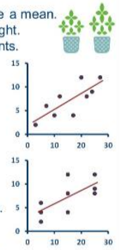
Two-factor ANOVA: introduction

Scenario: We want to know if values or levels of a factor influence a mean.
Examples: Do plants grow taller with different amounts of sunlight.
Do plants grow taller with different fertilizer treatments.

We *could* use correlation/regression techniques.

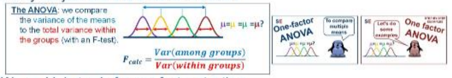
But there are many situations when data like this isn't possible or enough.



Very often we have data which comes from experiments with a limited set of different treatments, or observed sets of data, which can only be grouped as categories.

(1) Violates assumptions of regression/correlation.
(2) Not useful for qualitative categories (e.g., brands A/B/C).

Question: Why not just use a one-factor ANOVA?



Answer: We could, but only for one factor at a time.

Questions: (1) How might *multiple* factors influence, or be associated with, the means of the groups?
(2) How might these factors *interact* with each other (i.e., what if they're not independent)?

Answers: (1) We will, by considering each factor, independently, one at a time.
(2) We will estimate how much of the pattern isn't explainable by the factors independently and is therefore due an interaction.

One-factor ANOVA vs two-factor ANOVA

Similarities

The two-factor ANOVA uses many of the *same steps* as the one-factor ANOVA:
- sums of squares, df, mean sums, F tests, making a figure to interpret results.

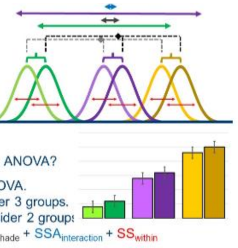
Differences

One-factor ANOVA: the *goal* is to determine if any *group means differ*.
- Solves the type I risk problem from multiple comparisons.

Two-factor ANOVA: the *goal* is to determine if there are any non-random *associations/relationships* between factor values and the data.
- Solves the type I risk problem from multiple comparisons.
- Allows discovery of non-independence (i.e., interactions) between factors.

Imagine a scenario in which we have 6 groups:
3 different colors (green/purple/brown) and
2 different shades (dark/light).

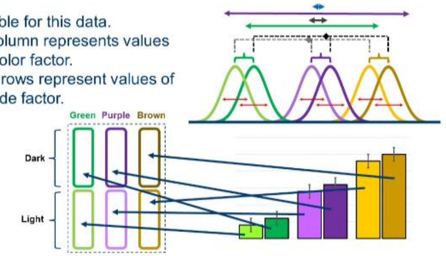
Visually we can see that:
green < purple < brown
light < dark



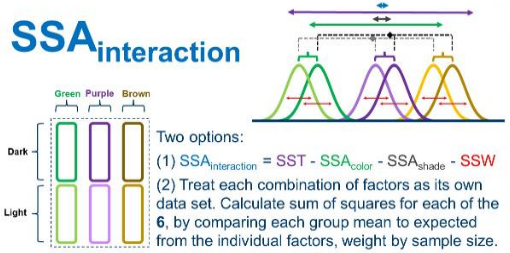
How to calculate the sums of squares for this ANOVA?

SS_{total} and SS_{within} same as one-factor ANOVA.
SSA_{color}: group colors together and consider 3 groups.
SSA_{shade}: group shades together and consider 2 groups.
$SSA_{interaction}$: from $SS_{total} = SSA_{color} + SSA_{shade} + SSA_{interaction} + SS_{within}$

Data table for this data.
Each column represents values of the color factor.
Sets of rows represent values of the shade factor.

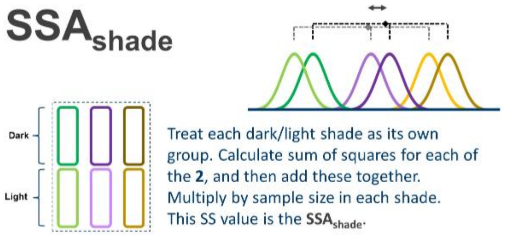


$SSA_{interaction}$



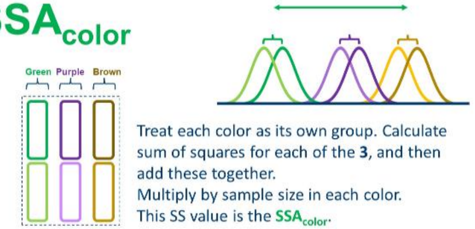
Two options:
(1) $SSA_{interaction} = SST - SSA_{color} - SSA_{shade} - SSW$
(2) Treat each combination of factors as its own data set. Calculate sum of squares for each of the **6**, by comparing each group mean to expected from the individual factors, weight by sample size.

