

## anova/one-way - 分散分析

anova は分散分析の機能を提供する汎用的なコマンドです。これに対し oneway は一元配置 ANOVA に特化し、多重比較機能も包含するなど、より使いやすさを追求したコマンドです。

1. 分散分析と多重比較
  2. 分散分析の前提条件
  3. 一元配置 ANOVA – oneway
  4. 一元配置 ANOVA – anova
  5. 線形回帰モデル
  6. 二元配置 ANOVA
  7. 二元配置 ANOVA 後の多重比較
  8. 反復測定 ANOVA
- 補足 1  
補足 2  
補足 3

## 1. 分散分析と多重比較

平均値が等しいと言えるかどうかを検定する場合、対象とする標本の数  $n$  が 2 つ以下の場合には通常  $t$  検定が用いられます (*mwp-041* 参照)。それでは 3 つの標本 A, B, C が与えられたとき、 $t$  検定を繰返し使用したら何が悪いのでしょうか？今、検定の有意水準  $\alpha$  を 5% とすると

- (1) A-B 間の比較で有意差が検出されない確率 = 0.95
- (2) A-C 間の比較で有意差が検出されない確率 = 0.95
- (3) B-C 間の比較で有意差が検出されない確率 = 0.95

ですから、3 回の検定を通じて有意差が検出されない確率は  $0.95^3 = 0.86$  となります。逆に言えば (1), (2), (3) のいずれかで有意差が検出される確率は  $1 - 0.86 = 0.14$  となり、正しい帰無仮説を棄却してしまう過誤 (第 1 種過誤) の確率が本来の 5% よりも大きく増大するといった問題が生じます。

このため標本数が3以上の場合には分散分析 (ANOVA: analysis of variance) という手法が用いられます。これはすべての平均値間に差がないことを  $F$  検定によって確認しようとするものです。しかしこの仮説が棄却された場合に、どの平均とどの平均の間に有意差が認められるかについて、分散分析自体は何ら情報をもたらしません。このため、有意確率の補正を伴う多重比較 (multiple comparison) 検定を併用することが通常行われます。

## 2. 分散分析の前提条件

分散分析の実行に際しては次の要件が満たされていることが前提となります。

- (1) 従属変数 (応答変数) は量的 (区間尺度) データであること
- (2) 従属変数 (応答変数) は正規分布に従うこと
- (3) 観測データは互いに独立であること
- (4) 各グループの分散は均一であること

 反復測定 (repeated-measures) ANOVA の場合には独立性に関する前提条件が成り立たなくなります。

## 3. 一元配置 ANOVA – oneway

要因として想定する因子 (factor) の数が1つの場合を一元配置 ANOVA (one-way ANOVA) と言います。一元配置 ANOVA に対しては `anova`, `oneway` コマンド双方を使用することができますが、本セクションでは `oneway` を用いて分析を行ってみます。使用するデータセットは `anova1.dta` です。

```
. use http://www.math-koubou.jp/stata/data12/anova1.dta, clear
```

全部で24個の血圧 (blood pressure) データが記録されていますが、ここではその一部のデータをリスト表示しておきます。

```
. list if _n <= 3 | _n >= 22, separator(3) *1
```

	bp	drug
1.	126	1
2.	121	1
3.	115	1
22.	137	4
23.	139	4
24.	123	4

\*1 メニュー操作 : Data ▷ Describe data ▷ List data

変数 `drug` は薬剤の種類を表すカテゴリ変数で 1, 2, 3, 4 という 4 つの値を取ります。この `drug` の値ごとにデータを整理し、平均値を算出すると次のようになります\*2。

drug	bp						平均値
1	126	121	115	123	125	113	120.5
2	112	123	115	129	106	108	115.5
3	123	112	133	124	130	121	123.8
4	122	132	125	137	139	123	129.7

