1. 多元配置分散分析

ここまでの章で，我々は多くの統計解析について学んできました。1 つの名義的な予測変数を用いて，2 つのグループの差 (e.g. t 検定，Chapter **??**) や，3 つ以上のグループの差 (e.g. 一元配置分散分析，Chapter **??**) について統計的検定を行う方法について見てきました。回帰分析の章 (Chapter **??**) では，*複数*の量的な予測変数を用いて単一の結果変数を説明するモデルを建てるという，強力で新しいアイデアが紹介されました。回帰モデルを用いることで，例えば，ある生徒のテスト勉強の時間や IQ テストの得点に基づいて，その生徒の読解テストの誤答数を予測することができます。

本章の目的は，複数の予測変数を使用するというアイデアを，分散分析の枠組みへと拡張することです。例えば，我々が，読解テストを用いて 3 つの異なる学校における生徒の成績を測定しようとしていると考えてみましょう。加えて，我々は，女子と男子が異なる速度で発達している (したがって，成績も平均的に異なることが予想される) と想定しています。各生徒は，彼／彼女らの性別と，所属する学校という 2 つの異なる変数によって分類されます。我々の目的は，これらのグループ化変数の *両方*に基づいて，読解テストの成績を分析することです。これを実現するための手法が，いわゆる**多元配置分散分析**です。ここでは 2 つのグループ化変数があるため，この手法は Chapter **??** で登場した一元配置分散分析に対して，二元配置分散分析と呼ばれることもあります。

1.1 多元配置分散分析 1: 釣合型デザイン，交互作用なし

分散分析について述べた Chapter **??**では，かなり単純な実験計画が想定されていました。各個人は特定のグループに属しており，我々の目的は，いくつかの結果変数について，これらのグループ間で平均値が異なるかどうかを明らかにすることでした。この節では，**多元配置デザイン**と呼ばれる，2 つ以上のグループ化変数を持つより広範な実験デザインについて見ていきます。先ほど，こうしたデザインが必要となるような例を 1 つ挙げました。Chapter **??**で登場した別の例では，各個人の経験した**気分の向上**に対する異なる薬の影響に注目しました。この例では薬の有意な効果が見出されま

したが，章の終盤では，それに加えてセラピーの効果を確認するための分析を行いました。セラピーの効果は見出されませんでしたが，同じ結果変数を予測する 2 つの分析を *別個に*行ったことに対する若干の懸念があります。おそらく，実際にはセラピーによる気分の向上効果は *あるのでしょうが*，その効果は薬の効果によって”隠されて”いたために見つけられなかったのではないでしょうか？言い換えれば，我々は**薬**と**セラピー**の *両方の*予測変数を含む，*単一の*分析を行う必要があります。この分析では，各個人は，投与された薬 (3 水準の要因) および受けたセラピーの種類 (2 水準の要因) という 2 つの要因によって分類されます。こうした分析は 3 × 2 要因デザインと呼ばれます。

JASP の’頻度’–’分割表’の分析を用いて**薬**と**セラピー**のクロス集計表を作成すると，Figure **??** のような表が得られます。

Contingency Tables

	therapy		
drug	CBT	no.therapy	Total
anxifree	3	3	6
joyzepam	3	3	6
placebo	3	3	6
Total	9	9	18

Figure1.1　**薬**と**セラピー**による分割表

集計表から，2 つの要因のあらゆる組み合わせに参加者が属している，すなわちこの分析が**完全交差**デザインであることだけでなく，各グループに同数の参加者が属していることが分かります。言い換えれば，この分析は**釣合型**デザインだということです。これは最も単純なケースであるため，この節では，釣合型デザインのデータをどのように分析するかを見ていきます。非釣合型デザインに関する説明はかなり冗長なので，ここでは一旦置いておくことにします。

1.1.1　検定したい仮説はどんなものか？

多元配置分散分析は，一元配置分散分析と同様に，母集団の平均値に関する仮説を検定するための手法ですしたがって，この分析の仮説が実際にはどのようなものであるかを明確にすることから始めるのが賢明でしょう。しかし，このことについて議論するにあたって，母集団の平均の簡潔な表記法があると非常に便利です。観測値は 2 つの異なる要因に応じて分類されているため，分析者が関心を持ちうる，非常に多くの平均値があります。これを確かめるために，今回のデザインにおいて計

算可能なあらゆるサンプル平均について考えてみましょう。まず，我々は明らかに，以下のようなグループごとの平均値に関心があります：

```
%drug      therapy      mood.gain
薬の種類       セラピーの種類       気分の向上
%placebo   no.therapy   0.300000
プラセボ   セラピーなし   0.300000
%anxifree no.therapy   0.400000
アンザイフリー セラピーなし   0.400000
%joyzepam no.therapy   1.466667
ジョイゼパム セラピーなし   1.466667
%placebo      CBT       0.600000
プラセボ       CBT       0.600000
%anxifree     CBT       1.033333
アンザイフリー      CBT       1.033333
%joyzepam     CBT       1.500000
ジョイゼパム      CBT       1.500000
```

この出力は，2 つの要因のあらゆる組み合わせ (e.g., プラセボ群でセラピーなし，プラセボ群で CBT を実施，など) におけるグループごとの平均値のリストになっています。これらの数値に加えて，行と列の平均および全体の平均を，以下のように 1 つの表で示しておくと便利です：

	セラピーなし	CBT	合計
プラセボ	0.30	0.60	0.45
アンザイフリー	0.40	1.03	0.72
ジョイゼパム	1.47	1.50	1.48
合計	0.72	1.04	0.88

これらの平均値のそれぞれは，当然ながらサンプル統計量です。これらの値は，我々の研究において行われた特定の観察に依存しています。我々が推定したいのは，これらの値と対応する母集団のパラメータです。すなわち，より広範な母集団の中に存在する真の平均です。これらの母平均も同様に表として整理することができますが，そのためには少々，数学的な表記が必要です。ここでは一般的な表記にしたがって，μ を母平均の記号として用います。ただし，表中には様々な平均値があるため，添字を使ってこれらを区別する必要があります。

表記法は次の通りです。この表は 2 つの要因によって構成されています。各行は要因 A(ここでは**薬**) のそれぞれの水準に対応し，各列は要因 B(ここでは**セラピー**) のそれぞれの水準に対応します。R が表中の行数を，C が列数を表すとき，この分析は $R \times C$ 要因の分散分析と表現することができます。ここでは $R = 3$，$C = 2$ となります。小文字を使って特定の行と列を表します。したがって，

μ_{rc} は要因 A の第 r 水準 (i.e. r 行目)，要因 B の第 c 水準 (c 行目) の母平均を表します。[*1]母平均は以下のように表すことができます：

	セラピーなし	CBT	合計
プラセボ	μ_{11}	μ_{12}	
アンザイフリー	μ_{21}	μ_{22}	
ジョイゼパム	μ_{31}	μ_{32}	
合計			

さて，残りの組み合わせについてはどうでしょうか？ 例えば，CBT を受けるかどうかに関わらず，今回のような実験においてジョイゼパムを投与される可能性のある (仮想的な) 母集団全体の平均的な気分の向上について，どのように記述すれば良いでしょうか。これは”ドット”記法によって表すことができます。先ほどのジョイゼパムの例に関しては，表の第 3 行目の値を平均すれば求められることが分かります。すなわち，2 つのセルの平均値 (i.e., μ_{31} および μ_{32}) を平均化するということです。この平均化の結果は**周辺平均**と呼ばれ，この場合には $\mu_{3.}$ と表記されます。CBT の周辺平均は，表の第 2 列目についての母平均と対応するため，$\mu_{.2}$ と表記されます。総平均は，行と列の両方を平均化 (周辺化[*2]) することによって得られる平均値であるため，$\mu_{..}$ と表記されます。母平均についての完全な表は，以下のように書くことができます：

	セラピーなし	CBT	合計
プラセボ	μ_{11}	μ_{12}	$\mu_{1.}$
アンザイフリー	μ_{21}	μ_{22}	$\mu_{2.}$
ジョイゼパム	μ_{31}	μ_{32}	$\mu_{3.}$
合計	$\mu_{.1}$	$\mu_{.2}$	$\mu_{..}$

この表記法によって，仮説を定式化して表現することが容易になります。以下の 2 点を明らかにすることを目指すと考えてみましょう。まず，薬の選択が気分に何らかの影響を及ぼすか？ 次に，CBT は気分に何らかの影響を及ぼすか？ もちろん，定式化することができる仮説はこれらだけではありません。Section **??**において，これらとは別の，非常に重要な仮説の例が示されます。しかし，これらは検定における最も単純な 2 つの仮説であるため，まずはこの 2 つから始めましょう。まず，最初の検定について考えます。もし薬が何の効果も持たないとすると，すべての行平均は同じになる

[*1]添字を使った表記法の良いところは，その一般化可能性です。もし，この実験に 3 つ目の要因が加わったとしても，単に 3 つ目の添字を追加するだけで済みます。原理的には，添字は実験に加えたい要因の数に応じていくつでも拡張することができますが，本書では 2 つ以上の要因を含む分析を扱うことはほとんどないため，添字が 3 つを超えることはありません

[*2]技術的には，周辺化は一般的な平均と全く同一ではありません。周辺化は，平均化しようとする様々なイベントの頻度を加味した加重平均です。しかし，釣合型デザインにおいては，すべてのセルの頻度が定義上等しいため，これらは同じ値になります。後に非釣合型デザインについて説明する際に，この計算が非常に頭痛の種になるものだということが分かるでしょう。ですが，今の所は忘れて構いません。

はずですね？ したがって，これが帰無仮説になります。一方で，薬が何らかの効果を持つとすると，行平均は異なるものになることが予想されます。正式には，これらの帰無仮説および対立仮説は，*周辺平均の等価性*の考え方に沿って書き表されます:

帰無仮説, H_0: 行平均が等しい, i.e., $\mu_{1.} = \mu_{2.} = \mu_{3.}$
対立仮説, H_1: 少なくとも 1 つの行平均が異なる

これらの統計的仮説が，Chapter **??**でこれらのデータに対して一元配置分散分析を行った際の仮説と 全く同じであることは注目に値します。その際には，プラセボ群の平均的な気分の向上を表す表記として μ_P を，2 つの薬のグループ平均を表す表記として μ_A と μ_J を用い，帰無仮説は $\mu_P = \mu_A = \mu_J$ で表されました。ここでも同じ仮説について説明しているのですが，複数のグループ化変数を持つより複雑な分散分析においては，より丁寧な表記が必要なため，ここでは帰無仮説は $\mu_{1.} = \mu_{2.} = \mu_{3.}$ と表されます。しかしながら，後述のように，仮説は同じであるものの，2 つ目のグループ化変数が存在することによって，仮説の検定の仕方は微妙に異なります。

もう一方のグループ化変数に話を移して，2 つ目の仮説検定も同様の方法で定式化できることに気付いたとしても，もはや驚かないでしょう。ただし，今度は薬の効果ではなく心理療法に注目するため，帰無仮説は列平均の等価性に対するものになります：

帰無仮説, H_0: 列平均は等しい, i.e., $\mu_{.1} = \mu_{.2}$
対立仮説, H_1: 列平均は異なる, i.e., $\mu_{.1} \neq \mu_{.2}$

1.1.2 JASP による分析の実行

先ほどの節で説明した帰無仮説と対立仮説には，随分と見覚えがあるように思えます。これらは基本的に，Chapter **??**の一元配置分散分析において検定した仮説と同じです。そのため，多元配置分散分析で用いられる仮説の *検定*も，Chapter **??**で登場した F 検定と本質的には同じであると期待しているのではないでしょうか。平方和 (SS)，平均平方 (MS)，自由度 (df)，そして最終的には p 値に変換することのできる F 統計量を参照する方法が，ここでも使えると思っているのではないですか？ まさにその通りです。そういうわけなので，ここでは前章までとは異なるアプローチを取りたいと思います。本書を通じて，まずは特定の分析の基礎となるロジック (およびある程度の数学的な記述) を説明し，その後に JASP による分析方法の解説を行うアプローチを取ってきました。今回は，これとは逆に，まず JASP でどのように分析を行うかを示します。その理由は，Chapter **??**で説明した単純な一元配置分散分析と，この章で使用するより複雑な分散分析との類似点を強調したいからです。

分析しようとしているデータが釣合型の要因計画に対応している場合，分散分析の実施は容易になります。どれほど容易であるかを確認するため，Chapter **??**で行った分析を再現することから始め

ましょう。忘れてしまった読者のために，この分析では 1 つの要因 (i.e., **薬**) によって結果変数 (i.e., **気分の向上**) を予測しようとし，Figure **??**のような結果を得ている。

ANOVA – mood.gain

Cases	Sum of Squares	df	Mean Square	F	p
drug	3.453	2.000	1.727	18.611	< .001
Residual	1.392	15.000	0.093		

Note. Type III Sum of Squares

Figure1.2　JASP による**気分の向上**を結果変数，**薬**を予測変数とする一元配置分散分析

. . . . . .

加えて，ここでは**セラピー**が**気分の向上**と関係しているかどうかも知りたいと考えます。Chapter **??** で行われた重回帰分析に関する議論を踏まえると，**セラピー**の変数を 2 番目の 'fixed effect' として加えるだけでこの分析ができると知っても驚かないでしょう。Figure **??**を見てください。

ANOVA – mood.gain

Cases	Sum of Squares	df	Mean Square	F	p
drug	3.453	2.000	1.727	26.149	< .001
therapy	0.467	1.000	0.467	7.076	0.019
Residual	0.924	14.000	0.066		

Note. Type III Sum of Squares

Figure1.3　JASP two way ANOVA of `mood.gain` by `drug` and `therapy`

Figure1.4　JASP による**気分の向上**を結果変数，**薬**および**セラピー**を予測変数とする二要因分散分析

. . . . . .

先ほどと同様に，この出力も非常に分かりやすくなっています。表の最初の行は，**薬**の要因に関する群間平方和 (SS) と，対応する群間の *df* を表しています。平均平方 (MS) と *F* 統計量および *p* 値も示されています。同様に，**セラピー**の要因に対応する行と，残差 (i.e., 群内変動) に対応する行があります。

これらの数値はそれぞれ見覚えのあるものでしょうし，これらの数値の関係もまた，一元配置分散分析を行ったときと変わっていません。平均平方は，SS を対応する *df* で割ることによって計算されていることに注意してください。したがって，まだ**薬**や**セラピー**，残差については言及していませ

んが，

$$\text{MS} = \frac{\text{SS}}{df}$$

という関係がここでも成り立ちます。これを確認するために，平方和がどのように計算されるかを気にかける必要はありません。代わりに，JASP が SS を正しく計算してくれたことを信じて，その他の数値の意味についても考えてみましょう。まず，**薬**の要因に関して，3.453 を 2 で割ると，平均平方は 1.727 という値になります。**セラピー**の要因に関しては，自由度が 1 しかなく，計算も容易になります：0.467(SS) を 1 で割ると，0.467(MS) が得られます。

F 統計量と p 値を見ると，それぞれ 2 つずつあることに気づきます；1 つは**薬**の要因に対応し，もう 1 つは**セラピー**の要因に対応しています。どちらの場合も，F 統計量は要因に対応する平均平方の値を残差に対応する平均平方の値で割ることで計算されます。最初の要因 (要因 A；今回の場合は**薬**) を表す省略表記として “A” を，残差を表す省略表記として “R” を用いる場合，要因 A に対応する F 統計量は F_A で表され，以下のように計算されます：

$$F_A = \frac{\text{MS}_A}{\text{MS}_R}$$

また，要因 B(i.e., **セラピー**) についても同様の計算ができます。先ほど表中の行数を表す文字としても R を使用しているので，残差の表記として “R” を用いるのは少し紛らわしいですが，SS_R や MS_R といった文脈でのみ，“R” を残差を表すものとして使用するので，混乱しないよう願います。ともかく，この式を**薬**の要因に適用すると，要因の平均平方 1.727 を残差の平均平方の値 0.066 で割ることになり，26.149 という F 統計量が得られます。**セラピー**の変数については，0.467 を 0.066 で割ることで，7.076 という F 統計量が計算されます。当然ですが，これらは先ほど JASP が分散分析表で報告した値と同じです。

分散分析表には p 値も含まれています。これもまた，特に目新しいことはありません。2 つの要因のそれぞれについて，要因と結果変数の間に関係は無いという帰無仮説を検定します (これについては後ほど詳しく説明します)。そのため，分散分析を行ったときと (明らかに) 同様の方法で，これらの仮説に関する F 統計量を計算しました。これらを p 値に変換するには，帰無仮説 (検討している要因の影響はない) のもとでの F *統計量*の度数分布である F *分布*が必要です。2 つの自由度の値は要因と残差にそれぞれ対応していることにも注目してください。**薬**の要因については，自由度 2 と 14 の F 分布を参照することになります (自由度については後ほど詳しく説明します)。一方，**セラピー**の要因については，自由度 1 と 14 の F 分布を参照することになります。

ここで，このより複雑な要因計画のもとでの分散分析表は，単純な一元配置の分散分析の分散分析表と，ほぼ同様の方法で読み取れることに気づくでしょう。要するに，3×2 要因の多元配置分散分析の結果，薬の有意な効果 ($F_{2,14} = 26.15, p < .001$) およびセラピーの有意な効果 ($F_{1,14} = 7.08, p = .02$) が見出されたことが分かります。あるいは，より専門的で正確な用語を用いると，薬とセラピーの効果という 2 つの**主効果**があるといえます。現時点では，これらを “主” 効果

