統計的仮説検定　その2

1　分散分析

2 つの平均値について，その差が有意であるかどうかは前述の t 検定を用いて行える．では 3 つ以上の平均値についてはどうすればよいか．それぞれ 1 対の平均値について t 検定を行うことも可能であるが，その場合にはあくまで対応する 2 つの平均値のみの有意性の議論となってしまうので，全体の中で 1 つだけ有意に差があるようなものを抽出できない．このような 3 つ以上の平均値についての統計的仮説検定が分散分析である．分散分析はその英語標記 (ANalysis Of VAriance) から ANOVA と呼ばれることもある．

例えば，3 つのクラスのテストの平均点を比べるような場合を考える．当然各クラスの平均点は異なる．このとき，どれかのクラスの成績が他に対して有意に異なっていることを示すためには，クラス内での得点のばらつきと，クラス間の得点のばらつきを比較し，クラス内よりもクラス間の分散がある程度大きいことが示されれば，有意かどうかが判断できる．そこで，分散の比較を行う必要があるが，2 つの正規母集団の分散違いを見るために，2 つの母分散の比を考える．このとき，分散の比の分布を表す F 分布を利用する．

1.1　F 分布

20 世紀初頭に分散比による統計を開発した Ronald A. Fisher に敬意を表して Snedecor が名づけたのが F 分布である．2 つの不偏分散 v_1^2 と v_2^2 があるとき，その比 v_1^2/v_2^2 は図 1 のようになる．図からわかるように曲線の形は不偏分散の自由度によって変わってくる．

F の値は 2 つの自由度を使用して求められるので，配布した F 分布表も 2 つの自由度に対して，指定した危険率 α に対応する値が表示される形式となっている．また，2 つの不偏分散 v_1^2 と v_2^2 が統計的に異なる分散といえるかどうかは，その比 F_0 が図 1 の右端の 5%未満の部分に入っているかで決まる．すなわち，次の不等式を満たしていれば異なる分散ということができる．なお，ここで α は設定した危険率に対応した値である．

$$\frac{v_1^2}{v_2^2} \geq \frac{1}{F_{\phi_1}^{\phi_2}}(\alpha)$$

1.2　1 元配置分散分析

k 人の被験者にそれぞれ n 回ある作業を行ってもらってその作業時間を測定するとし，その測定値を x_{ij} とする．ここで，変数 i は被験者を，変数 j は作業の順番を表すものとする．それらの作業にかかった時間が被験者によって有意に異なるのかを調べるときに，$k > 2$ であれば t 検定ではなく，分散分析を使うことになる．このとき，有意に異なるかどうかの**要因**としては被験者間の比較だけとすると，このような場合を **1 元配置分散分析**と呼ぶ．

被験者ごとの作業時間の平均値を $\bar{x}_i$，全平均を $\bar{x}$ とする．このとき，残差の平方和 S_T は当然であるが，以下の式となる．

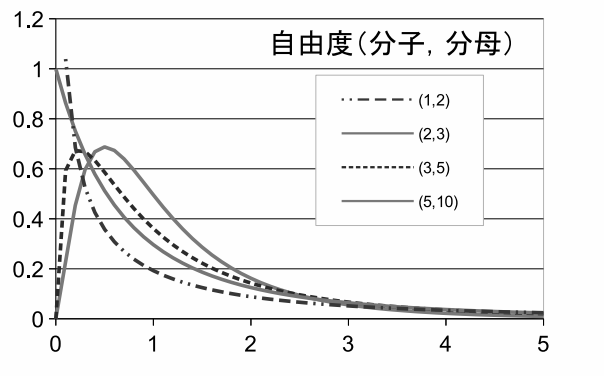

図 1: F 分布の例

$$S_T = \sum_i \sum_j (x_{ij} - \bar{x})^2 \tag{1}$$

分散分析の場合，このような残差を**変動**と呼ぶ．S_T は全変動という．さて，分析の目的は被験者によって作業時間が異なるかどうかであるので，被験者間の平均値の違いについて議論することとなる．その場合，被験者ごとのばらつきと被験者が行った作業時間自体のばらつきを考え，個々の被験者の平均値のばらつきと被験者間のばらつきの大きさを比較して検討することになる．そこで，式 (1) を以下のように，被験者間と被験者内のばらつきに分解する．

