

CSCI 3022

# intro to data science with probability & statistics

Lecture 26  
April 18, 2018

Forward & Backward Selection

+

Analysis of Variance (ANOVA)

Room for final has changed! FL MG 155



Department of Computer Science  
UNIVERSITY OF COLORADO BOULDER

# Practicum

- The **Practicum** is posted. It is due at 11:55pm on Wednesday May 2.
- **The Rules:**
  - All work must be your own. Collaboration of any kind is not permitted.
  - You may use any resources you like, but you may not post to message boards or other online resources asking for help.
  - We will answer general, clarifying questions in office hours.
  - If you have a question for us, post a **PRIVATE** message on Piazza.
  - Use pandas, numpy, scipy as you wish, plus matplotlib. No other packages pls.
  - See clarifying thread on Piazza. I'm adding any **clarifications text in blue**.

# Last time on CSCI 3022:

- Given data  $(x_{i1}, x_{i2}, \dots, x_{ip}, y_i)$  for  $i = 1, 2, \dots, n$  fit a MLR model of the form:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \epsilon_i \quad \text{where } \epsilon_i \sim N(0, \sigma^2)$$

- We can test if any of the features are important:

$$F = \frac{(SST - SSE)/p}{SSE/(n - p - 1)}$$

$$SST = \sum_{i=1}^n (y_i - \bar{y})^2$$

$$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

- The F-statistic follows an F-distribution
- Rejection Region:  $F \geq F_{\alpha, p, n-p-1}$  p-value:  $1 - \text{stats.f.cdf}(F, p, n-p-1)$

# Deciding on important variables

$i = 100$

- Suppose that we have 100 data points ( $n=100$ ), but we have 200 different features ( $p=200$ ). How can we learn which features are important and which are not?
- **Some options:**
  - Try all the possible combinations of features in models to see which gives the best fit.

Bad!

$p = 30$

$2^P$  different models.

$1,073,741,824$  models.  
yikes!

# Deciding on important variables

- Suppose that we have 100 data points ( $n=100$ ), but we have 200 different features ( $p=200$ ). How can we learn which features are important and which are not?
  - **Some options:**  $f_0$  do not include  $f_1, f_2, f_3 \dots$

# • Forward selection

1. fit null model with an intercept but no predictors.
  2. fit p-SLRs, 1 for each feature. Choose the one that gives the lowest SSE.  
lowest SSE
  3. fit p-1 MLRs. Choose that which gives lowest SSE...  
p-1 MLRs
  4. repeat. step 3, w/  $p-2$  MLRs, then  $p-3$  MLRs ...

"base line"

"baseline  
+ 1 feature"

For

do not include  
any  $\beta_1, \beta_2, \beta_3 \dots$

# Deciding on important variables

- Suppose that we have 100 data points ( $n=100$ ), but we have 200 different features ( $p=200$ ). How can we learn which features are important and which are not?
- **Some options:**

- **Backward selection:**

$p$  features

1. Fit model with *all* predictors

$p-1$  features

2. Remove the one with the largest  $p$ -value.

3. Fit model with  $p-1$  predictors.

$p-2$  features

4. Remove the one with the largest  $p$ -value...

# Quiz. Name That Computation!

1. **Advertising example.** I want to know if the set of {news,radio} together have at least one slope that is significantly different from 0.

F-test! For a subset of features.  $H_0: \beta_{\text{news}} = \beta_{\text{radio}} = 0$

2. **Home prices example.** I have 1000 data points and 30 features. I want to learn the 10 most predictive and significant features.

because  $n \gg p$  use forward selection to get 10 features, or backward selection.

3. **Home prices example.** I have 100 data points and 200 features. I want to learn the 20 most predictive features.

Forward only. If  $n < p$ , use forward selection.

4. **Shark attacks example.** I have 50 shark attacks, and I have 20 features. I fit my MLR model. Now I want to compute how well my model fits the data.

$R^2_a$  ← includes penalty for using so many features.

# Comparing multiple means

- We're often interested in comparing the means of a response from different groups
- **Example:** Suppose we are doing a study on the effect of diet on weight-loss. We have three different groups in the study:
  - **Control group:** exercise only
  - **Treatment A:** exercise plus Diet A
  - **Treatment B:** exercise plus Diet B
- We record the weight-loss of each participant after one week of the study and find the following results:

*Three participants per group.*

	Control	Diet A	Diet B
0	3	5	5
1	2	3	6
2	1	4	7

# Comparing multiple means

- We're often interested in comparing the means of a response from different groups
- **Example:** Suppose we are doing a study on the effect of diet on weight-loss. We have three different groups in the study:
  - **Control group:** exercise only
  - **Treatment A:** exercise plus Diet A
  - **Treatment B:** exercise plus Diet B
- We record the weight-loss of each participant after one week of the study and find the following results:

**Question:** Are the means of the different groups all the same?

What would we do if there were only **two** groups?

