USC
School of Business

Data Driven Decision Making: Categorical Data Analysis (χ^2 & ANOVA)

GSBA 545, Fall 2021

Professor Dawn Porter

Categorical Data Analysis

- χ^2 Testing for Nominal Categories
  - Goodness-of-Fit Tests
  - Test for Independence

- Analysis of Variance (ANOVA)
  - One-Way ANOVA
  - Two-Way ANOVA

USC School of Business

χ^2 *Test:* Goodness of Fit

Setting: There are k categories with n items, each classified into one and only one category. Is there a significant difference between the k groups?

H_0: probabilities are $p_1, p_2, \ldots p_k$
H_a: at least one probability differs from $p_1, p_2, \ldots p_k$

Test Statistic is:

$$\chi^2 = \sum_{j=1}^{k} \frac{(O_j - E_j)^2}{E_j}$$

where O_j is the observed and E_j is the expected value under H_0.

Reject H_0 if the p-value $< \alpha$.

*Note: χ^2 and the *p-value* are based on $(k - 1)$ degrees of freedom.

χ^2 *Tests:* Payment Methods

The *American Bankers Association* collects data regarding how consumers pay for in-store purchases. In 1999, the usages below were reported. A sample was taken in 2003 to assess the same information.

Q: Have the methods of payment significantly changed in those 5 years?

Purchase Method	**1999 Proportion**
Credit card	0.22
Debit card	0.21
Personal check	0.18
Cash	0.39

H_0: $p_{\text{credit}} = 0.22$, $p_{\text{debit}} = 0.21$, $p_{\text{check}} = 0.18$, $p_{\text{cash}} = 0.39$
H_a: at least one probability differs from H_0

χ^2 *Tests:* Payment Methods

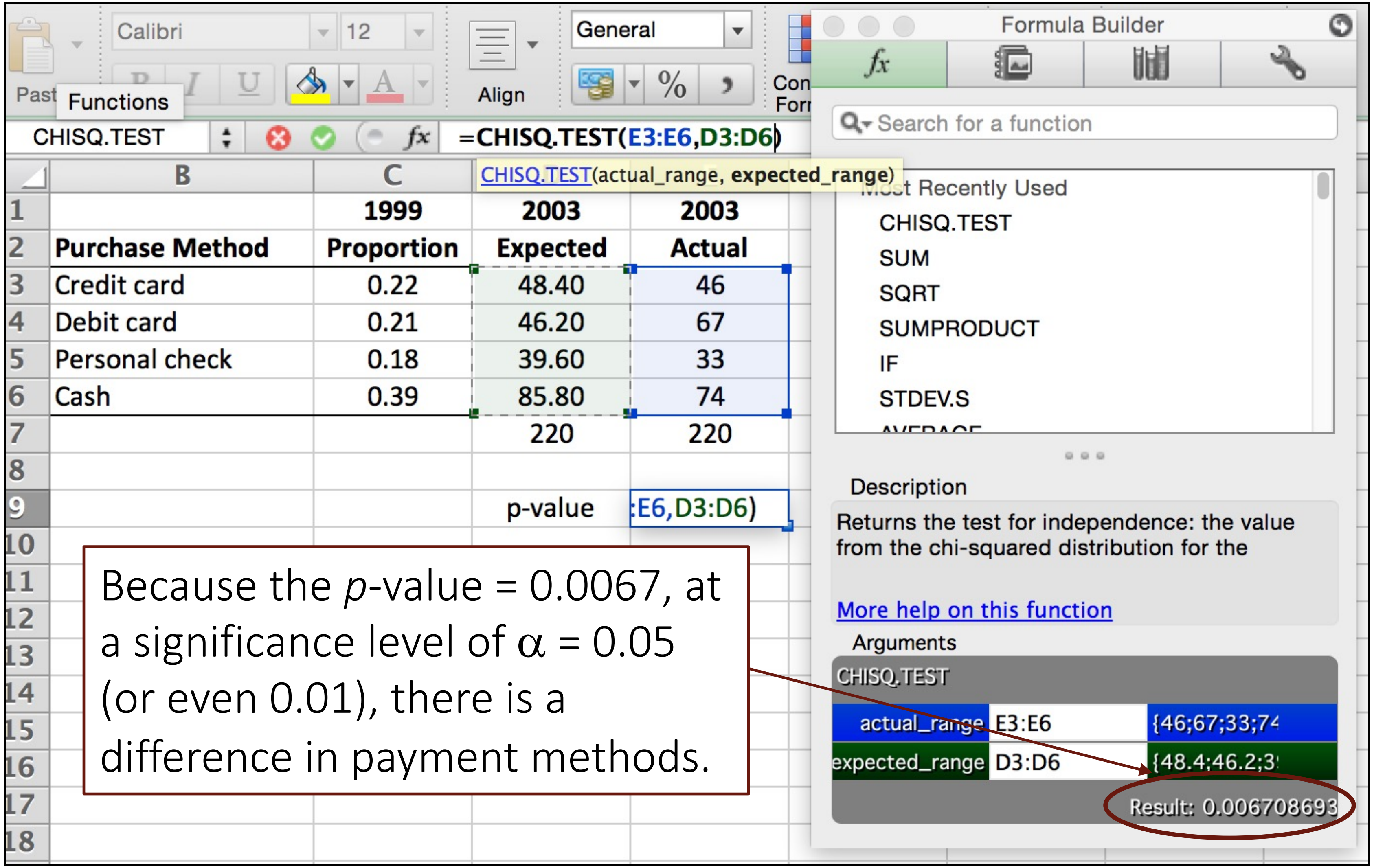
1. If proportions from 1999 still hold in 2003, how many, from a sample of n = 220, would we *expect* to see using each payment method?

2. How many *actually* used each method?

Purchase Method	1999 Proportion	2003 Expected	2003 Actual
Credit card	0.22	48.40	46
Debit card	0.21	46.20	67
Personal check	0.18	39.60	33
Cash	0.39	85.80	74
		220	220

$$E_{credit} = 220\,(0.22) = 48.40$$

χ^2 *Tests:* Payments (Excel)



Because the p-value = 0.0067, at a significance level of α = 0.05 (or even 0.01), there is a difference in payment methods.

χ^2 *Tests:* Health Promotion Campaign

A University survey of recent graduates revealed that a substantial proportion of students were not engaging in regular exercise. In fact, 60% of all graduates reported getting no regular exercise, 25% reported exercising sporadically and 15% reported exercising regularly as undergraduates. The next year the University launched a health promotion campaign on campus to try to increase healthy behaviors. To evaluate the impact of the program, the University again surveyed 470 graduates about their exercise levels.

	No exercise	Sporadic exercise	Regular exercise	Total
Observed	255	125	90	470

Is there evidence of a shift in exercise behavior at $\alpha = 0.05$ following the campaign?

a) What are the null & alternative hypotheses here?
b) What values would you *expect* to see if the null hypothesis still held true?
c) What is the p-value here? What is your conclusion?

χ^2 *Tests:* Health Promotion Campaign

A University survey of recent graduates revealed that a substantial proportion of students were not engaging in regular exercise. In fact, 60% of all graduates reported getting no regular exercise, 25% reported exercising sporadically and 15% reported exercising regularly as undergraduates. The next year the University launched a health promotion campaign on campus to try to increase healthy behaviors. To evaluate the impact of the program, the University again surveyed 470 graduates about their exercise levels.

	No exercise	Sporadic exercise	Regular exercise	Total
Observed	255	125	90	470

Is there evidence of a shift in exercise behavior at $\alpha = 0.05$ following the campaign?

a) What are the null & alternative hypotheses? Soln: H_0: $p_1 = 0.60$, $p_2 = 0.25$, and $p_3 = 0.15$ vs H_a: H_0 doesn't hold.

b) What values would you *expect* to see if the null hypothesis still held true? Soln:

	No exercise	Sporadic exercise	Regular exercise	Total
Observed	255	125	90	470
Expected	282	117.5	70.5	470

c) What is the p-value and your conclusion? Soln: Using Excel, p-value = 0.0146 < 0.05. The campaign had an effect.

χ^2 *Test:* Independence

Setting: There are n items being classified in an R by C contingency table.

Q: Is there a relationship between the rows (R) and the columns (C)?

H_0: the two classifications are statistically independent
H_a: the two classifications are *not* statistically independent

Test Statistic is:

$$\chi^2 = \sum_{\text{all cells}} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where O_j is the observed and E_j is the expected value under H_0.

Reject H_0 if the p-value < α.

*Note: χ^2 and the *p-value* are based on $(r-1)\ (c-1)$ degrees of freedom.

