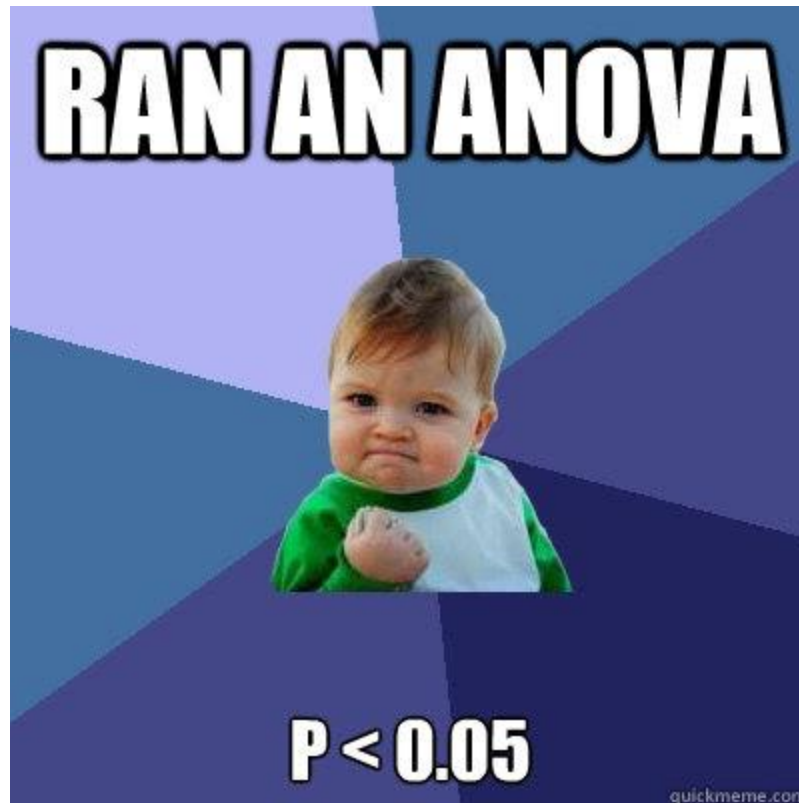


# Analysis of Variance



# Overview

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

Connections between ANOVA and linear regression

Paired tests after running an ANOVA

If there is time: explanation of experimental data used for homework

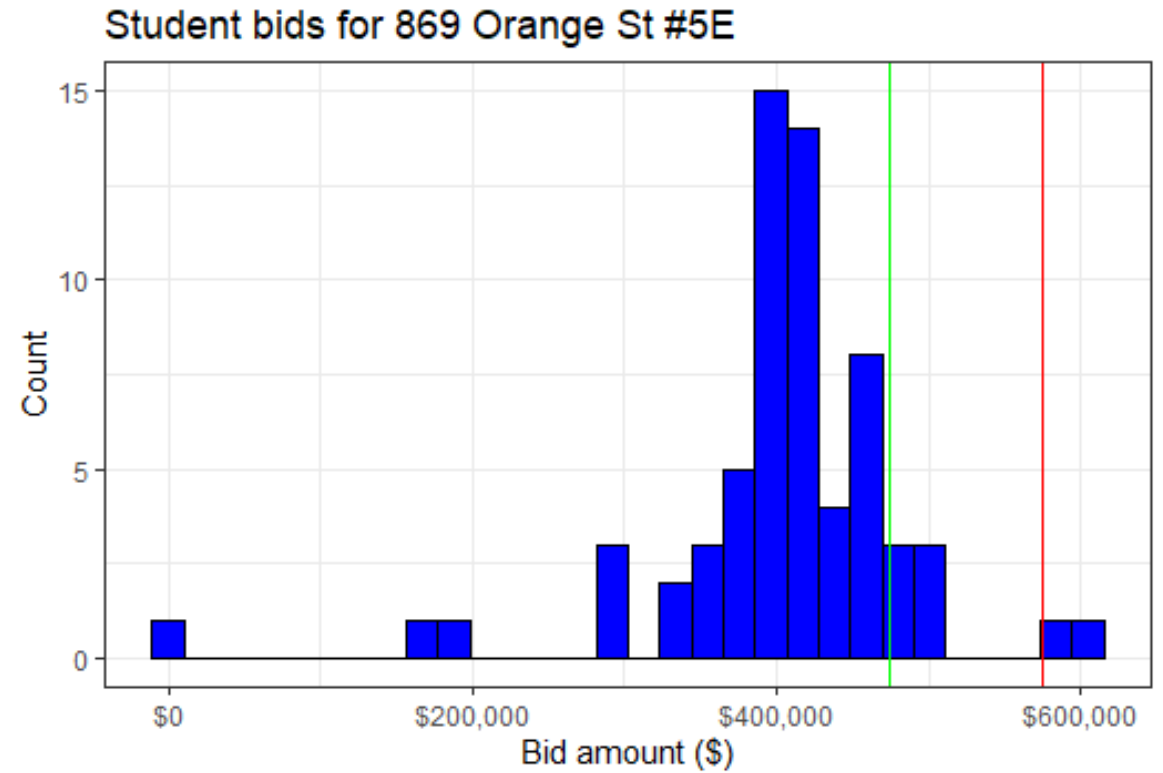
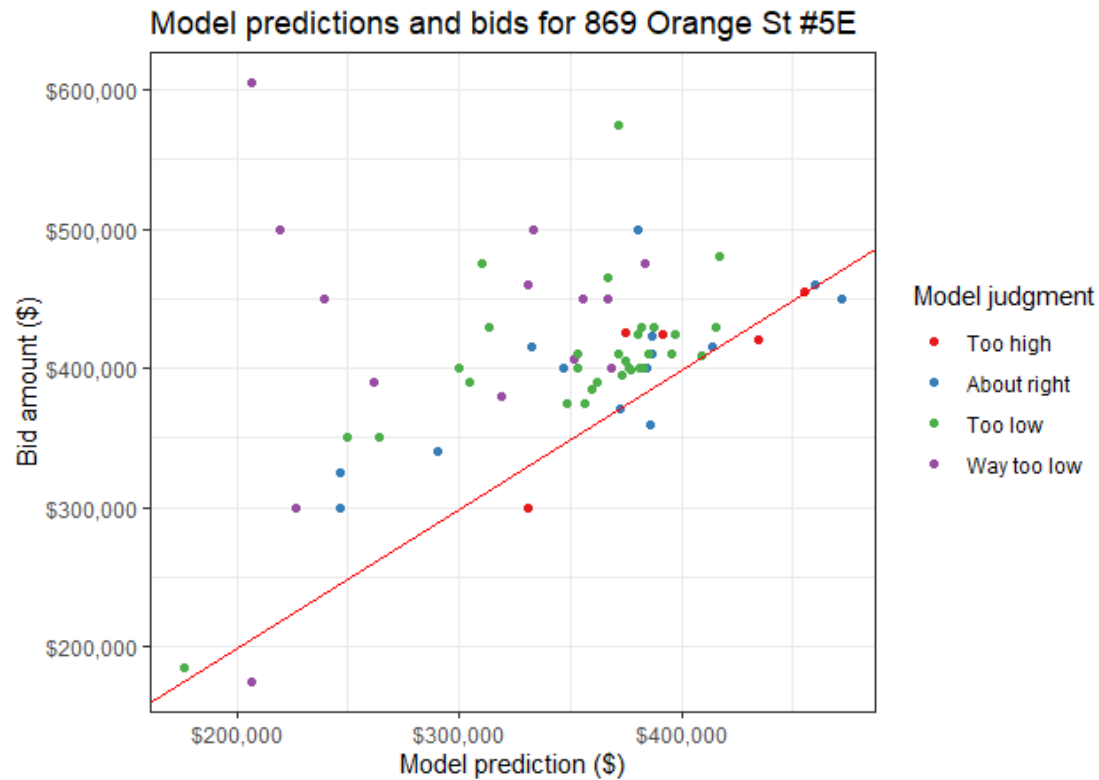
# Any questions about the final project?

Note: you can submit a collaborative project with one other person

Project should be twice as long and twice as impressive!

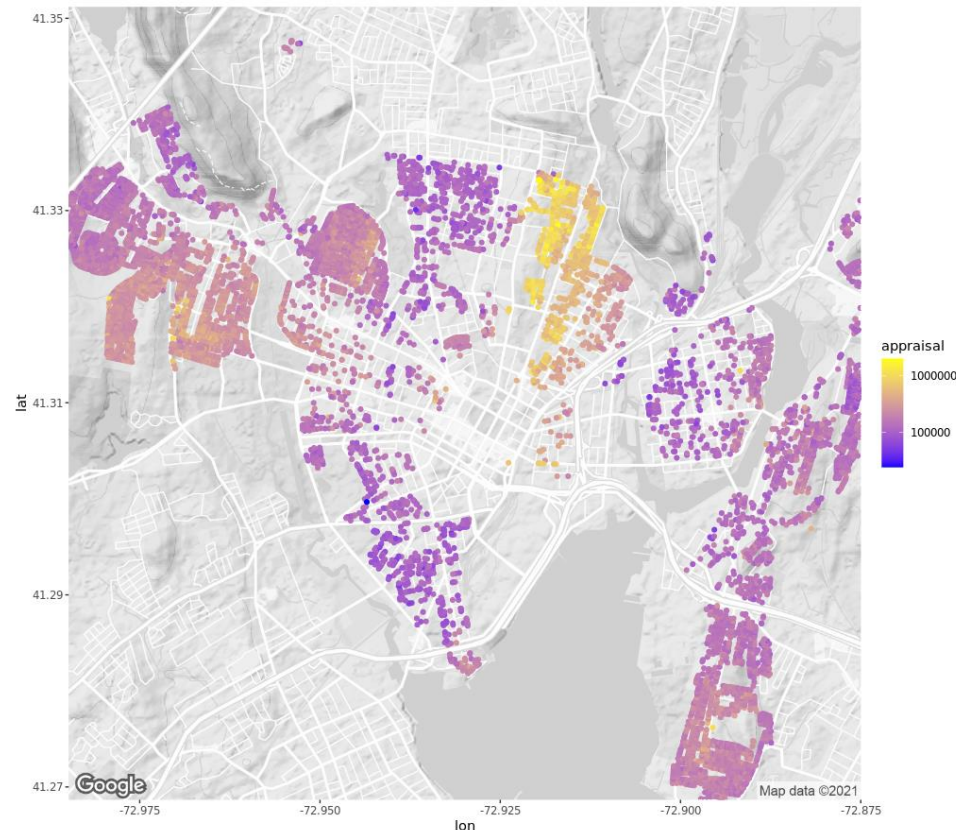
- i.e., 10-16 pages, more in depth analyses, etc.

