

# Factorial ANOVA

STA305 Spring 2014

More than one categorical  
explanatory variable

See last slide for  
copyright information

# Optional Background Reading

- Chapter 7 of *Data analysis with SAS*

# Factorial ANOVA

- More than one categorical explanatory variable.
- Categorical explanatory variables are called **factors**.
- More than one at a time
- If there are observations at all combinations of explanatory variable values, it's called a *complete* factorial design (as opposed to a fractional factorial).

# The potato study

- Cases are potatoes
- Inoculate with bacteria, store for a fixed time period.
- Response variable is rotten surface area in mm.
- Two explanatory variables, randomly assigned
  - Bacteria Type (1, 2, 3)
  - Temperature (1=Cool, 2=Warm)

# Two-factor design

	<b>Bacteria Type</b>		
<b>Temp</b>	1	2	3
1=Cool			
2=Warm			

Six treatment conditions

# Factorial experiments

- Allow more than one factor to be investigated in the same study: Efficiency!
- Allow the scientist to see whether the effect of an explanatory variable *depends* on the value of another explanatory variable: Interactions
- Thank you again, Mr. Fisher.

Normal with equal variance  
and conditional (cell) means  $\mu_{i,j}$

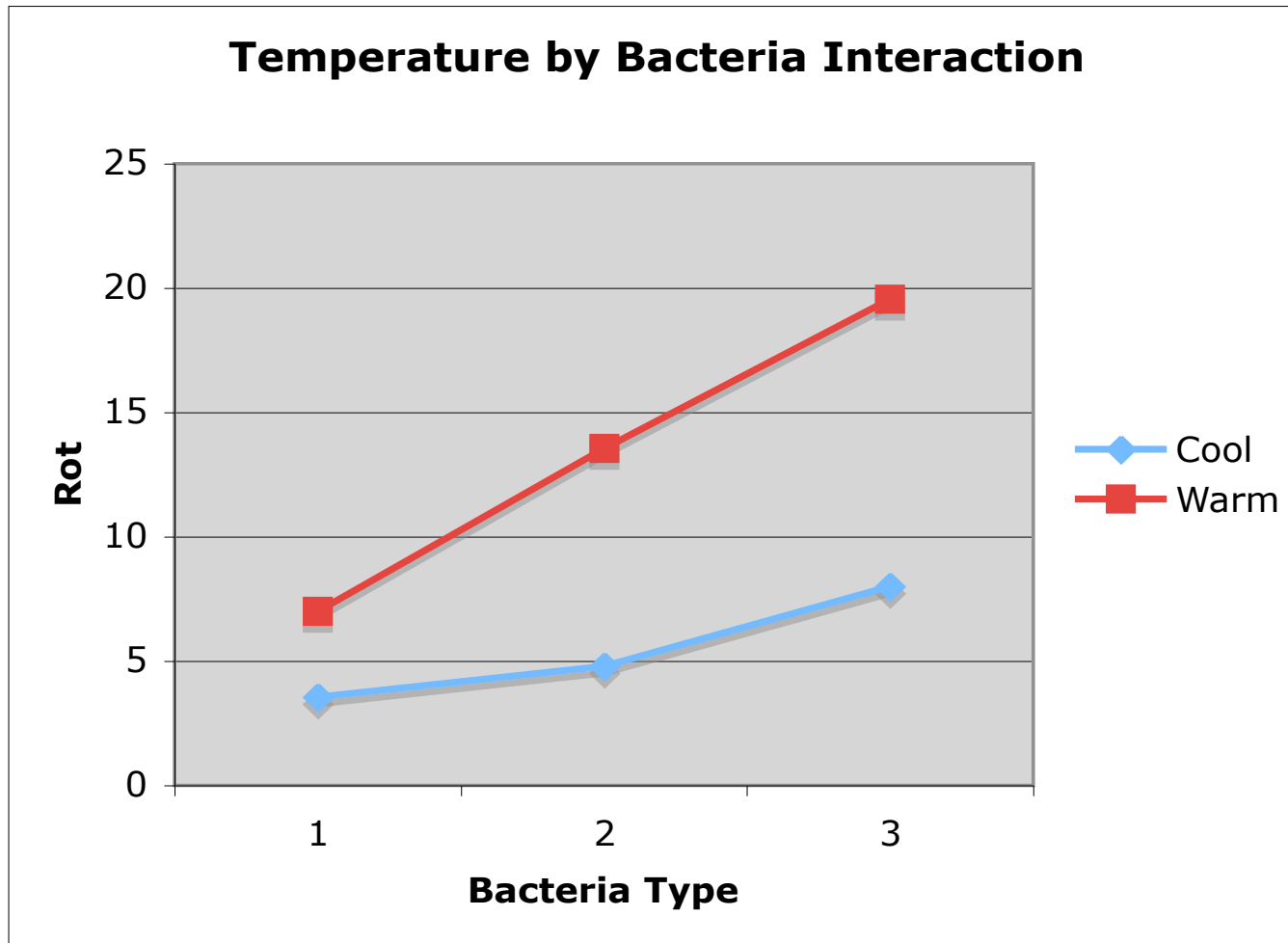
	Bacteria Type			
Temp	1	2	3	
1=Cool	$\mu_{1,1}$	$\mu_{1,2}$	$\mu_{1,3}$	$\frac{\mu_{1,1} + \mu_{1,2} + \mu_{1,3}}{3}$
2=Warm	$\mu_{2,1}$	$\mu_{2,2}$	$\mu_{2,3}$	$\frac{\mu_{2,1} + \mu_{2,2} + \mu_{2,3}}{3}$
	$\frac{\mu_{1,1} + \mu_{2,1}}{2}$	$\frac{\mu_{1,2} + \mu_{2,2}}{2}$	$\frac{\mu_{1,3} + \mu_{2,3}}{2}$	$\mu$

