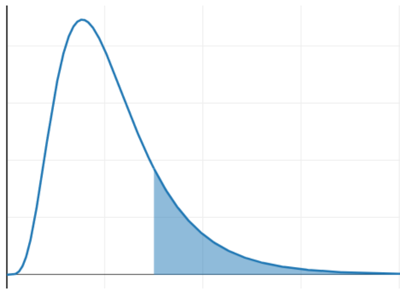
ANOVA

Laxminarayen

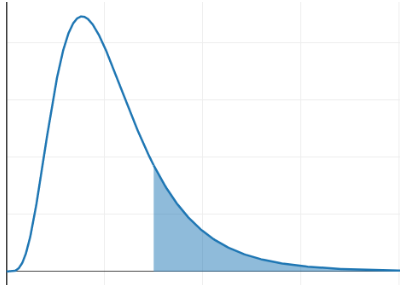
Analysis of Variance

- In this section we introduce a new distribution – the F-Distribution
- Used to answer the question "*What is the probability that two samples come from populations that have the same variance?*"



Analysis of Variance

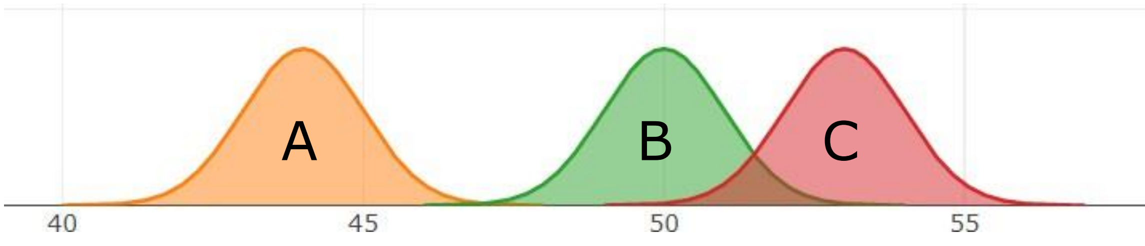
- In this section we introduce a new distribution – the F-Distribution
- Can also answer the question "*What is the probability that* ***three or more*** *samples come from the same population?*"



ANOVA
Analysis of Variance

ANOVA

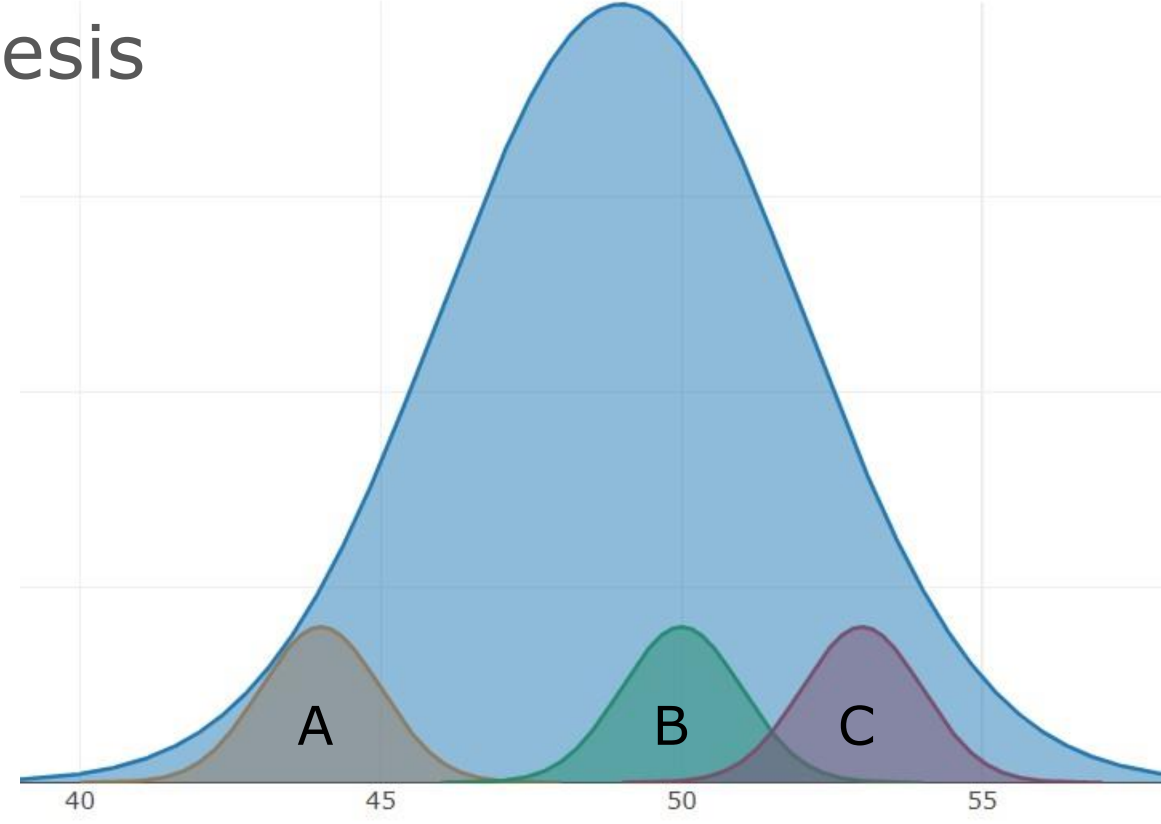
- In the previous section we tested samples to see if they are the representative of the population.
- What if we had three (or more) samples?
- Could we find if they are from same population?



