませんので、分散も標準偏差も0となります。ちなみにそれらがマイナスになることはありません。

(学歴主義に欠かせないテストの偏差値と、標準偏差は算出の仕方が異なり別のものですが、バラツキを考慮しているという点では目的は似ています。)

さて、自分で何かの集計表を作る場合、平均値を記入するのは簡単です。計算器を使えばよいし、表計算ソフト(Excel など)を使うのならば「=average( )」と関数を打てば良いわけです。実は分散も標準偏差もExcel を使えば平均値と同じ労力で求められます。

=average( ) ･･･平均値を求める。
=var( ) ･･･不偏分散を求める。通常、分散は不偏分散のことを指すので、これを利用する。
=varp( ) ･･･母分散を求める。ただし、これを利用することは少ない。
=stdev( ) ･･･標準偏差を求める。通常はこれを使用する。
=stdevp( ) ･･･母標準偏差を求める。ただし、これを利用することは少ない。

varとvarpはどちらも分散を求めます。varの後ろに付いている「p」は母数(Parameters)のことで、調査する対象が「母集団」という意味です(Excel の開発者に聞いたわけではありませんが、間違いないでしょう)。前の項の「母集団と標本」で記述したとおり、ほとんどの調査対象が母集団から抽出した標本であるので、通常はvarの「不偏分散」を使えば問題ないでしょう。また、十分な標本数であれば、不偏分散　も母分散とな

りますので、学術論文などの厳密さを要するものでない限り、どちらを選択するかで悩む必要はないと思います。

私がデータを表にして提示するときの例を示します。特に決まりはありませんが、参考にして下さい。

例）

| | 平均値 | 標準偏差 |
|---|---|---|
| Aグループ | 55.3 | 7.3 |
| Bグループ | 48.2 | 16.5 |

上のような資料から次のようなことを推測することができます。

Aグループは、とりわけ高い値や低い値もなく平均的である。
一方、Bグループの方は標準偏差が高いため、
①高い数値と低い数値の差が著しく大きい。
②全体的に数値がAグループよりも低い中で著しく高い値がごくわずか存在する。
③全体的にAグループと似通った数値の中で、著しく低い数値がわずかに存在する。
などの可能性がある。この他にもさまざまなパターンが考えられるので、分布を確認する必要がある。

### 4）自由度

統計処理で頻繁に登場するのがこの「自由度」です。自由度とはこれから処理しようとする標本数から「1」を引いた数です（データ数が10個だとすると、10個から「1」を引くので自由度は9になります）、と言いながら私が最も理解しにくいところがここです。様々な文献を見ると、

『ある3つの標本があったとき、2個のデータは選べるが、1個は自由に選べないため』
『平均値がわかれば、n個の試料のうちn－1個の値がわかると、残り1個の値も自動的に決まる』
『さまざまな変量に関与している要因を標本数から引くため』

などと記してあります。

私は統計学を専門的に学んできたわけではないので、どうしてもここのところが理解できないのですが、勝手に自分で考えをめぐらせた結果を述べると、

統計処理を行うためには母集団全体のデータを扱うことが望ましいのは先に述べたとおりですが、ほとんどの場合は母集団から一部のデータを抽出した標本で統計を行うことになります。ある母集団から標本をランダムに抽出するとします。このとき、抽出される標本数は何個でも良いのでしょうが、標本数は多いに越したことはありません。しかし、母集団から全部のデータを標本として抽出してしまっては、抽出する意味がありません（全部抽出するということは母集団で統計処理をし、抽出はしていない、ということになりますから）。したがって標本数として自由に抽出できる数は、母集団よりも1つ少ない数になると考えられます。同様に、ある母集団が3つだった場合、これから統計処理の対象となる「標本」として自由に抽出できる数は2つとなり、先の『　』の記述の通りになるのではないかと考えました。

あくまでもこれは私の推測です。まあ、理解できていなくても自由度の出し方を分かっていればそれほど困ることでもありませんが･･･。

サンプル数が少ない場合、母集団の数と自由度は多少異なりますが、母集団が大きくなればなるほど母集団と自由度は近づきます。すなわちサンプル数が多い方が、より確かなデータになりやすいということです。