χ^2 *Testing:* Friday the 13th

At α = 0.05, test H_0: Gender is independent of how one is killed

H_a: Gender is related to how one is killed

Actual Table

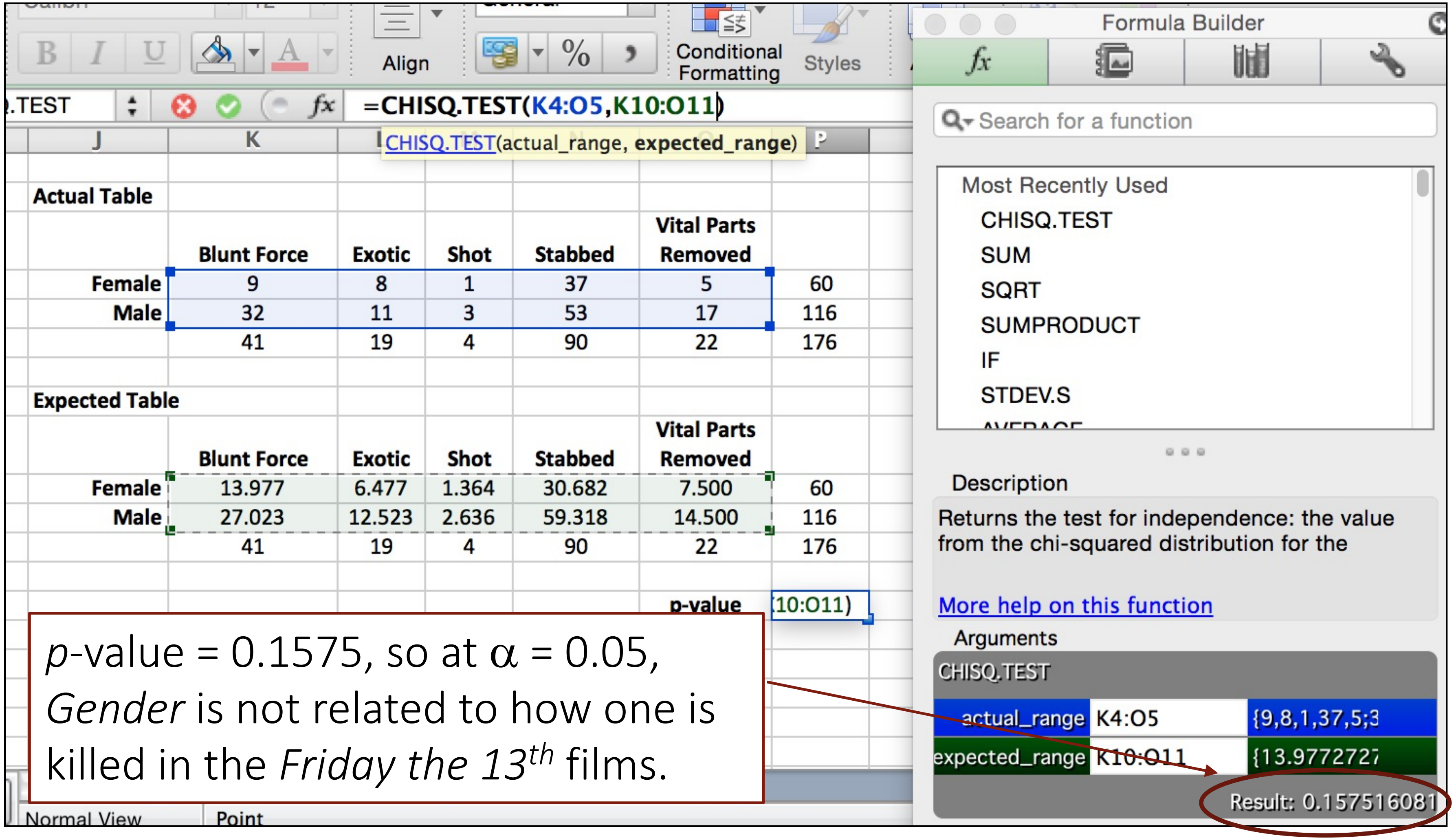
	Blunt Force	Exotic	Shot	Stabbed	Vital Parts Removed	
Female	9	8	1	37	5	60
Male	32	11	3	53	17	116
	41	19	4	90	22	176

$$E_{female,blunt\ force} = \frac{(c_{BluntForce})(r_{Female})}{N} = \frac{(41)(60)}{176} = \boxed{13.977}$$

Expected Table

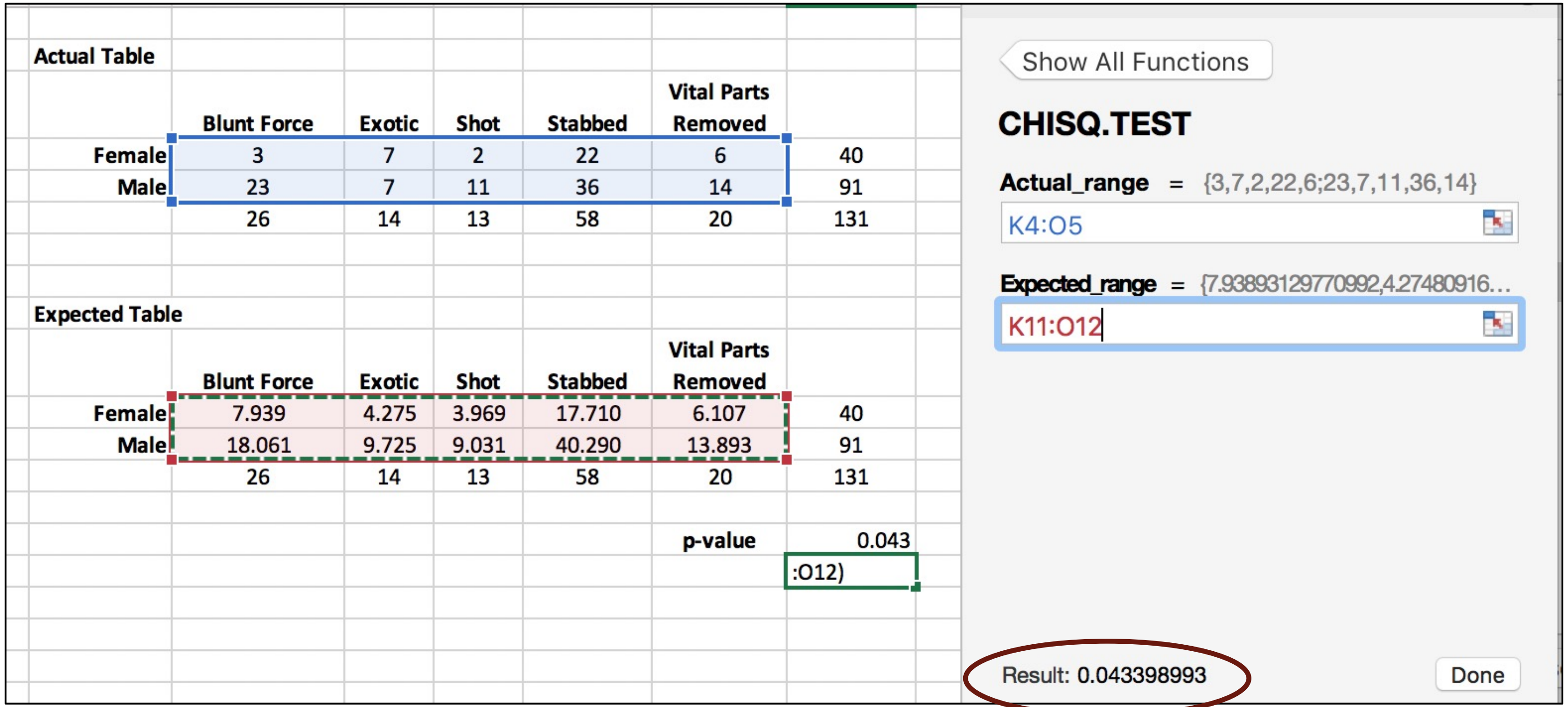
	Blunt Force	Exotic	Shot	Stabbed	Vital Parts Removed	
Female	13.977	6.477	1.364	30.682	7.500	60
Male	27.023	12.523	2.636	59.318	14.500	116
	41	19	4	90	22	176

χ^2 *Testing:* Friday the 13^{th}



p-value = 0.1575, so at α = 0.05, *Gender* is not related to how one is killed in the *Friday the 13th* films.

χ^2 *Testing:* Halloween



Actual Table

	Blunt Force	Exotic	Shot	Stabbed	Vital Parts Removed	
Female	3	7	2	22	6	40
Male	23	7	11	36	14	91
	26	14	13	58	20	131

Expected Table

	Blunt Force	Exotic	Shot	Stabbed	Vital Parts Removed	
Female	7.939	4.275	3.969	17.710	6.107	40
Male	18.061	9.725	9.031	40.290	13.893	91
	26	14	13	58	20	131
					p-value	0.043
						:O12)

p-value = 0.043, so at α = 0.05, *Gender* IS related to how one is killed in the *Halloween* films!