$$\begin{aligned}
\sum_i \sum_j (x_{ij} - \bar{x})^2 &= \sum_i \sum_j \{(x_{ij} - \bar{x}_i) + (\bar{x}_i - \bar{x})\}^2 \\
&= \sum_i \sum_j \left\{(x_{ij} - \bar{x}_i)^2 + 2(x_{ij} - \bar{x}_i)(\bar{x}_i - \bar{x}) + (\bar{x}_i - \bar{x})^2\right\} \\
&= \sum_i \sum_j (x_{ij} - \bar{x}_i)^2 + 2\sum_i \sum_j (x_{ij} - \bar{x}_i)(\bar{x}_i - \bar{x}) + \sum_i \sum_j (\bar{x}_i - \bar{x})^2 \\
&= \sum_i \sum_j (x_{ij} - \bar{x}_i)^2 + \sum_i \left\{(\bar{x}_i - \bar{x}) \sum_j (x_{ij} - \bar{x}_i)\right\} + n\sum_i (\bar{x}_i - \bar{x})^2 \\
&= \sum_i \sum_j (x_{ij} - \bar{x}_i)^2 + \sum_i \left\{(\bar{x}_i - \bar{x})(\sum_j x_{ij} - n\bar{x}_i)\right\} + n\sum_i (\bar{x}_i - \bar{x})^2 \\
&= \sum_i \sum_j (x_{ij} - \bar{x}_i)^2 + n\sum_i (\bar{x}_i - \bar{x})^2
\end{aligned} \tag{2}$$

式 (2) の右辺第 2 項は被験者間のばらつきに対応する残差平方和を表し，**群間変動**もしくは**級間変動**と呼ばれ，第 1 項は被験者内のばらつきということで**群内変動**もしくは**級内変動**と呼ばれ，それぞれ記号 S_B と S_W で表す．添え字の B と W はそれぞれ Between classes と Within classes から来ている．

$$S_W = \sum_i \sum_j (x_{ij} - \bar{x}_i)^2 \tag{3}$$

$$S_B = n\sum_i (\bar{x}_i - \bar{x})^2 \tag{4}$$

ここで，先に挙げた例に戻ると，個々の被験者の n 回の作業時間のばらつきに比べて，被験者ごとでの平均作業時間の差の方が大きくなるようなことになれば，被験者間の作業時間の違いに有意な差があるといえる．そこで，S_B と S_W からそれぞれの分散の推定値 v_B^2 と v_W^2 を求め，v_B^2 が v_W^2 に比べて有意に大きいかを F 分布を使って検定すれば良いということになる．

実際に 1 元配置分散分析を行うと以下の表 1 のようなデータ群を考えることになる．また，残差を計算する上ではこれまで行ってきたように変量の 2 乗とその和を知ることが必要であるので，表 2 も作成することになる．

表 1: 一元配置のデータ

標本（水準）	1	2	$\cdots$	k	和
変量 （繰り返し）	x_{11} x_{12} x_{1n_1}	x_{21} x_{22} $\cdots$ x_{2n_2}	$\cdots$ $\cdots$ $\cdots$	x_{k1} x_{k2} x_{kn_k}	
和	T_1	T_2	$\cdots$	T_k	T
平均	$\bar{x}_1.$	$\bar{x}_2.$	$\cdots$	$\bar{x}_k.$	$\bar{x}..$（総平均）

表 2: 平方和の計算用

標本（水準）	1	2	$\cdots$	k	和
変量の 2 乗	x_{11}^2	x_{21}^2	$\cdots$	x_{k1}^2	行および列全部の変量の和
	x_{12}^2	x_{22}^2	$\cdots$	x_{k2}^2	
		$\cdots$			$\sum_i^k \sum_j^{n_i} x_{ij}^2$
	$x_{1n_1}^2$	$x_{2n_2}^2$	$\cdots$	$x_{kn_k}^2$	
変量の和の 2 乗	T_1^2	T_2^2	$\cdots$	T_k^2	$T^2/N(N = n_1 + n_2 + \cdots + n_k)$
平均	T_1^2/n_1	T_2^2/n_2	$\cdots$	T_k^2/n_k	列の和 $\sum_i^k T_i^2/n_i$