例)
①のケース　母集団が3のとき、自由度は2(3－**1**)
②のケース　母集団が100のとき、自由度は99(100－**1**)
母集団と自由度の比率(母集団÷自由度)を求めると、②のケースの方が比率が1に近く、母集団と自由度が近い値になることがわかる。

### 5）母分散と不偏分散

先に説明したように、分散には「母分散」と「不偏分散」の2種類があります。例) 5, 7, 15の分散の例題で示した分散は、

$$\{(5-9)^2+(7-9)^2+(15-9)^2\}\div 3$$

のように最後に母集団の数「3」で割っています。これを母分散といいます。しかし、母集団で統計処理を行うことはまれで、数多いサンプルから標本を抽出して得たデータの場合は、

$$\{(5-9)^2+(7-9)^2+(15-9)^2\}\div(3-1)$$

のように、自由度2(標本数－1)で割ります。これを不偏分散といいます。特に断りがない場合、分散は不偏分散のことを示します。当然ながら、標本数が多ければ多いほど母分散と不偏分散は近似します。

### 6）正規分布

例えばクラスで試験などを行った場合、ほとんどの生徒がクラス平均値付近に集中します。平均値から離れるほど、すなわち、高得点または低得点であるほどその数は少なくなってきます。この分布の仕方は、試験の教科や難易度によっても多少異なりますが、特に大きな要因が働かなければおよそ富士山のような分布になります。統計学では、おおよそどの分布でも当てはまるように「標準正規分布図」というものが用意されています。それが次の通りです。

①平均値±標準偏差の範囲に全個体数の68.26%が分布している。
②平均値±2×標準偏差の範囲に全個体数の95.44%が分布している。
③平均値±3×標準偏差の範囲に全個体数の99.74%が分布している。

もしこのような標準正規分布からまったくはずれた分布をしている場合、実験をする前に何らかの作用があったと考えることができます。したがって、実験や統計処理の対象としてはふさわしくありません。

例えば定期試験などで、簡単過ぎる問題で満点者が続出したり、まったく勉強をせずに受験する生徒が

多数いたり、欠席などでまともに授業を受けていない生徒が多数いた場合などに正規分布をとらないケースが考えられます。

---

## ３．検定に入る前に

検定を用いて、「○○の方法を取り入れれば学習効率は2倍に高まる」「この薬には△△の作用が認められる」などという結論に結びつけていきますが、もちろん計算だけではそれらを求められるわけではありません。検定では「ある」「ない」を判断するだけなので、始めに実験者が仮説を立てなくてはなりません。そして、その仮説の可能性や信憑性の「ある」「ない」を統計処理によって判断します。

### １）仮説の立て方（帰無仮説と対立仮説）

AというグループとBというグループのそれぞれの平均値に、実験によって生じる差（人によって表現の仕方が多少異なることもありますが、ここでは「有意差」と表現します）の有無を証明するためには仮説を立てる必要があります。

仮説の立て方には、「対立仮説」と「帰無仮説」という2種類の方法があります。

| |
|---|
| ①対立仮説・・・『AとBの平均値に有意差がある』と仮定してその確率を求める方法。<br>②帰無仮説・・・『AとBの平均値に有意差はない』と仮定してその確率を求める方法。 |

一般に統計処理の場合、②の帰無仮説が用いられます。『AとBの平均値に有意差はない』という実験者の意に反した仮説をたて、統計処理によってそれを否定することで有意差を明らかにする手法です。

例えば、部活動を例にとりましょう。野球の打率において、素振りの練習量を2倍にして生徒の打率を他の生徒よりも高めようとしたときに、本当に練習量を増やした生徒の打率が上がったかどうかを調べるとします。通常練習量のグループをA群, 練習量2倍のグループをB群とします。

ここで帰無仮説を立てます。指導者としては練習量と打率が比例して欲しいでしょうからB群の打率が上がることを期待しますが、あえて『A群とB群に差はない』と仮定します。この仮説が検定によって否定されれば指導者の期待する結果となりますが、仮説を否定できないという結果であれば練習量は打率に影響しないとなります。