薬剤の種類によって平均値は微妙に異なるわけですが、これらの差が統計的に有意と言えるかどうかを `oneway` を使って検定してみます。なお、有意水準  $\alpha$  の値としてはデフォルトの 5% を用いることにします。また要約統計量を示す表の作成と多重比較検定の実行に関するオプションも指定することにします。

- Statistics ▸ Linear models and related ▸ ANOVA/MANOVA ▸ One-way ANOVA と操作
- Main タブ: Response variable: bp  
Factor variable: drug  
Multiple-comparison tests: Bonferroni  
Output: Produce summary table: ✓

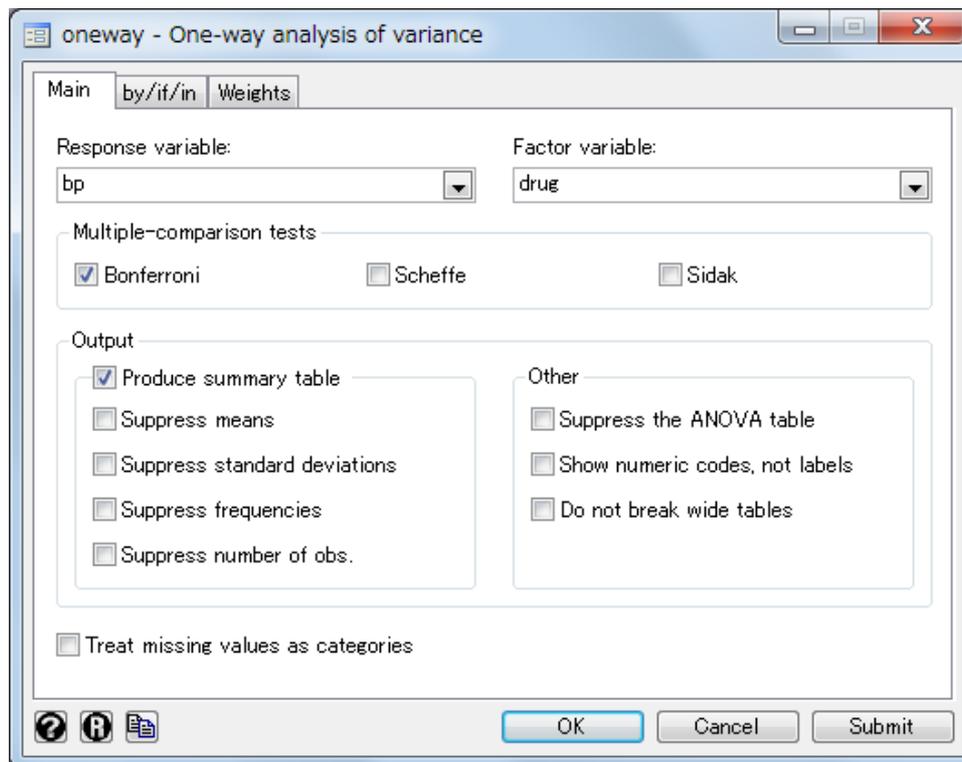


図 1 oneway ダイアログ- Main タブ

\*2 Stata の機能を用いてテーブルを作成する場合の操作方法については補足 1 を参照ください。

```
. oneway bp drug, bonferroni tabulate
```

drug	Summary of bp			Freq.
	Mean	Std. Dev.		
1	120.5	5.3572381		6
2	115.5	8.9162773		6
3	123.83333	7.3598007		6
4	129.66667	7.3665913		6
Total	122.375	8.6467511		24

Source	Analysis of Variance				
	SS	df	MS	F	Prob > F
Between groups	636.458333	3	212.152778	3.92	0.0237
Within groups	1083.16667	20	54.1583333		
Total	1719.625	23	74.7663043		

Bartlett's test for equal variances: chi2(3) = 1.1493 Prob>chi2 = 0.765

Comparison of bp by drug  
(Bonferroni)

Row Mean- Col Mean	1	2	3
2	-5 1.000		
3	3.33333 1.000	8.33333 0.383	
4	9.16667 0.260	14.1667 0.020	5.83333 1.000

