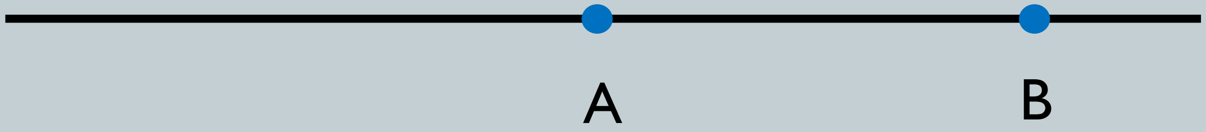
ANALYSIS OF VARIANCE

ANOVA

During your studies and this lecture, consider the next four ideas/analogies:

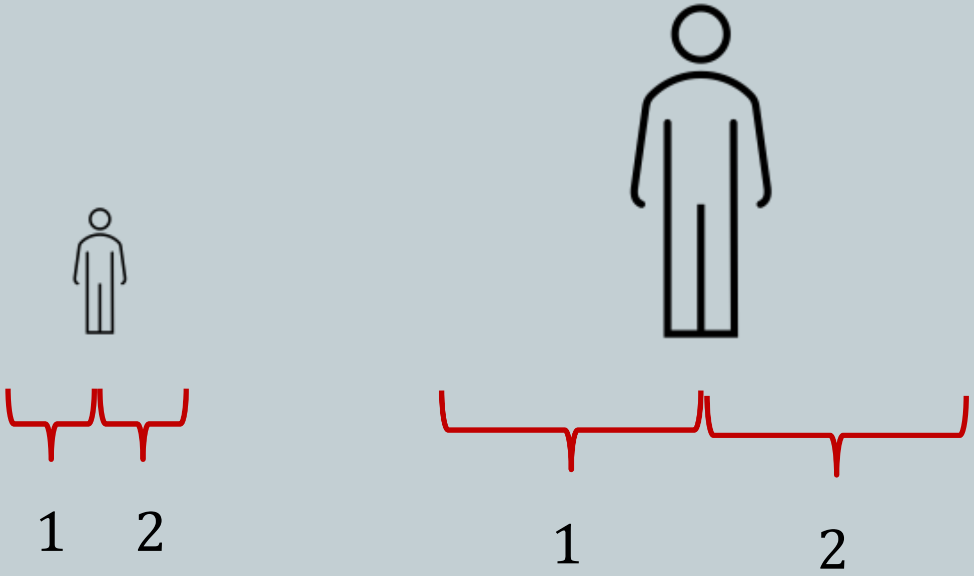
Idea 1

Can you get from point A to point B, or past, in 2 steps?



It depends…

How big are your ‘steps’ ?

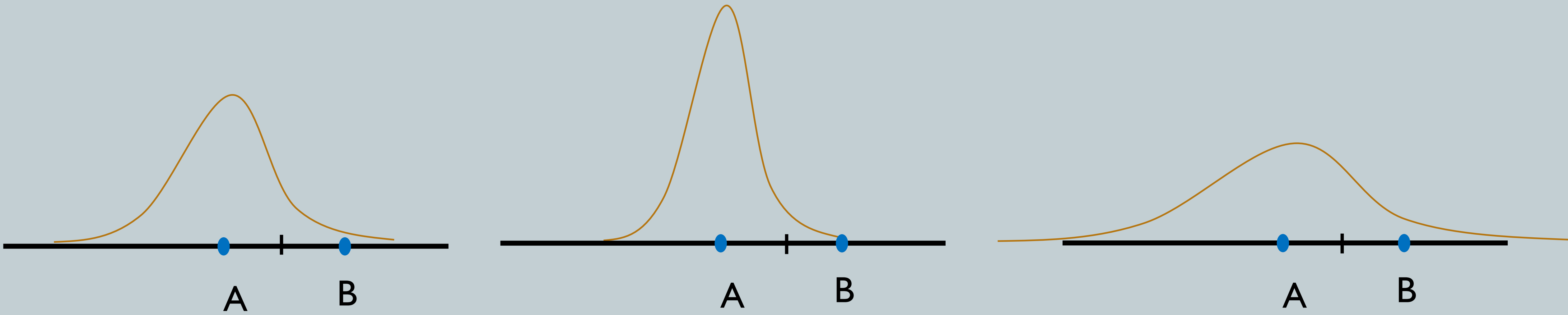


Idea 2

If you take two steps from point A, will there be 2.5% left in the tail?

It depends…

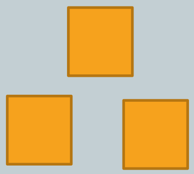
What is the shape of the distribution ? Mesokurtic, Leptokurtic, Platykurtic,…



Idea 3

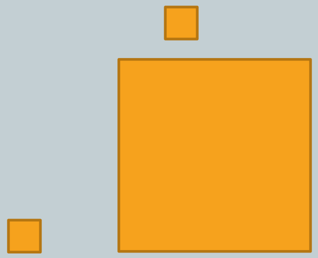
Which of these groups of squares add up to the smallest total area ?

[A]



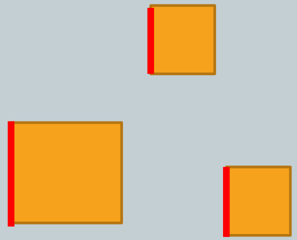
$2^2 + 2^2 + 2^2 = 12$

[B]

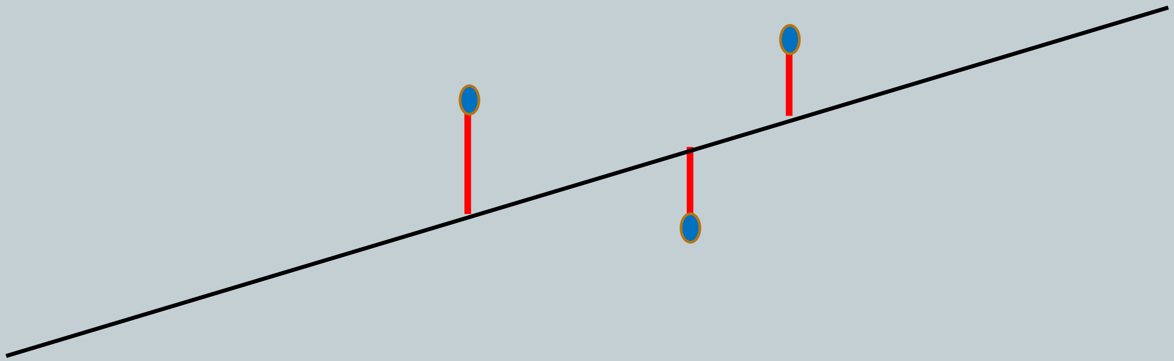


$1^2 + 1^2 + 4^2 = 18$

[C]



$2^2 + 2^2 + 3^2 = 17$



Idea 4

If a sample was taken from one of these groups and we calculated the sample mean, then how can we determine which group the sample came from?