# Tests

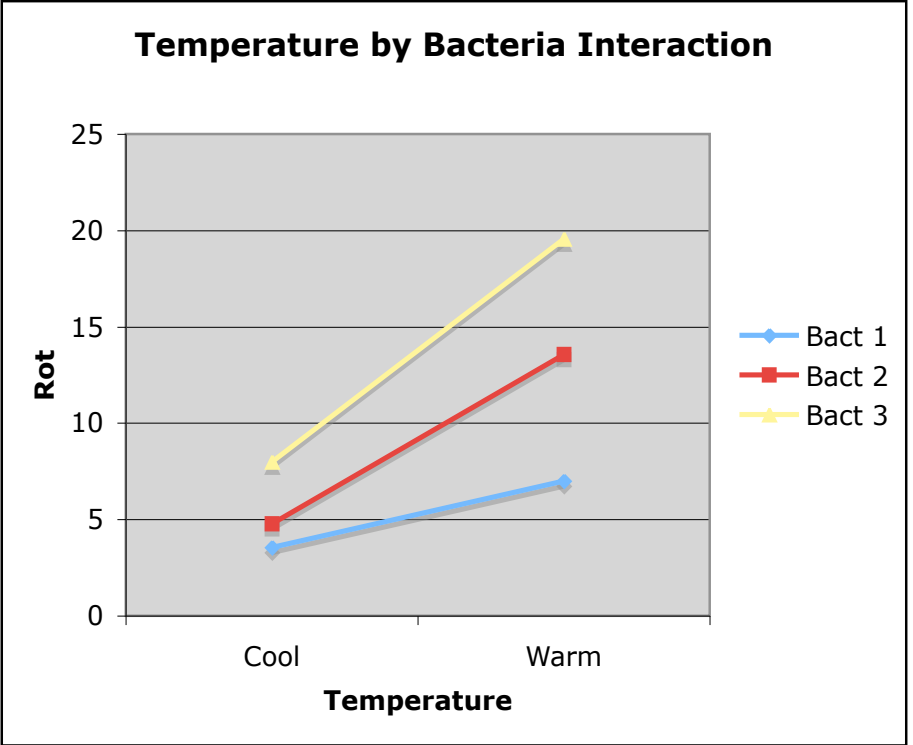
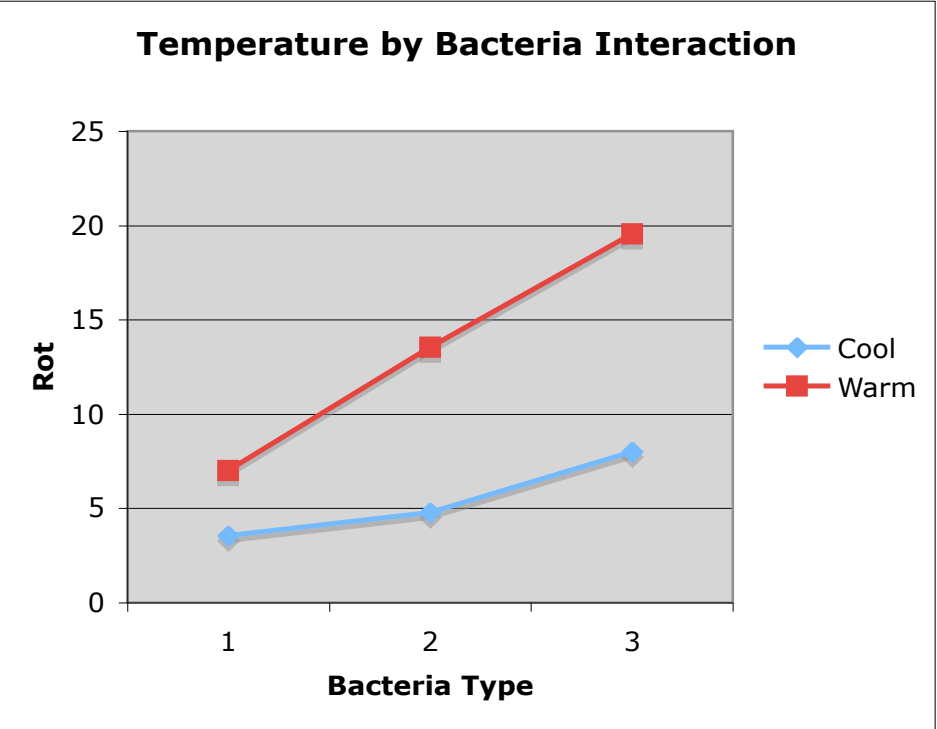
- Main effects: Differences among marginal means
- Interactions: Differences between differences (What is the effect of Factor A? **It depends** on the level of Factor B.)