統計処理は次の①～④の手順で行います。

| |
|---|
| ①仮説の設定<br>②検定法に従って検定値を算出<br>③検定値（数字）とP値（専門書には必ず一覧表がのっておりそこから該当する数値を見つける）を比較する。<br>④”$P<0.05$”，”$P>0.05$”という形で結果がでるので、それを“○○と××には有意差が認められる”などのように文章にする。 |

ここでは検定の過程は省略しますが、例えば仮説の可能性が $P<0.05$（詳細は次のP値の項参照）という

結果になったとすれば、『A群とB群に差がある』と結論付けられます。この”$P$<0.05”はおよそ確率のことで、『A群とB群に差はない、という可能性は5％未満』ということです。文章に直すと、

『A群とB群に差がない可能性は5％未満である』

＝『A群とB群に差がある可能性は95％以上である』

・・・『練習量を2倍にして成績が良くなる可能性は95％以上である。

と類推できます。逆に”$P$>0.05”になってしまった場合、『A群とB群に差はない、という可能性は5％以上』となり、仮説を捨てることはできず、結論は仮説の通りになってしまいます。

### 2）P値

前項で”$P$<0.05”というものが登場しましたが、これが「P値」です。$P$ とは Probability（確率）の略です。『A群の人とB群の人に差はない』という帰無仮説では、$P$<0.05 というのは100人中、5人未満は差がないということで、A群の人とB群の人に差はないことは珍しいことであるの意味です。一般的に $P$<0.05 のことを「5％水準」と呼び、この水準で判断します。もし、AB間に極めて差がある場合は1％水準（$P$<0.01）で検定し、『A群とB群に差がないことは100回に1回未満で、極めて珍しいことである』と力説しても構いません。

$P$についてはいくつか原則があります。

①P値の「P」は、国際基準ではイタリックの大文字で書く。
（人によって、あるいは科学雑誌によってはその通りになっていないこともある）
②P値を不等号で書くよりも実際の値を書くことが望ましい。
③P値の桁数は通常小数点以下第3位（$P$=0.048）で表すが、著しく有意の時には $P$<0.0001 でも良い。

### 3）片側検定と両側検定

例えばある学校で、テストの成績が芳しくないことから、補習を行い成績の向上を目指すとします。補習の効果があるかどうかを調べるためには、ほったらかしグループと補習グループの2つを作って試験の結果を見て点数に有意差があるかどうかを調べればよいわけです。当然指導者としては補習をすることによって学力アップを望んでいるわけですが、もしかすると補習をすることによって学力が低下する可能性がないとも言えません（実際に補習をして学力が下がったらシャレにならないですけどね）。

このとき、補習によって学力がアップすることを前提に有意差を調べることを「片側検定」といいます。それとは異なり、補習によって学力がアップする可能性とダウンする可能性の両方を含めて有意差を調べることを「両側検定」といいます。

一方的に補習＝学力アップと決めつけることは統計上望ましくありませんので、ほとんどの場合は両側検定を行いますし、以下は両側検定を前提に記述します。

### 4）棄却検定

正しい操作や環境で実験を行っても、ときに著しくかけ離れたデータが発生するときがあります。そのようなデータを破棄して良いものかどうかを検討する方法が「棄却検定」です。

代表的なものに、「THOMPSONの棄却検定」,「SMIRNOVの棄却検定」などがあります。また、棄却で

きる上限と下限を決めて棄却を行うという「増山の棄却限界判定法」などがあります。どれを選択して利用するかはそれほど重要でないと思います。また、詳しい計算の仕方については専門書を参考にして下さい。

ただし、怪しいデータが発生するときはむやみに棄却検定を行うのではなく、実験の性質や実験の操作手順に誤りがないかなどを見直すべきでしょう。また、実験は一度行えばよいというものではなく、当人が何度実験を行っても常に同じ結果になる、あるいは他の人が別の実験室などで行っても同じ結果がでることが大切です。これを実験の「再現性」といい、再現性が計れない実験は無意味です。

