

# Analysis of Variance



# Overview

One-way analysis of variance (ANOVA) concepts and R

Connections between ANOVA and linear regression

Factorial ANOVA

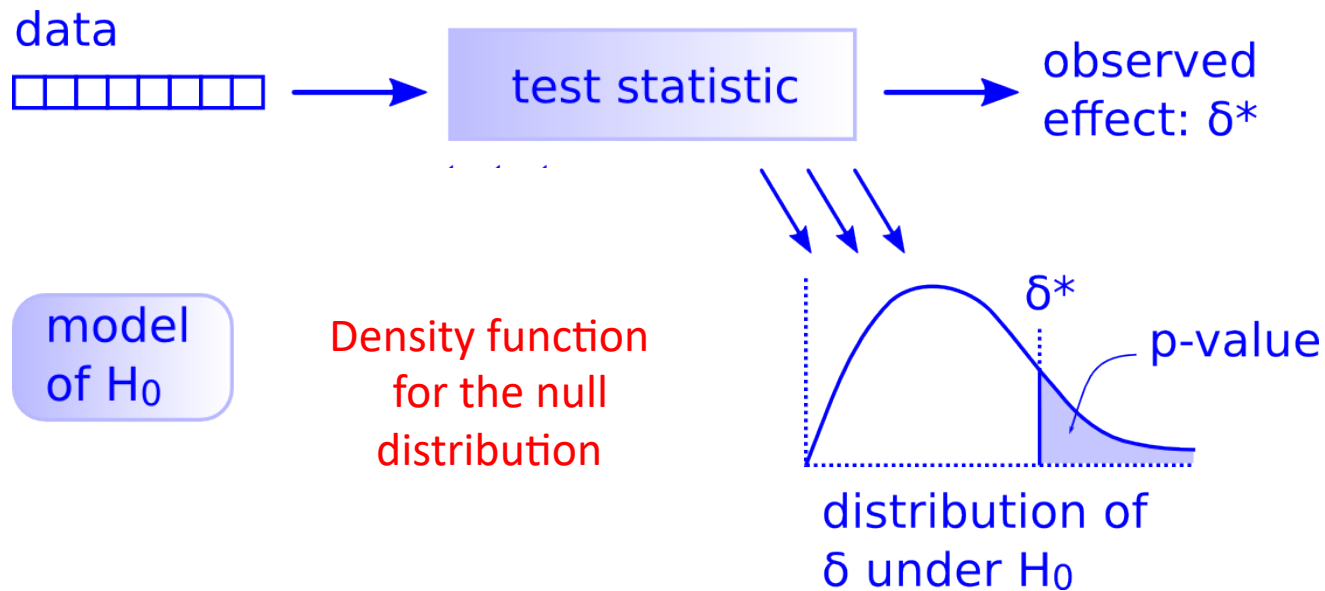
- 2-way ANOVA and interactions

# One-way analysis of variance (ANOVA)

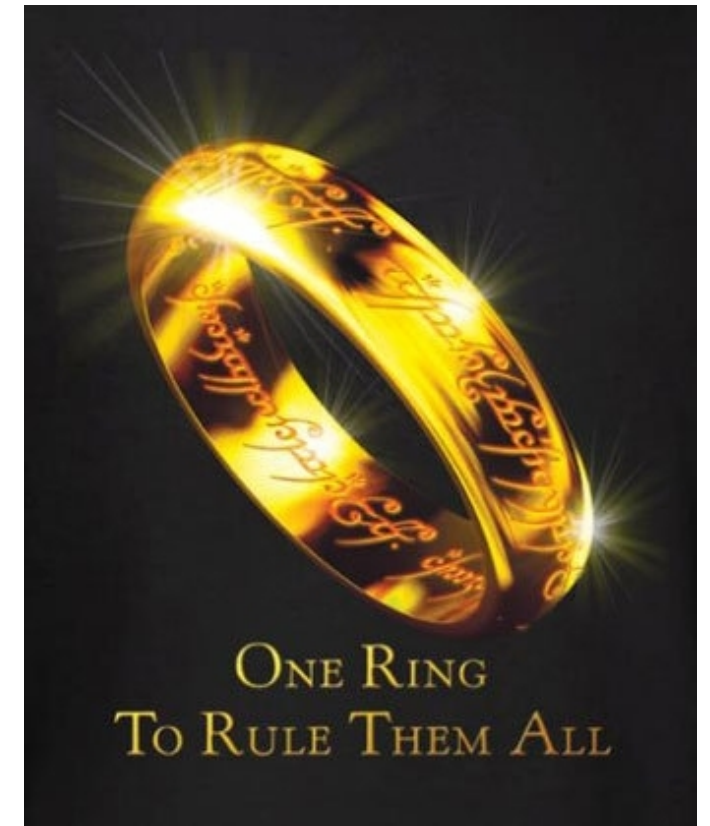
# One-way ANOVA

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.

There is only one [hypothesis test](#)!



Just follow the 5 hypothesis tests steps!



# 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.

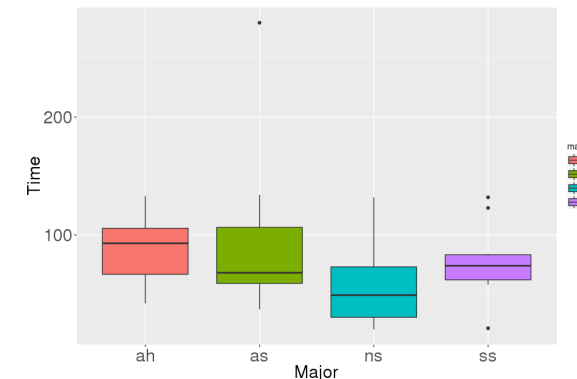
$$H_0: \mu_1 = \mu_2 = \dots = \mu_k$$

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

Q: Have we run a test comparing multiple means yet?

A: Yes! Hope College Sudokus!

	5	3	2	7			8
6		1	5				2
2			9	1	3		5
7	1	4	6	9	2		
	2						6
			4	5	1	2	9
	6		3	2	5		9
1					6	3	4
8			1	9	6	7	



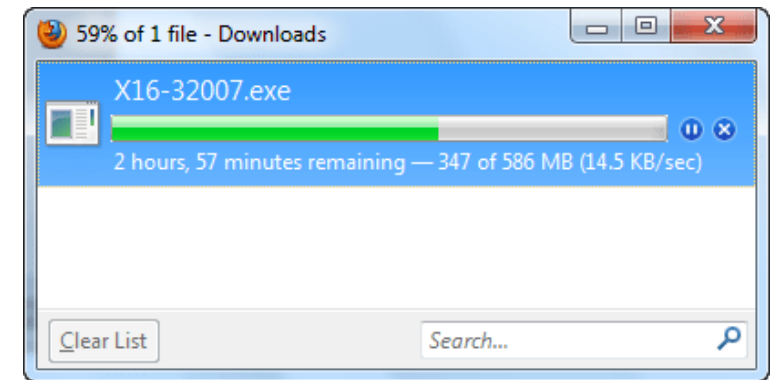
# How does the time of the day affect download speeds?

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.



