Analysis of Variance

Practicum

- The **Practicum** is posted. It is due at 11:59pm on Wednesday December 13th.
- **The Rules**:
    - All work must be your own. Collaboration of any kind is 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.

Previously on CSCI 3022

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

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

We can test if any of the features are important:

$$F = \frac{(SST - SSE)/p}{SSE/(n-p-1)} \qquad SST = \sum_{I=1}^{n} (y_i - \bar{y})^2 \qquad SSE = \sum_{I=1}^{n} (y_i - \hat{y}_i)^2$$

The F-statistic follows an F-distribution

Rejection Region: $F \geq F_{\alpha, p, n-p-1}$ p-value: 1 - stats.f.cdf(F, p, n-p-1)

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:

	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?

$H_0: \mu_1 = \mu_2$

$H_1: \mu_1 \neq \mu_2$

t OR Z - TEST FOR DIFF OF MEANS

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?

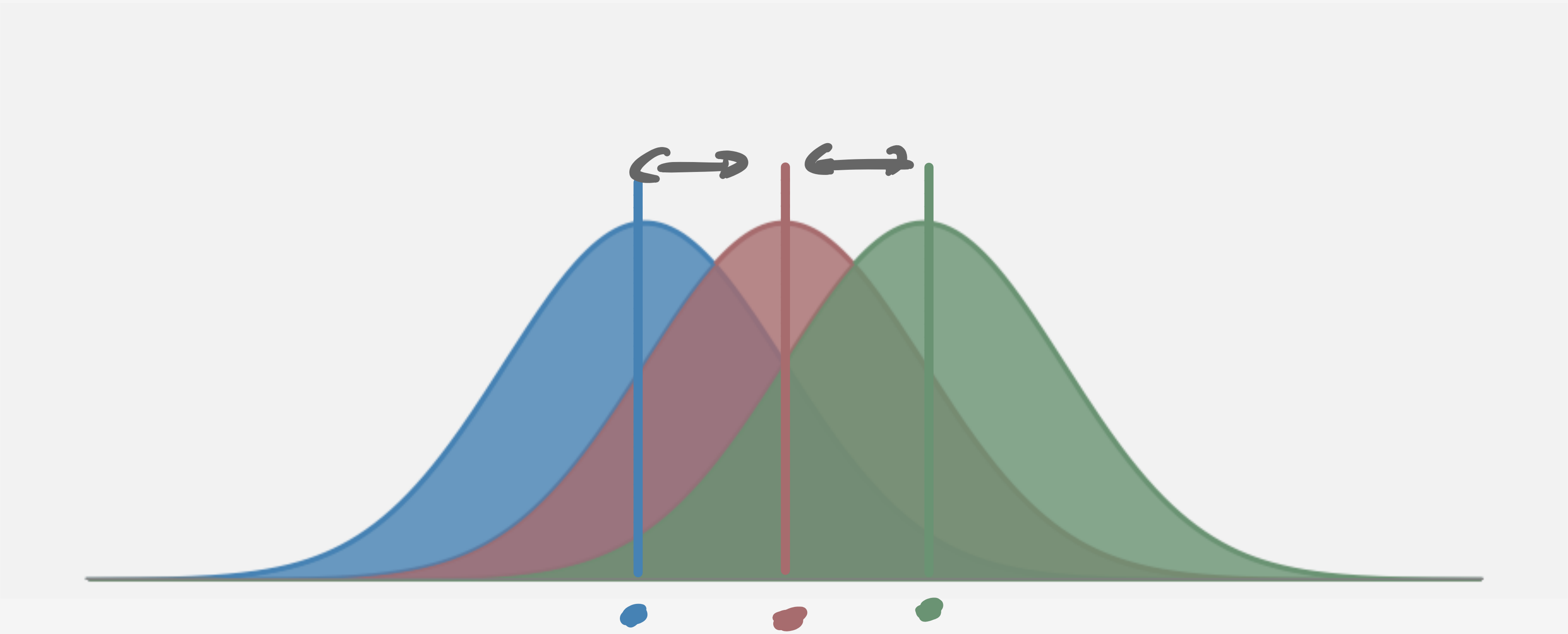
Why would a t- or z-test be problematic if we had many different groups?