# How did predicting condo process go?



# How did predicting condo process go?

Map of single-family house prices by location in New Haven





THE WALL STREET JOURNAL.

SUBSCRIBE

SIGN IN

BUSINESS | EARNINGS

## Zillow's Shuttered Home-Flipping Business Lost \$881 Million in 2021

Real-estate company says in a letter to shareholders that it is targeting revenue of \$5 billion by 2025

By [Will Parker](#) [Follow](#)

Updated Feb. 10, 2022 6:24 pm ET

[Share](#) [Resize](#)

[Listen](#) (1 min) [More](#)

The company shocked the market in November by [announcing it was closing Zillow Offers](#) because the [tech-powered platform failed](#) to accurately predict movements in home prices.

# Quick review of Poisson regression

# Quick review of Poisson regression

In Poisson regression we model counts

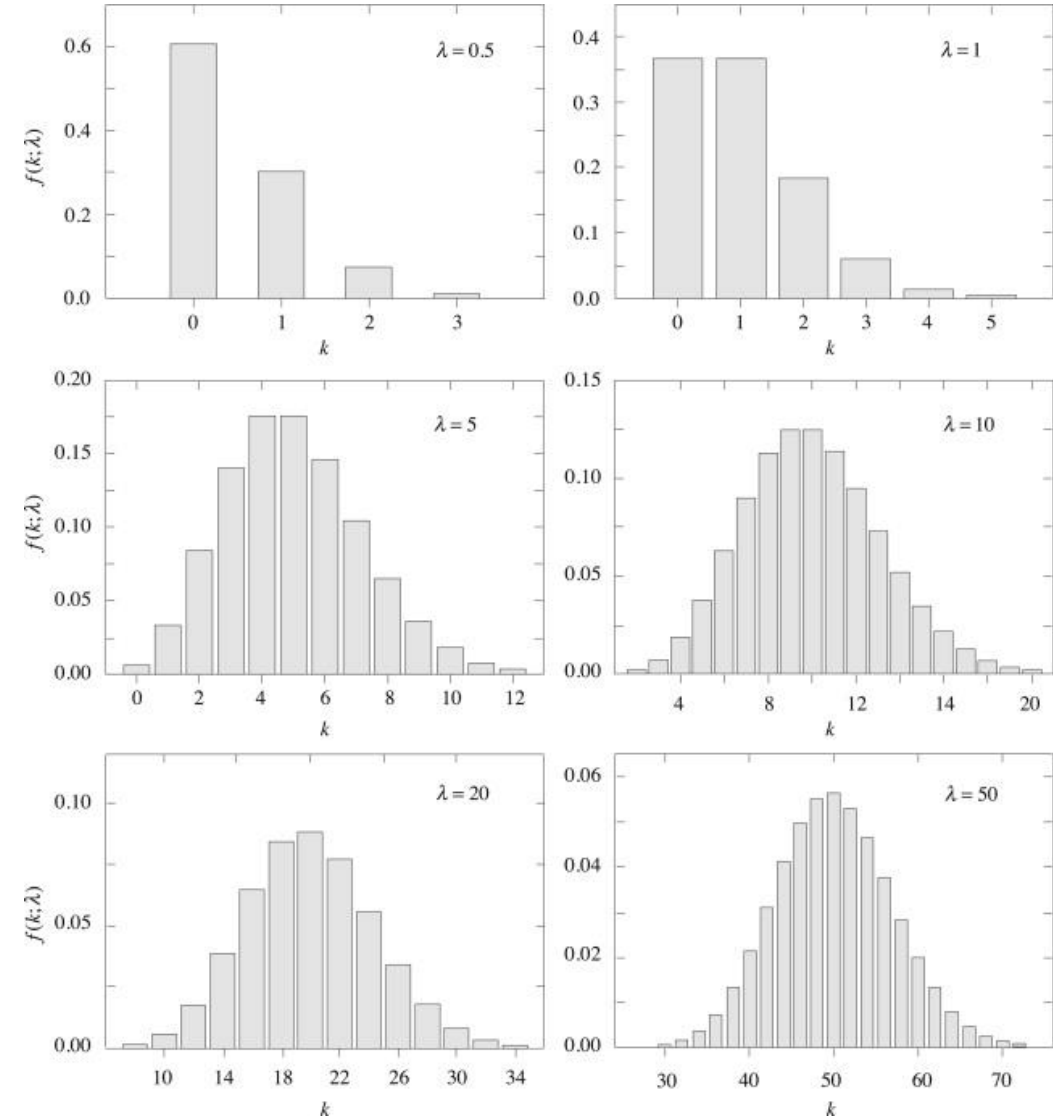
- i.e., integer values:  $k = 0, 1, 2, 3, \dots$
- The underlying probability model is a Poisson distribution

Poisson distributions are parametric distributions that have one parameter  $\lambda$

$$P(X = k|\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$$

We model the  $\log(\lambda)$  as a function of other explanatory variables  $x_i$

$$\log(\lambda) = \beta_0 + \beta_1 \cdot x \quad \lambda = e^{\beta_0 + \beta_1 \cdot x}$$



# Quick review of Poisson regression

Last class we used Poisson regression to assess if Roy had a tendency to say f#ck more when:

- He was **coaching** and/or
- He was **dating** Keeley



```
Call:
glm(formula = F_count_RK ~ Dating_flag + Coaching_flag, family = poisson,
    data = richmondway)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	1.72835	0.12625	13.690	< 2e-16 ***
Dating_flagYes	0.05337	0.11678	0.457	0.648
Coaching_flagYes	0.64330	0.13057	4.927	8.36e-07 ***

---  
signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1



# Quick review generalized linear model

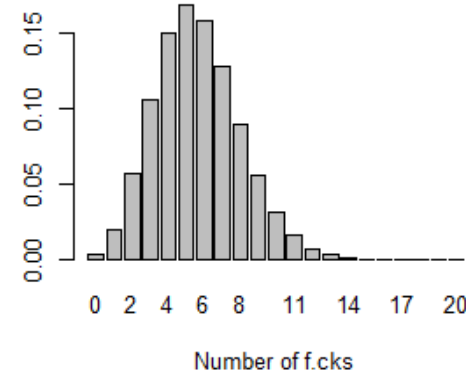
We can calculate the  $\lambda$ 's for each condition

```
# calculate the rate parameters lambda
the_coefs <- coef(glm_fit)

