特別寄稿（杉本典夫 先生のコメントまとめ）

対応のある2元配置分散分析の１例

An example of a 2-way ANOVA for 2 samples paired.

Two-way ANOVA compares the mean difference between the groups for the two independent variables.The first purpose of the two-way ANOVA is to investigate the effect of two different categories independent variable for continuous dependent variable.

前説：

本ブログにおいて、情報統計研究所は統計分析手法のチョットしたコツを分かり易くご紹介してまいりました。現在、統計分析には多くの商用ソフトを利用するのが一般的になっており、データを入力すれば、統計学的な知識・誤用に関係なく、容易にそれ相応の分析結果を得ることが出来ます。本ブログで紹介した数々の手法（Excel 、R 、商用ソフト）においても同様です。本ブログの投稿記事に対して杉本典夫先生から色々なコメントを頂いており、特に今回、杉本典夫先生のご了承を得ましたので、本題の内容をなるべく詳細にご紹介したいと思います。

臨床現場において、CPK活性値の測定は心筋梗塞などの診断と治療の指標として有用です。臨床検査では、患者から採血した血液の凝固を待ってから遠心分離（3000回転/分）し、上清の血清を別の試験官に採取し、CPK活性値の測定まで保管します。ここで、この血清を室温（日射下の窓際、28℃)に放置した場合と冷蔵庫(冷暗所)に保管した場合のCPK活性値にどのような影響を及ぼすか調べたデータについて考えて見たいと思います。

20検体の血清を冷暗所と室温に放置した場合のCPK活性値の変化は図１(5page)のとおりでした。1時間放置と2時間放置の検体は同じものではなく、2時間放置の検体の平均値の方が高いと言えます。そして、「対応のある平均値の差の検定」(t-test)では2時間放置の方が有意(p<0.001)となっており、CPK活性値の高い方が有意に減少すると思われます。

このCPK活性値の２つの差に注目すると、本題の「対応のある2元配置分散分析」の対象となります。そこで、表１(6page)のようなデータフォーマットを作り、2元配置分散分析に必要な計算を行います。この表1から平方和をもとに分散分析表を計算しますが、ここではSPSSでの分散分析表(図２, 7page)を示しておきます。

ここで、1時間放置と2時間放置のCPK活性値は図３(7page)のような相関関係があります。

以上の前説をもとに、杉本先生のコメントを原文のまま掲載しますので、ご参考になれば幸いです。

杉本先生コメント(原文)：

このように異なった条件における測定値を時系列的に分析するには測定条件×時間の二元配置デザインの試験を行うのが普通です。

例えば、このCPK活性のデータでは、同じ20例の被験者から採取した血液を4つに分割し、それを次のような4種類の方法で測定します。

　冷蔵庫−1時間放置後測定　　冷蔵庫−2時間放置後測定

　室温-1時間放置後測定　　　室温-2時間放置後測定

こうして測定した20例×4種類のデータについて、1hも2hも冷蔵庫と室温の差をデータにして、対応のある二元配置分散分析で解析すると、次のような結果になります。

要因A :No

要因B 1:B1h - A1h

要因B 2:B2h - A2h

時期　:時期(要因B)別・全個体(要因A)合計基礎統計量

------------------------------------------------------------------------------------------------

1　:例数=20　平均値=-14.4　標準偏差=32.8384　標準誤差=7.34288

2　:例数=20　平均値=-36.5　標準偏差=41.7177　標準誤差=9.32837

-----------------------------------------------------------------------------------------------

全体:例数=40　平均値=-25.45　標準偏差=38.7099　標準誤差=6.12058

・対応のある二元配置分散分析(two-way layout analysis of variance)4

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 要因 | 平方和 | 自由度 | 平均平方和 | F値 | 有意確率ｐ値 |
| 測定条件 | 25908.1 | 1 | 25908.1 | 11.23 | 0.00335538 \*\* |
| 個体残差 | 43833.9 | 19 | 2307.05 |  |  |
| 交互作用 | 4884.1 | 1 | 4884.1 | 9.54524 | 0.00603416 \*\* |
| 残差 | 9721.9 | 19 | 511.679 |  |  |
| 全体 | 84348 | 40 |  |  |  |

「測定条件」の検定結果は、

1hと2hのデータを平均してAとBで比較したものでありAよりもBの方が平均して-25.45だけ低いことが有意かどうかを検定しています。「交互作用」の検定結果は測定条件(A・B)と測定時間(1h・2h)の交互作用であり1hの平均低下値-14.4と2hの平均値低下値-36.5の差が有意かどうかを検定しています。

これらの結果から冷蔵庫保管と比較すると室温保管は低い値になりその低下値は1hよりも2hの方がより大きいことがわかります。この分析結果をより詳細に検討するため、級内相関係数、相関係数、回帰係数を求めるのが本来の手順です。

なおこのデータでは被験者、保存条件、放置時間という3種類の要因が血清CPK活性に影響を与えていると考えられます。そのため本来は被験者を要因A、保存条件を要因B、放置時間を要因Cとした三元配置分散分析を適用すべきです。しかし三元配置分散分析は主作用が3種類(A、B、C)、交互作用が4種類(A×B、A×C、B×C、A×B×C)も求められる複雑な手法であり、結果の解釈も複雑になります。

そこでこのような場合は、適当な要因を1つ潰して二元配置にすると便利です。例えば今回の場合は保存条件に関して、同一被験者の冷所保存データと室温保存データの差をデータにすることによって、要因を1つ潰して二元配置デザインに落としました。それが「対応のある二元配置分散分析」つまり「差をデータにした二元配置分散分析」です。

