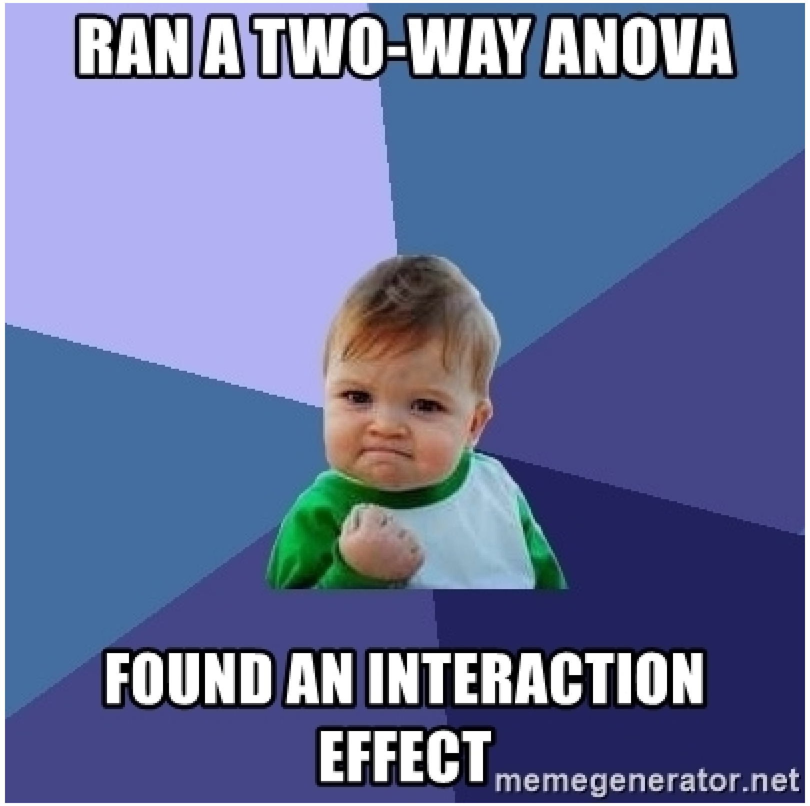
Analysis of Variance continued



Overview

Review/continuation of one-way ANOVA

Pairwise comparisons after running an ANOVA

Factorial ANOVAs and interaction effects

If there is time: string manipulation

One-way ANOVA

A **one-way analysis of variance (ANOVA)** is a parametric hypothesis test that can be used to examine if a set of means are all the same.

H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$

H_A: $\mu_i \neq \mu_j$ for some i, j

The statistic we use for a one-way ANOVA is the F-statistic

$$F = \frac{\text{between-group variability}}{\text{within-group variability}} = \frac{\frac{1}{K-1}\sum_{i=1}^{K} n_i(\bar{x}_i - \bar{x}_{tot})^2}{\frac{1}{N-K}\sum_{i=1}^{K}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2}$$

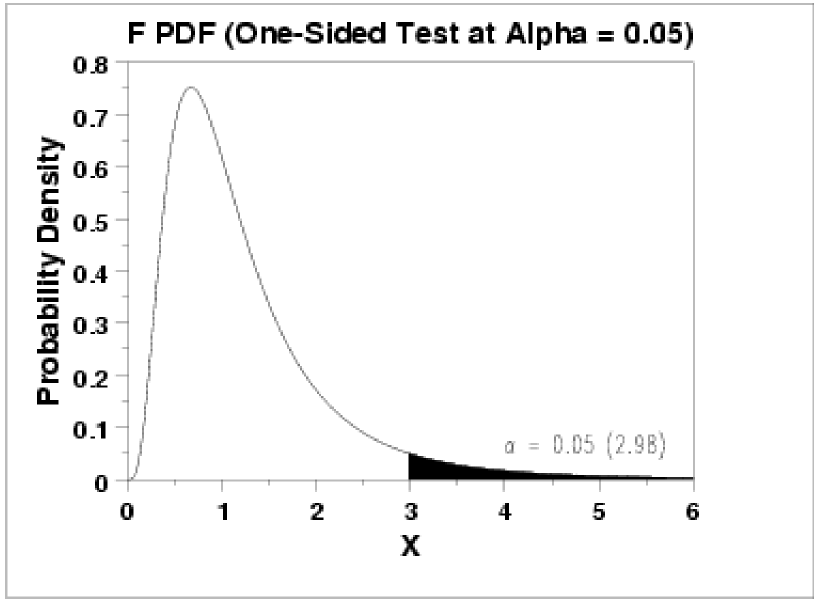
One-way ANOVA – the central idea

If H_0 is true, the F-statistic will come from an F distribution with parameters

- $df_1 = K - 1$
- $df_2 = N - K$

The F-distribution is valid if these conditions are met:

- The data in each group should follow a normal distribution
  - Check this with a Q-Q plot
- The variances in each group should be approximately equal
  - Check that $s_{max}/s_{min} < 2$



ANOVAs are robust to these assumptions, but what can we do if they are very badly violated?

