第十章　多標本の仮説検定と分散分析

3 種類以上の標本間の差を検定する場合を考える。例えば 3 種類の薬品 A, B, C の効果を t 検定する場合、A と B、A と C、B と C、の計 3 回 t 検定するとよさそうだが、これは正しくない。 α=0.05 で t 検定を行うとき(一検定あたり 20 回中 1 回のミスは受け入れる)、t 検定を 3 回繰り返すと危険率は $1-(1-\alpha)^3=0.14$ に増加してしまう。検定全体の危険率は標本数が増えると急増する（ex. 5 種類の比較で 0.40、10 種類で 0.90)。

多標本の平均の差を比較する場合には分散分析(ANOVA)を行う。

10.1　一要因分散分析(single-factor analysis of variance)

k 個の標本から平均の差の検定を行う場合、帰無仮説は H_0： $\mu_1=\mu_2=\mu_3=\cdots=\mu_k$ で、対立仮説 H_A：k 個の平均は等しくない、となる。

一要因分散分析(一元配置分散分析ともいう)：一つの要因(factor)が及ぼす影響を検定する分散分析。要因の種類を水準(level)といい、下の**例題 10.1** ではエサが要因でエサの種類 1〜4 が水準となる。標本は各水準に**無作為**に割り当てる(完全無作為化法、completely randomized design, CRD)。各水準の標本数は等しい方がよい(釣り合い型、balanced experimental design)。実験は、要因以外の様々な条件(性別・年齢・種類、成育環境など)をなるべく同じに行う。

例題 10.1　一要因分散分析（モデルＩ）

4 種類のエサで飼育したブタの体重増加量(kg)を比較する。以下の個体別調査データで、H_0： $\mu_1=\mu_2=\mu_3=\mu_4$ 、 H_A：各水準の平均は等しくない、を α=0.05 で検定する。

	Feed1	*Feed2*	*Feed3*	*Feed4*
	60.8	68.7	69.6	61.9
	67.0	67.7	77.1	64.2
	65.0	75.0	75.2	63.1
	68.6	73.3	71.5	66.7
	61.7	71.8		60.3
i	1	2	3	4
n_i	5	5	4	5
$\sum_{j=1}^{n_i} X_{ij}$	323.1	356.5	293.4	316.2
$\overline{X}_i$	64.62	71.30	73.35	63.24

水準 i の j 個目のデータを X_{ij} とすると、以下のようなデータ構造が考えられる。

$$X_{ij}=\overline{X}+\alpha_i+e_{ij}\quad（i=1〜k、j=1〜n_i）$$

ここで、要因の水準数を k、水準 i におけるデータ数を n_i としている。よって、全個体数 N

は、$N=\sum_{i=1}^{k} n_i$ であり、総平均 $\bar{X}=\sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij} \Big/ N$ である。$\bar{X}_i=\sum_{j=1}^{n_i} X_{ij} / n_i$ なので、要因の効果は $\alpha_i=\bar{X}_i-\bar{X}$ となる。e_{ij} は残差(誤差)と呼ばれ $e_{ij}=X_{ij}-\left(\bar{X}+\alpha_i\right)$ で計算できる。**例題** 10.1 の場合のデータ構造による分解を下に示すと、

生データ X_{ij}

Feed1	Feed2	Feed3	Feed4
60.8	68.7	69.6	61.9
67.0	67.7	77.1	64.2
65.0	75.0	75.2	63.1
68.6	73.3	71.5	66.7
61.7	71.8		60.3

＝ 総平均 $\bar{X}$

67.9	67.9	67.9	67.9
67.9	67.9	67.9	67.9
67.9	67.9	67.9	67.9
67.9	67.9	67.9	67.9
67.9	67.9		67.9

＋ 処理効果 α_i

α_1	α_2	α_3	α_4
−3.2	3.4	5.5	−4.6
−3.2	3.4	5.5	−4.6
−3.2	3.4	5.5	−4.6
−3.2	3.4	5.5	−4.6
−3.2	3.4		−4.6

＋ 残差(誤差) e_{ij}

−3.8	−2.6	−3.8	−1.3
2.4	−3.6	3.8	1.0
0.4	3.7	1.9	−0.1
4.0	2.0	−1.8	3.5
−2.9	0.5		−2.9

ex. X_{23} 73.3＝67.9+3.4+2.0。このデータ構造の個々の値を 2 乗すると下図になる。

3696.6	4719.7	4844.2	3831.6
4489.0	4583.3	5944.4	4121.6
4225.0	5625.0	5655.0	3981.6
4706.0	5372.9	5112.3	4448.9
3806.9	5155.2		3636.1

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij}^2 = 87955.3000 \cdots ①$$

4604.0	4604.0	4604.0	4604.0
4604.0	4604.0	4604.0	4604.0
4604.0	4604.0	4604.0	4604.0
4604.0	4604.0	4604.0	4604.0
4604.0	4604.0		4604.0