χ^2 *Tests:* Pharmaceutical Efficacy

The results of a random sample of children with pain from musculoskeletal injuries treated with acetaminophen, ibuprofen, or codeine are shown in the table.

Response	*Acetaminophen*	*Ibuprofen*	*Codeine*
Significant Improvement	58	81	61
Slight Improvement	42	19	39

At $\alpha = 0.01$, are you convinced the treatments and results are independent?

a) What are the null & alternative hypotheses here?
b) What values would you *expect* to see if the null hypothesis still held true?
c) What is the p-value and your conclusion?

χ^2 *Tests:* Pharmaceutical Efficacy

The results of a random sample of children with pain from musculoskeletal injuries treated with acetaminophen, ibuprofen, or codeine are shown in the table.

Response	*Acetaminophen*	*Ibuprofen*	*Codeine*
Significant Improvement	58	81	61
Slight Improvement	42	19	39

At $\alpha = 0.01$, are you convinced the treatments and results are independent?

a) What are the null & alternative hypotheses here?

Soln: H_0: Treatment and response are independent vs H_a: Treatment is related to response

b) What values would you *expect* to see if the null hypothesis still held true?

Soln:

Response	*Acetaminophen*	*Ibuprofen*	*Codeine*	*Total*
Significant Improvement	58 (66.7)	81 (66.7)	61 (66.7)	200
Slight Improvement	42 (33.3)	19 (33.3)	39 (33.3)	100

c) What is the p-value and your conclusion?

Soln: Using Excel, p-value = 0.0009 < 0.01. There is a difference between the medications.

Analysis of Variance

Setting: There are *quantitative* measurements divided into k different *categories* of data.

Q: Are the *categories* related to the quantitative outcomes?

H_0: $\mu_A = \mu_B = \ldots = \mu_k$
H_a: at least two of μ_A, μ_B,…, μ_k differ

***F*-Test Statistic** compares differences *between* the groups to the differences *within* the groups.

Reject H_0 if the p-value $< \alpha$, indicating there is a larger *between-group* difference than *within-groups,* and the results are significantly different between groups.

*Note: The F-value and the *p-value* use $(n – k)$ and $(k – 1)$ degrees of freedom.

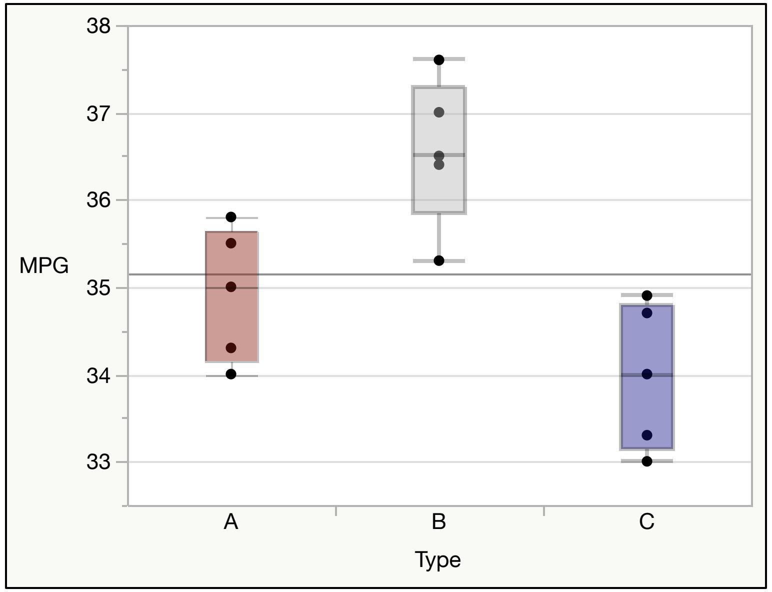
ANOVA: Gasoline Mileage

Is there a significant *mpg* difference between three *types of gasoline* (A, B, and C)? Each randomly selected car is test-driven using the chosen *gasoline type*.

- Response variable is *gasoline mileage*, in miles per gallon (*mpg*)
- *Gasoline types* (A, B, or C) are the treatments, or groups

Let X_{ij} denote the mileage x of the j^{th} car (j = 1,2, … , 5) using gasoline type i (i = A, B, or C).

Type A	*Type B*	*Type C*
$x_{A1} = 34.0$	$x_{B1} = 35.3$	$x_{C1} = 33.3$
$x_{A2} = 35.0$	$x_{B2} = 36.5$	$x_{C2} = 34.0$
$x_{A3} = 34.3$	$x_{B3} = 36.4$	$x_{C3} = 34.7$
$x_{A4} = 35.5$	$x_{B4} = 37.0$	$x_{C4} = 33.0$
$x_{A5} = 35.8$	$x_{B5} = 37.6$	$x_{C5} = 34.9$



ANOVA: Gasoline Mileage

Sample means estimate population type means

$\bar{x}_{\mathrm{A}}$ = 34.92 mpg estimates μ_{A}
$\bar{x}_{\mathrm{B}}$ = 36.56 mpg estimates μ_{B}
$\bar{x}_{\mathrm{C}}$ = 33.98 mpg estimates μ_{C}

Sample standard deviation are also calculated

s_{A} = 0.7662 mpg estimates σ_{A}
s_{B} = 0.8503 mpg estimates σ_{B}
s_{C} = 0.8349 mpg estimates σ_{C}

ANOVA: Assumptions

Completely randomized experimental design

- The sample has been selected randomly for each of *k* treatments

Constant variance

- The *k* population treatment responses have roughly the same variance

Normality

- The *k* treatment responses are all basically normally distributed

Independence

- The samples are randomly selected and independent

*Note: One-way ANOVA is not very sensitive to either the equal variances or normality assumptions.

Differences: Between Types

Is there a statistically significant difference *between* the types?

H_0: $\mu_A = \mu_B = \mu_C$
H_a: at least two of μ_A, μ_B, μ_C differ

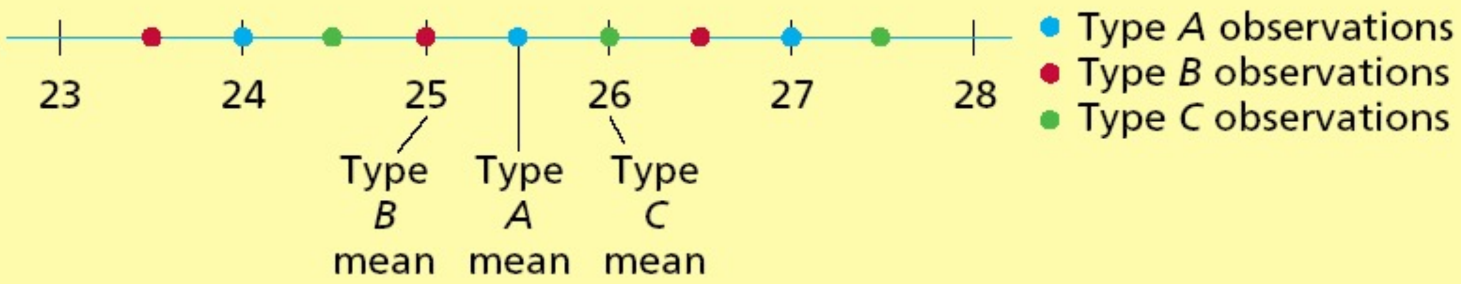
Compare variability **Between Types** (SS_{Type}) to **Within Types** (SS_{Error})

- **Between Types** is the variability of the means from group to group
- **Within Types** is the variability within each set of groups

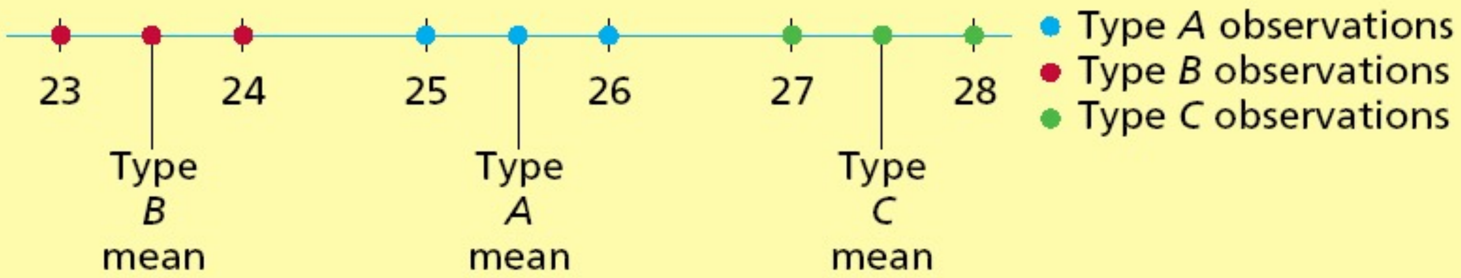
Possible results:

1. SS_{Type} isn't large compared to SS_{Error} → Cannot reject H_0: $\mu_A = \mu_B = \mu_C$
2. SS_{Type} IS large compared to SS_{Error} → Reject H_0 and presume at least two of μ_A, μ_B, μ_C differ

Differences: Gas Mileages



(a) Between-treatment variability is not large compared to within-treatment variability. Do not reject H_0: $\mu_A = \mu_B = \mu_C$



(b) Between-treatment variability is large compared to within-treatment variability. Reject H_0: $\mu_A = \mu_B = \mu_C$

F-Test: Type Differences

We want to compare *k* treatment means:

H_0: $\mu_A = \mu_B = \mu_C$
H_a: at least two of μ_A, μ_B, μ_C differ

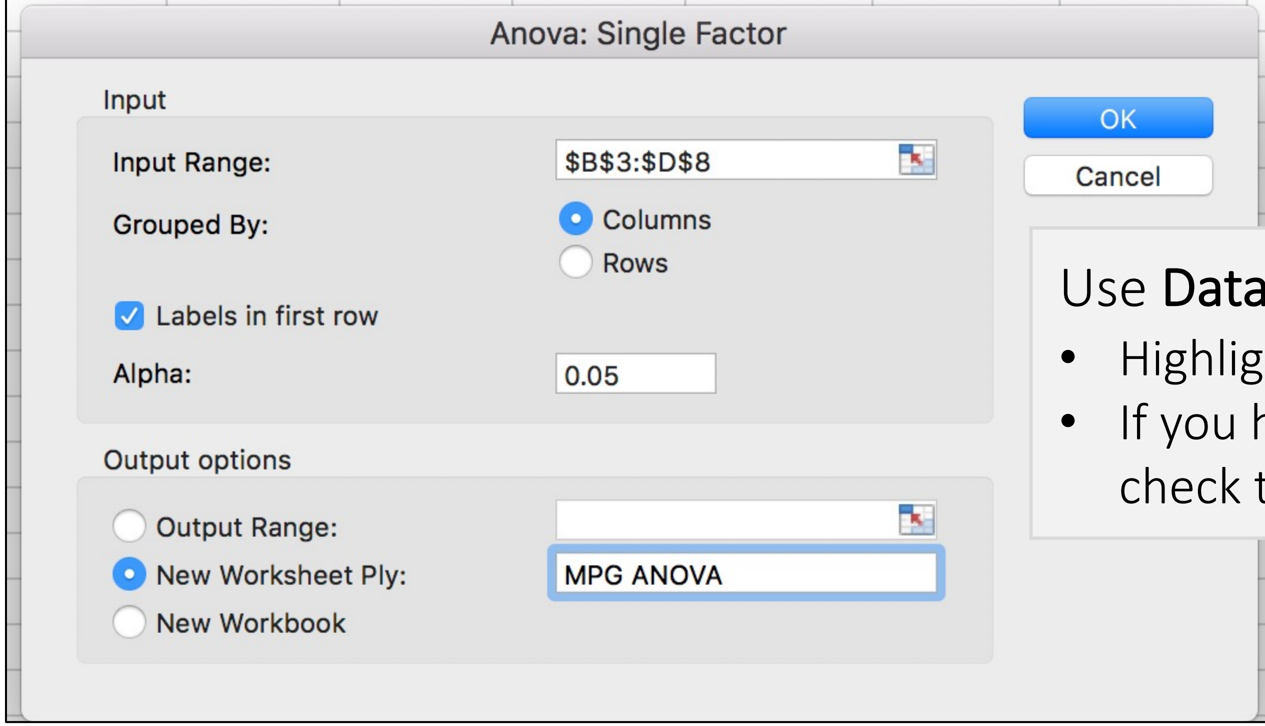
F statistic: $$F = \frac{\text{MST}}{\text{MSE}} = \frac{\text{SST}/(k-1)}{\text{SSE}/(n-k)}$$

with $df_1 = (k - 1)$ and $df_2 = (n - k)$ [here, $k = 3$]

As usual, the *p*-value is the area under the *F* curve to the right of *F*.

F-Test: Gas Mileages Excel

Type A	Type B	Type C
34	35.3	33.3
35	36.5	34
34.3	36.4	34.7
35.5	37	33
35.8	37.6	34.9



Use **Data – Data Analysis – ANOVA: Single Factor.**

- Highlight the matrix of values.
- If you highlighted the variable names, be sure to check the “Labels in First Row” box

F-Test: Gas Mileages Excel

Anova: Single Factor						
SUMMARY						
Groups	*Count*	*Sum*	*Average*	*Variance*		
Type A	5	174.6	34.92	0.587		
Type B	5	182.8	36.56	0.723		
Type C	5	169.9	33.98	0.697		
ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	17.049	2	8.525	12.742	0.001	3.885
Within Groups	8.028	12	0.669			
Total	25.077	14				

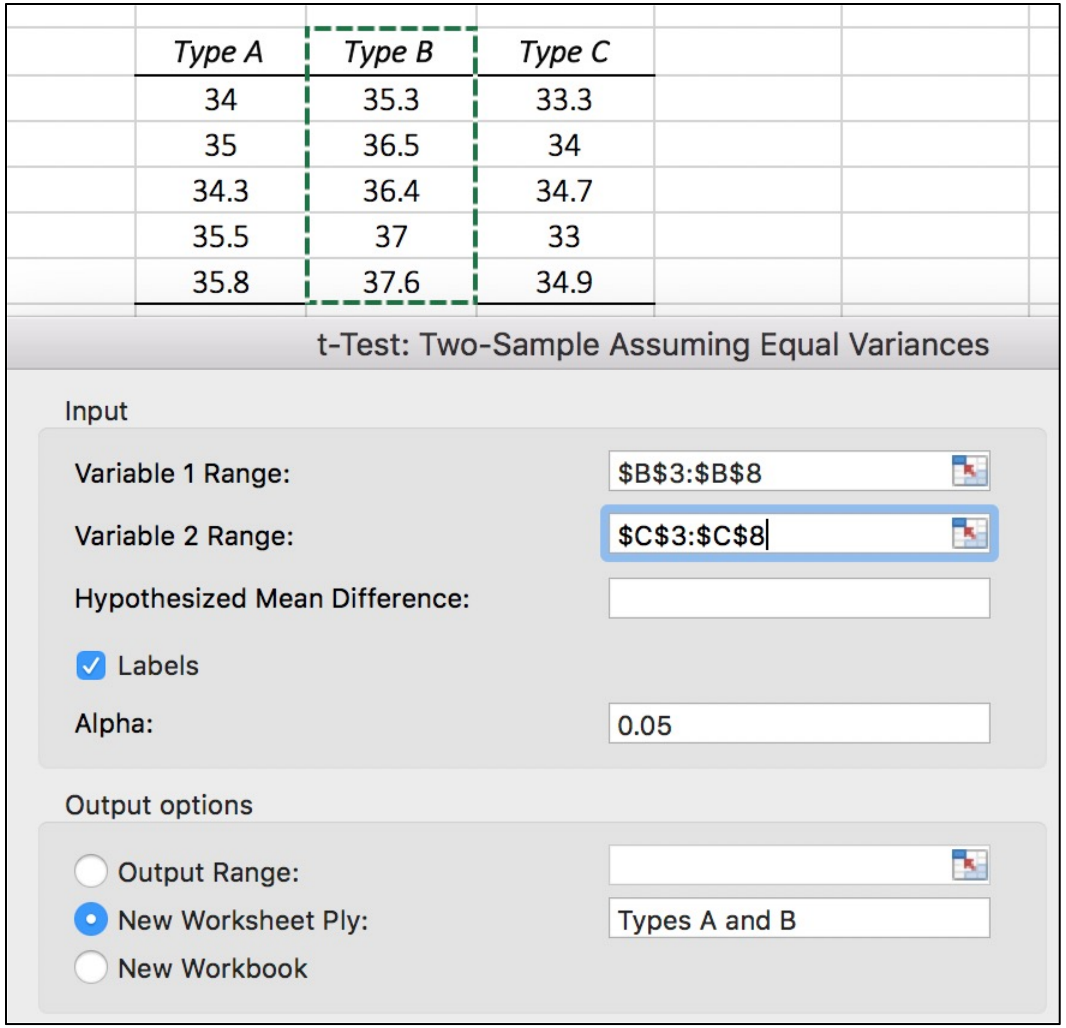
We can reject H_0 at the 0.05 significance level.

But where do the differences occur?

Gasoline Type A vs B: Excel

To test each *pair* of gasoline types, use

Data – Data Analysis – t-test: Two Sample Assuming Equal Variances.



