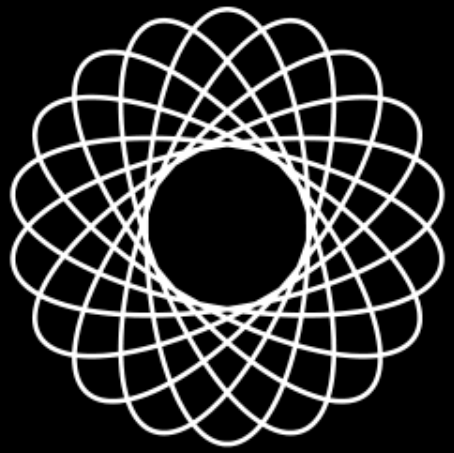
DATA SCIENCE



Anova in Excel

Anova: Single Factor						
SUMMARY						
Groups	*Count*	*Sum*	*Average*	*Variance*		
Shelf 1	10	1799.2	179.92	874.4529		
Shelf 2	10	1850.9	185.09	1401.637		
Shelf 3	10	1811.2	181.12	913.8396		
Shelf 4	10	1828.5	182.85	587.005		
Shelf 5	10	1787.2	178.72	82.51289		
ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	250.718	4	62.6795	0.081203	0.987743	2.578739
Within Groups	34735.02	45	771.8894			
Total	34985.74	49				

Anova

Total Variation = Between Variance + Within Variance

Anova

Total Variation = Between Variance + Within Variance

Another way of looking at total variation is:

1. Calculate Grand Mean of all observations
2. Calculate Difference between each observation and the Grand Mean
3. Square the differences
4. Add them all up

Anova

Total Variation = Between Variance + Within Variance

Another way of looking at total variation is:

1. Calculate Grand Mean of all observations
2. Calculate Difference between each observation and the Grand Mean
3. Square the differences
4. Add them all up

Should we check if total variation calculated this way will add up to the sum of between variance and within variance?

Anova

Total Variation = Between Variance + Within Variance

Another way of looking at total variation is:

1. Calculate Grand Mean of all observations
2. Calculate Difference between each observation and the Grand Mean
3. Square the differences
4. Add them all up

Should we check if total variation calculated this way will add up to the sum of between variance and within variance?

SST (Total Sum of Squares) = 459.775
SSB (Sum of Squares Between) = 13.875
SSW (Sum of Squares Within) = 445.9

Anova Assumptions

Anova Assumptions

1. The populations from which the samples were obtained must be normally or approximately normally distributed

Anova Assumptions

1. The populations from which the samples were obtained must be normally or approximately normally distributed

2. The samples must be independent

Anova Assumptions

1. The populations from which the samples were obtained must be normally or approximately normally distributed

2. The samples must be independent

3. The variances of the populations must be equal

Anova Assumptions

1. The populations from which the samples were obtained must be normally or approximately normally distributed

2. The samples must be independent

3. The variances of the populations must be equal

An ANOVA is used when the DV(outcome) is continuous, and the IVs (factors) are discrete

Agenda

Anova

- One Way
- **Two Way**
- Post Hoc Tests

Chi Square

- Association Tests
- Goodness-of-fit Tests

Chi Square Parametric

- Tests of Variance

Two Way Anova

Example – 2 Factors Influencing Outcome

Two Way Anova

Example – 2 Factors Influencing Outcome

- Let's say we are interested in understanding the impact of both shelf level as well as aisle placement on sales for Brand A

- That is, not only the height of the product placed, but also other brands / categories that the product is placed in are hypothesized to have an impact on Brand A sales

Two Way Anova

Example – 2 Factors Influencing Outcome

- Let's say we are interested in understanding the impact of both shelf level as well as aisle placement on sales for Brand A
- That is, not only the height of the product placed, but also other brands / categories that the product is placed in are hypothesized to have an impact on Brand A sales
- If there are three different aisles, we have 3*5 different placements for Brand A

Two Way Anova

Example – 2 Factors Influencing Outcome

- Let's say we are interested in understanding the impact of both shelf level as well as aisle placement on sales for Brand A
- That is, not only the height of the product placed, but also other brands / categories that the product is placed in are hypothesized to have an impact on Brand A sales
- If there are three different aisles, we have 3*5 different placements for Brand A
- How do we determine if mean sales rates are different between the groups?

Two Way Anova

A Two-Way ANOVA is useful when we desire to compare the effect of multiple levels of two factors and we have multiple observations at each level.

Two Way Anova

A Two-Way ANOVA is useful when we desire to compare the effect of multiple levels of two factors and we have multiple observations at each level.