ANOVA

- Our null hypothesis would look like:

$$H_0\text{: } \mu_A = \mu_B = \mu_C$$



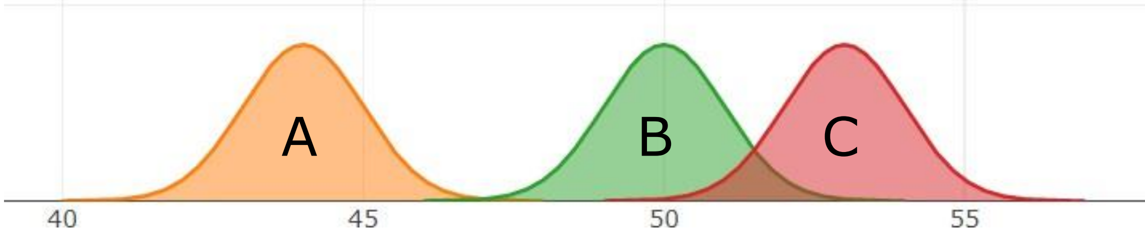
ANOVA

- We *could* test each pair:

$$H_0: \mu_A = \mu_B \qquad \alpha = 0.05$$

$$H_0: \mu_A = \mu_C \qquad \alpha = 0.05$$

$$H_0: \mu_B = \mu_C \qquad \alpha = 0.05$$



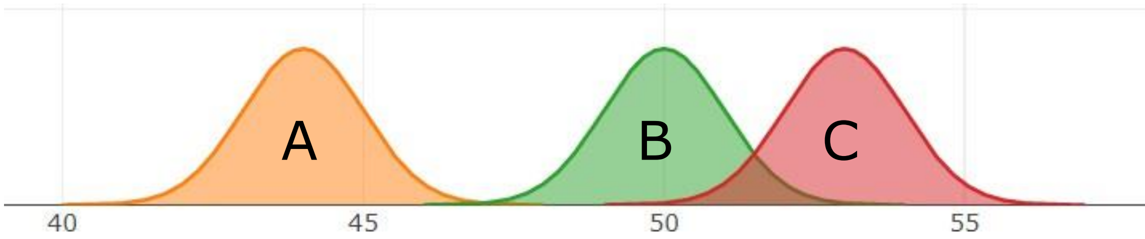
ANOVA

- The problem is, our overall confidence drops:

$$H_0: \mu_A = \mu_B \qquad \alpha = 0.05$$
$$H_0: \mu_A = \mu_C \qquad \alpha = 0.05$$
$$H_0: \mu_B = \mu_C \qquad \alpha = 0.05$$