上記のような値を準備した上で実際に変動を計算して分散分析表を作成する．式 (2) で用いたものは繰り返しの回数がすべて同じものであったので，実際には以下のように修正が必要となる．まず，全変動は以下のように求められる．

$$S_T = \sum_i^k \sum_j^{n_i} (x_{ij} - \bar{x}..)^2 = \sum_i^k \sum_j^{n_i} x_{ij}^2 - \frac{T^2}{N} \tag{5}$$

次に群間の変動 S_B を求める．

$$\begin{aligned} S_B &= \sum_i^k n_i(\bar{x}_i. - \bar{x}..)^2 \\ &= n_1(\bar{x}_1. - \bar{x}..)^2 + n_2(\bar{x}_2. - \bar{x}..)^2 + \cdots + n_k(\bar{x}_k. - \bar{x}..)^2 \\ &= \sum_i^k \frac{T_i^2}{n_i} - \frac{T^2}{N} \end{aligned} \tag{6}$$

群内変動 S_W は以下の式となる．

$$\begin{aligned} S_W &= \sum_i^k \sum_j^{n_i} (x_{ij} - \bar{x}_i.)^2 \\ &= (x_{11} - \bar{x}_1.)^2 + (x_{12} - \bar{x}_1.)^2 + \cdots + (x_{1n_1} - \bar{x}_1.)^2 \\ &\quad + (x_{21} - \bar{x}_2.)^2 + (x_{22} - \bar{x}_2.)^2 + \cdots + (x_{2n_2} - \bar{x}_2.)^2 \\ &\quad + \quad \cdots \\ &\quad + (x_{k1} - \bar{x}_k.)^2 + (x_{k2} - \bar{x}_k.)^2 + \cdots + (x_{kn_k} - \bar{x}_k.)^2 \\ &= \sum_i^k \sum_j^{n_i} x_{ij}^2 - \sum_i^k \frac{T_i^2}{n_i} \end{aligned} \tag{7}$$

当然であるが，$S_T = S_B + S_W$ である．

次に自由度であるが，全変動 S_T の自由度 ϕ_T についてはこれまで通りだが，S_B と S_W の自由度 ϕ_B と ϕ_W は注意が必要である．

$$\phi_T = N - 1$$

$$\phi_B = k - 1$$

$$\phi_W = N - k$$

ここでも，$\phi_T = \phi_B + \phi_W$ である．不偏分散は残差と自由度を用いて以下のように計算できる．

$$v_B^2 = \frac{S_B}{k-1} \tag{8}$$

$$v_W^2 = \frac{S_W}{N-k} \tag{9}$$

最後に， 不偏分散の比を上式から求める．

$$F_{N-k}^{k-1} = \frac{v_B^2}{v_W^2} \tag{10}$$

この比が自由度 $(k-1,\ N-k)$ の F 分布となることから有意性を検定する．分散分析表は以下のような項目からなる．

表 3: 分散分析表

要因	変動	自由度	不変分散	不偏分散比
群間変動	S_B	$k-1$	$v_B^2 = \frac{S_B}{k-1}$	$F_{N-k}^{k-1} = \frac{v_B^2}{v_W^2}$
群内変動	S_W	$N-k$	$v_W^2 = \frac{S_W}{N-k}$	
全変動	S_T	$N-1$		

【一元配置分散分析の実際】
配布資料に基づいて実際に分散分析を行ってみる．

1.3 検定結果の説明

以下のように検定結果の説明は行われる．

- 帰無仮説：得られた平均値の間に有意な差は無い．
- F_{N-k}^{k-1} を計算する．
- F 分布表で自由度 $(k-1,\ N-k)$ の値を確認する．
  - $F_{N-k}^{k-1} \geq F(\alpha, k-1, N-k)$ のとき
    - 帰無仮説を否定するときの危険率が α%未満であるので，帰無仮説は棄却できる．
    - よって，平均値の間に有意な差がある．$(p < \alpha)$
  - $F_{N-k}^{k-1} < F(.05, k-1, N-k)$ のとき
    - 帰無仮説を否定するときの危険率が 5%以上であるので，帰無仮説を棄却できない．
    - よって，平均値の間に有意な差があるとは言えない．