と呼ぶことはやや冗長に思えますが，これには意味があります。この後，2 つの要因の間に “交互作用” があるという可能性を検討するため，通常は主効果と交互作用効果を区別するのです。

1.1.3 平方和はどのように計算されるか？

ここまでの説明には 2 つの目的がありました。まず，多元配置分散分析を JASP で実行する方法は，一元配置分散分析とほとんど同じであることを示すことです。唯一の違いは，2 つ目の要因の追加です。次に，多元配置分散分析の分散分析表を参照することで，多元配置分散分析の背後にある基本的なロジックと構造が，一元配置分散分析の背後にあるものと同じであることを示すことです。その感覚を大切にしてください。まさしく，多元配置分散分析は，一元配置分散分析とほとんど同じ方法で構成されているのです。分析の詳細について掘り下げ始めると，この感覚は揺らいできます。得てして，この心地よい感覚は，次第に統計学の教科書の著者に対する恨み辛みへと変わっていきます。

それでは，詳細について見ていくことにしましょう。先ほどの節では，主効果 (ここでは薬およびセラピーの) に関する仮説検定が F 検定であることは説明しましたが，平方和 (SS) がどのように計算されるかは示されていませんでした。同様に，自由度 (df) の計算方法も説明されていませんが，こちらは比較的単純です。要因 A と要因 B の 2 つの予測変数があると仮定しましょう。結果変数を Y で表すとき，グループ rc(i.e., r は要因 A に対応する行の水準，c は要因 B に対応する列の水準) に属する i 番目の参加者の反応は Y_{rci} で表すことができます。したがって，$\bar{Y}$ を用いてサンプル平均を表す場合，同様の表記法でグループ平均，周辺平均，総平均を表すことができます。すなわち，$\bar{Y}_{rc}$ は要因 A の第 r 水準，要因 B の第 c 水準に対応するサンプル平均を表し，$\bar{Y}_{r.}$ は要因 A の第 r 水準に関する周辺平均を，$\bar{Y}_{.c}$ は要因 B の第 c 水準に関する周辺平均を，そして $\bar{Y}_{..}$ は総平均を表します。言い換えれば，サンプル平均は母平均と同様の表で整理することができます。今回のデータでは，以下のようになります：

	セラピーなし	CBT	合計
プラセボ	$\bar{Y}_{11}$	$\bar{Y}_{12}$	$\bar{Y}_{1.}$
アンザイフリー	$\bar{Y}_{21}$	$\bar{Y}_{22}$	$\bar{Y}_{2.}$
ジョイゼパム	$\bar{Y}_{31}$	$\bar{Y}_{32}$	$\bar{Y}_{3.}$
合計	$\bar{Y}_{.1}$	$\bar{Y}_{.2}$	$\bar{Y}_{..}$

先ほど示したサンプル平均は，$\bar{Y}_{11} = 0.30$，$\bar{Y}_{12} = 0.60$ などです。今回の例では，**薬**の要因には 3 つの水準が，**セラピー**の要因には 2 つの水準があるため，3×2 要因の多元配置分散分析を実行しようとしていました。より一般的な書き方では，要因 A(行方向の要因) が R 水準，要因 B(列方向の要

因) が C 水準を持ち，$R \times C$ 要因の多元配置分散分析を行うと表現できます。

表記が定まったことで，2 つの要因それぞれの平方和の値を比較的馴染みのある方法で計算することができます。要因 A についての群間の平方和は，(行の) 周辺平均 $\bar{Y}_{1.}$，$\bar{Y}_{2.}$ などが総平均 $\bar{Y}_{..}$ とどの程度異なるかを評価することで計算されます。これには一元配置分散分析と同様の方法が用いられます：$\bar{Y}_{i.}$ と $\bar{Y}_{..}$ の平方和の差を計算するのです。具体的には，各グループに N 人の参加者が属する場合，以下のように計算されます

$$\text{SS}_A = (N \times C) \sum_{r=1}^{R} \left(\bar{Y}_{r.} - \bar{Y}_{..}\right)^2$$

一元配置分散分析と同様に，この数式の中で最も興味深い[a]部分は $\left(\bar{Y}_{r.} - \bar{Y}_{..}\right)^2$ という部分であり，水準 r についての偏差の 2 乗に関連しています。この式が行なっているのは，要因の R 水準すべての偏差の二乗を計算し，足し合わせ，その結果を $N \times C$ に掛けるという計算です。最後の計算を行う理由は，このデザインでは要因 A において r 水準を持つセルが複数あるためです。実際に，要因 B のそれぞれの水準に対応する C 通りのセルがあります。例えば，この例では，薬の**アンザイフリー**という水準に対応する 2 つの異なるセルがあります：1 つは**セラピーなし**のグループ，もう 1 つは `CBT` のグループです。それだけでなく，これらのセルのそれぞれについて，N 個の観測値があります。したがって，SS の値を「観測値ごと」の群間の平方和を表す量に変換するためには，$N \times C$ を掛ける必要があるのです。要因 B についての式は，もちろん，いくつかの添え字が異なる点を除いて同じものになります

これらの式が得られたことで，先ほどの節の JASP の出力と照らし合わせることができます。繰り返しになりますが，こういった計算には専用のスプレッドシートプログラムが役立ちます。

まずは，**薬**の主効果について平方和を計算しましょう。各グループついて，合計 $N = 3$ の参加者がおり，$C = 2$ の異なる種類のセラピーがあります。見かたを変えると，特定の薬を投与された $3 \times 2 = 6$ の参加者がいることになります。スプレッドシートプログラムでこれらの計算を行うと，**薬**の主効果に関する平方和の値は 3.45 となります。驚くべきことではありませんが，これは先ほど Figure **??**で示した分散分析表における薬の要因の SS と同じ値です。

治療の効果についても，同様の計算を行うことができます。先ほどと同じく，各グループには $N = 3$ の参加者がいますが，今度は $R = 3$ の異なる種類の薬があるため，CBT を受けた $3 \times 3 = 9$ の参加者と，セラピーを受けなかった 9 名の参加者がいます。**セラピー**の主効果に関する平方和は，0.47 と計算されます。繰り返しになりますが，計算結果が Figure **??**の分散分析表と同じになることは驚くべきことではありません。

以上が，2 つの主効果の SS の値を計算する方法です。これらの SS の値は，Chapter **??**で一元配置分散分析を行ったときに計算した群間平方和の値と類似しています。ただし，今回は 2 つの異

なるグループ化変数があることで混乱しやすくなるため，それらを群間の SS 値として捉えることはお勧めできません。F 検定を行うためには，群内平方和も計算する必要があります。回帰分析の章 (Chapter **??**) で使用した用語と，そして JASP が分散分析表で出力する用語と合わせるため，群内 SS 値は *残差*平方和 SS_R で表すことにしましょう。

この文脈において，残差 SS 値について考える最も簡単な方法は，それを結果変数における周辺平均の違いを取り除いた (i.e., SS_A および SS_B を取り除いた) 後の，残りの変動として捉えることです。すなわち，SS_T というラベルの付いた，平方和の合計の計算から始めることになります。この計算式は，一元配置分散分析の場合とほぼ同じになります。各観測値 Y_{rci} と総平均 $\bar{Y}_{..}$ の差をとり，差の二乗を合計します

$$SS_T = \sum_{r=1}^{R} \sum_{c=1}^{C} \sum_{i=1}^{N} \left(Y_{rci} - \bar{Y}_{..} \right)^2$$

ここでの「三重総和」は実際以上に複雑に見えます。最初の 2 つの総和は，要因 A のすべての水準 (i.e., 表中の r のすべての行) および要因 B の全ての水準 (i.e., 表中の c のすべての列) を合計しています。各 rc の組み合わせは 1 つのグループに対応し，各グループには N 人の参加者が含まれているため，これらの参加者 (i.e., すべての i の値) を合計する必要があります。つまり，ここで行っているのは，データセット内の全ての観測値 ((i.e., rci の全ての組み合わせ) を合計することです。

ここで，結果変数の総合的な変動である SS_T が明らかになり，その変動のうちどれだけが，要因 A(SS_A) および要因 B(SS_B) に起因するかを知ることができます。したがって，残差平方和は 2 つの要因のいずれにも *起因しない* Y の変動であると定義されます。言い換えれば，

$$SS_R = SS_T - (SS_A + SS_B)$$

もちろん，残差 SS を直接計算するための公式もありますが，上記のように考えることには，より概念的な意味があります。残差という言葉は，それが変動の残りの部分であることを示しており，上記の式はそれを明確にします。「分散分析モデル」に起因する変動である $SS_A + SS_B$ を，(回帰分析の章で用いられていたように)SS_M と表記することも一般的であり，このことから，平方和の総和はモデルの平方和に残差の平方和を加えたものに等しい，という表現がよく使われます。この章の後半において，これは単なる表面的な類似性ではないことが分かります：分散分析と回帰分析が内部で行っていることは，実際に，同じなのです。

いずれにせよ，この式を用いて SS_R を計算し，JASP の出力した分差分析表と同じ答えが得られることを確認することには，時間を割くだけの価値があるでしょう。繰り返しますが，スプレッドシートを利用すると計算は非常に簡単です。上述の式を用いて SS の総和を算出し (SS の総和＝4.85 となります)，次に，残差の SS(＝ 0.92) を求めます。JASP の出力と同じ答えになるはずです。

[a]訳：「最も退屈な」

1.1.4 自由度はどのように求めるか？

自由度は，一元配置分散分析とほぼ同じ方法で計算されます。ある要因について，自由度は水準数から 1 を引いたものに等しくなります (i.e., 行方向の要因 A については $R - 1$，列方向の要因 B については $C - 1$)。したがって，**薬**の要因については $df = 2$，**セラピー**の要因については $df = 1$ となります。後ほど，回帰モデルとしての分散分析モデルの解釈について説明する際に (Section **??**を参照)，この数値の算出方法について詳しく説明します。当面の間は，自由度の単純な定義，すなわち，自由度は観測値の数から制約の数を引いたものに等しいという定義を利用できます。このことから，**薬**の要因については，3 つの個別のグループ平均値が観測されていますが，これらは 1 つの総平均によって制約されるため，自由度は 2 となります。残差の自由度の計算方法は，ロジックは似ていますが，全く同じではありません。今回の実験における観測値の総数は 18 です。制約は，総平均に関するものが 1 つ，**薬**の要因の追加のグループ平均に関するものが 2 つ，**セラピー**の要因の追加のグループ平均に関するものが 1 つあるため，この場合の自由度は 14 となります。式で表すと $N - 1 - (R - 1) - (C - 1)$ となり，$N - R - C + 1$ のように簡略化されます。

1.1.5 多元配置分散分析と一元配置分散分析

ここまで，多元配置分散分析が *どのように*行われるかについて見てきました。ここまでの経過を，一元配置分散分析の結果と比較することには，時間を割くだけの価値があります。そうすることで，*なぜ*多元配置分散分析を行わなければならないかが明らかになります。Chapter **??**では，まず使用した薬による差異を検討するための一元配置分散分析を実行し，次いでセラピーの違いによる差異を検討するための一元配置分散分析を実行しました。Section **??**で述べたように，一元配置分散分析で検定される帰無仮説および対立仮説は，多元配置分散分析で検定される仮説と全く同じです。分散分析表をさらに注意深く見ると，それぞれの分析において，要因に関する平方和の値 (**薬**の要因については 3.45，**セラピー**の要因については 0.92) および自由度の値 (**薬**の要因については 2，**セラピー**の要因については 1) が同じであることが分かります。しかしながら，結果は同じではありません！ 最も注目すべき点は，Section **??**において**セラピー**の要因について一元配置分散分析を行った際には，有意な効果は得られなかったことです (p 値は.21 でした)。

一方で，2 要因分散分析における**セラピー**の主効果に着目すると，有意な効果 ($p = .019$) が得られています。これら 2 つの分析は，明らかに同じではありません。

なぜこのようなことが起こるのでしょうか？ その答えは，*残差*の計算方法を理解することで明らかになります。F 検定の背後にある考え方は，特定の要因に起因する変動と，それらで説明できない変動 (残差) の比較であったことを思い出してください。**セラピー**についての一要因分散分析を実行することは，すなわち，**薬**の効果を無視することになり，薬の要因に由来する変動を残差へと放り込んでしまうことになります！ これによって，データは実際以上に煩雑になり，2 要因分散分析においては正しく有意な効果が見出されている**セラピー**の要因が，有意ではなくなってしまいます。

何かの影響を評価しようとするとき，他の重要な何か (e.g., 薬の要因) を無視してしまうと，分析が歪んでしまいます。もちろん，関心のある現象とはまったく関係のない変数は，無視してしまっても問題ありません。実験室の壁の色を記録しておいて，3 要因分散分析の結果，その要因が重要でないことが判明した場合には，その無関係な要因を除外した，より単純な 2 要因分散分析の結果を報告するだけで十分です。重要なのは，実際には差を生じさせる要因を，分析から除外しないことです！

1.1.6 この分析からどのような結果が得られるか？

ここまで説明してきた分散分析モデルは，我々がデータから発見する可能性のあるさまざまなパターンをカバーしています。例えば，2 要因分散分析デザインでは，4 通りの可能性があります：(a) 要因 A の効果のみがある場合，(b) 要因 B の効果のみがある場合，(c) 要因 A と要因 B の両方の効果がある場合，(d) どちらの要因の効果もない場合です。これら 4 つの可能性のそれぞれの例が，Figure **??**に示されています。

1.2 多元配置分散分析 2: 釣合型デザイン，交互作用あり

Figure **??**に示されている 4 つのパターンは，いずれも現実的なものです。これらのパターンを生じさせるようなデータセットも非常にたくさん存在します。しかしながら，生じうる結果のパターンはこれで全てではなく，また，ここまで説明してきた分散分析モデルは，あらゆるグループ平均のパターンを網羅しているわけではありません。何故でしょうか？ それは，これまでの説明では，薬が気分に影響を与える，セラピーが気分に影響を与える，ということについては議論できますが，両者の**交互作用**を扱うことができないからです。要因 A と要因 B の交互作用は，要因 A の効果が要因 B の水準に応じて *異なる* 場合には，いつでも生じうると言われています。2×2 要因の分散分析における，いくつかの交互作用効果の例を Figure **??**に示します。より具体的な例を挙げると，アンザイフリーとジョイゼパムの作用機序が，全く異なる生理学的メカニズムに依存していると仮定します。ここから，ジョイゼパムがセラピーの有無に関わらず気分に対してほぼ同じ影響をもたらす一方で，アンザイフリーは CBT と組み合わせて投与された場合にはるかに効果的であると考えます。前の章で説明した分散分析では，このアイデアを検討できません。交互作用が生じているかどうかを確かめるには，グループ平均を図示することが有効です。JASP では，分散分析の「Descriptives Plots」オプションを用いて行うことができます–単に，薬を「Horizontal axis」のボックスに，セラピーを「Separate Lines」のボックスに移動するだけです。Figure **??**と同様の図になるはずです。特に注目すべき点は，2 本の線が並行ではないということです。ジョイゼパムが投与された場合 (中央) の CBT の効果 (黒丸の線と白丸の線の差) はゼロに近く，プラセボが用いられた場合の CBT の効果 (右側) よりもさらに小さいようです。しかしながら，アンザイフリーが投与されると，CBT の効果はプラセボよりも大きくなります (左側)。この効果は真実でしょうか，それともランダムな変動に

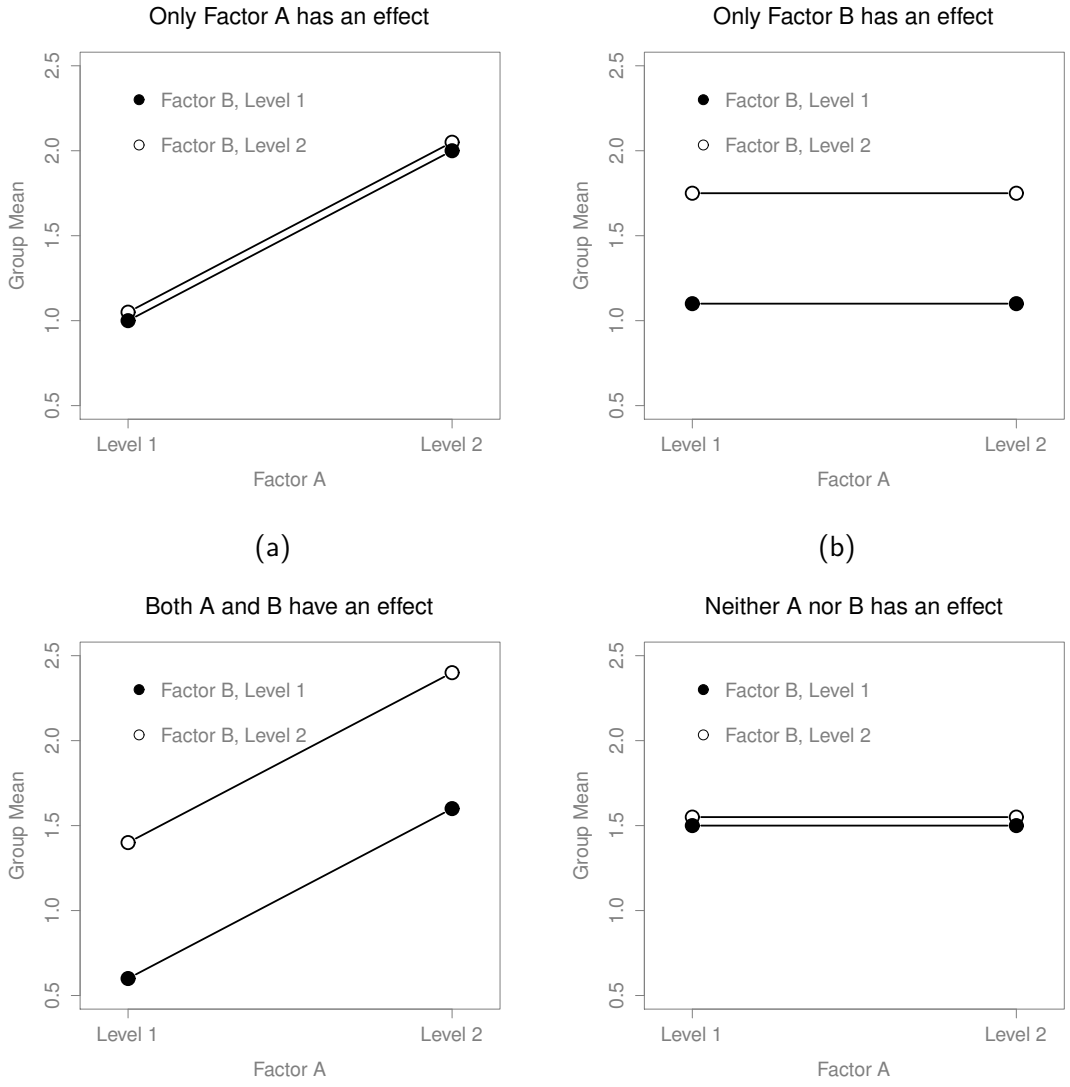

Figure1.5　交互作用のない 2 × 2 要因の分散分析における 4 つの出力。パネル (a) は要因 A の主効果があり，要因 B の主効果がない場合。パネル (b) は要因 A の主効果がなく，要因 B の主効果がある場合。パネル (c) は要因 A，要因 B のどちらの主効果もある場合。パネル (d) はどちらの要因の主効果もない場合。

よる単なる偶然なのでしょうか？ 前章までの分散分析では，この問いに答えることができません。なぜなら，交互作用が存在するというアイデアが含まれていないからです！ 本章では，この問題点を修正していきます。

1.2.1 相互作用とは正確にはどのようなものか?