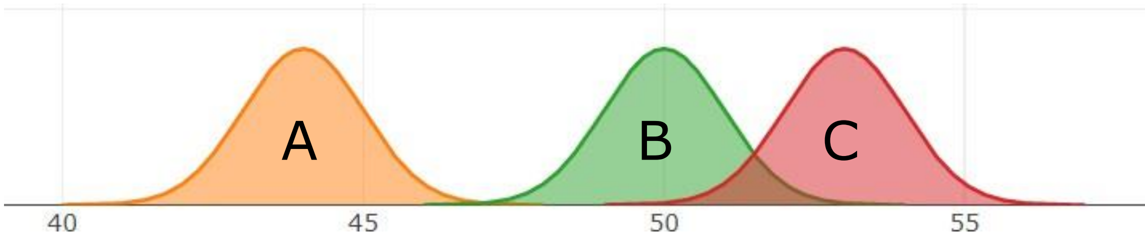
$$.95 \times .95 \times .95 = .857$$

$$85.7\%\ confidence\ level$$



ANOVA

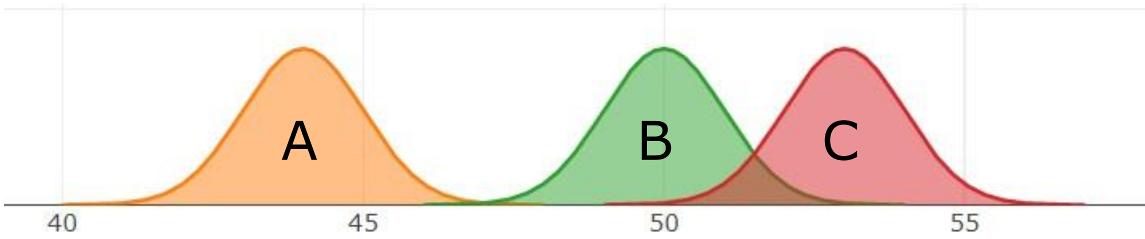
- This is where ANOVA comes in!
- We compute an F value, and compare it to a critical value determined by our degrees of freedom (the number of groups, and the number of items in each group)



ANOVA

Let’s work with some data:

GroupA	GroupB	GroupC
37	62	50
60	27	63
52	69	58
43	64	54
40	43	49
52	54	52
55	44	53
39	31	43
39	49	65
23	57	43



ANOVA

First calculate the sample means

Next calculate the overall mean

	GroupA	GroupB	GroupC
	37	62	50
	60	27	63
	52	69	58
	43	64	54
	40	43	49
	52	54	52
	55	44	53
	39	31	43
	39	49	65
	23	57	43
$\mu_{A,B,C}$	**44**	**50**	**53**
μ_{TOT}	**49**		

ANOVA

ANOVA considers two types of variance:

- Between Groups
  - how far group means stray from the total mean
- Within Groups
  - how far individual values stray from their respective group mean

ANOVA

The F value we're trying to calculate is simply the ratio between these two variances!

$$F = \frac{\text{Variance Between Groups}}{\text{Variance Within Groups}}$$

ANOVA

Recall the equation for variance:

$$s^2 = \frac{\Sigma(x-\bar{x})^2}{n-1} = \frac{SS}{df}$$

Here $\Sigma(x-\bar{x})^2$ is the “sum of squares” SS and $n-1$ is the “degrees of freedom” df

ANOVA

So the formula for the F value becomes:

$$F = \frac{Variance\ Between\ Groups}{Variance\ Within\ Groups} = \frac{\dfrac{SSG}{df_{groups}}}{\dfrac{SSE}{df_{error}}}$$

SSG = Sum of Squares Groups

SSE = Sum of Squares Error

df_{groups} = degrees of freedom (groups)

df_{error} = degrees of freedom (error)

$$SSG = 420$$

ANOVA

Sum of Squares Groups

$$(\mu_A - \mu_{TOT})^2 = (44 - 49)^2 = 25$$

$$(\mu_B - \mu_{TOT})^2 = (50 - 49)^2 = 1$$

$$(\mu_C - \mu_{TOT})^2 = (53 - 49)^2 = 16$$

$$42$$

Multiply by the number of items in each group:

$$42 \times 10 = 420$$

	GroupA	GroupB	GroupC
	37	62	50
	60	27	63
	52	69	58
	43	64	54
	40	43	49
	52	54	52
	55	44	53
	39	31	43
	39	49	65
	23	57	43
$\mu_{A,B,C}$	44	50	53
μ_{TOT}	49		

ANOVA

$$SSG = 420$$
$$df_{groups} = 2$$

GroupA	GroupB	GroupC
37	62	50
60	27	63
52	69	58
43	64	54
40	43	49
52	54	52
55	44	53
39	31	43
39	49	65
23	57	43
44	**50**	**53**

