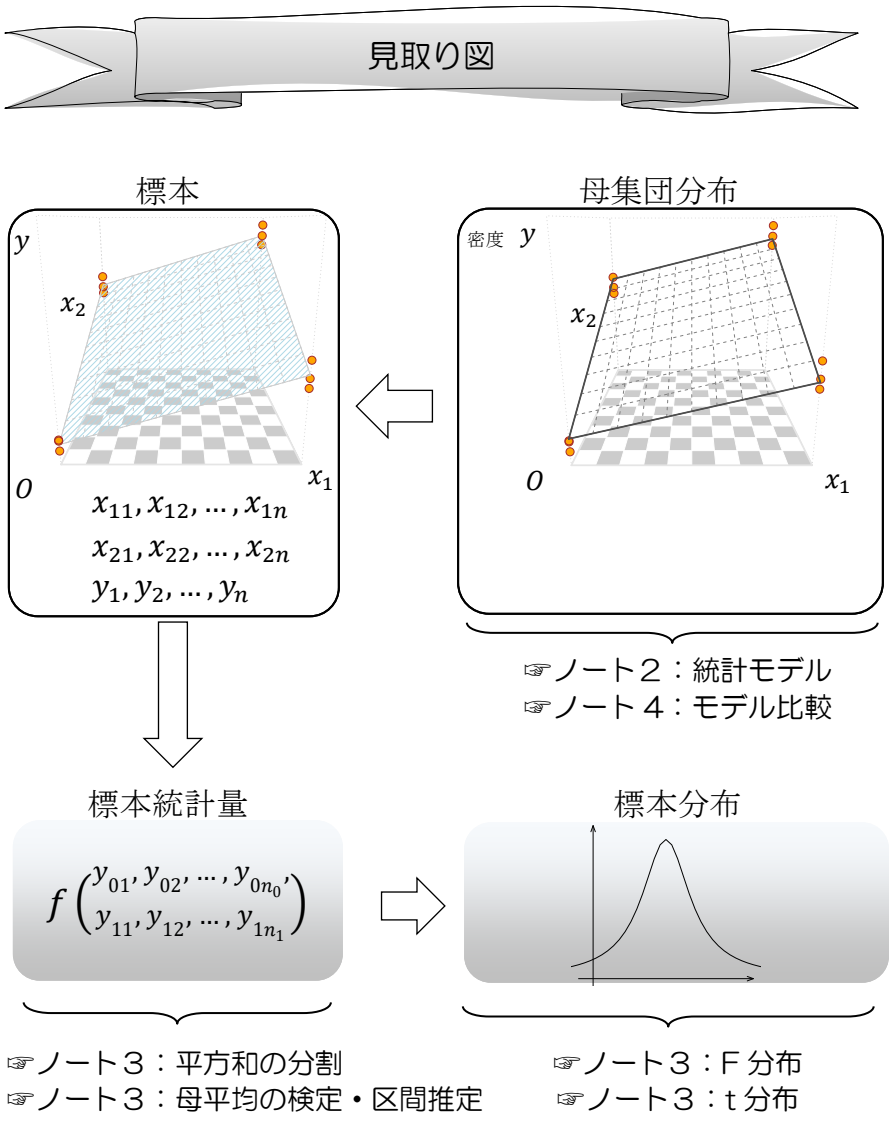
第6講　分散分析（対応なし）

学習の目標

- □ 回帰モデルにおいて独立変数の数が複数含まれるものを**重回帰モデル**ということが分かる。
- □ 回帰モデルを**数式や図によって表現**することができる。
- □ 研究者の興味や仮説に応じて母集団において柔軟なモデルを立てる意義ややり方についての基礎が理解できる。
- □ ある独立変数が別の変数を経由して従属変数に**間接効果**を持つとき仲介をなす変数を**媒介変数**ということが分かる。
- □ 複数の要因が連動して変化するためどちらに影響したか判断できない状況を**交絡**と呼ぶことが分かる。
- □ **調整変数**の存在によって独立変数単体の**主効果**では説明できない**交互作用効果**が生じることがあることが分かる。
- □ 独立性の仮定を乱すクラスターを持つデータ構造に対して**階層モデル**を想定する必要性が理解できる。
- □ 重回帰モデルの各独立変数にかかる係数を**偏回帰係数**と呼び、その解釈の仕方を理解することができる。
- □ 独立変数にかかる係数の大きさを比較する手段として**標準偏回帰係数**を用いることの意義が分かる。
- □ 推定した回帰モデルのデータへの適合度を測る基準として**重相関係数**と**決定係数**を用いることができる。
- □ **重回帰式の検定**および各**偏回帰係数の検定**、そして、**信頼区間**や**予測区間**についても理解をし、使用できる。
- □ **過学習**、**解釈可能性**、**多重共線性**などの問題が懸念される際に、独立変数を選択しモデルの改良を行うことができる。
- □ **罰則を設けた最小二乗推定**について理解することができ、**ラッソ回帰**を用いた**モデル縮約**を行うことができる。
- □ **モデル比較**の必要性を理解し、**情報量基準**や**クロスバリデーション**を用いてモデルを選択することができる。
- □ 多重共線性に解決の一つの対策として、**主成分分析**を援用し独立変数間の相関をゼロにしたモデルを提案できる。



データの形式

ID	独立変数１	独立変数２	…	予測変数ｐ	応答変数
1	1	0	…	2	2.1
2	0	1	…	1	3.2
⋮	⋮	⋮	⋮	⋮	⋮
n	1	0	…	0	1.5

（1）目的（リサーチクエスチョン）

三つ以上の群に差があるのかを見出したいとき、つまり複数の名義尺度からなる独立変数から従属変数を値を予測するモデルを作りたい際に実施する手法。

（2）考え方

前項で扱った重回帰分析は、複数の独立変数で従属変数の値を予測するモデルを作成する方法だったが、これらの独立変数が全て名義尺度だった際に、行う統計分析を分散分析と呼ぶ。

このため、分散分析は重回帰分析の特殊な場合としてみなすことができる。ただし、この分散分析は「実験 experiment」という研究手法と密接に結びついて発展してきたため、実験デザインを反映した特殊な情報を活用する。より正確に言えば、研究者は、理想的な分散分析が行えるよう効率的な実験デザインを組み立てて研究を行う。

例えば、分散分析は、独立変数の数で一元配置分散分析（独立変数が 1 個）、二元配置分散分析（独立変数が二個）などに分類される。一元配置とは異なり、二元配置では各独立変数の組み合わせ（独立変数 1 が 0、独立変数が 1、など）が同じサンプルサイズとなっているかどうかが大事になる。実験の組み立てにより、分散分析の調整がどのように必要となるかを中心に学びを進めてほしい。

（3）具体的なデータの例

ID	Item	R1	R2	R3	R4	独立変数 1	独立変数 2	実験協力者	従属変数
1	1	Since yesterday	I have been **walking**	with	my friends.	0	0	山田	
2	1	Yesterday	I have been **walking**	with	my friends.	1	0	山田	
3	1	Since yesterday	I **walked**	with	my friends.	0	1	山田	
4	1	Yesterday	I **walked**	with	my friends.	1	1	山田	
5	2	Since yesterday	I have been **cooking**	with	my friends.	0	0	山田	
6	2	Yesterday	I have been **cooking**	with	my friends.	1	0	山田	
7	2	Since yesterday	I **cooked**	with	my friends.	0	1	山田	
8	2	Yesterday	I **cooked**	with	my friends.	1	1	山田	
⋮									
93	24	Since yesterday	I have been **swimming**	with	my friends.	0	0	山田	
94	24	Yesterday	I have been **swimming**	with	my friends.	1	0	山田	
95	24	Since yesterday	I **swam**	with	my friends.	0	1	山田	
96	24	Yesterday	I **swam**	with	my friends.	1	1	山田	
97	Filler 1		I am exicted.					山田	
98	Filler 2		I am surprised.					山田	
⋮									
288	Filler 196		I am satisfied.					山田	

📖 ノート０ あらすじ：研究の流れ

○ 基礎：実験の流れ

（１）実験の実施

フェーズ１：実験の計画・実施

手順１：先行研究を踏まえ、問うべき**リサーチクエスチョン**を立てる。
手順２：リサーチクエスチョンに答えを出すのに必要な**証拠（データ）**がどのようなものかを考える。
手順３：独立変数と従属変数を決め、それらを測定する**実験**を作り上げる。