Kruskal-Wallis (non-parametric) test

There are also **non-parametric** tests which don't make assumptions about normality

The **Kruskal-Wallis** test compares several groups to see if one of the groups 'stochastically dominates' another

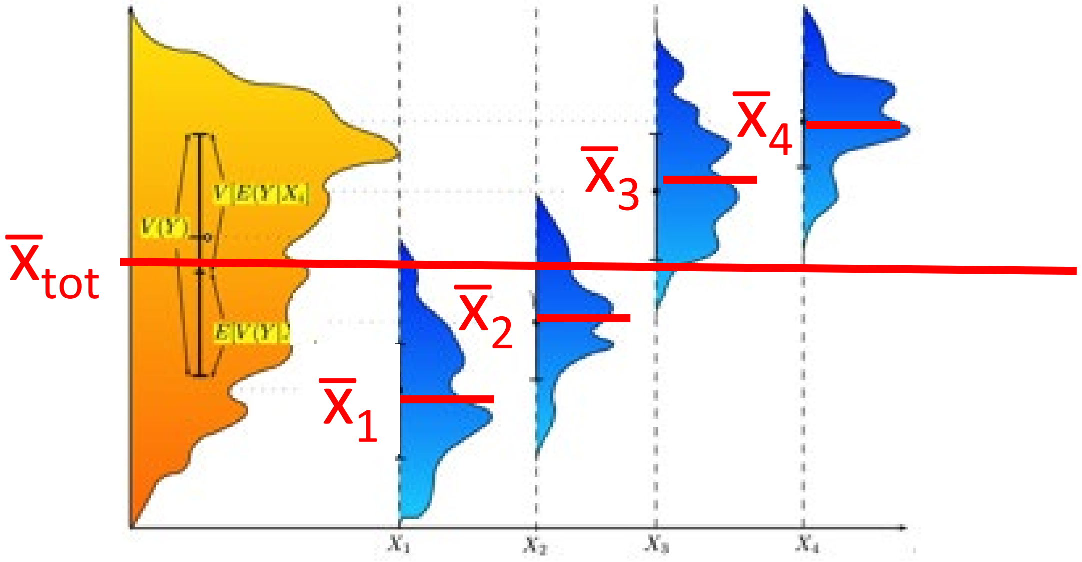
- Does not assume normality
- Tests if one group stochastically dominates another group
- Also tests whether the median for all the groups are the same
  - (if you assume groups have the same shaped and scale)
- The test is based on ranks so it is not influenced by outliers

The F-statistic

$$F = \frac{\text{between-group variability}}{\text{within-group variability}} = \frac{\frac{1}{K-1}\sum_{i=1}^{K} n_i(\bar{x}_i - \bar{x}_{tot})^2}{\frac{1}{N-K}\sum_{i=1}^{K}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2}$$

The F statistic measures a fraction of:

$$F = \frac{\text{variability between group means}}{\text{variability within each group}}$$



ANOVA table

$$F = \frac{\text{between-group variability}}{\text{within-group variability}} = \frac{\frac{1}{K-1}\sum_{i=1}^{K} n_i(\bar{x}_i - \bar{x}_{tot})^2}{\frac{1}{N-K}\sum_{i=1}^{K}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2}$$

Source	df	Sum of Sq.	Mean Square	F-statistic	p-value
Groups	$k-1$	SSG	$MSG = \frac{SSG}{k-1}$	$F = \frac{MSG}{MSE}$	Upper tail $F_{k-1,n-k}$
Error	$n-k$	SSE	$MSE = \frac{SSE}{n-k}$		
Total	$n-1$	$SSTotal$			

Where:

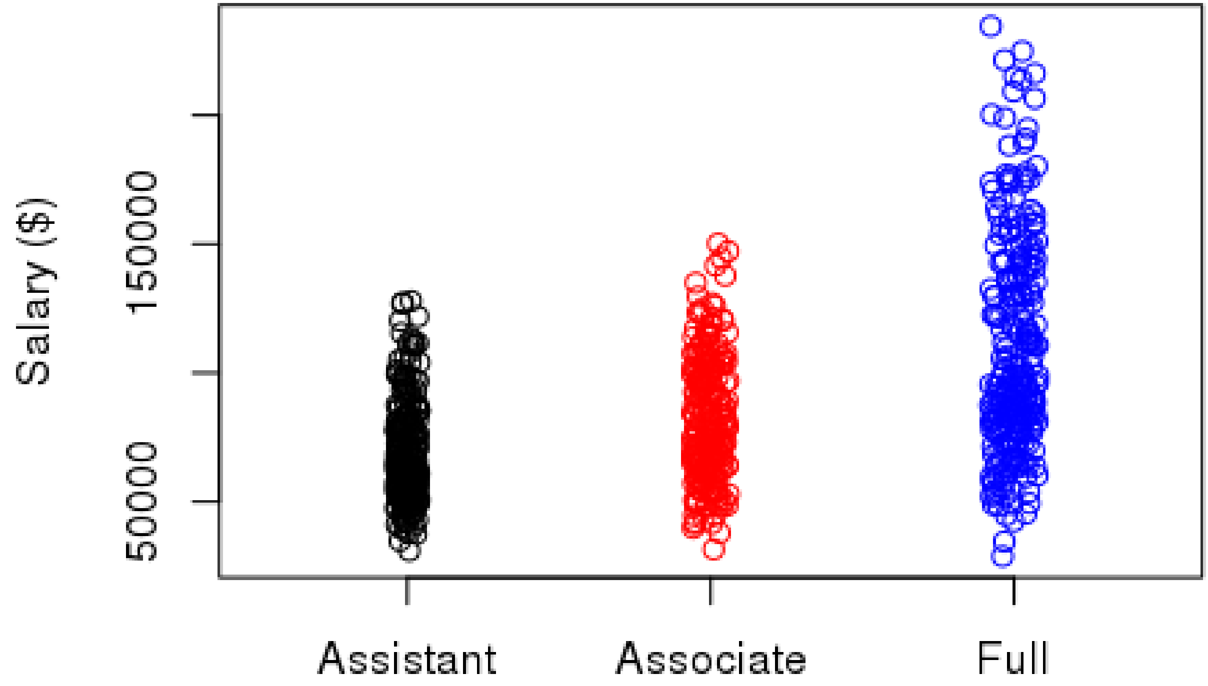
$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_{tot})^2$$