$\mu_{A,B,C}$	**44**, **50**, **53**
μ_{TOT}	**49**

Degrees of Freedom Groups

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

ANOVA

$$SSG = 420$$
$$df_{groups} = 2$$
$$SSE = 3300$$

Sum of Squares Error

$(x_A-\mu_A)^2$	$(x_A-\mu_A)^2$	$(x_B-\mu_B)^2$	$(x_B-\mu_B)^2$	$(x_C-\mu_C)^2$	$(x_C-\mu_C)^2$
49	64	144	16	9	1
256	121	529	36	100	0
64	25	361	361	25	100
1	25	196	1	1	144
16	441	49	49	16	100
	1062		**1742**		**496**

TOTAL	**3300**

$(37-44)^2$
$=(-7)^2$
$=49$

	GroupA	GroupB	GroupC
	37	62	50
	60	27	63
	52	69	58
	43	64	54
	40	43	49
	52	54	52
	55	44	53
	39	31	43
	39	49	65
	23	57	43
$\mu_{A,B,C}$	**44**	**50**	**53**
μ_{TOT}	**49**		

ANOVA

$$SSG = 420$$
$$df_{groups} = 2$$
$$SSE = 3300$$
$$df_{error} = 27$$

GroupA	GroupB	GroupC
37	62	50
60	27	63
52	69	58
43	64	54
40	43	49
52	54	52
55	44	53
39	31	43
39	49	65
23	57	43

$\mu_{A,B,C}$	**44**	**50**	**53**
μ_{TOT}	**49**		

Degrees of Freedom Error

$$df_{error} = (n_{rows} - 1) * n_{groups}$$
$$= (10 - 1) * 3$$
$$= 27$$

ANOVA

$$SSG = 420$$
$$df_{groups} = 2$$
$$SSE = 3300$$
$$df_{error} = 27$$

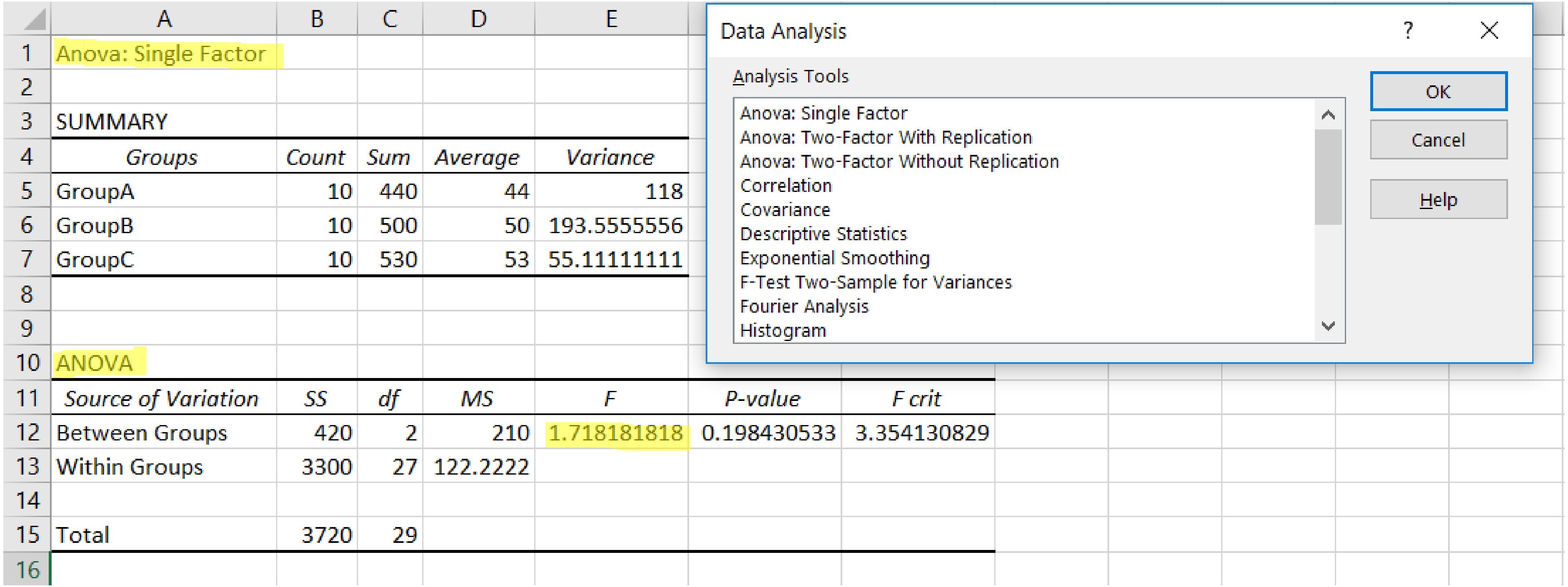
GroupA	GroupB	GroupC
37	62	50
60	27	63
52	69	58
43	64	54
40	43	49
52	54	52
55	44	53
39	31	43
39	49	65
23	57	43