フェーズ２：実験を実施する

手順１：大学における**研究倫理委員会**の承認を得る。
手順２：**実験協力者（被験者）**を集め、データを集める。

（２）予備解析

フェーズ３：予備解析

基本的な統計量の算出やグラフ化を行い、標本データへの予備的考察・分散分析の仮定等をチェック。

（３）推測統計学の枠組みに基づく分析

フェーズ４：モデル選択

手順１：母集団の構造を表したモデルを複数作る
手順２：モデル比較を行い、ベストなモデルを選択

フェーズ５：分散分析

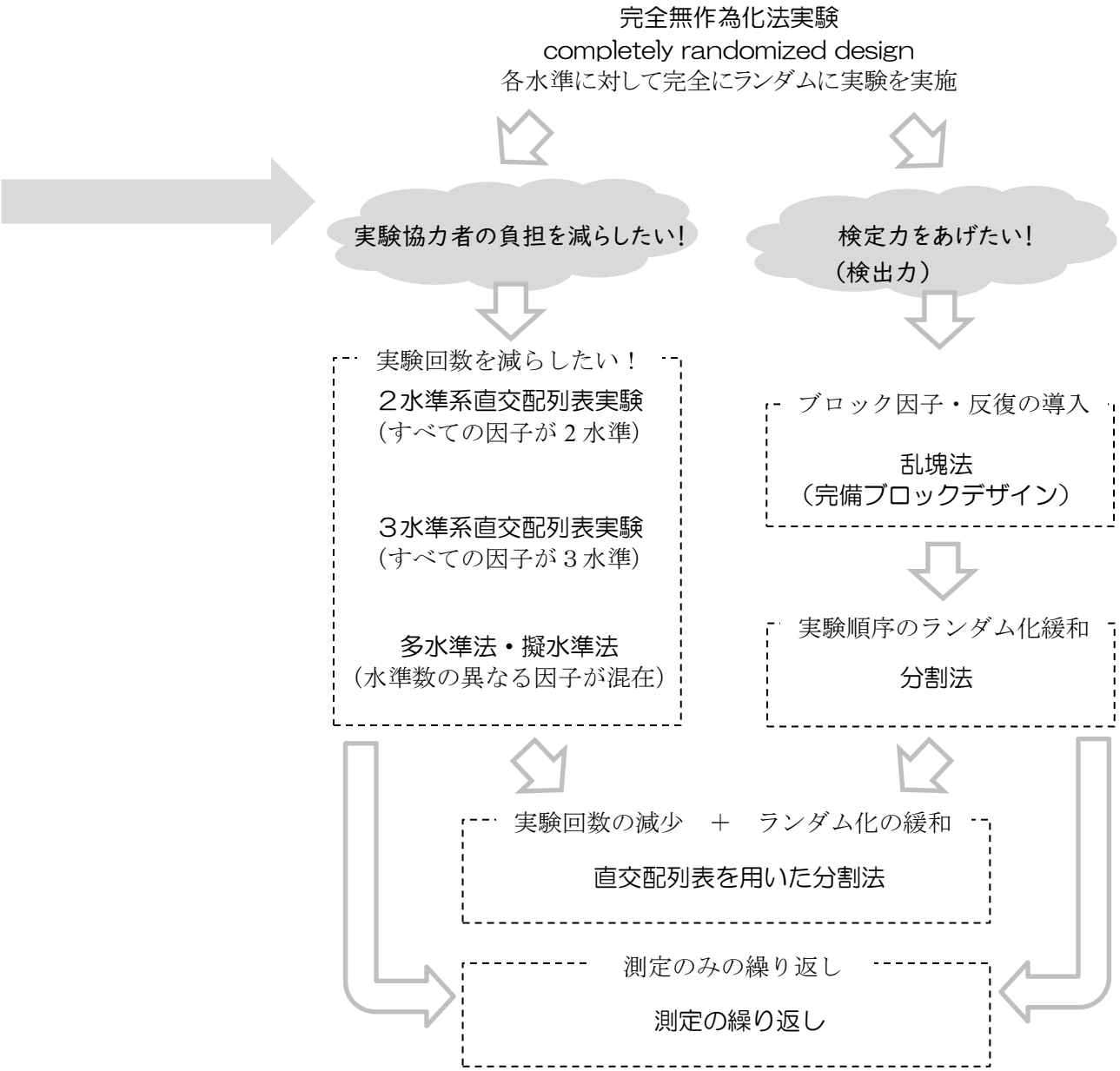
手順１：平方和を分解する
（例１）$ss_y = ss_A + ss_e$　一元配置
（例２）$ss_y = ss_A + ss_B + ss_{A\times B} + ss_e$　二元配置
手順２：残差平方和ss_eに比して独立変数の平方和（ss_A, ss_B, $ss_{A\times B}$）が大きいと言えるかF検定する

フェーズ６：パラメータへの統計的推測・評価

手順１：検討対象の要因における各水準の母平均を推測
手順２：検討対象の要因における水準間の差を推測

※点推定・信頼区間・予測区間などの検討

○ 発展：実験の改良



☞ 参考：永田靖 (2000)『入門実験計画法』東京：日科技連
長畑秀和 (2016)『R で学ぶ実験計画法』東京：朝倉書店
三輪哲久 (2015) 『実験計画法と分散分析』東京：朝倉書店

📖 ノート１ 実験デザイン：実験研究の基本事項

そもそも、実験って何ですか？
なんで必要なんですか？

実験とは、研究者が独立変数を統制して従属変数の挙動を調べる研究です。多くの場合、従属変数に影響を与える可能性のある要因（独立変数の候補）は無数に考えられます。しかし実際に研究で扱うのはそのごく一部です。**注目している要因以外の変数が従属変数に予期せぬ影響を与えてはいないことを保証する**ために、独立変数を統制しておくのです。

（１）要因 Factor

これは、質的な（＝離散値を取る）独立変数のこと。t 検定や分散分析では、従属変数に対する要因の効果を検定する。

※ 水準 Level

これは、要因のとる値のこと

例 : 実験協力者の性（要因）⇒男（水準１）女（水準２）

（２）交絡 Confounding

ある独立変数の効果を見たいのに他の要因が連動して絡んできてどの要因に効果があるかわからない状況。

例１ ミニマルペア　　容認度

[1] a. He is a good boy. 5
b. He am a good girl. 1

[2] a. He is a good boy. 5
b. He am a good boy. 1

例２ 被験者の特徴

[1] a. He is a good boy. 英語母語話者 5
b. He a good boy. 英語母語話者 3

[2] a. He is a good boy. DC 育ちの英語母語話者 5
b. He a good boy. DC 育ちの英語母語話者 1

（３）統制

① 操作可能性

研究者が独立変数の値を変化させられるかということ。

例１：性別　研究する側が変更することはできない。
例２：刺激文　研究する側が変更することができる。

② 統制 Control

交絡を避けるためターゲットの要因以外の要因が従属変数に影響を持たないように偏りを調整すること。

（統制１）**一定化**：変数値を固定し偏りを調整。
例：実験協力者を東京方言話者に限る

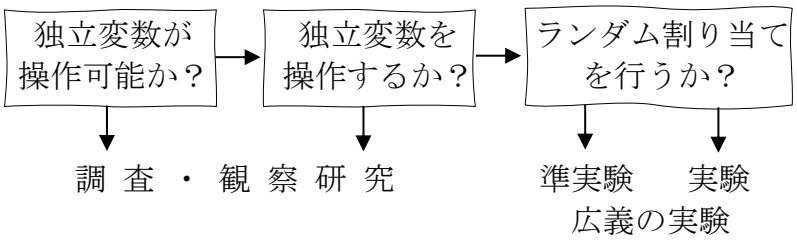
（統制２）**バランス化**：変数値を同数にし偏りを調整。
例：実験協力者を男女同数にする

（統制３）**ランダム化**：変数値を無作為にし偏りを調整。
例：無作為に発生させた番号にかける

③ 研究デザインの強弱

独立変数の操作を組み込んだ研究デザインを**強いデザイン**、そうでない研究デザインを**弱いデザイン**と呼ぶ。

※ 観察研究と実験研究



（３）対応のない要因 vs. 対応のある要因

① **対応のない要因**： 異なる水準に含まれる従属変数の値が互いに独立となるような要因のこと。

② **対応のある要因**： 異なる水準に含まれる従属変数の値に相関がある要因のこと。

（４）バランスデザイン vs. アンバランスデザイン

① **バランスデザイン**： 各水準のサンプル数が等しい実験デザイン（理想的なデザイン）。

② **アンバランスデザイン**： 各水準のサンプル数が異なる実験デザイン。

📖 増補： 直交実験デザイン

私の実験、参加者一人一人に 288 個の文を読んでもらうことになるんだけど…

「でも、そんなに読ませたら。かなり負担だろうし、何より実験の狙いに勘付かれてしまうかもしれない！」と不安になりますよね。このような不都合な状況を改善させるために、実験のサイズを小さくするような実験デザインが考案されています。