この節では，交互作用効果という重要なアイデアを紹介します。ここまで見てきた分散分析モデルでは，モデルに含まれる 2 つの *要因* (i.e., **薬**および**セラピー**) にしか着目していませんでした。交互作用を投入することで，モデルに新たな要素が追加されることになります：**薬** と **セラピー**の組み合わせです。直観的には，交互作用効果の背後にある考え方は非常に単純です。交互作用は単に，要因 A の影響が，要因 B の水準に応じて変化することを意味しています。しかし，このことは我々のデータに対してどのような意味を持っているのでしょうか？ Figure **??**に示したいくつかの図は，それぞれ全く見た目が異なりますが，すべて交互作用効果として扱われます。したがって，この定性的なアイデアを，統計学者が扱うような数学的な記述に変換することは非常に困難です。

結論として，交互作用効果の概念を，対立仮説と帰無仮説の観点から定式化することは困難であり，ましてやこの本の読者の多くは，それほど興味がないと思います。それでも，基本的なアイデアを示しておこうと思います。

まずは，主効果についてもう少し明示的にする必要があります。要因 A(今回の分析例では**薬**) の主効果について考えます。そもそも主効果は，2 つの周辺平均 $\mu_{r.}$ が全て等しいという帰無仮説の観点に基づいて定式化されていました。これらの周辺平均の全てが等しいとすると，それらは総平均 $\mu_{..}$ とも等しくなければなりませんね？ したがって，ここでは水準 r における要因 A の *効果* を，周辺平均 $\mu_{r.}$ と総平均 $\mu_{..}$ との差に等しいものとして定義します。

この効果を α_r で表し，以下のように表記します

$$\alpha_r = \mu_{r.} - \mu_{..}$$

ここで，周辺平均 $\mu_{r.}$ の平均値が総平均 $\mu_{..}$ になることと同様の理由で，定義上，すべての α_r の値は合計がゼロになる必要があります。同じように，水準 i における要因 B の効果を，列方向の周辺平均 $\mu_{.c}$ と総平均 $\mu_{..}$ の差として定義することができます。

$$\beta_c = \mu_{.c} - \mu_{..}$$

繰り返しになりますが，これらの β_c の合計はゼロにならなければなりません。統計学者が α_r や β_c の値を用いて主効果について説明することを好む理由は，交互作用効果がないということの意

味を正確に伝えることができるからです。交互作用がまったくない場合，α_r および β_c の値を用いて，グループ平均 μ_{rc} を完全に記述することができます。具体的には以下のようになります

$$\mu_{rc} = \mu_{..} + \alpha_r + \beta_c$$

これは，グループ平均に関して *特別*なことは何もない，すなわち，すべての周辺平均が明らかになっても，完全な予測ができないということを意味します。これはまさに，帰無仮説を表しています。対立仮説は，表中の少なくとも一つのグループ rc において

$$\mu_{rc} \neq \mu_{..} + \alpha_r + \beta_c$$

が成り立つこと，と表現できます。統計学者はしばしば，上記の式をやや異なる形式で表現します。彼らは通常，グループ rc に関連づけられた交互作用をいくつかの番号によって定義し，厄介なことに $(\alpha\beta)_{rc}$ と表現し，そして対立仮説を次のように表ます

$$\mu_{rc} = \mu_{..} + \alpha_r + \beta_c + (\alpha\beta)_{rc}$$

ここで，少なくとも 1 つのグループの $(\alpha\beta)_{rc}$ は非ゼロです。この表記法はやや見苦しいですが，次の節で説明するように，二乗和の計算方法を説明する際には便利です。

1.2.2 交互作用の二乗和の計算

交互作用項 $\mathrm{SS}_{A:B}$ の二乗和はどのように計算すれば良いでしょうか？ まず，先ほどの節において，実際のグループ平均が周辺平均から予測された値とどの程度異なるかという観点から，交互作用効果をどのように定義したかについて確認すると良いと思います。もちろん，これらの式はすべて，サンプル統計量ではなく母集団のパラメータに関するものであるため，実際にそれらがどのようなものであるかは分かりません。しかしながら，母平均の代わりにサンプル平均を用いることで，それらを推定することができます。要因 A に関して，水準 r における主効果を推定するための良い方法は，*サンプル*の周辺平均 $\bar{Y}_{rc}$ とサンプルの総平均 $\bar{Y}_{..}$ の差に着目することです。そこで，これを効果の推定値として用います

$$\hat{\alpha}_r = \bar{Y}_{r.} - \bar{Y}_{..}$$

同様に，水準 c における要因 B の主効果の推定値は，以下のように定義できます

$$\hat{\beta}_c = \bar{Y}_{.c} - \bar{Y}_{..}$$

ここで，2 つの主効果の SS の値について説明した式を改めて見てみると，それらの効果に関する項は二乗され足し合わされているということに気づくでしょう！ それでは，交互作用項ではどう

なっているでしょうか？ その答えは，以下のように，対立仮説のもとでグループ平均 μ_{rc} に関する式を再変形することで明らかになります

$$\begin{aligned}(\alpha\beta)_{rc} &= \mu_{rc} - \mu_{..} - \alpha_r - \beta_c \\ &= \mu_{rc} - \mu_{..} - (\mu_{r.} - \mu_{..}) - (\mu_{.c} - \mu_{..}) \\ &= \mu_{rc} - \mu_{r.} - \mu_{.c} + \mu_{..}\end{aligned}$$

そして，母平均の代わりにサンプル統計量を代入すると，グループ rc における交互作用効果の推定は以下のようになります

$$\hat{(\alpha\beta)}_{rc} = \bar{Y}_{rc} - \bar{Y}_{r.} - \bar{Y}_{.c} + \bar{Y}_{..}$$

ここで，これらの要因 A における R 水準，および要因 B における C 水準のすべての推定値を足し合わせることで，全体的な相互作用に関する二乗和の式が得られます

$$\mathrm{SS}_{A:B} = N \sum_{r=1}^{R} \sum_{c=1}^{C} \left(\bar{Y}_{rc} - \bar{Y}_{r.} - \bar{Y}_{.c} + \bar{Y}_{..}\right)^2$$

各グループについて N 個の観測値があるため，N が掛けられています。SS の値には，グループ間の変動ではなく，交互作用によって説明される *観測値*間の変動が反映されていると期待されます。

$\mathrm{SS}_{A:B}$ の計算式が準備できたので，交互作用項がモデルの一部であることを認識する必要があります（当然のことですが）。モデルにおける全体の平方和である SS_M は，関連する 3 つの SS 値の合計 $\mathrm{SS}_A + \mathrm{SS}_B + \mathrm{SS}_{A:B}$ に等しくなります。残差平方和 $\mathrm{SS}_T - \mathrm{SS}_M$ は，残りの変動，すなわち $\mathrm{SS}_T - \mathrm{SS}_M$ として定義されますが，交互作用項があることから，以下のようになります

$$\mathrm{SS}_R = \mathrm{SS}_T - (\mathrm{SS}_A + \mathrm{SS}_B + \mathrm{SS}_{A:B})$$

結果として，残差平方和 SS_R は，交互作用を含まない分散分析よりも小さくなります。

1.2.3 交互作用における自由度

交互作用における自由度の計算は，主効果における計算よりも少しだけ複雑です。まずは，分散分析モデルの全体について考えてみましょう。モデルに交互作用効果が含まれる場合，すべてのグループに，独自の平均 μ_{rc} を持つことが許されます。$R \times C$ 要因の分散分析の場合には，これはモデル中に $R \times C$ 通りの統計量と，総平均はすべてのグループ平均の平均値であるという，たった 1 つの制約があることを意味します。そのため，モデル全体としては $(R \times C) - 1$ の自由度が必要です。しかし，要因 A の主効果には $R - 1$ の自由度が，要因 B の主効果には $C - 1$ の自由度があります。このことは，交互作用に関する自由度が，

$$
\begin{aligned}
df_{A:B} &= (R \times C - 1) - (R - 1) - (C - 1) \\
&= RC - R - C + 1 \\
&= (R - 1)(C - 1)
\end{aligned}
$$

という式で表されるように，行の要因と列の要因の自由度の積にすぎないということを意味します。残差の自由度についてはどうでしょうか？ 交互作用項によってある程度の自由度が吸収されるため，残りの自由度は少なくなります。具体的には，交互作用を含むモデルが全部で $(R \times C) - 1$ の自由度を持ち，1 つの総平均を満たすよう制約されているデータセット内に N 個の観測値があるとき，残差の自由度は $N - (R \times C) - 1 + 1$，あるいは単に $N - (R \times C)$ となる。

1.2.4 JASP による分散分析の実行

JASP の分散分析モデルに交互作用項を加えることは難しくありません。というよりも，交互作用項は分散分析のデフォルトのオプションであるため，何もする必要はありません。すなわち，例えば**薬**と**セラピー**という 2 要因の分散分析を実行すると，これらの交互作用項- `drug*therapy` -が自動的にモデルに追加されるということです。[*3]交互作用項を含めた分散分析を実行すると，**??**のような結果が得られます。

結局，今回の分析では，有意な薬の主効果 ($F_{2,12} = 31.7, p < .001$) とセラピーの主効果 ($F_{1,12} = 8.6, p = .013$) が見出されますが，交互作用は有意ではありません ($F_{2,12} = 2.5, p = 0.125$)。

1.2.5 結果の解釈

多元配置分散分析の結果の解釈の際には，いくつかの非常に重要なポイントがあります。まずは，仮に (例えば) **薬**の有意な主効果が得られたとしても，どの薬が他と異なるかについては何も分からないという，一元配置分散分析と同様の問題です。これを明らかにするためには，追加の分析を行う必要があります。Sections **??**および**??**において，この追加の分析のいくつかを紹介します。交互作用効果についても，同様のことがいえます。有意な交互作用がみられたとしても，どのようなパターンの交互作用が存在しているかについては何も分かりません。ここでも，追加の分析を行う必要があります。

次に，有意な交互作用効果が得られているが，主効果が有意ではないという場合に，解釈が非常にややこしくなるという問題があります。このような結果はときどき生じます。例えば，Figure **??**で示されている交差したパターンの交互作用が，実際の結果でも生じる可能性があります。このケースでは，主効果はいずれも有意ではありませんが，交互作用効果が存在します。これは解釈が難しい状

[*3]先ほどの節で説明した，主効果に関する分析を実際に JASP で再現してみた読者は，すでにこの出力を目にしているかもしれません。説明を単純にするため，先述のモデルからは交互作用項を除外しています

況であり，多くの分析者を混乱させます。こうした状況に対して，統計学者が行いがちな一般的なアドバイスは，交互作用が存在する場合，主効果にはあまり注目すべきでないということです。彼らがこのように述べる理由は，主効果の検定は数学的観点から全く正しいものですが，有意な交互作用がみられる場合には，主効果が重要な仮説を検定していることは稀だからです。Section **??**を思い出してみると，主効果の帰無仮説は *周辺平均*が相互に等しいというものであり，周辺平均はいくつかの異なるグループ間の値を平均することで計算されていました。交互作用が有意であるということは，周辺平均を構成するグループが同種ではないということが *自明*になるため，これらの周辺平均について気にする必要は無くなります。

つまり，こういうことです。改めて，臨床的な例を用いて説明しましょう。2 × 2 要因デザインで，2 種類の恐怖症の治療法 (e.g., 系統的脱感作法と暴露療法)，および 2 種類の不安軽減薬 (e.g., アンザイフリーとジョイゼパム) の比較を行うと考えてみましょう。そして，治療として脱感作法を行った場合，アンザイフリーには効果がなく，治療として暴露療法を行った場合，ジョイゼパムには効果がないという結果が見出されたと仮定します。いずれの薬も，もう一方の治療においては効果的であるとします。これは古典的な交差パターンの交互作用であり，分散分析を実行すると，**薬**の主効果はありませんが，有意な交互作用が見出されます。さて，主効果がないということは，一体何を *意味*するのでしょうか？ もちろん，2 つの異なる治療法を平均すると，アンザイフリーとジョイゼパムの *平均的な*効果は同じであるということです。しかし，なぜ誰もがこのことを気にかけるのでしょうか？ 恐怖症の治療において，暴露療法と脱感作法を「平均的に」使用することなどできません。これはあまり意味のない考え方です。暴露療法か脱感作法のどちらかを選ぶことになります。一方の治療法では一方の薬が効果的であり，もう一方の治療法ではもう一方の薬が効果的なのです。ここで重要なのは交互作用であり，主効果はある意味で無関係です。

このような事態はしばしば生じます。主効果は周辺平均の検定であり，交互作用が存在する場合には，周辺平均には注目する必要がなくなることがよくあります。なぜなら，周辺平均が，本来平均すべきではないものを平均化してしまっていることが，交互作用によって示されるからです！ もちろん，交互作用が存在するからといって，主効果が無意味であるとは限りません。多くの場合，大きな主効果と，非常に小さな交互作用が得られます。このとき，「薬 A は一般に薬 B よりも効果的である」(大きな薬の主効果があるため) と主張することができますが，次のような表現を加えて主張を微修正する必要があります。「A と B の薬の効果の差は，治療法によって異なった」。いずれにせよ，ここでの主要なポイントは，有意な交互作用が得られたときには，常に一度立ち止まって，その分析の文脈において，主効果が実際には何を意味しているのかを *考える*ことです。主効果が重要であると自動的に思い込んではいけません。

1.3

効果量

多元配置分散分析の効果量の計算は，一元配置分散分析で用いられるものとかなり似ています (Section **??**を参照)。具体的には，特定の項に対する全体的な効果がどれほど大きいかを測る簡単な方法として，η^2 を利用できます。前回同様，η^2 はその項に関連づけられた平方和を，総平方和で割ることで定義されます。例えば，要因 A の主効果の効果量を求めるには，以下の式が用いられます：

$$\eta^2_A = \frac{\mathrm{SS}_A}{\mathrm{SS}_T}$$

この値が，回帰分析における R^2 とほぼ同様に解釈できるという点も，前回と同じです。[*4] この値は，要因 A の主効果によって説明することができる結果変数の分散の割合を表しています。したがって，この値は 0(影響なし) から 1(結果変数の変動の *全て*) の範囲を取ります。さらに，モデル内の全ての項から得られる全ての η^2 値の合計は，分散分析モデルの合計の R^2 と等しくなります。例えば，分散分析モデルが完全に適合している場合 (i.e., グループ内の変動が全くない場合！)，η^2 値は 1 になります。もちろん，このような事態が実生活で生じることは滅多にありません。

ただし，多元配置分散分析を行う際には，偏 η^2 という 2 つ目の効果量の指標が好んで報告されます。偏 η^2($_p\eta^2$ もしくは η^2_p と表記される場合もあります) の背後にある考え方は，特定の項についての効果量 (例えば，要因 A の主効果) を求める場合に，モデル内の他の効果 (e.g., 要因 B の主効果) を意図的に無視するというものです。すなわち，これらの他の全ての項の効果がゼロであると仮定して，η^2 値がどうなるかを計算するということです。これは実際には非常に計算が簡単です。分母から他の項に関する平方和を削除すればよいのです。つまり，要因 A の主効果についての偏 η^2 を求める際には，分母は要因 A の SS 値と残差の合計になります。

$$\text{partial } \eta^2_A = \frac{\mathrm{SS}_A}{\mathrm{SS}_A + \mathrm{SS}_R}$$

