Lecture 5: Two-Factor Studies with One Case per Treatment

STA 106: Analysis of Variance

Xiner Zhou
Department of Statistics, University of California, Davis

Example

(The Insurance Study)

The objective of the study:

City size (factor A)

Geographic region (factor B)

$\rightarrow$ Amount of premium for automobile insurance charged by an insurance company (for a specific type of coverage)

The study setup:

6 cities were selected to represent different regions of the state and different sizes of cities

(a) Premiums for Automobile Insurance Policy (in dollars)

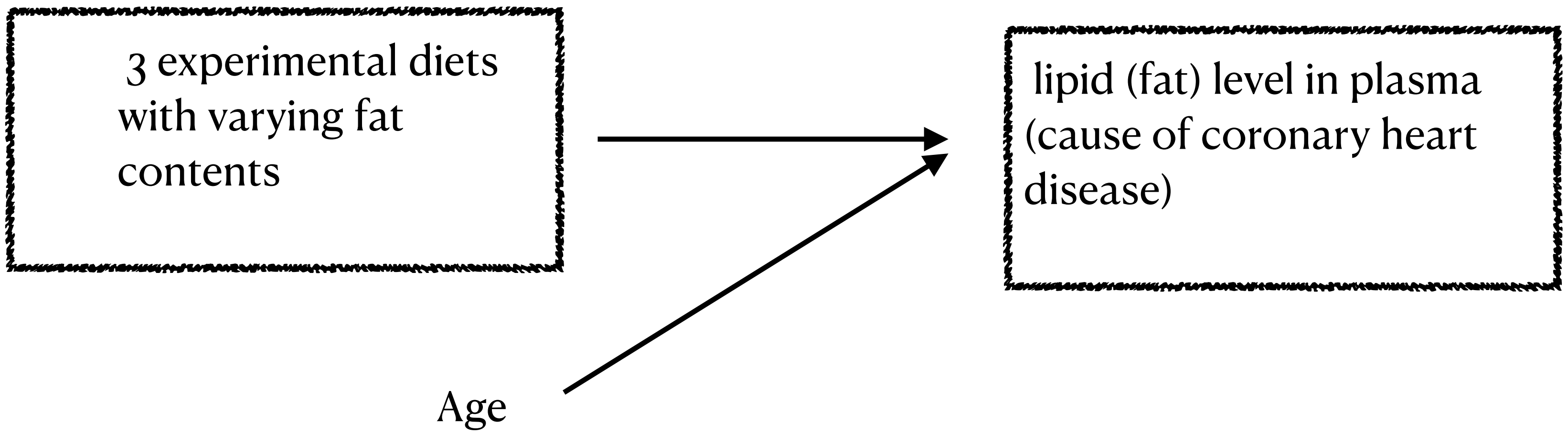
	Region (factor B)		
Size of City (factor A)	East ($j = 1$)	West ($j = 2$)	Average
Small ($i = 1$)	140	100	120
Medium ($i = 2$)	210	180	195
Large ($i = 3$)	220	200	210
Average	190	160	175

☑ Two-factor observational study

Example

(The Fat-in-Diet Study)

The objective of the study:



The study setup:

Within each block, 3 experimental diets were randomly assigned to the 3 subjects

Reduction in lipid level after some a certain period of time were recorded as the outcome

Block i		Fat Content of Diet $j = 1$ Extremely Low	$j = 2$ Fairly Low	$j = 3$ Moderately Low
1	Ages 15–24	.73	.67	.15
2	Ages 25–34	.86	.75	.21
3	Ages 35–44	.94	.81	.26
4	Ages 45–54	1.40	1.32	.75
5	Ages 55–64	1.62	1.41	.78

☑ Randomized Complete Block Design

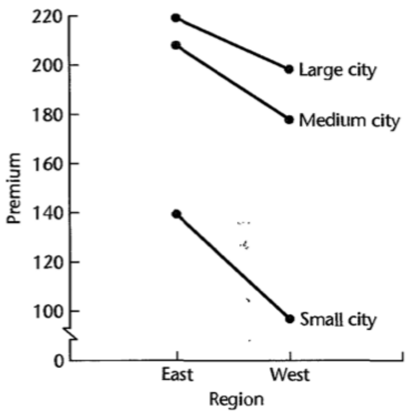
Two-Factor Studies with One Case per Treatment

Why and When would such cases occur?

- Constraints on cost, time, materials. …… severely limit the number of observations can be obtained
  - Only 1 subject (such as patient with certain characteristic) is available
- Outcome of interest is a single aggregated measure, there is no way to get more than 1 replicates for each treatment
  - Only aggregated measure at some large geographic regions, such as state, city…
  - Only aggregated measure at institutional level, such as hospitals, schools …
- Two very important kinds of studies
  - Randomized Complete Block Design
  - Observational studies with matched pairs

Two-Factor Studies with One Case per Treatment

What can go wrong with the two-factor factor effects model with interaction?



$$Y_{ij1} = \mu_{..} + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ij1}$$

$\rightarrow$ $\hat{Y}_{ij1} = \bar{Y}_{ij.} = Y_{ij1}$

$\rightarrow$ $e_{ij1} = 0$

$\rightarrow$ $MSE = 0$

With only one case per treatment, there is no way to estimate variability within treatments

there is no way to estimate error variance σ^2 by MSE, which is one parameter in the ANOVA model and the key to any inference .

Issue: over-parameterization

Two-Factor Studies with One Case per Treatment

Two-Way ANOVA Model without Interaction

Analysis of Variance

Test for Factor A and B Main Effects

Multiple Comparison Procedures

Tukey's test

Randomized Complete Block Designs

Two-Way ANOVA Model without Interaction

Assume factor A and B do not interact, i.e. all the interaction $\gamma_{ij} = 0$

$$Y_{ij} = \mu_{..} + \alpha_i + \beta_j + \varepsilon_{ij}$$ (Subscript k=1 dropped)

- $\mu_{..} = \frac{\sum_{i=1}^{a} \sum_{j=1}^{b} \mu_{ij}}{ab}$: overall mean

- $\alpha_i = \mu_{i.} - \mu_{..}$ main effect of factor A at ith level

  Subject to constraint $\sum \alpha_i = 0$

- $\beta_j = \mu_{.j} - \mu_{..}$ main effect of factor B at jth level

  Subject to constraint $\sum \beta_j = 0$

- $\gamma_{ij} = \mu_{ij} - \alpha_i - \beta_j = \mu_{ij} - \mu_{i.} - \mu_{.j} + \mu_{..}$ interaction effect of factor A at ith level with factor B at jth level

  Subject to a+b-1 constraints

$$\sum_i \gamma_{ij} = 0 \; j = 1,\ldots,b$$
$$\sum_j \gamma_{ij} = 0 \; i = 1,\ldots,a$$

- ε_{ij} are independent $N\left(0,\sigma^2\right)$ for $i = 1,\ldots,a; j = 1,\ldots,b$

Fitting the Two-Way ANOVA Model without interaction

Least squares estimates for parameters in factor effects parameterization:

$$\hat{\mu}_{..} = \frac{\sum_i \sum_j \hat{\mu}_{ij}}{ab} = \frac{\sum_i \sum_j Y_{ij}}{ab} = \bar{Y}_{..}$$

$$\hat{\alpha}_i = \hat{\mu}_{i.} - \hat{\mu}_{..} = \frac{\sum_j Y_{ij}}{b} - \bar{Y}_{..} = \bar{Y}_{i.} - \bar{Y}_{..}$$

$$\hat{\beta}_j = \hat{\mu}_{.j} - \hat{\mu}_{..} = \frac{\sum_i Y_{ij}}{a} - \bar{Y}_{..} = \bar{Y}_{.j} - \bar{Y}_{..}$$

Where:

$$\bar{Y}_{..} = \frac{\sum_i \sum_j Y_{ij}}{ab}$$

$$\bar{Y}_{i.} = \frac{\sum_j Y_{ij}}{b}$$

$$\bar{Y}_{.j} = \frac{\sum_i Y_{ij}}{a}$$

- fitted value for an observation Y_{ij}

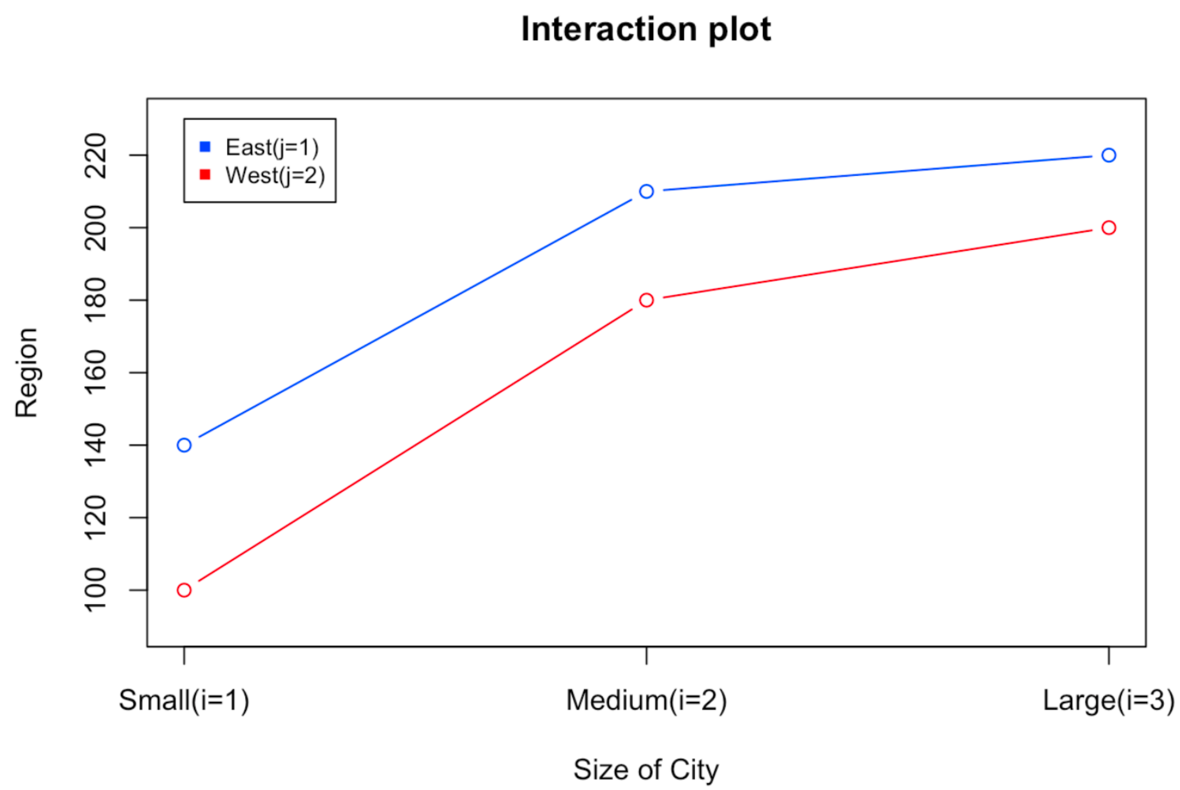
  ANOVA model's "best guess" or "best prediction"

  $$\hat{Y}_{ij} = \hat{\mu}_{..} + \hat{\alpha}_i + \hat{\beta}_j = \overline{Y_{..}} + \overline{Y_{i.}} - \overline{Y_{..}} + \overline{Y_{.j}} - \overline{Y_{..}} = \overline{Y_{i.}} + \overline{Y_{.j}} - \overline{Y_{..}}$$

Example

Size of City (factor A)	Region (factor B) East ($j = 1$)	West ($j = 2$)	Average
Small ($i = 1$)	140	100	120
Medium ($i = 2$)	210	180	195
Large ($i = 3$)	220	200	210
Average	190	160	175