# 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.

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

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

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

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

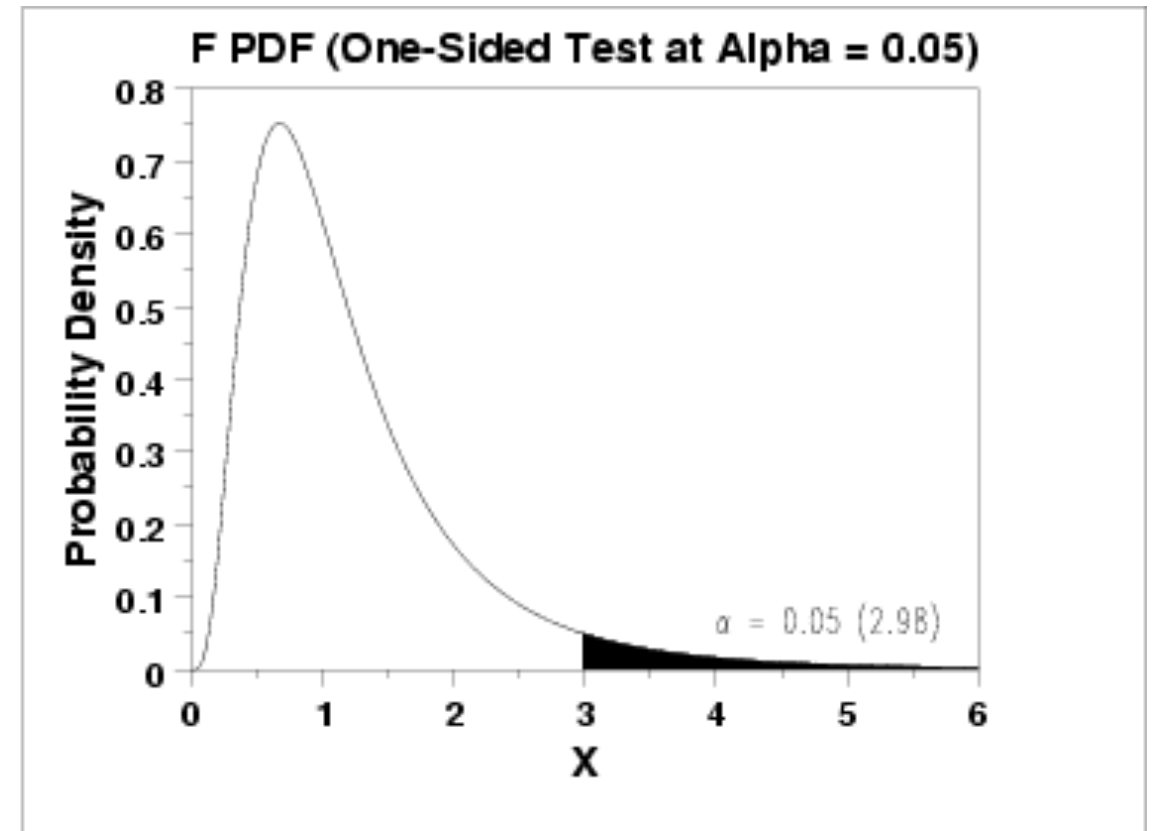
# One-way ANOVA – the central idea

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
- The variances in each group should be approximately equal





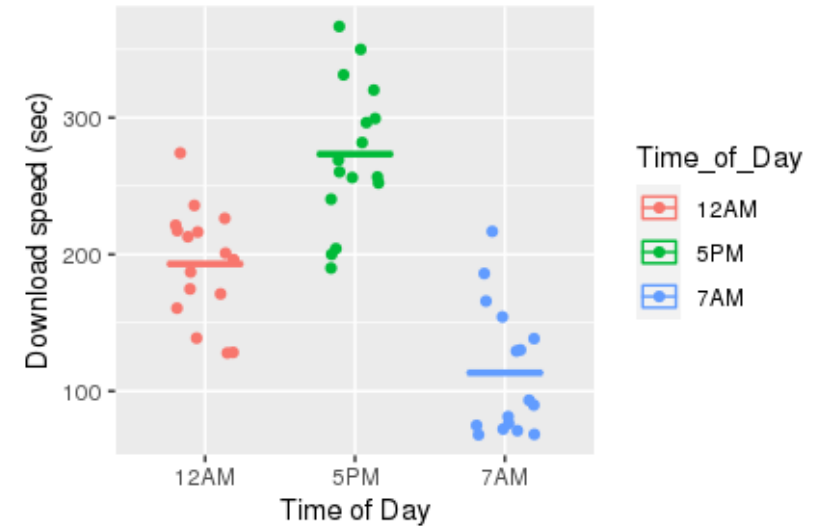
# Checking ANOVA conditions ('assumptions')

1. We can check if the data in each group is relatively normal by visually examining the residuals between each point and its group mean:

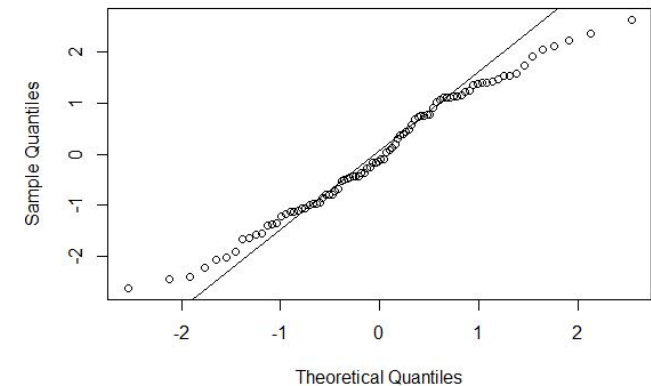
- Residuals as a function of the group mean
- Q-Q plots
- Histograms of residuals

2. We can check the equal variance condition by seeing if the ratio of the largest to smallest standard deviation is greater than 2

- $s_{\max}/s_{\min} < 2$



Normal Q-Q Plot



Note: the one-way ANOVA is fairly robust to violations of these conditions.

# Calculating the observed F-statistic

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

K: the number of groups

N: total number of points

$\bar{x}_{tot}$ : the mean across all the data

$\bar{x}_i$ : the mean of group i

$n_i$ : the number of points in group i

$x_{ij}$ : the  $j^{\text{th}}$  data point from group i

K = 3 different times of day

N = 48 total downloads (16 \* 3)

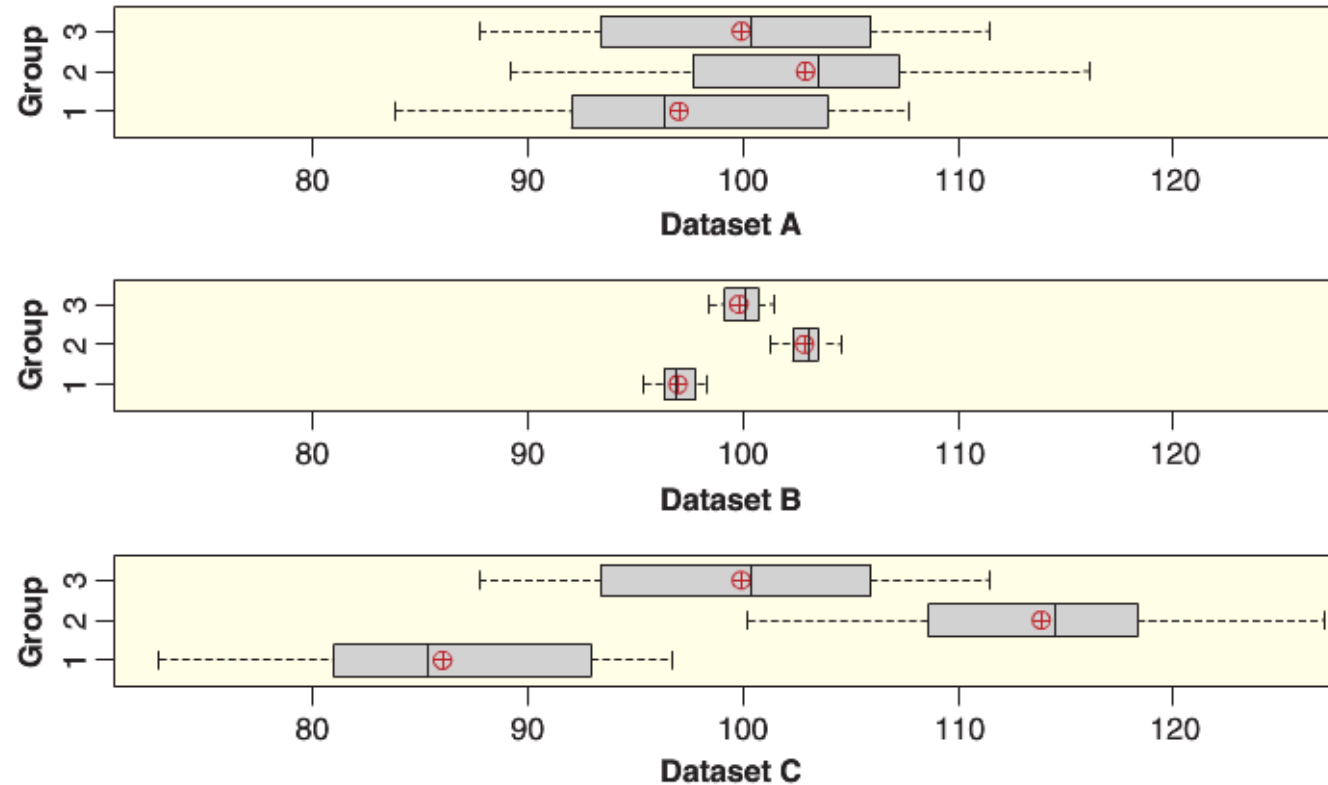
$\bar{x}_{tot}$ : the mean speed across all data