0	1	2	3	4	5	6	7	8	9	10	11	12	13

$\bar{x} = 6.5$

0	1	2	3	4	5

$\bar{x} = 2.5$

4	5	6	7	8	9

$\bar{x} = 6.5$

8	9	10	11	12	13

$\bar{x} = 10.5$

For example, if we draw a sample and it has a mean of **4.5**, then which group did it come from?

0	1	2	11	12	13

$\bar{x} = 6.5$

How different are the blue group and the orange group?

How different are the blue group and the red group?

How variable are the numbers within each group?

Let's review the single sample t-test and the 2-sample t-test as a lead-in to the question:
What do we do if we are making more than 1 or 2 comparisons?

t-test: How does this 1 single score compare to the population?

2-sample t-test: How do these 2 scores compare?

ANOVA: How do these 3 (or more) scores compare?

There are numerous types of ANOVA, it is a whole family of techniques.

Suppose there are 6000 freshmen at CU, and we want to know the group mean-score on a study skills test.

We can't give the exam to all 6000 students as it is simply impractical.
We generate an SRS of 20 freshmen, give them the test, and draw inferences from the data.

Mean score of sample:	$\bar{x} = 74.35$
Sample size:	$n = 20$
Sample standard deviation:	$s = 13.14$

With this information we can calculate a 95% confidence interval using a critical t-value:

$$t_{.975,\,19} = 2.09 \;\; \text{implies} \;\; \bar{x} \pm t \cdot \frac{s}{\sqrt{n}} = 74.35 \pm 2.09 \cdot \frac{13.14}{\sqrt{20}} = [68.21, 80.49]$$

Meaning, we are confident that 95% of intervals such as this will contain the population mean.

Mean score: $\bar{x} = 74.35$
Sample size: $n = 20$
Sample standard deviation: $s = 13.14$

Furthermore, if we know the historical mean of previous freshmen classes (perhaps the previous population mean is 78.5) then we could calculate a t-score and a p-value:

$t = \frac{\bar{x}-\mu}{s/\sqrt{n}} = \frac{74.35 - 78.5}{13.14/\sqrt{20}} = -1.412$ and `stats.t.cdf(-1.412, 19) = 0.0870594`

So, the probability that the mean score of these freshmen will be no more than 74.35 is about 8.7%.

The p-value indicates that if H_0: $\mu = 78.5$ and H_1: $\mu \neq 78.5$ we would fail to reject the null at $\alpha = .05$. (2*0.087=p-value)

We could also fail to reject H_0 because $-1.412 > t_{.025,19} = -2.09$.

Now, suppose there are 6000 freshmen and 5500 sophomores at CU.
We want to know if a particular population (freshmen or sophomores) scores significantly different than the other group on a study skills exam.

We could test all $11{,}500$ people but that is unrealistic.
Instead, we sample 10 freshmen and 10 sophomores.

$\bar{x}_{Fr} = 79.44$ with $s_{Fr} = 14.972$
$\bar{x}_{Sp} = 86.81$ with $s_{Sp} = 13.819$

We see the means obviously aren't the same, but how different are they? Are the Freshmen and Sophomores really two different populations (by study habits) or do they come from the same population ?

Are these averages "close enough" for us to conclude that mean scores are the same for the larger population of freshmen and sophomores at CU?

Or are the averages too different for us to make that conclusion?

Formulate an answer by looking at the difference in the means: $79 - 86 = -7$.

This difference in our samples estimates the difference between the population means for the two groups.

First, we decide on the risk we are willing to take for declaring that a significant difference exists.

Suppose we are willing to take a 5% risk of saying that the unknown population means for freshmen and sophomores are not equal when they really are.

This is what we mean by declaring $\alpha = .05$, it is the risk we are willing to take to commit a type 1 error; rejecting H_0 in favor of H_1 when we shouldn't have.

Then calculate the test statistic.

The standard error for each group gets us the test statistic:

$$t = \frac{\text{difference of group averages}}{\text{standard error of difference}} \quad \text{or} \quad t = \frac{(\overline{x_1} - \overline{x_2})}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = -4.343$$

Compare this test statistic to a theoretical value from the *t*-distribution.

Find the theoretical value from the *t*-distribution based on our null hypothesis which states that the means for freshmen and sophomores are equal.

$H_0\!: \mu_{Fr} = \mu_{Sp}$ and $H_1\!: \mu_{Fr} \neq \mu_{Sp}$

To find this value, we need both the significance level and the *degrees of freedom*.

The degrees of freedom (*df*) are based on the sample sizes of the two groups.
For the skills test data, this is:

$$df = 10 - 1 + 10 - 1 = 10 + 10 - 2 = 18$$

The critical *t*-value with $\alpha = .05$ and 18 degrees of freedom is 2.101

```
stats.t.ppf(.025, 18) = -2.10092204024096
```

We compare the value of our critical statistic -2.101 to the *t* value -4.343 .

Since -4.343 <-2.101 we reject the null hypothesis that the mean skills-test score for freshmen and sophomores are equal and conclude that we have evidence the score in the population is different between freshmen and sophomores.

`2*stats.t.cdf(-4.343, 18) = 0.0003919 < 0.05` also implies reject null.

Why do we use ANOVA?

What if we wish to compare the means of more than two populations?
(Freshmen, Sophomores, Juniors instead of just Freshmen and Sophomores.)

Now suppose you have data samples from several populations and are wondering whether the populations have different means.

One-way ANOVA answers that question.

Suppose we now want to compare three populations means (Freshmen, Sophomores, Juniors) to see if a difference exists somewhere among them.

What we are asking is:

Do all three of these means come from a common population?

Is one mean so far away from the other two that it is likely not from the same population?

Or are they all so far apart from each other that they all likely come from **unique** populations? (Study skills define the given population)

μ_{Fr} μ_{Sp} μ_{Jn}

Suppose there are 6000 Freshmen, 5500 Sophomores, and 5000 Juniors at CU. We want to compare the study skills of these three groups to see if they are the same or significantly different.

We can't test all 16,500 of them so we randomly sample 7 students from each group.

We are interested in whether or not a difference exists somewhere between the three different classes of students.