$$SSG = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x}_{tot})^2$$

SST = SSG + SSE

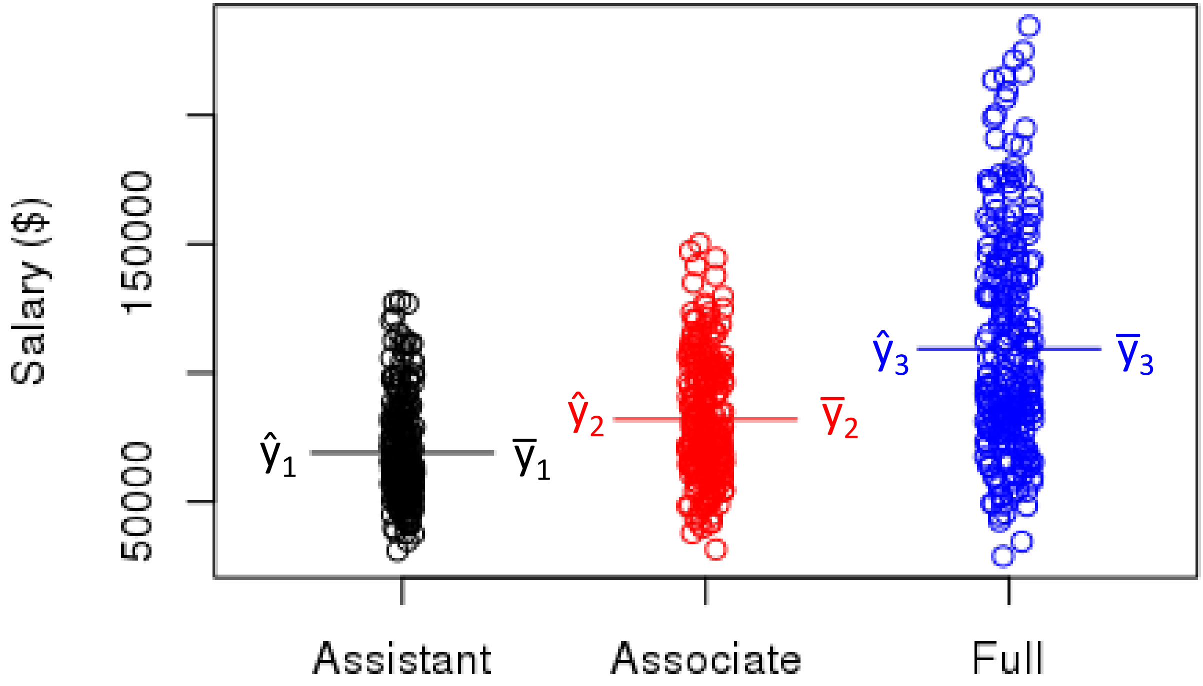
$$SSE = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2$$

ANOVA as regression with only categorical predictors



$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if Assistant Professor} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if Associate Professor} \\ \beta_0 + \epsilon_i & \text{if Full Professor} \end{cases}$$

Least squares prediction for $\hat{y}_i$ is $\bar{y}_k$



$$\hat{y}_i = \bar{y}_k = \begin{cases} \bar{y}_1 & \text{if Assistant professor} \\ \bar{y}_2 & \text{if Associate professor} \\ \bar{y}_3 & \text{if Full} \end{cases}$$

Planned comparisons/posthoc tests

Suppose we run a one-way ANOVA and we are able to reject the null hypothesis.

$$H_0: \mu_1 = \mu_2 = \ldots = \mu_k$$

$$H_A: \mu_i \neq \mu_j \text{ for some } i, j$$

Q: What else would we like to know?