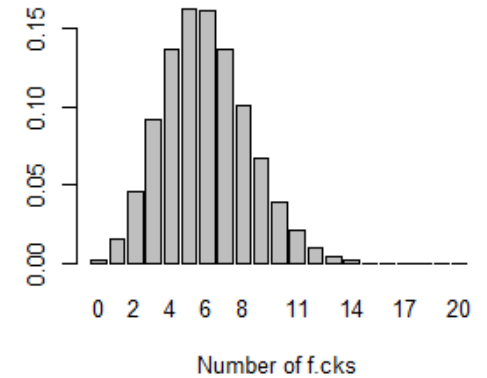
(lambda_baseline <- exp(the_coefs[1]))
(lambda_dating <- exp(the_coefs[1] + the_coefs[2]))
(lambda_coaching <- exp(the_coefs[1] + the_coefs[3]))
(lambda_both <- exp(the_coefs[1] + the_coefs[2] + the_coefs[3]))
```

We can then visualize the different Poisson distributions for each condition

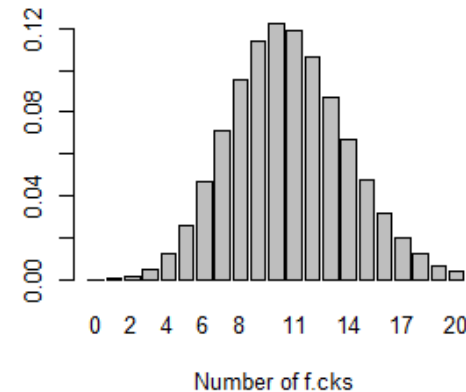
No dating or coaching ( $\lambda = 5.63$ )



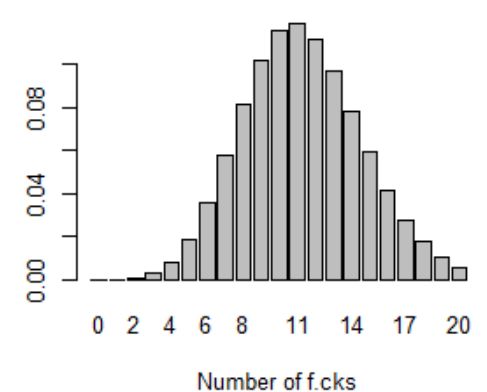
Dating only ( $\lambda = 5.94$ )



Coaching only ( $\lambda = 10.72$ )



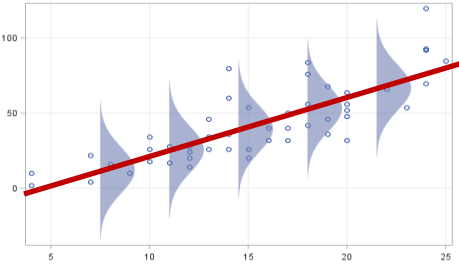
Dating and coaching ( $\lambda = 11.3$ )



# Quick Review: Generalized linear model

Logistic and Poisson regression are both ***generalized linear models***.  
They generalize the linear model to:

- Y values are not restricted to come from a normal distribution
  - So the y-values do not need to be real-values
- A linear prediction is put through an “inverse link” function to estimate the parameters of a probability model

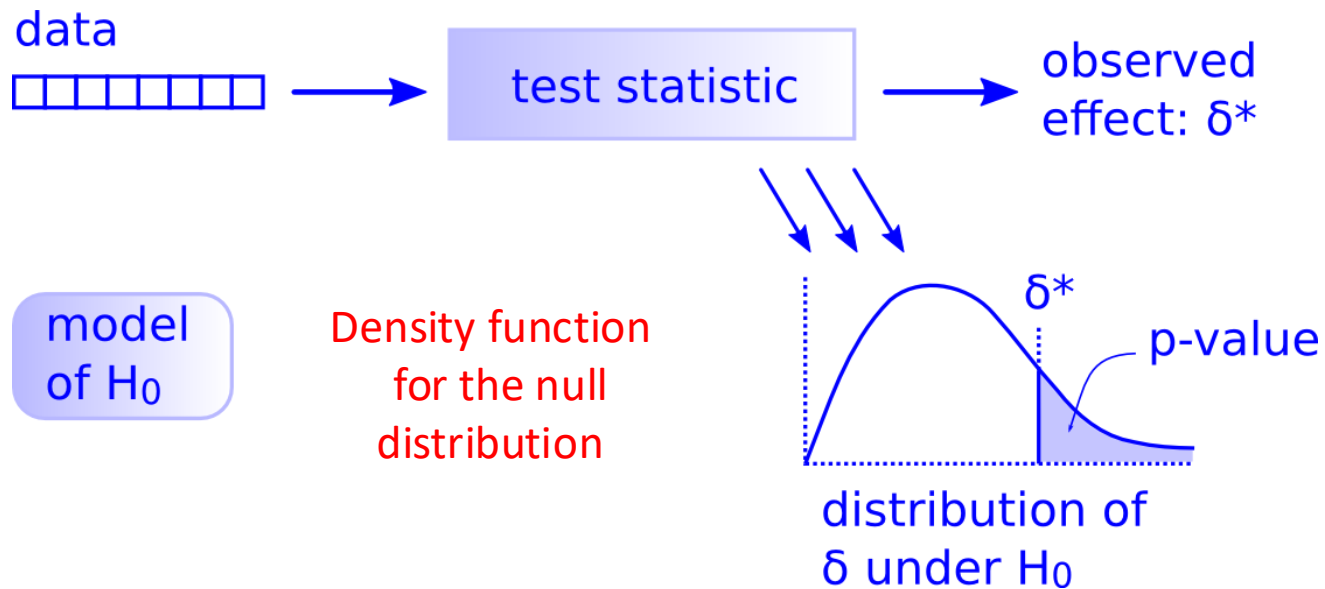
	Inverse link function	Probability model	
Linear regression	$\mu = \beta_0 + \beta_1 \cdot x$	$Y_i \sim N(\mu, \sigma_\epsilon)$	
Logistic regression	$P(Y = 1 x) = \frac{e^{\beta_0 + \beta_1 \cdot x}}{1 + e^{\beta_0 + \beta_1 \cdot x}}$	$Y_i \sim Bernoulli(\pi = P(Y = 1 x))$	
Poisson regression	$\lambda = e^{\beta_0 + \beta_1 \cdot x_1}$	$Y_i \sim Poisson(\lambda)$	

One-way analysis of variance (ANOVA)

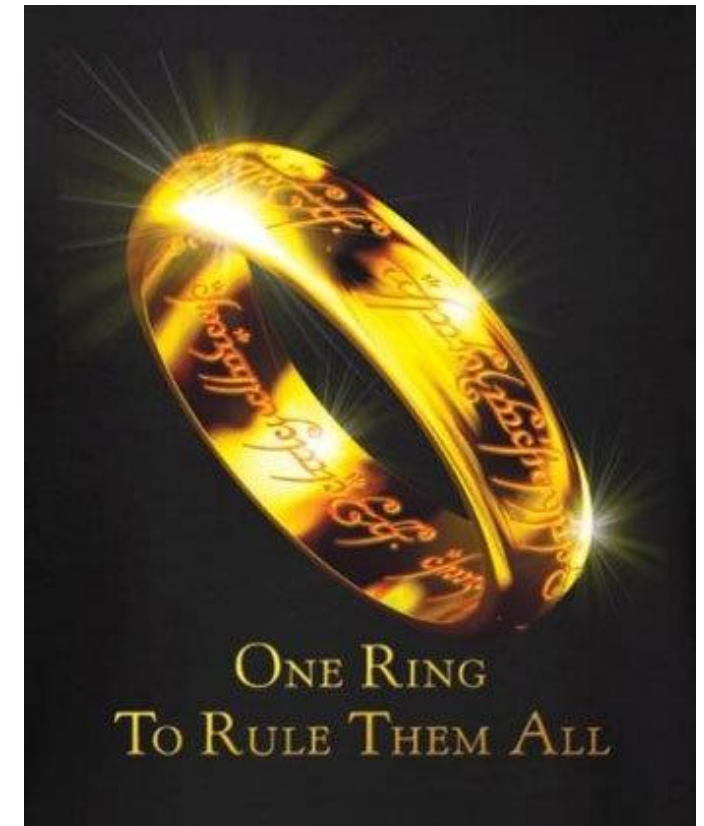
# One-way ANOVA

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

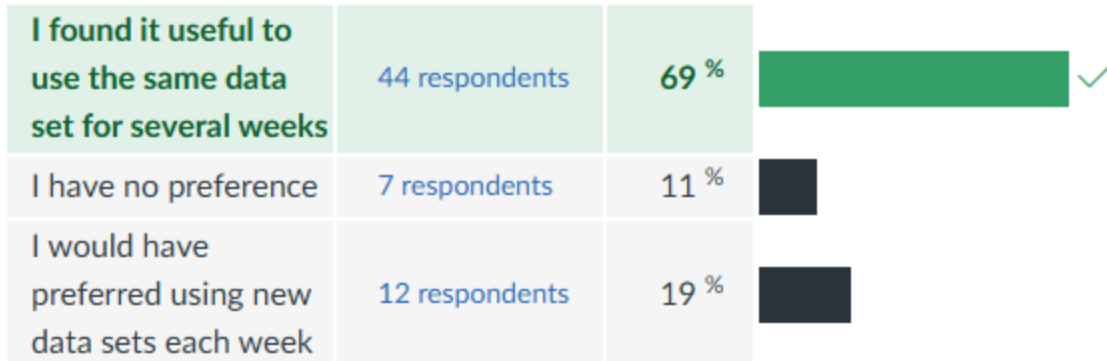
$$H_0: \mu_1 = \mu_2 = \dots = \mu_k$$

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

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

# Faculty salaries again...

Did you find it useful that used the same data sets for several weeks (i.e., the faculty salaries and car transactions data sets), or would you have preferred that we had used different data set each week?



Silly question: Do Assistant, Associate and Fully Professors get paid the same on average?