Type A	Type B	Type C
34	35.3	33.3
35	36.5	34
34.3	36.4	34.7
35.5	37	33
35.8	37.6	34.9

Gasoline Type A vs B: Excel

t-Test: Two-Sample Assuming Equal Variances		
	Type A	*Type B*
Mean	34.92	36.56
Variance	0.587	0.723
Observations	5	5
Pooled Variance	0.655	
Hypothesized Mean Difference	0	
df	8	
t Stat	-3.204	
P(T<=t) one-tail	0.006	
t Critical one-tail	1.860	
P(T<=t) two-tail	0.013	
t Critical two-tail	2.306	

For a two-sided test, gasoline types A & B are statistically different at $\alpha = 0.05$.

What about at $\alpha = 0.01$?

Gasoline Type A vs C: Excel

t-Test: Two-Sample Assuming Equal Variances		
	Type A	*Type C*
Mean	34.92	33.98
Variance	0.587	0.697
Observations	5	5
Pooled Variance	0.642	
Hypothesized Mean Difference	0	
df	8	
t Stat	1.855	
P(T<=t) one-tail	0.050	
t Critical one-tail	1.860	
P(T<=t) two-tail	0.101	
t Critical two-tail	2.306	

For a two-sided test, gasoline types A & C are NOT statistically different at $\alpha = 0.05$.

Gasoline Type B vs C: Excel

t-Test: Two-Sample Assuming Equal Variances		
	Type B	*Type C*
Mean	36.56	33.98
Variance	0.723	0.697
Observations	5	5
Pooled Variance	0.71	
Hypothesized Mean Difference	0	
df	8	
t Stat	4.841	
P(T<=t) one-tail	0.001	
t Critical one-tail	1.860	
P(T<=t) two-tail	0.001	
t Critical two-tail	2.306	

For a two-sided test, gasoline types B & C are statistically different at $\alpha = 0.05$.

What about at $\alpha = 0.01$?

ANOVA: Gourmet Cookie Case

Weight	Line
2	1
2.05	1
2.01	1
2	1
2.1	1
2.05	1
2.05	1
2.3	1
2	1
2.15	1
1.9	2
1.9	2
2.1	2
2	2
1.95	2
2.05	2
1.9	2
2.05	2
2.15	2
1.9	2
2	2
2.15	2
2.2	3
2.1	3
2.15	3
2.3	3
2.35	3
2.05	3
2.25	3
2.3	3
2.4	3
2.35	3

The Gourmet Cookie Company produces its best-selling cookie, Chocolate Macadamia Delites, on all three of its cookie production lines and ovens. The quality control manager is concerned that the cookies on one or more of the lines weigh more than they are supposed to. The average weight of the cookies on all three lines should be 2.0 ounces.

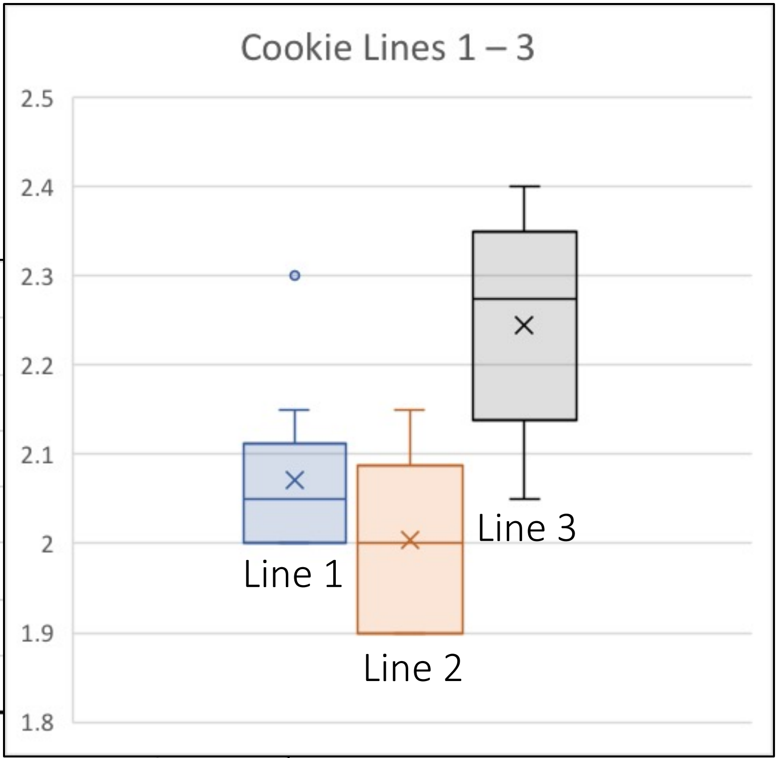
Test the hypothesis at the 0.05 level of significance that there is no difference in cookie weights between the three cookie production lines.

What is your conclusion?

Gourmet Cookie Case: Excel

There does appear to be a significant difference, but between which lines?

Anova: Single Factor						
SUMMARY						
Groups	*Count*	*Sum*	*Average*	*Variance*		
Line 1	10	20.71	2.0710	0.0089		
Line 2	12	24.05	2.0042	0.0093		
Line 3	10	22.45	2.2450	0.0136		
ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	0.329	2	0.1644	15.6741	2.42743E-05	3.328
Within Groups	0.304	29	0.0105			
Total	0.633	31				



Cookie Lines 1 vs 2: Excel

t-Test: Two-Sample Assuming Equal Variances		
	Line 1	*Line 2*
Mean	2.0710	2.0042
Variance	0.0089	0.0093
Observations	10	12
Pooled Variance	0.0091	
Hypothesized Mean Difference	0	
df	20	
t Stat	1.6363	
P(T<=t) one-tail	0.0587	
t Critical one-tail	1.7247	
P(T<=t) two-tail	0.1174	
t Critical two-tail	2.0860	

For a two-sided test, *gourmet cookie lines 1* and *2* do not seem to be statistically different at $\alpha = 0.05$.

Cookie Lines 1 vs 3: Excel

t-Test: Two-Sample Assuming Equal Variances		
	Line 1	*Line 3*
Mean	2.0710	2.2450
Variance	0.0089	0.0136
Observations	10	10
Pooled Variance	0.0112	
Hypothesized Mean Difference	0	
df	18	
t Stat	-3.6733	
P(T<=t) one-tail	0.0009	
t Critical one-tail	1.7341	
P(T<=t) two-tail	0.0017	
t Critical two-tail	2.1009	

For a two-sided test, *gourmet cookie lines 1* and *3* DO seem to be statistically different at $\alpha = 0.05$.

What about at $\alpha = 0.01$?

Cookie Lines 2 vs 3: Excel

t-Test: Two-Sample Assuming Equal Variances		
	Line 2	*Line 3*
Mean	2.0042	2.2450
Variance	0.0093	0.0136
Observations	12	10
Pooled Variance	0.0112	
Hypothesized Mean Difference	0	
df	20	
t Stat	-5.3084	
P(T<=t) one-tail	1.7E-05	
t Critical one-tail	1.7247	
P(T<=t) two-tail	3.4E-05	
t Critical two-tail	2.0860	

For a two-sided test, *gourmet cookie lines 2* and *3* DO seem to be statistically different at $\alpha = 0.05$.

What about at $\alpha = 0.01$?

USC School of Business

Two-Factor ANOVA

We want to study the effects of *two factors* on a response variable

- Perhaps there are two potential categorical factors related to *Y*
- **Two-Factor ANOVA** controls for the differences in these factors, much like Regression Analysis, as well as any potential interactions

The assessment is much the same as One-Way ANOVA, but there are more deviations calculated due to the added factors

- One row is devoted to variability due to *Factor A*
- One row is devoted to variability due to *Factor B*
- One row is devoted to variability due to the *interaction* between *A* and *B*
- One row is devoted to *Error* variability
- One row is devoted to the overall, or *Total*, variability

Two-Factor ANOVA

Source of Variation	*Degrees Of Freedom*	*Sum of Squares*	*Mean Squares*	*F*
Factor A	$r - 1$	SSA	MSA = SSA /(r – 1)	MSA/MSE
Factor B	$c - 1$	SSB	MSB = SSB /(c – 1)	MSB/MSE
AB (Interaction)	$(r - 1)(c - 1)$	SSAB	MSAB = SSAB /(r – 1)(c – 1)	MSAB/MSE
Error	$rc(n - 1)$	SSE	MSE = SSE/rc(n – 1)	
Total	$n - 1$	SST		

Two-Factor ANOVA: Juice Marketing

A new apple juice product was entering the marketplace. It had three distinct advantages relative to existing apple juices.

1. **Quality:** It was not a concentrate and was therefore considered to be of *higher quality* than many similar products.
2. **Price:** As one of the first juices packaged in cartons, it was *cheaper* than competing products.
3. **Convenience:** Partly because of the packaging, it was more *convenient*.