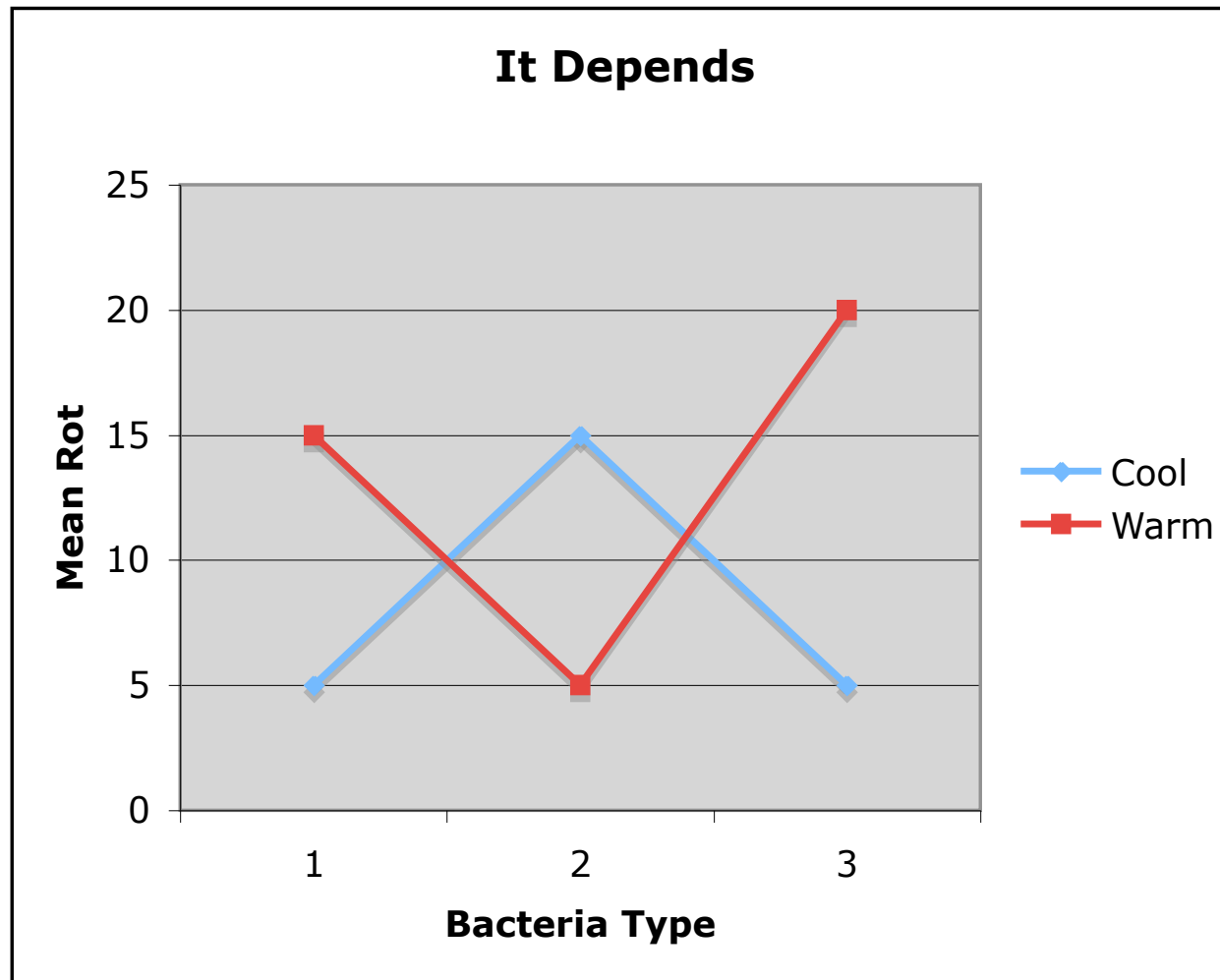
# To understand the interaction, plot the means



