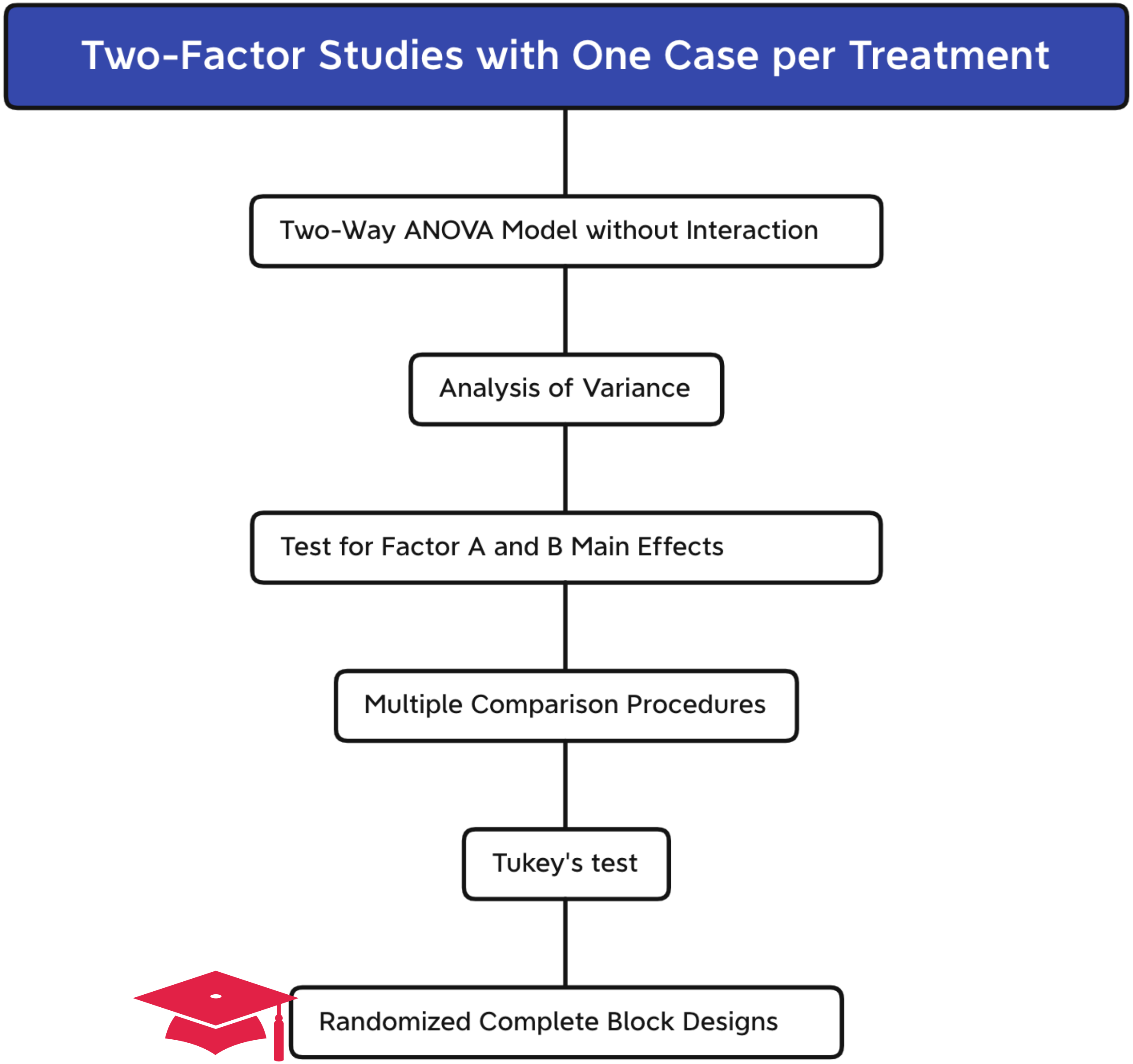
Lecture 5: Two-Factor Studies with One Case per Treatment