この結果は常に η^2 よりも大きな数値となります。私は皮肉屋なので，これが偏 η^2 の人気の理由だと考えています。偏 η^2 値もまた 0 から 1 の範囲となり，0 は効果がないことを表します。ただし，偏 η^2 値の大きさについての解釈は少々ややこしくなります。特に，各項の偏 η^2 値を直接比較することができない点には注意が必要です！ たとえば，グループ内の変動性が全くないとすると，$\mathrm{SS}_R = 0$ となります。このことが意味するのは，すべての項の偏 η^2 値が 1 になるということです。しかしながら，これはモデル内のすべての項が等しく重要である，あるいは，それらの大きさが等しいということではありません。これは，モデル内のすべての項のが，*残差の変動と比*べて大きな効果量を持つことを意味します。各項について比較することはできません。

[*4] この章は，文字 R で様々なものを表現することに関して，新記録を打ち立てているかもしれません。これまでのところ，ソフトウェアパッケージ，平均値の表における行の数，モデルの残差，そして回帰における相関係数を指して，R という文字を使っています。大変申し訳ないと思っています。アルファベットの文字数が十分ではないことは明らかです。我々も，R がそれぞれの文脈において指示しているものを明確にするために，かなりの努力を要しているということを申し添えます

このことは，具体例を見てみると分かりやすいでしょう。まず，Figure **??**において，交互作用項を含まない分散分析の効果量について見てみましょう：

	eta.sq	partial.eta.sq
%drug	0.71	0.79
薬	0.71	0.79
%therapy	0.10	0.34
セラピー	0.10	0.34

η^2 の値に着目すると，**薬**の要因が**気分の向上**の分散の 71%(i.e. $\eta^2 = 0.71$) を占めているのに対し，**セラピー**は 10% です。これにより，合計で 19% の変動が考慮されないままとなります (つまり，結果の変動の 19% が残差で構成されます)。全体的に見て，この結果は**薬**には非常に大きな効果[*5]があり，**セラピー**にはわずかな効果があったことを意味します。次に，Figure **??**に示されている偏 η^2 値を見てみましょう。

セラピーの効果はそれほど大きくないため，それを調整しても大きな違いはありません。したがって，**薬**の偏 η^2 は，${}_p\eta^2 = 0.79$ とそれほど増加しません。対照的に，**薬**の効果は非常に大きかったため，調整することで大きな違いが生じます。**セラピー**の偏 η^2 を計算すると，${}_p\eta^2 = 0.34$ まで上昇していることが分かります。自問しなければならないのは，これらの偏 η^2 の値が実際に何を *意味する* のか？ ということです。要因 A の主効果に対する偏 η^2 を解釈する一般的な方法は，それを要因 A のみを変化させた仮想実験の記述として解釈することです。*実際*の実験では要因 A と B の両方を変化させましたが，要因 A のみを変化させた実験についても簡単に想像することができます。偏 η^2 統計量は，そのような実験で得られることが予想される，結果変数の分散の量を表します。ただし，このような解釈は，その他の主効果に関する多くの事項と同様に，有意な交互作用効果が存在する場合には，あまり意味がないということに注意が必要です。

交互作用効果といえば，Figure **??**のように，交互作用項を含むモデルの効果量を計算したときに得られるものです。JASP では，単に ‘Additional Options’ - ‘Estimates of effect size’ を選択し，必要な変数を選ぶことで計算されます。見ての通り，主効果の η^2 値は変化しませんが，偏 η^2 値は変化します：

	eta.sq	partial.eta.sq
%drug	0.71	0.84
薬	0.71	0.84
%therapy	0.10	0.42
セラピー	0.10	0.42
%drug*therapy	0.06	0.29
薬*セラピー		

[*5] 信じられないほどの大きさです。このデータの不自然さが見えてきましたね！

1.3.1 推定グループ平均

多くの場合，分散分析の結果と，それに関連する信頼区間に基づいて，すべてのグループ平均の推定値を報告する必要があります。JASP では，Figure **??**にあるように，分散分析の 'Additional Options' - 'Marginal Means' の機能を用いてこれを行うことができます。交互作用項**薬*セラピー**を，アクションボックスに移動するだけです。実行した分散分析が**飽和モデル** (i.e., 考えられるすべての主効果と交互作用効果を含むモデル) である場合，グループ平均の推定値はサンプル平均とまったく同じになります。重要なのは，信頼区間は，グループごとの個別の標準誤差を使用するのではなく，プールされた標準誤差の推定値を用いるということです。

結果を見ると，セラピーを行わなかったプラセボ群における気分向上の推定平均は 0.300 であり，95% 信頼区間は 0.006 から 0.594 でした。各グループについて信頼区間を計算しても，同じ値にはならないということに注意してください。これは，分散分析モデルが分散の均一性を仮定しているので，プールされた標準偏差の推定値を使用するためです。

1.4 仮定の確認

一元配置分散分析と同様に，多元配置分散分析においても，分散の等質性 (すべての群の標準偏差が等しい) ，残差の正規性，観測の独立性の 3 つが主要な仮定となります。前の 2 つの仮定については，確認することができます。3 つ目の仮定については，測定値間に何らかの特別な関係性が存在するかどうか，自分自身で評価しなければなりません。例えば，時間を独立変数とする反復測定では，時点 1 と時点 2 の観測変数は *同じ*人物から測定されているため，関係があります。加えて，飽和モデルを使用していない場合 (例えば，交互作用項を省略している場合) には，省略されている項は重要ではないという仮定を置いていることになります。この最後の仮定については，省略された項を含めた分散分析を実行し，それらが有意であるかどうかを確認できるため，チェックすることは比較的容易です。分散の等質性と残差の正規性についてはどうでしょうか？ 結論からいうと，これらをチェックするのはとても簡単です。一元配置分散分析で行ったチェックの方法となんら変わりません。

1.4.1 分散の等質性

Section **??** で述べたように，異なる群やカテゴリ間で標準偏差のプロットを視覚的に比較し，Levene の検定の結果と一致するかどうかを確認するのは良いアイデアです。Levene の検定の理論については，Section **??** で説明したのでここでは触れません。この検定では，モデルが飽和モデル (i.e., すべての項を含む) であることが期待されています。なぜなら，この検定は主に群内の分散に関係しており，飽和モデル以外について適用しても，実際のところあまり意味がないからです。Levene の検定は，JASP の 'Assumption Checks' - 'Homogeneity tests' オプションで指定でき，その結果は Figure **??** のようになります。Levene の検定が有意でないということは，標準偏差のプロットの目視による確認との矛盾がなければ，分散の等質性の仮定には違反していないと考えて良いことになります。

1.4.2 残差の正規性

一元配置分散分析と同様に，残差の正規性を簡単な方法で検定できます (Section **??**を参照)。しかし，一般的には，QQ プロットを用いて残差を視覚的に調べるのが良いと思います。Figure **??** を見てください。

1.5 共分散分析 (ANCOVA)

共分散分析は，分散分析の応用として，従属変数に関連していると思われる，追加の連続変数があるような場合に用いられる分析です。この追加の変数は，共分散分析 (ANCOVA) という名が示す通り，共変量として分析に加えることができます。

共分散分析では，従属変数の値を共変量の影響力によって”調整”し，”調整後”の得点の平均値について，通常の方法で群間での検定を行います。この手法は，実験の精度を高めることができるため，従属変数における群の平均値の同等性について，より”強力な”検定を行うことができます。共分散分析はどのようにしてこれを行うのでしょうか？ 共変量そのものに関しては，通常，実験的にはなんの関心もありませんが，共変量を調整することにより，実験的な誤差の推定値を減少させることができます。そして，誤差分散を減少させることで，精度を高めることができるのです。これはすなわち，帰無仮説を棄却できないエラー (偽陰性あるいは第二種の過誤) が起こりにくいということを意味します。

このような利点に対して，共分散分析には群間の真の差を見えなくしてしまうというリスクもあ

り，これは避けなければなりません。例えば，年齢ごとに統計に対する不安をプロットした Figure **??** を見てみると，文系と理系という異なる背景あるいは専攻を持つ学生からなる，2 つの群があることが分かります。このとき，年齢を共変量とした共分散分析では，統計に対する不安は両群で差がないという結論になるかもしれません。この結論は妥当なものでしょうか？ 2 つの群の年齢は重なっておらず，分散分析は本質的に，”データのない範囲に外挿された”(**Everitt1996**) ことになるため，おそらく妥当ではないでしょう。

異なる群に対して共分散分析を適用する際には，慎重に検討する必要があることは明らかです。共分散分析は一元配置と多元配置のどちらのデザインにも使用できるため，どちらの計画でも注意が必要です。

1.5.1 JASP による共分散分析の実行

ある健康心理学者が，年齢を共変量として，日常的なサイクリングとストレスが幸福度に与える影響に関心を持っているとします。このデータセットは `ancova.csv` で見ることができます。このファイルを JASP で開き，共分散分析を行うには，‘ANOVA’ - ‘ANCOVA’ を選択して，共分散分析のウィンドウを開きます (Figure **??**)。従属変数‘happiness’を選択し，‘従属変数’ボックスに投入します。独立変数‘stress’と‘commute’を‘固定要因’ボックスに投入します。共変量‘age’を‘共変量’ボックスに投入します。次に，‘追加のオプション’ - ‘周辺平均’ をクリックし，交互作用項 `stress*commute` をボックスに投入します。

結果ウィンドウに共分散分析表が表示されます (Figure **??**)。共変量‘age’の F 値は $p = .023$ で有意であり，年齢が従属変数 `happiness` の重要な予測因子であることが示唆されました。推定された周辺平均値を見ると，今回の共分散分析には共変量に年齢が含まれているため，（共変量なしの分析と比較して）調整が行われています。結果のプロット（Figure **??**）は，有意な効果を視覚化して解釈するのに適した方法です。

主効果‘stress’の F 値 (52.61) に付随する確率は $p < .001$ です。主効果‘commute’の F 値 (42.33) に付随する確率は $p < .001$ となっています。これらはいずれも。検定結果が有意であるかどうかを判断する際に一般的に用いられる確率 ($p < .05$) よりも小さいため，ストレスの主効果 ($F(1, 15) = 52.61, p < .001$) と通勤方法の主効果 ($F(1, 15) = 42.33, p < .001$) が有意であったと結論づけることができます。また，ストレスと通勤方法の間には有意な交互作用も見られます ($F(1, 15) = 14.15, p = .002$)。

Figure **??** は，年齢を共変量とした共分散分析によって調整された，周辺化した平均幸福度の得点を示しています。この分析では，有意な交互作用効果があり，自転車で通勤するストレスの低い人は，車で通勤するストレスの低い人や，自転車でも車でもストレスの高い人よりも幸福度が高いという結果になりました。また，ストレスの主効果が大きく，ストレスの少ない人はストレスが多い人よ

り幸福であることが分かります。加えて，通勤方法の主効果も大きいことから，自転車で通勤する人は車で通勤する人よりも平均して幸福度が高いことも分かります。

1.6 線形モデルとしての分散分析

分散分析と回帰分析については，それらが基本的に同じものであるということを理解することが重要になります。表面的には，そんなことはないと思われるかもしれません。これまでの説明では，分散分析は主としてグループ間の差の検定を，回帰分析は主として変数間の相関関係の理解を目的としていました。これらの説明は完全に正しいといえます。しかし，フードの下を覗いてみれば，分散分析と回帰分析の基本的な仕組みは非常によく似通っています。実際，すでにその証拠については述べられています。分散分析と回帰分析はどちらも二乗和 (SS) に大きく依存し，またどちらも F 検定を利用しています。振り替えってみると，Chapters **??** および **??**の内容は，重複していた感じが否めません。

その理由は，分散分析も回帰分析も，どちらも**線形モデル**だからです。回帰分析については，このことは自明です。予測変数と結果変数の関係を定義するために用いる回帰式は線形方程式であり，これは明らかに線形モデルです。

$$Y_p = b_0 + b_1 X_{1p} + b_2 X_{2p} + \varepsilon_p$$

上記の式において，Y_p は p 番目の観測値 (e.g., p 人目の参加者) における結果変数の値，X_{1p} は p 番目の観測値における 1 つ目の予測変数の値，X_{2p} は p 番目の観測値における 2 つ目の予測変数の値，b_0, b_1, および b_2 の項は回帰係数，ε_p は p 番目の残差を表します。残差 ε_p を無視して，回帰直線そのものに着目すると，以下のようになります：

$$\hat{Y}_p = b_0 + b_1 X_{1p} + b_2 X_{2p}$$

ここで，$\hat{Y}_p$ は，実際に観測された値 Y_p ではなく，参加者 p について回帰直線に基づいて予測された Y の値になります。分散分析を線形モデルとして記述することもできます。少々イメージし難いかもしれませんが，実際には非常に簡単です。まず，極めて単純な例として，2×2 要因の分散分析を線形モデルに書き換えてみましょう。

1.6.1 データ

具体例として，私の講義での学生の**成績**が結果変数であるとしましょう。これは 0% から 100% までの比率尺度の変数です。関心のある予測変数は，学生が講義に出席したかどうか (**出席** 変数)，および学生が教科書を読んだかどうか (**予習** 変数) の 2 つです。ここでは，講義に出席した場合を**出席**

= 1，出席しなかった場合を**出席 = 0**とします。同様に，学生が教科書を読んでいれば**予習 = 1**，読んでいなければ**予習 = 0**とします。

さて，ここまでは順調ですね。ここから，いくらか数学的な処理を施す必要があります (すみません！) 今回の例では，Y_p がクラスの p 番目の学生の**成績**を表すとします。これは，本章の前半で使用した表記法とは全く異なります。以前は，予測変数 1(行の要因) の r 番目のグループと予測因子 2(列の要因) の c 番目のグループにおける i 番目の参加者を指して Y_{rci} という表記を使っていました。この拡張された表記法は，SS 値の算出方法を説明するのに非常に便利だったのですが，現在の文脈においては面倒なので，表記を切り替えます。さて，Y_p という表記は Y_{rci} よりも視覚的にシンプルですが，グループについての情報を含まないという欠点があります。例えば，$Y_{0,0,3} = 35$ という表記であれば，それは講義に出席せず (i.e., **出席 = 0**)，教科書を読まず (i.e., **予習 = 0**)，落第してしまった (**成績 = 35**) 学生のことである (そして，この学生は 3 人目である) とすぐに分かるでしょう。しかし，$Y_p = 35$ という表記では，p 番目の学生が良い成績を取れなかったということしか分かりません。重要な情報が失われてしまっているのです。これを解決することは，それほど難しくありません。この情報を記録するための，2 つの新しい変数 X_{1p} と X_{2p} を導入すればよいのです。この学生の場合，$X_{1p} = 0$ (i.e., **出席 = 0**)，$X_{2p} = 0$ (i.e., **予習 = 0**) であることが分かります。そのため，データは以下のようになります：

参加者, p	成績, Y_p	出席, X_{1p}	予習, X_{2p}
1	90	1	1
2	87	1	1
3	75	0	1
4	60	1	0
5	35	0	0
6	50	0	0
7	65	1	0
8	70	0	1

もちろん，これは特別なことではありません。まさに我々の期待するデータのフォーマットです！データファイルは `rtfm.csv` です。

1.6.2 二値の要因を用いた分散分析の回帰モデルとしての表現

さて，数学の話に戻りましょう。このデータには，連続変数 Y と 2 つの二値変数 X_1 および X_2 という 3 つの変数が含まれます。この 2×2 要因の分散分析は，まさに回帰モデルと同じであることに気付いていただきたいと思います。

$$Y_p = b_0 + b_1 X_{1p} + b_2 X_{2p} + \varepsilon_p$$

上記の式はもちろん，先ほど 2 つの予測変数の回帰モデルを表す際に用いたものと全く同じ式です。唯一の違いは，回帰分析では X_1 および X_2 は連続的な値ですが，この式では両者は *二値*変数 (i.e., 0 または 1 の値をとる) になっていることです。このことについて納得する方法はいくつかあります。1 つ目は，両者の式が等しいことを証明するための，長い長い数学的演習を行う方法です。ただ，おそらく本書の読者の多くは，それを参考にするどころか，むしろ迷惑に感じるのではないかと思います。その代わりに，基本的な考え方について説明したうえで，分散分析と回帰分析が似ているというだけでなく，どこから見ても同じであるということを，JASP に頼って説明することにしましょう。まず，このデータを用いて分散分析を行ってみましょう。`rtfm` データセットを使用して，JASP で分散分析を行った結果を Figure **??**に示します。

分散分析表と平均値から主要な数値を読み取ると，授業に出席した方が ($F_{1,5} = 21.6, p = .006$)，また教科書を読んでいた方が ($F_{1,5} = 52.3, p < 0.001$)，より高い成績を得られていることが分かります。この p 値と F 統計量をメモしておきましょう。

では，線形回帰の観点から同じ分析について考えてみましょう。`rtfm` データセットにおいて，**出席**および**予習**の変数は，あたかも数値的な予測変数であるかのようにコード化されています。今回の場合，これは全く問題ありません。授業に出席した学生 (i.e., `attend = 1`) は，そうでない学生 (i.e., `attend = 0`) と比べて，“より出席している” という意味になるのですから。したがって，これらを回帰モデルの予測変数として加えることは全く不合理ではありません。予測変数が 2 つの可能な値を取るため，少々変わってはいますが，線形回帰の仮定に反するものではありません。加えて，解釈も容易になります。**出席**の回帰係数が 0 より大きければ，講義に出席している学生の方が成績が良いことを意味します。もし 0 より小さければ，講義を受けている学生の方が成績が低くなります。**予習**の変数についても同様です。