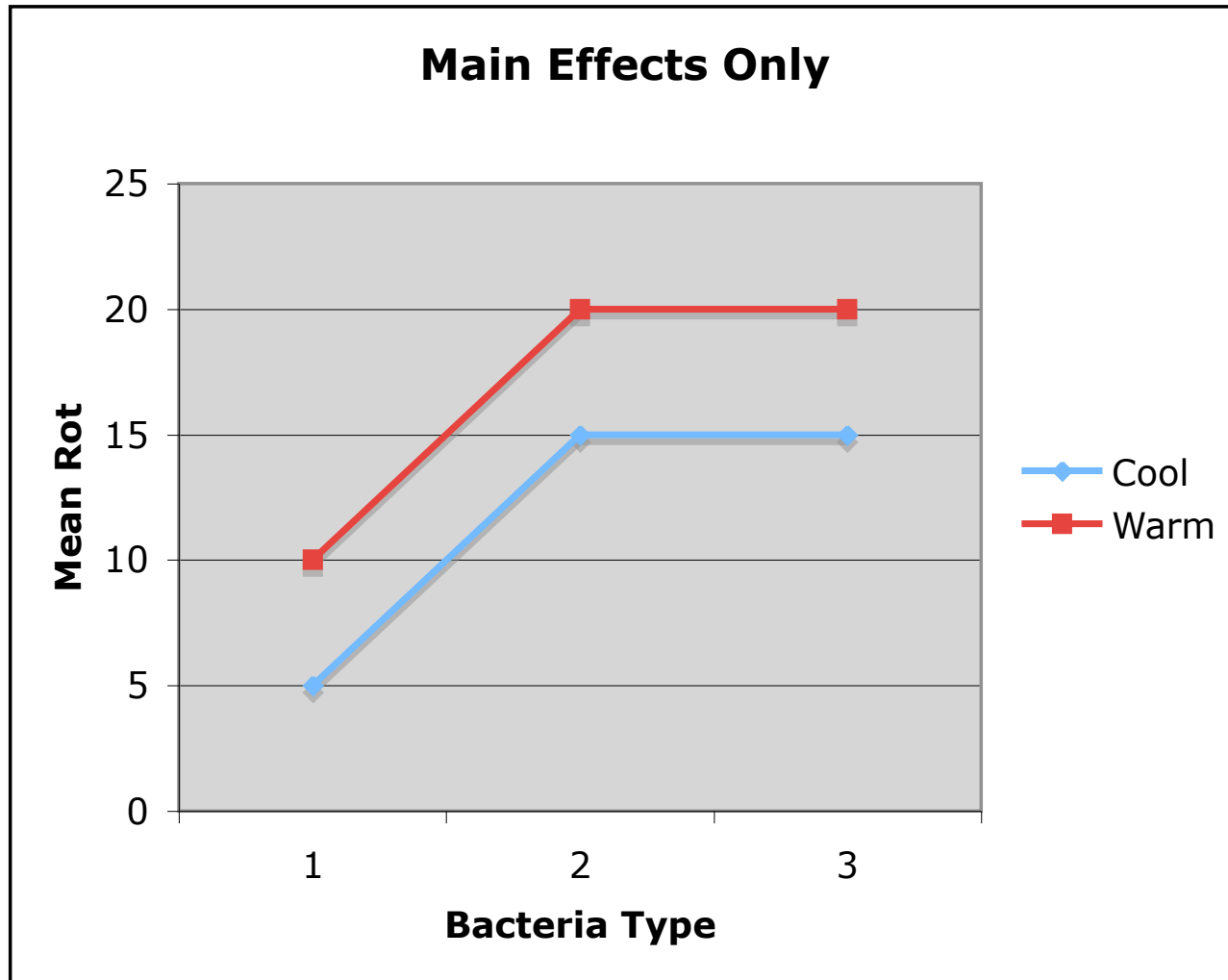
# Either Way



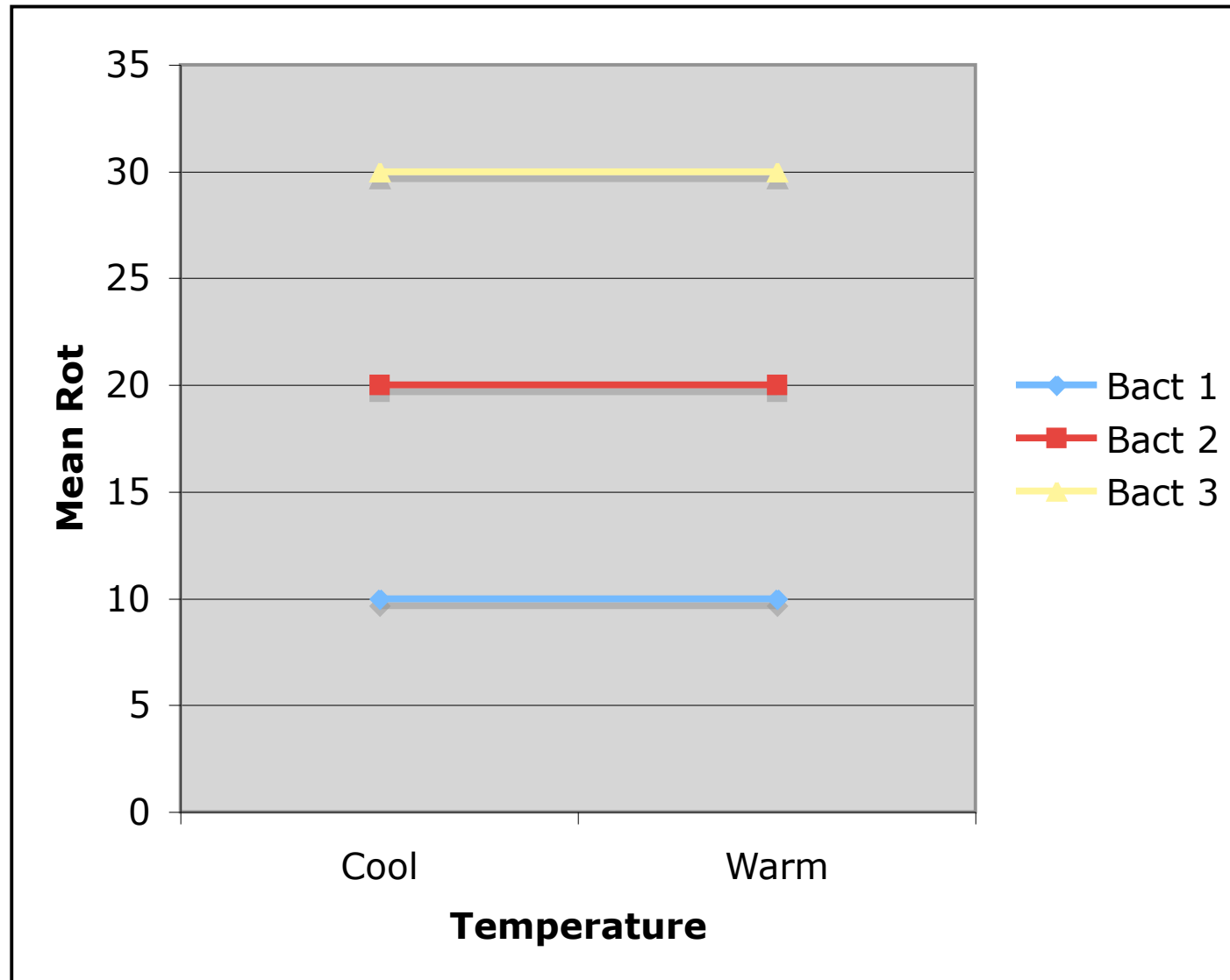
# Non-parallel