STA 106: Analysis of Variance

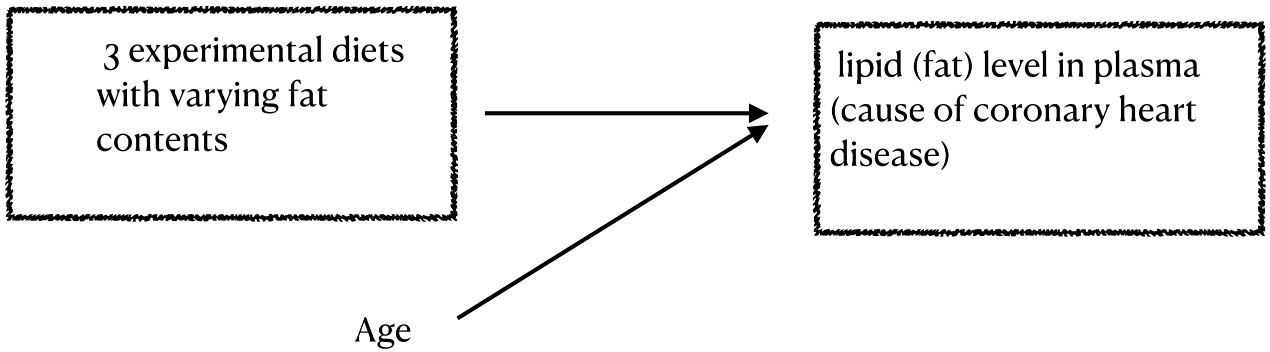
Xiner Zhou
Department of Statistics, University of California, Davis


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)

<u>The objective of the study:</u>



<u>The study setup:</u>

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)

- $\mu_{..} = \frac{\sum_i \sum_j \mu_{ij}}{ab}$: 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

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

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

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

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

Analysis of Variance

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

$df(SSTO) = n_b r - 1$

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

$df(SSBL) = n_b - 1$

$df(SSTR) = r - 1$

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

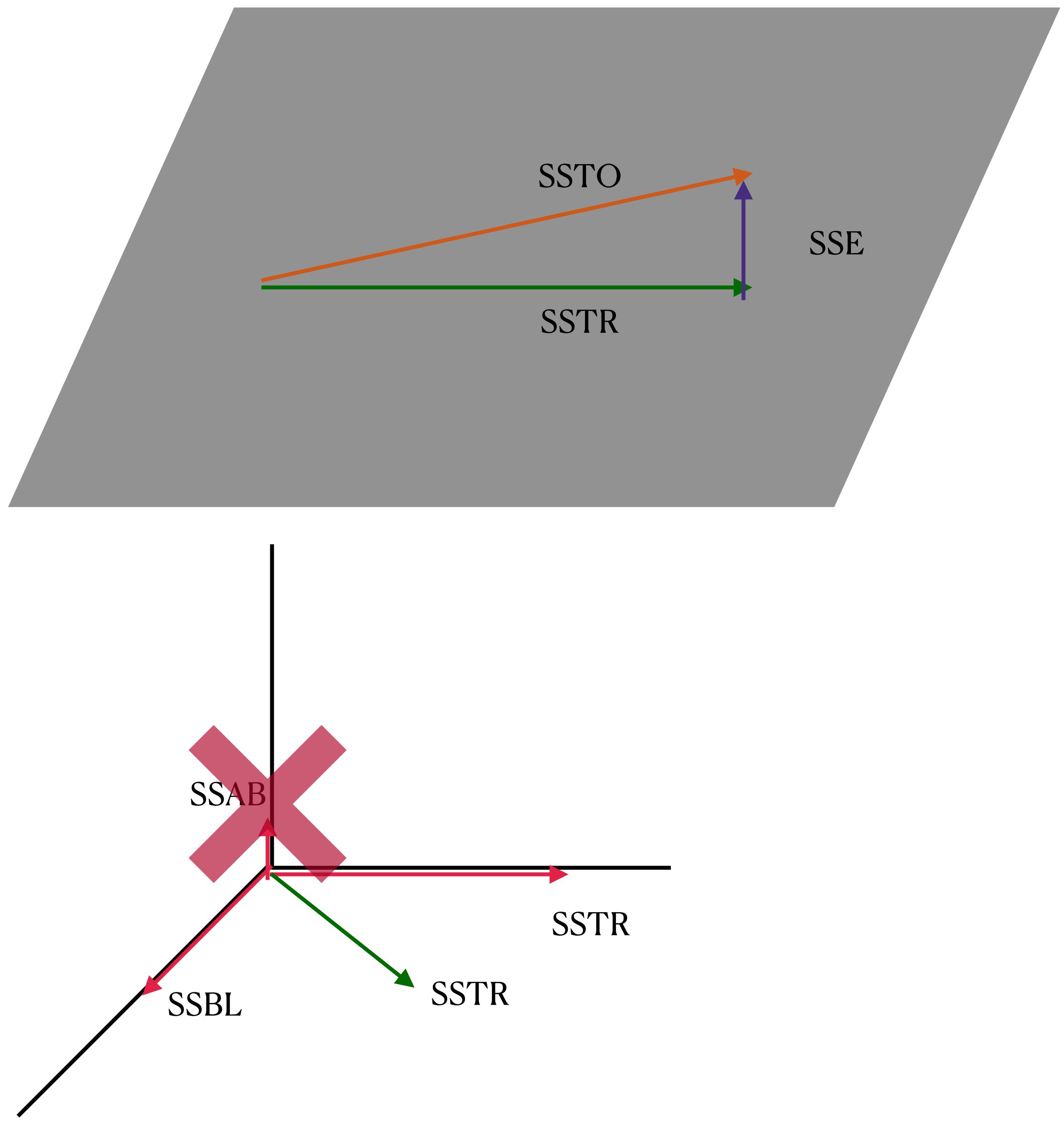
$$E[MSBL] = \sigma^2 + r\frac{\sum_i \rho_i^2}{n_b - 1} = \sigma^2 + r\frac{\sum_i \left(\mu_{i.} - \mu_{..}\right)^2}{n_b - 1}$$