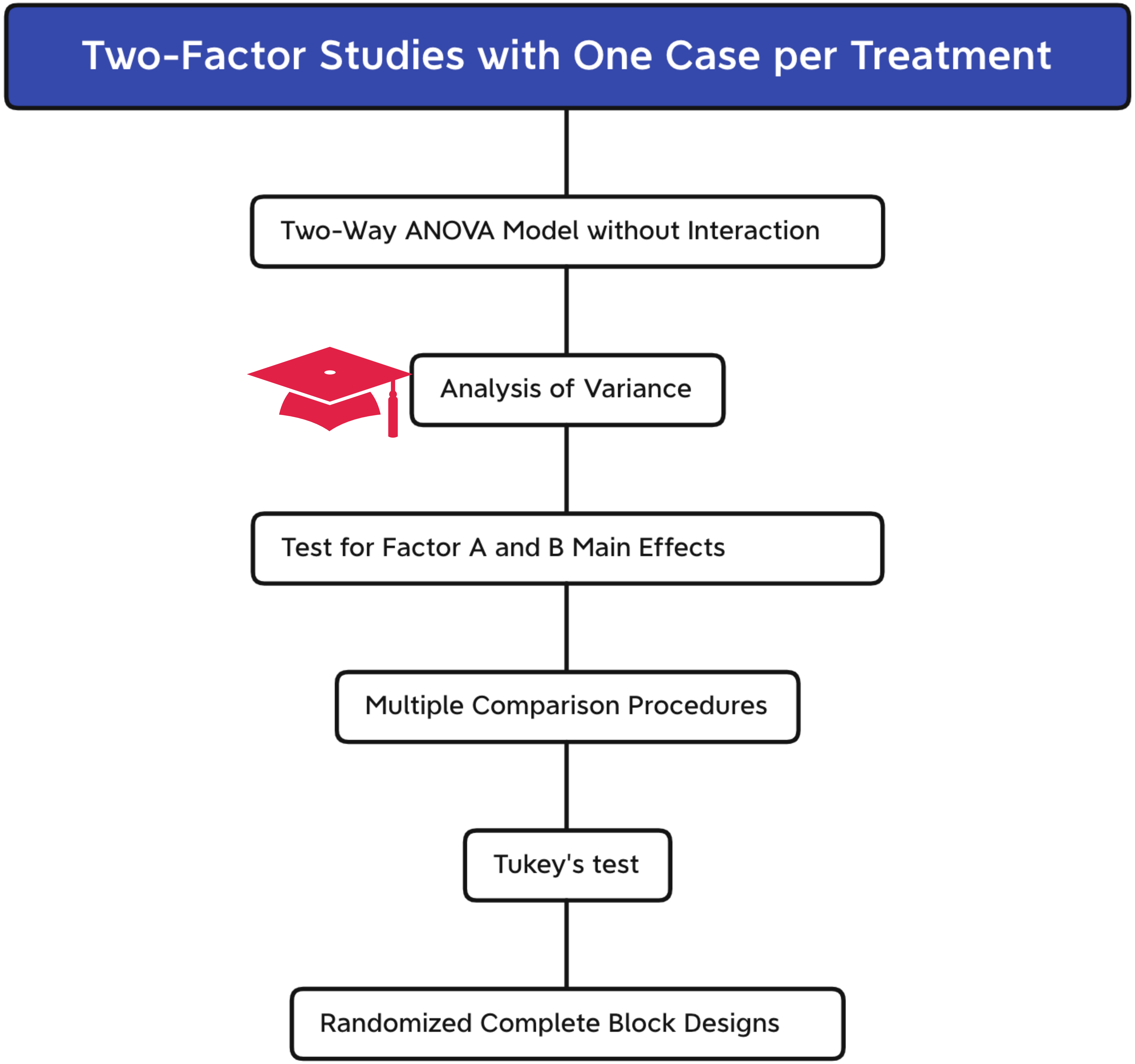
Plot the data. Does it appear that interaction effects are present? Does it appear that factor A and factor B main effects are present? Discuss.



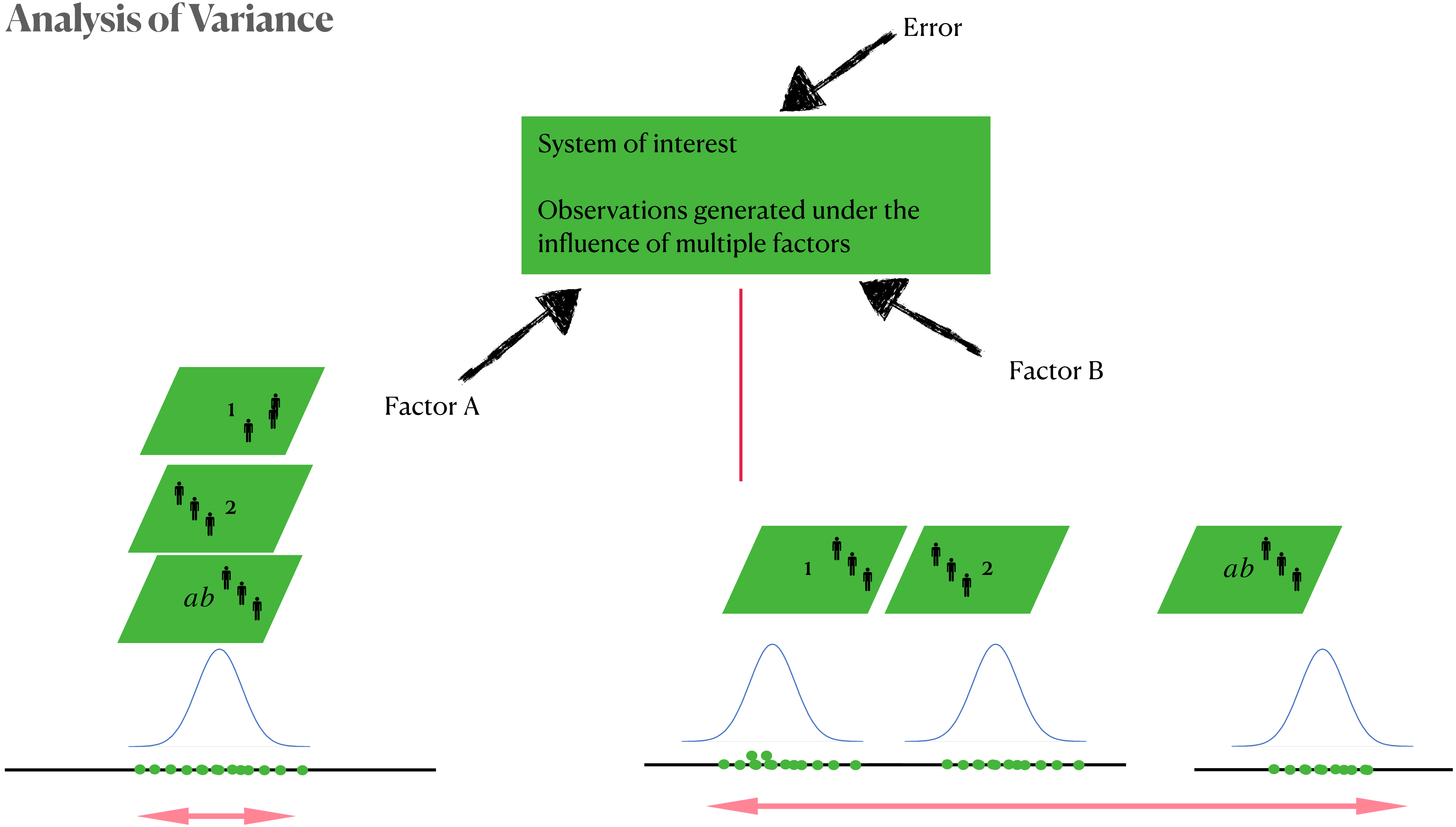
It appears that there could be a slight interaction between region and size of city in their effects on the premium. However, the lack of parallelism in the response lines could simply be the result of randomness.

It appears that factor A size of city do have effects, as we can see the premium increases as size increases.

It appears that factor B region also have effects, with east has higher premium than the west.


Two-Factor Studies with One Case per Treatment
Two-Way ANOVA Model without Interaction
Analysis of Variance
Test for Factor A and B Main Effects
Multiple Comparison Procedures
Tukey's test
Randomized Complete Block Designs

Analysis of Variance



Without factors A and B, the observations have some natural variation due to other extraneous factors, i.e. “error variance”

If some combinations of factor A and B indeed has some effects on the system, then we would expect more volatility .

Analysis of Variance

$$\underbrace{Y_{ij} - \bar{Y}_{..}}_{\text{Total Deviation}} = \underbrace{\left(Y_{ij} - \left(\bar{Y}_{i.} + \bar{Y}_{.j} - \bar{Y}_{..}\right)\right)}_{\text{Deviation due to extraneous factors}} + \underbrace{(\bar{Y}_{i.} - \bar{Y}_{..})}_{\text{A main effect}} + \underbrace{\left(\bar{Y}_{.j} - \bar{Y}_{..}\right)}_{\text{B main effect}}$$

$$\underbrace{\sum_i \sum_j \left(Y_{ij} - \bar{Y}_{..}\right)^2}_{\text{SSTO}} = \underbrace{\sum_i \sum_j \left(Y_{ij} - \left(\bar{Y}_{i.} + \bar{Y}_{.j} - \bar{Y}_{..}\right)\right)^2}_{\text{SSE}} + \underbrace{b \sum_i (\bar{Y}_{i.} - \bar{Y}_{..})^2}_{\text{SSA: factor A sum of squares}} + \underbrace{a \sum_j \left(\bar{Y}_{.j} - \bar{Y}_{..}\right)^2}_{\text{SSB: factor B sum of squares}}$$

$$\longrightarrow \quad \boxed{\text{SSTO} \quad = \quad \text{SSA + SSB +SSE}}$$

The partition of sum of squares is essentially the same as two-way ANOVA model with interaction, but let the $SSE = 0, SSAB = SSE, n = 1$

Analysis of Variance

$$\underbrace{\sum_i \sum_j \left(\bar{Y}_{ij} - \bar{Y}_{..}\right)^2}_{\text{SSTO}} = \underbrace{\sum_i \sum_j \left(Y_{ij} - \left(\bar{Y}_{i.} + \bar{Y}_{.j} - \bar{Y}_{..}\right)\right)^2}_{\text{SSE}} + \underbrace{b \sum_i \left(\bar{Y}_{i.} - \bar{Y}_{..}\right)^2}_{\text{SSA: factor A sum of squares}} + \underbrace{a \sum_j \left(\bar{Y}_{.j} - \bar{Y}_{..}\right)^2}_{\text{SSB: factor B sum of squares}}$$

$df(SSTO) = ab - 1$ $\quad df(SSE) = ab - (a + b - 1) = (a-1)(b-1)$ $\quad df(SSA) = a - 1$ $\quad df(SSB) = b - 1$

$$MSE = \frac{SSE}{(a-1)(b-1)} \qquad MSA = \frac{SSA}{a-1} \qquad MSB = \frac{SSB}{b-1}$$

$$E[MSE] = \sigma^2$$

$$E[MSA] = \sigma^2 + b\frac{\sum_i \alpha_i^2}{a-1} = \sigma^2 + b\frac{\sum_i \left(\mu_{i.} - \mu_{..}\right)^2}{a-1}$$