$\mu_{A,B,C}$	**44**	**50**	**53**
μ_{TOT}	**49**		

Plug these into our formula:

$$F = \frac{\frac{SSG}{df_{groups}}}{\frac{SSE}{df_{error}}} = \frac{\frac{420}{2}}{\frac{3300}{27}} = \frac{210}{122.2} \quad 2 = \mathbf{1.718}$$

ANOVA with Excel Data Analysis



Anova: Single Factor

SUMMARY

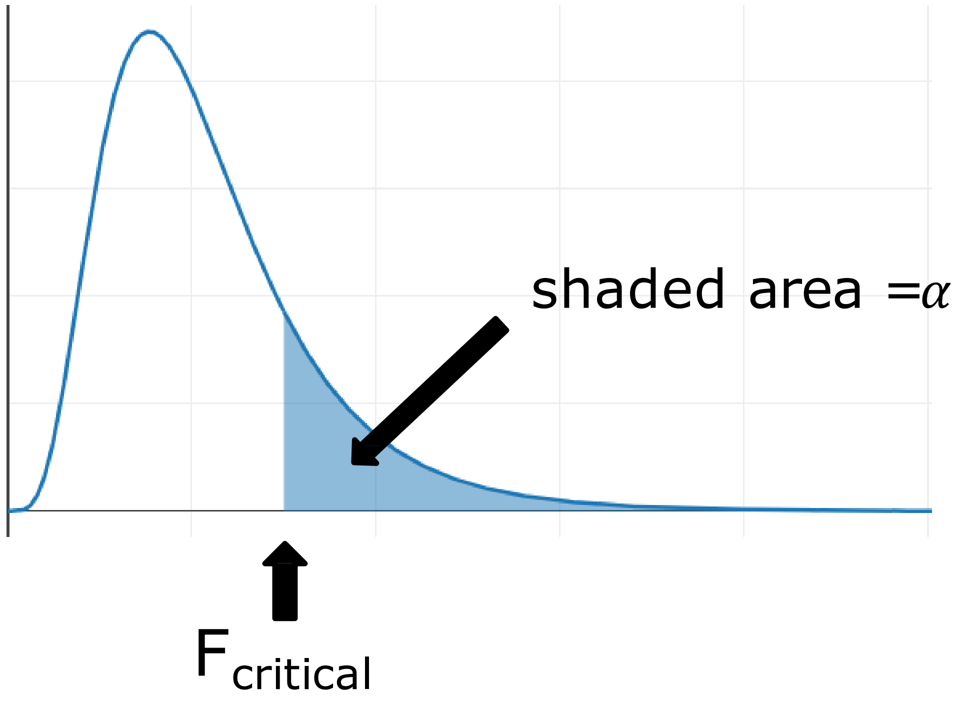
Groups	Count	Sum	Average	Variance
GroupA	10	440	44	118
GroupB	10	500	50	193.5555556
GroupC	10	530	53	55.11111111

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	420	2	210	1.718181818	0.198430533	3.354130829
Within Groups	3300	27	122.2222			
Total	3720	29				

F Distribution

F-Distribution



F-Distribution

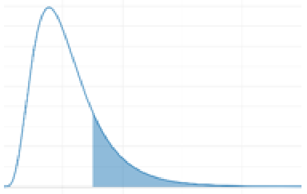
Look up our critical value from an F-table

use a table set for 95% confidence

find numerator df

find denominator df

critical value =3.35



	F-Table Upper Tail Area of 0.05				
	Numerator df				
denominator df	1	2	3	4	5
25	4.24	3.39	2.99	2.76	2.60
26	4.23	3.37	2.98	2.74	2.59
27	4.21	3.35	2.96	2.73	2.57
28	4.20	3.34	2.95	2.71	2.56
29	4.18	3.33	2.93	2.70	2.55
30	4.17	3.32	2.92	2.69	2.53