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

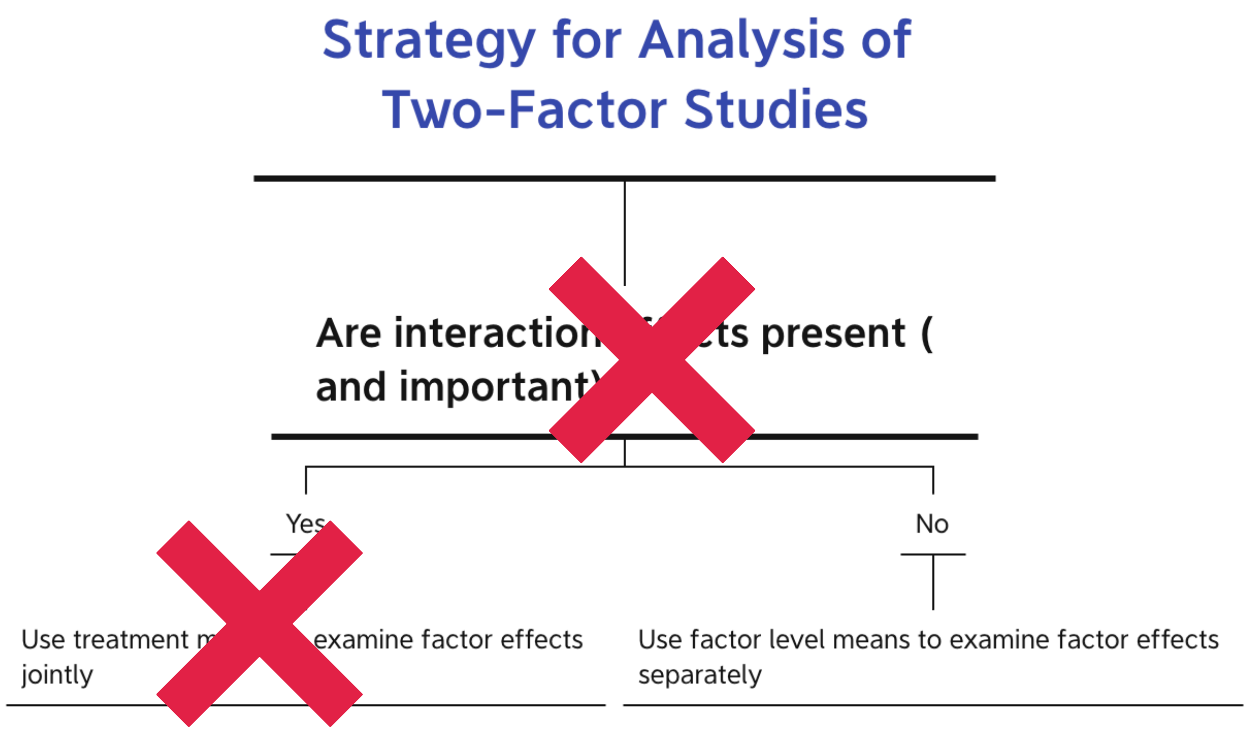
Analysis of Variance

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	$SSTR$	$r - 1$	$MSTR$	$\sigma^2 + n_b\frac{\sum \tau_i^2}{r - 1}$
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

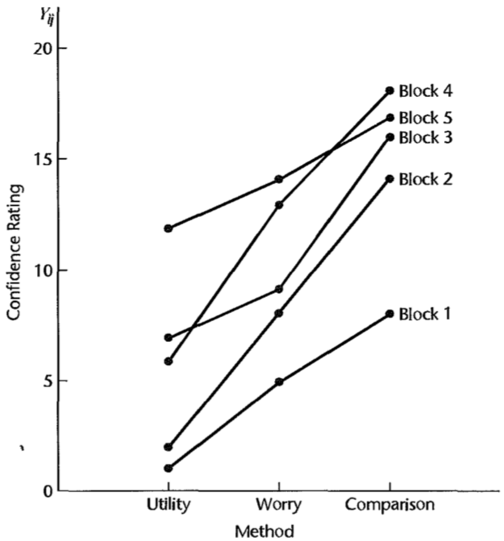
Test for Treatment (Main) Effects

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