この手法では保存条件の差をデータにするため、保存条件(要因B)は修正項(切片)になり、

被験者(A)は被験者×保存条件(A×B)という交互作用(被験者残差)になり、放置時間(C)は放置時間×保存条件(C×B)という交互作用になり、被験者×放置時間(A×C)は被験者×放置時間×保存条件(A×C×B)という交互作用(残差)になります。

そして保存条件(B)を検定する時は、被験者×保存条件(A×B)を検定誤差として用い、

放置時間×保存条件(B×C)という交互作用を検定する時は、被験者×放置時間×保存条件(A×C×B)を検定誤差として用います。

これは今回のような臨床検査値に対する測定条件の影響を検討する時だけでなく、医学・薬学分野の研究現場ではたまに用いられるデザインです。ところがSASやSPSSやRなどの既存の統計ソフトは、何故かこの手法をサポートしていません。

そのため「対応のある二元配置分散分析」を適用したい時は、複数の統計手法を組み合わせる必要があります。SPSSの二元配置分散分析を用いると、切片(保存条件)もSAMPLE(被験者×保存条件)もTIME(放置時間×保存条件)も、全て誤差(A×C×B)を検定誤差として用いてしまいます。

そのため保存条件の効果を検定するためには、切片の平均平方和をSAMPLEの平均値平方和で割ってF値を求め、手計算で保存条件の検定を行う必要があります。

これは被験者ごとに1hrと2hrのデータを平均し、それをデータにして冷所保存と室温保存の間で対応のあるt検定(1標本t検定)を行ったものに相当します。

SAMPLE(被験者×保存条件)の検定結果は、被験者と保存条件の間に交互作用があるかどうか、つまり保存条件の違いによるデータの変化量(差)が、被験者ごとにバラついているかどうかを検定したものです。

これは今回はあまり意味のない検定なので、僕は行いませんでした。それからSASやSPSSの分散分析では、平方和のタイプを4種類求めるようになっています。これらの平方和はバランスの取れた試験デザインを用いていれば、理論的に全て一致します。

そして分散分析は、本来はバランスの取れた実験デザインを組み立てるためにFisherが実験計画法を開発した時、それらの実験デザインから得られたデータを分析するために同時に開発した手法です。そのため4種類の平方和が一致しないような分散分析は、そもそも用いた実験デザインに欠陥があるのです。そのようなデザインでは各要因が直交せず、相関が生じます。その結果、各要因の平方和の間に相加性が成り立たず、計算方法や計算手順によって平方和が複数求められてしまいます。それが4種類の平方和です。

そのため本来は分散分析ではなく、一般的な線形モデルつまり重回帰分析のような多変量解析手法を適用すべきです。それにもかかわらず無理に分散分析を適用すると、4種類の平方和が一致しないということになってしまうわけです。したがってそういう時は、試験デザインを見直すのが本来の対処法ということになります。

薬効評価の場合は、投与前と投与後の変化量をデータにすることによって時間という要因を潰し、二元配置に落とすのが普通です。

今回のデータでも保存時間の差をデータにすることによって、別の形の二元配置にすることが可能です。今回はあえて保存条件を潰しましたが、保存時間を潰した二元配置も計算してみてください。

なおこのデータを三元配置分散分析で解析すると、次のような結果になります。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 要因 | 平方和 | 自由度 | 平均平方和 | F値 | 有意確率ｐ値 |
| 要因A:被験者 | 670212 | 19 | 35274.3 | 137.877 | 2.22045e-16\*\*\* |
| 要因B:保存条件 | 12954.1 | 1 | 2954.1 | 50.6335 | 9.1192e-07\*\*\* |
| 要因C:放置時間 | 50400.8 | 1 | 50400.8 | 197.002　 |  1.76071e-11\*\*\* |
| AxB | 21917 | 19 | 1153.52 | 4.50878 |  0.000951153\*\*\* |
| AxC | 29609.2 | 19 | 1558.38 | 6.09124 |  0.000123736\*\*\* |
| BxC | 2442.05 | 1 | 2442.05 | 9.54524 |  0.00603416\*\* |
| AxBxC:残差 | 4860.95 | 19 | 255.839 |  |  |
| 全体 | 792396 | 79 |  |  |  |

この結果とSPSSの結果と僕の対応のある二元配置分散分析の結果を見比べると、3者の関係がわかると思います。

謝辞：

　「統計のコツのこつ」は不定期ですがgooブログ（http://blog.goo.ne.jp/）に初心者にも分かりやすいように統計の手法をご紹介しております。統計手法のご紹介ですので統計学的な説明は割愛していますが、幸いにも、杉本典夫先生からそれを補って頂けるコメントがfacebook(情報統計研究所)に寄せられています。数々の貴重なコメントの中から筆者自身が同感した本題を編集し、ここに杉本先生のご厚意を得て投稿させて頂くと共に深甚なる謝意と表します。



 図１　冷暗所と室温に放置したCPK活性値の変化

表１対応のある2元配置分散分析のための計算表



B1h=室温１時間放置、A1h=冷暗所1時間放置

B2h=室温2時間放置、A2h=冷暗所２時間放置

(A)=(B1h＋A2h)、(A)^2＝(A)^2、(B)=(B1h－A2h)の合計、(B)^2=(B)^2

分散分析(平方和)の計算。

修正項：CT＝1018／40＝25908.1　、40＝行数×列数＝20×2

合　計：St＝Σ(B1h＋A2h)^2＋Σ(B1h－A2h)^2＝24636＋59712＝84348

行　　：Sa＝139484／2－CT＝43833.9　、2＝列数

列　　：Sb＝615844／20－CT＝4884.1　、20＝行数

誤差　：Se＝St－Sa－Sb＝84348－43833.9－4884.1＝9721.9



図２　SPSSによる分散分析表



図３　CPK活性値(1時間放置と2時間放置)の相関関係