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i} \bar{X}^2 = 87475.6126 \cdots ②$$

10.4	11.9	30.2	21.3
10.4	11.9	30.2	21.3
10.4	11.9	30.2	21.3
10.4	11.9	30.2	21.3
10.4	11.9		21.3

$$\sum_{i=1}^{k} n_i\left(\bar{X}_i-\bar{X}\right)^2 = 338.9374 \cdots ③$$

14.6	6.8	14.1	1.8
5.7	13.0	14.1	0.9
0.1	13.7	3.4	0.0
15.8	4.0	3.4	12.0
8.5	0.3		8.6

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i} e_{ij}^2 = 140.7500 \cdots ④$$

値は丸めて表示している、和を計算する場合の丸め誤差に注意!!

各平方和(sum of square, SS)は

全体の SS $=\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(X_{ij}-\bar{X}\right)^2 \quad \cdots$ ①−②

群間の SS $=\sum_{i=1}^{k} n_i\left(\bar{X}_i-\bar{X}\right)^2 \quad \cdots$ ③

残差の SS＝$\sum_{i=1}^{k}\left[\sum_{j=1}^{n_i}\left(X_{ij}-\bar{X}_i\right)^2\right] \quad \cdots$ ④

となり、

全体の SS = 群間 SS + 残差 SS

となる。各平方和に対して自由度(df)を考える(4 章 4.3, 4.4 を参照)。全体の df は全個体数-1 つまり N-1 で、群間の自由度は"水準数 k-1"、残差の自由度は"各水準の個体数-1 の和 n_1-1+n_2-1+n_3-1+n_4-1"である。自由度も、

全体の df = 群間 df + 残差 df

が成り立つ。平方和を自由度で割ると平均平方(=分散の不偏推定量)が計算できる。帰無仮

説を検定するための統計量は

$$F\text{比} = \text{群間 MS} / \text{残差 MS}$$

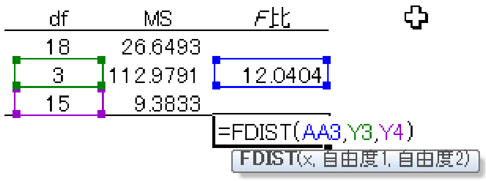
である。これを F 分布表の棄却域 $F_{\alpha(1),\nu_1,\nu_2}$ と比較する。つまり、有意水準 α の片側 F 検定(ν_1=群間 df、ν_2=残差 df)を行う。

平方和(=変動)と自由度、平均平方、F 比をまとめたものを分散分析表といい、**例題 10.1** の場合、下のようになる。

分散分析表(**例題 10.1**)

変動要因	平方和	自由度	平均平方	F比
全体	479.6874	18		
群	338.9374	3	112.9791	12.04
残差	140.7500	15	9.3833	

α=0.05 のとき、$F_{0.05(1),3,15}$=3.29 (附表もしくは Excel の FINV 関数で計算可)となり。F 比 > 3.29 であり、帰無仮説は棄却される。Excel の FDIST 関数より p 値は 0.00029 となる(右図)。つまり、4 種のエサの効果は等しくない、と結論できる(p = 0.00029)。



一要因分散分析はつまり、次の分散分析表にまとめることができる。

一要因の場合の分散分析表

変動要因	平方和(SS)	df	MS（分散）	F比
全体	$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(X_{ij}-\bar{X}\right)^2$	N-1		
群	$\sum_{i=1}^{k} n_i\left(\bar{X}_i-\bar{X}\right)^2$	k-1	$\frac{\text{群SS}}{\text{群df}}$	$\frac{\text{群MS}}{\text{残差MS}}$
残差	$\sum_{i=1}^{k}\left[\sum_{j=1}^{n_i}\left(X_{ij}-\bar{X}_i\right)^2\right]$	N-k	$\frac{\text{残差SS}}{\text{残差df}}$	

分散分析では、(k 群の)標本は等分散で正規分布する、と仮定している。仮定からの少々のズレに対して分散分析は頑健であるが、以下の点は解析精度に影響する(正規性からのズレや分散の差は大きくなるほど検定は不正確になる)。

・標本数は等しい方がよい

・標本数は多い方がよい

10.2　ウェルチの検定　分散が異なる場合の多標本平均の検定

標本間の分散や標本数が大きく異なる場合、ウェルチの検定が使える。k 種の標本に関する F統計量は、分子の自由度 $\nu_1=k-1$、分母の自由度 $\nu_2=(k^2-1)/(3A)$の F分布に従う。

$$F' = \frac{\sum_{i=1}^{k} c_i\left(\overline{X}_i - \overline{X}_w\right)^2}{(k-1)\left[1+\frac{2A(k-2)}{k^2-1}\right]}$$