（1）直交配列表実験

少ない実験回数で重要な主効果と交互作用が推定できるように、直交表に因子水準を割り当ててデザインする実験。

① 動機：実験回数を少なく抑えたい。

② 欠点：**交互作用の制限**
すべての交互作用は検討できない。
⇒あらかじめ「考慮しない交互作用」を決めておく。

（2）ラテン方格とグレコ・ラテン方格

① ラテン方格 Latin square
n 個のラテン文字を n 行×n 列に並べ、どの文字も各行・各列に一度ずつ現れるようにしたもの。

② グレコ・ラテン方格 Graeco-Latin square
ギリシャ文字とラテン文字が、単体でも、組み合わせでも、一度ずつ現れるようにしたもの。

（例）ラテン方格

	1	2	3	4
1	A	B	C	D
2	B	A	D	C
3	C	D	A	B
4	D	C	B	A

（例）グレコ・ラテン方格

	1	2	3	4
1	Aα	Bβ	Cγ	Dδ
2	Bγ	Aδ	Dα	Cβ
3	Cδ	Dγ	Aβ	Bα
4	Dβ	Cα	Bδ	Aγ

（３）直交表 Orthogonal table

どのような水準の組み合わせで実験を構築すると過不足なくデータを収集できるかを表した表。

記法：$L_8(2^7)$

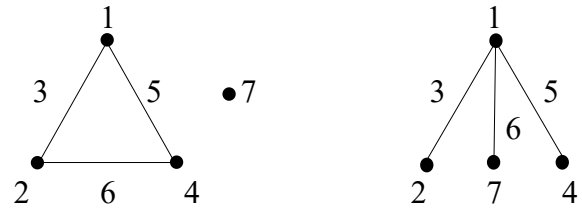
（例）二水準の直交配列実験 $L_8(2^7)$

	[1]	[2]	[3]	[4]	[5]	[6]	[7]	
1	0	0	0	0	0	0	0	
2	0	0	0	1	1	1	1	
3	0	1	1	0	0	1	1	
4	0	1	1	1	1	0	0	
5	1	0	1	0	1	0	1	
6	1	0	1	1	0	1	0	
7	1	1	0	0	1	1	0	
8	1	1	0	1	0	0	1	
	a		a		a		a	
		b	b			b	b	
				c	c	c	c	

（４）線点図 Linear graph

因子の割り当てに用いるグラフで、因子（主効果）を点で表し、二点を結んだ線分を交互作用に対応させたもの。

（例）$L_8(2^7)$



※ 各点・線は、直交表のどこかの列を表す。

🕮 ノート２ 母集団：統計モデル

（１）一元配置分散分析 one-way layout ANOVA model

分散分析の構造を表すモデルには複数の書き表し方がある。ここでは、よく使われる三つのモデルを紹介する。

> ☞ アドバイス
>
> モデル式の表していることが分からなくなったら、
> ①各記法が図中のどこを指しているか、考えよう！
> ②添え字が観測値の何を指しているか、考えよう！

① モデル１：グループの平均に注目したモデル

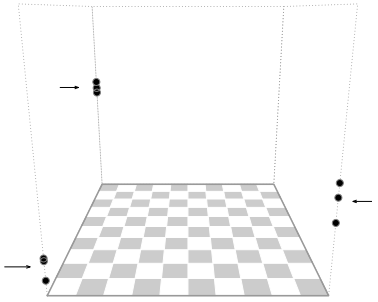
$$y_{ij} = \mu_j + \varepsilon_{ij}$$

$$\varepsilon_{ij} \sim N(0, \sigma^2)$$

ダミー変数で表現

$$y_{ij} = \mu_1 x_{i1} + \mu_2 x_{i2} + \cdots + \mu_J x_{iJ} + \varepsilon_{ij}$$

$$\varepsilon_{ij} \sim N(0, \sigma^2)$$



※記法

- y_{ij}　j 番目の水準（グループ）の i 番目のデータ
- ε_{ij}　j 番目の水準の i 番目のデータの残差
- μ　全体平均
- μ_j　j 番目の水準の平均
- α_j　j 番目の水準の処置効果
  - ① 全体平均からの差　⇒　モデル２
  - ② 基準となる水準からの差　⇒　モデル３

② モデル２：零和制約を持つモデル

$$y_{ij} = \mu + \alpha_j + \varepsilon_{ij}$$

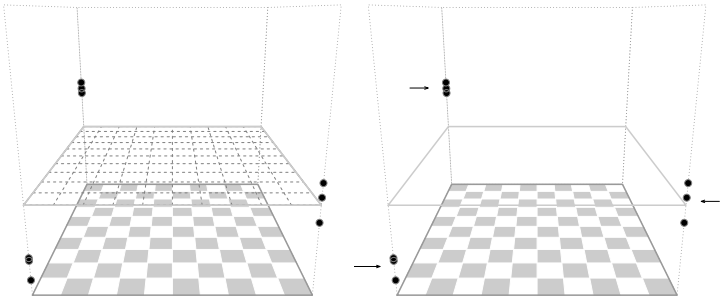
$$\varepsilon_{ij} \sim N(0, \sigma^2)$$

$$\sum_{j=1}^{J} \alpha_j = 0$$

ダミー変数で表現

$$y_{ij} = \mu + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \cdots + \alpha_J x_{iJ} + \varepsilon_{ij}$$

$$\varepsilon_{ij} \sim N(0, \sigma^2)$$



③ モデル３：端点制約を持つモデル

$$y_{ij} = \mu + \alpha_j + \varepsilon_{ij}$$

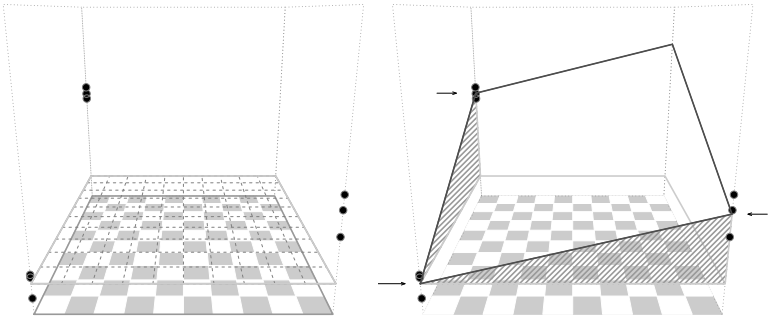
$$\varepsilon_{ij} \sim N(0, \sigma^2)$$

$$\alpha_1 = 0$$

ダミー変数で表現

$$y_{ij} = \mu + \alpha_2 x_{i2} + \cdots + \alpha_J x_{iJ} + \varepsilon_{ij}$$

$$\varepsilon_{ij} \sim N(0, \sigma^2)$$



（2）二元配置分散分析 two-way layout ANOVA model

① モデル１：グループの平均に注目したモデル

$$y_{ijk} = \mu_{jk} + \varepsilon_{ijk}$$

$$\varepsilon_{ijk} \sim N(0, \sigma^2)$$

ダミー変数で表現

$$\begin{aligned} y_{ijk} = {} & \mu_{11}x_{i1} + \mu_{12}x_{i2} + \cdots + \mu_{1K}x_{iK} \\ & + \mu_{21}x_{i\,K+1} + \mu_{22}x_{i\,K+2} + \cdots + \mu_{2K}x_{i\,K+K} \\ & \vdots \\ & + \mu_{J1}x_{i\,(J-1)K+1} + \mu_{J2}x_{i(J-1)K+2} + \cdots + \mu_{JK}x_{i(J-1)K+K} \\ & + \varepsilon_{ijk} \end{aligned}$$

$$\varepsilon_{ijk} \sim N(0, \sigma^2)$$

② モデル２：零和制約を持つモデル

$$y_{ijk} = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk} + \varepsilon_{ijk}$$

$$\varepsilon_{ijk} \sim N(0, \sigma^2)$$

$\sum_{j=1}^{J} \alpha_j = 0$

$\sum_{k=1}^{K} \beta_k = 0$

$\sum_{j=1}^{J} (\alpha\beta)_{j1} = 0$
$\sum_{j=1}^{J} (\alpha\beta)_{j2} = 0$
$\vdots$
$\sum_{j=1}^{J} (\alpha\beta)_{jK} = 0$

$\sum_{k=1}^{K} (\alpha\beta)_{1k} = 0$
$\sum_{k=1}^{K} (\alpha\beta)_{2k} = 0$
$\vdots$
$\sum_{k=1}^{K} (\alpha\beta)_{Jk} = 0$