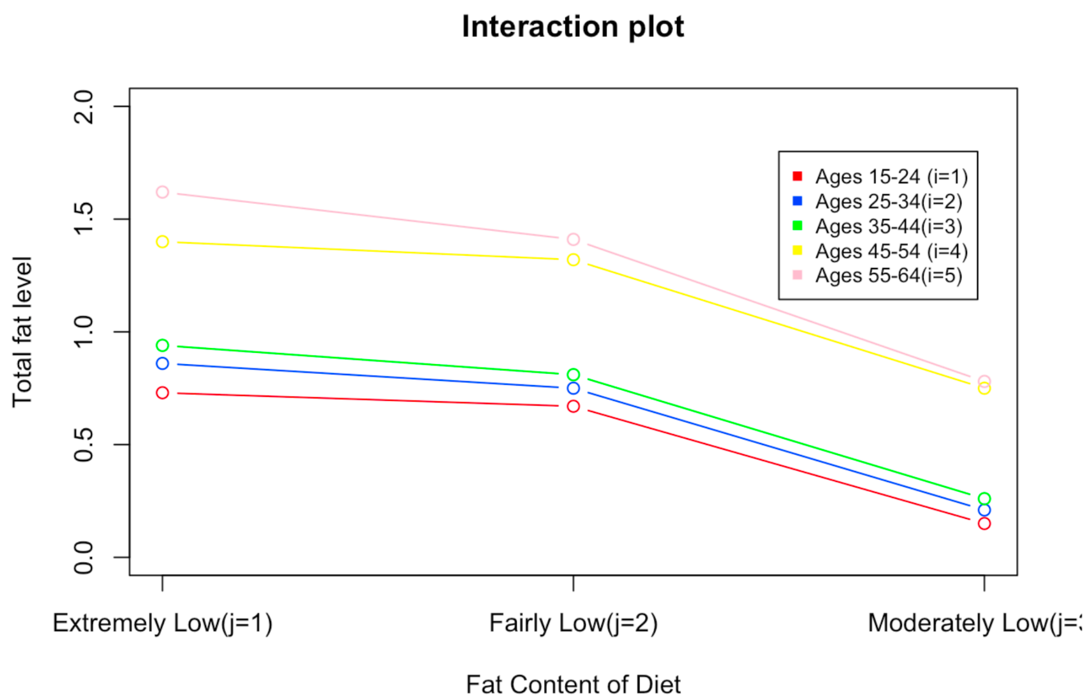
Block i		Fat Content of Diet $j = 1$ Extremely Low	$j = 2$ Fairly Low	$j = 3$ Moderately Low
1	Ages 15-24	.73	.67	.15
2	Ages 25-34	.86	.75	.21
3	Ages 35-44	.94	.81	.26
4	Ages 45-54	1.40	1.32	.75
5	Ages 55-64	1.62	1.41	.78

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 $\quad$ $H_a : D \neq 0$ 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.