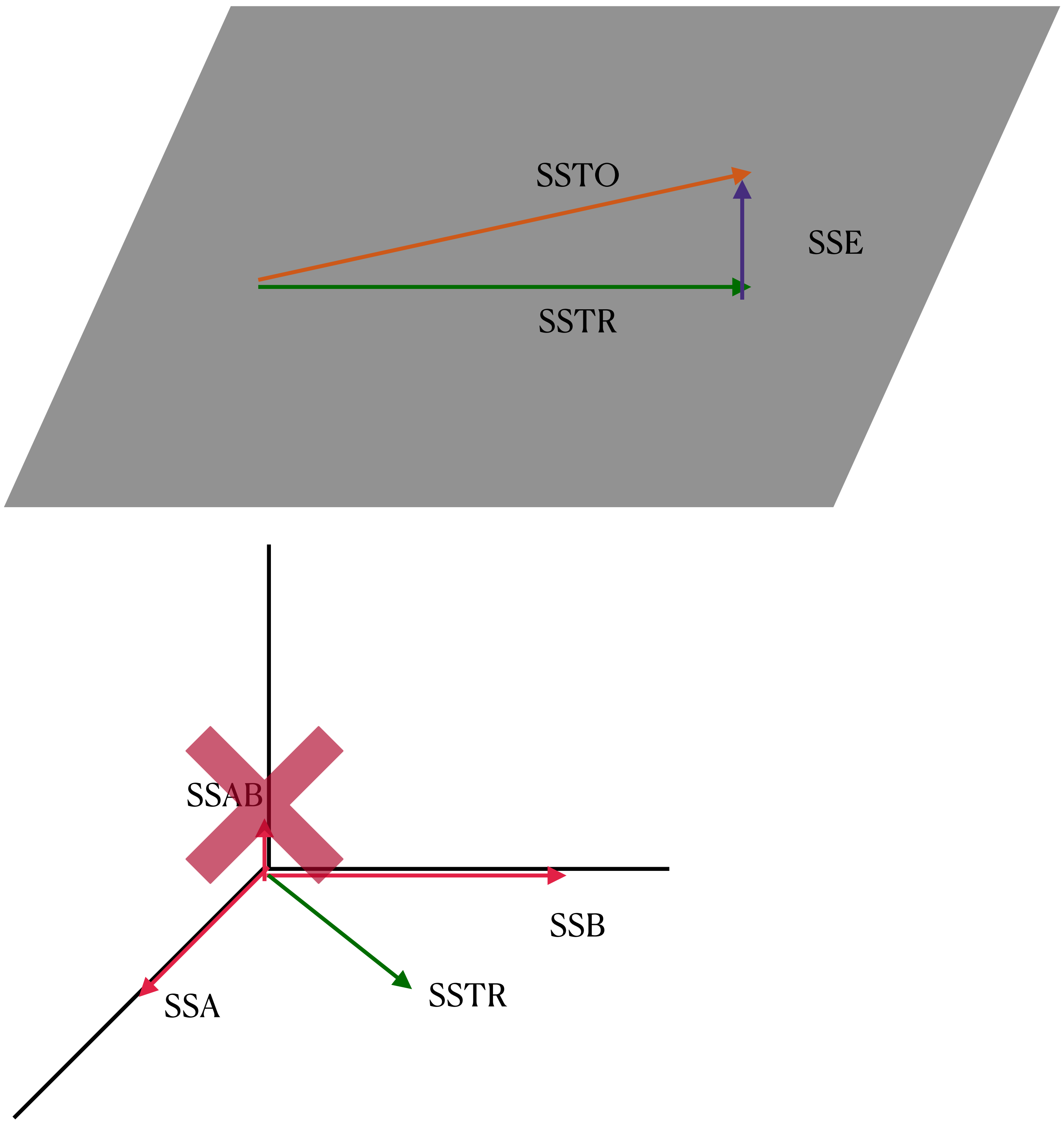
$$E[MSB] = \sigma^2 + a\frac{\sum_j \beta_j^2}{b-1} = \sigma^2 + a\frac{\sum_j \left(\mu_{.j} - \mu_{..}\right)^2}{b-1}$$

Analysis of Variance

TABLE 20.1 ANOVA Table for No-Interaction Two-Factor Model (20.1) with Fixed Factor Levels, $n = 1$.

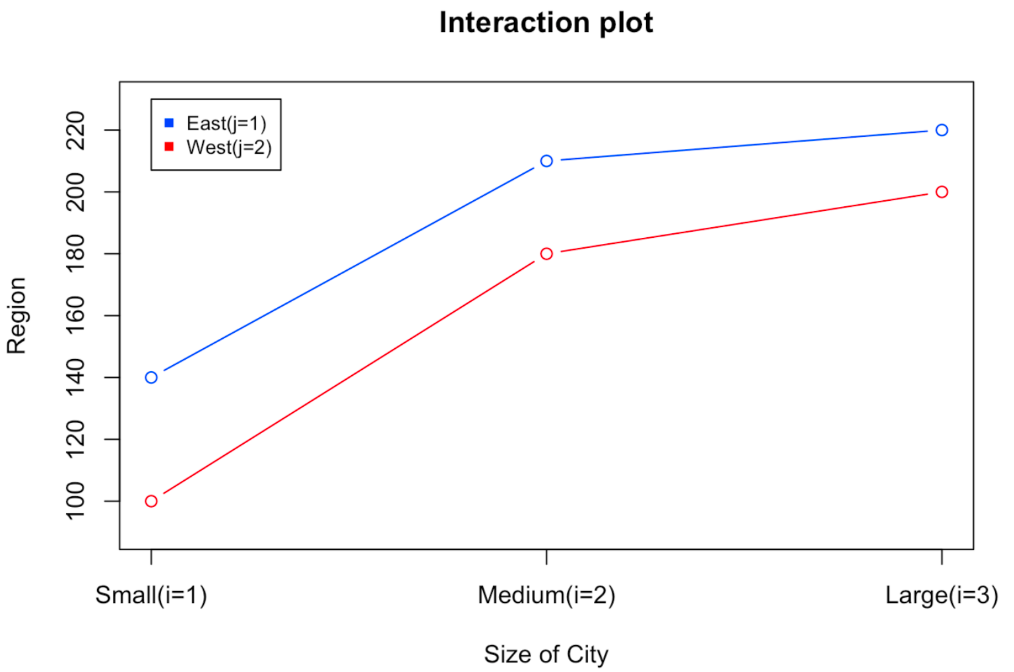
Source of Variation	SS	df	MS	E{MS}
Factor A	$SSA = b\sum(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2$	$a-1$	$MSA = \frac{SSA}{a-1}$	$\sigma^2 + b\frac{\sum(\mu_{i\cdot} - \mu_{\cdot\cdot})^2}{a-1}$
Factor B	$SSB = a\sum(\bar{Y}_{\cdot j} - \bar{Y}_{\cdot\cdot})^2$	$b-1$	$MSB = \frac{SSB}{b-1}$	$\sigma^2 + a\frac{\sum(\mu_{\cdot j} - \mu_{\cdot\cdot})^2}{b-1}$
Error	$SSAB = \sum\sum(Y_{ij} - \bar{Y}_{i\cdot} - \bar{Y}_{\cdot j} + \bar{Y}_{\cdot\cdot})^2$	$(a-1)(b-1)$	$MSAB = \frac{SSAB}{(a-1)(b-1)}$	σ^2
Total	$SSTO = \sum\sum(Y_{ij} - \bar{Y}_{\cdot\cdot})^2$	$ab-1$		

Geometry of Decomposition of Variance:



Example

	Region (factor B)		
Size of City (factor A)	East ($j = 1$)	West ($j = 2$)	Average
Small ($i = 1$)	140	100	120
Medium ($i = 2$)	210	180	195
Large ($i = 3$)	220	200	210
Average	190	160	175



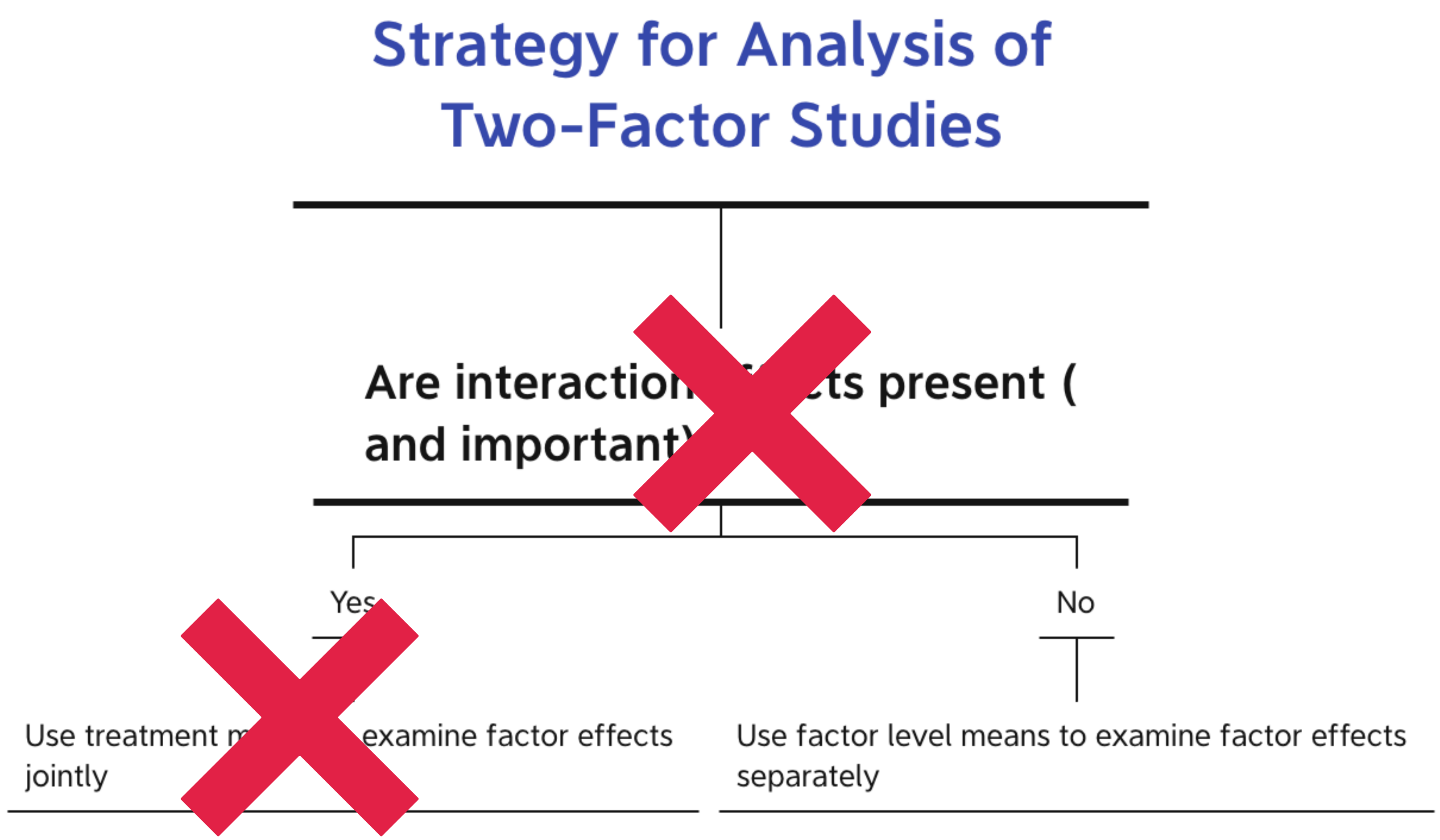
ANOVA Table

	SS	df	MS
factor A	9300	2	4650
factor B	1350	1	1350
Error	100	2	50
Totoal	10750	5	•

Two-Factor Studies with One Case per Treatment

- Two-Way ANOVA Model without Interaction
- Analysis of Variance
- Test for Factor A and B Main Effects
- Multiple Comparison Procedures
- Tukey's test
- Randomized Complete Block Designs

Strategy of Analysis



Inference for two-factor studies with one-case per treatment is the same as general two-factor studies, except that $df(MSE) = (a-1)(b-1)$

Test for Factor A and Factor B Main Effects

Compare E(MSA) vs E(MSE)

- If there are no factor A main effects
  - MSA has same expiation as MSE (they tend to be similar in numeric values)
- Otherwise
  - MSA tend to be larger than MSE

Compare E(MSB) vs E(MSE)

- If there are no factor B main effects
  - MSB has same expiation as MSE (they tend to be similar in numeric values)
- Otherwise
  - MSB tend to be larger than MSE

They suggest that ratios of Mean Squares provide evidence about the main effects, which will be the basis for F tests

Test for Factor A and Factor B Main Effects

To test whether or not factor A main effects are present:

$$H_0 : \alpha_1 = \ldots = \alpha_a = 0$$

$$H_a : \text{not all } \alpha_i = 0$$

Test statistic: $F^* = \dfrac{MSA}{MSE}$

Decision rule:

If $F^* \leq F_{1-\alpha}(a-1, (a-1)(b-1))$, then conclude H_0

If $F^* > F_{1-\alpha}(a-1, (a-1)(b-1))$, then conclude H_a

To test whether or not factor B main effects are present:

$$H_0 : \beta_1 = \ldots = \beta_b = 0$$

$$H_a : \text{not all } \beta_j = 0$$

Test statistic: $F^* = \dfrac{MSB}{MSE}$

Decision rule:

If $F^* \leq F_{1-\alpha}(b-1, (a-1)(b-1))$, then conclude H_0

If $F^* > F_{1-\alpha}(b-1, (a-1)(b-1))$, then conclude H_a

Example

Conduct separate tests for size and region main effects. In each test, use level of significance $\alpha = .05$ and state the alternatives, decision rule, and conclusion.