Pairwise comparisons

There are several tests that can be used to examine which pairs of means differed; i.e., to test:

- H_0: $\mu_i = \mu_j$
- H_A: $\mu_i \neq \mu_j$

These tests include:

- Fisher's Least Significant Difference
- Bonferroni procedure/correction
- Tukeys Honest significantly different

Fisher's Least Significant Difference (LSD)

1. Perform the ANOVA

2. If the ANOVA F-test is not significant, stop

3. If the ANOVA F-test is significant, then you can test H_0 for a pairwise comparisons using:

$$t = \frac{\bar{x}_i - \bar{x}_j}{\sqrt{MSE \cdot (\frac{1}{n_i} + \frac{1}{n_j})}}$$

Estimate of the SE
Uses the MSE as a pooled estimate of the σ^2

Use a t-distribution with n-k degrees of freedom

Very 'liberal' tests

- Likely to make Type I errors (lots of false rejections of H_0)
- Less likely to make Type II errors (highest chance of detecting effects)

Bonferroni correction

Controls for the ***family-wise error rate***

- i.e., α = 0.05 for making ***any*** Type I error ***over all pairs of comparisons***

1. Choose an α-level for the family-wise error rate α
2. Decide how many comparisons you will make. Call this m.
3. Reject any hypothesis tests that have p-values less than α/m
   - Pairwise tests typically done using a t-statistic, where the MSE is used in the estimate of the SE

$$t = \frac{\bar{x}_i - \bar{x}_j}{\sqrt{MSE \cdot (\frac{1}{n_i} + \frac{1}{n_j})}}$$

Use a t-distribution with n-k degrees of freedom

Very 'conservative' tests

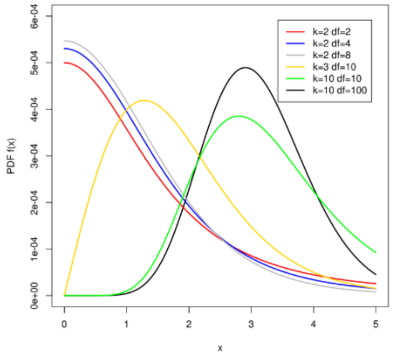
- Unlikely to make Type I errors (few false rejections of H_0)
- Likely to make Type II errors (insensitive at detecting real effects)

Tukey's Honest Significantly Different Test

Tukey's Honest Significantly Different test controls for the family-wise error rate but is less conservative than the Bonferroni correction

If the null hypothesis was true, q comes from a ***studentized range distribution***

$$q = \frac{\sqrt{2}(\bar{x}_{max} - \bar{x}_{min})}{\sqrt{MSE \cdot (\frac{1}{n_{max}} + \frac{1}{n_{min}})}}$$

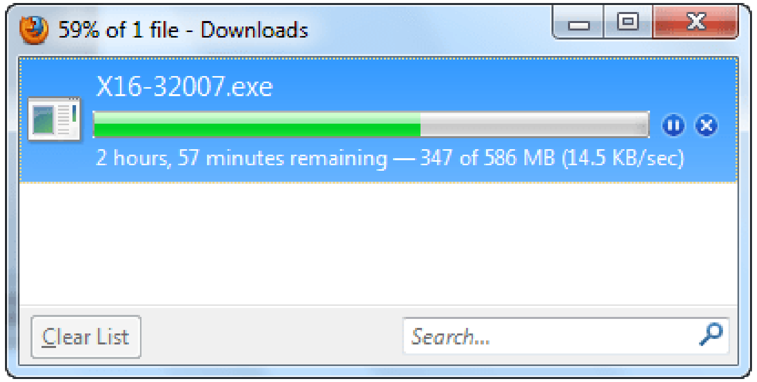


We can compare $q = \frac{\sqrt{2}(\bar{x}_i - \bar{x}_j)}{\sqrt{MSE \cdot (\frac{1}{n_i} + \frac{1}{n_j})}}$ for a pair of means i, j, to a studentized range distribution with parameters k, and N-k, to get a p-value

- Still based on assumptions that the data in each group is normal with equal variance

Motivating example: How does the time of the day affect download speeds?

A college sophomore was interested in knowing whether the time of day affected the speed at which he could download files from the Internet.