The Marketing Director wants to know:

a) Which **advantage** should be emphasized in advertisements?
b) Which **medium**, local television or newspapers, is better for sales?

Two-Factor ANOVA: Juice Marketing

Six cities with similar demographics are chosen, and a different combination of *Media* and *Marketing Strategy* is tried in each.

Sales of apple juice for ten weeks following the start of the ad campaigns are recorded for each city.

	Convenience	Quality	Price
Local Television	City 1	City 3	City 5
Newspaper	City 2	City 4	City 6

Differences: Between Factors

Are there statistically significant differences between the factors?

Factor *i*

- H_0: $\mu_{1j} = \mu_{2j} = \ldots = \mu_{kj}$
- H_a: at least two of $\mu_{1j}, \mu_{2j}, \ldots, \mu_{kj}$ differ

Factor *j*

- H_0: $\mu_{i1} = \mu_{i2} = \ldots = \mu_{im}$
- H_a: at least two of $\mu_{i1}, \mu_{i2}, \ldots, \mu_{im}$ differ

Interaction *ij*

- H_0: $\mu_{11} = \mu_{12} = \ldots = \mu_{im}$
- H_a: at least two of $\mu_{11}, \mu_{12}, \ldots, \mu_{im}$ differ

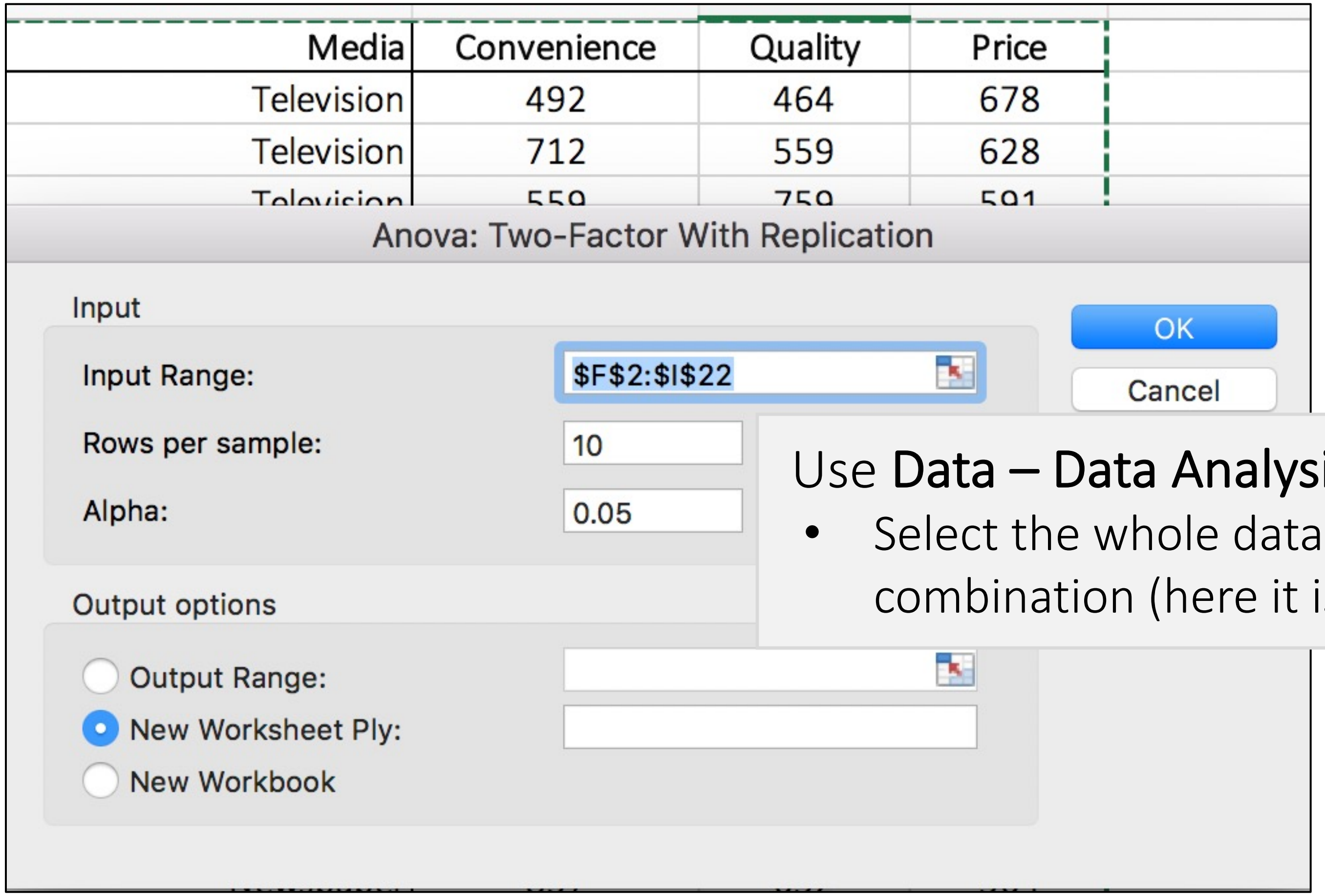
F-Test: Juice Marketing – Excel

Data needs to be in a *contingency table* format in Excel.

One column indicates the value of the first independent variable, then a separate column is needed for each *level* of the second variable (or factor).

	Advantage		
Media	Convenience	Quality	Price
Television	492	464	678
Television	712	559	628
Television	559	759	591
Television	447	558	633
Television	480	528	684
Television	624	670	761
Television	547	534	691
Television	444	657	549
Television	583	557	580
Television	672	474	645
Newspaper	690	577	803
Newspaper	650	616	583
Newspaper	705	708	525
Newspaper	653	486	497
Newspaper	577	480	813
Newspaper	837	652	564
Newspaper	629	585	708
Newspaper	799	538	545
Newspaper	498	581	616
Newspaper	842	797	586

F-Test: Juice Marketing – Excel



Use **Data – Data Analysis – Anova: Two-Factor with Replication.**
- Select the whole data table and input the number of rows per combination (here it is 10).

F-Test: Juice Marketing – Excel

ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Sample	31740	1	31740	3.4143	0.0701	4.0195
Columns	21720	2	10860	1.1682	0.3187	3.1682
Interaction	60760	2	30380	3.2680	0.0457	3.1682
Within	501998	54	9296			
Total	616218	59				

The p-value for *Sample* tells us about the significance of the different *Media* types. A value of 0.0701 is pretty close to significance at $\alpha = 0.05$.

The ANOVA output shows significance in the *Interaction* category, meaning a combination of *Media* and *Marketing Strategy* is related to *Sales*.

F-Test: Juice Marketing – Excel

SUMMARY	Convenience	Quality	Price	Total
Television				
Count	10	10	10	30
Sum	5560	5760	6440	17760
Average	556	576	644	592
Variance	8614.667	8541.778	3884.667	7997.586
Newspaper				
Count	10	10	10	30
Sum	6880	6020	6240	19140
Average	688	602	624	638
Variance	12566.889	9527.556	12642.000	12156.828

A closer look at the differences between the two *Media* types shows that the *Newspaper* option has a higher average *Sales.*

JMP Instructions for Categorical Analyses

χ^2 *Tests:* Payment Methods

The *American Bankers Association* collects data regarding how consumers pay for in-store purchases. In 1999, the usages below were reported. A sample was taken in 2003 to assess the same information.