ダミー変数で表現

$$\begin{aligned} y_{ijk} = {} & \mu + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \cdots + \alpha_J x_{iJ} \\ & + \beta_1 x_{iJ+1} + \beta_2 x_{iJ+2} + \cdots + \beta_K x_{iJ+K} \\ & + \alpha\beta_{11} x_{iJ+K+1} + \alpha\beta_{12} x_{iJ+K+2} + \cdots + \alpha\beta_{JK} x_{iJ+K+JK} \\ & + \varepsilon_{ijk} \end{aligned}$$

$$\varepsilon_{ijk} \sim N(0, \sigma^2)$$

③ モデル３：端点制約を持つモデル

$$y_{ijk} = \mu + \alpha_j + \beta_k + \alpha\beta_{jk} + \varepsilon_{ijk}$$

$$\varepsilon_{ijk} \sim N(0, \sigma^2)$$

$\alpha_1 = 0$　$\beta_1 = 0$

$$\alpha\beta_{11} = 0,\ \alpha\beta_{12} = 0,\ \cdots,\ \alpha\beta_{1K} = 0$$

$$\alpha\beta_{11} = 0,\ \alpha\beta_{21} = 0,\ \cdots,\ \alpha\beta_{J1} = 0$$

ダミー変数で表現

$$\begin{aligned} y_{ijk} = \mu &+ \alpha_2 x_{i1} + \alpha_3 x_{i2} + \cdots + \alpha_J x_{i\,J-1} \\ &+ \beta_2 x_{i\,J} + \beta_3 x_{i\,J+1} + \cdots + \beta_K x_{i\,J+K-2} \\ &+ \alpha\beta_{22} x_{i\,J+K-1} + \alpha\beta_{23} x_{i\,J+K} + \cdots + \alpha\beta_{2K} x_{i\,J+2K-2} \\ &+ \alpha\beta_{32} x_{i\,J+2K-1} + \alpha\beta_{33} x_{i\,J+2K} + \cdots + \alpha\beta_{3K} x_{i\,J+3K-3} \\ &\vdots \\ &+ \alpha\beta_{J2} x_{i\,K-K+1} + \alpha\beta_{J3} x_{i\,JK-K+2} + \cdots + \alpha\beta_{JK} x_{i\,JK-1} \\ &+ \varepsilon_{ijk} \end{aligned}$$

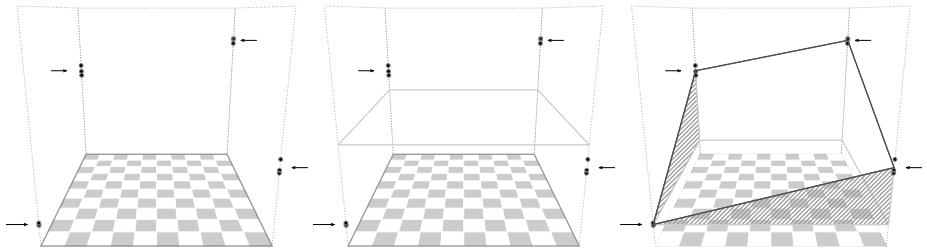
$$\varepsilon_{ijk} \sim N(0, \sigma^2)$$

※独立なパラメータの数：

$$1 + (J-1) + (K-1) + (J-1)(K-1) = JK$$

※　図による理解

モデル１　　モデル２　　モデル３



📖 増補：†共分散分析　ANCOVA

私の実験、各水準に被験者をランダムに割り当てられないんです！

「英会話スクールに通ったかどうか」を要因にしたいのだけど、「親の熱意」に関して無頓着となるように被験者をランダムに割り当てるのは難しい。こういうときは統制の一つの手段として「親の熱意」も変数としてモデルに取り込む方法がある。

（1）例：英語教育

小学校時代の英会話スクールへの通学が、中学校の英語のテストの成績に影響を与えているか？

モデル1

従属変数　：中学の共通テストの英語の成績
独立変数 1　：英会話スクールの有無
（要因）　（通った　vs. 通わなかった）

$$y_{ij} = \mu + \alpha_j + \varepsilon_{ij}$$

▼

モデル1の問題

「英会話スクールの有無」ではなくて、「親の教育への熱心さ」も影響を与えているのではないの？

⇒　モデル 1 では原因が交絡し、わからない。

▼

モデル2

「親の教育への熱心さ」を測定し、分散分析モデルに取り込む！

独立変数 2　：親の教育への熱心さ
（通った　vs. 通わなかった）

$$y_{ij} = \mu + \alpha_j + \beta w_i + \varepsilon_{ij}$$

（2）共分散分析 Analysis of Covariance

① 共変量／共変数 covariate

これは、モデルに組み入れられた要因のほかに観測値に影響を与える変数のこと。

② 共分散分析

これは、共変量と呼ばれる量的な変数をモデルに投入することで、残差を減らし、検定力を高める統計分析。

$$y_{ij} = \mu + \alpha_j + \beta w_i + \varepsilon_{ij}$$

⇒ 結局は特殊な場合の **重回帰モデル** に相当する。

> ☞ ポイント：βに添え字の jがない！
>
> これは添え字j（つまりグループ）によらず共変量が従属変数の予測に与える影響は一定だという仮定を表す。
>
> ⇒ すべてのグループが同じ傾きを持つということ。

③ 「回帰係数の等質性の仮定」の検定

グループごとの傾きはδ_jだけ違うという下のモデルを立て「δ_jが全て０である」という帰無仮説を検討。

$$y_{ij} = \mu + \alpha_j + (\beta + \delta_j)(x_{ij} - \bar{x}) + \varepsilon_{ij}$$

> ☞ ポイント：棄却できないことを願う変則的な検定
>
> 有意でない（帰無仮説を保持することになった場合）に、②の共分散分析のモデルを利用することになる。

※δ_jの意味

独立変数１と共変数（独立変数２）に影響があることを表す、独立変数１と２の交互作用を表している。

⇒ δ_jは固定効果として扱われる点でランダム係数モデルとは異なる。

📖 ノート３ 分析の評価１：作った統計モデルの正確さを評価

（参照：第 5 講ノート 4）

フェーズ５：分散分析

手順１：平方和を分解する

（例１）$ss_y = ss_A + ss_e$　　一元配置

（例２）$ss_y = ss_A + ss_B + ss_{A\times B} + ss_e$　二元配置

手順２：残差平方和ss_eに比して独立変数の平方和（ss_A, ss_B, $ss_{A\times B}$）が大きいと言えるか F 検定する

（１）回帰分析における平方和の分解

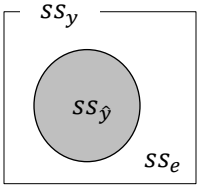
① 単回帰分析

$$y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i$$

$$\varepsilon_i \sim N(0, \sigma^2)$$

	平方和	自由度	平均平方和	F
$\hat{y}$	$ss_{\hat{y}}$	1	$MS_{\hat{y}}$	$MS_{\hat{y}}/MS_e$
e	ss_e	$n-1-1$	MS_e	
y	ss_y	$n-1$		

$ss_y = ss_{\hat{y}} + ss_e$



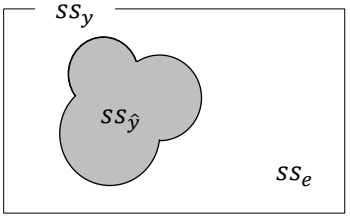
② 重回帰分析

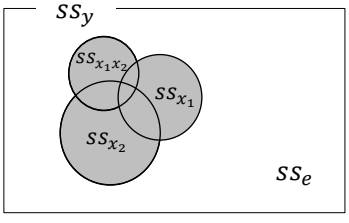
$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + \varepsilon_i$$

$$\varepsilon_i \sim N(0, \sigma^2)$$

	平方和	自由度	平均平方和	F
$\hat{y}$	$ss_{\hat{y}}$	p	$MS_{\hat{y}}$	$MS_{\hat{y}}/MS_e$
e	ss_e	$n-1-p$	MS_e	
y	ss_y	$n-1$		