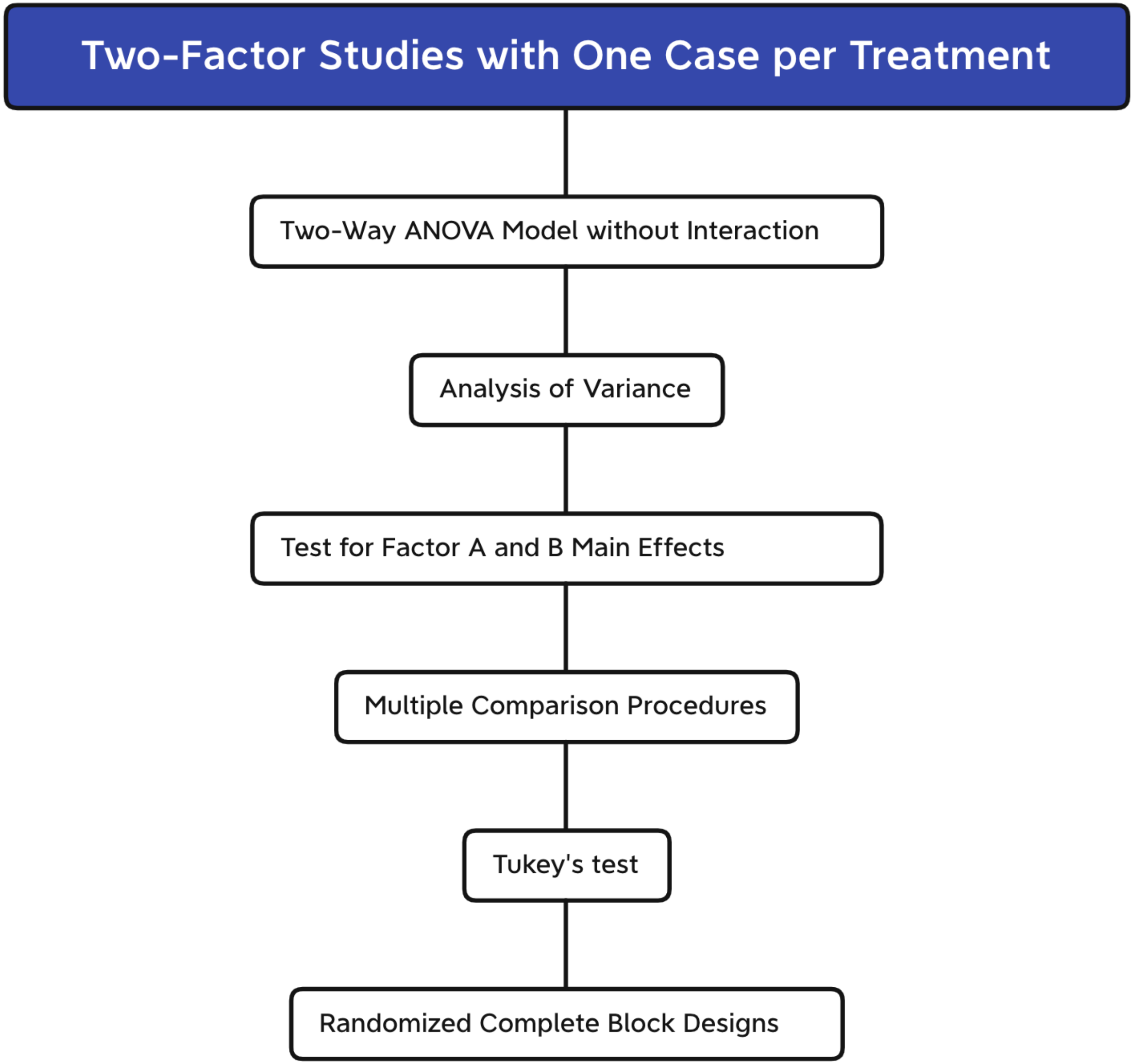
Example

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