F-Scores in MS Excel

- In Microsoft Excel, the following function returns an F-score:

α	df1	df2	Formula	Output Value
0.05	2	27	=FINV(A2,B2,C2)	3.3541308285292

F-Scores in Python

```
>>> from scipy import stats
>>> stats.f.ppf(1-.05,dfn=2,dfd=27)
3.3541308285291986
```

ANOVA

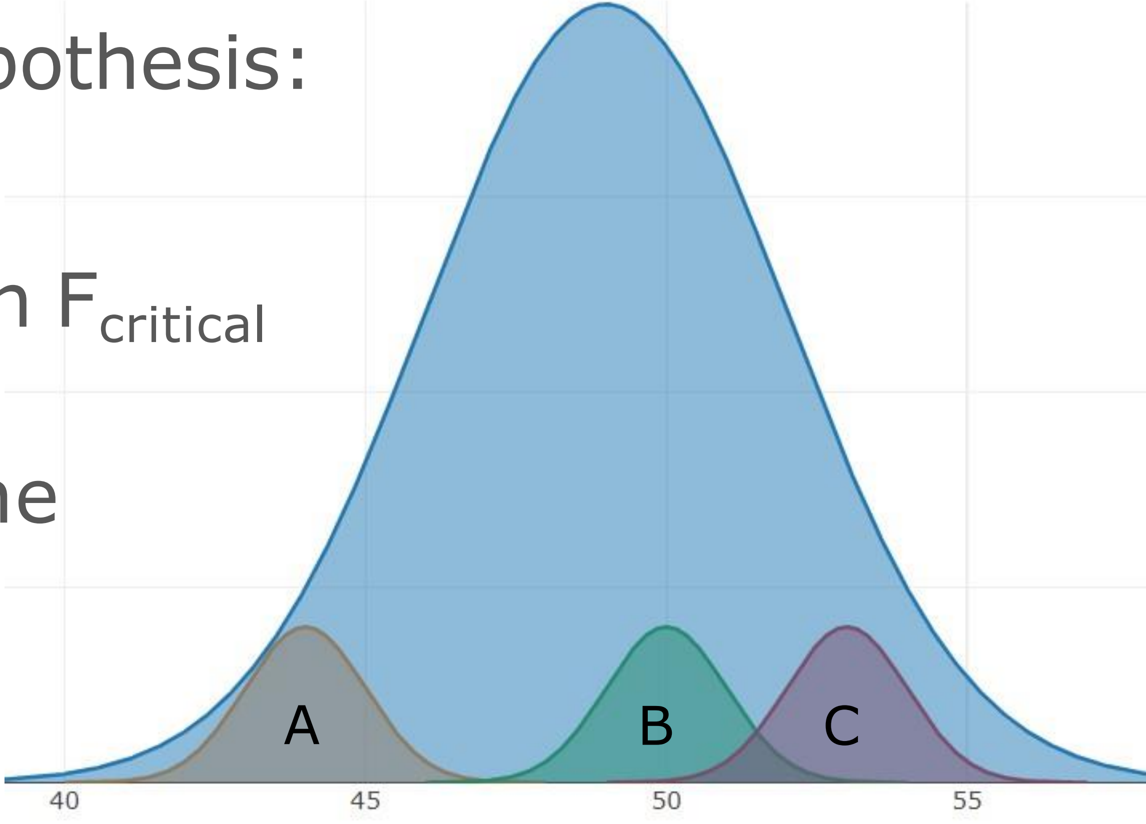
Recall our null hypothesis:

$$H_0: \mu_A = \mu_B = \mu_C$$

Since F is less than $F_{critical}$

$$1.718 < 3.354$$

we fail to reject the null hypothesis!



ANOVA Exercise #1

- In an effort to receive faster payment of invoices, a company introduces two discount plans
- One set of customers is given a 2% discount if they pay their invoice early
- Another set is offered a 1% discount
- A third set is not offered any incentive

ANOVA Exercise #1

- The results are as follows:
- Using ANOVA, can we say that the offers result in faster payments?