profiles = Interaction



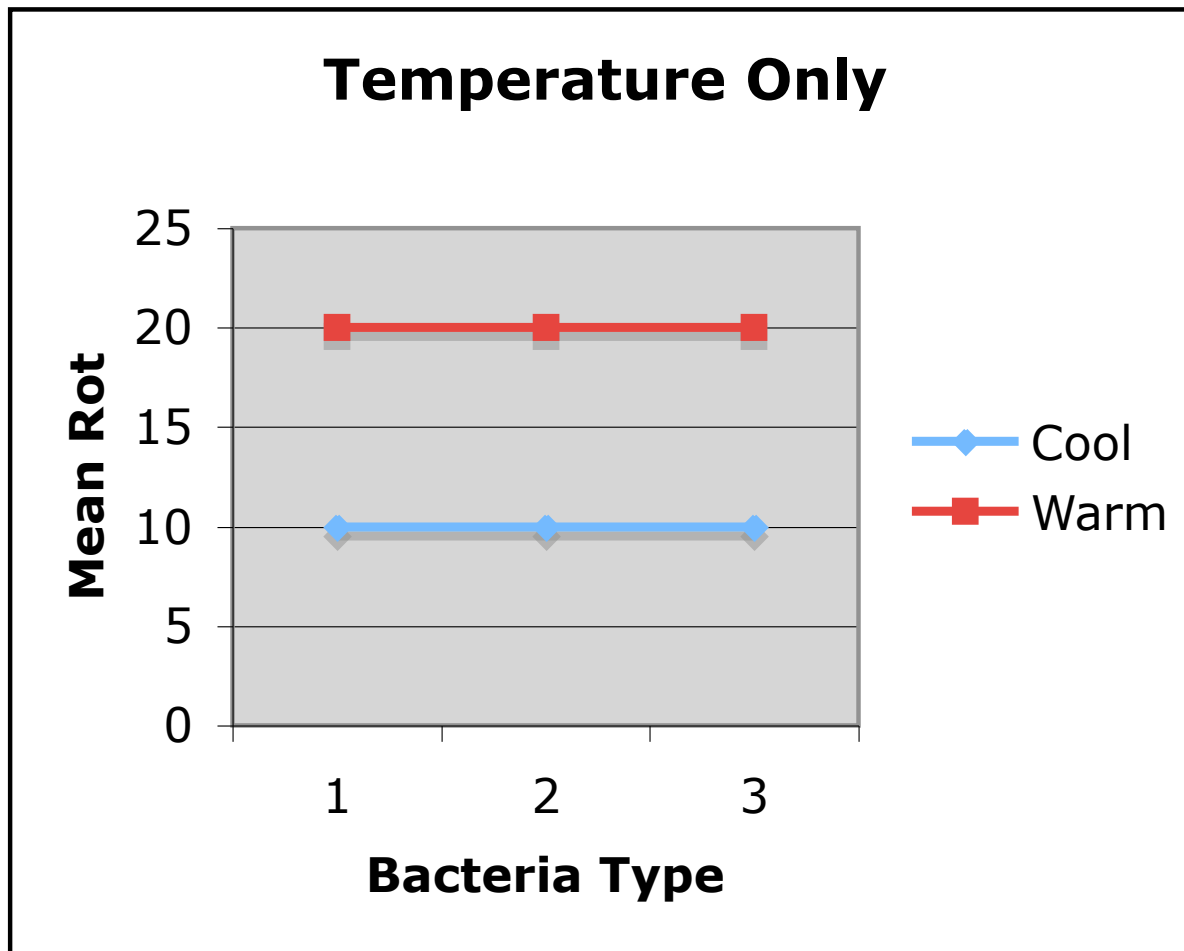
# Main effects for both variables, no interaction