SSA_{shade}



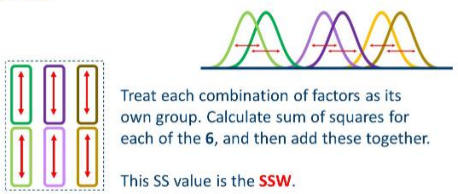
Treat each dark/light shade as its own group. Calculate sum of squares for each of the **2**, and then add these together.
Multiply by sample size in each shade.
This SS value is the SSA_{shade}.

SSA_{color}



Treat each color as its own group. Calculate sum of squares for each of the **3**, and then add these together.
Multiply by sample size in each color.
This SS value is the SSA_{color}.

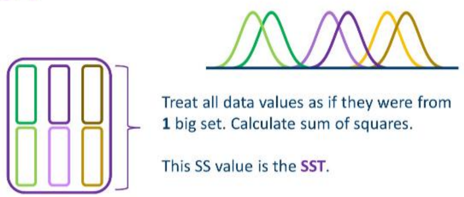
SSW



Treat each combination of factors as its own group. Calculate sum of squares for each of the **6**, and then add these together.

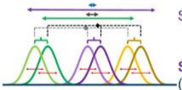
This SS value is the **SSW**.

SST



Treat all data values as if they were from **1** big set. Calculate sum of squares.

This SS value is the **SST**.



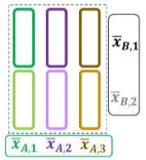
$SST = SSA_A + SSA_B + SSA_{int} + SSW$
SSA

SST: calculate sum of squares for all values (comparing to overall mean).

SSW: calculate the sum of squares values separately for each group using the group means, then sum them all.

SSA_A, SSA_B: calculate sum of squares values for each factor value (comparing to overall mean), then weight by factor sample size.

SSA_{int}: Via $SST = SSA_A + SSA_B + SSA_{int} + SSW$



Testing for significance

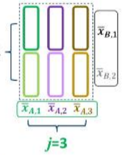
Just like with one factor ANOVA, we compare the mean squares for the factors and interaction to mean squares within with an F test.

$$MSW = \frac{SSW}{N - jk}$$

Factor A: $F_{calc,A} = \frac{MSA_A}{MSW}$ $\quad MSA_A = \frac{SSA_A}{j-1}$

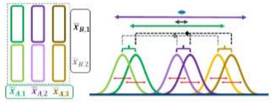
Factor B: $F_{calc,B} = \frac{MSA_B}{MSW}$ $\quad MSA_B = \frac{SSA_B}{k-1}$

Inter: $F_{calc,int} = \frac{MSA_{int}}{MSW}$ $\quad MSA_{int} = \frac{SSA_{int}}{(j-1)(k-1)}$



The ANOVA Table

Results are usually presented in an ANOVA Table



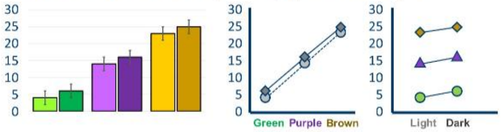
Source	df	SS	MS	F	p
Factor A	$j-1$	SSA_A	$MSA_A = \frac{SSA_A}{j-1}$	$F = \frac{MSA_A}{MSW}$	?
Factor B	$k-1$	SSA_B	$MSA_B = \frac{SSA_B}{k-1}$	$F = \frac{MSA_B}{MSW}$	?
Interaction	$(j-1)(k-1)$	SSA_{int}	$MSA_{int} = \frac{SSA_{int}}{(j-1)(k-1)}$	$F = \frac{MSA_{int}}{MSW}$	?
Within	$N-jk$	SSW	$MSW = \frac{SSW}{N-jk}$		
Total	$N-1$				

Indicate any significant associations between factors and the means, or a significant interaction.