To test the significance of Factor A main effect

$$H_0 : \alpha_i = 0, i = 1,2,3 \text{ vs } H_a : \text{ not all } \alpha_i \text{'s are } 0$$

Test statistic: $F^* = \frac{MSA}{MSE} = \frac{4650}{50} = 93$

Critical value $F(0.95,2,2) = 19$

Since the F test statistics of factor A is larger than the critical value, we reject the null hypothesis at 0.05 significance level and conclude that city size main effects are present.

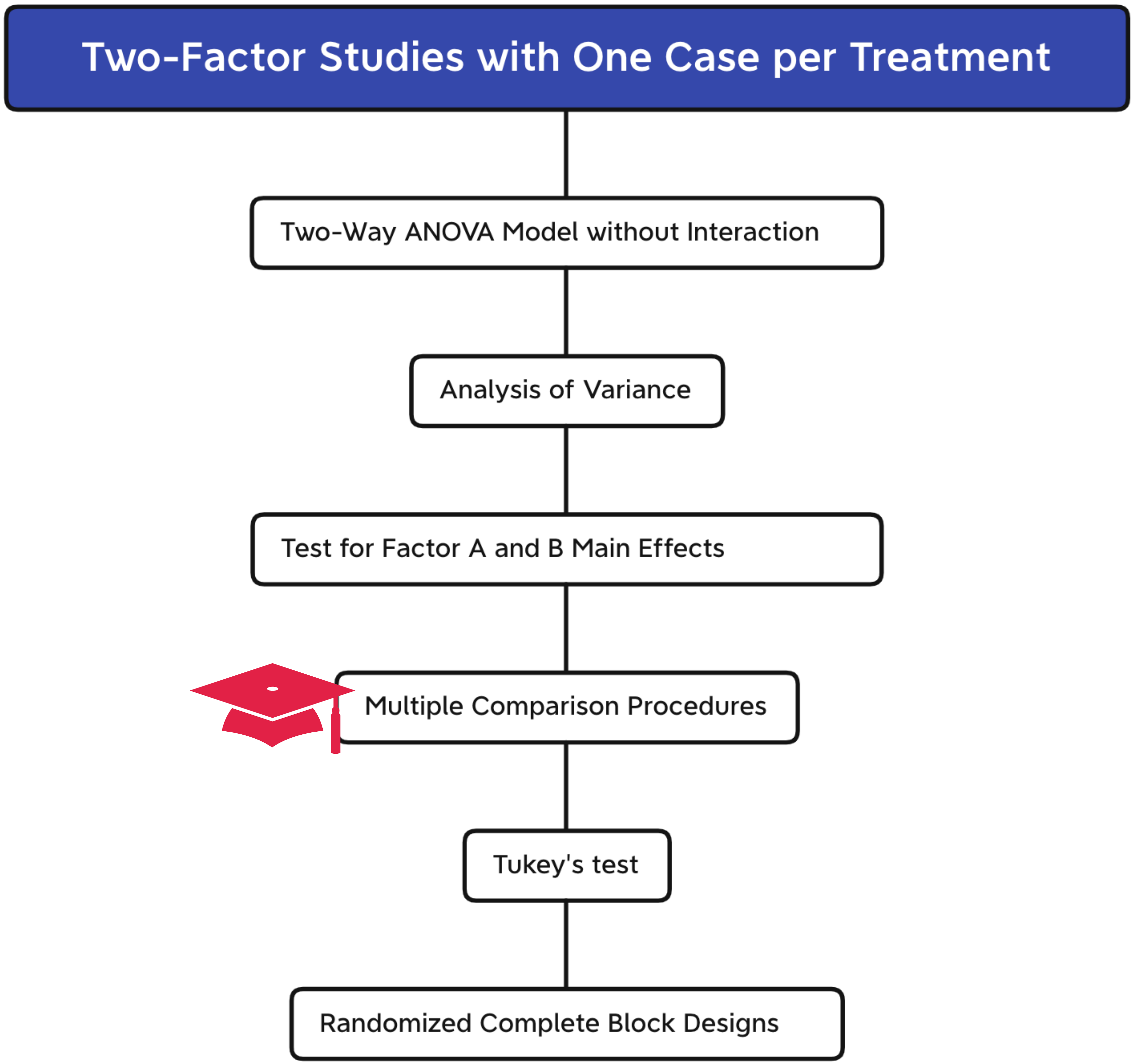
To test the significance of Factor B main effect

$$H_0 : \beta_j = 0, j = 1,2 \text{ vs } H_a : \text{ not all } \beta_j' \text{s are } 0$$

Test statistic: $F^* = \frac{MSB}{MSE} = \frac{1350}{50} = 27$

Critical value $F(0.95,1,2) = 18.5$

Since the F test statistic of factor B is larger than the critical value, we reject the null hypothesis at 0.05 significance level and conclude that geographic region main effects are present.


Two-Factor Studies with One Case per Treatment
Two-Way ANOVA Model without Interaction
Analysis of Variance
Test for Factor A and B Main Effects
Multiple Comparison Procedures
Tukey's test
Randomized Complete Block Designs

Analysis of Factor A (and B) Main Effects (When Factors Do Not Interact)

Multiple Comparison Procedure: Bonferroni

Suppose we're interested in making infernece about multiple quantities,
that are linear combinations of factor A level means (or factor B level means),
i.e., a family containing g linear combinations of factor level means

$$\mathscr{L} = \{L_1 = \sum_{i=1}^{r} c_{1i}\mu_{i\cdot}, \ldots, L_g = \sum_{i=1}^{r} c_{gi}\mu_{i\cdot}\}$$

$$\hat{L} = \sum_i c_i \bar{Y}_{i\cdot} \qquad s^2(\hat{L}) = \frac{MSE}{b} \sum_i c_i^2$$

Bonferroni's idea:
One very easy and conservative way to control family-wise error rate at α is to control individual test's significance level at $\alpha_0 = \dfrac{\alpha}{g}$

This procedure includes any inference about a single quantity as special case, just take g=1.

Analysis of Factor A (and B) Main Effects (When Factors Do Not Interact)

Multiple Comparison Procedure: Bonferroni

$(1-\alpha)100\,\%$ confidence interval for individual quantity in this family:

$$\hat{L}_i \pm Bs\left(\hat{L}_i\right) \text{ for } i = 1\ldots g$$

$$B = t\left(1 - \frac{\alpha}{2g}; (a-1)(b-1)\right)$$

Guarantee:
family-wise confidence coefficient is at least $(1-\alpha)100\,\%$

Meaning:
in at least $(1-\alpha)100\,\%$ of repetition of experiments, all the intervals in the family cover the true corresponding $L_i's$ $\alpha\,\%$ of repetition of experiments, some tests in the family made false discovery when the null hypothesis was true.

Hypothesis testing (t-test) for individual quantity in this family:

$$H_0^i : L_i = 0 \;\; H_a^i : L_i \neq 0$$

$$t^* = \frac{\hat{L}_i}{s(\hat{L}_i)} \sim t_{n_T - r} \text{if } H_0 \text{ is true}$$

If $|t^*| \leq B$, conclude H_0

If $|t^*| > B$, conclude H_a

Guarantee:
family-wise Type I error is at most α

Meaning:
in at most $\alpha\,\%$ of repetition of experiments, some tests in the family made false discovery when the null hypothesis was true.

Analysis of Factor A (and B) Main Effects (When Factors Do Not Interact)

Multiple Comparison Procedure: Sheffe

Suppose we're interested in making inference about <u>all possible contracts of factor A level means</u>

i.e., a family containing all possible contracts of factor A level means

$$\mathscr{L} = \{L = \sum_{i=1}^{r} c_i \mu_{i\cdot} \text{ where } \sum_{i=1}^{r} c_i = 0\}$$

Infinitely many claims or quantities

$$\hat{L} = \sum_i c_i \bar{Y}_{i\cdot\cdot} \qquad s^2(\hat{L}) = \frac{MSE}{b} \sum_i c_i^2$$

Analysis of Factor A (and B) Main Effects (When Factors Do Not Interact)

Multiple Comparison Procedure: Sheffe

$(1-\alpha)100\,\%$ confidence interval for individual quantity in this family:

$$\hat{L}_i \pm Ss\left(\hat{L}_i\right)$$

$$S = \sqrt{(a-1)F(1-\alpha;a-1,(a-1)(b-1))}$$

Guarantee:
family-wise confidence coefficient is at least $(1-\alpha)100\,\%$

Meaning:
in at least $(1-\alpha)100\,\%$ of repetition of experiments, all the intervals in the family cover the true corresponding $L_i's$ $\alpha\ \%$ of repetition of experiments, some tests in the family made false discovery when the null hypothesis was true.

Hypothesis testing (t-test) for individual quantity in this family:

$$H_0^i : L_i = 0 \ H_a^i : L_i \neq 0$$

$$t^* = \frac{\hat{L}_i}{s(\hat{L}_i)}$$

If $|t^*| \leq S$, conclude H_0

If $|t^*| > S$, conclude H_a

Guarantee:
family-wise Type I error is at most α

Meaning:
in at most $\alpha\ \%$ of repetition of experiments, some tests in the family made false discovery when the null hypothesis was true.

Analysis of Factor A (and B) Main Effects (When Factors Do Not Interact)

Multiple Comparison Procedure: Tukey

Suppose we're interested in making inference about all pairwise comparisons of factor level means
i.e., a family containing <u>all pairwise comparisons of factor level means</u>

$$\mathscr{L} = \{D_{ii'} = \mu_i - \mu_{i'} \text{ for } i \neq i'\}$$

$\frac{a(a-1)}{2}$ Pairwise comparisons

$$\hat{D}_{ii'} = \bar{Y}_{i\cdot\cdot} - \bar{Y}_{i'\cdot\cdot} \qquad s^2(\hat{D}_{ii'}) = MSE\frac{2}{b}$$

Analysis of Factor A (and B) Main Effects (When Factors Do Not Interact)

Multiple Comparison Procedure: Tukey

$(1-\alpha)100\ \%$ confidence interval for individual quantity in this family:

$$\hat{D}_{ii'} \pm Ts\left(\hat{D}_{ii'}\right)$$

$$T = \frac{1}{\sqrt{2}} q(1-\alpha; a, (a-1)(b-1))$$

Guarantee:
family-wise confidence coefficient is at least $(1-\alpha)100\ \%$

Meaning:
in at least $(1-\alpha)100\ \%$ of repetition of experiments, all the intervals in the family cover the true corresponding $L_i's$ $\alpha\ \%$ of repetition of experiments, some tests in the family made false discovery when the null hypothesis was true.