$ss_y = ss_{\hat{y}} + ss_e$　　$ss_y \neq ss_A + ss_B + ss_{AB} + ss_e$





（２）分散分析における平方和の分解

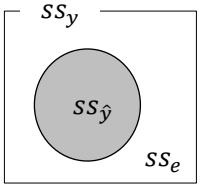
① 一元配置分散分析

$$y_{ij} = \mu + \alpha_j + \varepsilon_{ij}$$

$$\varepsilon_{ij} \sim N(0, \sigma^2)$$

	平方和	自由度	平均平方和	F
$\hat{y}$	$ss_{\hat{y}}$	$J-1$	$MS_{\hat{y}}$	$MS_{\hat{y}}/MS_e$
e	ss_e	$n-J$	MS_e	
y	ss_y	$n-1$		

$$ss_y = ss_{\hat{y}} + ss_e$$



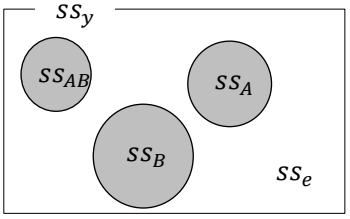
② 二元配置分散分析（バランスデザイン）

$$y_{ijk} = \mu + \alpha_j + \beta_k + \alpha\beta_{jk} + \varepsilon_{ijk}$$

$$\varepsilon_{ijk} \sim N(0, \sigma^2)$$

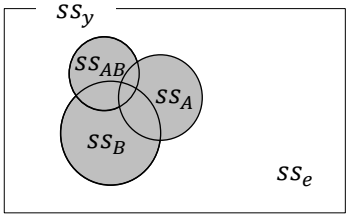
	平方和	自由度	平均平方和	F
要因A	ss_A	$J-1$	MS_A	MS_A/MS_e
要因B	ss_B	$K-1$	MS_B	MS_B/MS_e
交互作用	ss_{AB}	$(J-1)(K-1)$	MS_{AB}	MS_{AB}/MS_e
e	ss_e	$n-JK$	MS_e	
y	ss_y	$n-1$		

$$ss_y = ss_{\hat{y}} + ss_e$$
$$= ss_A + ss_B + ss_{AB} + ss_e$$



※アンバランスデザインでは平方和の分解は不成立！

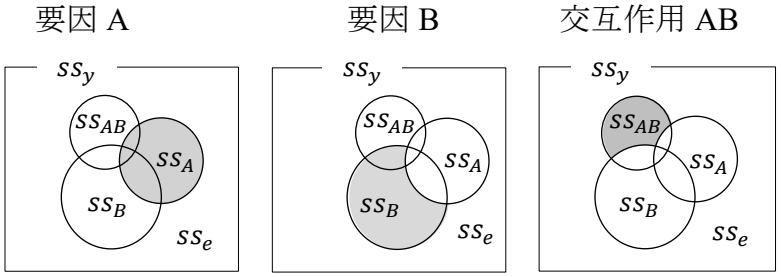
$$ss_y \neq ss_A + ss_B + ss_{AB} + ss_e$$



（3）平方和の種類

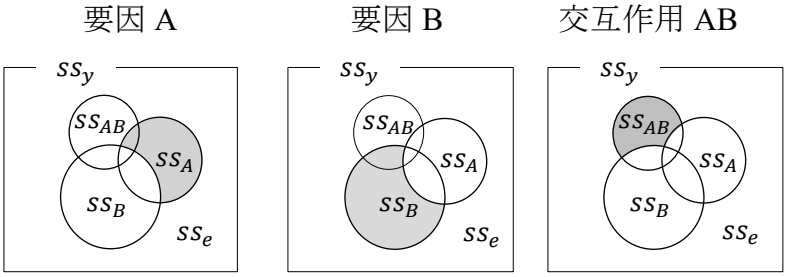
① タイプⅠの平方和

各独立変数（要因）の平方和を、モデルに組み込んだ順に評価するもの。



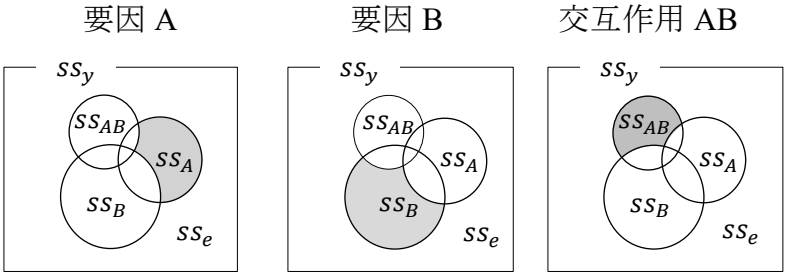
② タイプⅡの平方和

その独立変数（要因）でしか説明できない部分を、その変数の平方和と考えるもの。



③ タイプⅢの平方和

その独立変数（要因）でしか説明できない部分を、その変数の平方和と考えるもの。



（4）重相関係数と相関比

① 重相関係数 R Multiple correlation coefficient

これは実測値 y と 予測値 $\hat{y}$ の相関係数であり R と表す。

$$R = r_{y\hat{y}}$$

② 相関比 η Correlation ratio

これは実測値 y と 予測値 $\hat{y}$ の相関係数であり η と表す。

$$\eta = r_{y\hat{y}} = R$$

※ 分散分析では R の代わりに η と書く慣習がある。

（5）決定係数 Coefficient of determination

これは独立変数がどれだけ従属変数の値を決定するかを示す指標。別名：分散説明率 Proportion of variance accounted for

① R^2 を用いた表記

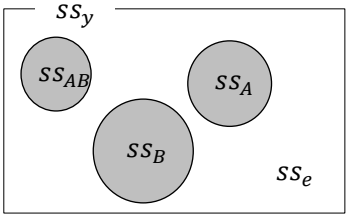
$$R^2 = \frac{s_{\hat{y}}^2}{s_y^2} = 1 - \frac{s_e^2}{s_y^2}$$

② η^2 を用いた表記

$$\eta^2 = \frac{s_{\hat{y}}^2}{s_y^2} = 1 - \frac{s_e^2}{s_y^2}$$

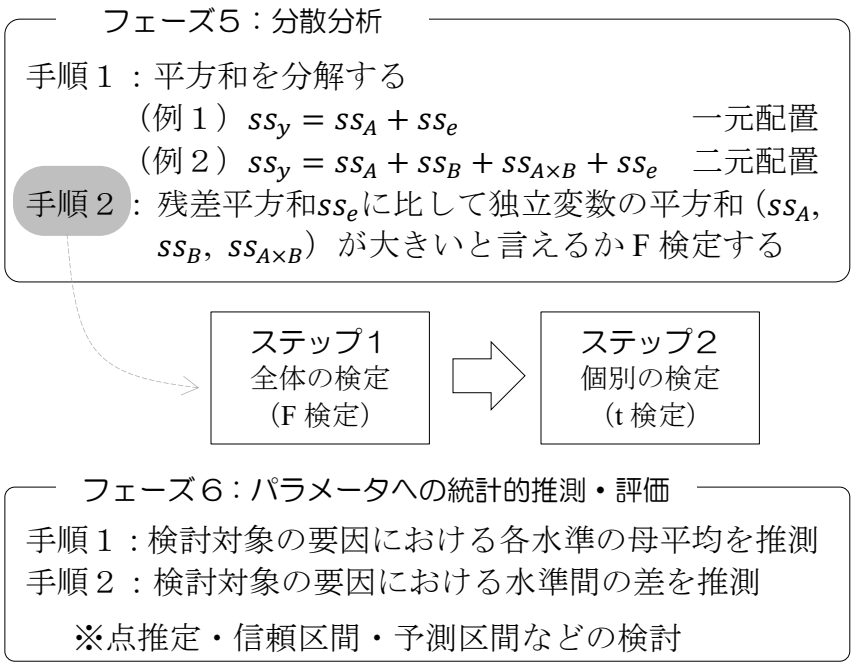
$$ss_y = ss_{\hat{y}} + ss_e$$

$$= ss_A + ss_B + ss_{AB} + ss_e$$



※ 分散分析では R の代わりに η^2 と書く慣習がある。

🕮 ノート４ 分析の評価２：係数の推定値の正確さの評価



（１）検定１：「全体」の検定　　ステップ1

バランスデザインの二元配置分散分析を例に概観する。

$$y_{ijk} = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk} + \varepsilon_{ijk}$$

① 帰無仮説と対立仮説

■ 要因 A の主効果に関する仮説
H0：「$\alpha_1 = \alpha_2 = \cdots = \alpha_J = 0$（すべての係数が０）」
H1：「$\alpha_1, \alpha_2, \ldots, \alpha_J$の少なくとも一つはゼロではない」

■ 要因 B の主効果に関する仮説
H0：「$\beta_1 = \beta_2 = \cdots = \beta_K = 0$（すべての係数が０）」
H1：「$\beta_1, \beta_2, \ldots, \beta_K$の少なくとも一つはゼロではない」

■ 交互作用効果に関する仮説
H0：「$(\alpha\beta)_{11} = (\alpha\beta)_{12} = \cdots = (\alpha\beta)_{JK} = 0$」
H1：「少なくとも一つの$(\alpha\beta)_{jk}$はゼロではない」

② 検定統計量 F

☞ F 値（要因 A の検定における検定統計量）

$$F_A = \frac{ss_A/(J-1)\text{「要因 A の水準}\hat{y}_{j\cdot}\text{のばらつき」}}{ss_e/JK(I-1)\text{「}y_{ijk}\text{の}\hat{y}_{ijk}\text{からのばらつき」}}$$

$$\mathrm{ss_A} = KI\sum_{j=1}^{J}\left(\bar{y}_{j\cdot}-\bar{y}\right)^2 \quad ss_e = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\left(y_{ijk}-\bar{y}_{jk}\right)^2$$