CI for  $\mu_1 - \mu_2$  ? includes 0

CIs  $\mu_1$  and  $\mu_2$  ? overlap

t-test  $H_0: \mu_1 = \mu_2$

z-test  $H_1: \mu_1 \neq \mu_2$

# Comparing multiple means

- We're often interested in comparing the means of a response from different groups
  - **Example:** Suppose we are doing a study on the effect of diet on weight-loss. We have three different groups in the study:
    - **Control group:** exercise only
    - **Treatment A:** exercise plus Diet A
    - **Treatment B:** exercise plus Diet B
  - We record the weight-loss of each participant after one week of the study and find the following results:

**Question:** Are the means of the different groups all the same?

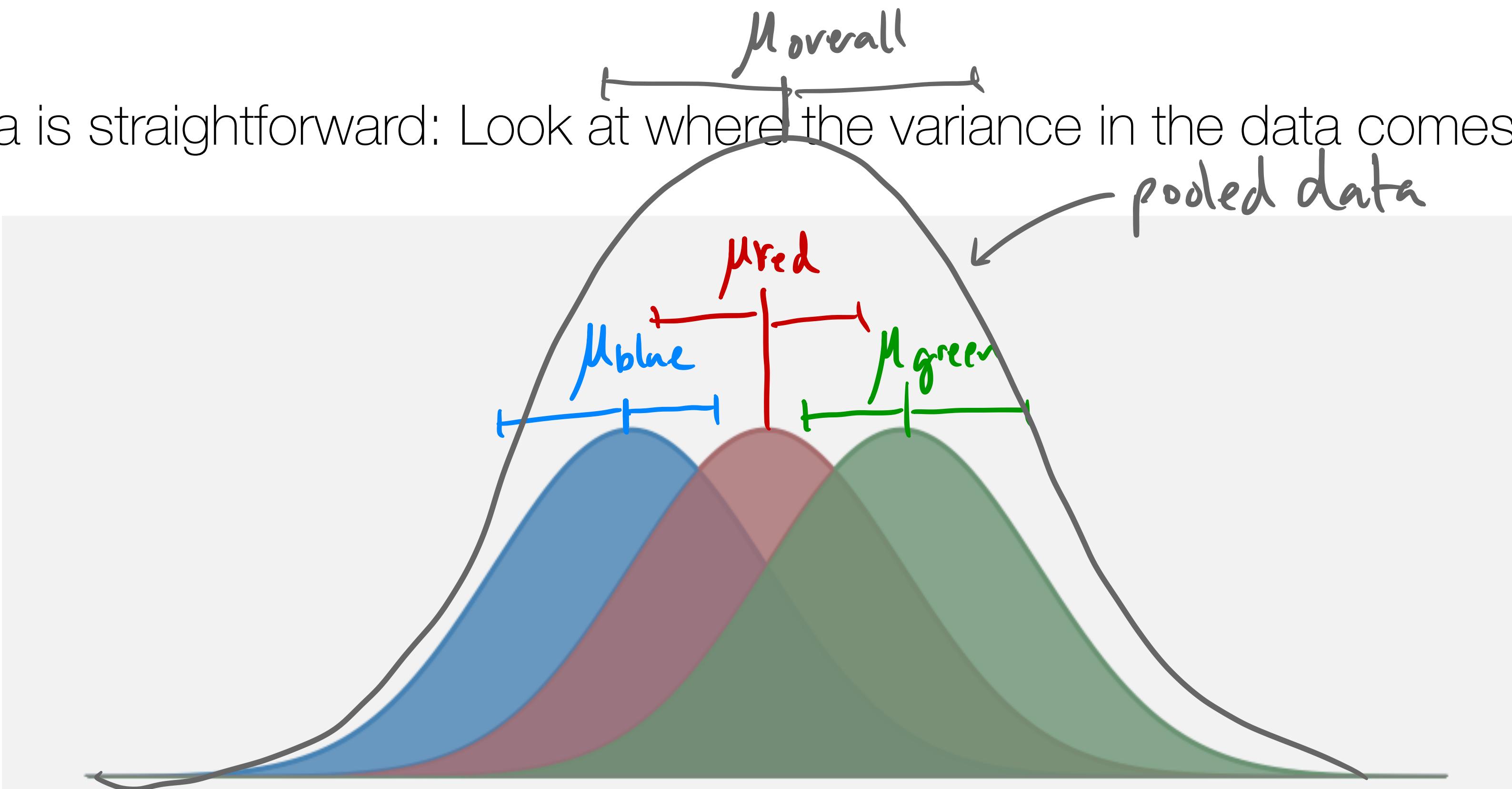
Why would a t- or z-test be problematic if we had many different groups?