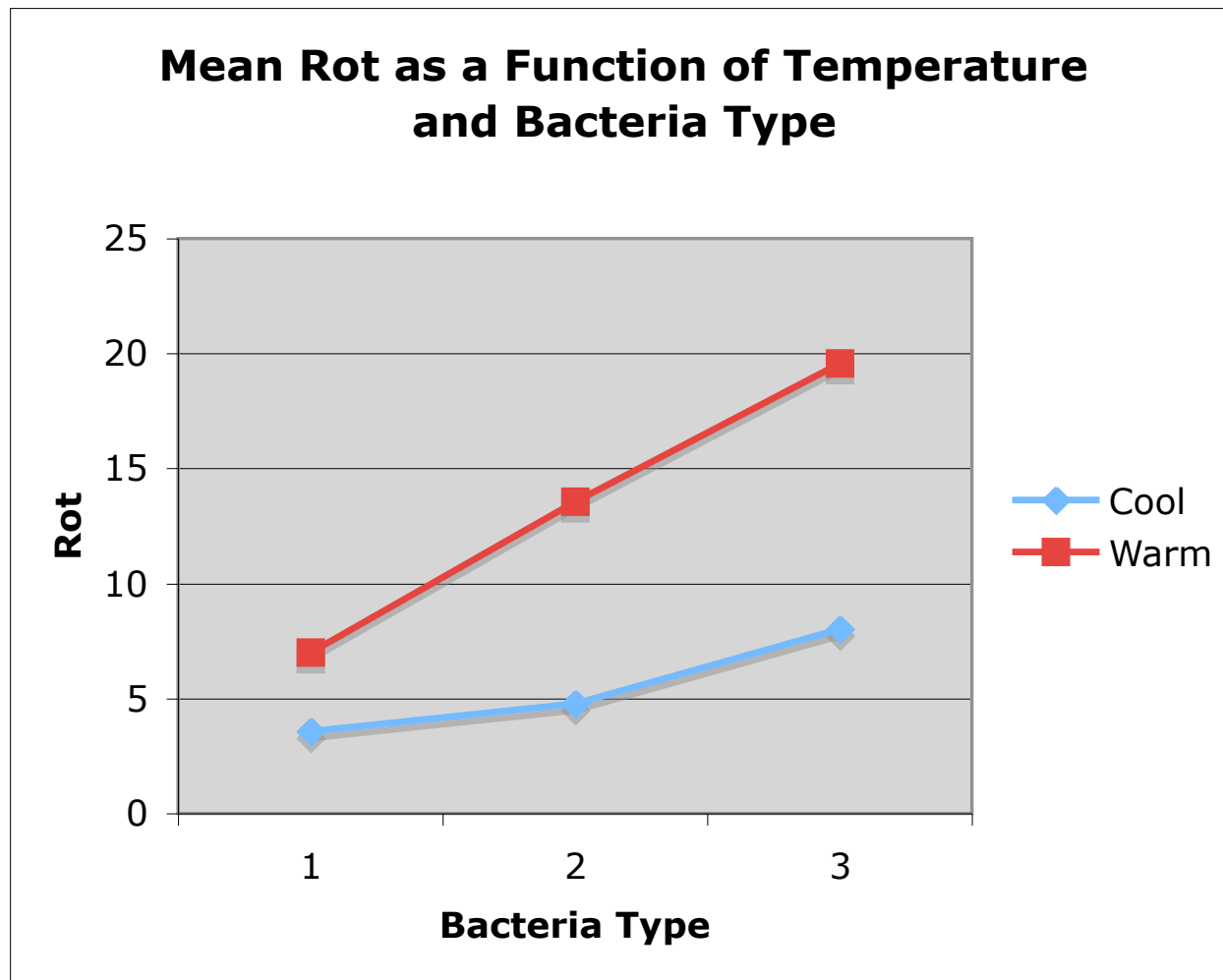
# Main effect for Bacteria only



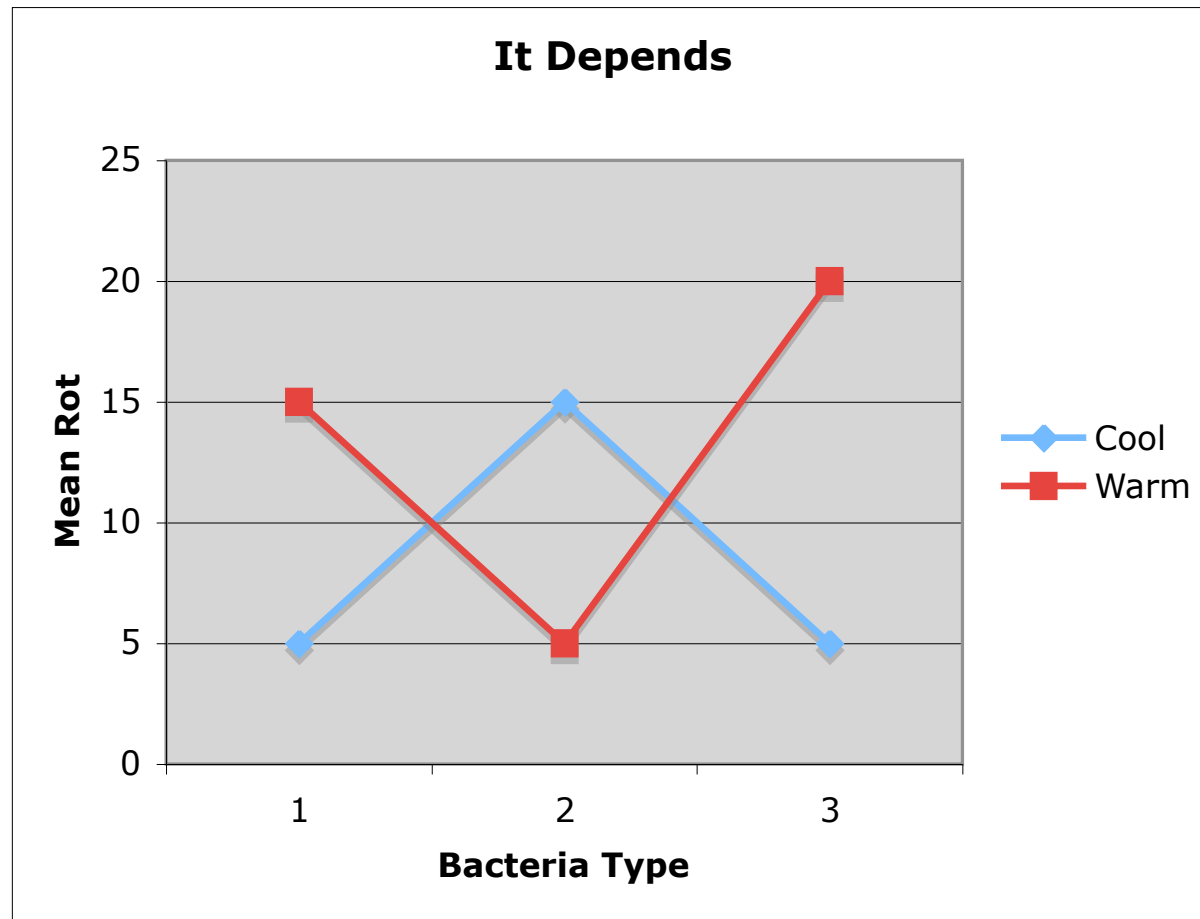
# Main Effect for Temperature Only



# Both Main Effects, and the Interaction



# Should you interpret the main effects?





# Testing Contrasts

	Bacteria Type			
Temp	1	2	3	
1=Cool	$\mu_{1,1}$	$\mu_{1,2}$	$\mu_{1,3}$	$\frac{\mu_{1,1} + \mu_{1,2} + \mu_{1,3}}{3}$
2=Warm	$\mu_{2,1}$	$\mu_{2,2}$	$\mu_{2,3}$	$\frac{\mu_{2,1} + \mu_{2,2} + \mu_{2,3}}{3}$
	$\frac{\mu_{1,1} + \mu_{2,1}}{2}$	$\frac{\mu_{1,2} + \mu_{2,2}}{2}$	$\frac{\mu_{1,3} + \mu_{2,3}}{2}$	$\mu$