We will conduct this analysis using a One-Way ANOVA technique.

The data from the three samples is as follows:

Group 1	Group 2	Group 3
Freshmen	Sophomores	Juniors
82	71	64
93	62	73
61	85	87
74	94	91
69	78	56
70	66	78
53	71	87
$\bar{x}_1 = 71.71$	$\bar{x}_2 = 75.29$	$\bar{x}_3 = 76.57$

How varied are these means?

The Grand Mean: The mean of all 21 scores taken together.

$\bar{\bar{x}} = 74.52$

A glance at the column means tells us something about freshmen… maybe.

ANOVA is an analysis of variance.

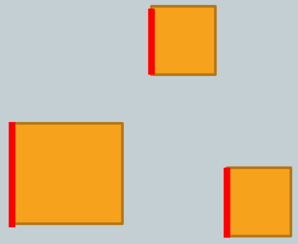
Variance is the average squared deviation (difference) of a data point from the distribution mean.

Take the distance of each data point from the mean, square each distance, add them together, and then find the average.

$$\textbf{Sample}\ \text{Variance} = s^2 = \frac{\sum(x-\mu)^2}{n-1}$$

Take out the phrase "find the average" part and we are left with just the Sum of Squares (SS).

$$\text{Sum of Squares} = \sum(x-\mu)^2$$



$$2^2 + 2^2 + 3^2 = 17$$

Sum of the squares of the difference of the dependent variable and its mean.

So, SS is variance without finding the average of the sum of the squared deviations.

This is what we are working up to:

SST is the total/overall sum of squares

SSB is the column/between/treatment/group sum of squares

SSW is the within/error sum of squares.

SST = SSB + SSW

$$F = \frac{\frac{SSB}{G-1}}{\frac{SSW}{N-G}} = \frac{MSB}{MSW}$$

F is a ratio of variances that we will analyze.

SST (Sum of Squares Total)

1. Find the difference between each data point and the overall mean, $\bar{\bar{x}} = 74.52$.
2. Square the difference.
3. Add them up.

Group 1	Group 2	Group 3
Freshmen	Sophomores	Juniors
82	71	64
93	62	73
61	85	87
74	94	91
69	78	56
70	66	78
53	71	87
$\bar{x}_1 = 71.71$	$\bar{x}_2 = 75.29$	$\bar{x}_3 = 76.57$

$$(82 - 74.52)^2 + (93 - 74.52)^2 + \cdots + (71 - 74.52)^2 + (62 - 74.52)^2 + \cdots + (64 - 74.52)^2 + (73 - 74.52)^2 + \cdots + (87 - 74.52)^2$$

SST

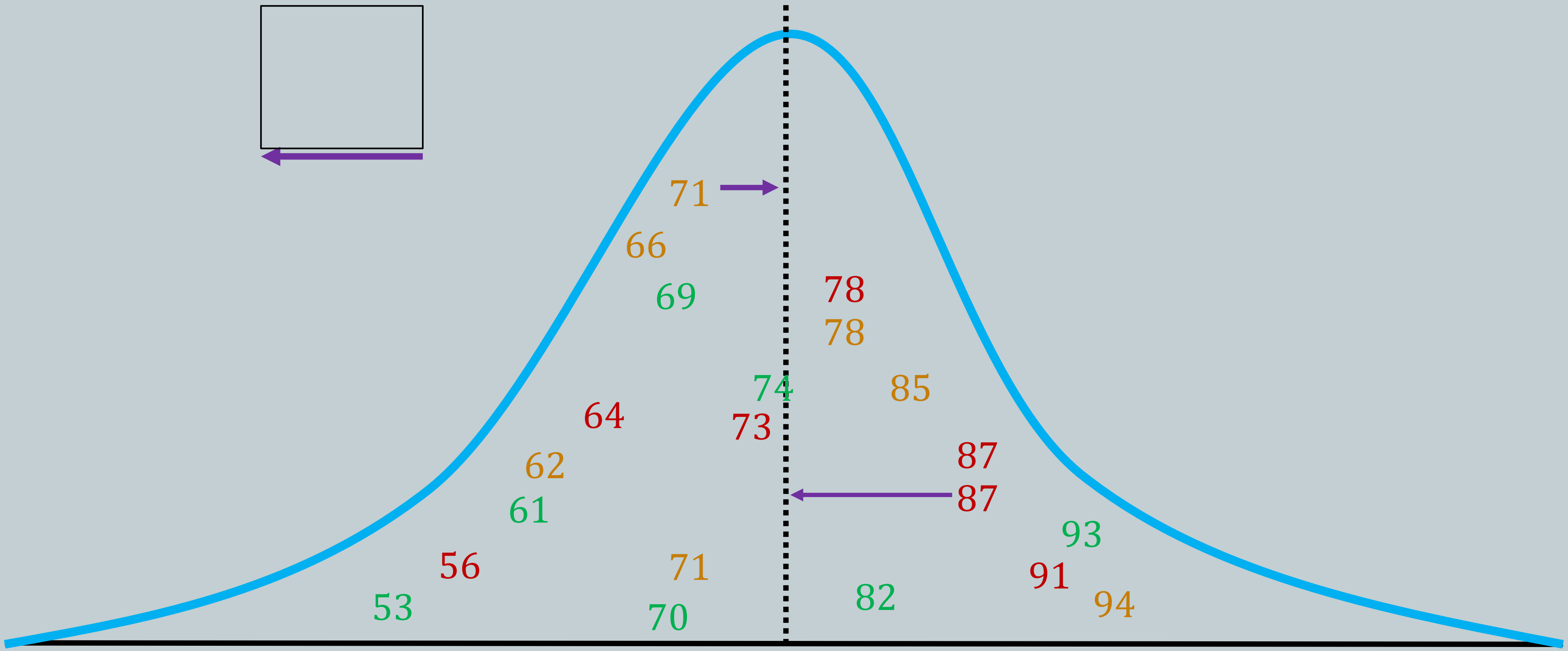
Find difference between each data point and the overall mean

Square the difference.

Add them up

In this case there would be 21 squared deviations.

$\bar{\bar{x}} = 74.52$



SSB (Sum of Squares of Columns)

Find difference between each group mean and the overall mean: $\bar{\bar{x}} = 74.52$

Square the deviations.

multiply by group size

Add them up.