① expensive  $\binom{k}{2}$  comparison =  $\frac{k(k-1)}{2} \sim k^2$       ② multiple comparisons  
k groups

# Analysis of variance

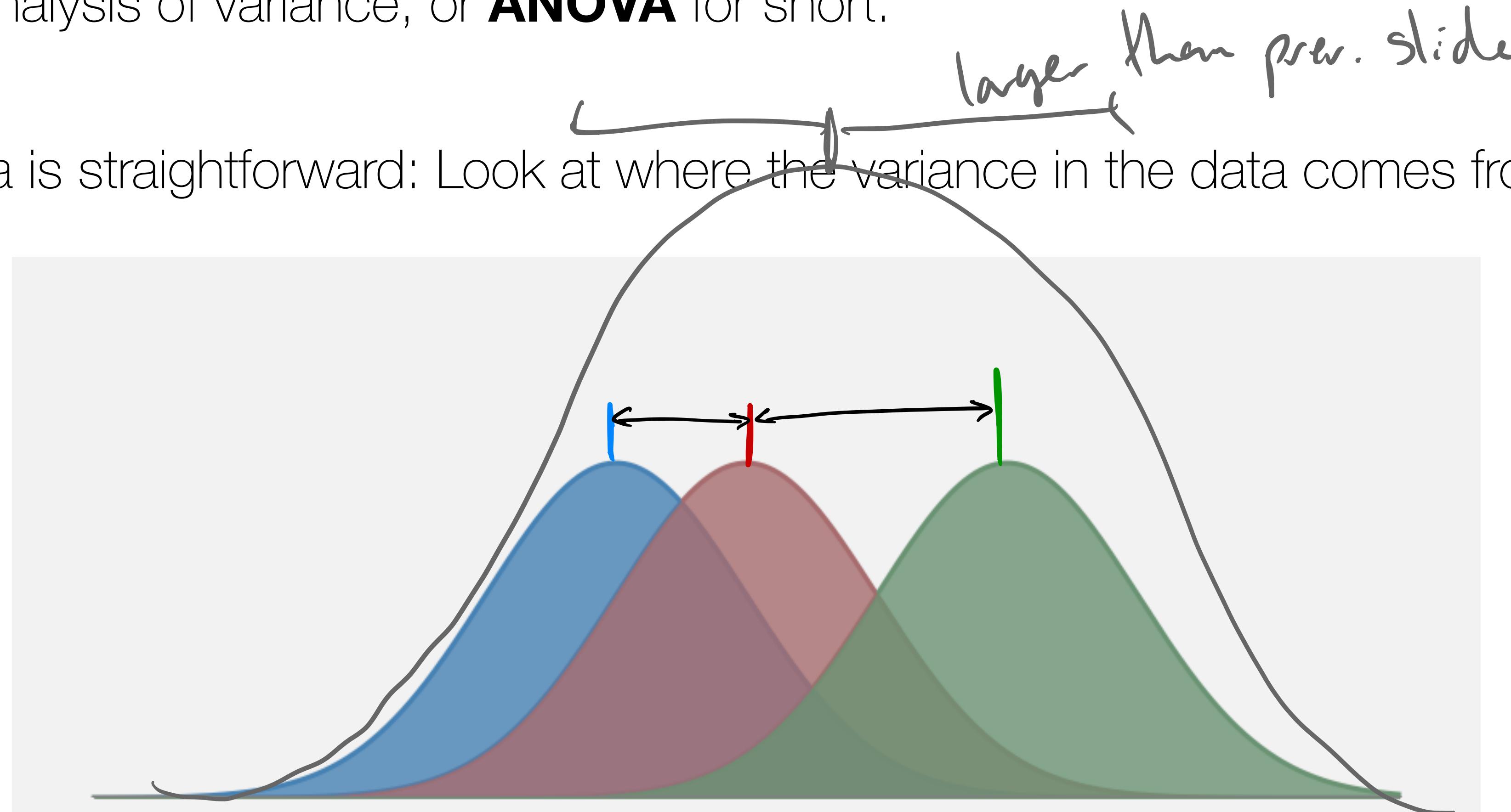
- We can answer the question **"Are any of the means different?"** using a procedure called analysis of variance, or **ANOVA** for short.

- The idea is straightforward: Look at where the variance in the data comes from.



# Analysis of variance

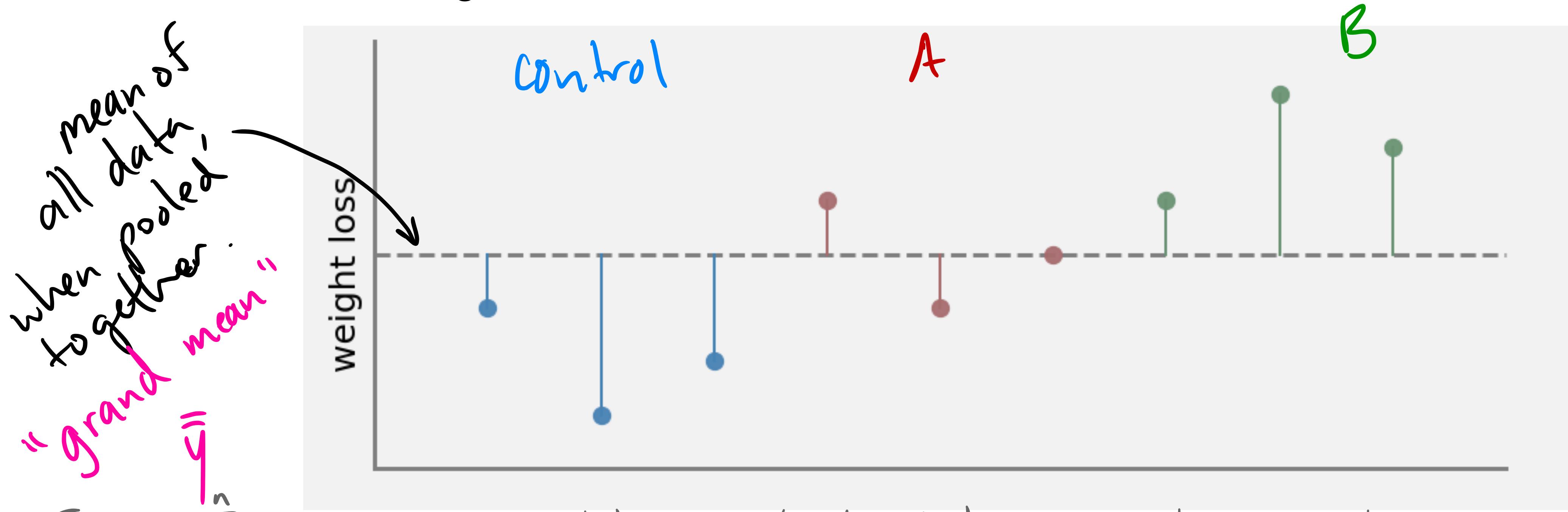
- We can answer the question “Are any of the means different?” using a procedure called analysis of variance, or **ANOVA** for short.
- The idea is straightforward: Look at where the variance in the data comes from.