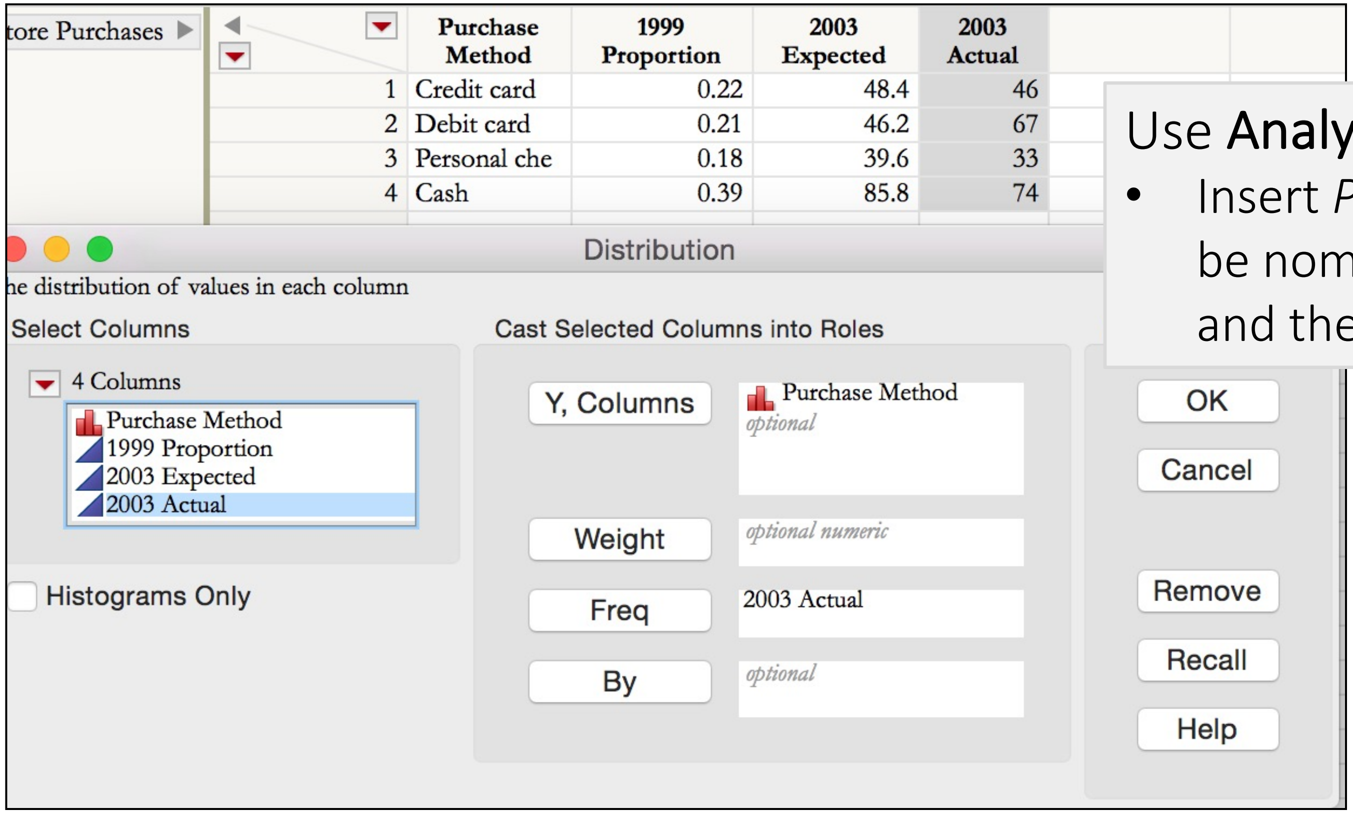
Q: Have the methods of payment significantly changed in those 5 years?

Purchase Method	1999 Proportion
Credit card	0.22
Debit card	0.21
Personal check	0.18
Cash	0.39

H_0: $p_{credit} = 0.22$, $p_{debit} = 0.21$, $p_{check} = 0.18$, $p_{cash} = 0.39$
H_a: at least one probability differs from H_0

χ^2 *Tests:* Payments (JMP)

H_0: $p_{credit} = 0.22$, $p_{debit} = 0.21$, $p_{check} = 0.18$, $p_{cash} = 0.39$
vs H_a: at least one probability differs from H_0



Use **Analyze – Distribution**

- Insert *Purchase Method* (must be nominal) as the *Y* variable and the *2003 Actual* as the **Freq.**

χ^2 *Tests:* Payments (JMP)

H_0: $p_{credit} = 0.22$, $p_{debit} = 0.21$, $p_{check} = 0.18$, $p_{cash} = 0.39$
vs H_a: at least one probability differs from H_0

ibution Fit Y by X Multivariate Graph Builder Bubble Plot

	Purchase Method	1999 Proportion	2003 Expected	2003 Actual
1	Credit card	0.22	48.4	46
2	Debit card	0.21	46.2	67
3	Personal check	0.18	39.6	33
4	Cash	0.39	85.8	74

Under the *Purchase Method* red triangle, select **Test Probabilities**. Here, insert the hypothesized proportions from 1999.

Distributions

Purchase Method

Frequencies

Level	Count	Prob
Cash	74	0.33636
Credit card	46	0.20909
Debit card	67	0.30455
Personal check	33	0.15000
Total	220	1.00000

N Missing 0
4 Levels

Test Probabilities

Level	Estim Prob	Hypoth Prob
Cash	0.33636	0.39000
Credit card	0.20909	0.22000
Debit card	0.30455	0.21000
Personal check	0.15000	0.18000

Click then Enter Hypothesized Probabilities.

χ^2 *Tests:* Payments (JMP)

H_0: $p_{credit} = 0.22$, $p_{debit} = 0.21$, $p_{check} = 0.18$, $p_{cash} = 0.39$
vs H_a: at least one probability differs from H_0

Test Probabilities

Level	Estim Prob	Hypoth Prob
Cash	0.33636	0.39000
Credit card	0.20909	0.22000
Debit card	0.30455	0.21000
Personal check	0.15000	0.18000

Test	ChiSquare	DF	Prob>Chisq
Likelihood Ratio	11.2001	3	0.0107*
Pearson	12.2064	3	0.0067*

Method: Fix hypothesized values, rescale omitted

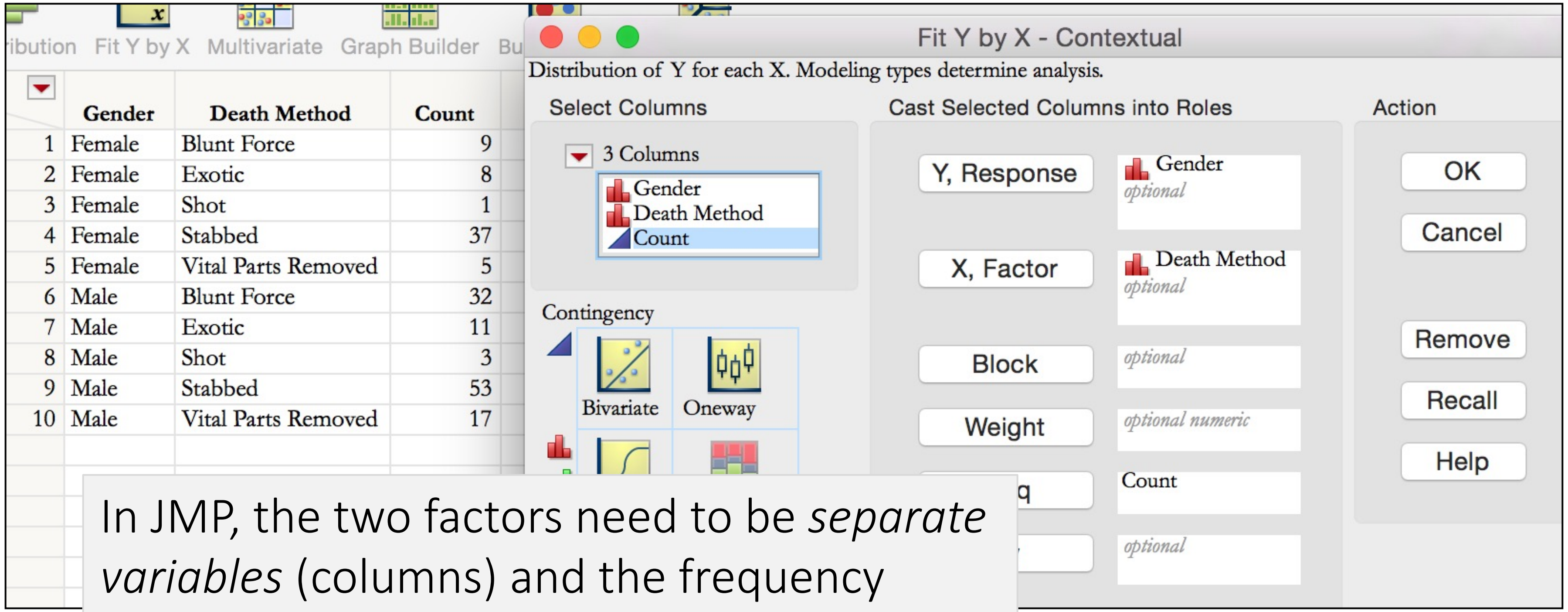
The **Pearson Test** *p*-value of 0.0067 indicates there IS a significant difference between 1999 and 2003.

χ^2 *Testing:* Friday the 13th

At α = 0.05, test

H_0: Gender is independent of how one is killed

H_a: Gender is related to how one is killed



	Gender	Death Method	Count
1	Female	Blunt Force	9
2	Female	Exotic	8
3	Female	Shot	1
4	Female	Stabbed	37
5	Female	Vital Parts Removed	5
6	Male	Blunt Force	32
7	Male	Exotic	11
8	Male	Shot	3
9	Male	Stabbed	53
10	Male	Vital Parts Removed	17

In JMP, the two factors need to be *separate variables* (columns) and the frequency column has the **Counts** for each combination. Use **Analyze – Fit *Y* by *X***.