しかし，ちょっと待ってください。これは *なぜ*正しいのでしょうか？ 統計学の講義を複数回受けたことがあり，数学に慣れている人にとっては直感的に分かることかもしれませんが，それ以外の人には初見では分かりません。その理由を知るために，特定の学生に目を向けてみましょう。まず，データセットの 6 番目と 7 番目の学生 (i.e., $p = 6$ および $p = 7$) について考えてみましょう。どちらの学生も教科書を読んでいないので，**予習** `= 0` となっています。このことを数学的に表記すると，$X_{2,6} = 0$ および $X_{2,7} = 0$ となります。しかしながら，7 番の学生は講義に出席している (i.e., `attend = 1`, $X_{1,7} = 1$) のに対し，6 番の学生は出席していません (i.e., `attend = 0`, $X_{1,6} = 0$)。それでは，これらの数値を回帰直線の一般式に当てはめるとどうなるか見てみましょう。6 番の学生については，回帰式は以下のように予測します。

$$
\begin{aligned}
\hat{Y}_6 &= b_0 + b_1 X_{1,6} + b_2 X_{2,6} \\
&= b_0 + (b_1 \times 0) + (b_2 \times 0) \\
&= b_0
\end{aligned}
$$

したがって，この学生は切片項 b_0 の値に対応する成績を取ることが予測されます。7 番の学生につ

いてはどうでしょうか？ 回帰式に数値を代入すると，今度は以下のようになります。

$$\begin{aligned}\hat{Y}_7 &= b_0 + b_1 X_{1,7} + b_2 X_{2,7} \\ &= b_0 + (b_1 \times 1) + (b_2 \times 0) \\ &= b_0 + b_1\end{aligned}$$

この学生は授業に出席しているので，予測される成績は切片項 b_0 に**出席**変数 b_1 の係数を *加えた*ものに等しくなります。つまり，b_1 が 1 より大きければ，講義に出る学生は出ない学生より成績が良いと予測されるわけです。逆に，係数が負であれば，講義に出た学生の方が成績が悪くなることが予測されます。このことについて，もう少し掘り下げてみましょう。講義にも出席し $(X_{1,1} = 1)$，*さらに*教科書も読んできた $(X_{2,1} = 1)$ 1 番の学生の場合はどうなるでしょうか？ これらの数値を回帰式に代入すると，次のようになります。

$$\begin{aligned}\hat{Y}_1 &= b_0 + b_1 X_{1,1} + b_2 X_{2,1} \\ &= b_0 + (b_1 \times 1) + (b_2 \times 1) \\ &= b_0 + b_1 + b_2\end{aligned}$$

すなわち，授業に出席すると良い成績が取れ (i.e., $b_1 > 0$)，教科書を読んでおくことでも良い成績が取れるため (i.e., $b_2 > 0$)，1 番の学生は 6 番や 7 番の学生よりも高い成績を取ることが予測されます。

ここまで来れば，教科書は読んでいるが講義を受けなかった 3 番の学生の成績が $b_2 + b_0$ となるという回帰モデルの予測についても全く驚かないでしょう。わざわざ回帰式を書いて読者を退屈させるようなことはしません。その代わりに，以下のような *予測された成績*の表をお示しします：

		read textbook?	
		no	yes
attended?	no	b_0	$b_0 + b_2$
	yes	$b_0 + b_1$	$b_0 + b_1 + b_2$

このように，切片項 b_0 は，講義に出席したり教科書を読んだりする時間を取らない学生から期待される，“ベースライン” としての成績のような役割を果たします。同様に，b_1 は講義に出席すれば得られると期待される成績の向上を，b_2 は教科書を読むことで得られる成績の向上を表しています。これが分散分析であれば，b_1 を出席の主効果，b_2 を予習の主効果として扱いたいと思うかもしれません！ 実際に，単純な 2×2 の分散分析では，*まさに*これと同様なことが起こります。

さて，分線分析と回帰分析が，基本的に同じものであることが分かってきたところで，実際に JASP で `rtfm` データを用いた回帰分析を行い，本当にそうなのか納得してみましょう。通常の方法で回帰分析を実行すると，Figure **??**のような結果が得られます。

いくつか興味深い点があります。まず，切片項は 43.5 であり，教科書を読まず，講義にも出席しなかった 2 名の学生の “グループ” 平均である 42.5 という値に近いことに注目してください。次に，

出席変数の回帰係数が $b_1 = 18.0$ であることに注目してください。これは，講義に出席した学生は，出席しなかった学生よりも成績が 18 点高いということを示唆しています。つまり，教科書を読まずに講義に参加した学生の成績は $b_0 + b_1$ であり，$43.5 + 18.0 = 61.5$ となることが予測されます。教科書を読んでいる学生についても，同様のことが生じていることが確認できます。

実は，分散分析と回帰分析の等価性を確かめるために，もう少し踏み込むことができます。回帰分析の出力において，**出席**変数と**予習**変数の p 値を見てみましょう。これらは，先ほど分散分析を実行したときに見たものと同じ値です。回帰モデルの検定では統計量 t が，分散分析では統計量 F が計算されるため，このことは少し意外に思われるかもしれません。しかし，思い返せば，Chapter **??**において t 分布と F 分布の間に関係があることは既に述べられています。自由度 k の t 分布に従って分布する量があるとき，それを二乗した量は自由度 1 と k の F 分布に従うことになります。このことは，回帰モデルにおける統計量 t について確認することができます。**出席**変数の t 値は 4.65 です。これを二乗すると 21.6 となり，分散分析における対応する F 統計量と一致します。

1.6.3 対比を用いた非二値の要因の変換

ここまで，2×2 の分散分析を線形モデルとして見る方法について述べてきました。このことは，$2 \times 2 \times 2$ の分散分析や，$2 \times 2 \times 2 \times 2$ の分散分析に対しても簡単に一般化することができます。ここまでの話と全く同じです。各要因に対応した新たな二値変数を追加するだけです。ただし，2 水準以上の要因について考えると，少々ややこしくなります。例えば，本章の前半で行った，`clinicaltrial.csv` データを用いた 3×2 の分散分析について考えてみましょう。3 つの水準をもつ**薬**要因を，回帰モデルに適した数値に変換するにはどうすればよいでしょうか？

この問に対する答えは，実はとてもシンプルです。3 水準の要因が，*2* つの二値変数によって記述しなおせることに気付きさえすれば良いのです。例えば，**薬：アンザイフリー**という新たな二値変数と作るとしましょう。**薬**変数が**"アンザイフリー"**と等しいときに，**薬：アンザイフリー** `= 1` となります。そうでない場合には**薬：アンザイフリー** `= 0` となります。この場合，この変数はアンザイフリーと他の 2 つの薬との間に**対比**を設定します。もちろん，**薬：アンザイフリー**の対比だけでは**薬**変数に含まれるすべての情報を捉えるのに十分ではありません。ジョイゼパムとプラセボを区別できるような追加の対比が必要です。これを実現するために，薬がジョイゼパムであれば 1，そうでなければ 0 となるような，**薬：ジョイゼパム**という 2 つ目の二値変数を作成することができます。これら 2 つの対比により，3 種類の薬を完全に識別することが可能になります。下記の表はその様子を表したものです：

薬	薬：アンザイフリー	薬：ジョイゼパム
"プラセボ"	0	0
"アンザイフリー"	1	0
"ジョイゼパム"	0	1

もし患者に投与された薬がプラセボであれば，2 つの対比変数はいずれも 0 になります。薬がアンザイフリーの場合，**薬：アンザイフリー**変数は 1 に，**薬：ジョイゼパム**は 0 になります。ジョイゼパムの場合はその逆です。**薬：ジョイゼパム**が 1 に，**薬：アンザイフリー**は 0 になります。

対比変数を作成することは，JASP の ‘列の作成’ 機能を使えば，さほど難しいことではありません。例えば，数式ボックスに次のような簡単な R コードを入力すれば，二値の**薬：アンザイフリー**変数を作成することができます：`ifelse(drug == 'anxifree', 1, 0)` 同様に，新しく**薬：ジョイゼパム**変数を作成するには，次のコードを使用します：ifelse(drug == 'joyzepam', 1, 0) **セラピー：CBT** 変数を作成するコードは次の通りです：ifelse(therapy == 'CBT', 1, 0) これらの新しい変数は，JASP のデータファイル `clinicaltrial2.jasp` で見ることができます。こでまで通り，“f_x” 記号をクリックすると，R コードも表示されます。

ここまで，3 水準の要因を 2 つの二値変数で再コード化する方法を見てきました。更に，分散分析と回帰分析が，二値変数に関して同じように動作することも確認済みです。しかし，今回の場合，さらに複雑な問題が発生します。その問題については次章で説明します。

1.6.4 非二値の要因における分散分析と回帰分析の等価性

ここまでの処理で，同じデータセットに 2 種類のバージョンができました。**薬**変数を 1 つの 3 水準の要因として表現している `clinicaltrial.csv` ファイルのオリジナルデータ，およびこれを 2 つの二値の対比に変換した `clinicaltrial2.omv` ファイルの拡張データです。改めて強調しますが，ここで説明したいのは，最初に示した 3×2 の分散分析が，対比変数を用いた回帰モデルと等価であるということです。まずは，分散分析を再度実行してみると，Figure **??**のような結果が得られます。交互作用は除外しておきましょう。‘Model’ に移動し，`drug*therapy` の項を ‘Model Terms’ ボックスから取り除きます。

当然，この出力は驚くべきものではありません。先ほどの分散分析と全く同じです。次に，予測変数として `CBTtherapy` を，説明変数として `druganxifree`, `drugjoyzepam` を投入した回帰分析を実行してみましょう。結果は Figure **??**のようになります。

ふーむ。前回とはかなり違った印象になっていると思います。当然のことですが，回帰分析の出力は，それぞれ別個に回帰分析を行った場合と同様に，3 つの予測変数のそれぞれについて別々の結果が出力されています。`CBTtherapy` 変数の p 値は，元々の分散分析における**セラピー**要因の値と全く同じなので，回帰モデルと分散分析が同様のことを行っているのだと安心することができます。一方で，この回帰モデルは `druganxifree` および `drugjoyzepam` の対比を，あたかも全く関係のない 2 つの変数のように検定しています。もちろん，これは仕方のないことです。この可哀そうな回帰分析は，`drugjoyzepam` と `druganxifree` が，実際には我々が 3 水準の**薬**要因をコード化するために用いた 2 つの異なる対比であることを知る由もないのですから。それが知る限りでは，`drugjoyzepam` お

よび druganxifree は，drugjoyzepam および therapyCBT と同じくらい相互に無関係なのです。しかし，我々はこのことをよく知っています。この段階では，これら 2 つの対比がそれぞれ有意であるかどうかを判断することには全く意味がありません。我々が知りたいのは，薬物による “総合的な” 効果があるかどうかです。すなわち，*我々が* JASP に望むのは，2 つの “薬物に関連した” 対比を検定の目的に沿って一まとめにした，ある種の “モデル比較” のための検定です。見覚えはないでしょうか？ 今回の場合，Figure **??**のように，予測変数に CBTtherapy を含み，薬物に関連した 2 つの変数を除外した “ヌルモデル” を指定すれば良いのです。これを実行するためには，‘Model’ ボタンをクリックし，2 つの薬物に関連する項について ‘Add to null model’ をチェックします。そして，両モデルについて AIC を求めることで，予測変数として**薬**をモデルに含めるべきかどうかを評価できます (Section **??**を参照) Remember, smaller is better! 値が小さい方がよい，ということを思い出してください。

1.7 対比の指定を行う別の方法

前節では，要因をいくつかの対比に変換する方法を紹介しました。この方法では，2 値変数の組み合わせによって，以下の表のような定義を行いました

drug	druganxifree	drugjoyzepam
"プラセボ"	0	0
"アンザイフリー"	1	0
"ジョイゼパム"	0	1

表の各行が要因の各水準に対応し，各列がそれぞれの対比に対応しています。この表は，常に列より行が 1 つ多く，特別な名前を持っています。その名前とは，**対比行列**です。しかし，対比行列の指定には様々な方法があります。本節では，統計学者が用いるいくつかの標準的な対比行列について，そしてそれらを JASP でどのように使用するかについて説明します。この後の非釣り合い型の分散分析の章 (Section **??**) を読むのであれば，本章によく目を通しておくことをお勧めします。そうでなければ，読み飛ばしても構いません。釣り合い型デザインでは，対比の選択はあまり重要ではないからです。

1.7.1 処理対比

先に述べたような特殊な対比では，要因のある 1 つの水準が，ある種の “ベースライン” カテゴリー (i.e., 今回の例では **プラセボ**) として特別な意味を持ち，これに対応して他の 2 つの水準が定義されます。このような対比のことを**処理対比**といいます。“ダミーコーディング” とも呼ばれていま

す。この対比では，要因の各水準はベースとなる水準と比較されます。ベースとなる水準は切片の値です。

この名称は，要因の中のあるカテゴリーが実際に特別な，ベースラインを表すものである場合に，対比が極めて自然かつ感覚的であるという事実を反映しています。臨床試験の例では，このことは理にかなっています。**プラセボ**条件は被験者に投薬をしていない状況に相当する，特別な条件です。他の 2 つの条件は，プラセボ条件との関係によって定義されています。すなわち，プラセボをアンザイフリーに置き換える場合と，ジョイゼパムに置き換える場合です。

上記の表は，3 つの水準を持つ要因に対する処理対比行列です。もし 5 つの水準を持つ要因の処理対比行列が必要な場合には，以下のように設定することができます：

```
%Level   2 3 4 5
水準   2 3 4 5
1     0 0 0 0
2     1 0 0 0
3     0 1 0 0
4     0 0 1 0
5     0 0 0 1
```

この例では，1 つ目の対比が水準 2 と水準 1 の対比，2 つ目の対比が水準 3 と水準 1 の対比，という具合になっています。デフォルトでは，要因の *最初の*水準が常にベースラインカテゴリーとなることに注意してください (i.e., すなわち，その水準は，それ自身と対応づけられた明示的な対比を持たず，値はすべての水準について 0 となります)。JASP では，‘Data Variable’ ウィンドウに表示される変数の水準の順序を操作することで，どのカテゴリが要因の最初の水準になるかを変更することができます（スプレッドシートの変数名をダブルクリックして，‘Data Variable’ ビューを表示する必要があります)。

1.7.2 ヘルマート対比

処理対比は，様々なシーンで活躍します。しかし，この方法が最も有効なのは，本当の意味でベースとなるカテゴリが存在し，そのカテゴリとの関係によって他のすべての群を評価したい場合です。そのようなベースラインカテゴリが存在しないような状況では，各群を他の群の平均と比較する方が，より理にかなっているかもしれません。ここで，JASP の ‘ANOVA’ - ‘Contrasts’ 選択ボックスの ‘helmert’ オプションで実行される**ヘルマート対比**を紹介しましょう。ヘルマート対比は，各群を “その前の” 群の平均と比較するというアイデアが根底にあります。すなわち，第 1 の対比は 2 つ目の群と 1 つ目の群の差，第 2 の対比は 3 つ目の群と 1 つ目および 2 つ目の群の平均値との差，といった具合です。5 つの水準を持つ要因の場合，以下のような対比行列に変換されます：

```
1   -1   -1   -1   -1
2    1   -1   -1   -1
3    0    2   -1   -1
4    0    0    3   -1
5    0    0    0    4
```

ヘルマート対比の利点として，すべての対比の和が 0 になる (i.e., すべての列の和が 0 になる) という性質があります。この性質は，ヘルマート対比を利用すれば，分散分析を回帰として解釈する際に，切片項が大平均 $\mu_{..}$ に対応する，ということを意味します。処理対比においては，切片項はベースラインカテゴリの群平均に対応していました。この特性は，状況によっては非常に有効です。これまで仮定してきたような，釣り合い型デザインにおいてはそれほど重要ではありませんが，後に Section **??**で非釣り合い型デザインについて考える時，その有用性が明らかになります。実を言えば，わざわざ本章を設けているのは，非釣り合い型の分散分析を理解する際に，対比が重要になるからなのです。

1.7.3 Sum to zero contrasts

1.7.4 零和対比

3 つ目のオプションとして，JASP では “単純” 対比と呼ばれ，グループ間の一対比較を構成する “零和” 対比について簡単に紹介します。具体的には，各対比はあるグループとベースラインカテゴリ（今回の場合は第 1 グループに相当）との差を符号化したものになります：

```
1   -1   -1   -1   -1
2    1    0    0    0
3    0    1    0    0
4    0    0    1    0
5    0    0    0    1
```

ヘルマート対比と同様に，各列の和が 0 になることが分かります。これは，分散分析を回帰モデルとして扱ったとき，切片項が大平均に対応することを意味しています。これらの対比を解釈する際に認識すべきことは，これらの対比がそれぞれ，グループ 1 と他の 4 つのグループのうちの 1 つとの一対比較であるということです。具体的には，対比 1 は “グループ 2 − グループ 1” の対比，対比 2 は “グループ 3 − グループ 1” の対比，といった具合です。[*6]

[*6] 処理対比と単純対比の違いは何かと聞かれることがあります。基本的な例として，男性=0，女性=1 とする性別の主効果を考えます。処理対比の場合，係数は女性と男性の平均の差の指標となり，切片は男性の平均値を表します。一方で，単純対比，すなわち男性=-1，女性=1 の場合，切片は平均値の平均であり，主効果は切片と各グループ平均の差となります。

1.7.5 JASP におけるその他の対比

JASP には，他にも様々な種類の対比を生成できるオプションが用意されています。これらは，分散分析のメインウィンドウの ‘対比’ オプションにあり，以下のようなタイプの対比がリストアップされています。

対比のタイプ	
偏差	各水準（基準カテゴリを除く）の平均と全水準の平均（大平均）を比較します。
シンプル	単純対比では，処理対比と同様に各水準の平均を指定された水準の平均と比較します。このような対比は，対照群がある場合に有効です。デフォルトでは，最初のカテゴリが指定されています。ただし，単純対比では，切片は要因におけるすべての水準の大平均となります。
差	各水準（最初の水準を除く）の平均を，それ以前の水準の平均と比較します。(逆ヘルマート対比と呼ばれることもあります)
Helmert	要因の各水準（最後の水準を除く）の平均を，それ以降の水準の平均と比較します。
繰り返し	各水準（最後の水準を除く）の平均を，その後の水準の平均と比較します。
多項	線形効果と二次効果を比較します。1 つ目の自由度は全カテゴリを通しての線形効果を含んでおり，2 つ目の自由度は二次効果です。この対比は，多項式のトレンドを推定するためにしばしば用いられます。

1.8 Post hoc tests

Time to switch to a different topic. Rather than pre-planned comparisons that you have tested using contrasts, let's suppose you've done your ANOVA and it turns out that you obtained some significant effects. Because of the fact that the F-tests are "omnibus" tests that only really test the null hypothesis that there are no differences among groups, obtaining a significant effect doesn't tell you which groups are different to which other ones. We discussed this issue back in Chapter **??**, and

in that chapter our solution was to run t-tests for all possible pairs of groups, making corrections for multiple comparisons (e.g., Bonferroni, Holm) to control the Type I error rate across all comparisons. The methods that we used back in Chapter **??** have the advantage of being relatively simple and being the kind of tools that you can use in a lot of different situations where you're testing multiple hypotheses, but they're not necessarily the best choices if you're interested in doing efficient post hoc testing in an ANOVA context. There are actually quite a lot of different methods for performing multiple comparisons in the statistics literature (**Hsu1996**), and it would be beyond the scope of an introductory text like this one to discuss all of them in any detail.