EXPENSIVE + PROBLEM OF MULT COMPARISONS

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.

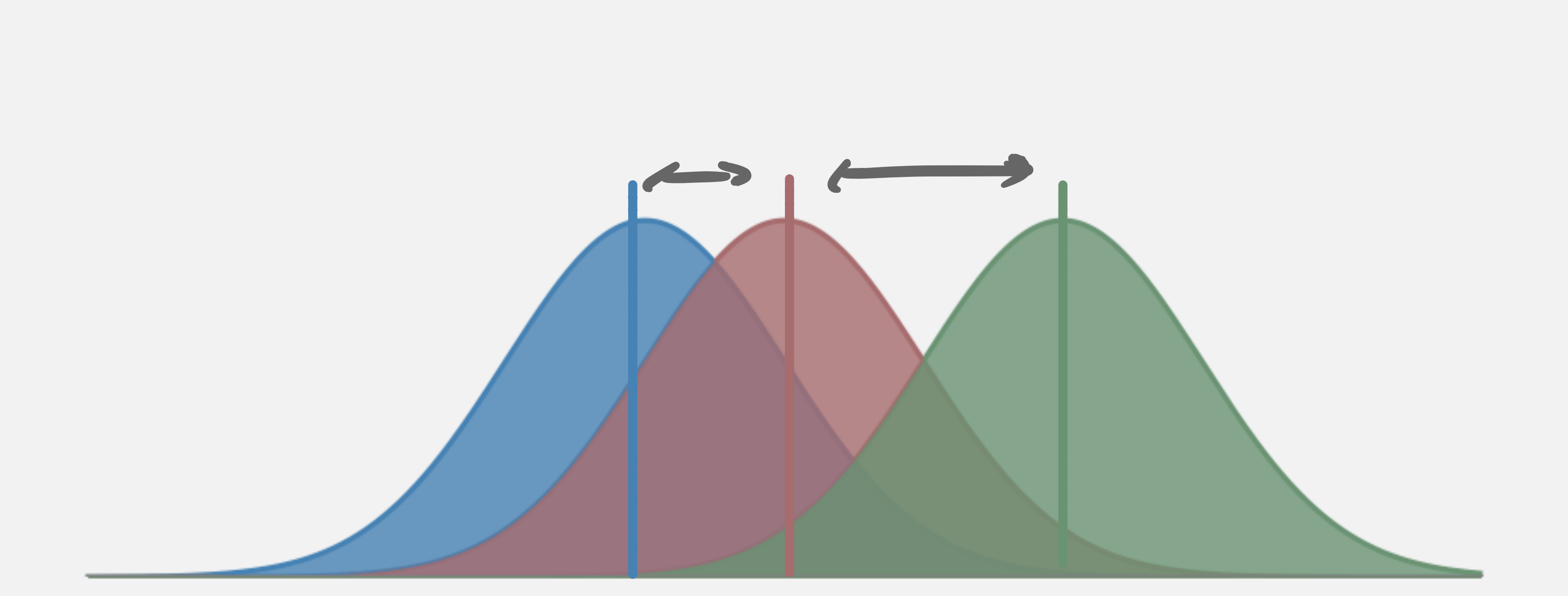
The **idea** is straightforward: Look at where the variance in the data comes from.



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.