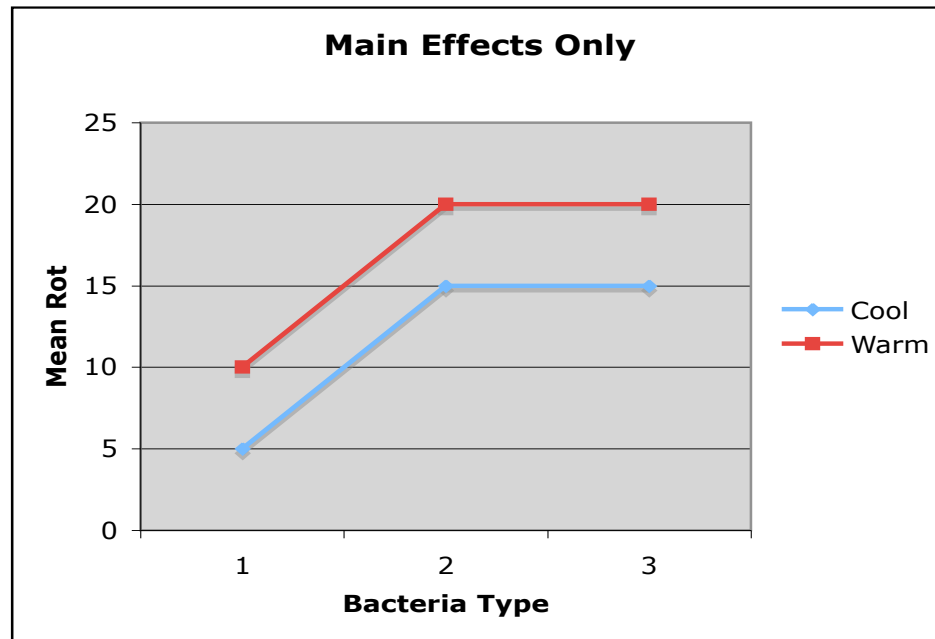
- Differences between marginal means are definitely contrasts
- Interactions are also sets of contrasts

# Interactions are sets of Contrasts

	Bacteria Type			
Temp	1	2	3	
1=Cool	$\mu_{1,1}$	$\mu_{1,2}$	$\mu_{1,3}$	$\frac{\mu_{1,1} + \mu_{1,2} + \mu_{1,3}}{3}$
2=Warm	$\mu_{2,1}$	$\mu_{2,2}$	$\mu_{2,3}$	$\frac{\mu_{2,1} + \mu_{2,2} + \mu_{2,3}}{3}$
	$\frac{\mu_{1,1} + \mu_{2,1}}{2}$	$\frac{\mu_{1,2} + \mu_{2,2}}{2}$	$\frac{\mu_{1,3} + \mu_{2,3}}{2}$	$\mu$

- $H_0 : \mu_{1,1} - \mu_{2,1} = \mu_{1,2} - \mu_{2,2} = \mu_{1,3} - \mu_{2,3}$
- $H_0 : \mu_{1,2} - \mu_{1,1} = \mu_{2,2} - \mu_{2,1}$  and  
 $\mu_{1,3} - \mu_{1,2} = \mu_{2,3} - \mu_{2,2}$

# Interactions are sets of Contrasts



- $H_0 : \mu_{1,1} - \mu_{2,1} = \mu_{1,2} - \mu_{2,2} = \mu_{1,3} - \mu_{2,3}$
- $H_0 : \mu_{1,2} - \mu_{1,1} = \mu_{2,2} - \mu_{2,1}$  and  
 $\mu_{1,3} - \mu_{1,2} = \mu_{2,3} - \mu_{2,2}$

# Equivalent statements

- The effect of A depends upon B
- The effect of B depends on A

$$H_0 : \mu_{1,1} - \mu_{2,1} = \mu_{1,2} - \mu_{2,2} = \mu_{1,3} - \mu_{2,3}$$

$$H_0 : \mu_{1,2} - \mu_{1,1} = \mu_{2,2} - \mu_{2,1} \text{ and}$$

$$\mu_{1,3} - \mu_{1,2} = \mu_{2,3} - \mu_{2,2}$$

# Three factors: A, B and C

- There are three (sets of) main effects: One each for A, B, C
- There are three two-factor interactions
  - A by B (Averaging over C)
  - A by C (Averaging over B)
  - B by C (Averaging over A)
- There is one three-factor interaction:  $A \times B \times C$

# Meaning of the 3-factor interaction

- The form of the  $A \times B$  interaction depends on the value of  $C$
- The form of the  $A \times C$  interaction depends on the value of  $B$
- The form of the  $B \times C$  interaction depends on the value of  $A$
- These statements are equivalent. Use the one that is easiest to understand.

# To graph a three-factor interaction

- Make a separate mean plot (showing a 2-factor interaction) for each value of the third variable.
- In the potato study, a graph for each type of potato

# Four-factor design

- Four sets of main effects
- Six two-factor interactions
- Four three-factor interactions
- One four-factor interaction: The nature of the three-factor interaction depends on the value of the 4th factor
- There is an F test for each one
- And so on ...