## (1) ANOVA 表の解釈

ここでの操作では tabulate オプションを指定しているので、最初にまず因子水準ごとに要約統計量を整理した表が出力されています。この例では度数 (frequency) がいずれも 6 ということでバランスの取れたデータとなっていますが、oneway にしろ anova にしろ、バランスの取れていないデータ (unbalanced data) も扱うことができます。

要約統計量の表に続く形で出力されているのが ANOVA 表です。

Analysis of Variance					
Source	SS	df	MS	F	Prob > F
Between groups	636.458333	3	212.152778	3.92	0.0237
Within groups	1083.16667	20	54.1583333		
Total	1719.625	23	74.7663043		
Bartlett's test for equal variances: chi2(3) = 1.1493 Prob>chi2 = 0.765					

SS は平方和 (sum of squares) を意味します。regress の項 (*mwp-037* 参照) では  $y$  の変動を

$$\sum (y_i - \bar{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2$$

のように分解し

$\sum (y_i - \bar{y})^2$  を TSS (total sum of squares)

$\sum (\hat{y}_i - \bar{y})^2$  を MSS (model sum of squares)

$\sum (y_i - \hat{y}_i)^2$  を RSS (residual sum of squares)

と表現しましたが、今の場合、MSS に相当するのが群間変動 (between groups) 636.46、RSS に相当するのが群内変動 (within groups) 1083.17、TSS に相当するのが全変動 (total) 1719.63 です。このとき、それぞれの変動を自由度 (df: degrees of freedom) で割ることによって MS (mean square) すなわち平均平方が各々 212.15, 54.16 と算出されます。これらの比を取ったものが  $F$  値で  $212.15/54.16 = 3.92$  という値になります。この  $F$  値を使った  $F$  検定の結果が  $p$  値 0.0237 と示されているわけですが、この場合の帰無仮説はすべての平均値が等しいとするものです。今の場合、 $p$  値は  $< 0.05$  ですから帰無仮説は棄却されることとなります。

なお、ANOVA 表の末尾に Bartlett 検定の結果が表示されていますが、これは等分散性に関するものです。等分散性は ANOVA の前提条件の一つであるわけですが、今の場合  $p$  値は  $\gg 0.05$  ですから、この前提条件に関する問題はないと言えます。

## (2) 多重比較

上記 ANOVA 表からは  $\mu_1 = \mu_2 = \mu_3 = \mu_4$  が主張できないということはわかったわけですが、どのペア間に有意差があるかを見るためには多重比較検定のプロセスが必要となります。上記操作では bonferroni というオプションを指定しているので、Bonferroni 補正を施した形の多重比較検定の結果が ANOVA 表に引き続く形で出力されています。

		Comparison of bp by drug (Bonferroni)		
Row Mean- Col Mean		1	2	3
2		-5 1.000		
3		3.33333 1.000	8.33333 0.383	
4		9.16667 0.260	14.1667 0.020	5.83333 1.000

表示されている行列要素を  $M_{ij}$  と表記することにすれば、 $M_{ij}$  には  $\mu_i - \mu_j$  の値とそれが 0 と言えるか否かに関する Bonferroni 補正後の  $p$  値が示されています。この結果からすると  $\mu_4 - \mu_2$  間の差のみ有意と判定されています。多重比較補正の手法としては Bonferroni の他に Scheffe, Šidák が選択できます。

## 4. 一元配置 ANOVA – anova

評価版では割愛しています。

## 5. 線形回帰モデル

評価版では割愛しています。

## 6. 二元配置 ANOVA

評価版では割愛しています。

## 7. 二元配置 ANOVA 後の多重比較

評価版では割愛しています。

## 8. 反復測定 ANOVA

評価版では割愛しています。

## 補足 1 – 表の作成

評価版では割愛しています。

## 補足 2 – 繰返しのない二元配置 ANOVA

評価版では割愛しています。

## 補足 3 – 二元配置データのグラフ化

評価版では割愛しています。