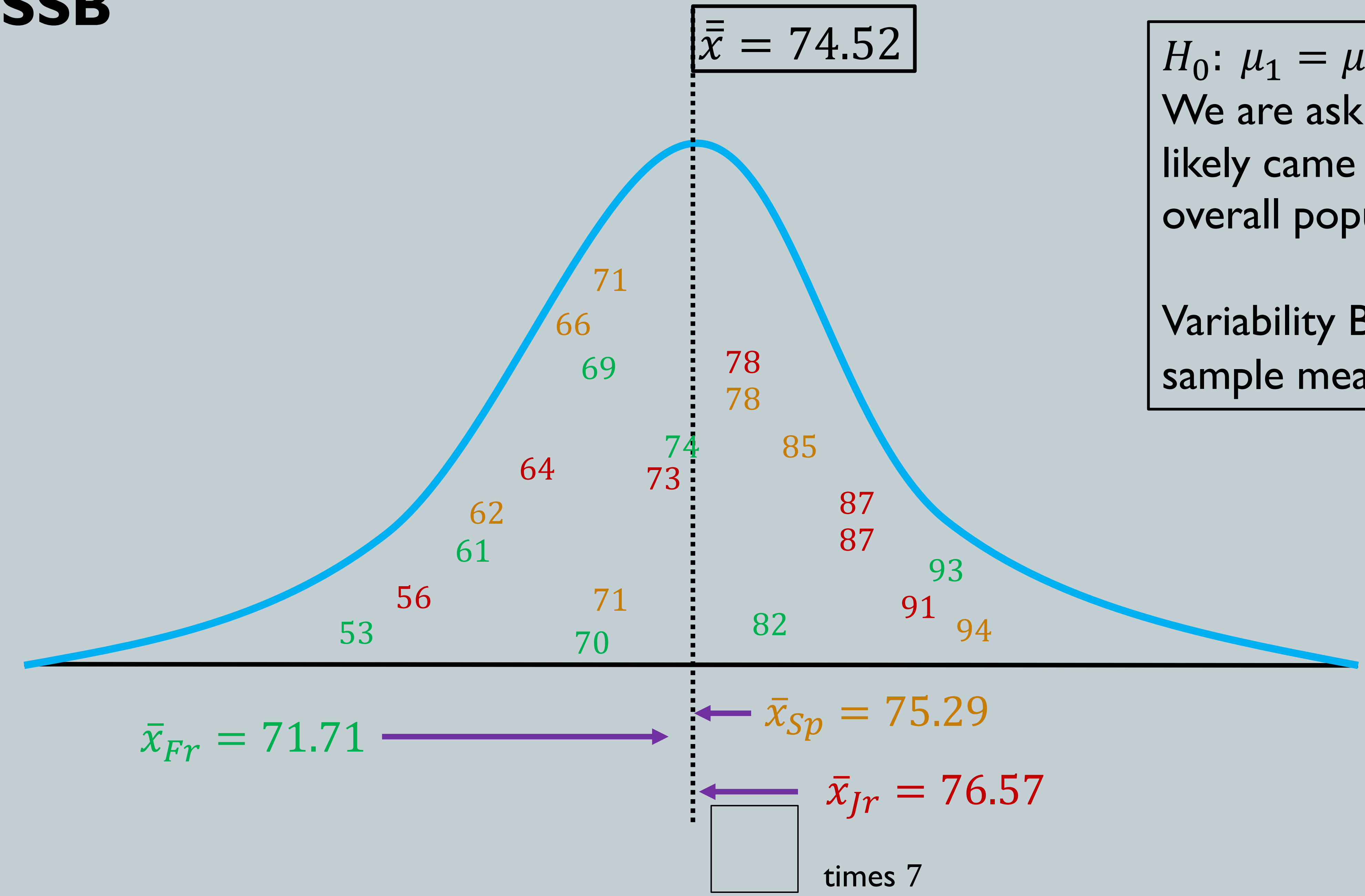
In this case we would have 3 squared deviations 7 times.

(We multiply by 7 since we are doing this computation for each data point)

Group 1	Group 2	Group 3
Freshmen	Sophomores	Juniors
82	71	64
93	62	73
61	85	87
74	94	91
69	78	56
70	66	78
53	71	87
$\bar{x}_1 = 71.71$	$\bar{x}_2 = 75.29$	$\bar{x}_3 = 76.57$

$$SSB = 7 \cdot (71.71 - 74.52)^2 + 7 \cdot (75.29 - 74.52)^2 + 7 \cdot (76.57 - 74.52)^2$$

SSB



H_0: $\mu_1 = \mu_2 = \mu_3$
We are asking if each mean likely came from the larger overall population.

Variability BETWEEN the sample means

SSW (Sum of Squared estimate of Errors)

Find difference between each data point and its column mean

Square each deviation.

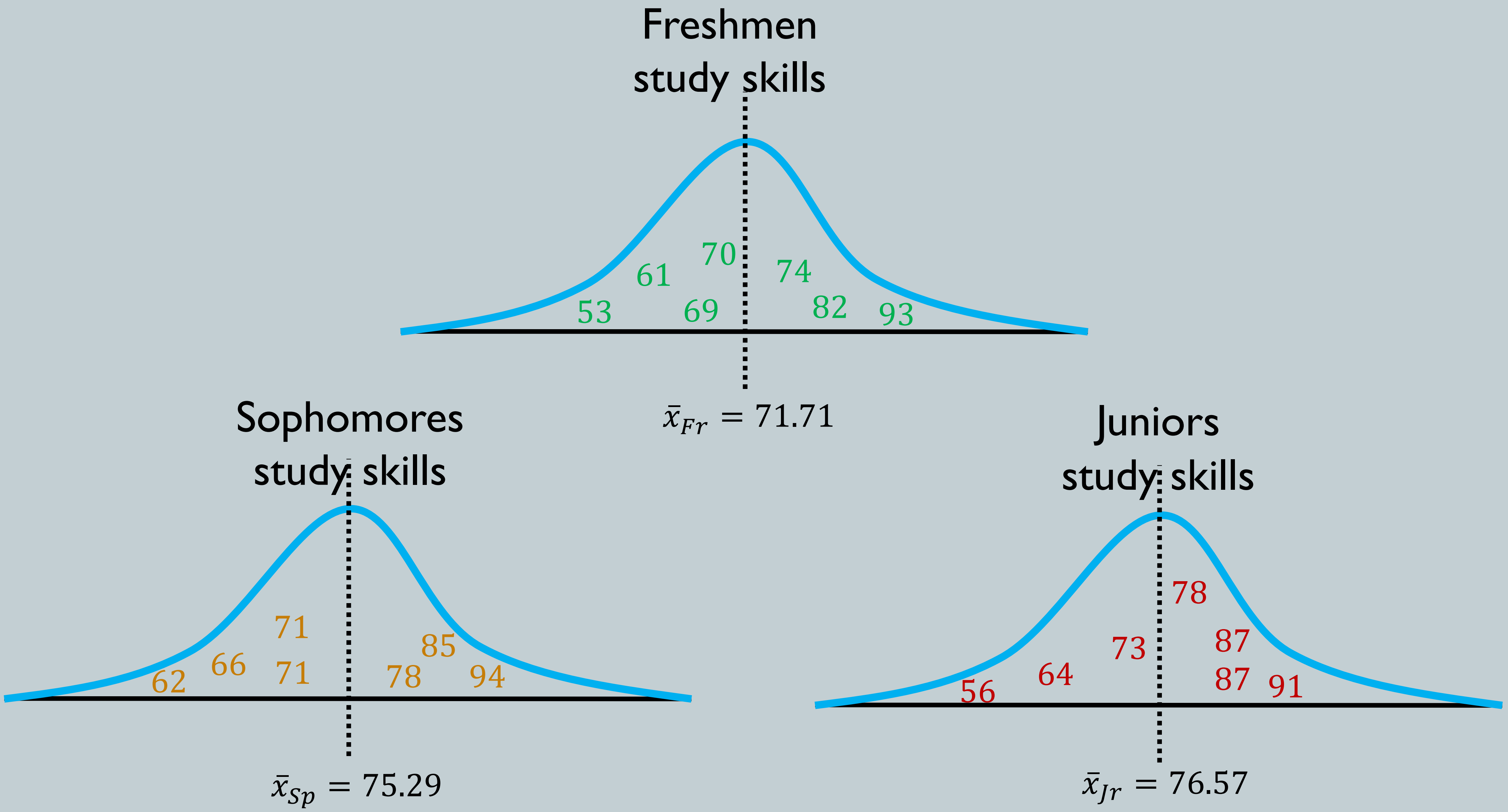
Add up the squared deviations.