# 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

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

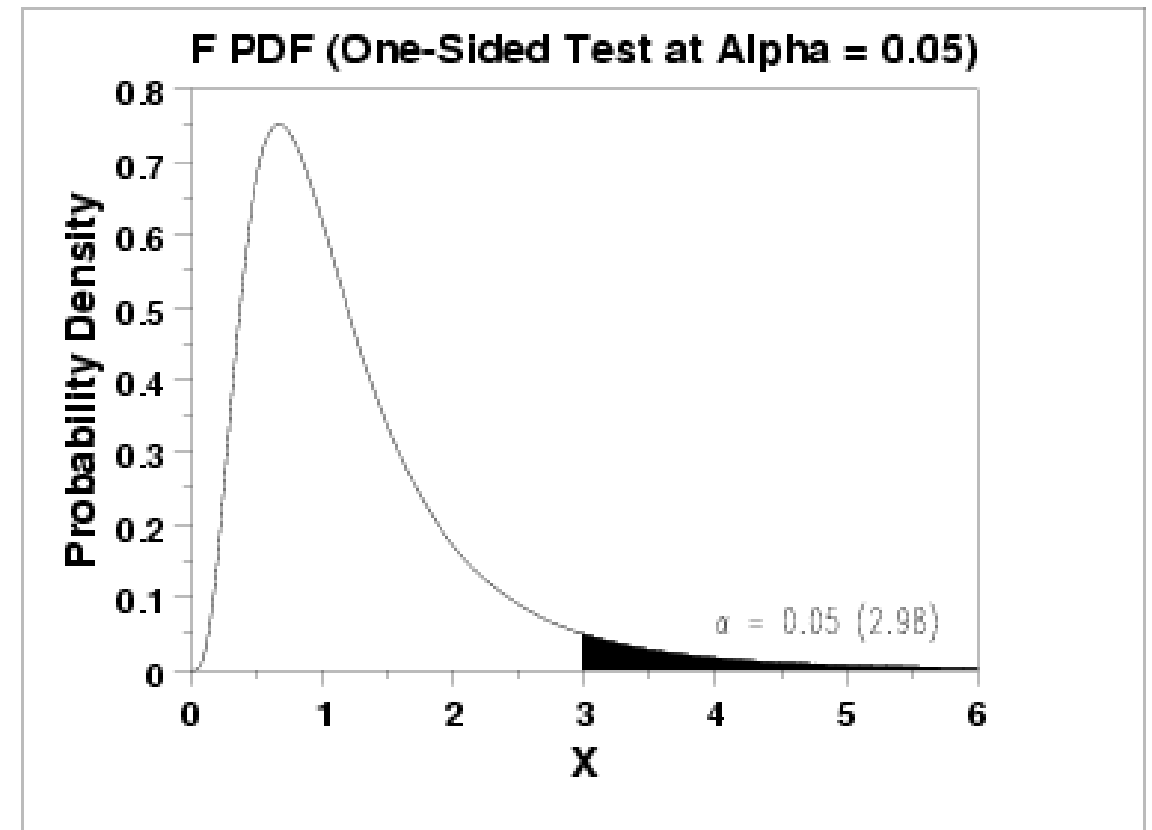
# One-way ANOVA – the central idea

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





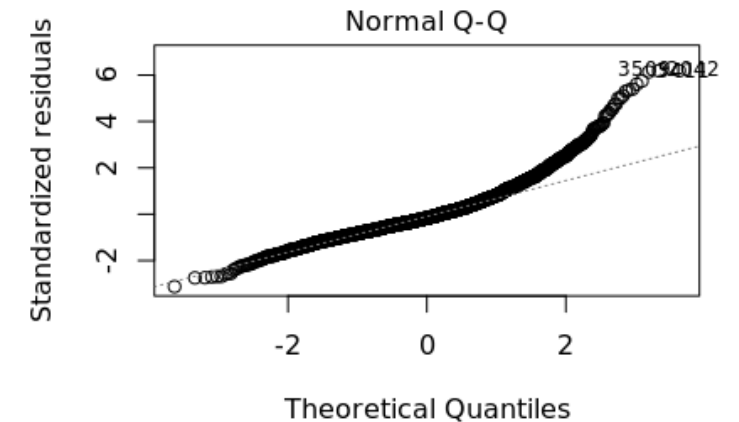
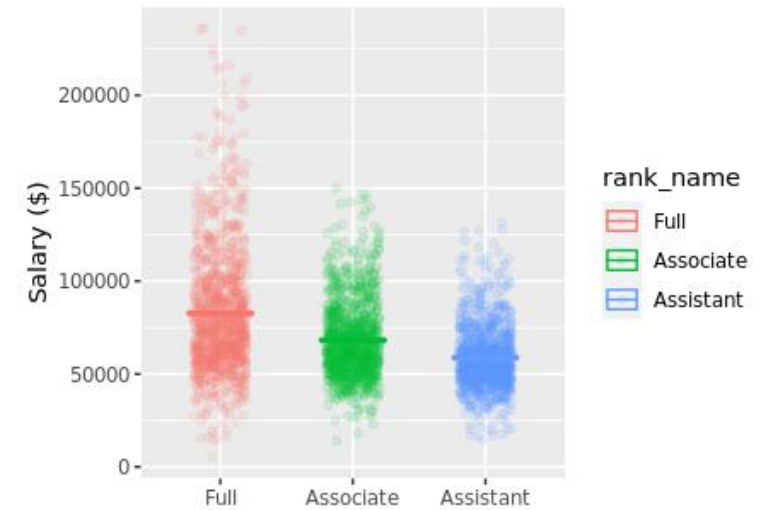
# Checking ANOVA conditions ('assumptions')

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

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



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

# Calculating the observed F-statistic

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

K: the number of groups

N: total number of points

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

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

$n_i$ : the number of points in group i

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

K = 3 different ranks

N = 690 total salaries

$\bar{y}_{tot}$ : the mean salary over all professors

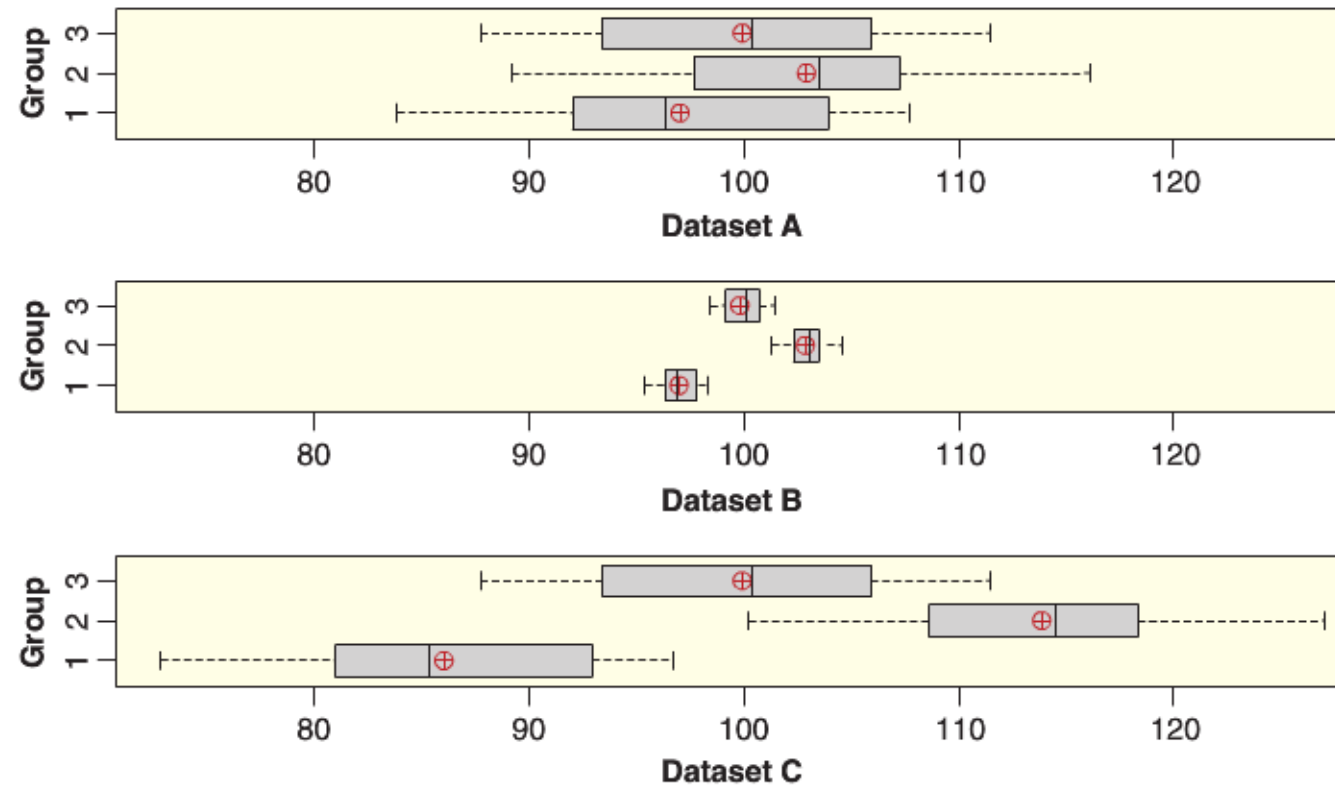
$\bar{y}_i$ : the mean salary for the  $i^{\text{th}}$  professor rank

$n_i$  = num professors of the  $i^{\text{th}}$  rank (330)

$y_{ij}$ : the  $j^{\text{th}}$  professor salary in the  $i^{\text{th}}$  rank

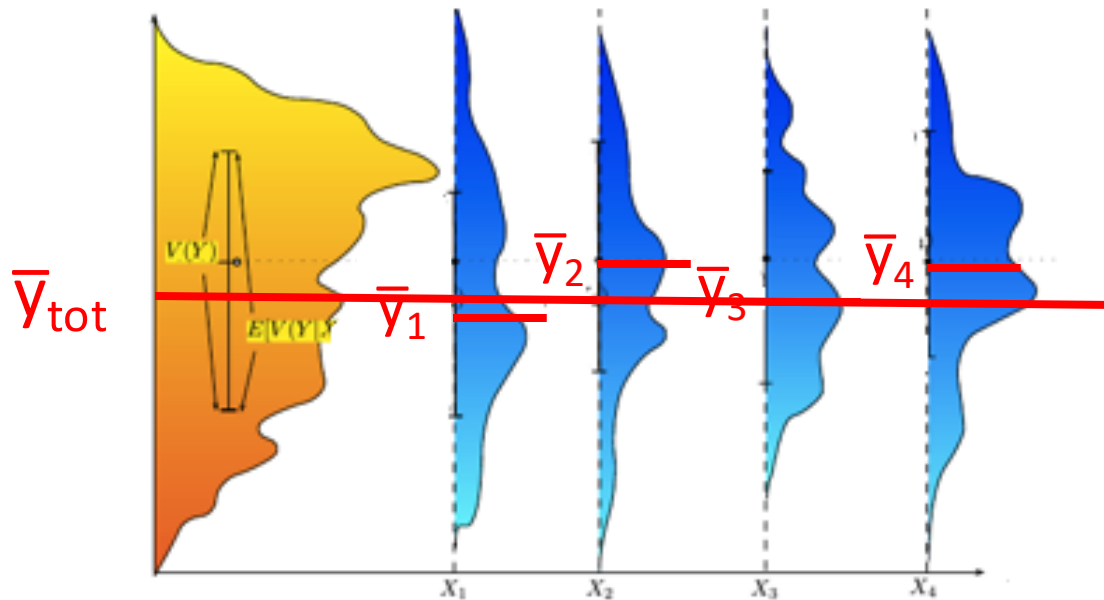
# Why use the F-Statistic?

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



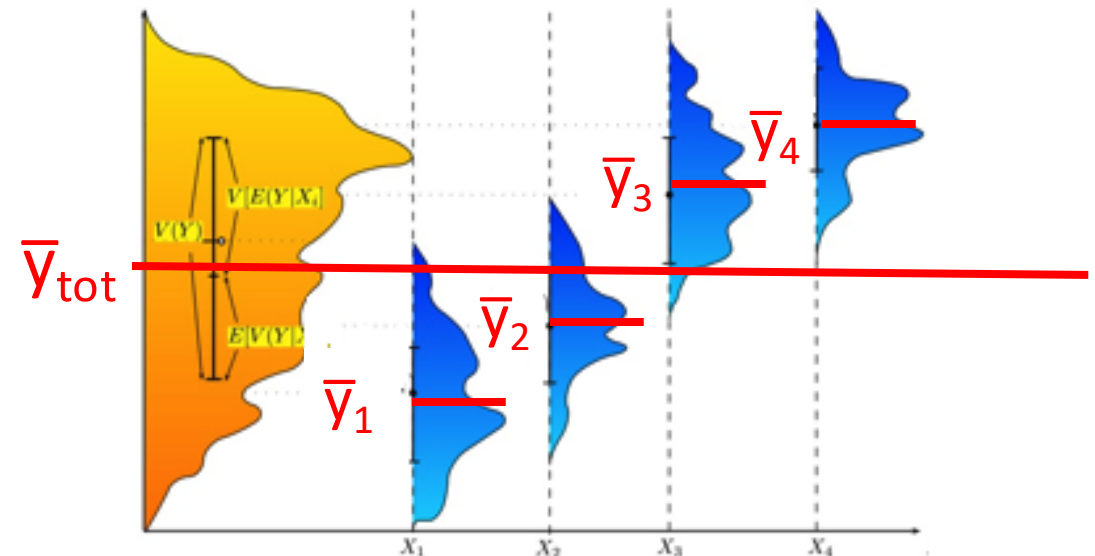
# The F-Statistic

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



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

If  $H_0$  is **not true**, the data from all groups **do not** have the same mean



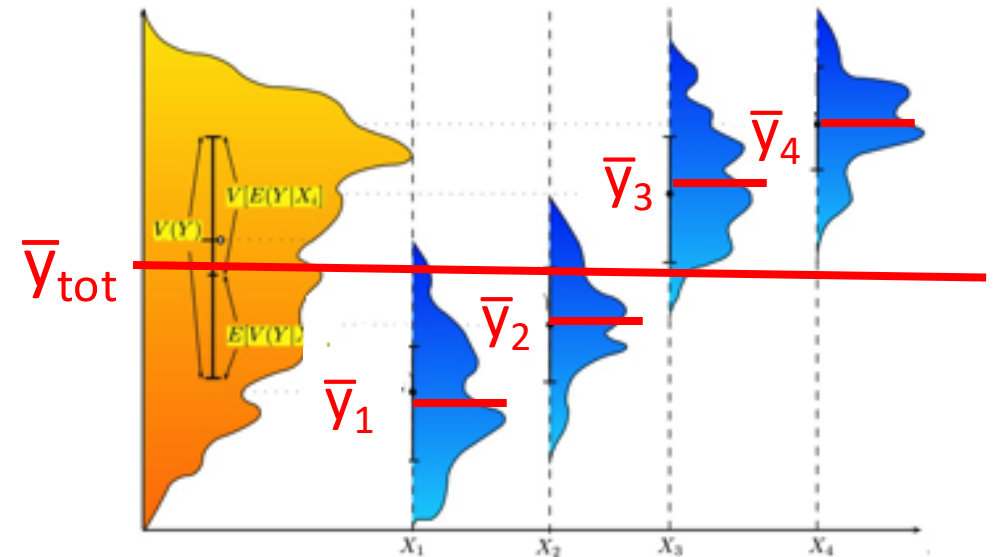
- Different means  $\bar{y}_i$
- Smaller spreads  $s_i$

# The F-statistic

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

The F statistic measures a fraction of:

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



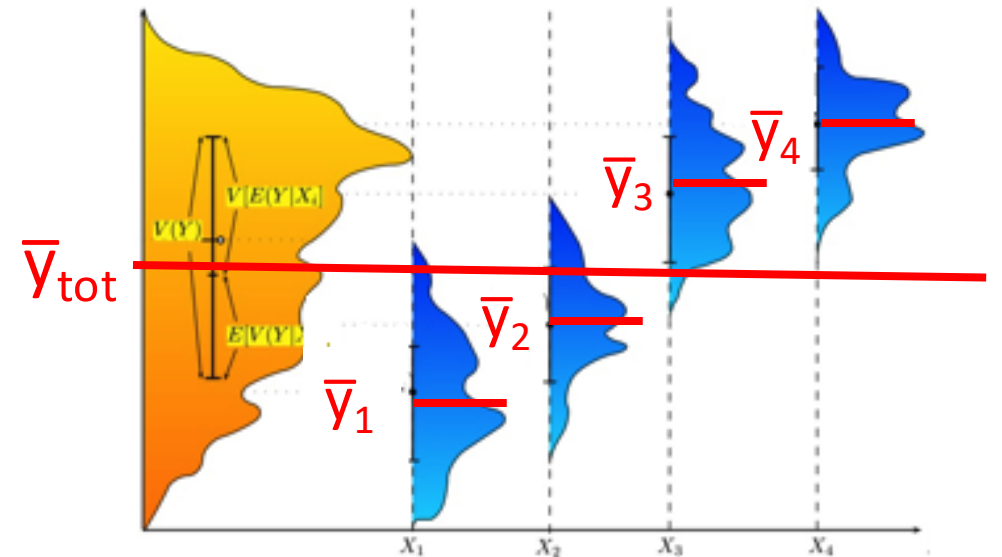
# The F-statistic

Sum of Squares Group (SSG)

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

The F statistic measures a fraction of:

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



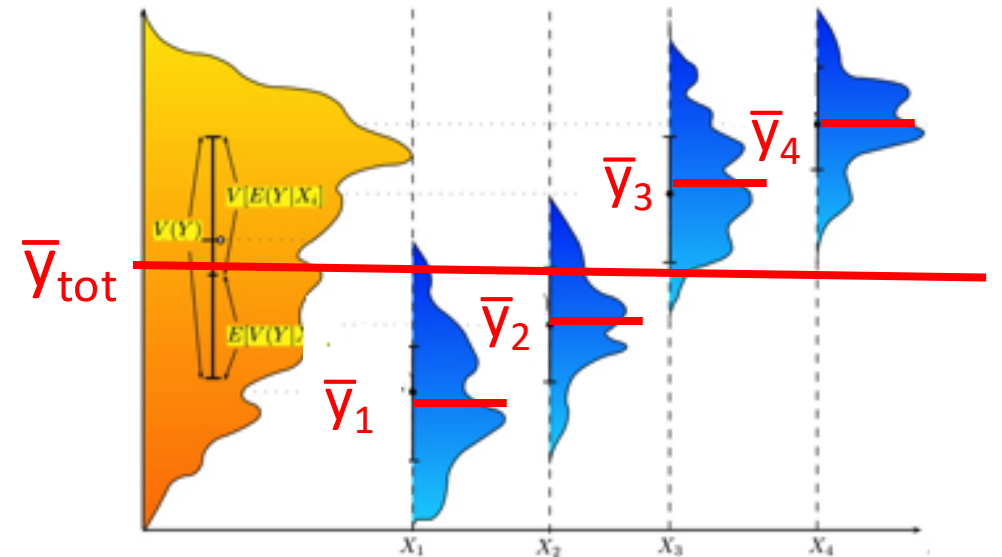
# The F-statistic

Mean Squares Group (MSG)

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

The F statistic measures a fraction of:

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



# The F-statistic

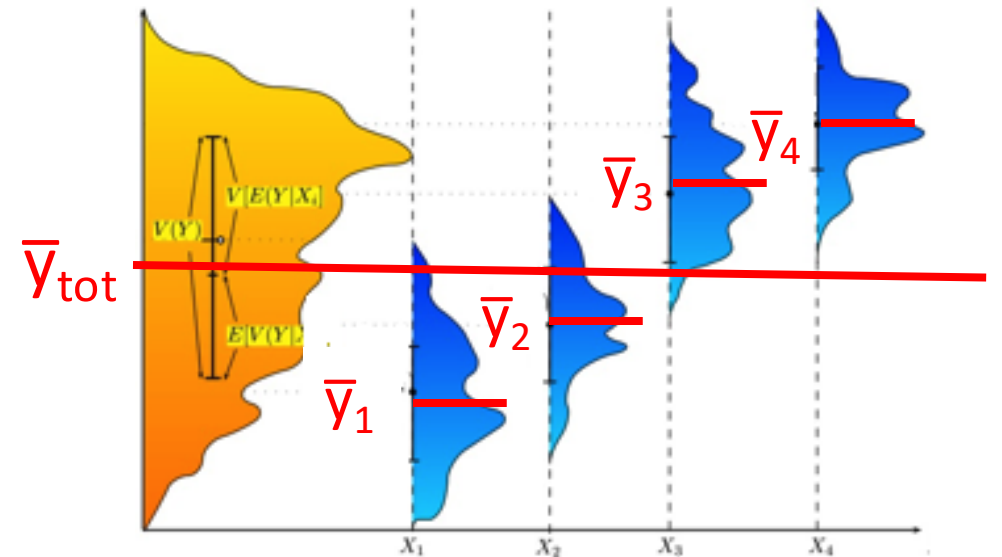
$$F = \frac{\text{between-group variability}}{\text{within-group variability}} = \frac{\frac{1}{K-1} \sum_{i=1}^K n_i (\bar{y}_i - \bar{y}_{tot})^2}{\frac{1}{N-K} \sum_{i=1}^K \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2}$$

Sum of Squares Error (SSE)

↑

The F statistic measures a fraction of:

$$F = \frac{\text{Mean Squares Group (MSG)}}{\text{variability within each group}}$$





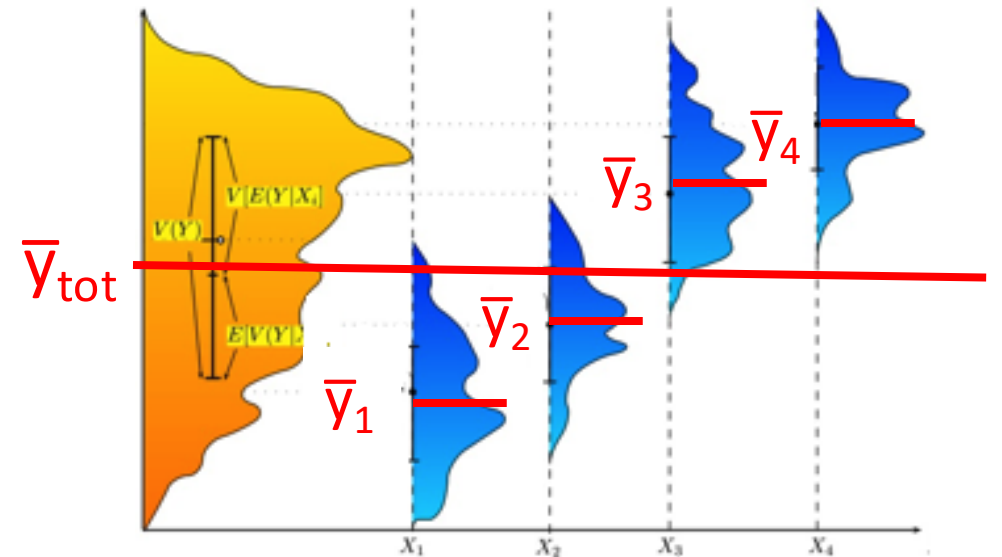
# The F-statistic

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

Mean of Squares Error (MSE)

The F statistic measures a fraction of:

$$F = \frac{\text{Mean Squares Group (MSG)}}{\text{variability within each group}}$$



# The F-statistic

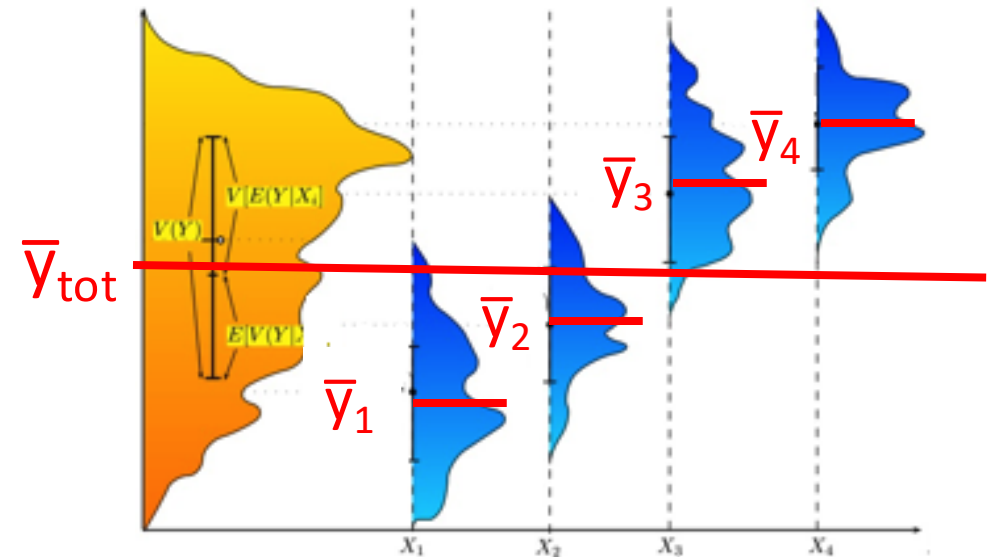
$$F = \frac{\text{between-group variability}}{\text{within-group variability}} = \frac{\frac{1}{K-1} \sum_{i=1}^K n_i (\bar{y}_i - \bar{y}_{tot})^2}{\frac{1}{N-K} \sum_{i=1}^K \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2}$$

Mean of Squares Error (MSE)

↑

The F statistic measures a fraction of:

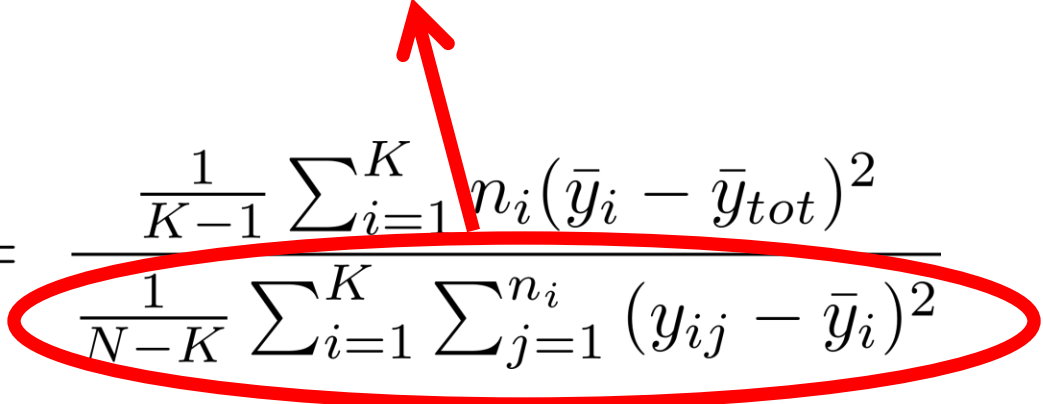
$$F = \frac{\text{Mean Squares Group (MSG)}}{\text{Mean of Squares Error (MSE)}}$$



# The F-statistic

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

Mean of Squares Error (MSE)



The F statistic measures a fraction of:

$$F = \frac{\text{Mean Squares Group (MSG)}}{\text{Mean of Squares Error (MSE)}}$$

If the null hypothesis is true, the F-statistic will be around 1

Larger values, are stronger evidence against the null

# 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:

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

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

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