ここで、$c_i = \frac{n_i}{s_i^2}$、$C = \sum_{i=1}^{k} c_i$、$\overline{X}_w = \frac{\sum_{i=1}^{k} c_i \overline{X}_i}{C}$、$A = \sum_{i=1}^{k} \frac{\left(1 - c_i/C\right)^2}{\nu_i}$ である。

問題 10.2 で計算してみよう。

10.3　分散分析の使い方

多標本データをレポートや論文へのまとめ方

・標本数・平均・標準偏差 or 標準誤差でデータを示す

・正規分布する場合は、信頼限界も示すとよい

多標本データ別の、検定方法の使い分け

・多標本が正規分布し等分散の場合(ズレが少ない)、通常の ANOVA で検定

・多標本は正規分布(近似)するが、分散が異なる場合、ウェルチの検定

・多標本は非正規だが分布や分散が近似の場合、Kruskal-Wallis 検定ノンパラメトリック

・多標本が非正規で分布や分散も異なる場合、

①データ変換を試す、②仮説検定せず、平均や分散のみ図表で報告

[注意]仮説検定全般に当てはまることだが、$\alpha=0.05$ の棄却限界値に拘りすぎるのはよくない。標本(から推測される母集団)が仮定を満たさない時、α はタイプ I エラー確率とズレる(仮定からのズレに応じてズレは大きくなる)。(計算した)統計量が棄却限界値と近い場合は($\alpha=0.05$ に対し $p=0.047$ など)、検定結果の解釈には注意が必要である。可能であれば実験を繰り返す方がよい。

10.4　多標本の等分散性の検定

多標本の等分散性の検定には、バーレットの検定が使える。比較する群が k 種類の場合、i 番目の群の個体数を n_i、自由度を ν_i、平方和を SS_i とすると、各群が等分散であるという帰無仮設のもとでは k 群の分散をプールでき、$s_p^2 = \sum_{i=1}^{k} \mathrm{SS}_i / \sum_{i=1}^{k} \nu_i$ と書ける。

バーレットの検定に用いる B 統計量は、

$$B = \left(\ln s_p^2\right)\left(\sum_{i=1}^{k}\nu_i\right) - \sum_{i=1}^{k}\nu_i \ln s_i^2$$

と書ける。B 統計量には補正係数 C があり、補正後の $B_C = B/C$ は、自由度 k-1 の χ^2 分布することが分かっている。

$$C = 1 + \frac{1}{3(k-1)}\left(\sum_{i=1}^{k}\frac{1}{\nu_i} - 1\Big/\sum_{i=1}^{k}\nu_i\right)$$

例題 10.13　例題 10.1 のデータの等分散性をバートレットの検定で評価する。

H_0 : $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$、 H_A : 各水準の分散は等しくない、を α=0.05 で検定する。

	Feed1	*Feed2*	*Feed3*	*Feed4*
	60.8	68.7	69.6	61.9
	67.0	67.7	77.1	64.2
	65.0	75.0	75.2	63.1
	68.6	73.3	71.5	66.7
	61.7	71.8		60.3
i	1	2	3	4
n_i	5	5	4	5
ν_i	4	4	3	4
SS_i	44.768	37.660	34.970	23.352
s_i^2	11.192	9.415	11.657	5.838
$\ln s_i^2$	2.4152	2.2423	2.4559	1.7644
$\nu_i \ln s_i^2$	9.6608	8.9692	7.3676	7.0576
$1/\nu_i$	0.250	0.250	0.333	0.250

$s_p^2 =$ (1) $\boxed{9.3883}$ kg^2、B= (2) $\boxed{0.529}$、C= (3) $\boxed{1.109}$、B_C= (3) $\boxed{0.476}$、$\chi^2_{0.05,3}$ = (4) $\boxed{7.815}$、p= (5) $\boxed{0.923}$

ここで(4)、(5)の計算には、Excel の CHIINV 関数、CHIDIST 関数を用いるとよい。

(注意)バートレットの検定は標本が正規分布することを仮定している。

［レポート］

［問題 10.1］ pp2 の、

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(X_{ij}-\overline{X}\right)^2=\sum_{i=1}^{k}\sum_{j=1}^{n_i}X_{ij}^2-\sum_{i=1}^{k}\sum_{j=1}^{n_i}\overline{X}^2$$ を証明せよ。

［問題 10.2］ 下のデータは 3 品種のコムギ(G,A,L)からそれぞれ 5 個体ずつを無作為抽出して測定したカリウム濃度(単位組織あたり mg)である。3 品種間のカリウム濃度の平均に差があるかウェルチの検定(α=0.05)で検定せよ。

	品種G	品種A	品種L
	27.9	24.2	29.1
	27.0	24.7	27.7
	26.0	25.6	29.9
	26.5	26.0	30.7
	27.0	27.4	28.8
	27.5	26.1	31.1
i	1	2	3
n_i	6	6	6
ν_i	5	5	5
$\overline{X}_i$	27.0	25.7	29.6
s_i^2	0.462	1.279	1.607