# Analysis of variance

$$\text{Overall variance} = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2$$

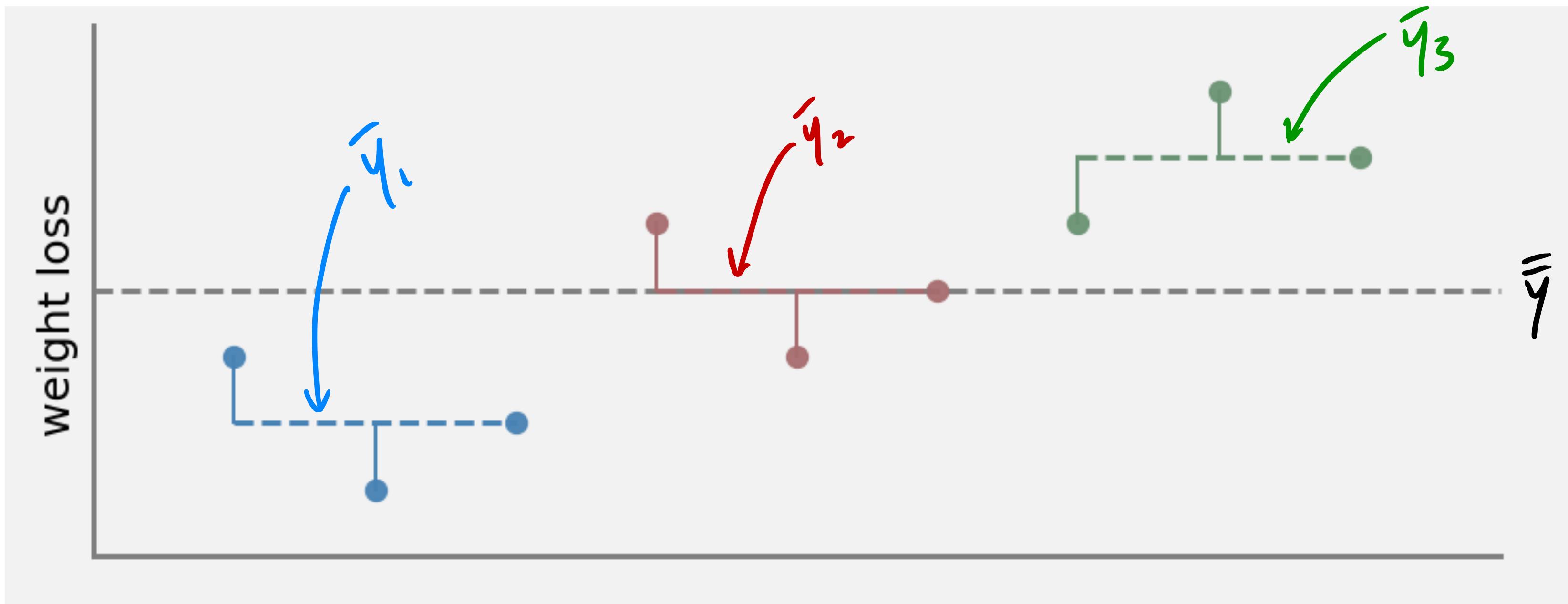
- We can answer the question “Are any of the means different?” using a procedure called analysis of variance, or **ANOVA** for short.
- The idea is straightforward: Look at where the variance in the data comes from.



$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$  ← calculation pretends that group labels don't exist.

# Analysis of variance

- We can answer the question “Are any of the means different?” using a procedure called analysis of variance, or **ANOVA** for short.
- The idea is straightforward: Look at where the variance in the data comes from.