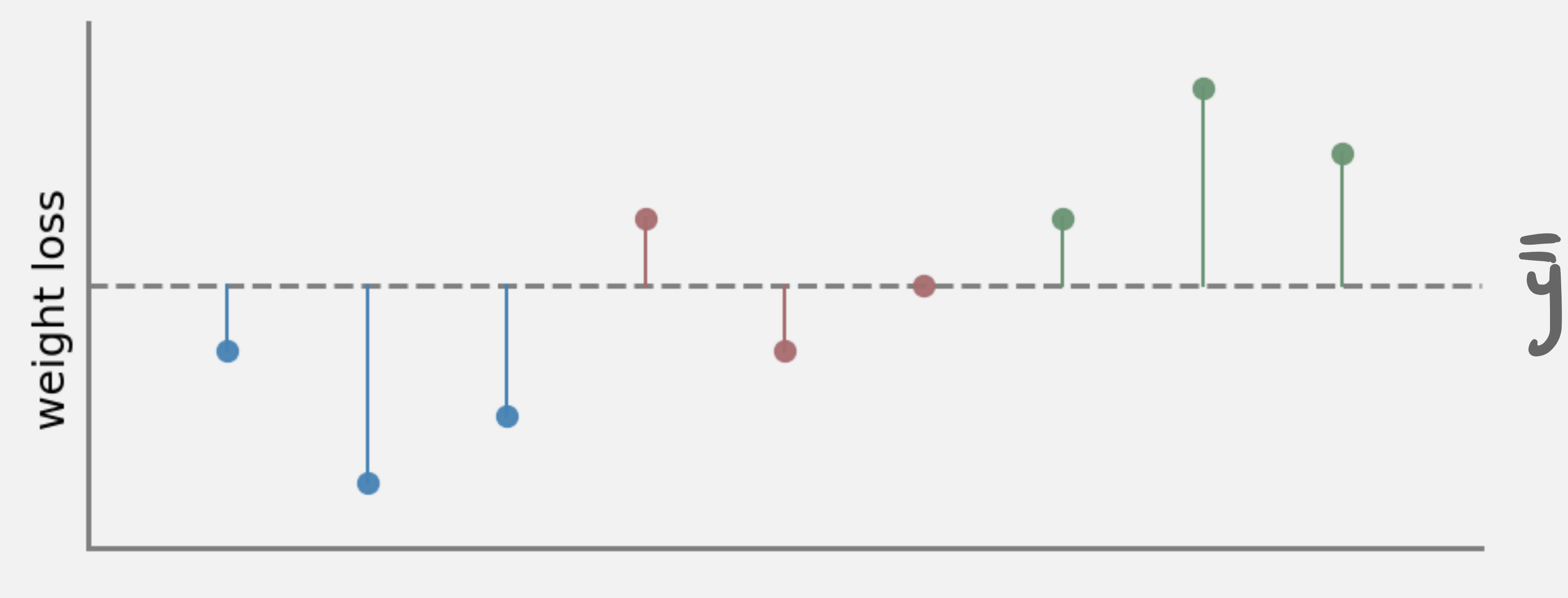
The **idea** is straightforward: Look at where the variance in the data comes from.



Analysis of Variance

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.