# As the number of factors increases

- The higher-way interactions get harder and harder to understand
- All the tests are still tests of sets of contrasts (differences between differences of differences ...)
- But it gets harder and harder to write down the contrasts
- Effect coding becomes easier

# Effect coding

<b>Bact</b>	<b>B<sub>1</sub></b>	<b>B<sub>2</sub></b>
1	1	0
2	0	1
3	-1	-1

<b>Temperature</b>	<b>T</b>
1=Cool	1
2=Warm	-1

$$E(Y|\mathbf{X} = \mathbf{x}) = \beta_0 + \beta_1 B_1 + \beta_2 B_2 + \beta_3 T + \beta_4 B_1 T + \beta_5 B_2 T$$

# Interaction effects are products of dummy variables

$$E(Y|\mathbf{X} = \mathbf{x}) = \beta_0 + \beta_1 B_1 + \beta_2 B_2 + \beta_3 T + \beta_4 B_1 T + \beta_5 B_2 T$$

- The A x B interaction: Multiply each dummy variable for A by each dummy variable for B
- Use these products as additional explanatory variables in the multiple regression
- The A x B x C interaction: Multiply each dummy variable for C by each product term from the A x B interaction
- Test the sets of product terms simultaneously

# Make a table

$$E(Y|\mathbf{X} = \mathbf{x}) = \beta_0 + \beta_1 B_1 + \beta_2 B_2 + \beta_3 T + \beta_4 B_1 T + \beta_5 B_2 T$$

Bact	Temp	$B_1$	$B_2$	$T$	$B_1 T$	$B_2 T$	$E(Y \mathbf{X} = \mathbf{x})$
1	1	1	0	1	1	0	$\beta_0 + \beta_1 + \beta_3 + \beta_4$
1	2	1	0	-1	-1	0	$\beta_0 + \beta_1 - \beta_3 - \beta_4$
2	1	0	1	1	0	1	$\beta_0 + \beta_2 + \beta_3 + \beta_5$
2	2	0	1	-1	0	-1	$\beta_0 + \beta_2 - \beta_3 - \beta_5$
3	1	-1	-1	1	-1	-1	$\beta_0 - \beta_1 - \beta_2 + \beta_3 - \beta_4 - \beta_5$
3	2	-1	-1	-1	1	1	$\beta_0 - \beta_1 - \beta_2 - \beta_3 + \beta_4 + \beta_5$

# Cell and Marginal Means

	<b>Bacteria Type</b>			
<b>Tmp</b>	<b>1</b>	<b>2</b>	<b>3</b>	
<b>1=C</b>	$\beta_0 + \beta_1 + \beta_3 + \beta_4$	$\beta_0 + \beta_2 + \beta_3 + \beta_5$	$\beta_0 - \beta_1 - \beta_2$ $+ \beta_3 - \beta_4 - \beta_5$	$\beta_0$ $+ \beta_3$
<b>2=W</b>	$\beta_0 + \beta_1 - \beta_3 - \beta_4$	$\beta_0 + \beta_2 - \beta_3 - \beta_5$	$\beta_0 - \beta_1 - \beta_2$ $- \beta_3 + \beta_4 + \beta_5$	$\beta_0$ $- \beta_3$
	$\beta_0 + \beta_1$	$\beta_0 + \beta_2$	$\beta_0 - \beta_1 - \beta_2$	$\beta_0$

# We see

- Intercept is the grand mean
- Regression coefficients for the dummy variables are deviations of the marginal means from the grand mean
- What about the interactions?

$$E(Y|\mathbf{X} = \mathbf{x}) = \beta_0 + \beta_1 B_1 + \beta_2 B_2 + \beta_3 T + \beta_4 B_1 T + \beta_5 B_2 T$$

**A bit of algebra shows**

$$\mu_{1,1} - \mu_{2,1} = \mu_{1,2} - \mu_{2,2} \text{ is equivalent to } \beta_4 = \beta_5$$

$$\mu_{1,2} - \mu_{2,2} = \mu_{1,3} - \mu_{2,3} \text{ is equivalent to } \beta_4 = -\beta_5$$

$$\text{So } \beta_4 = \beta_5 = 0$$

# Factorial ANOVA with effect coding is pretty automatic

- You don't have to make a table unless asked.
- It always works as you expect it will.
- Hypothesis tests are the same as testing sets of contrasts.



# Again

- Intercept is the grand mean.
- Regression coefficients for the dummy variables are deviations of the marginal means from the grand mean.
- Test of main effect(s) is test of the dummy variables for a factor.
- Interaction effects are products of dummy variables.

# Balanced vs. Unbalanced Experimental Designs

- Balanced design: Cell sample sizes are proportional (maybe equal).
- Explanatory variables have zero relationship to one another.
- Numerator SS in ANOVA are independent (because contrasts are orthogonal).
- Everything is nice and simple
- Most experimental studies are designed this way.
- As soon as somebody drops a test tube, it's no longer balanced.

# Analysis of unbalanced data

- When explanatory variables are related, there is potential ambiguity.
- A is related to Y, B is related to Y, and A is related to B.
- Who gets credit for the portion of variation in Y that could be explained by either A or B?
- With a regression approach, whether you use contrasts or dummy variables (equivalent), the answer is **nobody**.
- Think of full, reduced models.
- Equivalently, general linear test

# Some software is designed for balanced data

- The special purpose formulas are much simpler.
- Very useful *in the past*.
- Since most data are at least a little unbalanced, a recipe for trouble.
- Most textbook data are balanced, so they cannot tell you what your software is really doing.
- R's `anova` and `aov` functions are designed for balanced data, though `anova` applied to `lm` objects can give you what you want if you use it with care.
- SAS `proc glm` is much more convenient. SAS `proc anova` is for balanced data.

# Copyright Information

This slide show was prepared by Jerry Brunner, Department of Statistics, University of Toronto. It is licensed under a Creative Commons Attribution - ShareAlike 3.0 Unported License. Use any part of it as you like and share the result freely. These PowerPoint slides will be available from the course website:

<http://www.utstat.toronto.edu/brunner/oldclass/305s14>