In this case we would have 21 squared deviations.

Group 1	Group 2	Group 3
Freshmen	Sophomores	Juniors
82	71	64
93	62	73
61	85	87
74	94	91
69	78	56
70	66	78
53	71	87
$\bar{x}_1 = 71.71$	$\bar{x}_2 = 75.29$	$\bar{x}_3 = 76.57$

$$SSW = (82 - 71.71)^2 + (93 - 71.71)^2 + \cdots + (53 - 71.71)^2 +$$
$$(71 - 75.29)^2 + (62 - 75.29)^2 + \cdots + (71 - 75.29)^2 +$$
$$(64 - 76.57)^2 + (73 - 76.57)^2 + \cdots + (87 - 76.57)^2 .$$

SSW


Freshmen
study skills
53 61 69 70 74 82 93
$\bar{x}_{Fr} = 71.71$
Sophomores
study skills
62 66 71 71 78 85 94
$\bar{x}_{Sp} = 75.29$
Juniors
study skills
56 64 73 78 87 87 91
$\bar{x}_{Jr} = 76.57$

Formulas for one-way ANOVA

$$MSB = \frac{SSB}{df_{groups}}$$

$$df_{groups} = G - 1$$

G is the number of columns

N is the number of observations

$$MSW = \frac{SSW}{df_{within}}$$

$$df_{within} = N - G$$

$$SST$$

$$df_{total} = N - G + G - 1 = N - 1$$

$$F = \frac{MSB}{MSW}$$

Formulas for one-way ANOVA

$MSB = \frac{SSB}{2}$

$df_{groups} = 3 - 1 = 2$

$MSW = \frac{SSW}{18}$

$df_{within} = 21 - 3 = 18$

G is the number of columns

N is the number of observations

SST

$df_{total} = 21 - 1 = 20$

$F = \frac{MSB}{MSW}$

We are analyzing 3 groups (Freshmen, Sophomores, Juniors)

There are 7 people in each group, $3 \cdot 7 = 21$

Formulas for one-way ANOVA

$$MSB = \frac{88.67}{2} = 44.33$$

$$df_{groups} = 2$$

G is the number of columns

N is the number of observations

$$MSW = \frac{2812.57}{18} = 156.25$$

$$df_{within} = 18$$

$$SST$$

$$df_{total} = 20$$

$$F = \frac{44.33}{156.25} = 0.28$$

The F-statistic is ratio of variances and guides our decision concerning whether or not the groups are from different populations.

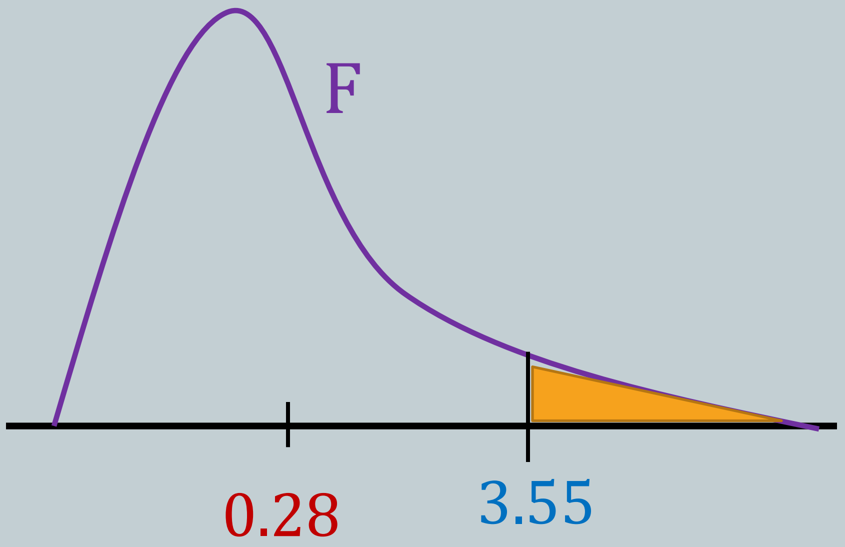
Source of Variance	df	SS	MS	F
Between (columns, groups)	2	88.67	44.33	0.28
within (error)	18	2812.57	156.25	
total	20	2901.24		

$$F = \frac{MSB}{MSW} = 0.28$$

$$F_{critical} = F_{\alpha, df_B, df_W} = F_{0.05,2,18} = 3.55$$

F-statistic is not larger than F-critical.

```
stats.f.ppf(.95,2,18) = 3.55
```



Fail to reject H_0.

No significant difference in mean test score by class rank (Freshmen, Sophomore, Junior).

Place the scores in a dataframe and call `stats.f_oneway` on the dataframe.