☞ F 値（要因 B の検定における検定統計量）

$$F_B = \frac{ss_B/(K-1)\text{「要因 B の水準}\hat{y}_{\cdot k}\text{のばらつき」}}{ss_e/JK(I-1)\text{「}y_{ijk}\text{の}\hat{y}_{ijk}\text{からのばらつき」}}$$

$$\mathrm{ss_B} = JI\sum_{k=1}^{K}(\bar{y}_{\cdot k}-\bar{y})^2$$

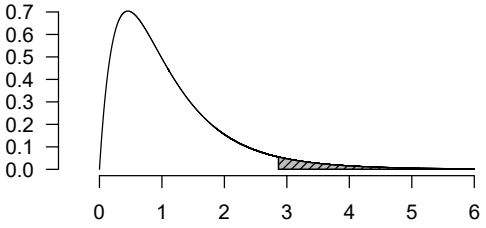
☞ F 値（交互作用の検定における検定統計量）

$$F_{\mathrm{A\times B}} = \frac{ss_{A\times B}/(J-1)(K-1)\text{「交互作用のばらつき」}}{ss_e/JK(r-1)\text{「}y_{ijk}\text{の}\hat{y}_{\mathrm{ijk}}\text{からのばらつき」}}$$

$$\mathrm{ss_{A\times B}} = I\sum_{j=1}^{J}\sum_{k=1}^{K}\left(\bar{y}_{jk}-\bar{y}_{j\cdot}-\bar{y}_{\cdot k}+\bar{y}\right)^2$$

③ F 分布

H0 が真なら F 値は F 分布に従う。



④ 分散分析表

	平方和	自由度	平均平方和	F 値
要因 A	ss_A	$J-1$	$MS_A = \frac{ss_A}{J-1}$	$F = \frac{MS_A}{MS_e}$
要因 B	ss_B	$K-1$	$MS_B = \frac{ss_B}{K-1}$	$F = \frac{MS_B}{MS_e}$
交互作用	$ss_{A\times B}$	$(J-1)(K-1)$	$MS_{A\times B} = \frac{ss_{A\times B}}{(J-1)(K-1)}$	$F = \frac{MS_{A\times B}}{MS_e}$
残差	ss_e	$JK(I-1)$	$MS_e = \frac{ss_e}{JK(I-1)}$	
全体	ss_y	$n-1$	$MS_y = \frac{ss_y}{n-1}$	

（２）検定２：「個別」の回帰係数の検定　　ステップ２

① 帰無仮説と対立仮説

H0：$\beta_2 = 0$

H1：$\beta_2 \neq 0$

② 検定統計量 t

☞ t 値（水準の検定における検定統計量）

$$t = \frac{\hat{\beta}_i \text{「}\beta_i\text{の推定値」}}{\hat{\sigma}_{\hat{\beta}_i} \text{「}\beta_i\text{の推定値」の標準誤差の推定値}}$$

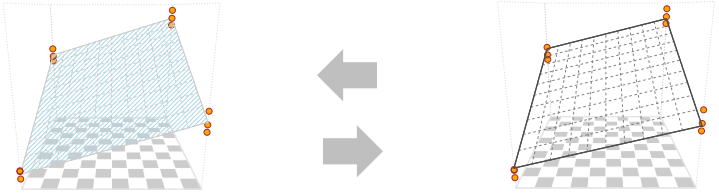
	点推定値	標準誤差	t value	Pr(>|t|)	
$\hat{\mu}$	-5.03	0.29	-16.9	2e-16	***
$\hat{\alpha}_2$	0.47	0.04	11.18	1.93e-14	***
$\hat{\beta}_2$	0.99	0.04	23.48	2e-16	***
$\widehat{\alpha\beta}_{22}$	0.00	0.01	1.50	0.14	

F-statistic: 36.57 on 3 and 98 DF, p-value: 6.032e-16

（３）予測値（グループ平均）の点推定

母集団モデルが正しいという仮定の下、最小二乗法を用いて行った点推定の結果は次のように与えられる。

標本　　母集団



$y_{ijk} = \hat{\mu} + \hat{\alpha}_j + \hat{\beta}_k + \left(\widehat{\alpha\beta}\right)_{jk} + e_{ijk}$　　$y_{ijk} = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk} + \varepsilon_{ijk}$

$$(\alpha\beta)_{jk} = \bar{y}_{ij\cdot} - \bar{y}_{\cdot j\cdot} - \bar{y}_{\cdot\cdot k} + \bar{y}_{\cdots}$$

$$\hat{\beta}_k = \bar{y}_{\cdot\cdot k} - y_{\cdots}$$

$$\hat{\alpha}_j = \bar{y}_{\cdot j\cdot} - y_{\cdots}$$

$$\hat{\mu} = \bar{y}_{\cdots}$$

$$\hat{\mu}_{jk} = \hat{\mu} + \hat{\alpha}_j + \hat{\beta}_k$$

$$= \bar{y}_{\cdot j\cdot} + \bar{y}_{\cdot\cdot k} - y_{\cdots}$$

（4）予測値（グループ平均）の信頼区間

標本を繰り返し抽出した時、この範囲を設けておけば 100 回に 95 回、真のμ_{jk} を含むだろう、という区間。

$$\left[\hat{\mu}_{jk} - t_c \times \sqrt{\frac{1}{n_e}\sigma^2}, \quad \hat{\mu}_{jk} + t_c \times \sqrt{\frac{1}{n_e}\sigma^2}\right]$$

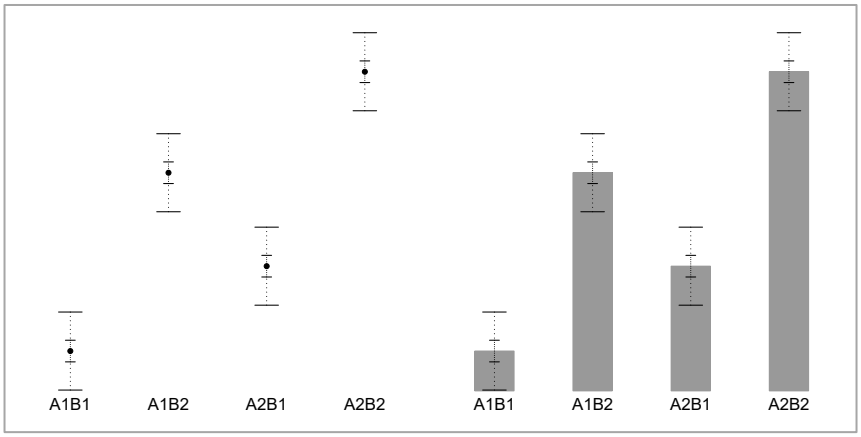
（5）データの 95%予測区間

将来データを採っても、この範囲を設けておけば 100 回に 95 回データを含むだろう、という区間。

$$\left[\hat{\mu}_{jk} - t_c \times \sqrt{\left(1+\frac{1}{n_e}\right)\sigma^2}, \quad \hat{\mu}_{jk} + t_c \times \sqrt{\left(1+\frac{1}{n_e}\right)\sigma^2}\right]$$

※n_eは反復有効数と呼ばれ、伊奈の式／田口の式で求める。

（6）結果の図示（信頼区間と予測区間）



※論文によっては信頼区間だけ表示している場合もある。

Ｒでの実装

長畑秀和 (2016)『R で学ぶ実験計画法』東京 : 朝倉書店に様々ケースのとても実践的なコードが存在する。

📖 増補１：多重比較

（１）多重比較の問題

① 多重比較

分散分析で有意差が見つかった要因に対し**二群の差の検定**を各水準の組み合わせに対し何度も検討すること。

② 多重比較の問題（多重性）

同じデータに対して検定を繰り返すことで、**第一種の過誤**の確率が上がってしまう。

> ☞ 第一種の過誤を犯す確率
>
> 「k 回中少なくとも一回は第一種の過誤を犯す確率」
>
> $= 1 -$「k 回中一回も第一種の過誤を犯さない確率」
>
> $= 1 - (1 - \alpha)^k$

例：1 要因（3 水準）

水準１

水準２　　水準３

③ 多重比較法（多重比較検定）の系統

- 多重比較法
  - 分布調整型
    - t 分布や F 分布の代わりに第一種の過誤に強い分布を採用
    - （例１）Tukey の方法
    - （例２）Hochberg の方法
    - （例３）Gabriel の方法
    - （例４）Dunnett の方法
  - 統計量調整型
    - （例）　Scheffe の方法
  - 有意水準調整型
    - （例１）Bonferroni の方法
    - （例２）Holm の方法
    - （例３）Sidak の方法
    - （例４）P-E-G-W の方法
    - （例５）Tamhane の方法