---

## ４．検定の方法

統計学の本を手に取ると、なんだか分からないことばっかり大量に記述されていることがほとんどです。なぜ大量に書いてあるのかというと、目的や方法によって統計処理の方法が異なり、それだけ多くの統計手法が存在するからです。例えば『2つのグループの平均値に差があるのか』と『3つのグループの平均値に差があるのか』というのは同じ比較であっても、グループ数の違いによって生じる誤差をできるだけ少なくするために、同じ手法で統計処理してはいけません。それに見合った方法を選択する必要があります。以下に代表的な事例を挙げ、適当な統計処理の方法を紹介します。どのような検定法でも手順にしたがって平均値や分散などを算出したり、P値などの表と見比べたりと、ほとんどが機械的な作業なので、計算方法は専門書を参考にして下さい。

### 例題1　2つのグループの比較

人間の左脳は論理的思考, 右脳は芸術的感性に優れているという。人間の神経は脳から脊髄に達するまでに延髄で交叉しているため、右利きの人は左脳、左利きの人は右脳に特に刺激を与えていることになる。したがって右利きの人は数学の点数が高くなると言えるかもしれない。本当にそうなのか調査したい。データは次の通りである。

| | | 平均 | 標本数 | 標準偏差 |
|---|---|---|---|---|
| 右利き | 80 95 63 73 62 88 42 92 51 63 | 70.9 | 10 | 16.85 |
| 左利き | 63 82 53 46 | 61.0 | 4 | 13.55 |

具体的な統計処理に入る前に、まず、その標本が統計処理をするのに相応しいかどうかを判断しなくてはなりません。そのため、はじめに正規分布, 分散比のチェックを行います。

正規分布のチェックでは、Shapiro-wilk 検定により統計処理に相応しい分布かどうかを判断します。

次に分散比のチェックでは、標本が同一の環境または管理条件下にあったかどうかを分散比(F)によって調べます。これをF検定と言います。分散比は二つのグループの分散の比率で求めます。分散比＝分散／分散で求め、グループの種類にかかわらず分散の大きい方が分子になるので必ず分散比は1以上となります。同一の環境, 管理条件下にあった標本の場合、実験を施したとしても大きく分散が異なることは考えにくく、もし分散比が余りにも大きいとすれば、異なる母集団から得た標本であるか、または、何らかの要因があって純粋に目的の実験が行われたとは考えにくくなります。分散比がおよそ P=0.001 の値よりも大きい場合、P<0.001 と表記し、その後の統計処理を行うことは無意義となります。もし分散比が P<0.001、または、正規分布と見なせないという結果がでた場合でも、別の統計処理の手法(不等分散のt検定)が考えられていま

すのでがっかりしないで下さい。

F検定によって同一母集団に属していると結論づけられたら、実際に平均値に差があるのかどうかを調べます。使用される検定はt検定です。ただし、t検定にもいくつかありますので、必ず適当なt検定を選んで下さい。種類は次の通りです。

①関連するデータのt検定(Paired t-test)

例えば、ある個体の実験前と実験後のデータなどです。2つのデータが同一個体からのものであるとか、何らかの対応がある場合です。当然ながら比較するそれぞれのグループの標本数は同一になります。

②独立したデータで、標本数が同じときのt検定

例えば、A群には水だけ、B群には薬剤を添加して実験したというように、両群のおおもとの標本に対応がなく、標本数が同じ場合です。

③独立したデータで、標本数が異なるときのt検定

④不等分散のt検定

F検定によって同一母集団のデータではないと判断されたデータのt検定です。

いずれの場合でもt検定で有意性がなかった場合は、有意差無しとしてあきらめて下さい。

**例題1の統計処理の手順**

ⅰ) Shapiro-wilk 検定(正規分布のチェック)

Shapiro-wilk 検定により　右利き 0.9502　　左利き 0.9500

５%水準の値　　右利き 0.842　　左利き 0.748

…それぞれ5%水準値よりも大きいので、右利きも左利きも正規分布とみて問題ない。

(Shapiro-wilk 検定の結果と5%水準値を比較した結果)