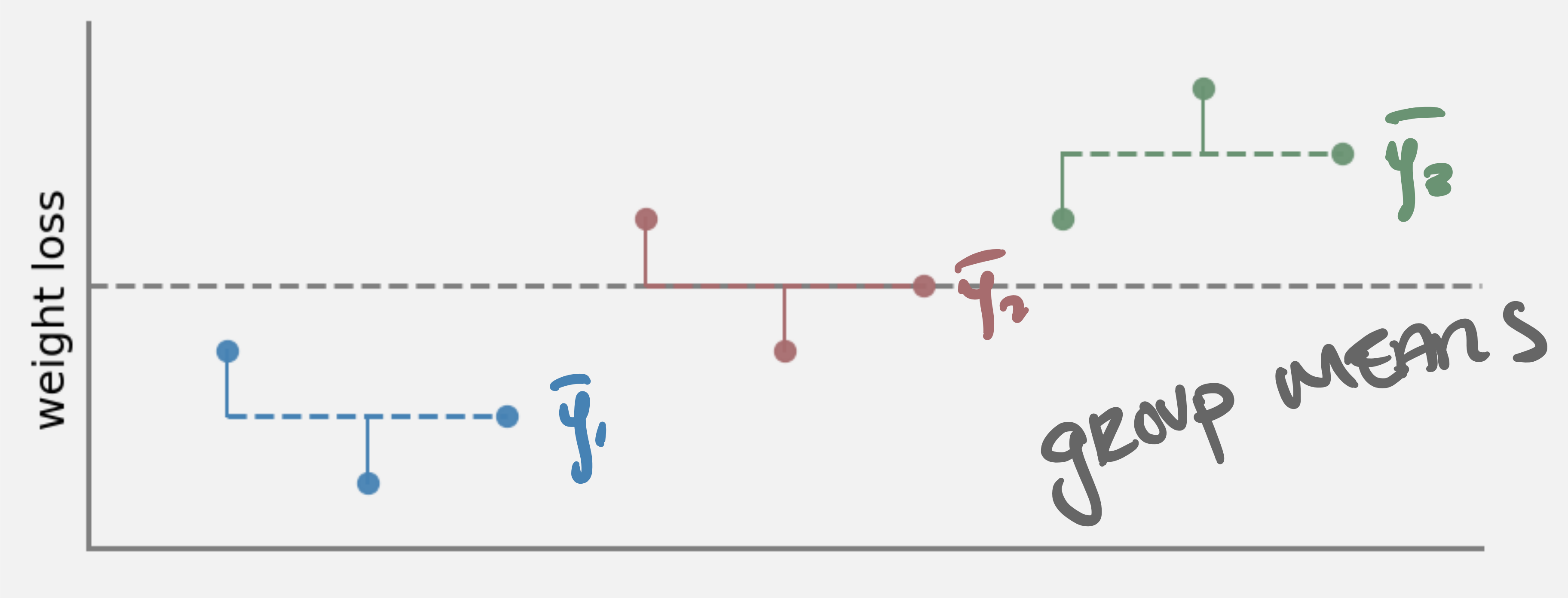


$\bar{\bar{y}}$ = GRAND MEAN

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

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

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

← RANDOM ERROR $\sim N(0, \sigma^2)$

POP MEAN OF GROUP i

j^{th} DATUM FROM GROUP i

Assumptions:

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

The One-Way ANOVA Model

Let's compute some means!

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

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

$$\bar{y} = (\underbrace{3+2+1}_{6} + \underbrace{5+3+4}_{12} + \underbrace{5+6+7}_{18})/9 = 4$$

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

$$\bar{y}_1 = (3+2+1)/3 = 2, \quad \bar{y}_2 = (5+3+4)/3 = 4, \quad \bar{y}_3 = (5+6+7)/3 = 6$$

NOTE: $N = n_1 + n_2 + n_3 = 3+3+3 = 9$

$$\bar{y} = \frac{n_1\bar{y}_1 + n_2\bar{y}_2 + n_3\bar{y}_3}{N} = \frac{3\cdot 2 + 3\cdot 4 + 3\cdot 6}{9} =$$

It's the Variances, Stupid

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

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

A helpful decomposition: DEVIATION FROM $\bar{\bar{y}}$ FOR single point

$$y_{ij} - \bar{\bar{y}} = (y_{ij} - \bar{y}_i) + (\bar{y}_i - \bar{\bar{y}})$$

WITHIN BETWEEN group

Then, a minor (mathematical) miracle occurs:

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

$$= SSW + SSB$$

The One-Way ANOVA Model $\bar{\bar{y}} = 4$

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

	Control	Diet A	Diet B
0	3	5	5
1	2	3	6
2	1	4	7
$\bar{y}_i$	2	4	6

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

$$SSB = 3(2-4)^2 + 3(4-4)^2 + 3(6-4)^2$$
$$= 3\cdot 4 + 3\cdot 0 + 3\cdot 4 = 24$$