In [8]:
```
dfD = pd.DataFrame({"Freshmen": [82, 93, 61, 74, 69, 70, 53],
                    "Sophomores": [71, 62, 85, 94, 78, 66, 71],
                    "Juniors": [64, 73, 87, 91, 56, 78, 87] })
dfD.head(20)
```

Out[8]:

	Freshmen	Sophomores	Juniors
0	82	71	64
1	93	62	73
2	61	85	87
3	74	94	91
4	69	78	56
5	70	66	78
6	53	71	87

In [3]:
```
F, pval = stats.f_oneway(dfD["Freshmen"], dfD["Sophomores"], dfD["Juniors"])
print("F = {:.5f}".format(F))
print("pval = {:.5f}".format(pval))
```

```
F = 0.28373
pval = 0.75628
```

Why not just run multiple t-tests?

Pairwise comparison for this example means three t-tests ALL with an $\alpha = 0.05$ type 1 error rate at 95% confidence.

$$H_0: \bar{x}_1 = \bar{x}_2 \;; \quad \alpha = 0.05$$

$$H_0: \bar{x}_1 = \bar{x}_3 \;; \quad \alpha = 0.05$$

$$H_0: \bar{x}_2 = \bar{x}_3 \;; \quad \alpha = 0.05$$

This error compounds with each t-test:

$(.95)(.95)(.95) = .857$ which means $\alpha = 1 - 0.857 = 0.143$

The type 1 error rate then rises to 0.143.

The probability that you incorrectly reject the null is going up!

The Tukey's honestly significant difference test (Tukey's HSD)

This is a test used to check differences among all the sample means for significance.
The Tukey's HSD tests all pairwise differences while controlling the probability of making one or more Type 1 errors!

The Tukey's HSD test is one of several tests designed for this purpose and fully controls this Type 1 error rate.

This test (Tukeys) compares all possible pairs of means.

The Tukey test is invoked when you need to determine if the interaction among three or more variables is mutually statistically significant, which unfortunately is not simply a sum or product of the individual levels of significance.

The Tukey's honestly significant difference test (Tukey's HSD)

An ANOVA test can tell you if your results are significant overall, but it won't tell you which pairs are different.

After you have run an ANOVA and found significant results (i.e., at least one mean is different from the others), then run Tukey's HSD.

A common mistaken belief is that the Tukey HSD should only be used following a significant ANOVA. Actually, the ANOVA is not necessary because the Tukey test controls the Type 1 error rate on its own.

NoteBook #24

```
from statsmodels.stats.multicomp import MultiComparison
mc = MultiComparison(data, labels)
result = mc.tukeyhsd()
```

Like ANOVA's `.summary`, we will simply grab our numbers for Tukey's HSD from `.tukeyhsd()` , but the idea behind the Tukey HSD test is to focus on the largest value of the difference between two group means.

The relevant statistics is: $q = \frac{\bar{x}_{max} - \bar{x}_{min}}{\sqrt{MS_W/_n}}$

$n =$ the size of each of the group samples.
q has a distribution called the studentized range q.

The critical values for this distribution are presented in the 'Studentized Range q table" based on, a, the values of the number of groups and df_W.

If $q > q_{critical}$ then the two means are significantly different.