ⅱ) F 検定(きちんとした実験環境で実験されたデータであるかを大まかにチェック)

右利きの分散　315.66

左利きの分散　244.67

分散比　1.290　　5%水準値　8.812

…分散比が5%水準値よりも小さいので安全に同一母集団に属していると考えられる。

ⅲ) t 検定(有意差の有無を調査)

対応のないt検定の結果、5%水準で有意差が見出せない。

したがって、『右利きの人は左利きの人よりも数学の点数が高くなることはない』という仮説は捨てられない。最終的な結論は次の通り。

**「右利きも左利きも点数に差はない。」**

（これは例題です。実際に調査したわけではありません）

私は、実際のところ、ⅰ）の正規分布のチェックなどを必ずしているという人を見かけたことはありません。だからと言ってしなくて良いと断言できるものでもありません。以上の手順は丁寧な方法と思って下さい。

また、ⅱ）とⅲ）の検定は、Excel の関数で処理できます。

=Ftest(　) ･･･F 検定。
=Ttest(　) ･･･t 検定。対応や標本数などによって複数あるので、適当な引数を当てる。
※使用に際してはエクセルの統計関数ヘルプをご覧下さい。

**例題2　2つのグループのアンケート結果の比較**

ある2つのカレー専門店（ヒーヒーハウスとピリ辛工房）の激辛カレーを購入し、テイクアウトしてきた。これを8名の人に試食してもらい、カレーの激辛度を5段階評価してもらった。回答してもらった評価の5は最も辛いことを表す。回答結果ではヒーヒーハウスの点数の方が高かったが、果たしてヒーヒーハウスの方が“辛い”と言い切れるのだろうか？

（もちろん先入観を防ぐために試食時は店名などをわからないようにしてある。）

| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| ヒーヒーハウス | 5 | 3 | 4 | 5 | 4 | 3 | 2 | 4 |
| ピリ辛工房 | 3 | 1 | 4 | 1 | 3 | 3 | 3 | 1 |

この例では、アンケートの結果にある数的データは絶対ではなく、その順番が大きな意味を持っています。例題中のAとBのように、Aの人はピリ辛工房を3とし、それよりもヒーヒーハウスの方が辛いと思って5にしたとします。一方Bの人はヒーヒーハウスを3とし、ピリ辛工房はそれよりも辛くはないので1としたかもしれません。このような場合AとBに点数の違いはありますが、両者ともヒーヒーハウスの方が辛いということについては一致しています。したがって点数よりもどちらの方が上かという順位に注目すべきです。

データの最大値と最小値がまったく見当がつかず、数的データがその順番に意味を持つ、あるいは、正規分布を取らないようなデータについては、データを小さい順に並べ替えて順位データにしてから検定を行う方が良いのです。このように一度順位データに並べ替えてから行う検定をノンパラメトリック検定といいます。逆に、数的データをそのまま扱うような検定をパラメトリック検定といいます。

順位付けの方法は次の通りです。注意したいのは、同じ値のデータがあった場合は平均順位にするということです。

Excel で順位を出す場合は「=rank(　)」と入力します。ただし、同じ順位のデータは平均順位ではないので注意して下さい。

例）

| | | | | | | |
|---|---|---|---|---|---|---|
| データ | 72.5 | 75.1 | 76.3 | 76.3 | 80.1 | 81.9 |
| 順位 | 1 | 2 | **3.5** | **3.5** | 5 | 6 |

・・・(3＋4) /2 = **3.5**

2群間の比較については次のノンパラメトリック検定が一般的です。

①Wilcoxon の順位和検定
　対応のない独立したデータの有意差を判定するために使われます。
②Wilcoxon の符号付き順位和検定
　対応のあるデータの有意差を判定するために使われます。
　したがって比較する2群の標本数は一致しているはずです。