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

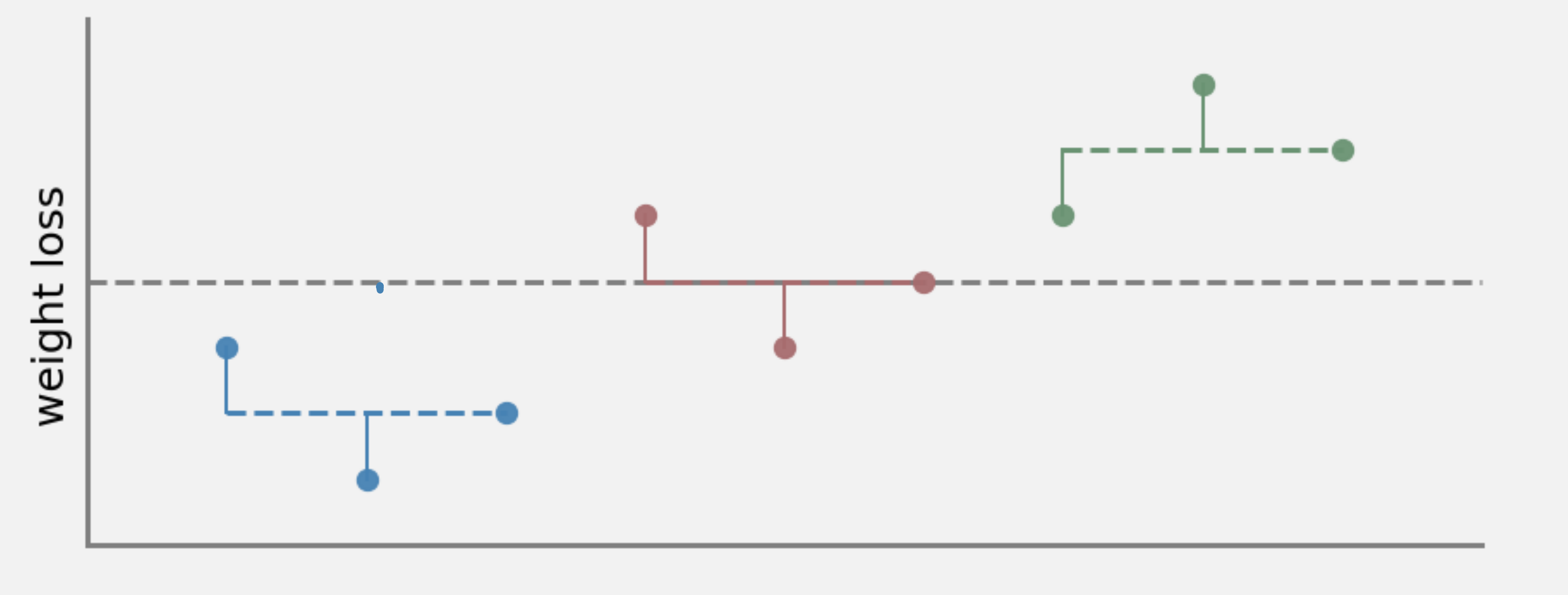
$$(3-2)^2 + (2-2)^2 + (1-2)^2 + (5-4)^2 + (3-4)^2 + (4-4)^2 + (5-6)^2 + (6-6)^2 + (7-6)^2 = 1+0+1+1+1+0+1+0+1 = 6$$

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

$$SST = SSB + SSW = 24 + 6 = 30$$

The One-Way ANOVA Model

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?

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

- The **BETWEEN** group degrees of freedom is:

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

$$SSB_{DF} = 3 - 1 = 2$$

$SSB\_df = I - 1$ (DATA IS $\bar{y}_i$, ESTIMATE $\bar{\bar{y}}$)

- The **WITHIN** group degrees of freedom is:

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

$$SSW_{DF} = 9 - 3 = 6$$

$SSW\_df = N - I$ (DATA IS y_{ij}, ESTIMATE I $\bar{y}_i$'s)

A Hypothesis Test

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

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

H_1 : $\mu_i \neq \mu_j$ FOR SOME (i,j)-PAIR

Our test statistic will be:

$$F = \frac{SSB/SSB_{DF}}{SSW/SSW_{DF}} = \frac{SSB/(I-1)}{SSW/(N-I)} \sim F_{I-1,N-I}$$

REJECTION REGION: $F \geq F_{\alpha, I-1, N-I}$

p-val = 1 − stats.f.cdf(F, 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

ANOVA	SS	DF	SS/DF	F
BETWEEN	24	2 3-1	$\frac{24}{2} = 12$	$\frac{12}{1} = 12$
WITHIN	6	6 9-3	$\frac{6}{6} = 1$	pval = 0.008
TOTAL	30	8		

ANOVA as Multiple Linear Regression

Interestingly, there is a very close relationship between One-Way ANOVA and MLR.