Back to ANOVA, our variability ratio

Variability Between the means

———

Variability Within the means

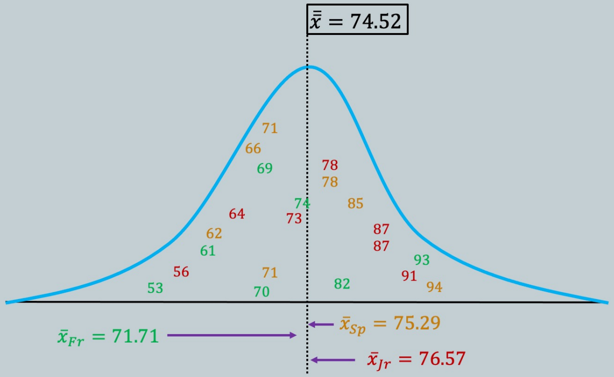
=

Distance from overall mean

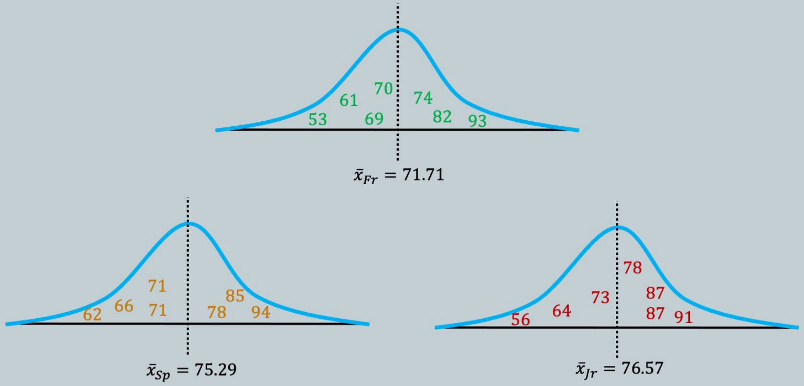
———

Internal Spread

=



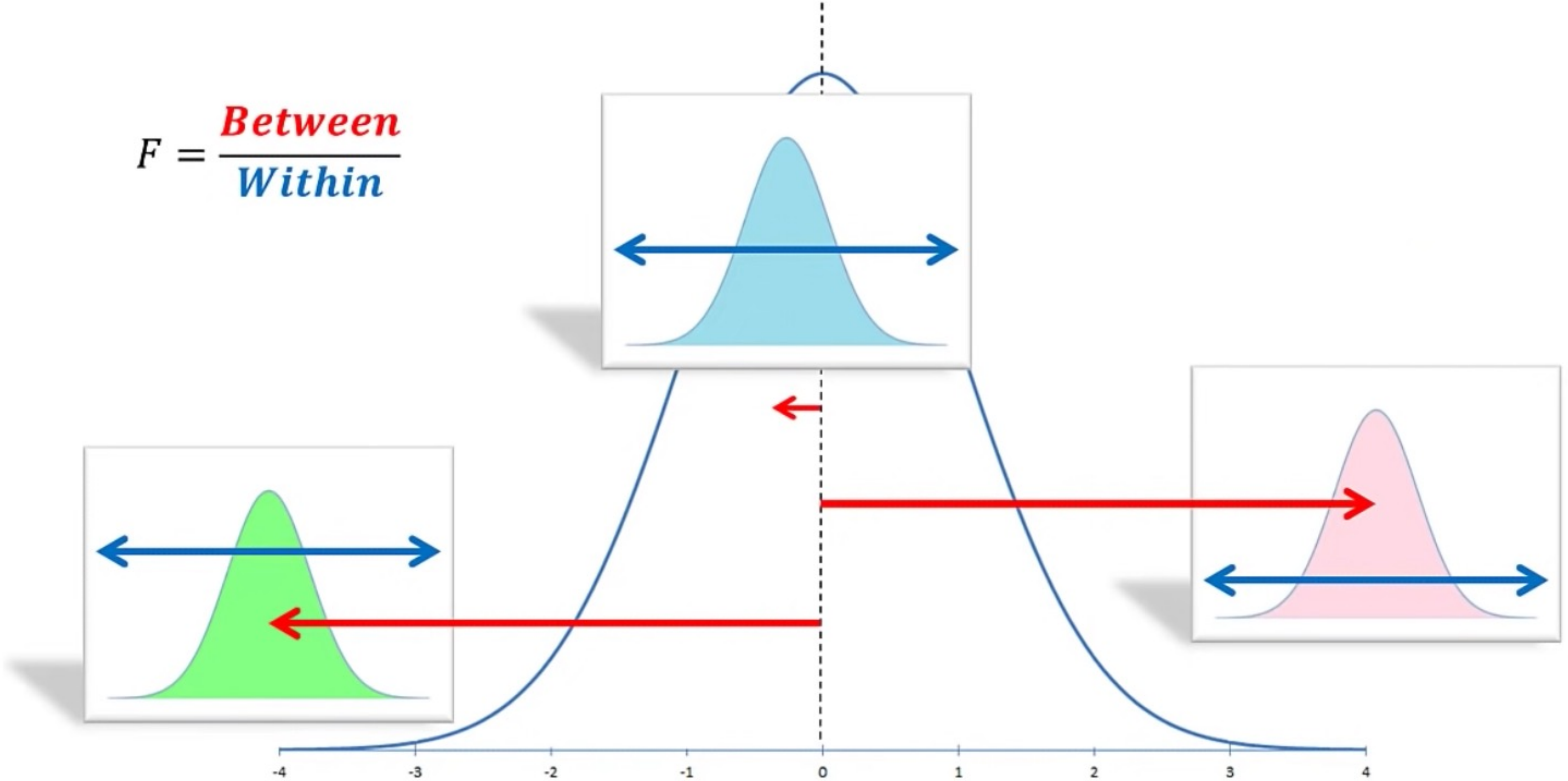
———



ANOVA: Analysis of Variance is a *variability ratio*

$$Variance\ Between + Variance\ Within = Total\ Variance$$

$$F = \frac{Between}{Within}$$



We know $Variance\ Between + Variance\ Within = Total\ Variance,$

but how do we interpret this $\frac{Variance\ Between}{Variance\ Within}$?

If the variability BETWEEN the means (distance from overall mean) in the numerator is relatively large compared to the variance WITHIN the samples (internal spread) in the denominator, the ratio will be much larger than 1.

Therefore, the samples most likely do NOT come from a common population;
Hence, we reject the null hypothesis that means are equal.

$$\frac{\text{Variance Between}}{\text{Variance Within}} = \text{some ratio}$$

$$\frac{LARGE}{small} = Reject\ H_0$$

At least one mean is an outlier and each distribution is narrow; distinct from each other

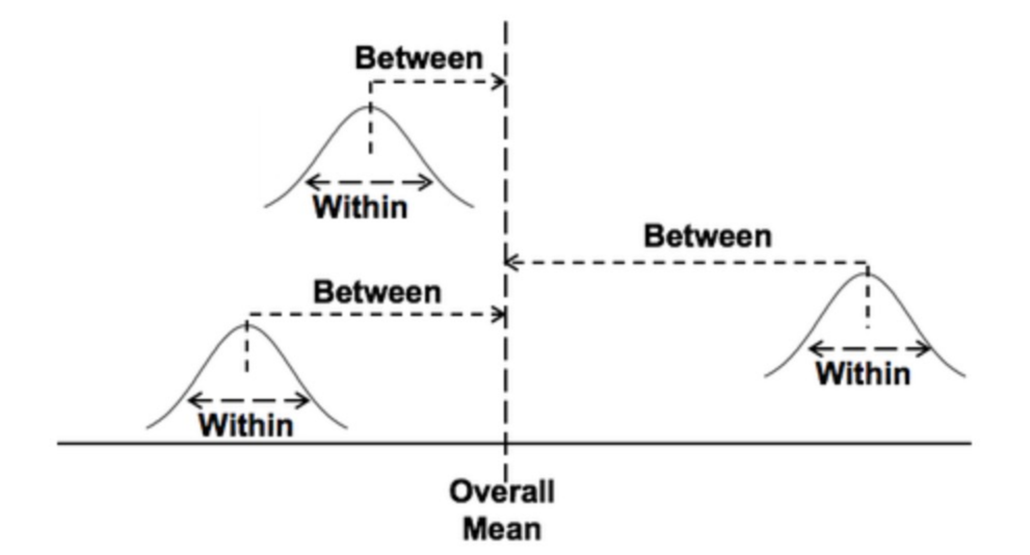
$$\frac{similar}{similar} = Fail\ to\ Reject\ H_0$$

Means are fairly close to overall mean and/or distributions overlap a bit; hard to distinguish

$$\frac{small}{Large} = Fail\ to\ Reject\ H_0$$

The means are very close to overall mean and/or distributions "melt" together.

In Summary:



Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	F
Within	$SSW = \sum_{k=1}^{g}\sum_{j=1}^{s}(x_j - \bar{x}_k)^2$	$N - g$ or $N - c$	$MSW = \frac{SSW}{N-g}$	$F = \frac{MSB}{MSW}$
Between	$SSB = \sum_{k=1}^{g}(\bar{x}_k - \bar{\bar{x}})^2$	$g - 1$ or $c - 1$	$MSB = \frac{SSB}{g-1}$	
Total	$SST = \sum_{i=1}^{n}(x_i - \bar{\bar{x}})^2$ $SST = SSB + SSW$			N is number of data points in total g number of groups (columns) s is amount in group sample

ANOVA

Imagine three groups of people.

All of them exercise.

Group A is on a vegetable diet.

Group B is on a grapefruit diet.

Weekly weight loss is then measured.

Control	**Group A**	**Group B**
3	5	5
2	3	6
1	4	7
$\bar{x} =$	$\bar{x} =$	$\bar{x} =$

$$F = \frac{MSB}{MSW}$$

ANOVA is also valuable in analyzing the MLR models.