# ANOVA table

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

$$SST = SSG + SSE$$

Where:

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

$$SSG = \sum_{i=1}^k n_i (\bar{y}_i - \bar{y}_{tot})^2$$
$$SSE = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$$

# 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



Let's try it out in R!



# Connections between regression and ANOVAs



# 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

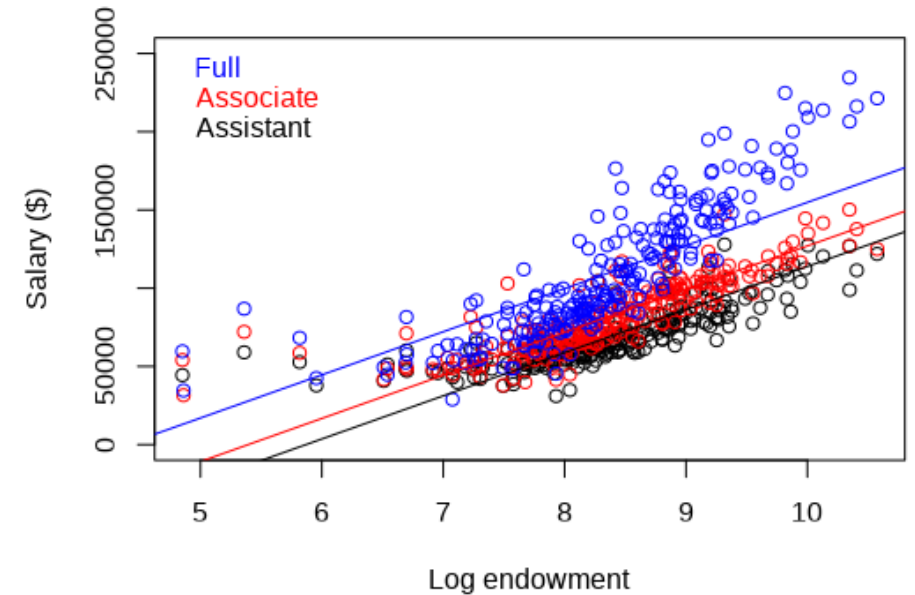
Suppose we want to predict faculty salary  $y$  as a function of endowment  $x_1$ , with separate intercepts for faculty rank

$$x_{i1} = \log(\text{endowment})$$

$$x_{i2} = \begin{cases} 1 & \text{if assistant professor} \\ 0 & \text{otherwise} \end{cases}$$

$$x_{i3} = \begin{cases} 1 & \text{if associate professor} \\ 0 & \text{otherwise} \end{cases}$$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1}$$

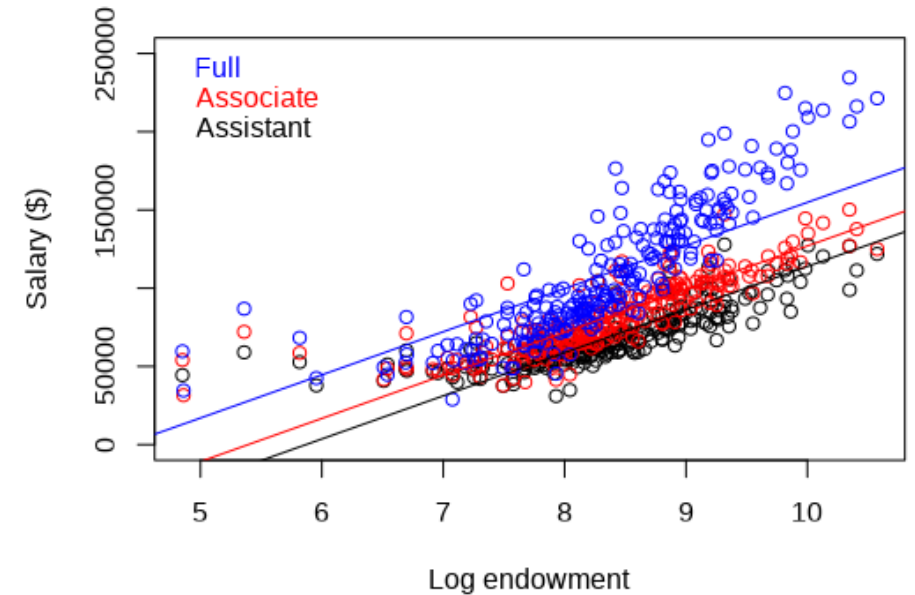


$$= \begin{cases} \hat{\beta}_0 + \hat{\beta}_1 x_{i1} & \text{if full professor} \end{cases}$$

# ANOVA as regression with only categorical predictors

We can view running an ANOVA as creating a model that **only has categorical predictors**

Common slope for  
log endowment

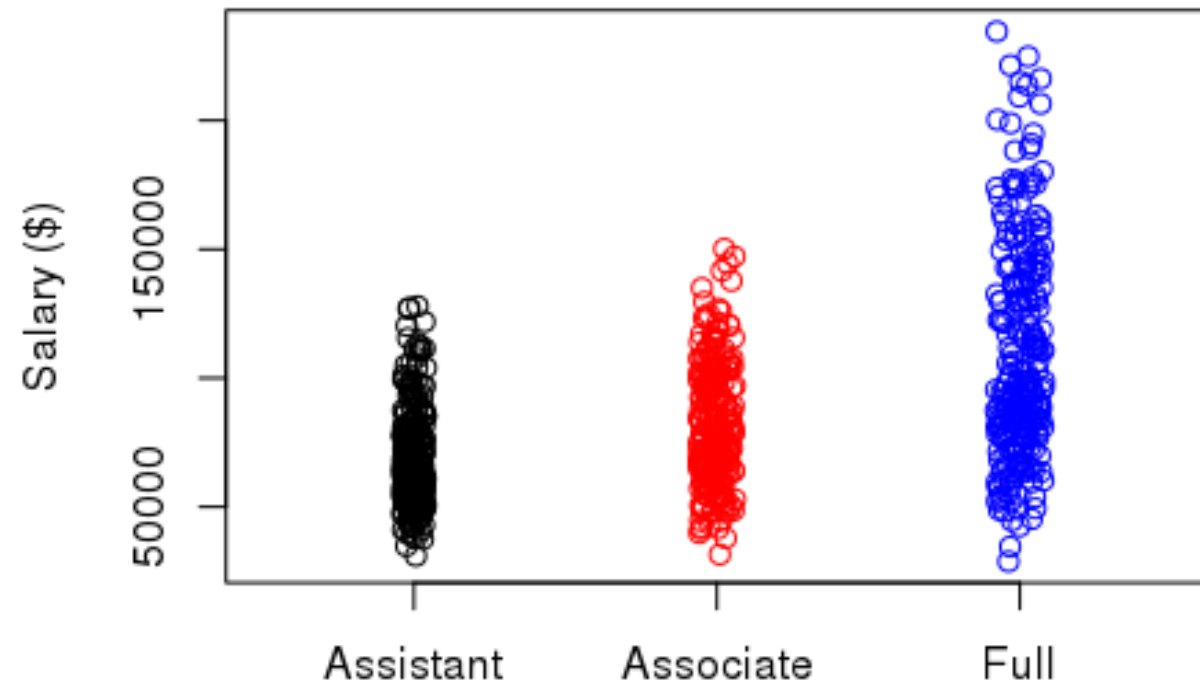


$$x_{i2} = \begin{cases} 1 & \text{if assistant professor} \\ 0 & \text{otherwise} \end{cases}$$

$$x_{i3} = \begin{cases} 1 & \text{if associate professor} \\ 0 & \text{otherwise} \end{cases}$$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3} = \begin{cases} \hat{\beta}_0 + \hat{\beta}_2 & \text{if assistant professor} \\ \hat{\beta}_0 + \hat{\beta}_3 & \text{if associate professor} \\ \hat{\beta}_0 & \text{if full professor} \end{cases}$$

# ANOVA as regression with only categorical predictors

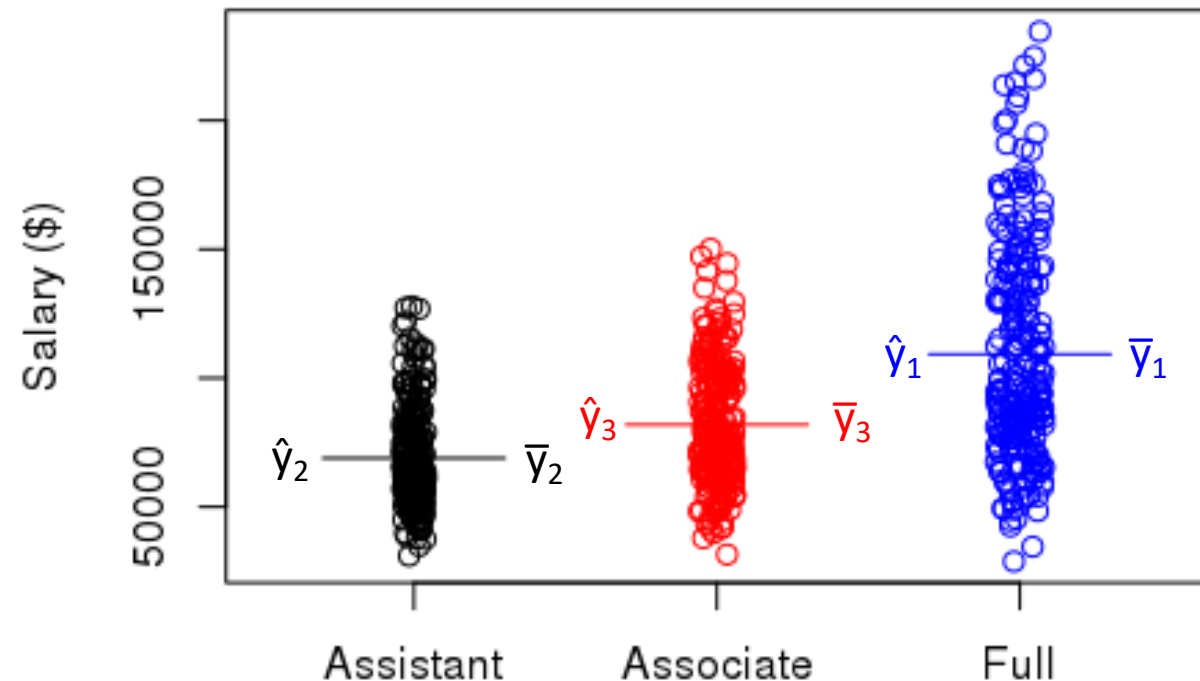


$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if Assistant Professor} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if Associate Professor} \\ \beta_0 + \epsilon_i & \text{if Full Professor} \end{cases}$$