Suppose you have I groups that you want to compare. A random sample of size n_i is taken from the i^{th} group. Then

- CHOOSE ONE GROUP AS THE CONTROL

Model: $y_{ij} = \mu_0 + \tau_1 X_{1j} + \tau_2 X_{2j} + \cdots + \tau_{I-1} X_{I-1,j} + \epsilon_{ij}$

- y_{ij} IS j^{th} RESPONSE FOR i^{th} GROUP
- $X_{ij} = \begin{cases} 1 & \text{IF } j^{th} \text{ RESP FROM } i^{th} \text{ GROUP} \\ 0 & \text{ELSE} \end{cases}$

ANOVA as Multiple Linear Regression

Interestingly, there is a very close relationship between One-Way ANOVA and MLR.

Suppose you have I groups that you want to compare. A random sample of size n_i is taken from the i^{th} group. Then

		X_1	X_2
	Control	Diet A	Diet B
0	3	5	5
1	2	3	6
2	1	4	7

y_{ij}	x_{1j}	x_{2j}
3	0	0
2	0	0
1	0	0
5	1	0
4	1	0
3	1	0
5	0	1
6	0	1
7	0	1

ANOVA as Multiple Linear Regression

Interestingly, there is a very close relationship between One-Way ANOVA and MLR.

Suppose you have I groups that you want to compare. A random sample of size n_i is taken from the i^{th} group. Then

$$y_{ij} = \mu_0 + \tau_1 x_{1j} + \tau_2 x_{2j} + \epsilon_{ij}$$

MEAN RESPONSE FOR control

$x_{1j} = x_{2j} = 0$ $\quad y_{ij} = \mu_0$

MEAN RESPONSE FOR DIET A

$x_{1j} = 1$ & $x_{2j} = 0$ $\quad y_{ij} = \mu_0 + \tau_1$

MEAN RESPONSE FOR DIET B

$x_{ij} = 0$ & $x_{2j} = 1$ $\quad y_{ij} = \mu_0 + \tau_2$

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

ANOVA as Multiple Linear Regression

Interestingly, there is a very close relationship between One-Way ANOVA and MLR.

Suppose you have I groups that you want to compare. A random sample of size n_i is taken from the i^{th} group. Then

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

CONTROL: μ_0 (μ_1)

DIET A : $\mu_0 + \tau_1$ (μ_2)

DIET B = $\mu_0 + \tau_2$ (μ_3)

THINK OF τ_1 & τ_2 AS THE TREATMENT EFFECTS FOR THE TWO DIETS

ANOVA as Multiple Linear Regression

Interestingly, there is a very close relationship between One-Way ANOVA and MLR.

Suppose you have I groups that you want to compare. A random sample of size n_i is taken from the i^{th} group. Then

$$y = \mu_0 + \tau_1 X_1 + \tau_2 X_2$$

MLR F-TEST:

$H_0: \tau_1 = \tau_2 = 0$

$H_1: \tau_k \neq 0$ FOR SOME k

$\Longleftrightarrow$ ANOVA EQUIV

H_0: All MEANS $= \mu_0$

H_1: At LEAST ONE MEAN $\neq \mu_1$

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

Tukey's Honest Significance Test

Suppose that we determine that some of the means are different.

How can we tell which ones? TUKEY'S HSD OR TUKEY'S RANGE TEST

HYPOTHESIS TESTS FOR PAIRWISE COMPARISON OF MEANS.

FIXES PROBLEM OF MC'S.

ADJUSTS SO THAT MAKING A TYPE I ERROR OVER ALL PAIRWISE-COMPARISONS IS α.

OK! Let's Go to Work!

Get in groups, get out laptop, and open the Lecture 24 In-Class Notebook

Let's:

- Figure out how to do ANOVA in Python
- See the connection between ANOVA and MLR
- See how to do Tukey's HSD in Python