（２）有意水準調整型

■ ボンフェロニ（Bonferroni）の方法

① 方法

検定「全体」の第一種の過誤を犯す確率をαとするとき個々の検定の有意水準を $\frac{\alpha}{k}$ にする。

② 利点

（利点１）各検定が独立でない場合にも利用可能
（利点２）正規分布以外の分布でも利用可能

③ 欠点

kが大きくなると検出力が落ちてしまう。

検出力の向上

■ ホルム（Holm）の方法

① 方法

(Step 1) k個の各検定をp値が小さい順に並べる。

(Step 2) i番目の検定の有意水準を$\frac{\alpha}{k-i+1}$ とする。

(Step 3) 頭から順番に検定を行い、m番目で初めて H0 が「保持」されたとする。このとき、

(i) 1番目から$m-1$番目までの検定の H0 を「棄却」

(ii) m番目からの k 個の各検定は H0 を「保持」する

② 利点

（利点１）各検定が独立でない場合にも利用可能

（利点２）正規分布以外の分布でも利用可能

（利点３）ボンフェロニよりも検出力が高い。

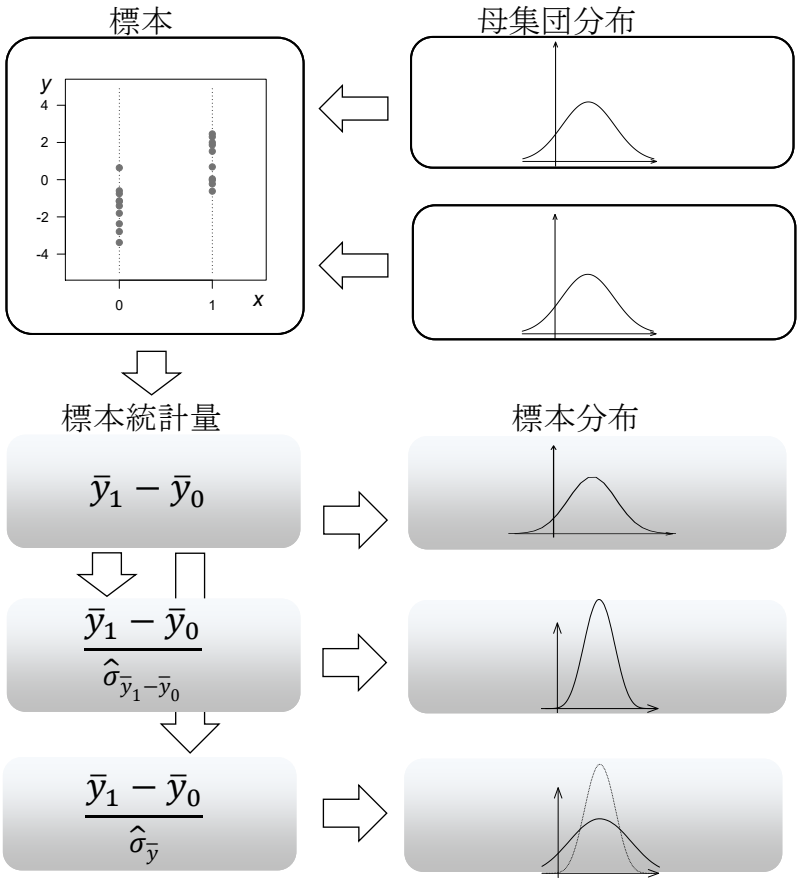
（３）分布調整型

■ テューキー（Tukey）の方法

① 方法

t 値を修正した q 値を用いて多重比較を行う。

$$t = \frac{\bar{y}_i - \bar{y}_j}{\hat{\sigma}_{\bar{y}_i - \bar{y}_j}} \qquad q = \frac{\bar{y}_i - \bar{y}_j}{\hat{\sigma}_{\bar{y}}}$$



② 利点

t 分布に比べ、検定を繰り返しても棄却されにくい分布。

③ 欠点

（欠点１）バランスデータにしか使えない。

※ アンバランスなデータに適用できるように改良された Tukey-Kramer の方法というものもある。

（欠点２）対応のない分析にしか使えない。

🕮 ノート５分析の評価３：想定した統計モデルの適切さの評価

（参照：第 5 講ノート 6）

（１）†正規性の仮定の検証

H0：母集団分布が正規分布だ
H1：「母集団分布が正規分布だ」は偽

① コロモゴロフ＝スミルノフ（K-S）の検定
理論的な正規分布と、観測データのヒストグラムを比べて、両者の差を比較して H0 の妥当性を考える方法。
② シャピロ＝ウィルク（Shapiro-Wilk） の検定
ケース数が少ない場合に利用される。

（２）†等分散性の仮定の検証

H0：全ての水準の分散は等しい
H1：「全ての水準の分散は等しい」は偽

① バートレット（Bartlett）の検定
ルビーンの検定と比べ、正規分布／それに正規分布の標本の等分散性の検定での検出に優れるとされる。

② ルビーン（Levene）の検定
バートレットの検定よりもよく用いられる。

※あまり使われない
極端にサンプルサイズが少ないとき（例：25 個以下など）にのみに検討するのが普通。

（理由１）分散分析は正規性・等分散性に頑健。

（理由２）検定の多重性の問題を招く。

（理由３）棄却されないことを願う奇妙な検定。

SPSS コミュニティでの使用が多い。詳しい説明が欲しい場合 SPSS のマニュアルを見るといい。

https://www.spss-tutorials.com/

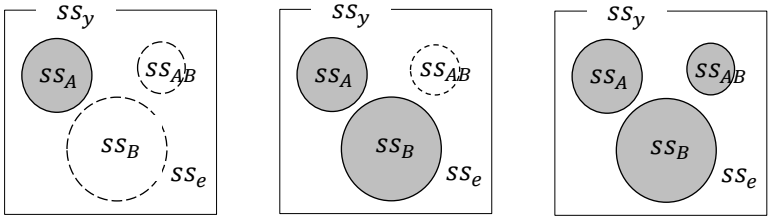
（3）検定に基づく変数選択

分散分析では第5章で見たモデル選択手法が一般的に使われるように前から「検定による変数選択」が使われていた。

- モデル選択
  - 情報量基準
  - クロスバリデーション
  - （ラッソ回帰）
  - （検定による変数選択）

① 基本的なアイデア

新しく変数を増やした時の**決定係数の増加量**に注目。



モデル１　　モデル２　　モデル３

② 帰無仮説と対立仮説

q個の独立変数からなるモデルに新たに$p-q$個変数を加え、全部でp個の独立変数を持つモデルを作るとき：

H0:前者の決定係数$R^2_{y\cdot 1\cdots q}$と後者の決定係数に差がない。

H1: H0 は成り立たない。

③ 検定統計量 F

決定係数の増加量を次の検定統計量の増加で調べる。

$$F=\frac{\left(R^2_{y\cdot 1\cdots p}-R^2_{y\cdot 1\cdots q}\right)/\left(p-q\right)}{\left(1-R^2_{y\cdot 1\cdots p}\right)/\left(n-p-1\right)}$$

	残差自由度	残差平方和	自由度	平方和	F	Pr(>F)
M1	46	1182.8				
M2	45	39.6	1	1143.1	1295.0	<2e-16
M3	44	38.8	1	0.84	0.9	0.3348

④ 弱点

（問題点１）検定の多重性の問題

（問題点２）ネストしているモデルしか使えない。

資料

資料6-1　　平方和の分解（バランスデザインの二元配置）

$$\sum_{i=1}^{I}\sum_{i=1}^{J}\sum_{k=1}^{K}\left(y_{ijk}-\bar{y}\right)^2$$

$$=\sum_{i=1}^{I}\sum_{i=1}^{J}\sum_{k=1}^{K}\left(y_{\cdot j\cdot}-\bar{y}\right)^2$$
$$+\sum_{i=1}^{I}\sum_{i=1}^{J}\sum_{k=1}^{K}\left(y_{\cdot\cdot k}-\bar{y}\right)^2$$
$$+\sum_{i=1}^{I}\sum_{i=1}^{J}\sum_{k=1}^{K}\left(y_{\cdot jk}-\bar{y}_{\cdot j\cdot}-\bar{y}_{\cdot\cdot k}+\bar{y}\right)^2$$
$$+\sum_{i=1}^{I}\sum_{i=1}^{J}\sum_{k=1}^{K}\left(y_{ijk}-\bar{y}_{\cdot jk}\right)^2$$

$$=\underbrace{IK\sum_{j=1}^{J}\hat{\alpha}_j^{\,2}}_{ss_A}+\underbrace{IJ\sum_{k=1}^{K}\hat{\beta}_i^2}_{ss_B}+\underbrace{I\sum_{j=1}^{J}\sum_{k=1}^{K}\widehat{\alpha\beta}_{jk}^2}_{ss_{AB}}+\underbrace{\sum_{i=1}^{I}\sum_{i=1}^{J}\sum_{k=1}^{K}e_{ijk}^2}_{ss_e}$$

$$ss_y=ss_A+ss_B+ss_{AB}+ss_e$$

	平方和	自由度	平均平方和	F
要因A	ss_A	$J-1$	MS_A	MS_A/MS_e
要因B	ss_B	$K-1$	MS_B	MS_B/MS_e
交互作用	ss_{AB}	$(J-1)(K-1)$	MS_{AB}	MS_{AB}/MS_e
e	ss_e	$N-JK$	MS_e	
y	ss_y	$N-1$		