$\bar{x}_i$ : the means for the  $i^{\text{th}}$  time of day

$n_i$  = 16 downloads for each time of day

$x_{ij}$ : the  $j^{\text{th}}$  download at the  $i^{\text{th}}$  time of day

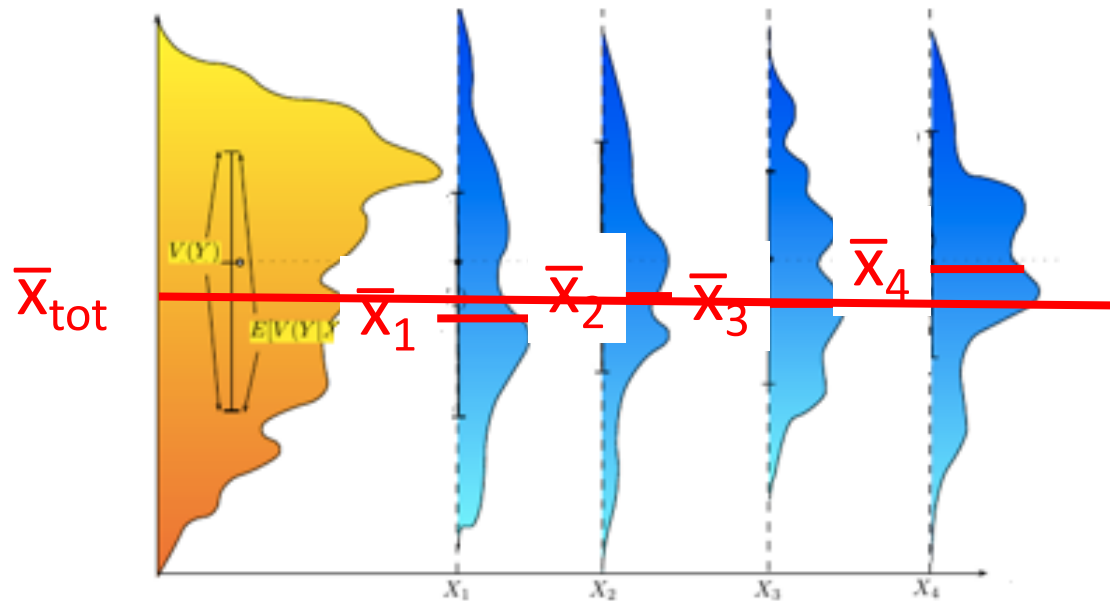
# Why use the F-Statistic?



Which dataset gives the strongest evidence that there is a difference in population means?

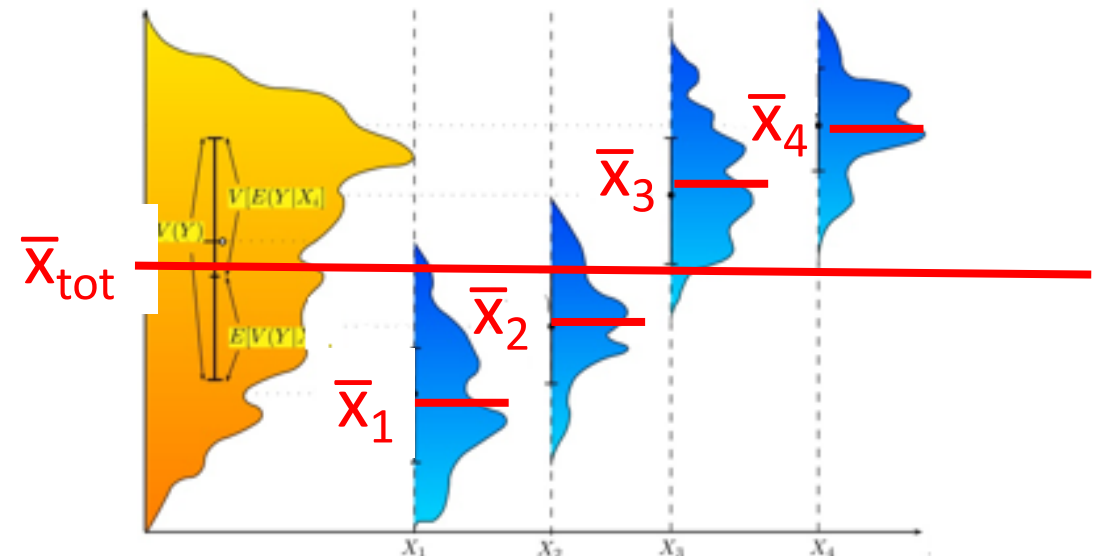
# The F-Statistic

If  $H_0$  is **true**, the data from all groups have **the same means**



- Similar means  $\bar{x}_i$
- Similar spread  $s_i$

If  $H_0$  is **not true**, the data from all groups have **different means**



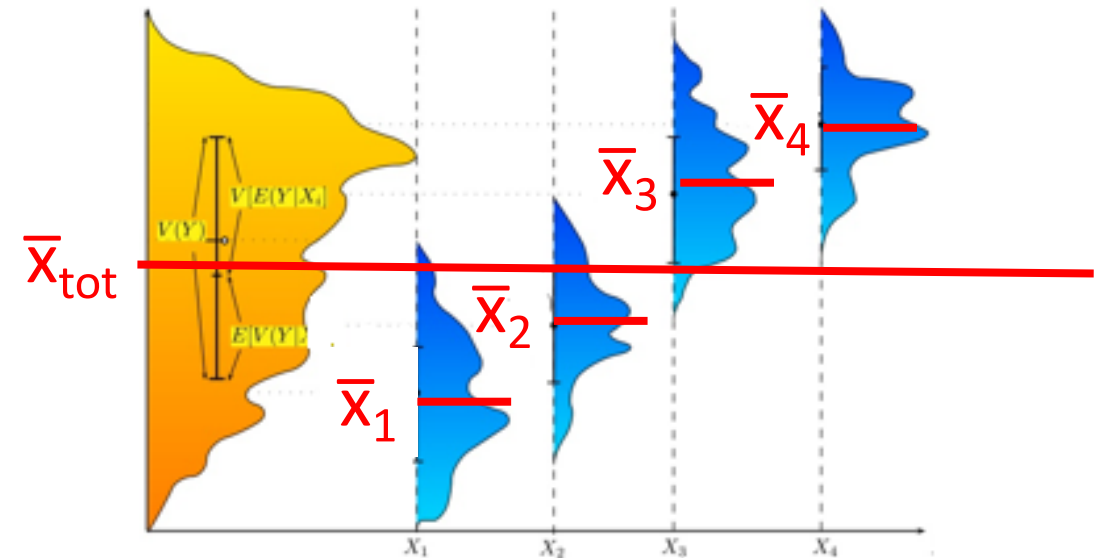
- Different means  $\bar{x}_i$
- Smaller spreads  $s_i$

# 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:

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



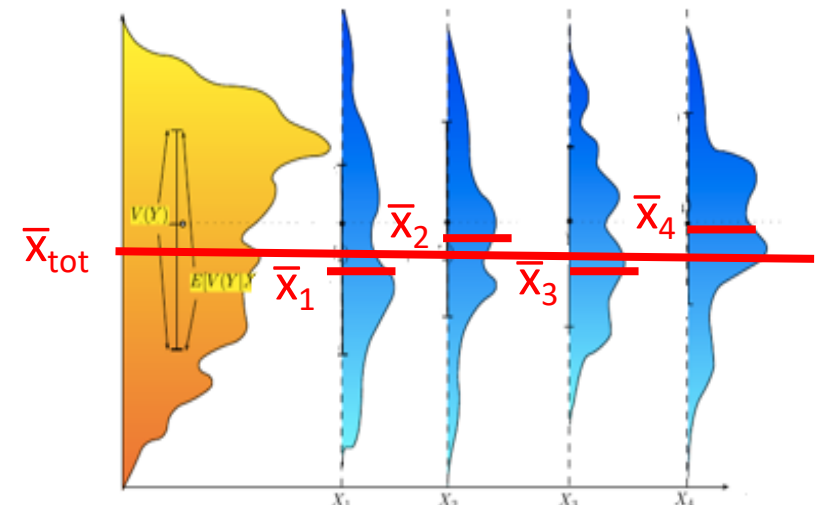
# The F-statistic

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

The F statistic measures a fraction of:

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

$$s_i^2 \approx \sigma^2$$



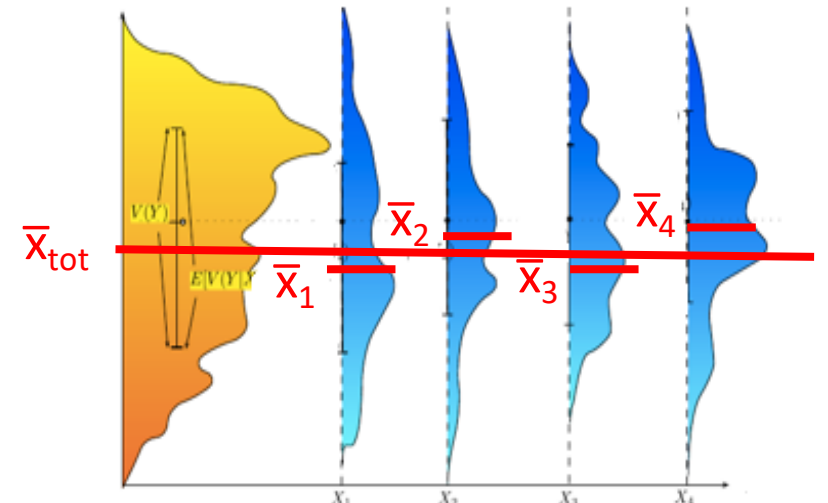
# The F-statistic

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

The F statistic measures a fraction of:

$$F = \frac{\text{variability between group means}}{\approx \sigma^2}$$

$$s_i^2 \approx \sigma^2$$



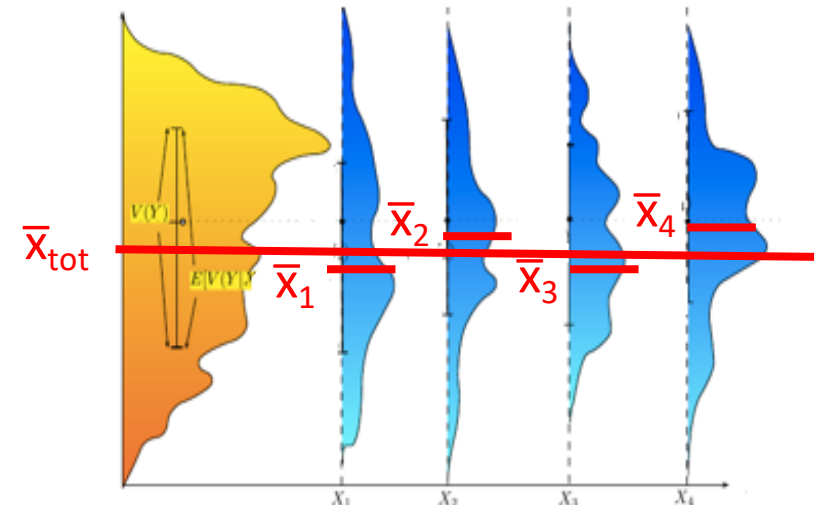
# The F-statistic

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

The F statistic measures a fraction of:

$$F = \frac{\text{variability between group means}}{\approx \sigma^2}$$

SE<sup>2</sup>





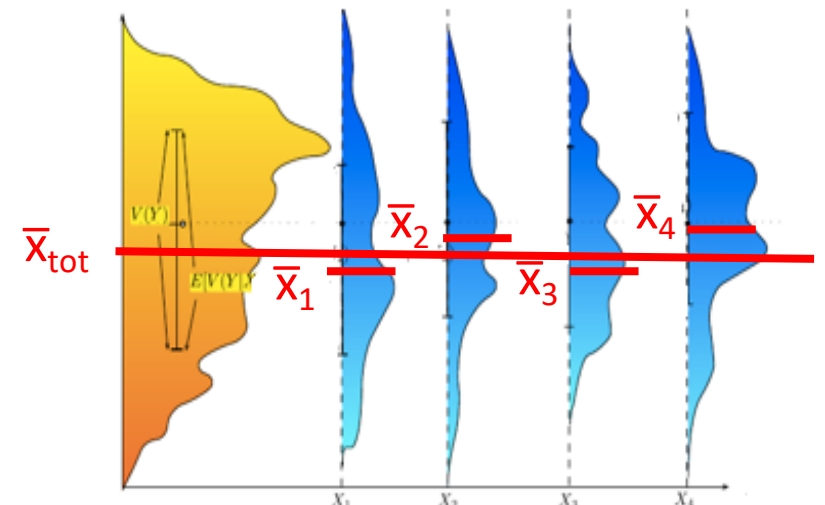
# The F-statistic

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

The F statistic measures a fraction of:

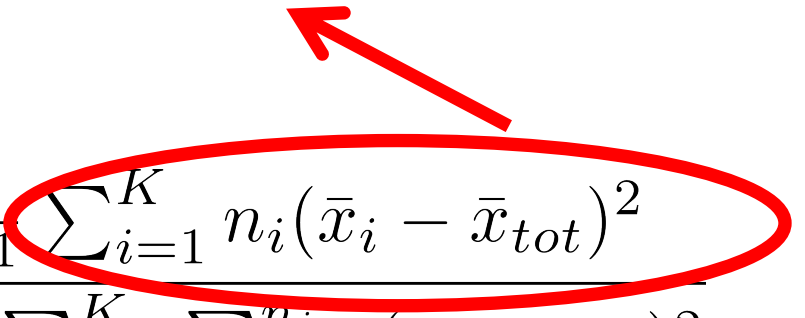
$$F = \frac{\approx \sigma^2}{\approx \sigma^2} \approx 1$$

$$SE^2 \approx \sigma^2/n$$



# The F-statistic

Sum of Squares Group (SSG)

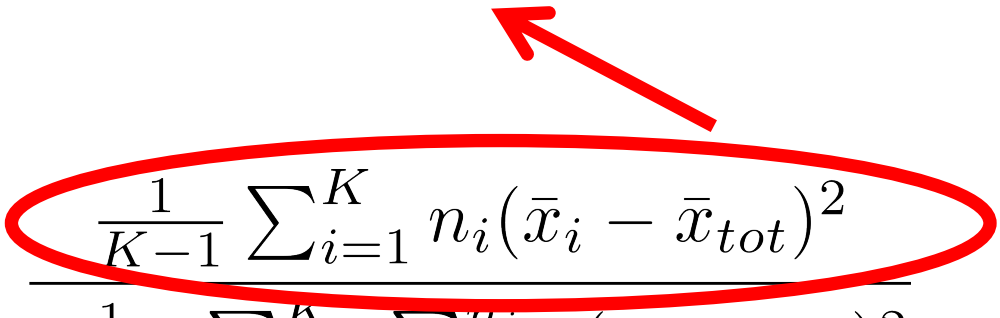
$$F = \frac{\text{between-group variability}}{\text{within-group variability}} = \frac{\frac{1}{K-1} \sum_{i=1}^K n_i (\bar{x}_i - \bar{x}_{tot})^2}{\frac{1}{N-K} \sum_{i=1}^K \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2}$$


The F statistic measures a fraction of:

$$F = \frac{\approx \sigma^2}{\approx \sigma^2} \approx 1$$

# The F-statistic

Mean Squares Group (MSG)