**例題3　3つ以上のグループの比較**

ある植物の成長が音楽を聴かせることによって促されるのかどうかを調査した。A～Cの3群にはそれぞれ、ベートーベン, 宇多田ヒカル, 民謡を聴かせた。なお対照群は日常の雑音以外は特に音楽を聴かせていない。果たして音楽の効果はあるといえるのだろうか。

| | | | | | | |
|---|---|---|---|---|---|---|
| 対照区 | 34.3 | 35.1 | 37.1 | 30.9 | 36.5 | 31.4 |
| ベートーベン | 36.1 | 34.5 | 38.6 | 37.6 | 35.6 | 34.1 |
| 宇多田ヒカル | 35.4 | 36.6 | 31.9 | 38.3 | 37.8 | 34.8 |
| 民謡 | 36.4 | 38.2 | 34.3 | 37.1 | 38.4 | 36.8 |

比較するグループが3群以上の場合には、多重分析の方法に従って検定をします。統計学の専門書を読むのが面倒だからといってt検定のような2群間の比較を総当たりで行ってはいけません。統計はあくまでも確率ですから、必ず検定結果には過誤が生じます。正しい仮説を捨ててしまう過ちを第1種の過誤、正しくない仮説を取ってしまう過ちを第2種の過誤と呼びます。このような過誤は群数や要因が多くなるほど複雑になり、とんでもない結論を生み出すことも考えられます。そのために統計学者が様々な検定法を考えています。私たちは実験の目的や条件にあった検定法を選択しなくてはなりません。

代表的な検定法を紹介する前に、まずは2群間比較の時のように分散に注目して適正な実験データであるかどうかを確かめます。用いるのは「Bartlett」検定です。これによって、同一母集団から抽出したデータであるかどうかを判断します。

（もし、同一母集団から得た標本とは言えないときは、3群以上のときに使われるノンパラメトリック検定を行うと良いでしょう。これについては後述します。）

次に簡単に有意差があるかどうかを分散分析という方法で調べます。特に要因が1つである場合、一元配置法と呼ばれる分散分析を行います。要因が1つというのは、例えば、摂食した食事の影響を調べる場合など、実験結果に影響する原因が1つの場合です。要因が2つというのは、例えば、摂食した食事と男女の性

別による影響があるなど、実験結果に影響する原因が2つの場合です。実験を計画するときは要因が1つだけ（あるいは少なく）になるようにするのが鉄則です。

ここでは初級者向けということで要因が1つである場合についてのみ記述します。申し訳ありませんが要因が複数のときの統計処理はご自分で勉強して下さい。

一元分散分析によって有意差があると判断された場合は、さらにどこに有意差があるのかを実験の目的や条件を考えながら適切な検定法を選択し、有意差を判定します。以下に代表的な検定法を紹介します。

### A）対照区との比較

例題のように、対照区と比較して各群に有意差があるかどうかを判断します。対照区と各群の標本数の異同によって次の2つの検定法を使い分けて下さい。

①Dunnett の検定
　標本数が同じ場合
②Scheffe の検定
　標本数が異なる場合

### B）すべての区との比較

特に対照区が設けられていない場合です。例えば10回の漢字テストの結果から、『○組は□組よりも漢字テストに対する意識が高い』などと結論づけられるかどうかを判定する場合などです。ちなみに、「高いから何だ」って言われると少々困ってしまいますが、担任としては統計からも明らかに自分のクラスよりも上だとわかれば、自分のやり方を考え直し、意識付けの高いクラスの担任のやり方などを参考にするといった方向転換のきっかけになるかもしれません。まあ、ギスギスした例になってしまいますが、話を戻します。

①Tukey の検定
②Bonferroni の検定
③Duncan の多重範囲検定（MRT）
④最小有意差法（LSD）

Tukey と Bonferroni の方法にはそれほど違いはありません。専門書などで多く紹介されているのは Tukey の方です。LSD は、検出力が高い方法で（言い方を変えると”判定が甘い”）、もし、①～③の方法で有意差を見つけられなかった場合、④のLSDで試して見ても良いでしょう。しかし、その場合LSDで有意差があったからと言って一方的に結論付けるのではなく、LSD は甘いということを踏まえながら総合的に考察をするべきです。Duncan の MRT は、Tukey や LSD の折衷的な方法と言われています。