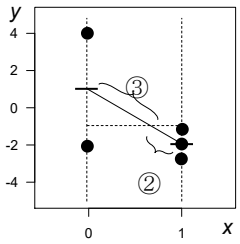
資料6-2　　F 値と t 値の関係

予測値の平方和の自由度が１のとき、すなわち二群の差の検定のとき、$F = t^2$となる。

（１）全体平均とグループ平均の関係

$\bar{y}_0$と $\bar{y}_1$を$n_1 : n_0$に内分する点が$\bar{y}$となっている。

$$\bar{y} = \frac{n_0}{n_0 + n_1}\ \bar{y}_0 + \frac{n_1}{n_0 + n_1}\ \bar{y}_1$$



（２）予測値の平方和

$$\begin{aligned}
ss_{\hat{y}} &= \sum_{i=1}^{n_0}(\hat{y}_0 - \bar{y})^2 + \sum_{i=1}^{n_1}(\hat{y}_1 - \bar{y})^2 \\
&= n_0(\hat{y}_0 - \bar{y})^2 + n_1(\hat{y}_1 - \bar{y})^2 \\
&= n_0\left(\hat{y}_0 - \frac{n_0}{n_0+n_1}\ \bar{y}_0 - \frac{n_1}{n_0+n_1}\ \bar{y}_1\right)^2 + n_1(\hat{y}_1 - \bar{y})^2 \\
&= n_0\left(\frac{n_0+n_1-n_0}{n_0+n_1}\ \bar{y}_0 - \frac{n_1}{n_0+n_1}\ \bar{y}_1\right)^2 + n_1(\hat{y}_1 - \bar{y})^2 \\
&= n_0\left(\frac{n_1}{n_0+n_1}\ \bar{y}_0 - \frac{n_1}{n_0+n_1}\ \bar{y}_1\right)^2 + n_1(\hat{y}_1 - \bar{y})^2 \\
&= n_0\left\{\frac{n_1}{n_0+n_1}\ (\bar{y}_0 - \ \bar{y}_1)\right\}^2 + n_1(\hat{y}_1 - \bar{y})^2 \\
&= n_0\left(\frac{n_1}{n_0+n_1}\right)^2\ (\bar{y}_0 - \ \bar{y}_1)^2 + n_1(\hat{y}_1 - \bar{y})^2 \\
&= \frac{n_0 n_1^2}{(n_0+n_1)^2}\ (\bar{y}_0 - \ \bar{y}_1)^2 + n_1(\hat{y}_1 - \bar{y})^2 \\
&= \frac{n_0 n_1^2}{(n_0+n_1)^2}\ (\bar{y}_0 - \ \bar{y}_1)^2 + \frac{n_0^2 n_1}{(n_0+n_1)^2}\ (\bar{y}_0 - \ \bar{y}_1)^2 \\
&= \frac{n_0 n_1 (n_0+n_1)}{(n_0+n_1)^2}\ (\bar{y}_0 - \ \bar{y}_1)^2 \\
&= \frac{n_0 n_1}{n_0+n_1}\ (\bar{y}_0 - \ \bar{y}_1)^2
\end{aligned}$$

（３）　残差の平方和

$$\begin{aligned}
ss_e &= \sum_{i=1}^{n_0}(\hat{y}_{i0} - \bar{y}_0)^2 + \sum_{i=1}^{n_1}(\hat{y}_{i1} - \bar{y}_1)^2 \\
&= n_0 s_1^2 + n_1 s_1^2
\end{aligned}$$

（4）F 値

これは、平方和を自由度で割ったものの比を表す。

$$F = \frac{\chi_1^2/df_1}{\chi_2^2/df_2}$$

★ 分子の平方和に$ss_{\hat{y}}$を代入（自由度１）

$$F = \frac{ss_{\hat{y}}/1}{\chi_2^2/df_2}$$

★ 分母の平方和にss_eを代入（自由度$n-2$）

$$F = \frac{ss_{\hat{y}}/1}{ss_e/(n-2)}$$

★（２）と（３）の結果を代入

$$F = \frac{\frac{n_0 n_1}{n_0+n_1}(\bar{y}_0-\bar{y}_1)^2}{(n_0 s_1^2+n_1 s_1^2)/(n-2)}$$

★ ss_eを$n-2$で割った分母は不偏分散となっている。これを$s'^2 = \frac{1}{n-2}(n_0 s_1^2 + n_1 s_1^2)$とおく。

$$F = \frac{\frac{n_0 n_1}{n_0+n_1}(\bar{y}_0-\bar{y}_1)^2}{s'^2}$$

$$= \frac{(\bar{y}_0-\bar{y}_1)^2}{s'^2} \times \frac{n_0 n_1}{n_0+n_1}$$

$$= \frac{(\bar{y}_0-\bar{y}_1)^2}{s'^2 \times \left(\frac{n_0 n_1}{n_0+n_1}\right)^{-1}}$$

$$= \frac{(\bar{y}_0-\bar{y}_1)^2}{s'^2 \times \left(\frac{1}{n_0}+\frac{1}{n_1}\right)}$$

★ このルートを取ると t 値となる。

$$t = \frac{\bar{y}_1-\bar{y}_0}{\sqrt{s'^2\left(\frac{1}{n_0}+\frac{1}{n_1}\right)}} = \sqrt{F}$$

$$= \frac{\bar{y}_1-\bar{y}_0}{s_{\bar{y}_1-\bar{y}_0}}$$

$$= \frac{\bar{y}_1-\bar{y}_0}{\hat{\sigma}^2_{\bar{y}_1-\bar{y}_0}}$$

資料6-3　乱塊法 Randomized block design

ブロック因子を導入し、ブロックごとで一通りの水準の組み合わせで実験を行い、それをブロックの数繰り返す実験。

① 動機：残差を小さくしたい

② 実験順序

完全にランダムな順序

因子 A	ID		
A_1（投与なし）	2	4	7
A_2（プラセボ）	1	5	8
A_3（薬 X 投与）	3	10	12
A_4（薬 Y 投与）	6	9	11

乱塊法

因子 A	ID		
A_1（投与なし）	2	5	12
A_2（プラセボ）	1	7	11
A_3（薬 X 投与）	4	6	9
A_4（薬 Y 投与）	3	8	10

③ モデル式：

$$y_{ij} = \mu + a_i + r_j + \varepsilon_{ij}$$

$$\varepsilon_{ij} \sim N(0, \sigma^2)$$

$$r_j \sim N(0, \sigma_R^2)$$

資料6-4　分割法 Split-plot design

これは、多元配置モデルにおいて、無作為化（ランダム化）を段階的に行う方法のこと。

ブロック因子を導入し、ブロックごとで一通りの水準の組み合わせで実験を行い、それをブロックの数繰り返す実験。

① 動機：ランダムな順番で実験を行うことが難しい

② 実験順序

完全にランダムな順序

因子 A	ID		
A_1（投与なし）	2	4	7
A_2（プラセボ）	1	5	8
A_3（薬 X 投与）	3	10	12
A_4（薬 Y 投与）	6	9	11

乱塊法

因子 A	ID		
A_1（投与なし）	2	5	12
A_2（プラセボ）	1	7	11
A_3（薬 X 投与）	4	6	9
A_4（薬 Y 投与）	3	8	10

③ モデル式：

$$y_{ijk} = \mu + a_i + r_k + \varepsilon_{(1)ik} + b_j + (ab)_{ij} + \varepsilon_{(2)ijk}$$

$$\varepsilon_{(2)ijk} \sim N\left(0, \sigma^2_{(2)}\right)$$

$$\varepsilon_{(1)ij} \sim N\left(0, \sigma^2_{(1)}\right)$$

$$r_j \sim N(0, \sigma^2_R)$$