# The one-way ANOVA model

I

- Suppose that we have  $I$  groups that we want to compare, each with  $n_i$  data points
- We model the relationship between responses and group means as follows:

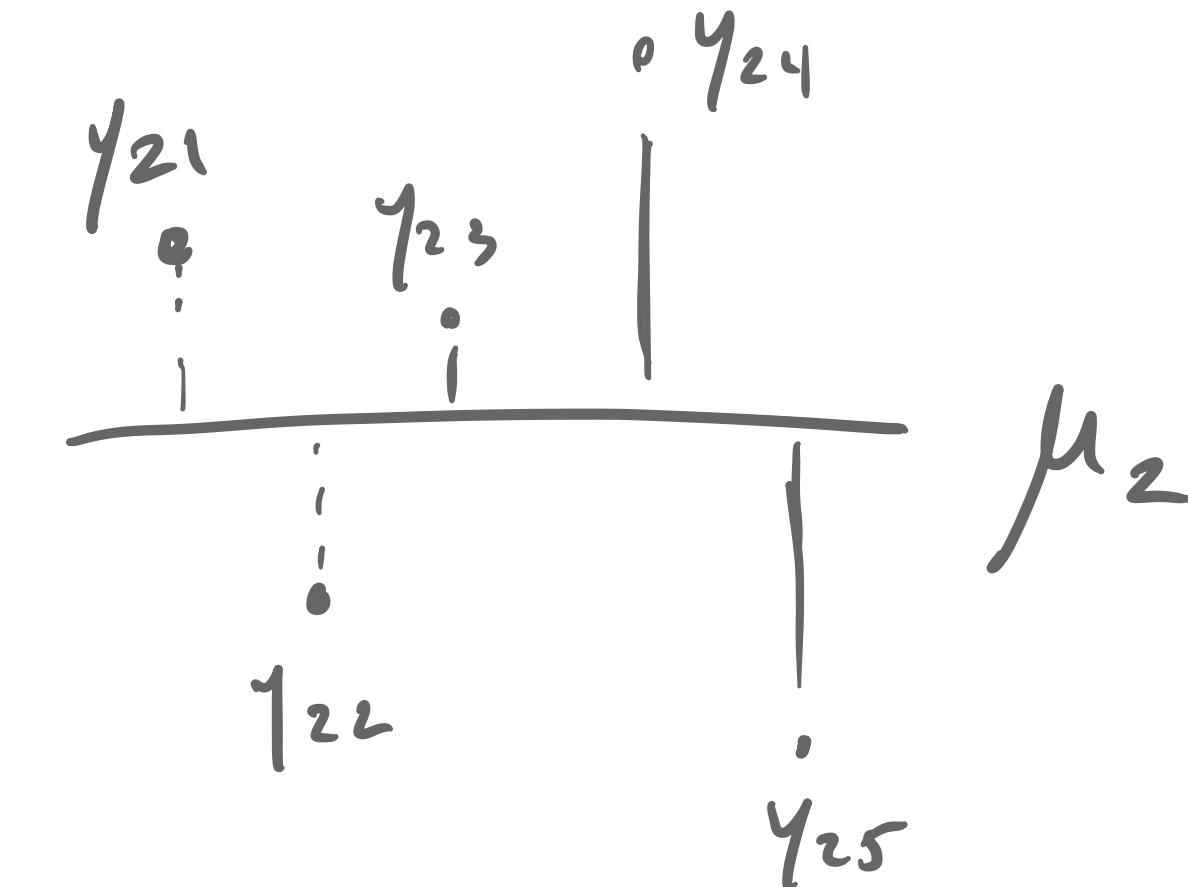
$$y_{ij} = \mu_i + \epsilon_{ij}$$

group  $i$

$j^{\text{th}}$  data point  
in that group

mean of  
group  $i$

noise, error  
 $\epsilon_{ij} \sim N(0, \sigma^2)$



prev example  
 $n_1 = 3$   
 $n_2 = 3$   
 $n_3 = 3$   
points

## Assumptions:

- the responses are i.i.d. samples from normally distributed groups
- the variance of each group is the same

ANOVA is "pretty robust"  
to violations of these assumptions.  
• non-normal data  
• slightly non-equal variances.

# The one-way ANOVA model

Let's compute some means!

- The **grand mean** is the sample mean of all responses.

$$\bar{y} = \frac{1}{9} [(3+2+1) + (5+3+4) + (5+6+7)]$$

$$= \frac{1}{9} [6 + 12 + 18] = \frac{1}{9} \cdot 36 = 4$$

$$\boxed{\bar{y} = 4}$$

- The **group means** are the sample means within each group.