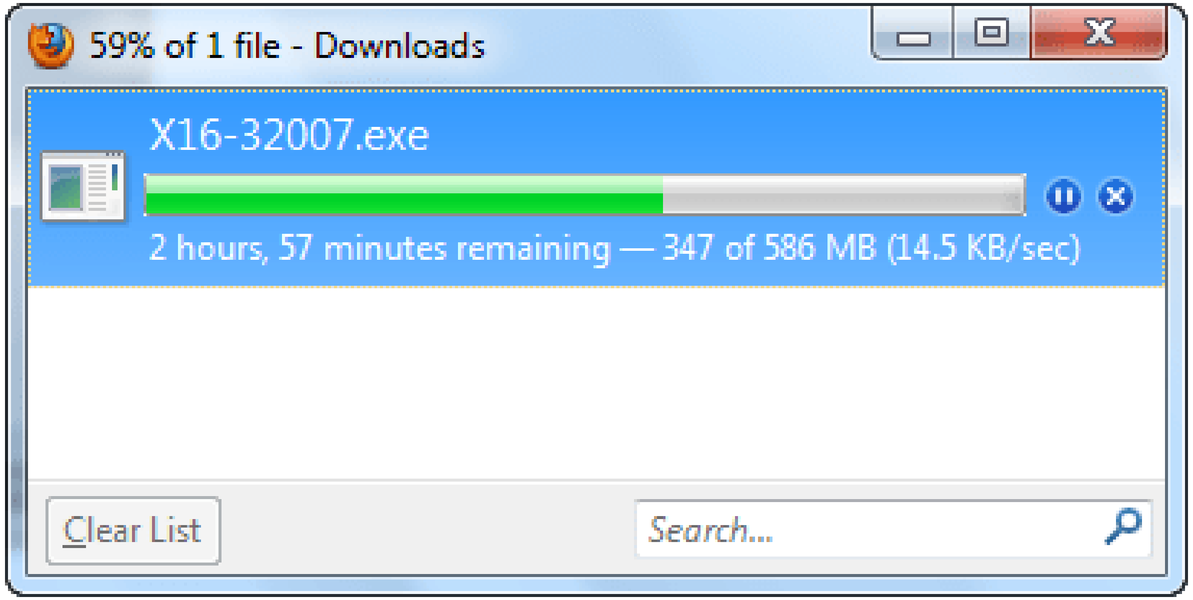


To address this question, he placed a file on a remote server and then proceeded to download it at three different time periods of the day:

- 7AM, 5PM, 12AM

He downloaded the file 48 times in all, 16 times at each time of day, and recorded the time in seconds that the download took.

Let’s try the Kruskal-Wallis test and pairwise comparisons in R…

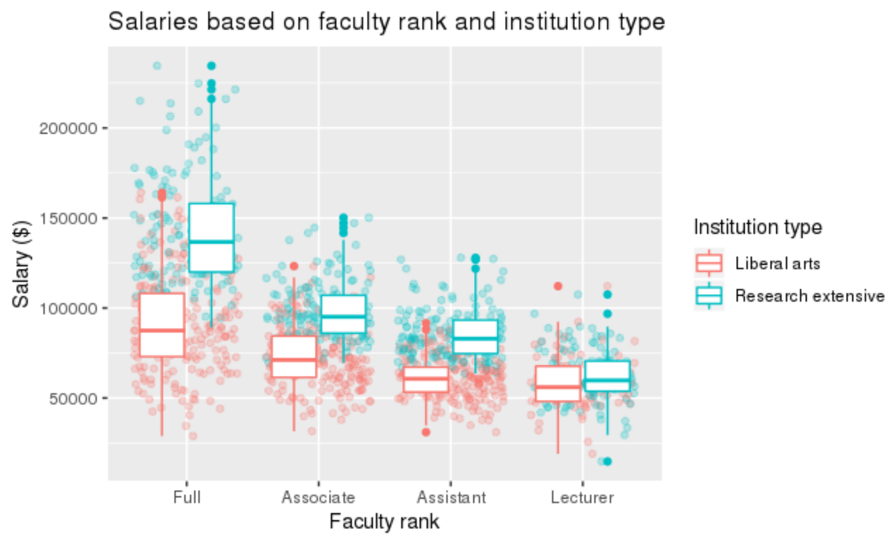


Factorial ANOVA

In a **factorial ANOVA**, we model the response variable y as a function of **more than one** categorical predictor

Example 1: Do faculty salaries depend on faculty rank, and the type of college/university

- Factors he looked at were:
  - **Rank**: Lecturer, Assistant, Associate, Full
  - **Institute**: liberal arts college, research university
  - 4 x 2 design



Factorial ANOVA

Example 2: A student at Queensland University of Technology conducted an experiment to determine what types of sandwiches ants prefer.

- Factors he looked at were:
  - **Bread**: rye, whole wheat multigrain, white
  - **Filling**: peanut better, ham and pickle, and vegemite
  - 4 x 3 design

The student creating 4 sandwiches of all combinations of bread and filling (48 sandwiches total) and randomly left pieces in front of ant nests.

He then measured how many ants were on the sandwiches 5 minutes later.

Two-way ANOVA hypotheses

Main effect for A (bread type doesn't matter or faculty rank doesn't matter)

H_0: $\alpha_1 = \alpha_2 = \ldots = \alpha_J = 0$

H_A: $\alpha_j \neq 0$ for some j

Main effect for B (filling doesn't matter)

H_0: $\beta_1 = \beta_2 = \ldots = \beta_K = 0$

H_A: $\beta_k \neq 0$ for some k

Interaction effect:

H_0: All $\gamma_{jk} = 0$

H_A: $\gamma_{jk} \neq 0$ for some j, k

Where:

α_j: is the "effect" for factor A at level j

β_k: is the "effect" for factor B at level k

γ_{jk} : is the interaction between level j of factor A, and level k of factor B.