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


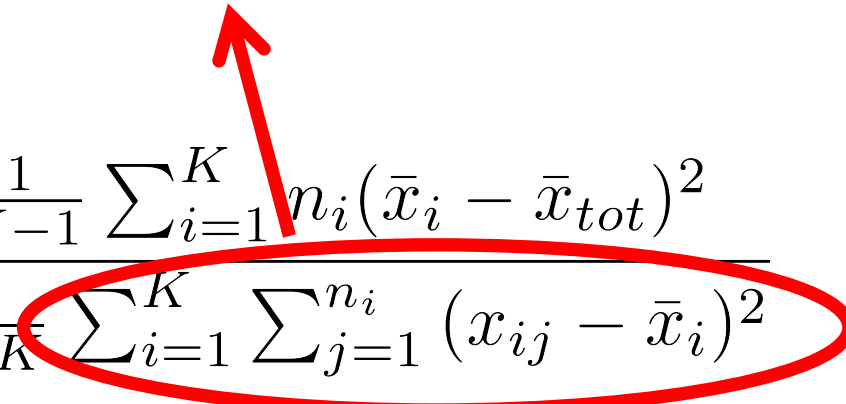
The F statistic measures a fraction of:

$$F = \frac{\text{MSG}}{\approx \sigma^2} \approx 1$$

# The F-statistic

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

Sum of Squares Error (SSE)



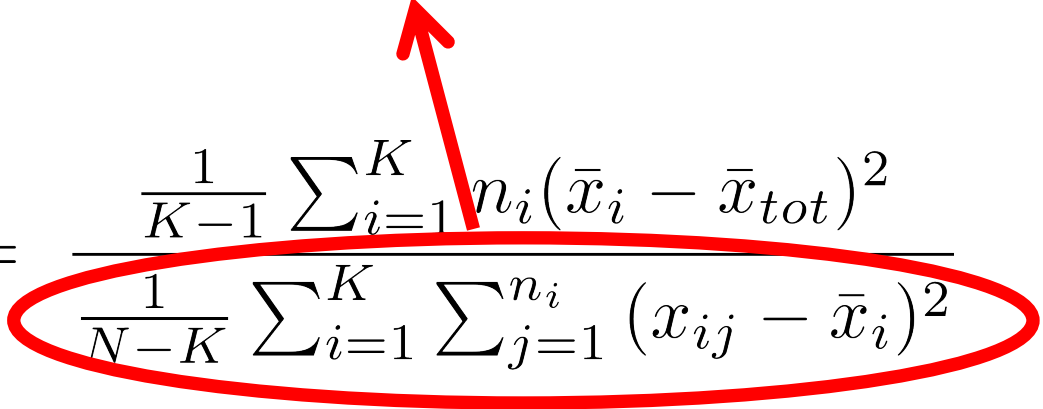
The F statistic measures a fraction of:

$$F = \frac{\text{MSG}}{\approx \sigma^2} \approx 1$$

# The F-statistic

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

Mean of Squares Error (MSE)



The F statistic measures a fraction of:

$$F = \frac{\text{MSG}}{\text{MSE}} \approx 1$$

Same as what we saw in regression:  $SSTotal = SSG + SSE$

# ANOVA table

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:

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

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

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

# ANOVA table

Just as we saw for linear regression, we have the relationship:

$$SST = SSG + SSE$$

Where:

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

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

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

# Running a one-way ANOVA

Step 1: State the null and alternative hypothesis

Step 2: Calculate the F-statistic on using actual data

Step 3: Create the appropriate F-distribution

Step 4: Calculate the p-value

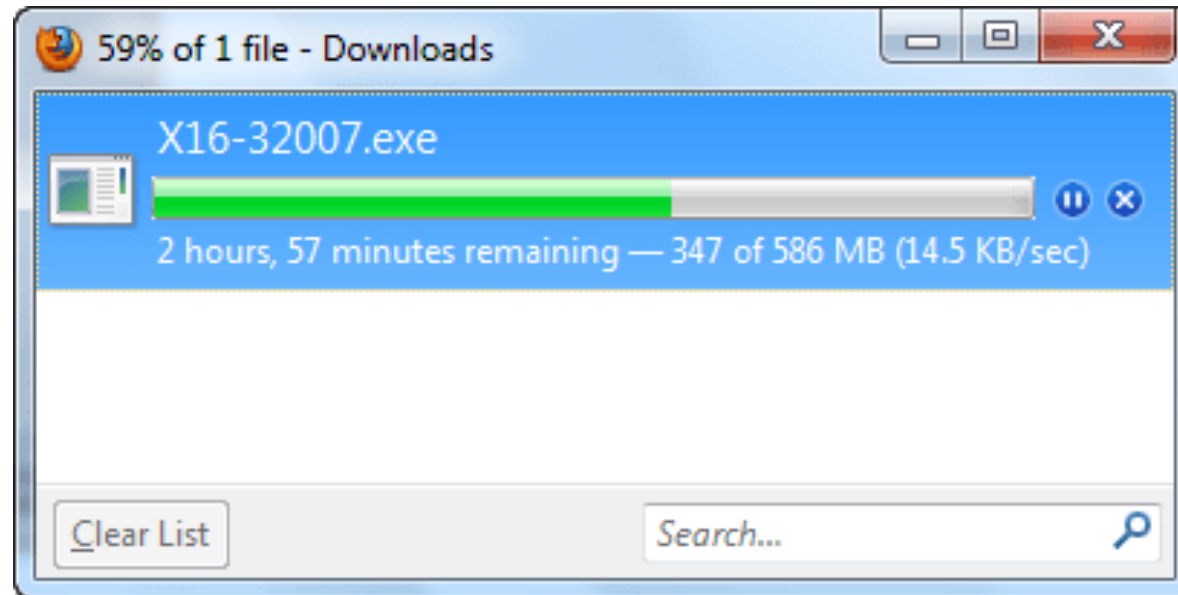
Step 5: Make a decision

Check our underlying assumptions  
were met

Two red arrows originate from the text 'Check our underlying assumptions were met'. One arrow points diagonally upwards and to the left towards 'Step 2: Calculate the F-statistic on using actual data'. The other arrow points diagonally upwards and to the left towards 'Step 4: Calculate the p-value'.



# Let's try it out in R!



# Connections regression and categorical ANOVA

# ANOVA as regression with only categorical predictors

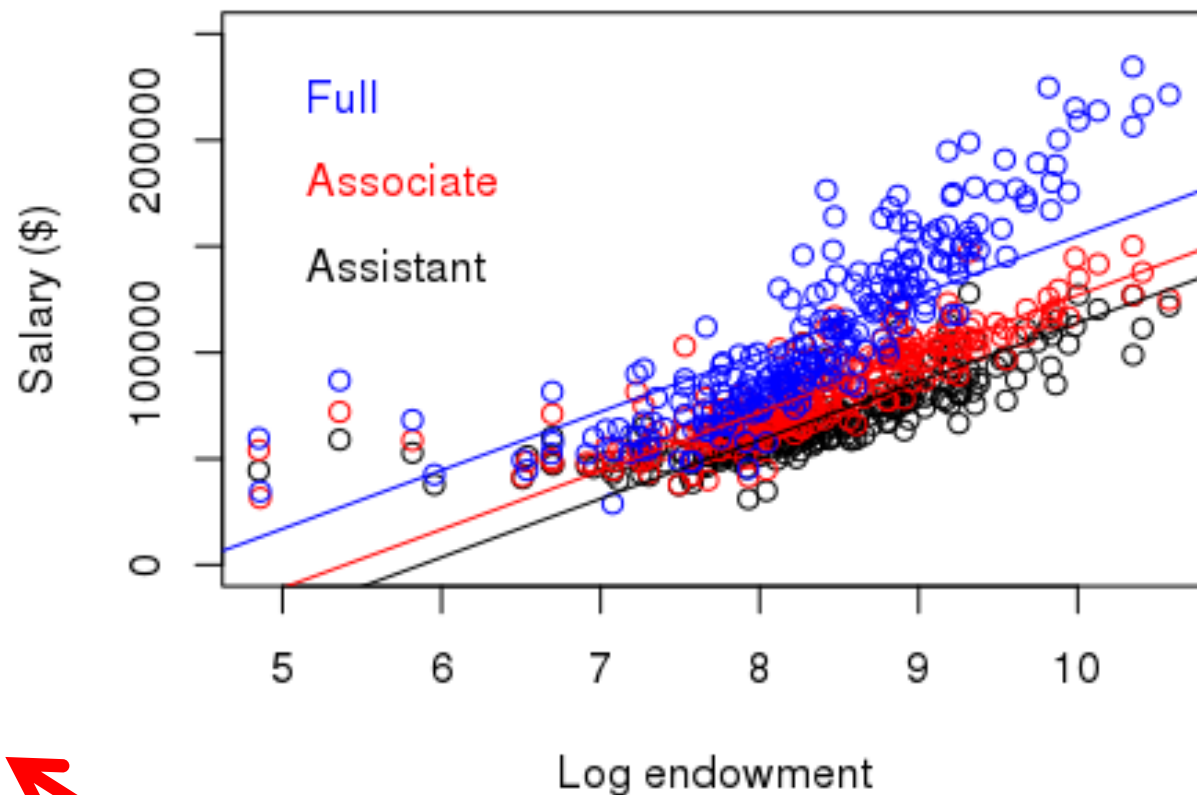
Recall we can have categorical predictors with k levels in a regression model by using k -1 dummy variables:

- e.g., we would need two dummy variables to have different intercepts for Assistant, Associate and Full Professors

$$x_{i1} = \begin{cases} 1 & \text{if Assistant Professor} \\ 0 & \text{if Full Professor} \end{cases} \quad x_{i2} = \begin{cases} 1 & \text{if Associate Professor} \\ 0 & \text{if Full Professor} \end{cases}$$

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

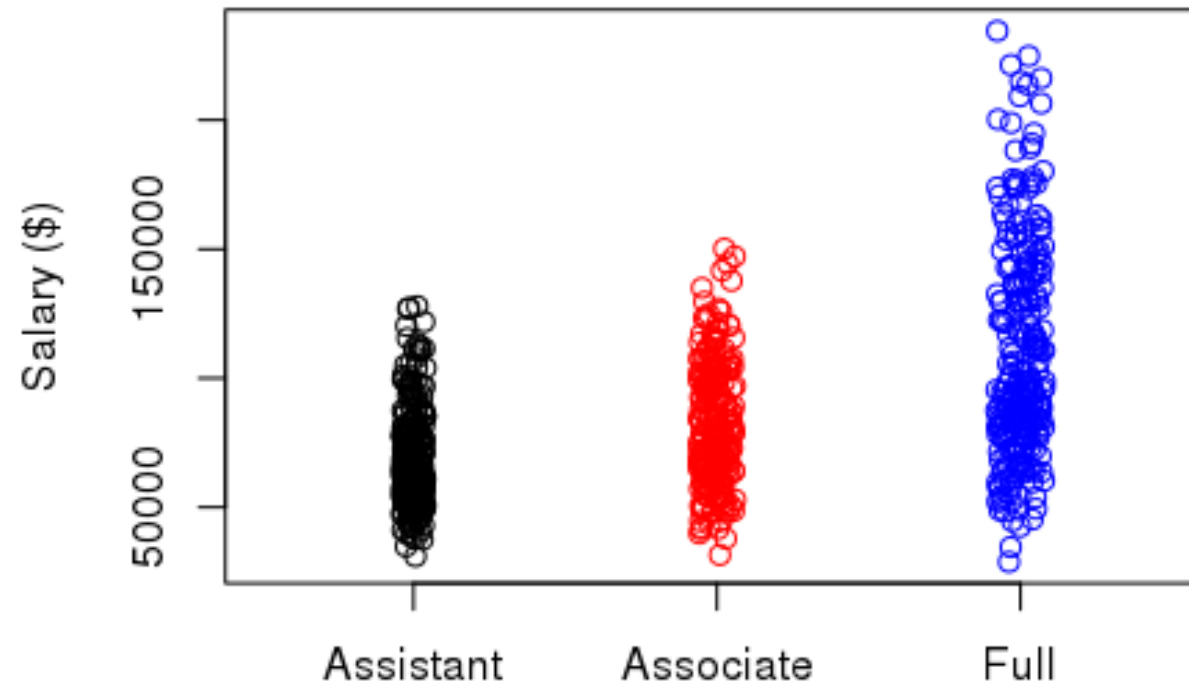
# ANOVA as regression with only categorical predictors



Common slope for  
log endowment

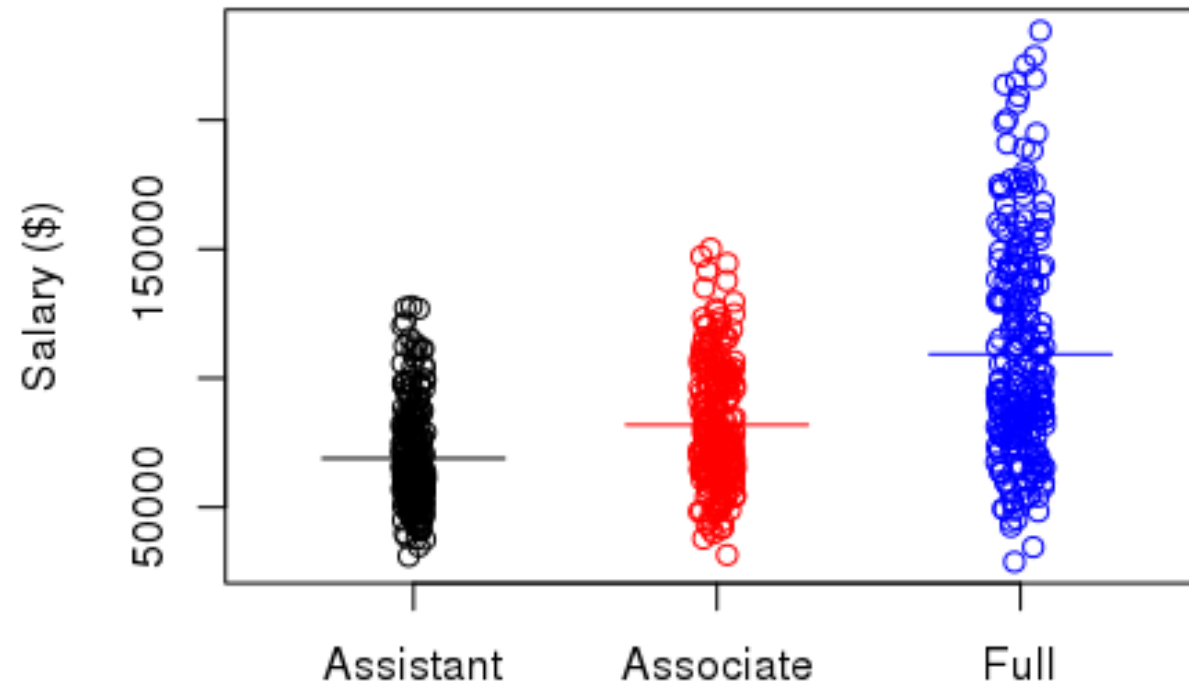
$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \beta_3 x_i + \epsilon_i & \text{if Assistant Professor} \\ \beta_0 + \beta_2 + \beta_3 x_i + \epsilon_i & \text{if Associate Professor} \\ \beta_0 + \beta_3 x_i + \epsilon_i & \text{if Full Professor} \end{cases}$$