$$\bar{y}_1 = \frac{1}{3} (3+2+1) = \frac{1}{3} 6 = 2$$

$$\bar{y}_3 = \frac{1}{3} (5+6+7) = \frac{1}{3} 18 = 6$$

$$\bar{y}_2 = \frac{1}{3} (5+3+4) = \frac{1}{3} 12 = 4$$

$$N = n_1 + n_2 + n_3 = 9$$

$$\bar{y} = \frac{n_1 \bar{y}_1 + n_2 \bar{y}_2 + n_3 \bar{y}_3}{N}$$

# It's the variances, stupid

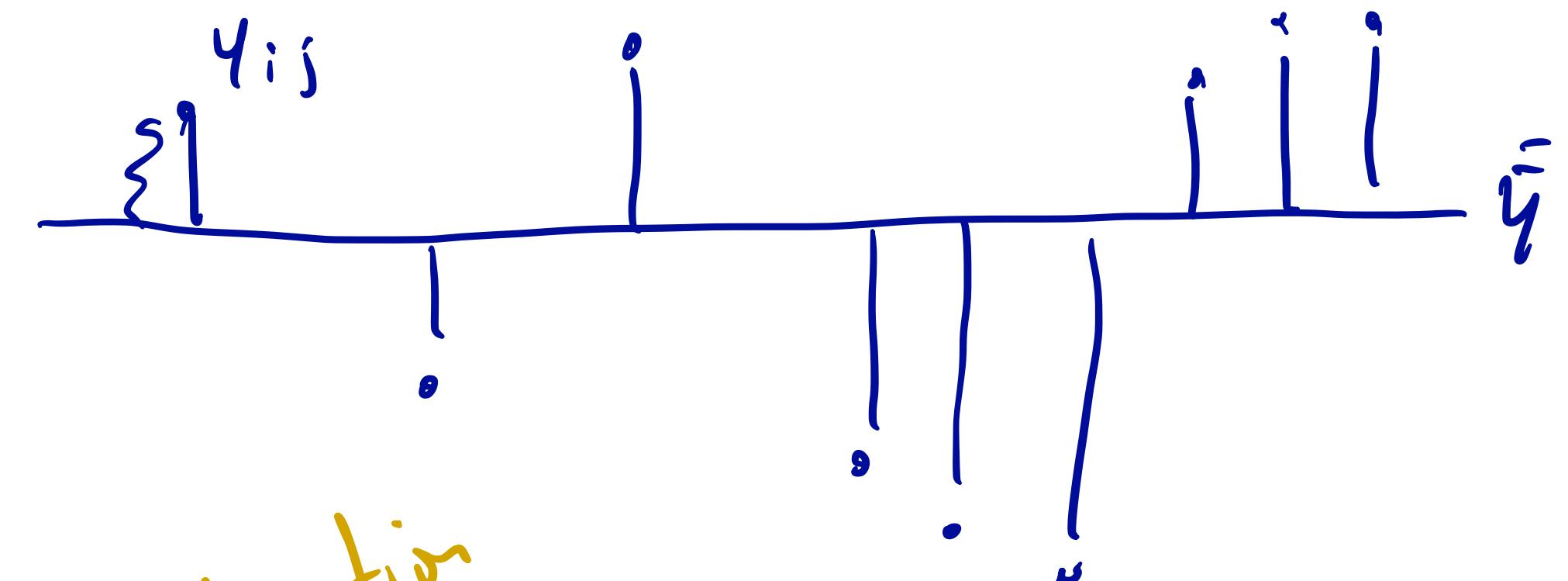
cf "It's the economy, stupid"

1992 Carville.

- Where does the total variation in the data come from? Remember linear regression:

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

over all groups      within each group



- A helpful decomposition:

$$y_{ij} - \bar{y} + \bar{y}_i - \bar{\bar{y}}_i = (\underbrace{y_{ij} - \bar{y}_i}_{\text{within group deviation}}) + (\underbrace{\bar{y}_i - \bar{\bar{y}}_i}_{\text{between group deviation}})$$

overall deviation

- Then, a minor (mathematical) miracle occurs:

$$SST = \sum_{i=1}^I \sum_{j=1}^{n_i} \left[ (y_{ij} - \bar{y}_i)^2 + (\bar{y}_i - \bar{\bar{y}}_i)^2 \right]$$

SSWithin      SSBetween

# The one-way ANOVA model

Let's compute some variances (or at least, sums of squares)!

- The **BETWEEN** group sum of squares is:

$$SSB = \sum_{i=1}^I \sum_{j=1}^{n_i} (\bar{y}_i - \bar{\bar{y}})^2 = \sum_{i=1}^I n_i (\bar{y}_i - \bar{\bar{y}})^2$$

$$= 3(2-4)^2 + 3(4-4)^2 + 3(6-4)^2 = 24$$

$$\bar{y}_i = \begin{matrix} 2 \\ 4 \\ 6 \end{matrix}$$

- The **WITHIN** group sum of squares is:

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

$(3-2)^2$	$1$	$(6-4)^2$	$1$	$(5-6)^2$	$1$	}
$(2-2)^2$	$0$	$(3-4)^2$	$1$	$(6-6)^2$	$0$	
$(1-2)^2$	$1$	$(4-4)^2$	$0$	$(7-6)^2$	$1$	