Two-way ANOVA in R with interaction

Source	df	Sum of Sq.	Mean Square	F-stat	p-value
Factor A	K - 1	SSA	MSA = SSA/(K-1)	MSA/MSE	$F_{K-1,KJ(c-1)}$.
Factor B	J - 1	SSB	MSB = SSB/(J-1)	MSB/MSE	$F_{J-1,KJ(c-1)}$
A x B	(K-1)(J-1)	SSAB	MSAB = SSAB/(K-1)(J-1)	MSAB/MSE	$F_{(K-1)(J-1),KJ(c-1)}$
Error	KJ(c - 1)	SSE	MSE = SSE/(K-1)(J-1)		
Total	N - 1	SSTotal			

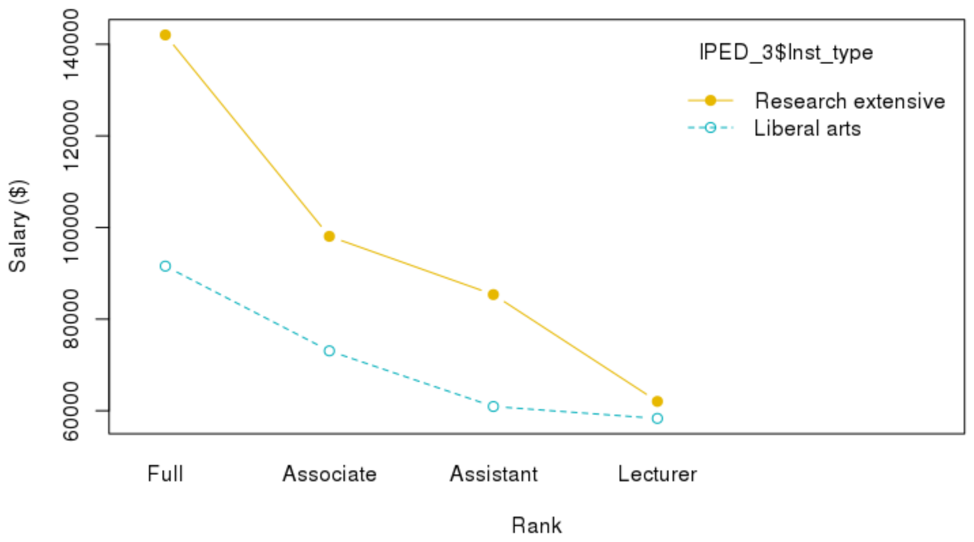
For balanced design: SSTotal = SSA + SSB + SSAB + SSE

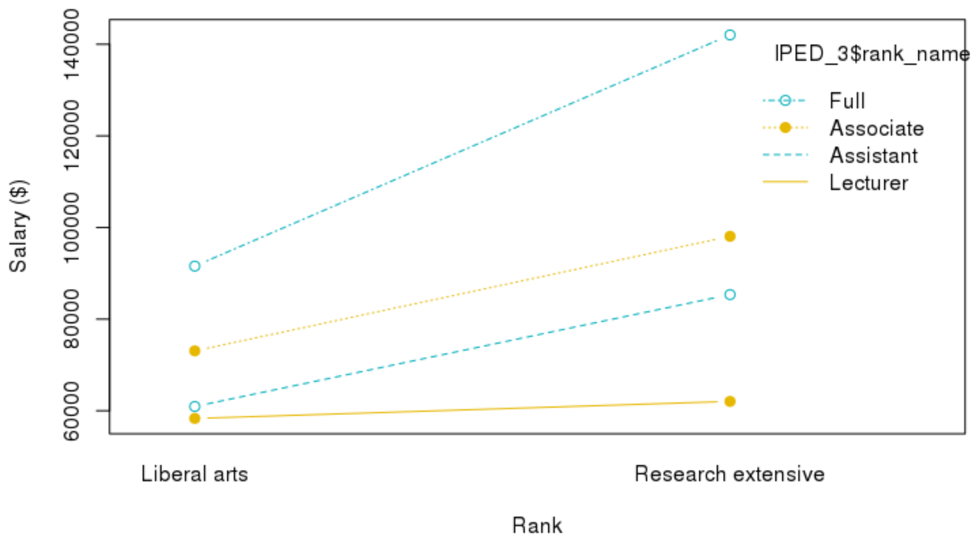
ANOVA table for a balanced design with c replicates in each group