Hypothesis testing (t-test) for individual quantity in this family:

$$H_0^i : D_{ii'} = 0 \ H_a^i : D_{ii'} = \neq 0$$

$$q^* = \frac{\hat{D}_{ii'}}{s(\hat{D}_{ii'})}$$

If $|q^*| \leq T$, conclude H_0

If $|q^*| > T$, conclude H_a

Guarantee:
family-wise Type I error is at most α

Meaning:
in at most $\alpha\ \%$ of repetition of experiments, some tests in the family made false discovery when the null hypothesis was true.

Example

Make all pairwise comparisons for different sizes of city and regions; use the Bonferroni procedure with a 90 percent family confidence coefficient. State your findings.

There are 3 pairwise comparison for factor A and 1 pairwise comparisons for factor B, 4 in total.

$$B = t(1 - \alpha/8, (a-1)(b-1))$$

```
## [1] 6.205347
```

For factor A, $\hat{D}_{12} = \bar{Y}_{1.} - \bar{Y}_{2.}$

$$\bar{Y}_{1.} - \bar{Y}_{2.} \pm B\sqrt{\frac{2MSE}{b}}$$

Code

```
## [1] -76.97457 313.02543
```

$\hat{D}_{13} = \bar{Y}_{1.} - \bar{Y}_{3.}$

$$\bar{Y}_{1.} - \bar{Y}_{3.} \pm B\sqrt{\frac{2MSE}{a}}$$

Code

```
## [1] -91.97457 328.02543
```

$\hat{D}_{23} = \bar{Y}_{2.} - \bar{Y}_{3.}$

$$\bar{Y}_{2.} - \bar{Y}_{3.} \pm B\sqrt{\frac{2MSE}{b}}$$

Code

```
## [1] -16.97457 403.02543
```

For factor B, $\hat{D}_{12} = \bar{Y}_{.1} - \bar{Y}_{.2}$

$$\bar{Y}_{.1} - \bar{Y}_{.2} \pm B\sqrt{\frac{2MSE}{a}}$$

Code

```
## [1]  28.38777 348.38777
```

For this family of confidence intervals, the following conclusions may be drawn with family confidence coefficient of 90 percent:

- The average premium for different city sizes do not differ
- But the average premium for Eastern cities is higher than Western cities.

Two-Factor Studies with One Case per Treatment

Two-Way ANOVA Model without Interaction

Analysis of Variance

Test for Factor A and B Main Effects

Multiple Comparison Procedures

Tukey's test

Randomized Complete Block Designs

How do we know "No Interaction" assumption is correct or wrong?

Tukey Test for Additivity (Tukey one degree of freedom test)

Problem: can't allow arbitrary forms of interaction because of limited data

Tukey's idea: allow some restricted form of interaction

ij th interaction effect is proportional to the product of the main effects

$$\gamma_{ij} = D\alpha_i\beta_j$$

Motivation:

if in fact, the interaction effect γ_{ij} depends on main effects α_i, β_j in a relatively simple way, then the Turkey's assumption can be shown to be accurate.

How do we know "No Interaction" assumption is correct or wrong?

Two-way ANOVA model with Turkey's interaction:

$$Y_{ij} = \mu_{..} + \alpha_i + \beta_j + D\alpha_i\beta_j + \varepsilon_{ij}$$

Least squares estimates :

$$\hat{\mu}_{..} = \frac{\sum_i \sum_j \hat{\mu}_{ij}}{ab} = \frac{\sum_i \sum_j Y_{ij}}{ab} = \bar{Y}_{..}$$

$$\hat{\alpha}_i = \hat{\mu}_{i.} - \hat{\mu}_{..} = \frac{\sum_j Y_{ij}}{b} - \bar{Y}_{..} = \bar{Y}_{i.} - \bar{Y}_{..}$$

$$\hat{\beta}_j = \hat{\mu}_{.j} - \hat{\mu}_{..} = \frac{\sum_i Y_{ij}}{a} - \bar{Y}_{..} = \bar{Y}_{.j} - \bar{Y}_{..}$$

$$\hat{D} = \frac{\sum_i \sum_j \hat{\alpha}_i \hat{\beta}_j Y_{ij}}{\sum_i \hat{\alpha}_i^2 \sum_j \hat{\beta}_j^2} = \frac{\sum_i \sum_j (\bar{Y}_{i.} - \bar{Y}_{..})(\bar{Y}_{.j} - \bar{Y}_{..})Y_{ij}}{\sum_i (\bar{Y}_{i.} - \bar{Y}_{..})^2 \sum_j (\bar{Y}_{i.} - \bar{Y}_{..})^2}$$

How do we know "No Interaction" assumption is correct or wrong?

$$Y_{ij} - \bar{Y}_{..} = \left(Y_{ij} - \hat{Y}_{ij}\right) + \hat{\alpha}_i + \hat{\beta}_j + \hat{D}\hat{\alpha}_i\hat{\beta}_j$$

Total Deviation — Deviation due to extraneous factors — A main effect — B main effect — AB interaction effect

$$\sum_i \sum_j \left(\bar{Y}_{ij} - \bar{Y}_{..}\right)^2 = \sum_i \sum_j e_{ij}^2 + \sum_i \hat{\alpha}_i + \sum_j \hat{\beta}_j + \sum_i \sum_j \left(\hat{D}\hat{\alpha}_i\hat{\beta}_j\right)^2$$

SSTO — SSE — SSA — SSB — SSAB

$df(SSE) = ab - a - b$

$df(SSAB) = 1$

$$\longrightarrow \boxed{\text{SSTO} = \text{SSA} + \text{SSB} + \text{SSAB} + \text{SSE}}$$

How do we know "No Interaction" assumption is correct or wrong?

Tukey Test for Additivity (Tukey one degree of freedom test)

$H_0 : D = 0$ no interaction present $\quad$ $H_a : D \neq 0$ interaction is present

Test statistic: $F^* = \dfrac{MSAB}{MSE}$

Large value of F^* support H_a

Small value, when $F^* \approx 1$ support H_0

—> We reject H_0 for large value of F^*, i.e. $F^* \geq c$

$\longrightarrow$ Decision rule:

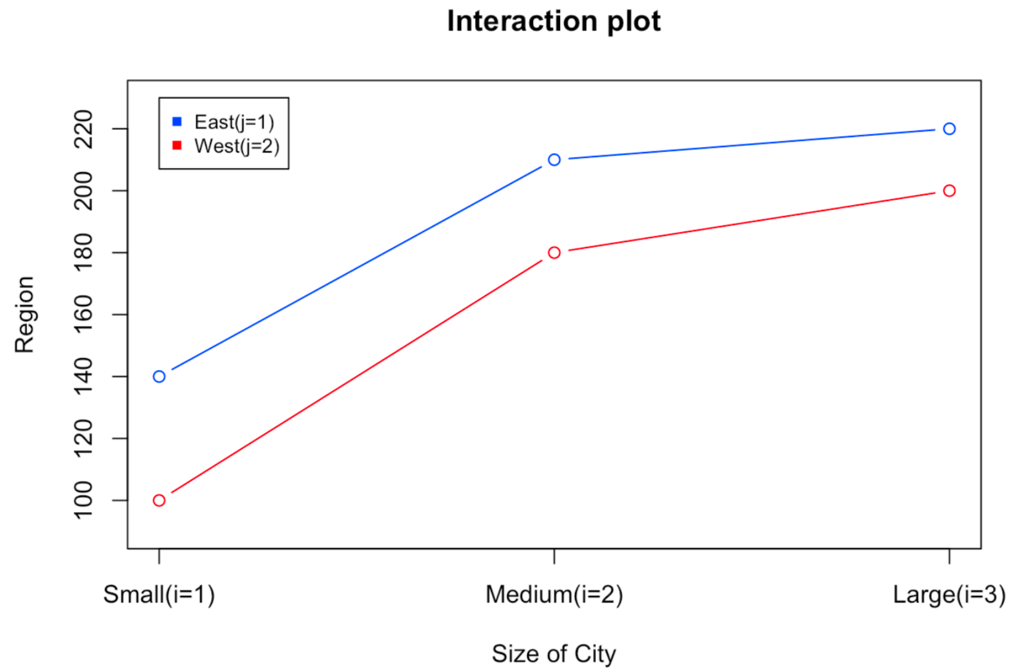
If $F^* \leq F_{1-\alpha}(1, ab - a - b)$, then conclude H_0

If $F^* > F_{1-\alpha}(1, ab - a - b)$, then conclude H_a

Example

Conduct the Tukey test for additivity; use $\alpha = .1$. State the alternatives, decision rule, and conclusion. If the additive model is not appropriate, what might you do?

$H_0 : D = 0$ no interaction present $\qquad H_a : D \neq 0$ interaction is present



Test statistic: $F^* = \dfrac{MSAB}{MSE}$

$$MSAB = \sum_i \sum_j \left(\hat{D}\hat{\alpha}_i\hat{\beta}_j\right)^2$$

$$MSE = \sum_i \sum_j \left(Y_{ij} - \hat{Y}_{ij}\right)^2$$

```
## [1] 6.75
```

Code

```
## [1] 39.86346
```

For $\alpha = .10$, we require $F(.90; 1, 1) = 39.9$. Since $F^* = 6.8 \leq 39.9$, we conclude that region and size of city do not interact. Use of the no-interaction model for the data therefore appears to be reasonable.

How do we know "No Interaction" assumption is correct or wrong?

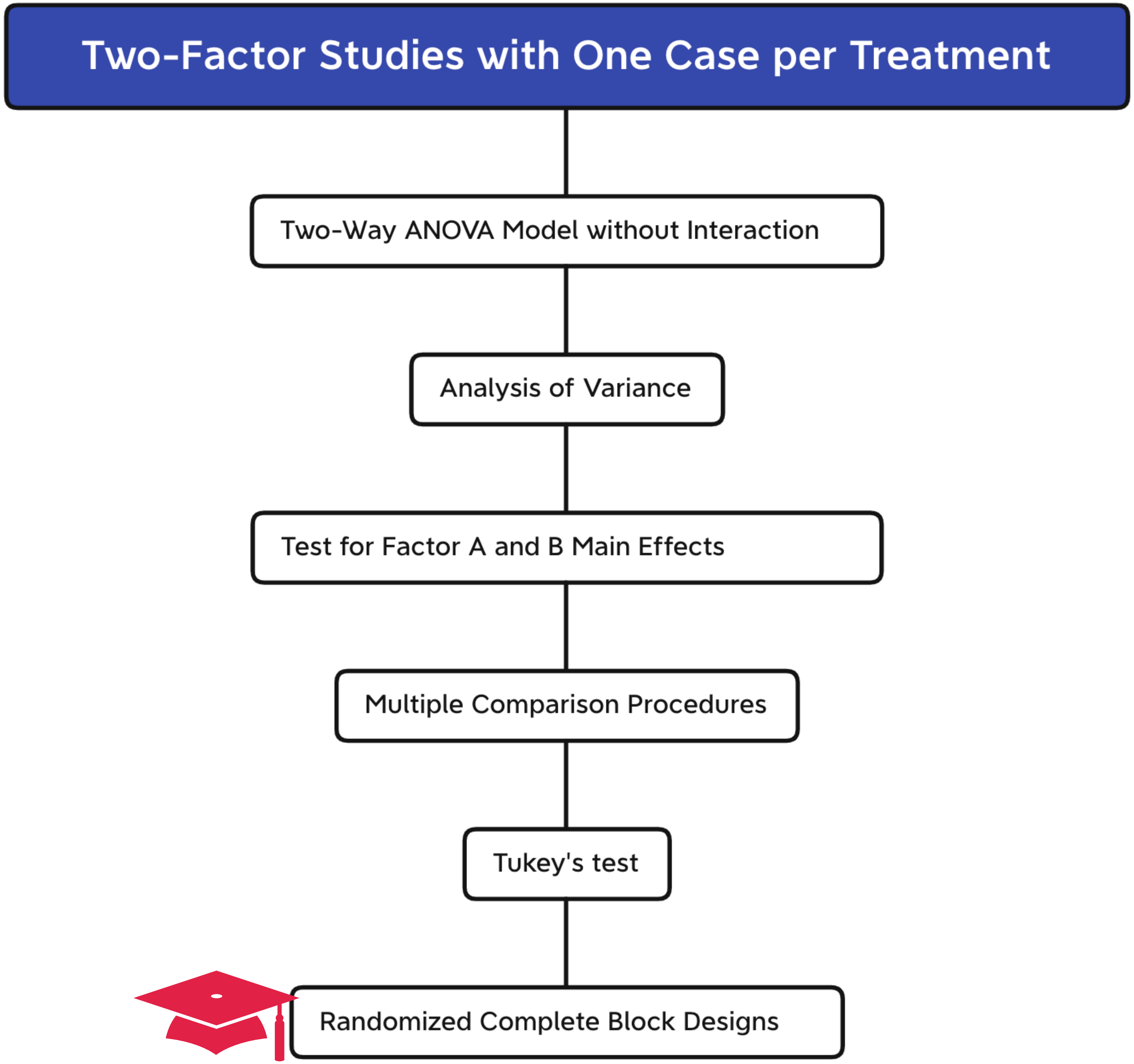
Tukey Test for Additivity (Tukey one degree of freedom test)