- The **TOTAL** sum of squares is:

$$SST = SSW + SS B = 6 + 24 = 30$$

$$n_1 = 3 \quad n_2 = 3 \quad n_3 = 3$$

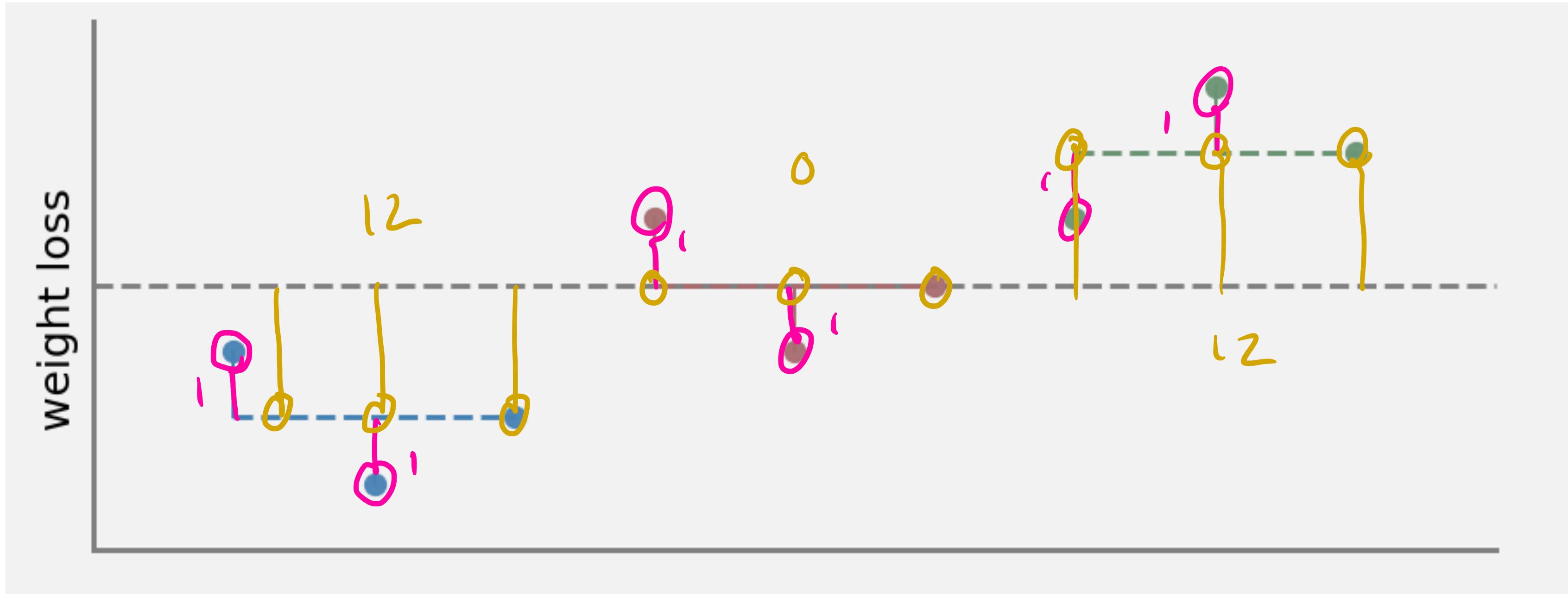
Control	Diet A	Diet B	
0	3	5	5
1	2	3	6
2	1	4	7

# The one-way ANOVA model

$$SSB = 24$$

$$SSW = 6$$

- Compare these results to the original picture:



	Control	Diet A	Diet B
0	3	5	5
1	2	3	6
2	1	4	7

# The one-way ANOVA model



What about degrees of freedom?

- The **BETWEEN** group degrees of freedom is (are?):

$$SSB = \sum_{i=1}^I \sum_{j=1}^{n_i} (\bar{y}_{ij} - \bar{\bar{y}})^2 = \sum_{i=1}^I n_i (\bar{y}_i - \bar{\bar{y}})^2$$

treat  
this like data  
computed from the data

	Control	Diet A	Diet B
0	3	5	5
1	2	3	6
2	1	4	7

$$SSB_{df} = I - 1 = 3 - 1 = 2$$

$N$  = total data pts

- The **WITHIN** group degrees of freedom is (are?):

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

data  
computed from the data

$$N = \sum_{i=1}^I n_i$$

$$SSW_{df} = N - I = 9 - 3 = 6$$

# A hypothesis test

i.e. = id est → that is, specifically  
e.g. = exempla gratia → for example.

- We want to perform a hypothesis test to determine if the group means are equal. We have

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

$$H_1 : \mu_i \neq \mu_j \text{ for some } (i, j) \text{ pair}$$

i.e. two of the group means are different.

- Our test statistic will be:

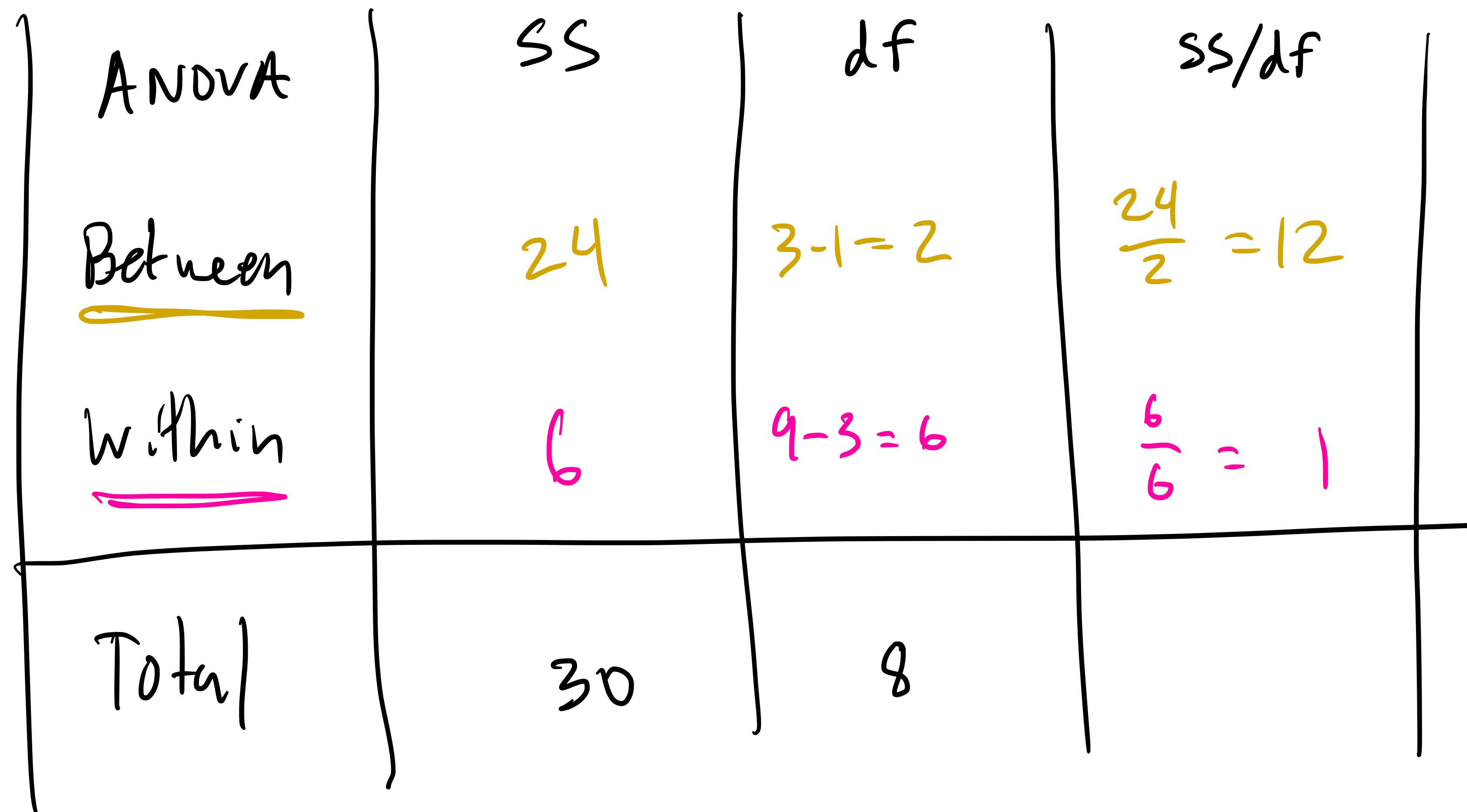
$$F = \frac{\frac{SSB}{SSB_{df}}}{\frac{SSW}{SSW_{df}}} = \frac{\frac{SSB}{I-1}}{\frac{SSW}{N-I}} = \frac{SSB(N-I)}{SSW(I-1)}$$

Rej. Region Test:  $F \geq F_{\alpha, I-1, N-I}$

p-val : 1 - stats.f.cdf( $F_{21}, I-1, N-I$ )

# The ANOVA Table

- It is common practice to organize all computations into an ANOVA table



	Control	Diet A	Diet B
0	3	5	5
1	2	3	6
2	1	4	7

$$F = \frac{12}{1} = 12$$

$$p\text{-val} = 0.008$$

# ANOVA as multiple linear regression

- Interestingly, there is a very close relationship between One-Way ANOVA and MLR!
- Suppose you have  $l$  groups that you want to compare. A random sample of size  $n_i$  is taken from the  $i^{\text{th}}$  group. Then

# ANOVA as multiple linear regression

- Interestingly, there is a very close relationship between One-Way ANOVA and MLR!
- Suppose you have  $l$  groups that you want to compare. A random sample of size  $n_i$  is taken from the  $i^{\text{th}}$  group. Then

# Tukey's honest significance test

- Suppose that we determine that some of the means are different.
- How can we tell which ones?

Tukey's HST or Tukey's Range Test

Hypothesis Test for pair-wise comp. of means.

Fixes problem of multiple comparisons.

→ Adjusts so that making a Type-I error over all possible pair-wise comparisons =  $\alpha$

same Tukey  
as S-number  
Summary  
EDA

rejecting  $H_0$   
even though  
 $H_0$  was true.

# Tukey's honest significance test

- Suppose that we determine that some of the means are different.
- How can we tell which ones?

# Tukey's honest significance test

- Suppose that we determine that some of the means are different.
- How can we tell which ones?

# Tukey's honest significance test

- Suppose that we determine that some of the means are different.
- How can we tell which ones?

# Tukey's honest significance test

- Suppose that we determine that some of the means are different.
- How can we tell which ones?

# Tukey's honest significance test

- Suppose that we determine that some of the means are different.
- How can we tell which ones?

# Tukey's honest significance test

- Suppose that we determine that some of the means are different.
- How can we tell which ones?