Interaction plots

Interaction plots can help us visualize main effects and interactions

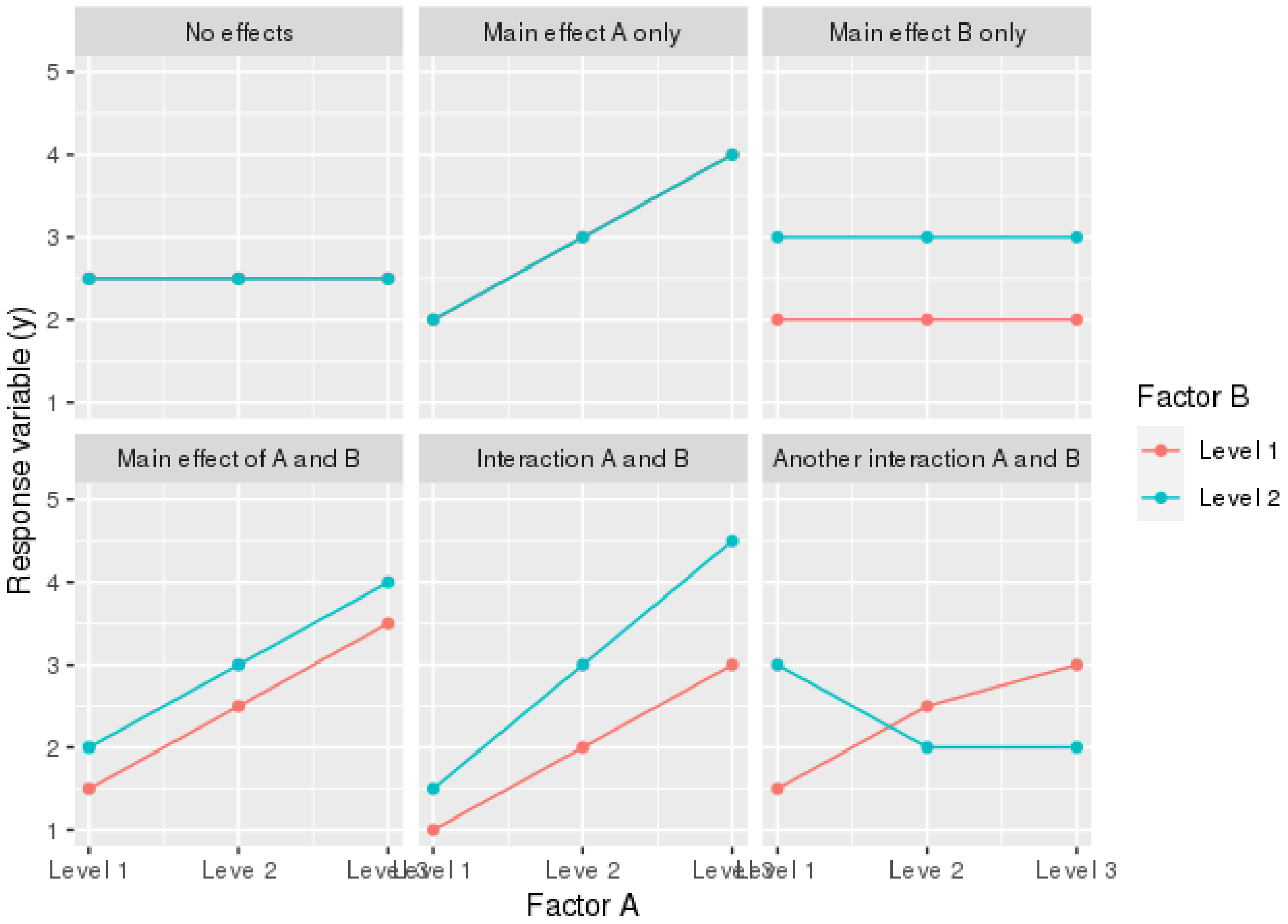
- Plot the levels of one of the factors on the x-axis
- Plot the levels of the other factor as separate lines





Either factor can be on the x-axis although sometimes there is a natural choice

Interpreting interaction plots



Interpreting interactions

When interactions are present, one must be careful interpreting main effects

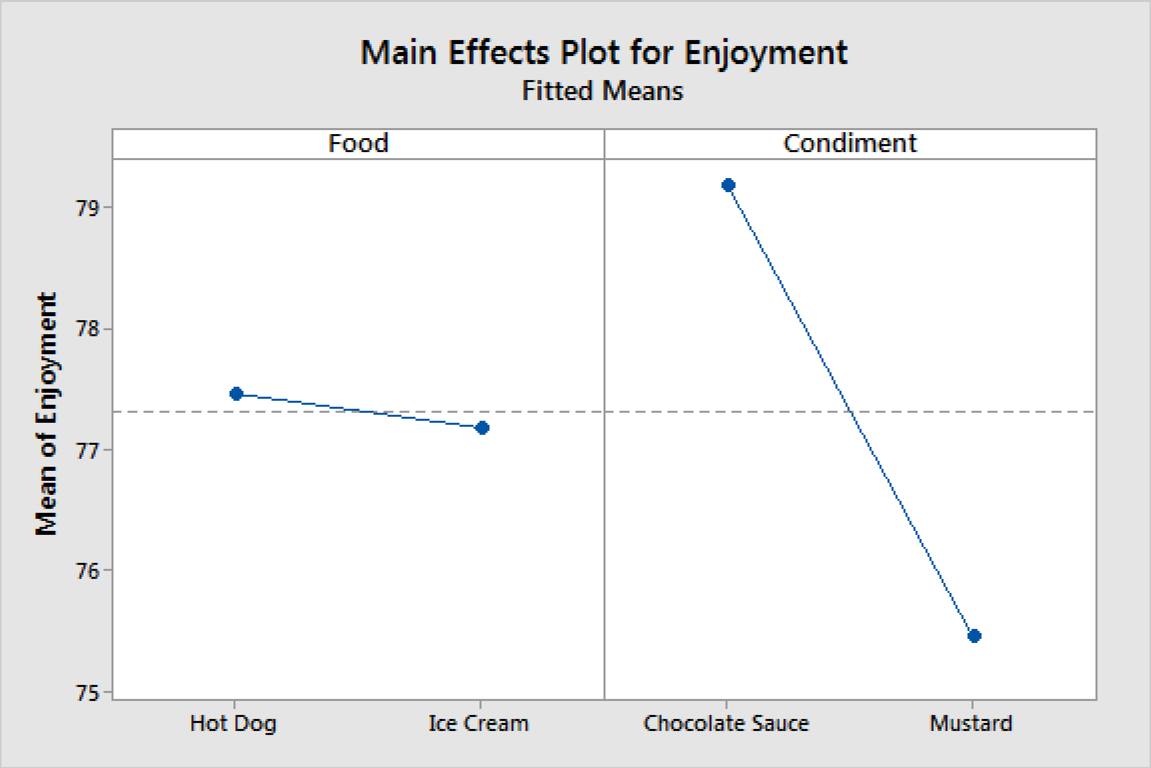
- i.e., the value of one factor A, depends on the value of second factor B