- Effective in detecting the interactions that are "simple" and approximately in the form $D\alpha_i\beta_j$
- Remedial actions are needed if interaction effects are present by Tukey's test

  Transformation of Y to remove interaction effects

  $\sqrt{Y}$, $logY$, Box-Cox transformation Y^{λ}

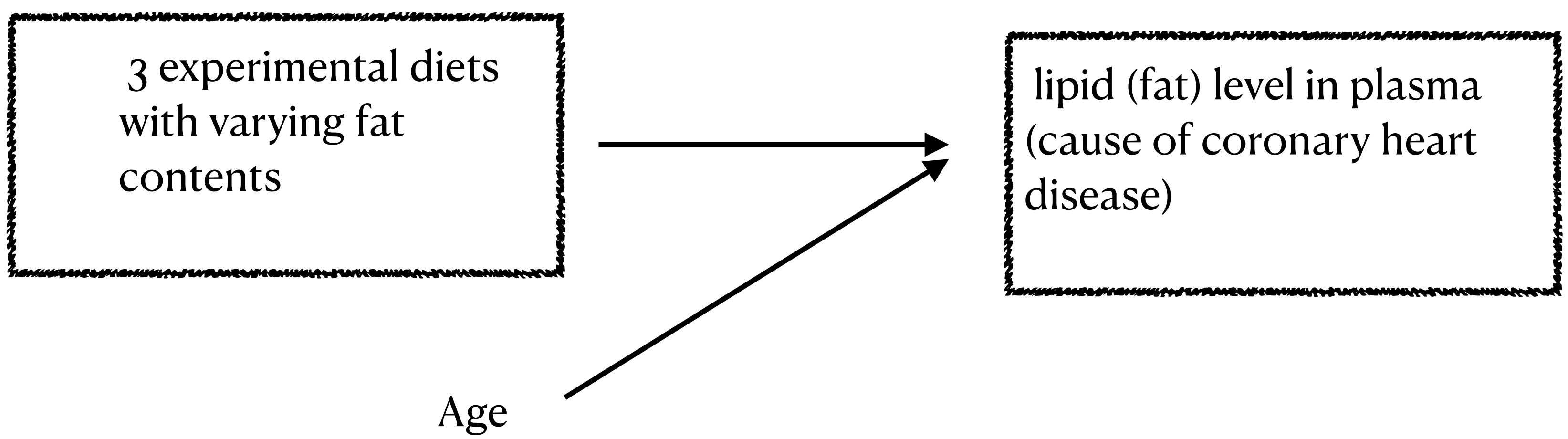
  If no such transformation can remove the interaction, then be cautious about the reliability of model result


Two-Factor Studies with One Case per Treatment
Two-Way ANOVA Model without Interaction
Analysis of Variance
Test for Factor A and B Main Effects
Multiple Comparison Procedures
Tukey's test
Randomized Complete Block Designs

Example

(The Fat-in-Diet Study)

The objective of the study:



The study setup:

Within each block, 3 experimental diets were randomly assigned to the 3 subjects

Reduction in lipid level after some a certain period of time were recorded as the outcome

Block i		Fat Content of Diet $j = 1$ Extremely Low	$j = 2$ Fairly Low	$j = 3$ Moderately Low
1	Ages 15–24	.73	.67	.15
2	Ages 25–34	.86	.75	.21
3	Ages 35–44	.94	.81	.26
4	Ages 45–54	1.40	1.32	.75
5	Ages 55–64	1.62	1.41	.78

☑ Randomized Complete Block Design

Randomized Complete Block Designs

Randomized Block Designs

is used primarily to reduce the variance of error terms, so that more precise inference of treatment effects can be made, compared to completely randomized design.

> units are divided into blocks defined by some nuisance factor(s) that affects the outcome ,
> separate randomization are conducted in each block,
> effect of experimental factor is obtained by combining the estimated effects from all blocks

When each treatment only has 1 case within a block —> **Randomized Complete Block Design**

☑ Two-factor study

  Experimental factor + block as observational factor

☑ What Models should be used?

When each treatment has multiple replicates within a block —> **Randomized Block Design**

☑ Two-factor study

  Experimental factor + block as observational factor

☑ What Models should be used?

" Why would anyone use a randomized complete block design that requires the assumption that block and treatment effects do not interact, when this assumption can be avoided and checked by randomized block design?"

Criteria for Blocking

Characteristics associated with the unit:

- If subjects are persons:
  - gender, age income, intelligence, education, job experience.....
- If subjects are geographic areas:
  - population size, average household income, average education level ......

Characteristics associated with the experimental setting:

- Observer
- Time of processing
- Machine
- Measuring instrument
- ......

Experience in the subject matter field

Two-Way ANOVA Model without Interaction for RCBD

RCBD may be viewed as a spacial case of the two-factor study with 1 case per treatment,
where blocks are factor A (observational factor) , and treatments are factor B (experimental factor).

Assume: no interaction effects between blocks and treatments,
that is, treatment effects do not differ across blocks

$$Y_{ij} = \mu_{..} + \rho_i + \tau_j + \varepsilon_{ij}$$

(Subscript k=1 dropped)

- $\mu_{..} = \dfrac{\sum_i \sum_j \mu_{ij}}{ab}$: overall mean
- $\rho_i = \mu_{i.} - \mu_{..}$ main effect of factor A at ith level
  Subject to $n_b - 1$ constraints $\sum \alpha_i = 0$
- $\tau_j = \mu_{.j} - \mu_{..}$ main effect of factor B at jth level
  Subject to $r - 1$ constraints $\sum \tau_j = 0$
- ε_{ij} are independent $N\left(0,\sigma^2\right)$ for $i = 1,\ldots,n_b; j = 1,\ldots,r$

Fitting the Two-Way ANOVA Model without interaction

Least squares estimates for parameters in factor effects parameterization:

$$\hat{\mu}_{..} = \frac{\sum_i \sum_j \hat{\mu}_{ij}}{n_b r} = \frac{\sum_i \sum_j Y_{ij}}{n_b r} = \bar{Y}_{..}$$

$$\hat{\rho}_i = \hat{\mu}_{i.} - \hat{\mu}_{..} = \frac{\sum_j Y_{ij}}{r} - \bar{Y}_{..} = \bar{Y}_{i.} - \bar{Y}_{..}$$

$$\hat{\tau}_j = \hat{\mu}_{.j} - \hat{\mu}_{..} = \frac{\sum_i Y_{ij}}{n_b} - \bar{Y}_{..} = \bar{Y}_{.j} - \bar{Y}_{..}$$

Where:

$$\bar{Y}_{..} = \frac{\sum_i \sum_j Y_{ij}}{n_b r}$$

$$\bar{Y}_{i.} = \frac{\sum_j Y_{ij}}{r}$$

$$\bar{Y}_{.j} = \frac{\sum_i Y_{ij}}{n_b}$$

- fitted value for an observation Y_{ij}

  ANOVA model's "best guess" or "best prediction"

  $$\hat{Y}_{ij} = \hat{\mu}_{..} + \hat{\rho}_i + \hat{\tau}_j = \overline{Y_{..}} + \overline{Y_{i.}} - \overline{Y_{..}} + \overline{Y_{.j}} - \overline{Y_{..}} = \overline{Y_{i.}} + \overline{Y_{.j}} - \overline{Y_{..}}$$

- residual e_{ij}

$$e_{ij} = Y_{ij} - \hat{Y}_{ij} = Y_{ij} - \left(\overline{Y_{i.}} + \overline{Y_{.j}} - \overline{Y_{..}}\right)$$

Analysis of Variance

$$\underbrace{Y_{ij} - \bar{Y}_{..}}_{\text{Total Deviation}} = \underbrace{\left(Y_{ij} - \left(\bar{Y}_{i.} + \bar{Y}_{.j} - \bar{Y}_{..}\right)\right)}_{\text{Deviation due to extraneous factors}} + \underbrace{\left(\bar{Y}_{i.} - \bar{Y}_{..}\right)}_{\text{Block main effect}} + \underbrace{\left(\bar{Y}_{.j} + \bar{Y}_{..}\right)}_{\text{Treatment main effect}}$$

$$\underbrace{\sum_{i}\sum_{j}\left(\bar{Y}_{ij} - \bar{Y}_{..}\right)^{2}}_{\text{SSTO}} = \underbrace{\sum_{i}\sum_{j}\left(Y_{ij} - \left(\bar{Y}_{i.} + \bar{Y}_{.j} - \bar{Y}_{..}\right)\right)^{2}}_{\text{SSE}} + \underbrace{r\sum_{i}\left(\bar{Y}_{i.} - \bar{Y}_{..}\right)^{2}}_{\text{SSBL: Block sum of squares}} + \underbrace{n_b\sum_{j}\left(\bar{Y}_{.j} - \bar{Y}_{..}\right)^{2}}_{\text{SSTR: Treatment sum of squares}}$$

$$\longrightarrow \boxed{\text{SSTO} \quad = \quad \text{SSBL} + \text{SSTR} + \text{SSE}}$$

The partition of sum of squares is exactly the same as two-way ANOVA model without interaction, just with different notation.