# ANOVA as regression with only categorical predictors



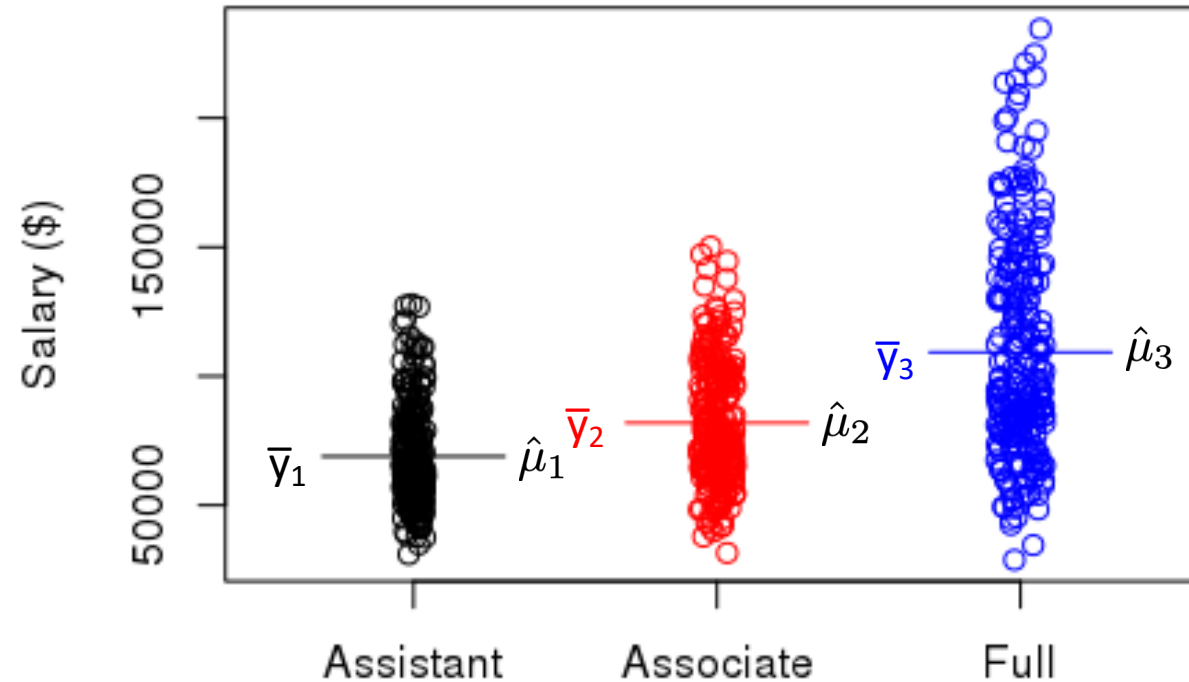
$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if Assistant Professor} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if Associate Professor} \\ \beta_0 + \epsilon_i & \text{if Full Professor} \end{cases}$$

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



$$y_i = \mu_k + \epsilon_i = \begin{cases} \mu_1 + \epsilon_i & \text{if Assistant Professor} \\ \mu_2 + \epsilon_i & \text{if Associate Professor} \\ \mu_3 + \epsilon_i & \text{if Full Professor} \end{cases}$$

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



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

# ANOVA decomposition

$$F = \frac{\frac{1}{K-1} \sum_{i=1}^K n_i (\bar{y}_i - \bar{y}_{tot})^2}{\frac{1}{N-K} \sum_{i=1}^K \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2}$$

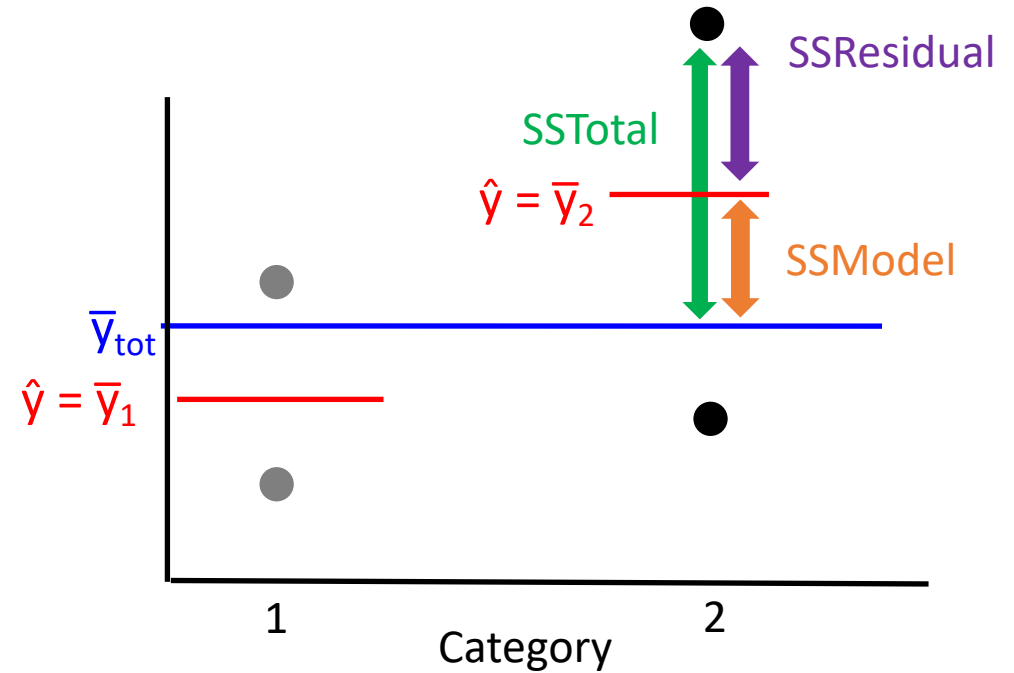
The ANOVA decomposes the variance as:

- $SS_{Total} = SS_{Model} + SS_{Residual}$

$$y_{ij} - \bar{y}_{tot} = (\hat{y}_{ij} - \bar{y}_{tot}) + (y_{ij} - \hat{y}_{ij})$$

$$(y_{ij} - \bar{y}_{tot})^2 = (\hat{y}_{ij} - \bar{y}_{tot})^2 + (y_{ij} - \hat{y}_{ij})^2$$

$$(y_{ij} - \bar{y}_{tot})^2 = (\bar{y}_i - \bar{y}_{tot})^2 + (y_{ij} - \bar{y}_i)^2$$



$\hat{y}_{ji} = \bar{y}_i$   
(the prediction for each class is the group mean)



Let's examine these relationships in R...

# Factorial ANOVA

In a **one-way ANOVA** we model the response variable  $y$  as a function of **one** categorical predictor

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

- Analogous to extending simple linear regression to multiple regression

# Factorial ANOVA

For example, we could use a **two-way ANOVA** to assess how salaries differ for:

a. Faculty ranks:

- 1. Full
- 2. Associate
- 3. Assistant
- 4. Lecturer

b. Institution types:

- 1. Extensive research institutions
- 2. Liberal arts colleges

This would be called a 4 x 2 design

- 2 factors: the first has 4 levels and the second has 2 levels

# Writing a one-way ANOVA in terms of effects

For a **one-way ANOVA**, we can state the null and alternative hypotheses as:

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

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

We can also write values as our response variable  $y$  as:

$$y_i = \mu_k + \varepsilon_i$$

$$y_i = \mu + \alpha_k + \varepsilon_i$$

Where:

$\mu$ : is the overall mean

$\alpha_k$ : is the “effect” for level  $k$

# Writing a one-way ANOVA in terms of effects

For a **one-way ANOVA**, we can state the null and alternative hypotheses as:

$$H_0: \alpha_1 = \alpha_2 = \dots = \alpha_k = 0$$

$$H_A: \alpha_i \neq 0 \text{ for some } i$$

We can also write values as our response variable  $y$  as:

$$y_i = \mu_k + \varepsilon_i$$

$$y_i = \mu + \alpha_k + \varepsilon_i$$

Where:

$\mu$ : is the overall mean

$\alpha_k$ : is the “effect” for level  $k$

# Two-way ANOVA hypotheses

Our model for a two-way ANOVA is:  $y_i = \mu + \alpha_j + \beta_k + \varepsilon_i$

For a **two-way ANOVA** (without interactions)  
there are two sets of hypotheses we can assess:

Main effect for A:

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

$$H_A: \alpha_j \neq 0 \text{ for some } j$$

Where:

$\alpha_j$ : is the “effect” for factor A at level j

$\beta_k$ : is the “effect” for factor B at level k

Main effect for B:

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

$$H_A: \beta_k \neq 0 \text{ for some } k$$

# Two-way ANOVA in R

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

ANOVA table for a balanced design with 1 replicate in each group

# Interactions

We can also examine whether there is an interaction between rank and institution type

- i.e., does the difference in salaries between faculty ranks differ across institution types?