INTERACTION PLOTS

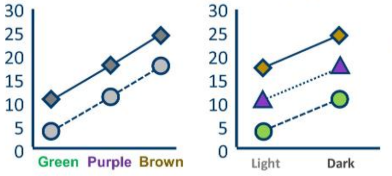
The ANOVA analysis provides information about whether factors (and interaction) are significant, but not what those effects are.

To visualize this, we make a pair of interaction plots using the mean values of the groups. Each plot uses the values of one of the factors on the X-axis and plots the group means on the Y-axis. The second factor is represented by symbols within the figure.



INTERACTION PLOTS

The plots suggest associations and interactions.
The ANOVA table reveals whether they are statistically significant.

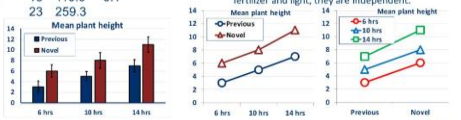


Naive guesses
Color: sig
Shade: sig
Inter: NS

Example ANOVA Table: Imagine a company develops a novel type of fertilizer. They raise plants under 3 light levels with 2 types of fertilizer, the new fertilizer and a previous formula. Heights of 4 plants with each combination after 30 days.

Source	df	SS	MS	F	p
Fertilizer	1	66.7	66.7	10.9	0.004
Light Lvl	2	81.3	40.7	6.6	0.007
Interact.	2	1.3	0.7	0.1	0.897
Within	18	110.0	6.1		
Total	23	259.3			

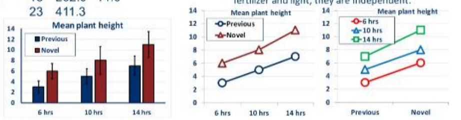
There **is a significant association** between height and the fertilizer treatment.
There is **a significant association** between height and the amount of light.
There is **no significant interaction** between fertilizer and light, they are independent.



Example ANOVA Table: Imagine a company develops a novel type of fertilizer. They raise plants under 3 light levels with 2 types of fertilizer, the new fertilizer and a previous formula. Heights of 4 plants with each combination after 30 days.

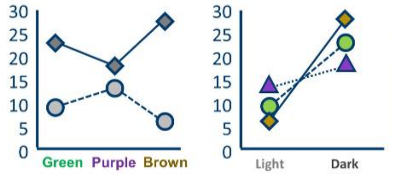
Source	df	SS	MS	F	p
Fertilizer	1	66.7	66.7	4.58	0.046
Light Lvl	2	81.3	40.7	2.79	0.088
Interact.	2	1.3	0.7	0.05	0.955
Within	18	262.0	14.6		
Total	23	411.3			

There **is a significant association** between height and the fertilizer treatment.
There is **no significant association** between height and the amount of light.
There is **no significant interaction** between fertilizer and light, they are independent.



INTERACTION PLOTS

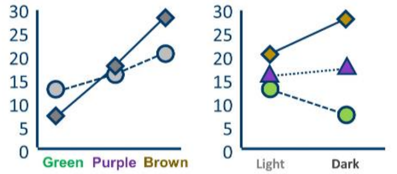
The plots suggest associations and interactions.
The ANOVA table reveals whether they are statistically significant.



Naive guesses
Color: NS
Shade: sig
Inter: sig

INTERACTION PLOTS

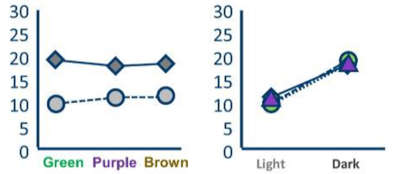
The plots suggest associations and interactions.
The ANOVA table reveals whether they are statistically significant.



Naive guesses
Color: sig
Shade: ?
Inter: sig

INTERACTION PLOTS

The plots suggest associations and interactions.
The ANOVA table reveals whether they are statistically significant.

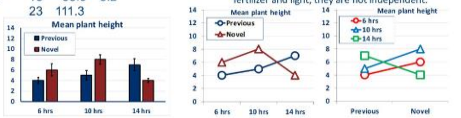


Naive guesses
Color: NS
Shade: sig
Inter: ?

Example ANOVA Table: Imagine a company develops a novel type of fertilizer. They raise plants under 3 light levels with 2 types of fertilizer, the new fertilizer and a previous formula. Heights of 4 plants with each combination after 30 days.