Analysis of Variance

$$\underbrace{\sum_i \sum_j \left(\bar{Y}_{ij} - \bar{Y}_{..}\right)^2}_{\text{SSTO}} = \sum_i \sum_j \underbrace{\left(Y_{ij} - \left(\bar{Y}_{i.} + \bar{Y}_{.j} - \bar{Y}_{..}\right)\right)^2}_{\text{SSE}} + r \underbrace{\sum_i \left(\bar{Y}_{i.} - \bar{Y}_{..}\right)^2}_{\text{SSBL: Block sum of squares}} + n_b \underbrace{\sum_j \left(\bar{Y}_{.j} - \bar{Y}_{..}\right)^2}_{\text{SSTR: Treatment sum of squares}}$$

$df(SSTO) = n_b r - 1$

$df(SSE) = n_b r - (n_b + r - 1) = (n_b - 1)(r - 1)$

$df(SSBL) = n_b - 1$

$df(SSTR) = r - 1$

$$E[MSE] = \sigma^2$$

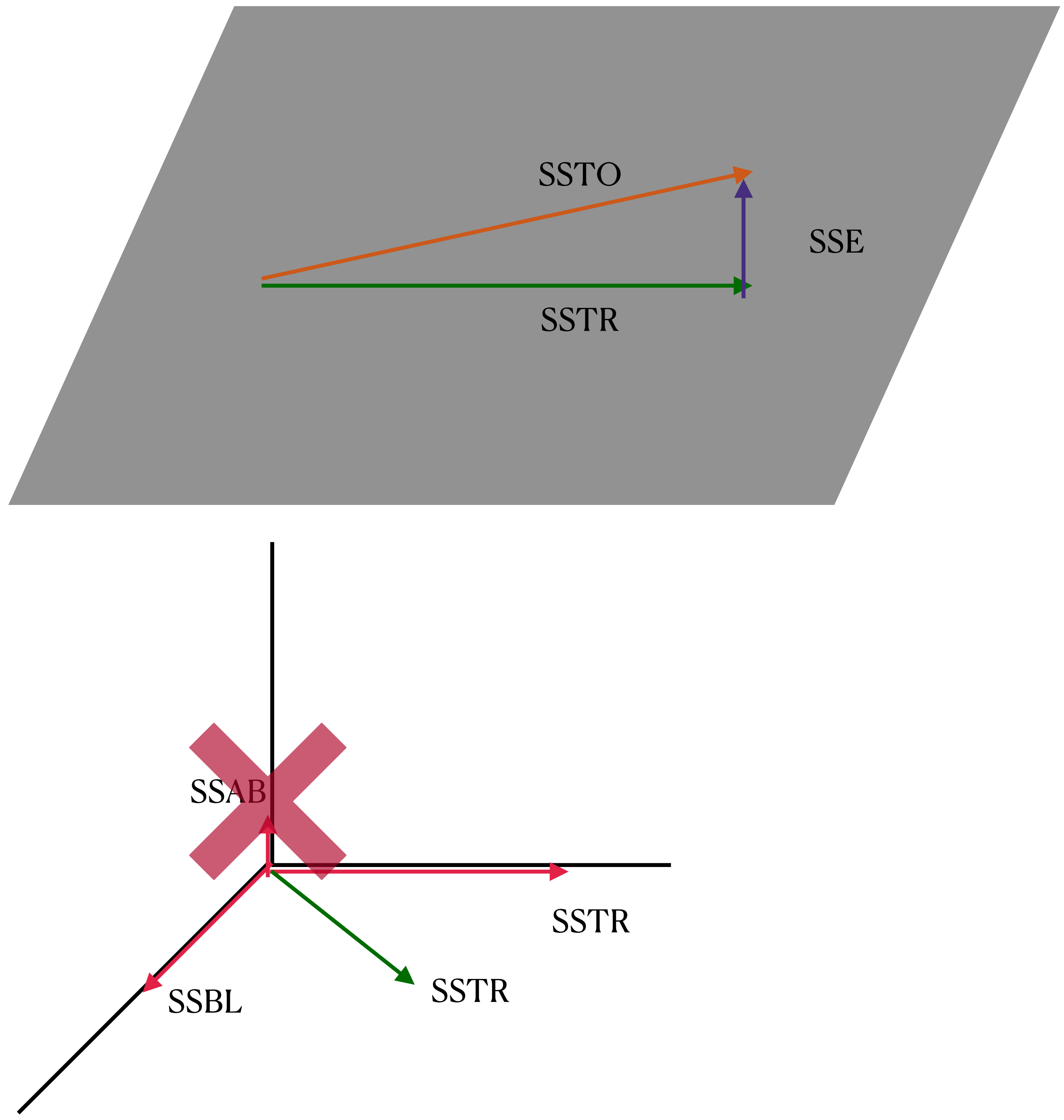
$$E[MSBL] = \sigma^2 + r\frac{\sum_i \rho_i^2}{n_b - 1} = \sigma^2 + r\frac{\sum_i \left(\mu_{i.} - \mu_{..}\right)^2}{n_b - 1}$$

$$E[MSTR] = \sigma^2 + n_b\frac{\sum_j \tau_j^2}{r - 1} = \sigma^2 + n_b\frac{\sum_j \left(\mu_{.j} - \mu_{..}\right)^2}{r - 1}$$

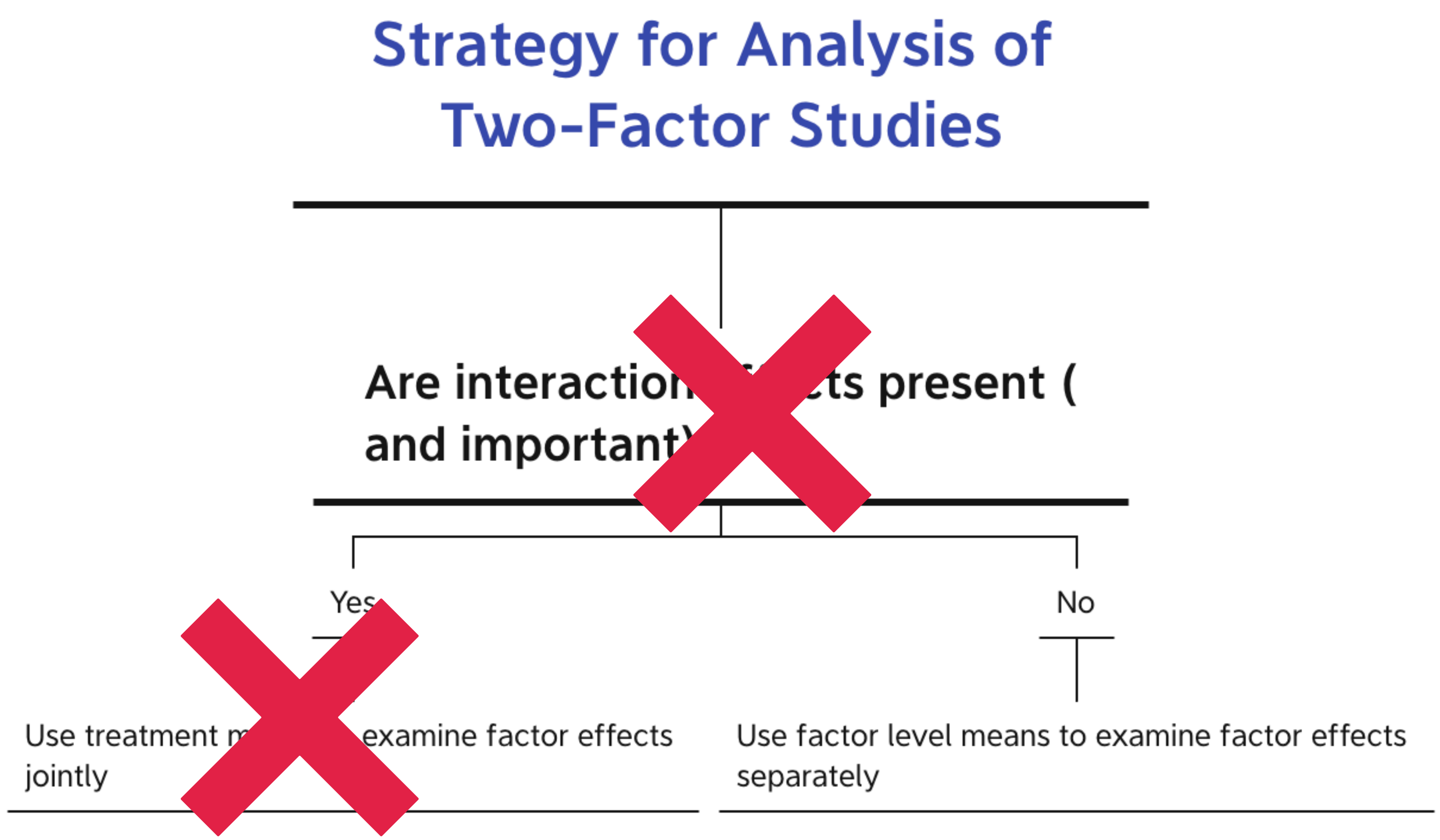
Analysis of Variance

Source of Variation	SS	df	MS	$E\{MS\}$
Blocks	$SSBL$	$n_b - 1$	$MSBL$	$\sigma^2 + r\frac{\sum \rho_i^2}{n_b - 1}$
Treatments	$SSTR$	$r - 1$	$MSTR$	$\sigma^2 + n_b\frac{\sum \tau_j^2}{r - 1}$
Error	$SSBL.TR$	$(n_b - 1)(r - 1)$	$MSBL.TR$	σ^2
Total	$SSTO$	$n_b r - 1$		

Geometry of Decomposition of Variance:



Strategy of Analysis



Inference for RCBD is the same as two-factor studies with one-case per treatment, except that $df(MSE) = (n_b - 1)(r - 1)$

Test for Treatment (Main) Effects

The primary purpose of including the blocking factor is to increase precision of inference and estimation of treatment effects, not to discover its relationship with the outcome.

Therefore, Investigations are not concerned with making any inference about block effects.

To test whether or not treatment main effects are present:

$$H_0 : \tau_1 = \ldots = \tau_r = 0$$

$$H_a : \text{not all } \tau_i = 0$$

Test statistic: $F^* = \dfrac{MSTR}{MSE}$

Decision rule:

If $F^* \leq F_{1-\alpha}(r-1,(n_b-1)(r-1))$, then conclude H_0

If $F^* > F_{1-\alpha}(r-1,(n_b-1)(r-1))$, then conclude H_a

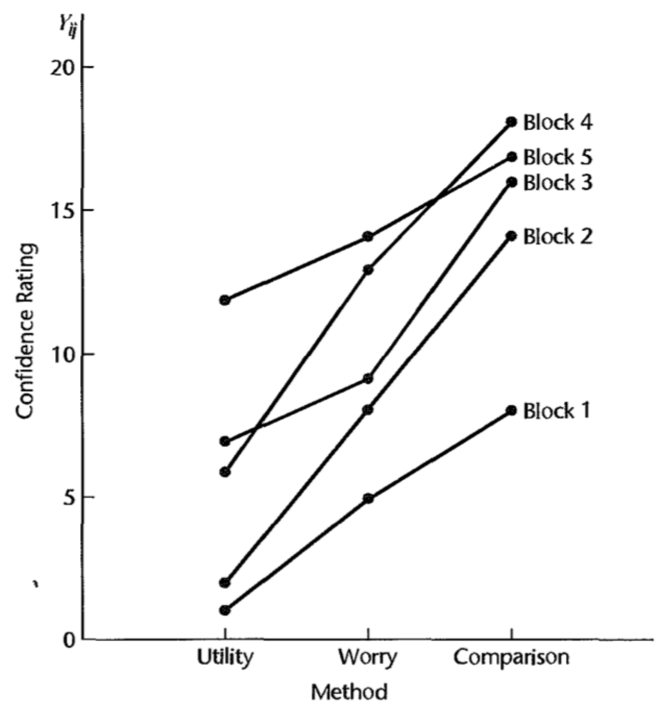
Test for Treatment (Main) Effects

Once the existence of treatment effects established by the F-test, interests often focus on multiple comparisons of the treatment means $\mu_{\cdot j}$s, where $\mu_{\cdot j}$ is the mean response for jth treatment averaged over all blocks.