### 例題4　3つ以上のグループのアンケート結果の比較

ある学校でスピーチコンテストへの参加を募ったところ、希望者が殺到した。学校側としてはより多くの人数を参加させたいが、主催者の意向もあり、スピーチが特に優秀だと思われる生徒を選抜し、参加させることにした。一次審査でA～Gの7名を選抜した。さらに厳選するために二次審査を開き、スピーチの内容を判定するために5名の人がパネラーとなりスピーチを判定した。しかし、最高得点の生徒と2位, 3位の生徒の点数が僅差であり、果たしてスピーチの実力差があるのかどうかわからない。もし、1位の生徒と差がないのであれば主催者にお願いして複数名参加させたいと考えている。そこで、統計処理を行い、最高得点の生徒と有意差がないものを参加させることにした。

| | パネラー1 | パネラー2 | パネラー3 | パネラー4 | パネラー5 |
|---|---|---|---|---|---|
| 生徒A | 4 | 5 | 3 | 3 | 3 |
| 生徒B | 4 | 4 | 3 | 2 | 4 |
| 生徒C | 5 | 5 | 3 | 4 | 4 |
| 生徒D | 3 | 2 | 1 | 2 | 2 |
| 生徒E | 2 | 2 | 1 | 1 | 2 |
| 生徒F | 3 | 3 | 3 | 3 | 4 |
| 生徒G | 4 | 5 | 3 | 2 | 4 |

今回のような場合、数的データよりもその順位を重要視しなくてはなりませんので、3群以上のノンパラメトリック検定、すなわち Kuruskal-Wallis 検定を行います。

はじめに Kruskal-Wallis 検定によって有意差があるかどうかを簡単に調べます。有意差が認められたら、さらにどの群間に有意差があるかを調べます。順位データを用いれば、パラメトリック検定で紹介した方法で検定しても良いでしょう。

**検定法に関する注意**

これまで紹介してきた統計手法はおおよそ間違いないと思いますが、検定法にはいろいろな性質があります。例えば、

『この検定法は標本数が同じ場合に適している』

と記述していても、別の文献には

『標本数が異なっていてもほぼ同じ結果を示す』

などと記述してあったりします。ですから、すべての人が同じ検定の仕方とは限りません。

これまで紹介してきた検定の方法で特に問題はないと思いますので参考にして頂ければ幸いです。また、統計処理を持ってしても100%こうだと言い切ることはできません。あくまでも有意性がある可能性が高いかそうでないかを示すだけですので、実験の性質や目的、内容等を総合して結論に結びつけられることが重要です。

---

## 5．おわりに

私が大学生のとき、統計学に対する学習意欲は極めて低く、おそらくここまで述べてきたことは講義でやっ

ていたのでしょうが、その当時のことはまったく覚えていません。ようやく卒業論文等で実験を行うようになり、データを処理するときになって、統計処理がわからないのではまずいと思い、勉強し始めました。

大学院生時代に Lotus で処理できるように自分でプログラムを組んだのを皮切りに、いくつかの専門書を読みあさり、書物を中心に統計学を学びました。専門書にも具体的な計算方法が載っていないようなレアな検定法も、英語文献を取り寄せて翻訳したことがあります。

今年数年ぶりに統計処理が必要となり、引き出しの奥にしまってあったプログラムを利用しました。今では Excel を中心にデータを処理していることもあり、この度 Excel 版に作り直しました。

高校では、実験の結論に至るまでに専門的な統計処理は行いません。しかし、統計処理は軽んじられるべきではありません。何らかの機会を得て、生徒に実験, 結果, 統計処理を経て結論までたどり着き、実験というものを教えられればいいなぁと思っています。いつか実験に必要になるかもしれないことを期待しながら、Excel 版のプログラムとこの統計学のススメを作成しました。

ここに紹介した統計手法は、絶対ではなく人によって若干異なるかもしれませんが、およそ間違いはないと思います。しかし、これによる一切の責任は負えませんのでご了承ください。もし間違い等がありましたらご連絡を頂けると幸いです。

高梨和幸 〈E-mail〉nacchi@helen.ocn.ne.jp