That being said, there's one tool that I do want to draw your attention to, namely Tukey's "Honestly Significant Difference", or **Tukey's HSD** for short. For once, I'll spare you the formulas and just stick to the qualitative ideas. The basic idea in Tukey's HSD is to examine all relevant pairwise comparisons between groups, and it's only really appropriate to use Tukey's HSD if it is *pairwise* differences that you're interested in.[*7] For instance, earlier we conducted a factorial ANOVA using the `clinicaltrial.csv` data set, and where we specified a main effect for drug and a main effect of therapy we would be interested in the following four comparisons:

- The difference in mood gain for people given Anxifree versus people given the placebo.
- The difference in mood gain for people given Joyzepam versus people given the placebo.
- The difference in mood gain for people given Anxifree versus people given Joyzepam.
- The difference in mood gain for people treated with CBT and people given no therapy.

For any one of these comparisons, we're interested in the true difference between (population) group means. Tukey's HSD constructs **simultaneous confidence intervals** for all four of these comparisons. What we mean by 95% "simultaneous" confidence interval is that, if we were to repeat this study many times, then in 95% of the study results the confidence intervals would contain the relevant true value. Moreover, we can use these confidence intervals to calculate an adjusted p value for any specific comparison.

The `TukeyHSD` function in JASP is pretty easy to use. You simply specify the ANOVA model term that you want to run the post hoc tests for. For example, if we were looking to run post hoc tests for the main effects but not the interaction, we would open up the 'Post Hoc Tests' option in the ANOVA analysis screen, move the `drug` and `therapy` variables across to the box on the right, and then select the 'Tukey' checkbox in the list of possible post hoc corrections that could be applied.

[*7] If, for instance, you actually find yourself interested to know if Group A is significantly different from the mean of Group B and Group C, then you need to use a different tool (e.g., Scheffe's method, which is more conservative, and beyond the scope of this book). However, in most cases you probably are interested in pairwise group differences, so Tukey's HSD is a pretty useful thing to know about.

This, along with the corresponding results table, is shown in Figure **??**

The output shown in the 'Post Hoc Tests' results table is (I hope) pretty straightforward. The first comparison, for example, is the Anxifree versus placebo difference, and the first part of the output indicates that the observed difference in group means is .27. The next number is the standard error for the difference, from which we could calculate the 95% confidence interval (you can actually display this 95% CI if you select the 'Confidence intervals' box). Then there is a column with the degrees of freedom, a column with the t-value, and finally a column with the p-value. For the first comparison the adjusted p-value is .21. In contrast, if you look at the next line, we see that the observed difference between joyzepam and the placebo is 1.03, and this result is significant ($p < .001$).

So far, so good. What about the situation where your model includes interaction terms? For instance, the default option in JASP is to allow for the possibility that there is an interaction between drug and therapy. If that's the case, the number of pairwise comparisons that we need to consider starts to increase. As before, we need to consider the three comparisons that are relevant to the main effect of `drug` and the one comparison that is relevant to the main effect of `therapy`. But, if we want to consider the possibility of a significant interaction (and try to find the group differences that underpin that significant interaction), we need to include comparisons such as the following:

- The difference in mood gain for people given Anxifree and treated with CBT, versus people given the placebo and treated with CBT
- The difference in mood gain for people given Anxifree and given no therapy, versus people given the placebo and given no therapy.
- etc

There are quite a lot of these comparisons that you need to consider. So, when we run the Tukey post hoc analysis for this ANOVA model, we see that it has made a *lot* of pairwise comparisons (19 in total), as shown in Figure **??**. You can see that it looks pretty similar to before, but with a lot more comparisons made.

1.9

The method of planned comparisons

Following on from the previous sections on contrasts and post hoc tests in ANOVA, I think the method of planned comparisons is important enough to deserve a quick discussion. In our discussions of multiple comparisons, in the previous section and back in Chapter **??**, I've been assuming that the tests you want to run are genuinely post hoc. For instance, in our drugs example above, maybe you thought that the drugs would all have different effects on mood (i.e., you hypothesised a main effect of drug), but you didn't have any specific hypothesis about how they would be different, nor did you have any real idea about *which* pairwise comparisons would be worth looking at. If that is the case, then you really have to resort to something like Tukey's HSD to do your pairwise comparisons.

The situation is rather different, however, if you genuinely did have real, specific hypotheses about which comparisons are of interest, and you *never ever* have any intention to look at any other comparisons besides the ones that you specified ahead of time. When this is true, and if you honestly and rigorously stick to your noble intentions to not run any other comparisons (even when the data look like they're showing you deliciously significant effects for stuff you didn't have a hypothesis test for), then it doesn't really make a lot of sense to run something like Tukey's HSD, because it makes corrections for a whole bunch of comparisons that you never cared about and never had any intention of looking at. Under those circumstances, you can safely run a (limited) number of hypothesis tests without making an adjustment for multiple testing. This situation is known as the **method of planned comparisons**, and it is sometimes used in clinical trials. However, further consideration is out of scope for this introductory book, but at least you know that this method exists!

1.10

Factorial ANOVA 3: unbalanced designs

Factorial ANOVA is a very handy thing to know about. It's been one of the standard tools used to analyse experimental data for many decades, and you'll find that you can't read more than two or three papers in psychology without running into an ANOVA in there somewhere. However, there's one huge difference between the ANOVAs that you'll see in a lot of real scientific articles and the ANOVAs that I've described so far. In in real life we're rarely lucky enough to have perfectly balanced designs. For one reason or another, it's typical to end up with more observations in some cells than in others. Or, to put it another way, we have an **unbalanced design**.

Unbalanced designs need to be treated with a lot more care than balanced designs, and the

statistical theory that underpins them is a lot messier. It might be a consequence of this messiness, or it might be a shortage of time, but my experience has been that undergraduate research methods classes in psychology have a nasty tendency to ignore this issue completely. A lot of stats textbooks tend to gloss over it too. The net result of this, I think, is that a lot of active researchers in the field don't actually know that there's several different "types" of unbalanced ANOVAs, and they produce quite different answers. In fact, reading the psychological literature, I'm kind of amazed at the fact that most people who report the results of an unbalanced factorial ANOVA don't actually give you enough details to reproduce the analysis. I secretly suspect that most people don't even realise that their statistical software package is making a whole lot of substantive data analysis decisions on their behalf. It's actually a little terrifying when you think about it. So, if you want to avoid handing control of your data analysis to stupid software, read on.

1.10.1 The coffee data

As usual, it will help us to work with some data. The `coffee.csv` file contains a hypothetical data set that produces an unbalanced 3×2 ANOVA. Suppose we were interested in finding out whether or not the tendency of people to `babble` when they have too much coffee is purely an effect of the coffee itself, or whether there's some effect of the `milk` and `sugar` that people add to the coffee. Suppose we took 18 people and gave them some coffee to drink. The amount of coffee / caffeine was held constant, and we varied whether or not milk was added, so `milk` is a binary factor with two levels, `"yes"` and `"no"`. We also varied the kind of sugar involved. The coffee might contain `"real"` sugar or it might contain `"fake"` sugar (i.e., artificial sweetener) or it might contain `"none"` at all, so the `sugar` variable is a three level factor. Our outcome variable is a continuous variable that presumably refers to some psychologically sensible measure of the extent to which someone is "babbling". The details don't really matter for our purpose. Take a look at the data in the JASP spreadsheet view, as in Figure **??**.

Looking at the table of means in Figure **??** we get a strong impression that there are differences between the groups. This is especially true when we look at the standard deviations. Across groups, this standard deviation varies quite a lot.[*8] Whilst this at first may seem like a straightforward factorial ANOVA, a problem arises when we look at how many observations we have in each group. See the different Ns for different groups shown in Figure **??**. This violates one of our original

[*8] This discrepancy in standard deviations might (and should) make you wonder if we have a violation of the homogeneity of variance assumption. I'll leave it as an exercise for the reader to double check this using the Levene test option.

assumptions, namely that the number of people in each group is the same. We haven't really discussed how to handle this situation.

1.10.2 "Standard ANOVA" does not exist for unbalanced designs

Unbalanced designs lead us to the somewhat unsettling discovery that there isn't really any one thing that we might refer to as a standard ANOVA. In fact, it turns out that there are *three* fundamentally different ways[*9] in which you might want to run an ANOVA in an unbalanced design. If you have a balanced design all three versions produce identical results, with the sums of squares, F-values, etc., all conforming to the formulas that I gave at the start of the chapter. However, when your design is unbalanced they don't give the same answers. Furthermore, they are not all equally appropriate to every situation. Some methods will be more appropriate to your situation than others. Given all this, it's important to understand what the different types of ANOVA are and how they differ from one another.

The first kind of ANOVA is conventionally referred to as **Type I sum of squares**. I'm sure you can guess what the other two are called. The "sum of squares" part of the name was introduced by the SAS statistical software package and has become standard nomenclature, but it's a bit misleading in some ways. I think the logic for referring to them as different types of sum of squares is that, when you look at the ANOVA tables that they produce, the key difference in the numbers is the SS values. The degrees of freedom don't change, the MS values are still defined as SS divided by df, etc. However, what the terminology gets wrong is that it hides the reason *why* the SS values are different from one another. To that end, it's a lot more helpful to think of the three different kinds of ANOVA as three different *hypothesis testing strategies*. These different strategies lead to different SS values, to be sure, but it's the strategy that is the important thing here, not the SS values themselves. Recall from Section **??** that any particular F-test is best thought of as a comparison between two linear models. So, when you're looking at an ANOVA table, it helps to remember that each of those F-tests corresponds to a *pair* of models that are being compared. Of course, this leads naturally to the question of *which* pair of models is being compared. This is the fundamental difference between ANOVA Types I, II and III: each one corresponds to a different way

[*9] Actually, this is a bit of a lie. ANOVAs can vary in other ways besides the ones I've discussed in this book. For instance, I've completely ignored the difference between fixed-effect models in which the levels of a factor are "fixed" by the experimenter or the world, and random-effect models in which the levels are random samples from a larger population of possible levels (this book only covers fixed-effect models). Don't make the mistake of thinking that this book, or any other one, will tell you "everything you need to know" about statistics, any more than a single book could possibly tell you everything you need to know about psychology, physics or philosophy. Life is too complicated for that to *ever* be true. This isn't a cause for despair, though. Most researchers get by with a basic working knowledge of ANOVA that doesn't go any further than this book does. I just want you to keep in mind that this book is only the beginning of a very long story, not the whole story.

of choosing the model pairs for the tests.

1.10.3 Type I sum of squares

The Type I method is sometimes referred to as the "sequential" sum of squares, because it involves a process of adding terms to the model one at a time. Consider the coffee data, for instance. Suppose we want to run the full 3×2 factorial ANOVA, including interaction terms. The full model contains the outcome variable `babble`, the predictor variables `sugar` and `milk`, and the interaction term `sugar*milk`. This can be written as `babble ~ sugar + milk + sugar*milk`. The Type I strategy builds this model up sequentially, starting from the simplest possible model and gradually adding terms.

The simplest possible model for the data would be one in which neither milk nor sugar is assumed to have any effect on babbling. The only term that would be included in such a model is the intercept, written as `babble ~ 1`. This is our initial null hypothesis. The next simplest model for the data would be one in which only one of the two main effects is included. In the coffee data, there are two different possible choices here, because we could choose to add milk first or to add sugar first. The order actually turns out to matter, as we'll see later, but for now let's just make a choice arbitrarily and pick sugar. So, the second model in our sequence of models is `babble ~ sugar`, and it forms the alternative hypothesis for our first test. We now have our first hypothesis test:

Null model:	`babble ~ 1`
Alternative model:	`babble ~ sugar`

This comparison forms our hypothesis test of the main effect of `sugar`. The next step in our model building exercise is to add the other main effect term, so the next model in our sequence is `babble ~ sugar + milk`. The second hypothesis test is then formed by comparing the following pair of models:

Null model:	`babble ~ sugar`
Alternative model:	`babble ~ sugar + milk`

This comparison forms our hypothesis test of the main effect of `milk`. In one sense, this approach is very elegant: the alternative hypothesis from the first test forms the null hypothesis for the second one. It is in this sense that the Type I method is strictly sequential. Every test builds directly on the results of the last one. However, in another sense it's very inelegant, because there's a strong asymmetry between the two tests. The test of the main effect of `sugar` (the first test) completely ignores `milk`, whereas the test of the main effect of `milk` (the second test) does take `sugar` into account. In any case, the fourth model in our sequence is now the full model,

`babble ~ sugar + milk + sugar*milk`, and the corresponding hypothesis test is:

Null model:	`babble ~ sugar + milk`
Alternative model:	`babble ~ sugar + milk + sugar*milk`

Type III sum of squares is the default hypothesis testing method used by JASP ANOVA, so to run a Type I sum of squares analysis we have to select 'Type 1' in the 'Sum of squares' selection box in the JASP 'ANOVA' - 'Model' options. This gives us the ANOVA table shown in Figure **??**.

The big problem with using Type I sum of squares is the fact that it really does depend on the order in which you enter the variables. Yet, in many situations the researcher has no reason to prefer one ordering over another. This is presumably the case for our milk and sugar problem. Should we add milk first or sugar first? It feels exactly as arbitrary as a data analysis question as it does as a coffee-making question. There may in fact be some people with firm opinions about ordering, but it's hard to imagine a principled answer to the question. Yet, look what happens when we change the ordering, as in Figure **??**.

The *p*-values for both main effect terms have changed, and fairly dramatically. Among other things, the effect of `milk` has become significant (though one should avoid drawing any strong conclusions about this, as I've mentioned previously). Which of these two ANOVAs should one report? It's not immediately obvious.

When you look at the hypothesis tests that are used to define the "first" main effect and the "second" one, it's clear that they're qualitatively different from one another. In our initial example, we saw that the test for the main effect of `sugar` completely ignores `milk`, whereas the test of the main effect of `milk` does take `sugar` into account. As such, the Type I testing strategy really does treat the first main effect as if it had a kind of theoretical primacy over the second one. In my experience there is very rarely if ever any theoretically primacy of this kind that would justify treating any two main effects asymmetrically.

The consequence of all this is that Type I tests are very rarely of much interest, and so we should move on to discuss Type II tests and Type III tests.