The multiple comparison procedure is same as two-factor studies with one-case per treatment, except that $df(MSE) = (n_b - 1)(r - 1)$

Evaluation of Appropriateness of No Block-Treatment Interactions

Graphical way:



A severe lack of parallelism is a strong indication that blocks and treatments interact in their joint effect on the outcome, that is, when they affect the outcome simultaneously

Formal test: Tukey's test for additivity

Example

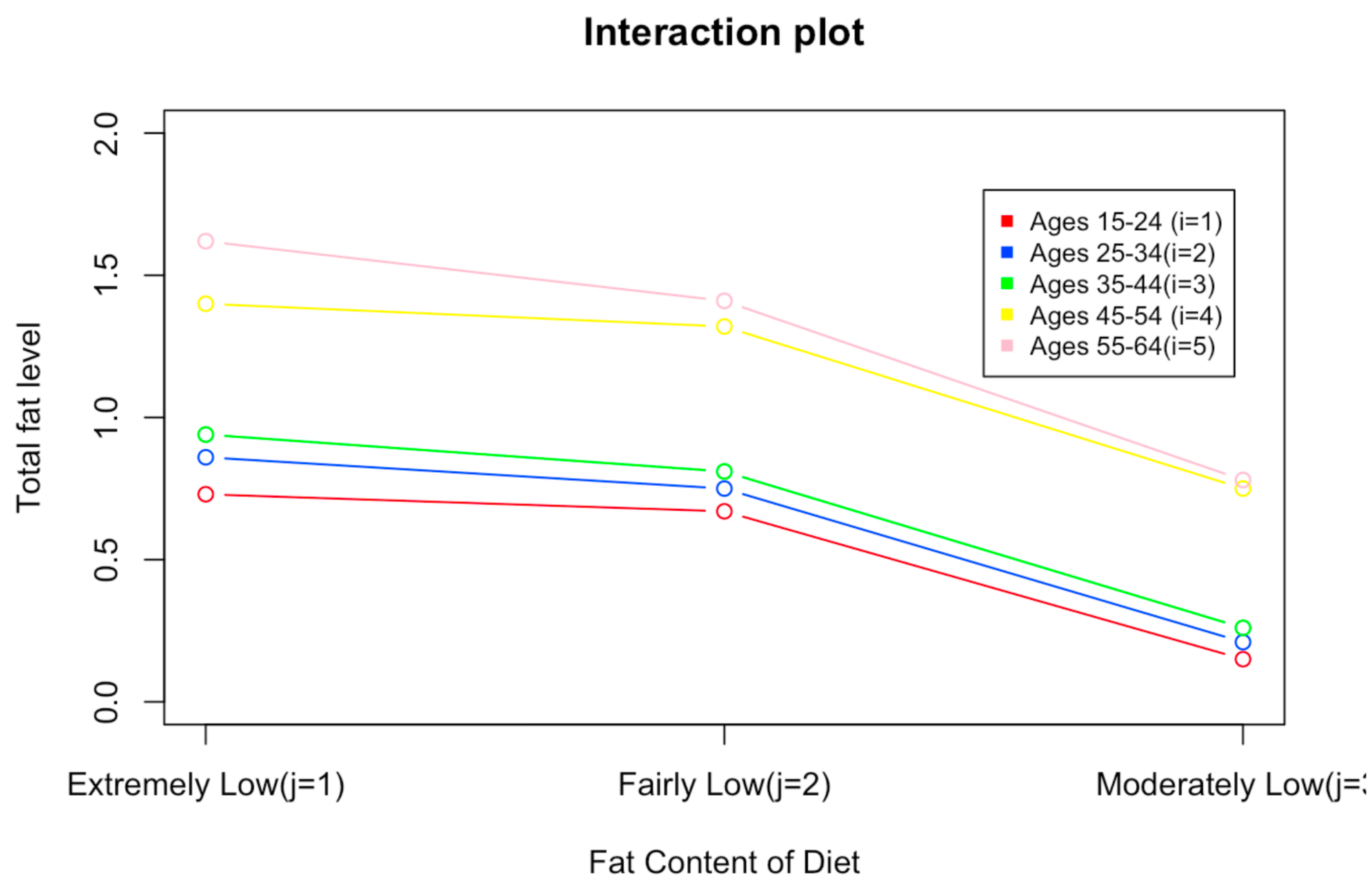
Block		Fat Content of Diet		
		$j = 1$	**$j = 2$**	**$j = 3$**
i		Ex tremely Low	Fairly Low	Moderately Low
1	Ages 15-24	.73	.67	.15
2	Ages 25-34	.86	.75	.21
3	Ages 35-44	.94	.81	.26
4	Ages 45-54	1.40	1.32	.75
5	Ages 55-64	1.62	1.41	.78

Why do you think that age of subject was used as a blocking variable?

age is predictive of or associated with lipid level, so blocking on age is likely to reduce the error term variability and thus increase the precision of treatment effect estimations.

Example

Plot the data. What does this plot suggest about the appropriateness of the no-interaction assumption here? Does it appear that factor A and factor B main effects are present? Discuss.



The no-interaction assumption appears to be appropriate.

The fat content in diet factor appears to have effects on total lipid level, with the extremely low fat diet tend to have higher total lipid level and moderately low fat diet tend to have lower total lipid level.

The blocking factor, Age, appears to have effects on total lipid level, but the difference appears only between people younger than 44 and people older than 44 .

Example

Conduct the Tukey test for additivity: use $\alpha = .01$.
State the alternatives, decision rule, and conclusion. If the additive model is not appropriate, what might you do?

Two-way ANOVA model with Turkey's interaction:

$$Y_{ij} = \mu_{..} + \rho_i + \tau_j + D\rho_i\tau_j + \varepsilon_{ij}$$

$H_0 : D = 0$ no interaction present $\qquad H_a : D \neq 0$ interaction is present

Test statistic: $F^* = \dfrac{MSAB}{MSE}$

$$MSAB = \sum_i \sum_j \left(\hat{D}\hat{\rho}_i\hat{\tau}_j\right)^2$$

$$MSE = \sum_i \sum_j \left(Y_{ij} - \hat{Y}_{ij}\right)^2$$

For $\alpha = .01$, we require $F(.99; 1,7) = 12.246$.

Since $F^* = 6.4 \leq 12.246$, we conclude that fat content in diet and age do not interact.

Use of the no-interaction model for the data therefore appears to be reasonable.

Example

Assume that randomized block model is appropriate. Obtain the analysis of variance table.

ANOVA Table

	SS	df	MS
~~factor A~~ Blocks	1.41896	4	0.35474
~~factor B~~ Treatments	1.32028	2	0.66014
Error	0.01932	8	0.002415
Totoal	2.75856	14	

Example

Test whether or not the mean reductions in lipid level differ for the three diets; use $\alpha = .05$.

$H_0 : \tau_1 = \tau_2 = \tau_3 = 0$ $\quad$ $H_a :$ not all β_j equal zero

Since $F^* = 1113.823 > 4.45897$, we reject the null and conclude H_a, that fat content in diet effects are present.

Example

Estimate $L_1 = \mu_{.1} - \mu_{.2}$ and $L_2 = \mu_{.2} - \mu_{.3}$using the Bonferroni procedure with a 95 percent family confidence coefficient. State your findings.

There are 2 pairwise comparison for factor B.

Bonferroni method: $B = t(1 - \alpha/4, 2, 8)$

Code

```
## [1] 2.306004
```

$\hat{L}_1 = \bar{Y}_{.1} - \bar{Y}_{.2}$

$$\bar{Y}_{.1} - \bar{Y}_{.2} \pm B\sqrt{\frac{2MSE}{a}}$$

Code

```
## [1] 0.06271481 2.04671481
```

$\hat{L}_2 = \bar{Y}_{.2} - \bar{Y}_{.3}$

$$\bar{Y}_{.2} - \bar{Y}_{.3} \pm B\sqrt{\frac{2MSE}{a}}$$

Code

```
## [1] 0.5067148 1.3667148
```

For this family of confidence intervals, the following conclusions may be drawn with family confidence coefficient of 90 percent:

- The average total lipid level for extremely low fat content in diet is higher than that for fairly low fat content in diet
- The average total lipid level for fairly low fat content in diet is higher than that for moderately low fat content in diet
- Therefore, moderately low fat content in diet group has the lowest total lipid level, whereas extremely low fat content in diet group has the highest total lipid level.

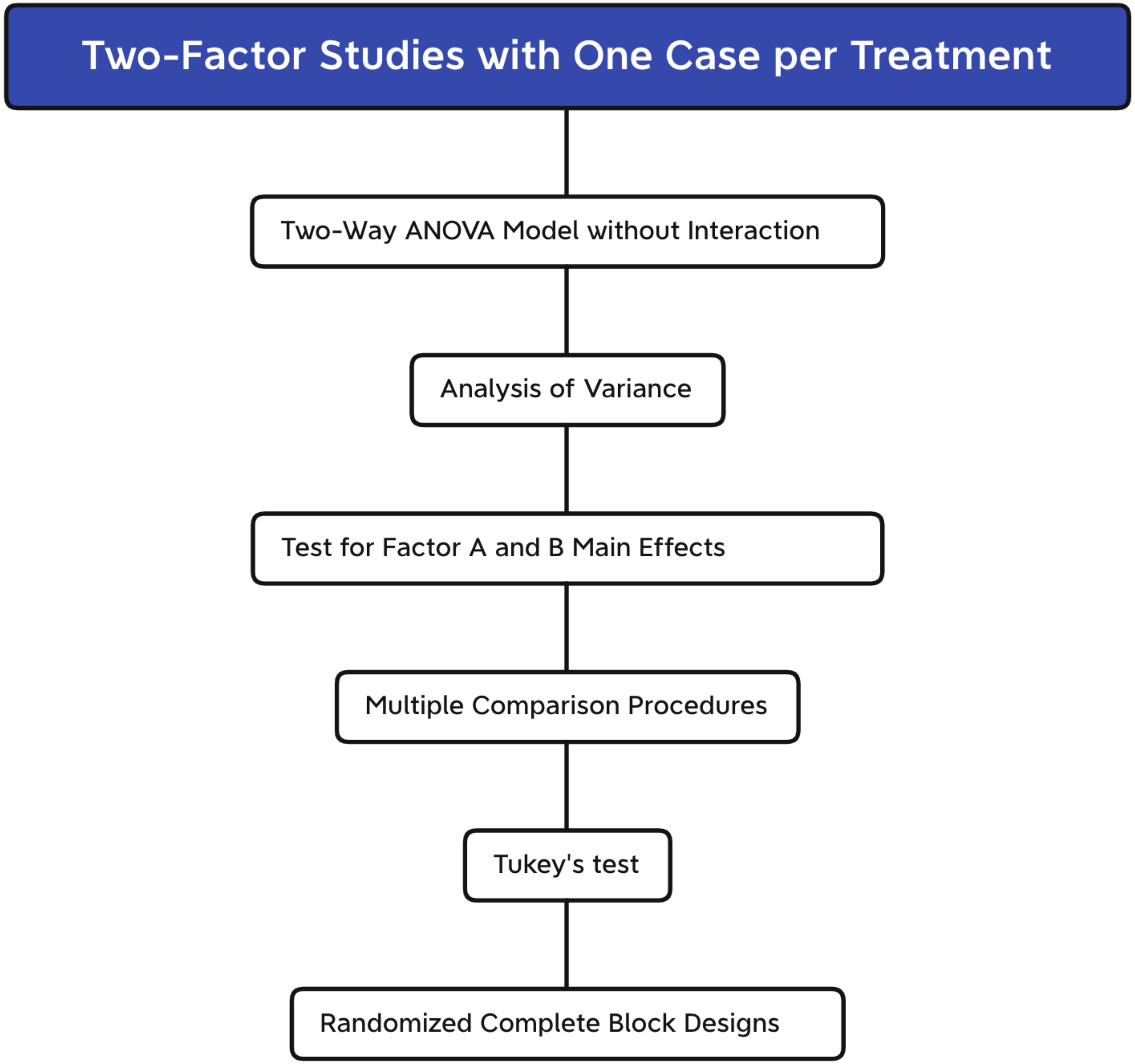
Example

Test whether or not blocking effects are present; use $\alpha = .05$.
(not really an interesting question to ask...)

$H_0 : \rho_1 = \ldots = \rho_5 = 0$ $\quad$ $H_a :$ not all $\rho_i' s$ equal zero

Since $F^* = 307.1072 > 3.837853$, we conclude H_a, that age blocking effects are present.

Summary


Two-Factor Studies with One Case per Treatment
Two-Way ANOVA Model without Interaction
Analysis of Variance
Test for Factor A and B Main Effects
Multiple Comparison Procedures
Tukey's test
Randomized Complete Block Designs