Source	df	SS	MS	F	p
Fertilizer	1	2.7	2.7	0.83	0.375
Light Lvl	2	9.3	4.7	1.45	0.261
Interact.	2	41.3	20.7	6.41	0.006
Within	18	58.0	3.2		
Total	23	111.3			

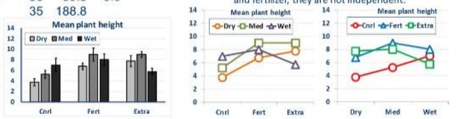
There **is no significant association** between height and the fertilizer treatment.
There is **no significant association** between height and the amount of light.
There is **a significant interaction** between fertilizer and light, they are not independent.



Example ANOVA Table: Imagine a study on the effects of different amounts of fertilizer. Plants are raised under 3 water levels with no fertilizer, some fertilizer and extra fertilizer. Heights of 4 plants with each combination are measured after 30 days.

Source	df	SS	MS	F	p
Water	2	16.7	8.3	2.79	0.078
Fertilizer	2	46.6	23.1	7.72	0.002
Interact.	4	36.2	9.0	3.02	0.031
Within	30	89.8	3.0		
Total	35	188.8			

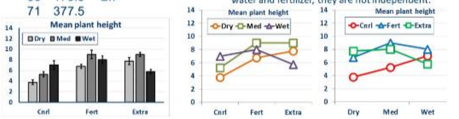
There **is no significant association** between height and the water treatment.
There is **a significant association** between height and the amount of fertilizer.
There is **a significant interaction** between water and fertilizer, they are not independent.



Example ANOVA Table: Imagine a study on the effects of different amounts of fertilizer. Plants are raised under 3 water levels with no fertilizer, some fertilizer and extra fertilizer. Heights of 8 plants with each combination are measured after 30 days.

Source	df	SS	MS	F	p
Water	2	33.3	16.7	6.13	0.004
Fertilizer	2	92.3	46.2	16.97	1.1×10^{-6}
Interact.	4	72.3	18.1	6.65	0.0001
Within	66	179.5	2.7		
Total	71	377.5			

There **is no significant association** between height and the water treatment.
There is a **highly significant association** between height and the amount of fertilizer.
There is a **highly significant interaction** between water and fertilizer, they are not independent.



IMPORTANT CAUTIONS

1. ANOVA techniques are homoscedastic, the variances in each group must be equal. Do an F_{max} test before beginning (just like one-factor ANOVA).

2. A significant association does not prove direct causation.

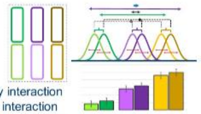
"correlation doesn't imply causation"

2a. This ***DOES NOT MEAN*** that a significant association is useless. (this is a common amateur/beginner mistake)

2b. If other factors *have not* been controlled (e.g., a field or epidemiology study), the ANOVA can demonstrate a connection of some kind. It may be one factor directly causing the other, indirect causation, or a third factor influencing both.

2c. If all other factors *have* been controlled (i.e., an experiment), the ANOVA can demonstrate causation, but this may be indirect.

OVERALL SUMMARY



Is there an association between factors and means?
Approach: **AN**alysis **O**f **VA**riance = **ANOVA**
Prerequisite = equality of population variances

Conceptual hypotheses: H_0: no associations or any interaction
H_A: 1+ associations or an interaction

Practical hypotheses: H_0: $MSA_A \leq MSW$, H_0: $MSA_B \leq MSW$, H_0: $MSA_{int} \leq MSW$
H_A: $MSA_A > MSW$, H_A: $MSA_B > MSW$, H_A: $MSA_{int} > MSW$

1. Perform an F_{max} test to ensure equality of variances.
2. Calculate SST, SSW, SSA_A, SSA_B, SSA_{int}, and the degrees of freedom.
3. Complete the AVOVA table: calculate the MSA_A, MSA_B, MSA_{int}, MSW, $F_A = MSA_A/MSW$, $F_B = MSA_B/MSW$, $F_{int} = MSA_{int}/MSW$, and the p values to "reject" or "fail to reject" each H_0.
4. If any H_0 is rejected: create a pair of interaction plots to interpret.
5. Be careful with interpretations.