1.10.4 Type III sum of squares

Having just finished talking about Type I tests, you might think that the natural thing to do next would be to talk about Type II tests. However, I think it's actually a bit more natural to discuss Type III tests (which are simple and the default in JASP) before talking about Type II tests (which are trickier). The basic idea behind Type III tests is extremely simple. Regardless of which term you're trying to evaluate, run the *F*-test in which the alternative hypothesis corresponds to the full ANOVA model as specified by the user, and the null model just deletes that one term that you're testing. For

instance, in the coffee example, in which our full model was `babble ~ sugar + milk + sugar*milk`, the test for a main effect of `sugar` would correspond to a comparison between the following two models:

Null model:	`babble ~ milk + sugar*milk`
Alternative model:	`babble ~ sugar + milk + sugar*milk`

Similarly the main effect of `milk` is evaluated by testing the full model against a null model that removes the `milk` term, like so:

Null model:	`babble ~ sugar + sugar*milk`
Alternative model:	`babble ~ sugar + milk + sugar*milk`

Finally, the interaction term `sugar*milk` is evaluated in exactly the same way. Once again, we test the full model against a null model that removes the `sugar*milk` interaction term, like so:

Null model:	`babble ~ sugar + milk`
Alternative model:	`babble ~ sugar + milk + sugar*milk`

The basic idea generalises to higher order ANOVAs. For instance, suppose that we were trying to run an ANOVA with three factors, `A`, `B` and `C`, and we wanted to consider all possible main effects and all possible interactions, including the three way interaction `A*B*C`. The table below shows you what the Type III tests look like for this situation:

Term being tested is	Null model is `outcome ~ ...`	Alternative model is `outcome ~ ...`
`A`	`B + C + A*B + A*C + B*C + A*B*C`	`A + B + C + A*B + A*C + B*C + A*B*C`
`B`	`A + C + A*B + A*C + B*C + A*B*C`	`A + B + C + A*B + A*C + B*C + A*B*C`
`C`	`A + B + A*B + A*C + B*C + A*B*C`	`A + B + C + A*B + A*C + B*C + A*B*C`
`A*B`	`A + B + C + A*C + B*C + A*B*C`	`A + B + C + A*B + A*C + B*C + A*B*C`
`A*C`	`A + B + C + A*B + B*C + A*B*C`	`A + B + C + A*B + A*C + B*C + A*B*C`
`B*C`	`A + B + C + A*B + A*C + A*B*C`	`A + B + C + A*B + A*C + B*C + A*B*C`
`A*B*C`	`A + B + C + A*B + A*C + B*C`	`A + B + C + A*B + A*C + B*C + A*B*C`

As ugly as that table looks, it's pretty simple. In all cases, the alternative hypothesis corresponds to the full model which contains three main effect terms (e.g. `A`), three two-way interactions (e.g. `A*B`) and one three-way interaction (i.e., `A*B*C`). The null model always contains 6 of these 7 terms, and the missing one is the one whose significance we're trying to test.

At first pass, Type III tests seem like a nice idea. Firstly, we've removed the asymmetry that caused us to have problems when running Type I tests. And because we're now treating all terms the same way, the results of the hypothesis tests do not depend on the order in which we specify

them. This is definitely a good thing. However, there is a big problem when interpreting the results of the tests, especially for main effect terms. Consider the coffee data. Suppose it turns out that the main effect of `milk` is not significant according to the Type III tests. What this is telling us is that `babble ~ sugar + sugar*milk` is a better model for the data than the full model. But what does that even *mean*? If the interaction term `sugar*milk` was also non-significant, we'd be tempted to conclude that the data are telling us that the only thing that matters is `sugar`. But suppose we have a significant interaction term, but a non-significant main effect of `milk`. In this case, are we to assume that there really is an "effect of sugar", an "interaction between milk and sugar", but no "effect of milk"? That seems crazy. The right answer simply *must* be that it's meaningless[*10] to talk about the main effect if the interaction is significant. In general, this seems to be what most statisticians advise us to do, and I think that's the right advice. But if it really is meaningless to talk about non-significant main effects in the presence of a significant interaction, then it's not at all obvious why Type III tests should allow the null hypothesis to rely on a model that includes the interaction but omits one of the main effects that make it up. When characterised in this fashion, the null hypotheses really don't make much sense at all.

Later on, we'll see that Type III tests can be redeemed in some contexts, but first let's take a look at the ANOVA results table using Type III sum of squares, see Figure **??**.

But be aware, one of the perverse features of the Type III testing strategy is that typically the results turn out to depend on the *contrasts* that you use to encode your factors (see Section **??** if you've forgotten what the different types of contrasts are).[*11]

Okay, so if the *p*-values that typically come out of Type III analyses are so sensitive to the choice of contrasts, does that mean that Type III tests are essentially arbitrary and not to be trusted? To some extent that's true, and when we turn to a discussion of Type II tests we'll see that Type II analyses avoid this arbitrariness entirely, but I think that's too strong a conclusion. Firstly, it's important to recognise that some choices of contrasts will always produce the same answers (ah, so this is what is happening in JASP). Of particular importance is the fact that if the columns of our contrast matrix are all constrained to sum to zero, then the Type III analysis will always give the same answers.

1.10.5 Type II sum of squares

Okay, so we've seen Type I and III tests now, and both are pretty straightforward. Type I tests are

[*10] Or, at the very least, rarely of interest.

[*11] However, in JASP the results for Type III sum of squares ANOVA are the same regardless of the contrast selected, so JASP is obviously doing something different!

performed by gradually adding terms one at a time, whereas Type III tests are performed by taking the full model and looking to see what happens when you remove each term. However, both can have some limitations. Type I tests are dependent on the order in which you enter the terms, and Type III tests are dependent on how you code up your contrasts. Type II tests are a little harder to describe, but they avoid both of these problems, and as a result they are a little easier to interpret.

Type II tests are broadly similar to Type III tests. Start with a "full" model, and test a particular term by deleting it from that model. However, Type II tests are based on the **marginality principle** which states that you should not omit a lower order term from your model if there are any higher order ones that depend on it. So, for instance, if your model contains the two-way interaction `A*B` (a 2nd order term), then it really ought to contain the main effects `A` and `B` (1st order terms). Similarly, if it contains a three-way interaction term `A*B*C`, then the model must also include the main effects `A`, `B` and `C` as well as the simpler interactions `A*B`, `A*C` and `B*C`. Type III tests routinely violate the marginality principle. For instance, consider the test of the main effect of `A` in the context of a three-way ANOVA that includes all possible interaction terms. According to Type III tests, our null and alternative models are:

Null model:	`outcome ~ B + C + A*B + A*C + B*C + A*B*C`
Alternative model:	`outcome ~ A + B + C + A*B + A*C + B*C + A*B*C`

Notice that the null hypothesis omits `A`, but includes `A*B`, `A*C` and `A*B*C` as part of the model. This, according to the Type II tests, is not a good choice of null hypothesis. What we should do instead, if we want to test the null hypothesis that `A` is not relevant to our `outcome`, is to specify the null hypothesis that is the most complicated model that does not rely on `A` in any form, even as an interaction. The alternative hypothesis corresponds to this null model plus a main effect term of `A`. This is a lot closer to what most people would intuitively think of as a "main effect of `A`", and it yields the following as our Type II test of the main effect of `A`:[*12]

Null model:	`outcome ~ B + C + B*C`
Alternative model:	`outcome ~ A + B + C + B*C`

Anyway, just to give you a sense of how the Type II tests play out, here's the full table of tests that would be applied in a three-way factorial ANOVA:

[*12]Note, of course, that this does depend on the model that the user specified. If the original ANOVA model doesn't contain an interaction term for `B*C`, then obviously it won't appear in either the null or the alternative. But that's true for Types I, II and III. They never include any terms that you *didn't* include, but they make different choices about how to construct tests for the ones that you did include.

Term being tested is	Null model is `outcome ~ ...`	Alternative model is `outcome ~ ...`
`A`	`B + C + B*C`	`A + B + C + B*C`
`B`	`A + C + A*C`	`A + B + C + A*C`
`C`	`A + B + A*B`	`A + B + C + A*B`
`A*B`	`A + B + C + A*C + B*C`	`A + B + C + A*B + A*C + B*C`
`A*C`	`A + B + C + A*B + B*C`	`A + B + C + A*B + A*C + B*C`
`B*C`	`A + B + C + A*B + A*C`	`A + B + C + A*B + A*C + B*C`
`A*B*C`	`A + B + C + A*B + A*C + B*C`	`A + B + C + A*B + A*C + B*C + A*B*C`

In the context of the two way ANOVA that we've been using in the coffee data, the hypothesis tests are even simpler. The main effect of `sugar` corresponds to an F-test comparing these two models:

Null model: `babble ~ milk`
Alternative model: `babble ~ sugar + milk`

The test for the main effect of `milk` is

Null model: `babble ~ sugar`
Alternative model: `babble ~ sugar + milk`

Finally, the test for the interaction `sugar*milk` is:

Null model: `babble ~ sugar + milk`
Alternative model: `babble ~ sugar + milk + sugar*milk`

Running the tests are again straightforward. Just select 'Type 2' in the 'Sum of squares' selection box in the JASP 'ANOVA' - 'Model' options, This gives us the ANOVA table shown in Figure **??**.

Type II tests have some clear advantages over Type I and Type III tests. They don't depend on the order in which you specify factors (unlike Type I), and they don't depend on the contrasts that you use to specify your factors (unlike Type III). And although opinions may differ on this last point, and it will definitely depend on what you're trying to do with your data, I do think that the hypothesis tests that they specify are more likely to correspond to something that you actually care about. As a consequence, I find that it's usually easier to interpret the results of a Type II test than the results of a Type I or Type III test. For this reason my tentative advice is that, if you can't think of any obvious model comparisons that directly map onto your research questions but you still want to run an ANOVA in an unbalanced design, Type II tests are probably a better choice than Type I or Type

III.[*13]

1.10.6 Effect sizes (and non-additive sums of squares)

JASP also provides the effect sizes η^2 and partial η^2 when you select these options. However, when you've got an unbalanced design there's a bit of extra complexity involved.

If you remember back to our very early discussions of ANOVA, one of the key ideas behind the sums of squares calculations is that if we add up all the SS terms associated with the effects in the model, and add that to the residual SS, they're supposed to add up to the total sum of squares. And, on top of that, the whole idea behind η^2 is that, because you're dividing one of the SS terms by the total SS value, an η^2 value can be interpreted as the proportion of variance accounted for by a particular term. But this is not so straightforward in unbalanced designs because some of the variance goes "missing".

This seems a bit odd at first, but here's why. When you have unbalanced designs your factors become correlated with one another, and it becomes difficult to tell the difference between the effect of Factor A and the effect of Factor B. In the extreme case, suppose that we'd run a 2×2 design in which the number of participants in each group had been as follows:

	sugar	no sugar
milk	100	0
no milk	0	100

Here we have a spectacularly unbalanced design: 100 people have milk and sugar, 100 people have no milk and no sugar, and that's all. There are 0 people with milk and no sugar, and 0 people with sugar but no milk. Now suppose that, when we collected the data, it turned out there is a large (and statistically significant) difference between the "milk and sugar" group and the "no-milk and no-sugar" group. Is this a main effect of sugar? A main effect of milk? Or an interaction? It's impossible to tell, because the presence of sugar has a perfect association with the presence of milk.

[*13] I find it amusing to note that the default in R is Type I and the default in SPSS, JASP, and jamovi is Type III. Neither of these appeals to me all that much. Relatedly, I find it depressing that almost nobody in the psychological literature ever bothers to report which Type of tests they ran, much less the order of variables (for Type I) or the contrasts used (for Type III). Often they don't report what software they used either. The only way I can ever make any sense of what people typically report is to try to guess from auxiliary cues which software they were using, and to assume that they never changed the default settings. Please don't do this! Now that you know about these issues make sure you indicate what software you used, and if you're reporting ANOVA results for unbalanced data, then specify what Type of tests you ran, specify order information if you've done Type I tests and specify contrasts if you've done Type III tests. Or, even better, do hypotheses tests that correspond to things you really care about and then report those!

Now suppose the design had been a little more balanced:

	sugar	no sugar
milk	100	5
no milk	5	100

This time around, it's technically possible to distinguish between the effect of milk and the effect of sugar, because we have a few people that have one but not the other. However, it will still be pretty difficult to do so, because the association between sugar and milk is still extremely strong, and there are so few observations in two of the groups. Again, we're very likely to be in the situation where we *know* that the predictor variables (milk and sugar) are related to the outcome (babbling), but we don't know if the *nature* of that relationship is a main effect of one or the other predictor, or the interaction.

This uncertainty is the reason for the missing variance. The "missing" variance corresponds to variation in the outcome variable that is clearly attributable to the predictors, but we don't know which of the effects in the model is responsible. When you calculate Type I sum of squares, no variance ever goes missing. The sequential nature of Type I sum of squares means that the ANOVA automatically attributes this variance to whichever effects are entered first. However, the Type II and Type III tests are more conservative. Variance that cannot be clearly attributed to a specific effect doesn't get attributed to any of them, and it goes missing.

1.11 Summary

- Factorial ANOVA with balanced designs, without interactions (Section **??**) and with interactions included (Section **??**)
- Effect size, estimated means, and confidence intervals in a factorial ANOVA (Section **??**)
- Checking assumptions in ANOVA (Section **??**)
- Analysis of Covariance (ANCOVA) (Section **??**)
- Understanding the linear model underlying ANOVA, including different contrasts (Section **??** and **??**)
- Post hoc testing using Tukey's HSD (Section **??**) and a brief commentary on planned comparisons (Section **??**)
- Factorial ANOVA with unbalanced designs (Section **??**)

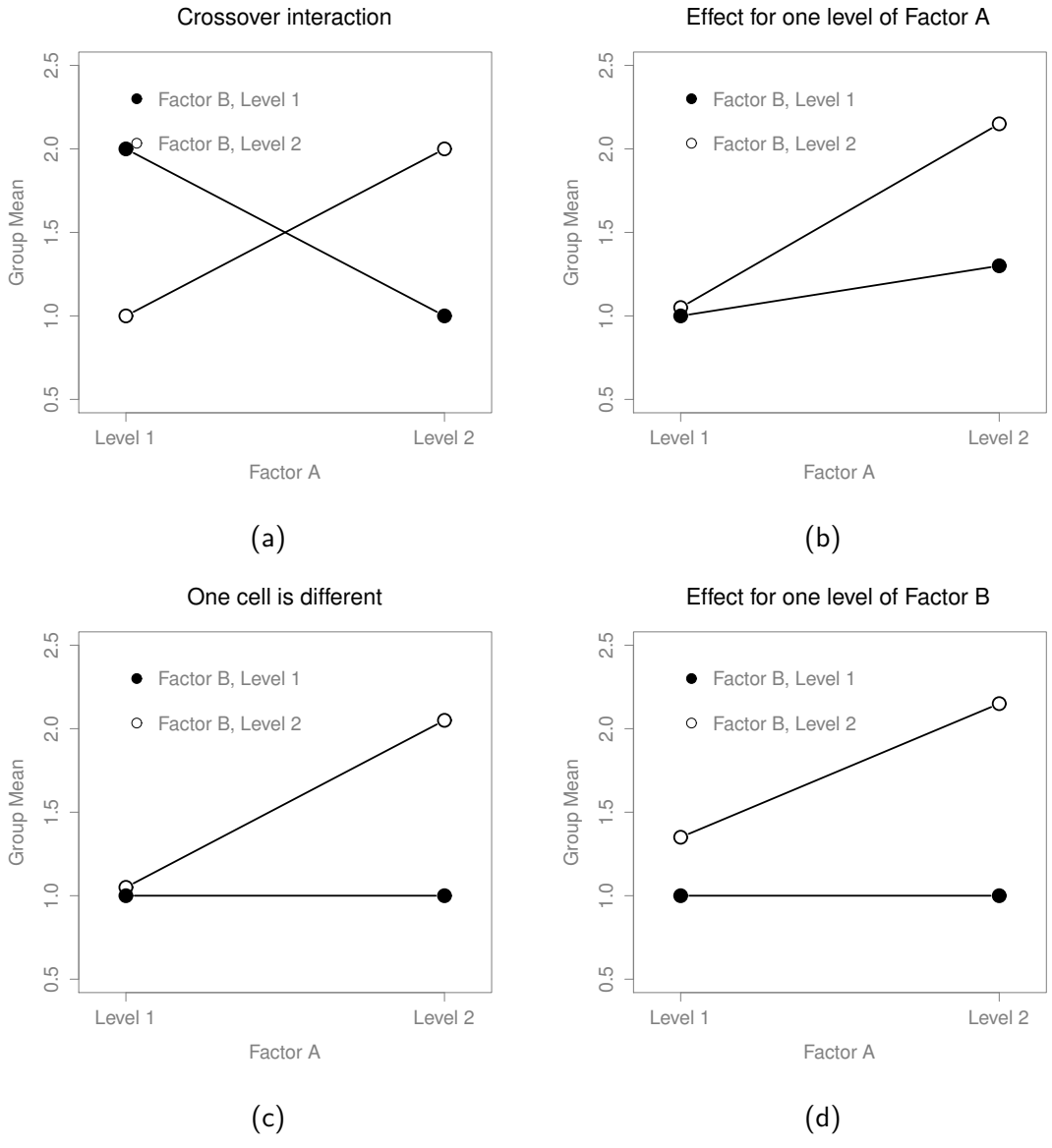

Figure1.6　2 × 2 要因の分散分析における様々な交互作用

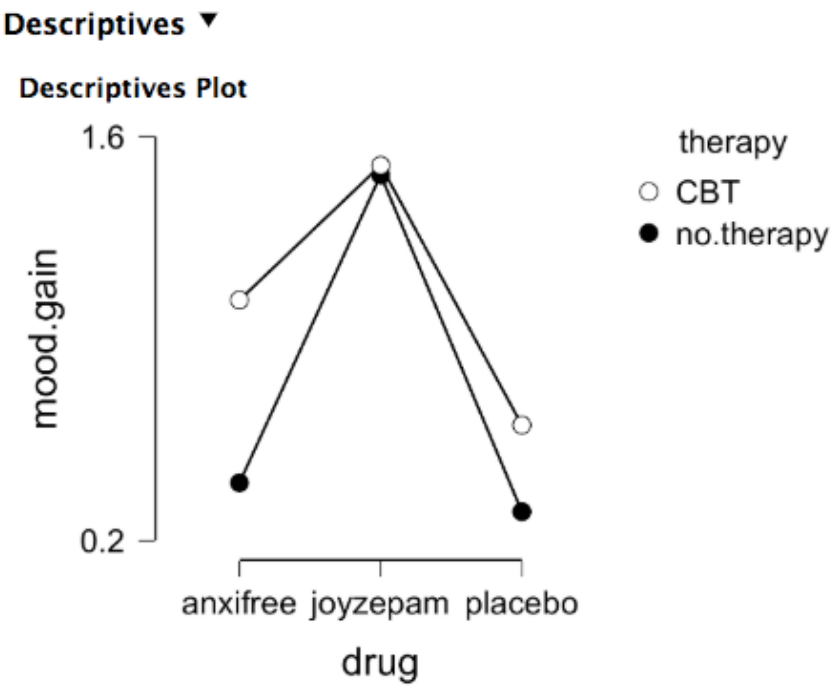

Figure1.7　臨床試験データに対して分散分析の「Descriptives Plot」オプションを使用した際の JASP の出力

ANOVA – mood.gain

Cases	Sum of Squares	df	Mean Square	F	p
drug	3.453	2.000	1.727	31.714	< .001
therapy	0.467	1.000	0.467	8.582	0.013
drug ∗ therapy	0.271	2.000	0.136	2.490	0.125
Residual	0.653	12.000	0.054		