Previously, we had been making ‘models’ to guide our predictive efforts.

Models quantify the relationships between our chosen variables.

We started with SLR: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$

We are given the data for x and y.

Our goal is to fit the model which will give us the best estimates for β_0 and β_1.

This concept generalizes naturally to MLR,

where we have multiple variables on the righthand side of the relationship:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i + \cdots + \beta_p x_i + \epsilon_i$$

The great part of modeling with Python is that anyone can build these linear models.

The beta coefficients are provided along with regression statistics. Great!

The scary part of modeling with Python is that anyone can build these models.

Nothing requires you to check that the model is reasonable, much less statistically significant. So, before you blindly believe a model, check it.

Answer questions like:

Is the model statistically significant?

Check the F-statistic.

Are the coefficients significant?

Check the coefficients t-statistics and p-values in the summary or check their confidence intervals.

Is the model useful?

Check R^2

Does the model fit the data well?

Plot the residuals and check the regression diagnostics.

Does the data satisfy assumptions behind linear regression?

Check whether the diagnostics confirm that a linear model is reasonable for your data.

Regression creates a model.

ANOVA is a method of evaluating such models.

Regression and ANOVA are closely linked.

When you add or delete a predictor variable from a linear regression, you want to know whether that change did nor did not improve the model.

ANOVA compares two regression models and reports whether they are significantly different.

What if we coded the study skills data as seen here:

The 1's and 0's essentially indicate group membership, and treat the freshmen as a 'control'.

We can perform MLR on this dataset to see how the classes compare.

$$y_{ij} = \mu_0 + \beta_1 x_{1j} + \beta_2 x_{2j}$$

For instance: $y_{2,3} = 85 = \mu_0 + \beta_1 \cdot 1 + \beta_2 \cdot 0$

μ_0	The sample mean of the Freshmen
$\mu_1 = \mu_0 + \beta_1$	The sample mean of the Sophomores
$\mu_2 = \mu_0 + \beta_2$	The sample mean of the Juniors

82	0	0
93	0	0
61	0	0
74	0	0
69	0	0
70	0	0
53	0	0
71	1	0
62	1	0
85	1	0
94	1	0
78	1	0
66	1	0
71	1	0
64	0	1
73	0	1
87	0	1
91	0	1
56	0	1
78	0	1
87	0	1

Look for this data structure in NoteBook #24.

we are changing from one chart to another.

Freshmen	Sophomores	Juniors
82	71	64
93	62	73
61	85	87
74	94	91
69	78	56
70	66	78
53	71	87

82	0	0
93	0	0
61	0	0
74	0	0
69	0	0
70	0	0
53	0	0
71	1	0
62	1	0
85	1	0
94	1	0
78	1	0
66	1	0
71	1	0
64	0	1
73	0	1
87	0	1
91	0	1
56	0	1
78	0	1
87	0	1

$$y_{ij} = \mu_0 + \beta_1 x_{1j} + \beta_2 x_{2j}$$

μ_0 The sample mean of the Freshmen

$\mu_1 = \mu_0 + \beta_1$ The sample mean of the Sophomores

$\mu_2 = \mu_0 + \beta_2$ The sample mean of the Juniors

The study skills data, and the data structure of the dataframe

In [8]:
```
dfD = pd.DataFrame({"Freshmen": [82, 93, 61, 74, 69, 70, 53],
                    "Sophomores": [71, 62, 85, 94, 78, 66, 71],
                    "Juniors": [64, 73, 87, 91, 56, 78, 87] })
dfD.head(20)
```

Out[8]:

	Freshmen	Sophomores	Juniors
0	82	71	64
1	93	62	73
2	61	85	87
3	74	94	91
4	69	78	56
5	70	66	78
6	53	71	87

In [3]:
```
F, pval = stats.f_oneway(dfD["Freshmen"], dfD["Sophomores"], dfD["Juniors"])
print("F = {:.5f}".format(F))
print("pval = {:.5f}".format(pval))
```

```
F = 0.28373
pval = 0.75628
```

H_0: $\mu_0 = \mu_1 = \mu_2$

Fail to reject H_0

MLR instead of ANOVA

```python
y = dfR.loc[:,"Score"]                                   #
X = dfR.loc[:,["Sophomores", "Juniors"]]
X = sm.add_constant(X)                                   #
#print(X)
model = sm.OLS(y, X).fit()
model.summary()
```

Dep. Variable:	Score	**R-squared:**	0.031
Model:	OLS	**Adj. R-squared:**	-0.077
Method:	Least Squares	**F-statistic:**	0.2837
Date:	Wed, 01 Dec 2021	**Prob (F-statistic):**	0.756
Time:	14:00:11	**Log-Likelihood:**	-81.220
No. Observations:	21	**AIC:**	168.4
Df Residuals:	18	**BIC:**	171.6
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	71.7143	4.725	15.179	0.000	61.788	81.640
Sophomores	3.5714	6.682	0.535	0.600	-10.466	17.609
Juniors	4.8571	6.682	0.727	0.477	-9.180	18.895

Omnibus:	0.820	**Durbin-Watson:**	1.815
Prob(Omnibus):	0.664	**Jarque-Bera (JB):**	0.703
Skew:	0.051	**Prob(JB):**	0.703
Kurtosis:	2.109	**Cond. No.**	3.73

	Score	Sophomores	Juniors
0	82	0.0	0.0
1	93	0.0	0.0
2	61	0.0	0.0
3	74	0.0	0.0
4	69	0.0	0.0
5	70	0.0	0.0
6	53	0.0	0.0
7	71	1.0	0.0
8	62	1.0	0.0
9	85	1.0	0.0
10	94	1.0	0.0
11	78	1.0	0.0
12	66	1.0	0.0
13	71	1.0	0.0
14	64	0.0	1.0
15	73	0.0	1.0
16	87	0.0	1.0
17	91	0.0	1.0
18	56	0.0	1.0
19	78	0.0	1.0
20	87	0.0	1.0

Recall the means of each group:

Freshmen: 71.71

Sophomores: 75.29

Juniors: 76.57

H_0: $\mu_0 = \mu_1 = \mu_2$

Fail to reject H_0

Next Time: Logistic Regression

82	0	0
93	0	0
61	0	0
74	0	0
69	0	0
70	0	0
53	0	0
71	1	0
62	1	0
85	1	0
94	1	0
78	1	0
66	1	0
71	1	0
64	0	1
73	0	1
87	0	1
91	0	1

k