This is similar to using the same slope vs. different slopes model for an interaction between a quantitative and categorical variable

$$y_i = \mu + \alpha_j + \beta_k + \gamma_{jk} + \varepsilon_i$$



# Two-way ANOVA hypotheses

## Main effect for A:

$$H_0: \alpha_1 = \alpha_2 = \dots = \alpha_j = 0$$

$$H_A: \alpha_j \neq 0 \text{ for some } j$$

Where:

$\alpha_j$ : is the “effect” for factor A at level j

$\beta_k$ : is the “effect” for factor B at level k

## Main effect for B:

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

$$H_A: \beta_k \neq 0 \text{ for some } k$$

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

## Interaction effect:

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

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

# Two-way ANOVA in R with interaction

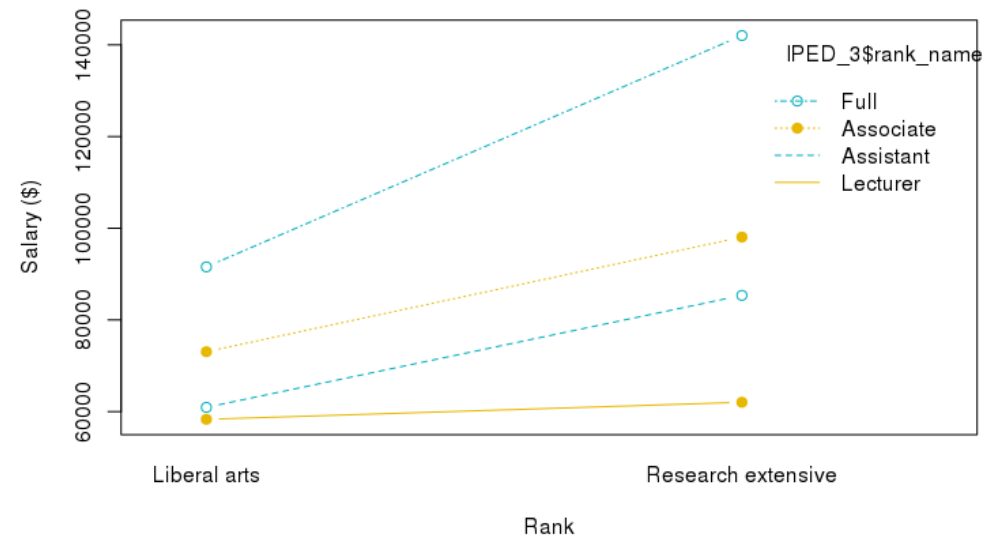
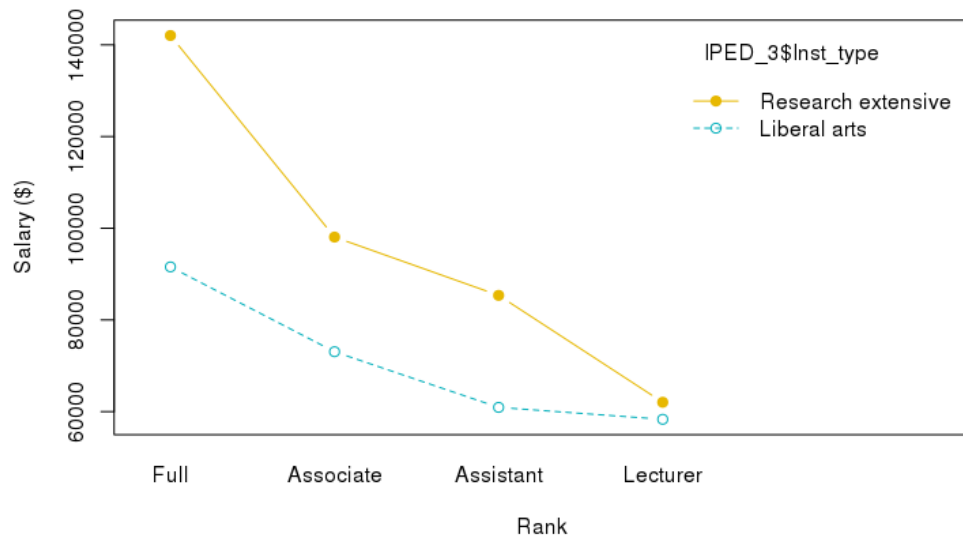
Source	df	Sum of Sq.	Mean Square	F-stat	p-value
Factor A	K - 1	SSA	$MSA = SSA/(K-1)$	$MSA/MSE$	$F_{K-1, KJ(c-1)}$
Factor B	J - 1	SSB	$MSB = SSB/(J-1)$	$MSB/MSE$	$F_{J-1, KJ(c-1)}$
A x B	(K-1)(J-1)	SSAB	$MSAB = SSAB/(K-1)(J-1)$	$MSAB/MSE$	$F_{(K-1)(J-1), KJ(c-1)}$
Error	KJ(c - 1)	SSE	$MSE = SSE/(K-1)(J-1)$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

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

# Interaction plots

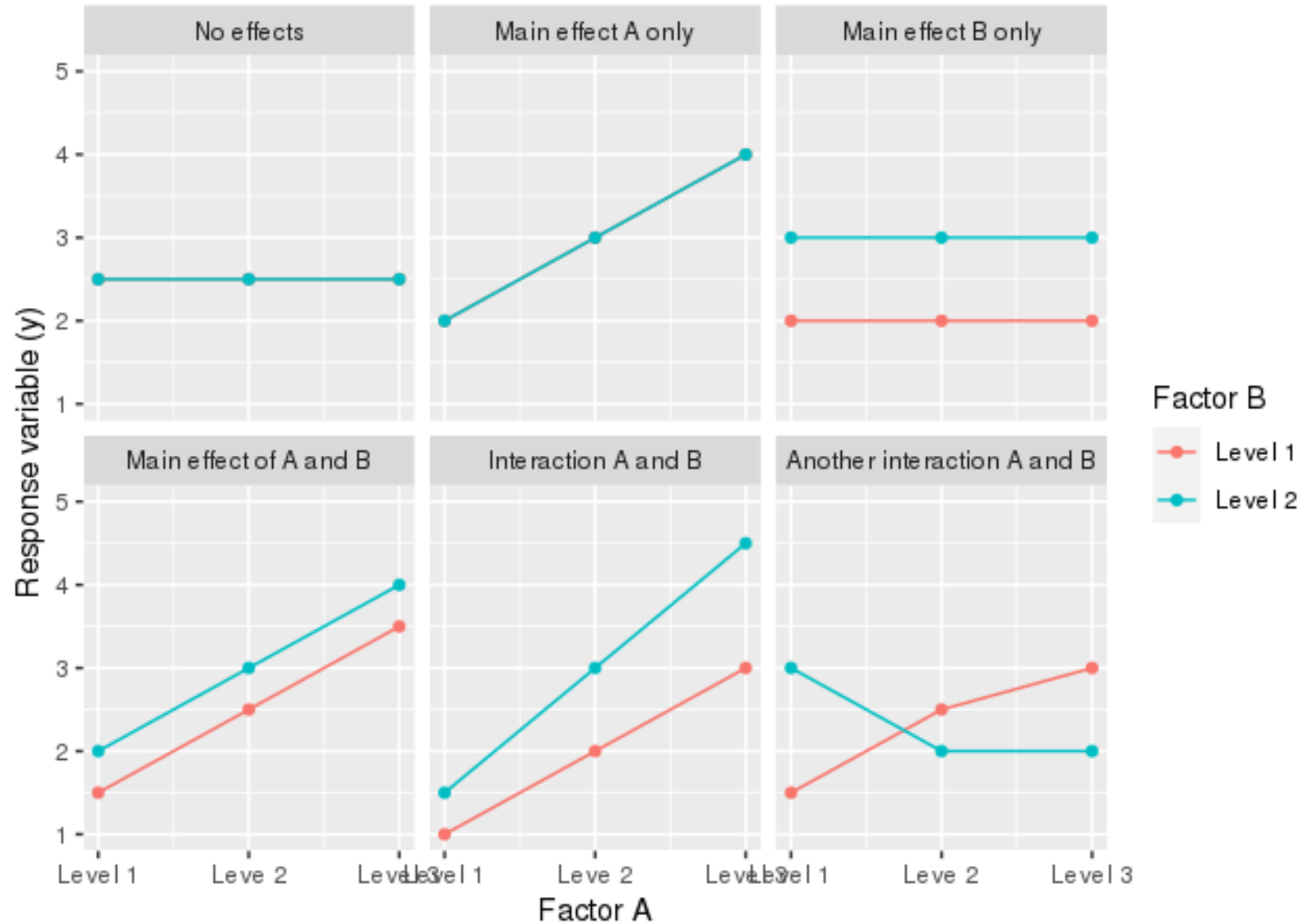
Interaction plots can help us visualize main effects and interactions

- 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



# 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.

Let's examine this in R...