Note. Type III Sum of Squares

Figure1.8　交互作用項 `drug*therapy` を含む，完全な多元配置モデルの出力

Marginal Means

Marginal Means – drug ✻ therapy

				95% CI	
drug	therapy	Marginal Mean	SE	Lower	Upper
anxifree	CBT	1.033	0.135	0.740	1.327
	no.therapy	0.400	0.135	0.106	0.694
joyzepam	CBT	1.500	0.135	1.206	1.794
	no.therapy	1.467	0.135	1.173	1.760
placebo	CBT	0.600	0.135	0.306	0.894
	no.therapy	0.300	0.135	0.006	0.594

Figure1.9　飽和モデルの周辺平均を示す JASP のスクリーンショット, i.e. `clinicaltrial` データセットの交互作用コンポーネントを含む

Assumption Checks ▼

Test for Equality of Variances (Levene's)

F	df1	df2	p
0.206	5.000	12.000	0.954

Q-Q Plot ▼

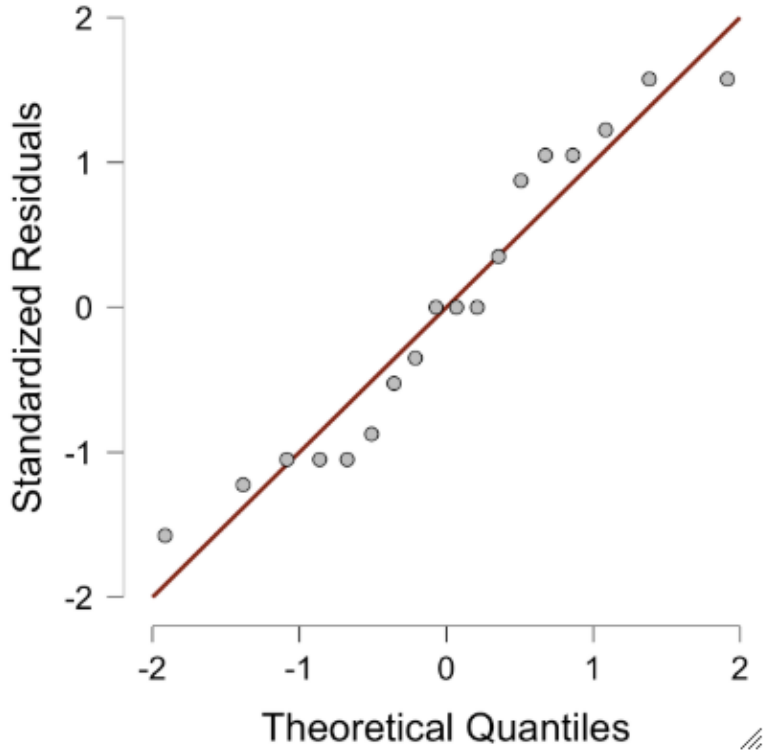

Figure1.10　分散分析モデルの仮定の確認

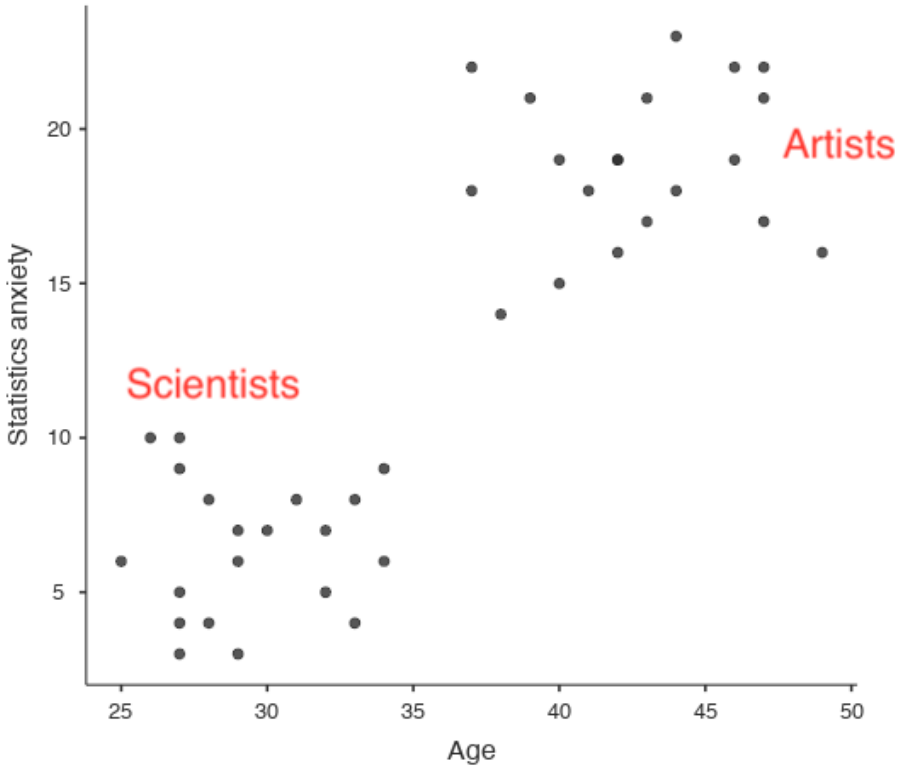

Figure1.11　2 つの群の年齢ごとの統計に対する不安のプロット

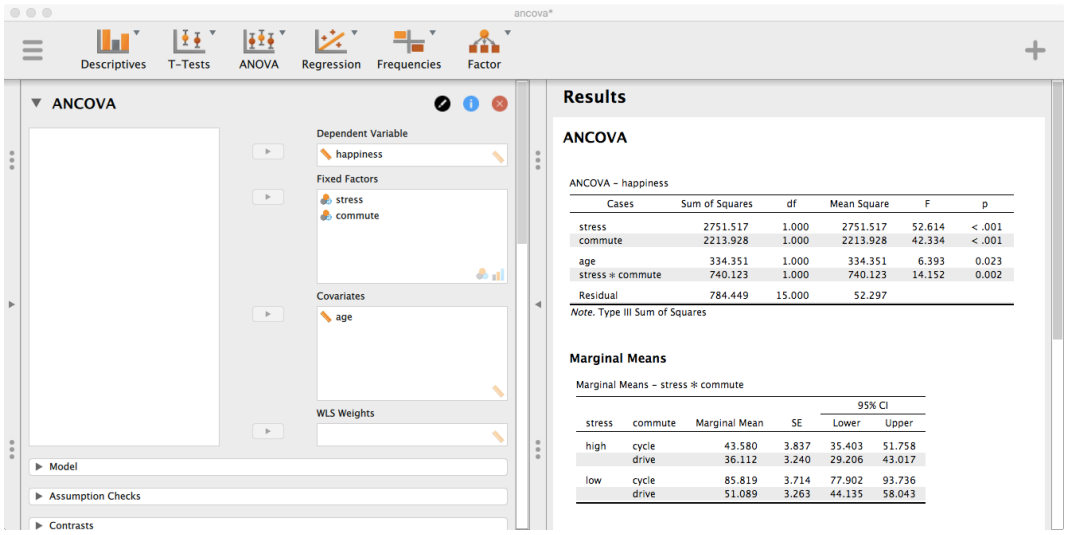

Figure1.12　JASP による共分散分析の実行画面

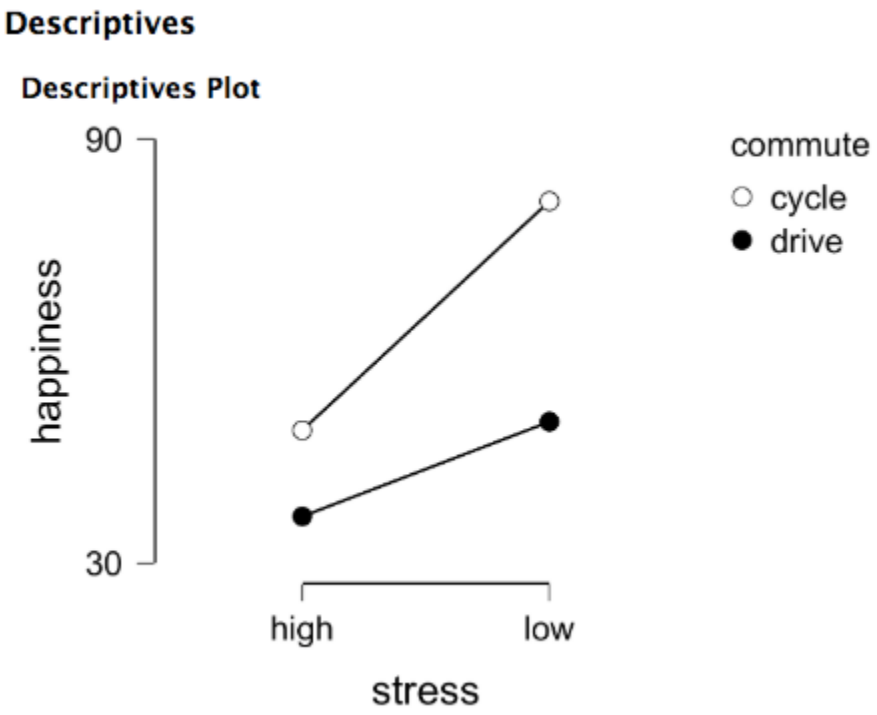

Figure1.13　ストレスと通勤方法の関数としての平均幸福度のプロット

ANOVA – grade

Cases	Sum of Squares	df	Mean Square	F	p
attend	648.000	1.000	648.000	21.600	0.006
reading	1568.000	1.000	1568.000	52.267	< .001
Residual	150.000	5.000	30.000		

Note. Type III Sum of Squares

Figure1.14　JASP の `rtfm.csv` データセットを用いた交互作用項のない分散分析

Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p
1	(Intercept)	43.500	3.354		12.969	< .001
	attend	18.000	3.873	0.523	4.648	0.006
	reading	28.000	3.873	0.814	7.230	< .001

Figure1.15　JSAP の `rtfm.csv` データセットを用いた交互作用項を含まない回帰分析

ANOVA – mood.gain

Cases	Sum of Squares	df	Mean Square	F	p
drug	3.453	2.000	1.727	26.149	< .001
therapy	0.467	1.000	0.467	7.076	0.019
Residual	0.924	14.000	0.066		

Note. Type III Sum of Squares

Figure1.16　JASP による交互作用を含まない分散分析の出力

Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p
1	(Intercept)	0.289	0.121		2.385	0.032
	druganxifree	0.267	0.148	0.242	1.797	0.094
	drugjoyzepam	1.033	0.148	0.939	6.965	< .001
	therapyCBT	0.322	0.121	0.311	2.660	0.019

Figure1.17　JASP による対比変数 `druganxifree` および `drugjoyzepam` を用いた回帰分析, with contrast variables `druganxifree` and `drugjoyzepam`

Linear Regression ▼

Model Summary ▼

Model	R	R^2	Adjusted R^2	RMSE
0	0.844	0.713	0.674	0.305
1	0.900	0.809	0.768	0.257

Note. Null model includes druganxifree, drugjoyzepam

ANOVA

Model		Sum of Squares	df	Mean Square	F	p
0	Regression	3.453	2	1.727	18.611	< .001
	Residual	1.392	15	0.093		
	Total	4.845	17			
1	Regression	3.921	3	1.307	19.791	< .001
	Residual	0.924	14	0.066		
	Total	4.845	17			

Note. Null model includes druganxifree, drugjoyzepam

Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p
0	(Intercept)	0.450	0.124		3.619	0.003
	druganxifree	0.267	0.176	0.242	1.516	0.150
	drugjoyzepam	1.033	0.176	0.939	5.876	< .001
1	(Intercept)	0.289	0.121		2.385	0.032
	druganxifree	0.267	0.148	0.242	1.797	0.094
	drugjoyzepam	1.033	0.148	0.939	6.965	< .001
	therapyCBT	0.322	0.121	0.311	2.660	0.019

Figure1.18　JASP の回帰分析におけるモデル比較。0 はヌルモデル，1 は対立モデル

Post Hoc Tests ▼

Post Hoc Comparisons - drug ▼

		Mean Difference	SE	t	p_{tukey}
placebo	joyzepam	−1.033	0.148	−6.965	< .001
	anxifree	−0.267	0.148	−1.797	0.206
joyzepam	anxifree	0.767	0.148	5.168	< .001

Post Hoc Comparisons - therapy

		Mean Difference	SE	t	p_{tukey}
CBT	no.therapy	0.322	0.121	2.660	0.019

Figure1.19 Tukey HSD post hoc test in JASP

Post Hoc Tests ▼

Post Hoc Comparisons - drug ✻ therapy ▼

		Mean Difference	SE	t	p_{tukey}
placebo,CBT	joyzepam,CBT	−0.900	0.191	−4.724	0.005
	anxifree,CBT	−0.433	0.191	−2.275	0.275
	placebo,no.therapy	0.300	0.191	1.575	0.628
	joyzepam,no.therapy	−0.867	0.191	−4.549	0.007
	anxifree,no.therapy	0.200	0.191	1.050	0.892
joyzepam,CBT	anxifree,CBT	0.467	0.191	2.449	0.214
	placebo,no.therapy	1.200	0.191	6.299	< .001
	joyzepam,no.therapy	0.033	0.191	0.175	1.000
	anxifree,no.therapy	1.100	0.191	5.774	< .001
anxifree,CBT	placebo,no.therapy	0.733	0.191	3.849	0.022
	joyzepam,no.therapy	−0.433	0.191	−2.275	0.275
	anxifree,no.therapy	0.633	0.191	3.324	0.053
placebo,no.therapy	joyzepam,no.therapy	−1.167	0.191	−6.124	< .001
	anxifree,no.therapy	−0.100	0.191	−0.525	0.994
joyzepam,no.therapy	anxifree,no.therapy	1.067	0.191	5.599	0.001

Figure1.20 Tukey HSD post hoc test in JASP factorial ANOVA with an interaction term

Results

Descriptive Statistics

Descriptive Statistics

	babble	
	no	yes
Valid	10	8
Missing	0	0
Mean	5.320	4.750
Std. Deviation	0.796	0.962
Minimum	3.900	3.500
Maximum	6.600	5.900

Descriptive Statistics

Descriptive Statistics

	babble		
	fake	none	real
Valid	6	5	7
Missing	0	0	0
Mean	5.033	4.440	5.543
Std. Deviation	0.814	1.038	0.637
Minimum	3.900	3.500	4.600
Maximum	5.900	5.800	6.600

Figure1.21 Descriptives for the `coffee.csv` data set, separately split by `milk` and `sugar`, respectively.

ANOVA

ANOVA – babble

Cases	Sum of Squares	df	Mean Square	F	p
sugar	3.558	2.000	1.779	6.749	0.011
milk	0.956	1.000	0.956	3.628	0.081
sugar ∗ milk	5.944	2.000	2.972	11.277	0.002
Residual	3.162	12.000	0.264		

Note. Type I Sum of Squares

Figure1.22 ANOVA results table using Type I sum of squares in JASP

ANOVA ▼

ANOVA – babble

Cases	Sum of Squares	df	Mean Square	F	p
milk	1.444	1.000	1.444	5.479	0.037
sugar	3.070	2.000	1.535	5.824	0.017
milk ∗ sugar	5.944	2.000	2.972	11.277	0.002
Residual	3.163	12.000	0.264		

Note. Type I Sum of Squares

Figure1.23 ANOVA results table using Type I sum of squares in JASP, but with factors entered in a different order (`milk` first)

ANOVA

ANOVA – babble

Cases	Sum of Squares	df	Mean Square	F	p
milk	1.004	1.000	1.004	3.810	0.075
sugar	2.132	2.000	1.066	4.045	0.045
milk ✻ sugar	5.944	2.000	2.972	11.277	0.002
Residual	3.163	12.000	0.264		

Note. Type III Sum of Squares

Figure1.24 ANOVA results table using Type III sum of squares in JASP

ANOVA

ANOVA – babble

Cases	Sum of Squares	df	Mean Square	F	p
milk	0.956	1.000	0.956	3.628	0.081
sugar	3.070	2.000	1.535	5.824	0.017
milk ✻ sugar	5.944	2.000	2.972	11.277	0.002
Residual	3.163	12.000	0.264		

Note. Type II Sum of Squares

Figure1.25 ANOVA results table using Type II sum of squares in JASP