There are three null hypothesis that can be tested in a two-way ANOVA

Two Way Anova

A Two-Way ANOVA is useful when we desire to compare the effect of multiple levels of two factors and we have multiple observations at each level.

There are three null hypothesis that can be tested in a two-way ANOVA

- The population means of the first factor are equal. This is like the one-way ANOVA for the row factor.

Two Way Anova

A Two-Way ANOVA is useful when we desire to compare the effect of multiple levels of two factors and we have multiple observations at each level.

There are three null hypothesis that can be tested in a two-way ANOVA

- The population means of the first factor are equal. This is like the one-way ANOVA for the row factor.
- The population means of the second factor are equal. This is like the one-way ANOVA for the column factor.

Two Way Anova

A Two-Way ANOVA is useful when we desire to compare the effect of multiple levels of two factors and we have multiple observations at each level.

There are three null hypothesis that can be tested in a two-way ANOVA

- The population means of the first factor are equal. This is like the one-way ANOVA for the row factor.
- The population means of the second factor are equal. This is like the one-way ANOVA for the column factor.
- There is no interaction between the two factors. This is similar to performing a test for independence with contingency tables.

Output Interpretation

Look at interaction p-value first:

Output Interpretation

Look at interaction p-value first:

- If significant, implies impact of factor 1 depends on levels of factor 2

Output Interpretation

Look at interaction p-value first:

- If significant, implies impact of factor 1 depends on levels of factor 2
- May not then matter if individual impacts are significant

Output Interpretation

Look at interaction p-value first:

- If significant, implies impact of factor 1 depends on levels of factor 2
- May not then matter if individual impacts are significant

If interaction p value is NS:

Output Interpretation

Look at interaction p-value first:

- If significant, implies impact of factor 1 depends on levels of factor 2
- May not then matter if individual impacts are significant

If interaction p value is NS:

- Re run ANOVA dropping the interaction term

Two Way Anova

Example:

Is there a difference in energy expended (calories burned) based on stretching before exercise and weights during exercise?

PreStretch	AnkleWeights	Energy
No stretch	No weights	106.9
No stretch	No weights	84
No stretch	No weights	97.5
No stretch	No weights	97.1
No stretch	No weights	99.5
No stretch	Weights	100.2
No stretch	Weights	101
No stretch	Weights	118.5
No stretch	Weights	104.5
No stretch	Weights	111.2
Stretch	No weights	82.8
Stretch	No weights	80.4
Stretch	No weights	95.6
Stretch	No weights	82
Stretch	No weights	83.2
Stretch	Weights	89.1
Stretch	Weights	106.4
Stretch	Weights	98.3
Stretch	Weights	89.2
Stretch	Weights	104.6

Two Way Anova

Example:

Is there a difference in energy expended (calories burned) based on stretching before exercise and weights during exercise?

Factors (IVs) – 2: Pre Stretch and Ankle Weights

PreStretch	AnkleWeights	Energy
No stretch	No weights	106.9
No stretch	No weights	84
No stretch	No weights	97.5
No stretch	No weights	97.1
No stretch	No weights	99.5
No stretch	Weights	100.2
No stretch	Weights	101
No stretch	Weights	118.5
No stretch	Weights	104.5
No stretch	Weights	111.2
Stretch	No weights	82.8
Stretch	No weights	80.4
Stretch	No weights	95.6
Stretch	No weights	82
Stretch	No weights	83.2
Stretch	Weights	89.1
Stretch	Weights	106.4
Stretch	Weights	98.3
Stretch	Weights	89.2
Stretch	Weights	104.6

Two Way Anova

Example:

Is there a difference in energy expended (calories burned) based on stretching before exercise and weights during exercise?

Factors (IVs) – 2: Pre Stretch and Ankle Weights

Two levels in Each Factor:

Pre Stretch: Yes, No

Ankle Weights: Yes, No

PreStretch	AnkleWeights	Energy
No stretch	No weights	106.9
No stretch	No weights	84
No stretch	No weights	97.5
No stretch	No weights	97.1
No stretch	No weights	99.5
No stretch	Weights	100.2
No stretch	Weights	101
No stretch	Weights	118.5
No stretch	Weights	104.5
No stretch	Weights	111.2
Stretch	No weights	82.8
Stretch	No weights	80.4
Stretch	No weights	95.6
Stretch	No weights	82
Stretch	No weights	83.2
Stretch	Weights	89.1
Stretch	Weights	106.4
Stretch	Weights	98.3
Stretch	Weights	89.2
Stretch	Weights	104.6

Two Way Anova

Example:

Is there a difference in energy expended (calories burned) based on stretching before exercise and weights during exercise?