# ANOVA as regression with only categorical predictors

If we use least squares, our predicted value  $\hat{y}_i$  is  $\bar{y}_k$

- i.e., if  $x_i$  belongs to category  $k$ , our prediction is the mean of the  $y$ -values of points in category  $k$



$$\hat{y}_i = \bar{y}_k = \begin{cases} \bar{y}_2 = \hat{\beta}_0 + \hat{\beta}_1 & \text{if Assistant professor} \\ \bar{y}_3 = \hat{\beta}_0 + \hat{\beta}_2 & \text{if Associate professor} \\ \bar{y}_1 = \hat{\beta}_0 & \text{if Full} \end{cases}$$

# ANOVA decomposition

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

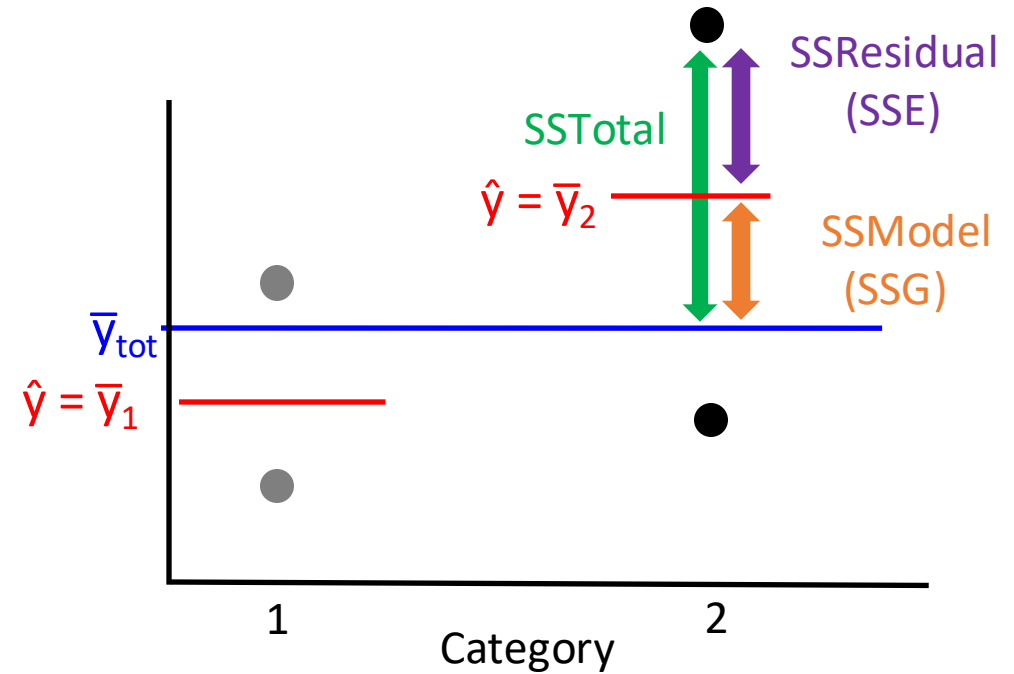
The ANOVA decomposes the variance as:

- **SSTotal** = **SSModel (SSG)** + **SSResidual (SSE)**

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

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

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



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

Let's examine these relationships in R...

# Planned comparisons/posthoc tests

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

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

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

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 SE

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.,  $\alpha = 0.05$  for making **any** Type I error **over all pairs of comparisons**

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

$$t = \frac{\bar{x}_i - \bar{x}_j}{\sqrt{MSE \cdot (\frac{1}{n_i} + \frac{1}{n_j})}} \quad \text{Use a t-distribution with } n-k \text{ degrees of freedom}$$

Very 'conservative' tests

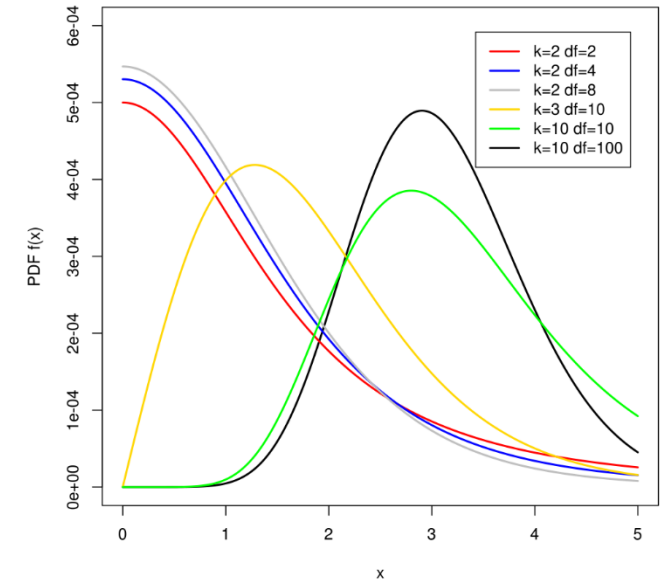
- Unlikely to make Type I errors (few false rejections of  $H_0$ )
- Likely to make Type II errors (insensitive at detecting real effects)

# Tukey's Honest Significantly Different Test

Controls for the family-wise error rate

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

Where  $q$  comes from a ***studentized range distribution***



The test is based on the distribution of  $|\bar{x}_{max} - \bar{x}_{min}|$  that would be expected under the null hypothesis that none of the pairs of means are different