For example, suppose you want to determine which condiment is the most enjoyable

- You might really like chocolate sauce, but your enjoyment will depend on the type of food you are eating

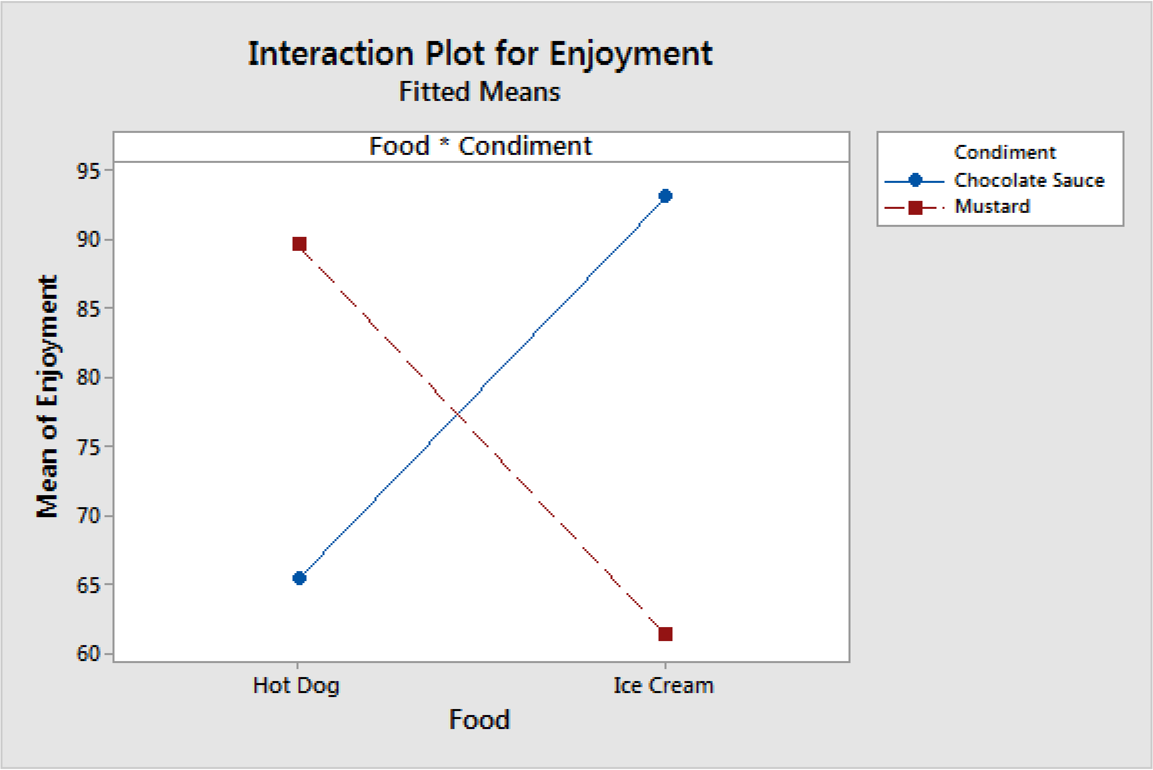
Example taken from [Statistics by Jim]

Interpreting interactions



Example taken from [Statistics by Jim]

Interpreting interactions



Example taken from [Statistics by Jim]

Let’s examine two-way ANOVAs in R…

Complete and balanced designs

Complete factorial design: at least one measurement for each possible combination of factor levels

- E.g., in a two-way ANOVA for factors A and B, if there are K levels for factor A, and J levels for factor B, then there needs to be at least one measurement for each of the KJ levels

Balanced design: the sample size is the same for all combination of factor levels

- E.g., there are the same number of samples in each of the KJ level combinations.
- The computations and interpretations for non-balanced designs are a bit harder.

Unbalanced designs

For unbalanced designs, there are different ways to computer the sum of squares, and hence one can get different p-values

- The problem is analogous to multicollinearity. If two explanatory variables are correlated either can account for the variability in the response data.

Type I sum of squares, (also called sequential sum of squares) the order that terms are entered in the model matters.

- `anova(lm(y ~ A*B))` gives different results than using `anova(lm(y ~ B*A))`
- SS(A) is taken into account before SS(B) is considered etc.

Type III sum of squares, the order that that terms are entered into the model does not matter.

- `Car::Anova(lm(y ~ A*B) , type = "III")` is the same as `car::Anova(lm(y ~ B*A) , type = "III")`
- For each factor, SS(A), SS(B), SS(AB) is taken into account after all other factors are added

Let’s examine it R…