Factors (IVs) – 2: Pre Stretch and Ankle Weights

Two levels in Each Factor:

Pre Stretch: Yes, No

Ankle Weights: Yes, No

2 Way ANOVA: With Replication

PreStretch	AnkleWeights	Energy
No stretch	No weights	106.9
No stretch	No weights	84
No stretch	No weights	97.5
No stretch	No weights	97.1
No stretch	No weights	99.5
No stretch	Weights	100.2
No stretch	Weights	101
No stretch	Weights	118.5
No stretch	Weights	104.5
No stretch	Weights	111.2
Stretch	No weights	82.8
Stretch	No weights	80.4
Stretch	No weights	95.6
Stretch	No weights	82
Stretch	No weights	83.2
Stretch	Weights	89.1
Stretch	Weights	106.4
Stretch	Weights	98.3
Stretch	Weights	89.2
Stretch	Weights	104.6

Two Way Anova

Example:

Is there a difference in energy expended (calories burned) based on stretching before exercise and weights during exercise?

Factors (IVs) – 2: Pre Stretch and Ankle Weights

Two levels in Each Factor:

Pre Stretch: Yes, No

Ankle Weights: Yes, No

2 Way ANOVA: With Replication

Multiple observations for same combination of factors

PreStretch	AnkleWeights	Energy
No stretch	No weights	106.9
No stretch	No weights	84
No stretch	No weights	97.5
No stretch	No weights	97.1
No stretch	No weights	99.5
No stretch	Weights	100.2
No stretch	Weights	101
No stretch	Weights	118.5
No stretch	Weights	104.5
No stretch	Weights	111.2
Stretch	No weights	82.8
Stretch	No weights	80.4
Stretch	No weights	95.6
Stretch	No weights	82
Stretch	No weights	83.2
Stretch	Weights	89.1
Stretch	Weights	106.4
Stretch	Weights	98.3
Stretch	Weights	89.2
Stretch	Weights	104.6

Two Way Anova - Excel

We do not need to get into all the calculation details:

- Run a two-way analysis in Excel:

Two Way Anova - Excel

We do not need to get into all the calculation details:

- Run a two-way analysis in Excel:
  Data Analysis \ 2 Factor ANOVA with replication

Two Way Anova - Excel

We do not need to get into all the calculation details:

- Run a two-way analysis in Excel:
  Data Analysis \ 2 Factor ANOVA with replication

- Data has to be arranged in a specific manner

	No weights	Weights
No stretch	106.9	100.2
	84	101
	97.5	118.5
	97.1	104.5
	99.5	111.2
Stretch	82.8	89.1
	80.4	106.4
	95.6	98.3
	82	89.2
	83.2	104.6

Two Way Anova - Excel

Anova: Two-Factor With Replication			
SUMMARY	No weights	Weights	Total
No stretch			
Count	5	5	10
Sum	485	535.4	1020.4
Average	97	107.08	102.04
Variance	68.38	59.587	85.09822
Stretch			
Count	5	5	10
Sum	424	487.6	911.6
Average	84.8	97.52	91.16
Variance	37.6	67.427	91.62267
Total			
Count	10	10	
Sum	909	1023	
Average	90.9	102.3	
Variance	88.44667	81.83778	

Two Way Anova - Excel

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Sample	591.872	1	591.872	10.16115	0.005724	4.493998
Columns	649.8	1	649.8	11.15565	0.004154	4.493998
Interaction	8.712	1	8.712	0.149566	0.704045	4.493998
Within	931.976	16	58.2485			
Total	2182.36	19				

Agenda

Anova

- One Way
- Two Way
- **Post Hoc Tests**

Chi Square

- Association Tests
- Goodness-of-fit Tests

Chi Square Parametric

- Tests of Variance

Post Hoc Tests

Note that the ANOVA only tells us if at least one group mean is unequal, but not which one

Post Hoc Tests

Note that the ANOVA only tells us if at least one group mean is unequal, but not which one

To determine which means are unequal we use post hoc tests:

Post Hoc Tests

Note that the ANOVA only tells us if at least one group mean is unequal, but not which one

To determine which means are unequal we use post hoc tests:

- LSD Tests
- Tukey Tests
- Scheffe Tests

Post Hoc Tests

Note that the ANOVA only tells us if at least one group mean is unequal, but not which one

To determine which means are unequal we use post hoc tests:

- LSD Tests
- Tukey Tests
- Scheffe Tests

http://pages.uoregon.edu/stevensj/posthoc.pdf

Recap

Anova

- One Way
- Two Way
- Post Hoc Tests

Coming Up

Chi Square Tests

THANK YOU