- Controls for the familywise error rate but less conservative than the Bonferroni correction
- Still based on assumptions that the data in each group is normal with equal variance

Let's try pairwise comparisons in R...

# Homework 10

You will analyze data from a psychophysics experiment that explored popout attention

- Study done at Hampshire College by Jacob Prescott, Tapujit Debnath Tapu, Julian Oks, Kirsten Lydic

Background:

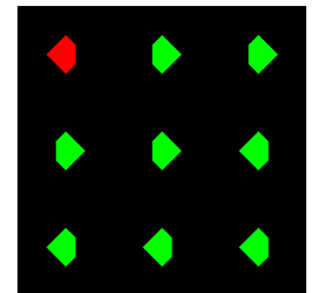
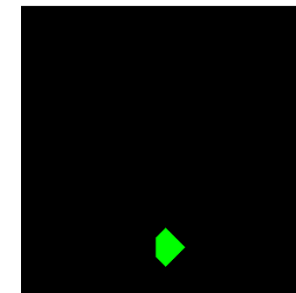


Exogenous attention



Endogenous attention

Single item      Multiple items

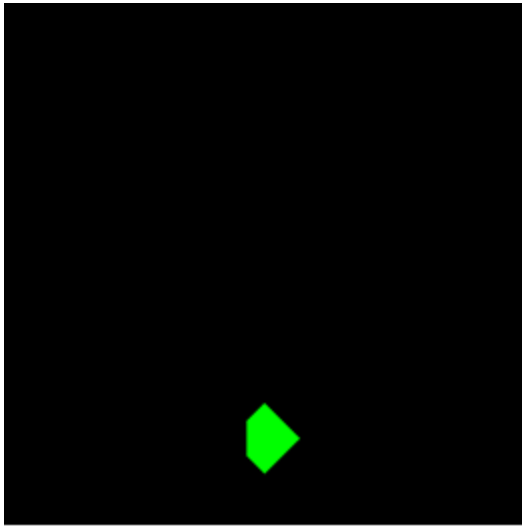


Do reaction times differ for:

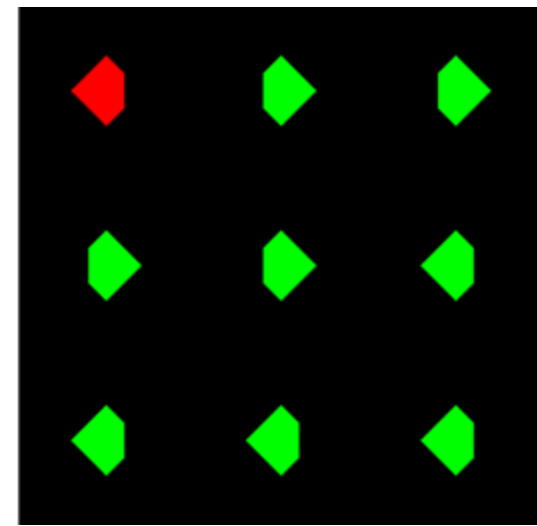
1. Position of target stimulus
2. Single vs. multiple item displays

# Homework 10

Participants engaged in a reaction time task where they needed to respond as *quickly* and as *accurately* as possible

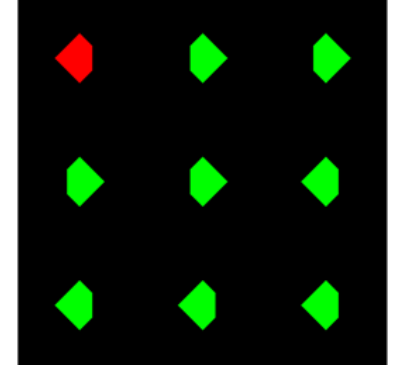


Press "z" because left side is cut off



Press "/" because right side is cut off

# Homework 10



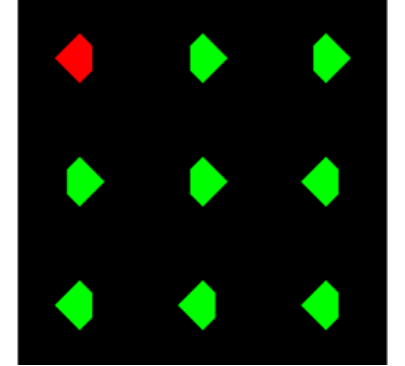
The experiment had a  $9 \times 2 \times 2 \times 2$  **factorial** design:

1. Position (9 levels): 9 locations where the target stimulus could appear
2. Isolated/distractor condition (2 levels): isolated or cluttered display
3. Target color (2 levels): red or green target
  - For cluttered displays, the distractors always had the opposite color of the target
4. Cut direction (2 levels): left or right side of the target diamond was cut off
  - Corresponds to pressing the "z" or "/" key

The experiment had 10 blocks where all 72 ( $9 \times 2 \times 2 \times 2$ ) stimuli were shown

8 volunteer participants participated in the experiment

# Homework 10



On homework 10 you will run:

- A one-way ANOVA to see if the mean reaction time is the same at all target positions
- A two-way ANOVA to look at how both position and isolated/cluttered displays affect mean reaction times.
- Explore another question using this data

Questions?