2% disc	1% disc	no disc
11	21	14
16	15	11
9	23	18
14	10	16
10	16	21

ANOVA Exercise #1

1. Calculate the means

	2% disc	1% disc	no disc
	11	21	14
	16	15	11
	9	23	18
	14	10	16
	10	16	21
$\mu_{2,1,0}$	**12**	**17**	**16**
μ_{TOT}	**15**		

$SSG = 70$

ANOVA Exercise #1

2. Find Sum of Squares Groups

$(\mu_2 - \mu_{TOT})^2 = (12 - 15)^2 = 9$

$(\mu_1 - \mu_{TOT})^2 = (17 - 15)^2 = 4$

$(\mu_0 - \mu_{TOT})^2 = (16 - 15)^2 = 1$

14

Multiply by the number of items in each group:

$14 \times 5 = 70$

	2% disc	1% disc	no disc
	11	21	14
	16	15	11
	9	23	18
	14	10	16
	10	16	21
$\mu_{2,1,0}$	**12**	**17**	**16**
μ_{TOT}	**15**		

ANOVA Exercise #1

$$SSG = 70$$
$$df_{groups} = 2$$

3. Degrees of Freedom Groups

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

	2% disc	1% disc	no disc
	11	21	14
	16	15	11
	9	23	18
	14	10	16
	10	16	21
$\mu_{2,1,0}$	**12**	**17**	**16**
μ_{TOT}	**15**		

ANOVA Exercise #1

$SSG = 70$
$df_{groups} = 2$
$SSE = 198$

4. Sum of Squares Error

$(x_2-\mu_2)^2$	$(x_1-\mu_1)^2$	$(x_0-\mu_0)^2$
1	16	4
16	4	25
9	36	4
4	49	0
4	1	25
34	**106**	**58**

	TOTAL	**198**

	2% disc	1% disc	no disc
	11	21	14
	16	15	11
	9	23	18
	14	10	16
	10	16	21
$\mu_{2,1,0}$	**12**	**17**	**16**
μ_{TOT}	**15**		

ANOVA Exercise #1

$SSG = 70$
$df_{groups} = 2$
$SSE = 198$
$df_{error} = 12$

5. Degrees of Freedom Error

$$df_{error} = (n_{rows} - 1) * n_{groups}$$
$$= (5 - 1) * 3$$
$$= 12$$

	2% disc	1% disc	no disc
	11	21	14
	16	15	11
	9	23	18
	14	10	16
	10	16	21
$\mu_{2,1,0}$	**12**	**17**	**16**
μ_{TOT}	**15**		

ANOVA Exercise #1

$SSG = 70$
$df_{groups} = 2$
$SSE = 198$
$df_{error} = 12$

	2% disc	1% disc	no disc
	11	21	14
	16	15	11
	9	23	18
	14	10	16
	10	16	21
$\mu_{2,1,0}$	**12**	**17**	**16**
μ_{TOT}	**15**		

6. Calculate F value:

$$F = \frac{\frac{SSG}{df\,groups}}{\frac{SSE}{df\,error}} = \frac{\frac{70}{2}}{\frac{198}{12}} = \frac{35}{16.5} = \mathbf{2.121}$$

7. Look up $F_{critical}$: **3.885**

ANOVA Exercise #1

$SSG = 70$

$df_{groups} = 2$

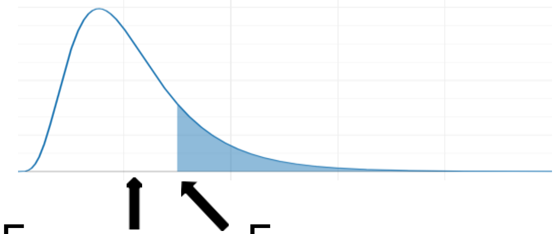
$SSE = 198$

$df_{error} = 12$

Since F falls to the left of $F_{critical}$

$2.121 < 3.885$

we fail to reject the null hypothesis!



$F_{calculated} = 2.121$

$F_{critical} = 3.885$

ANOVA Exercise #1

We don’t have enough to support the idea that our offers changed the average number of days that customers took to pay their invoices!