USC School of Business

χ^2 *Testing:* Friday the 13th

Contingency Table

Death Method	Count / Expected	Gender: Female	Gender: Male	
	Blunt Force	9	32	41
		13.9773	27.0227	
	Exotic	8	11	19
		6.47727	12.5227	
	Shot	1	3	4
		1.36364	2.63636	
	Stabbed	37	53	90
		30.6818	59.3182	
	Vital Parts Removed	5	17	22
		7.5	14.5	
		60	116	176

Tests

N	DF	-LogLike	RSquare (U)
176	4	3.4243080	0.0303

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	6.849	0.1441
Pearson	6.618	0.1575

At $\alpha = 0.05$, test

H_0: Gender is independent of how one is killed in *Friday the 13th* movies

H_a: Gender is related to how one is killed in *Friday the 13th* movies

The **Pearson Test** *p*-value of 0.1575 indicates *Gender* is not related to how one is killed in the *Friday the 13th* films.

χ^2 *Testing:* Halloween

USC School of Business

Contingency Table

Death Method	Count / Expected	Female	Male	
	Blunt Force	3 7.93893	23 18.0611	26
	Exotic	7 4.27481	7 9.72519	14
	Shot	2 3.96947	11 9.03053	13
	Stabbed	22 17.7099	36 40.2901	58
	Vital Parts Removed	6 6.10687	14 13.8931	20
		40	91	131

(Columns: Gender)

Tests

N	DF	-LogLike	RSquare (U)
131	4	5.3105252	0.0659

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	10.621	0.0312*
Pearson	9.830	0.0434*

At α = 0.05, test

H_0: Gender is independent of how one is killed in *Halloween* movies

H_a: Gender is related to how one is killed in *Halloween* movies

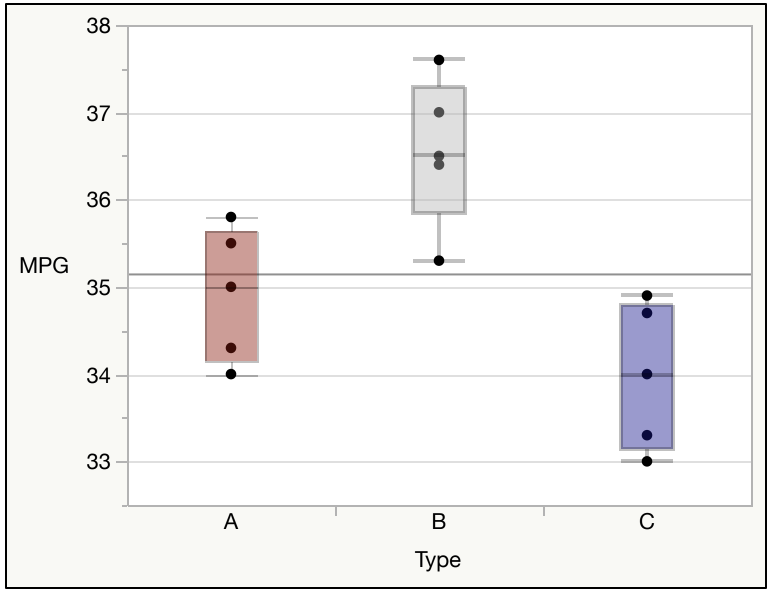
The **Pearson Test** *p*-value of 0.0434 indicates *Gender* IS related to how one is killed in the *Halloween* films!

ANOVA: Gasoline Mileage

Is there a significant *mpg* difference between three *types of gasoline* (A, B, and C)? Each randomly selected car is test-driven using the chosen *gasoline type*.

- Response variable is *gasoline mileage*, in miles per gallon (*mpg*)
- *Gasoline types* (A, B, or C) are the treatments, or groups

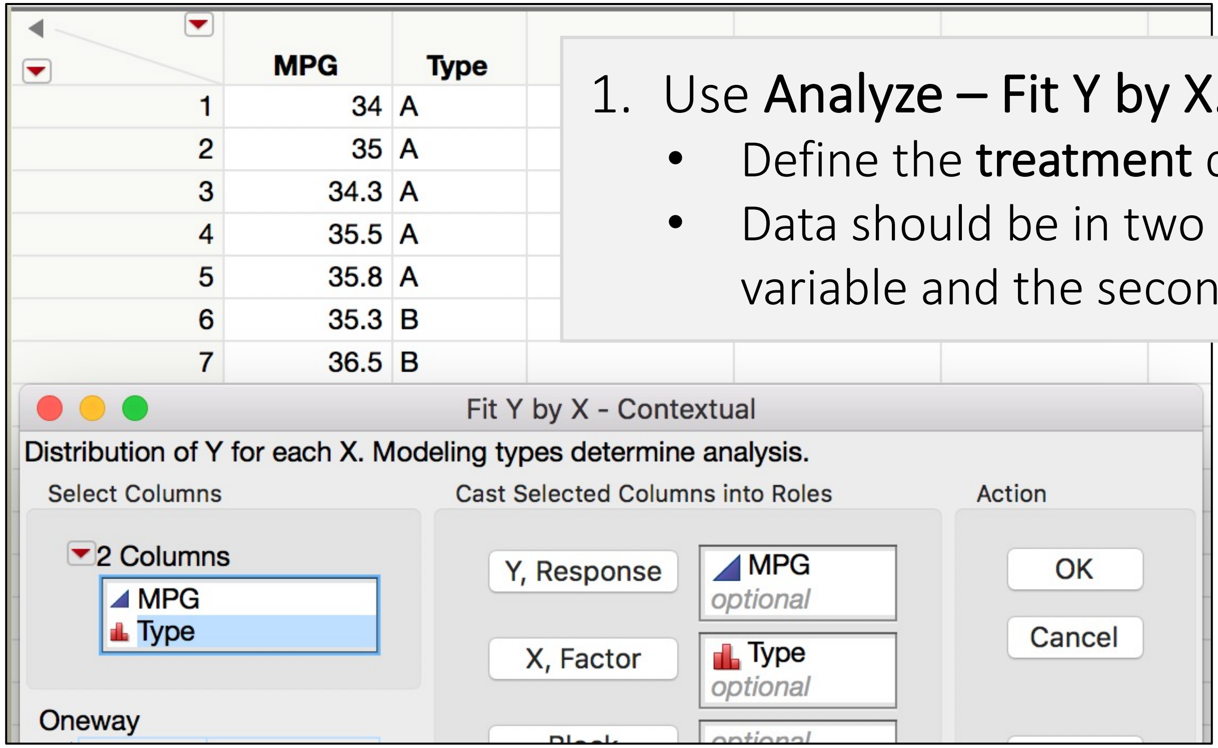
Type A	*Type B*	*Type C*
x_{A1} = 34.0	x_{B1} = 35.3	x_{C1} = 33.3
x_{A2} = 35.0	x_{B2} = 36.5	x_{C2} = 34.0
x_{A3} = 34.3	x_{B3} = 36.4	x_{C3} = 34.7
x_{A4} = 35.5	x_{B4} = 37.0	x_{C4} = 33.0
x_{A5} = 35.8	x_{B5} = 37.6	x_{C5} = 34.9



Let X_{ij} denote the mileage *x* of the j^{th} car (j = 1,2, … , 5) using gasoline type *i* (*i* = A, B, or C).

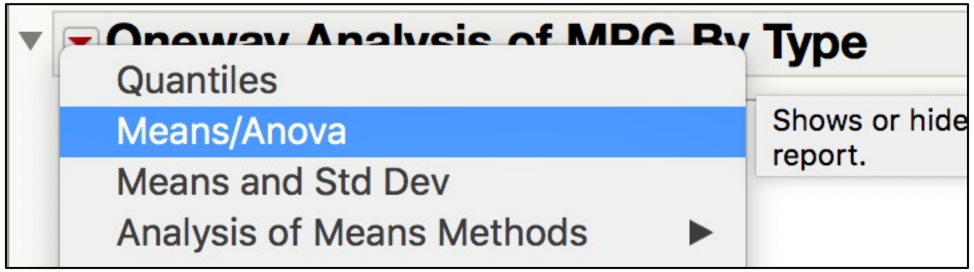
USC School of Business

F-Test: Gas Mileages JMP



1. Use **Analyze – Fit Y by X.**
   - Define the **treatment** column as *Character, Nominal*.
   - Data should be in two columns: One column for the *Y* variable and the second column for *X* (treatment).

2. Click on the red triangle and select the **Means/Anova** option.



F-Test: Gas Mileages JMP

We can reject H_0 at the 0.05 significance level.

But where do the differences occur?

Oneway Anova

Summary of Fit

Rsquare	0.680
Adj Rsquare	0.627
Root Mean Square Error	0.818
Mean of Response	35.153
Observations (or Sum Wgts)	15.000

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Type	2	17.049	8.525	12.742	0.0011*
Error	12	8.028	0.669		
C. Total	14	25.077			

Means for Oneway Anova

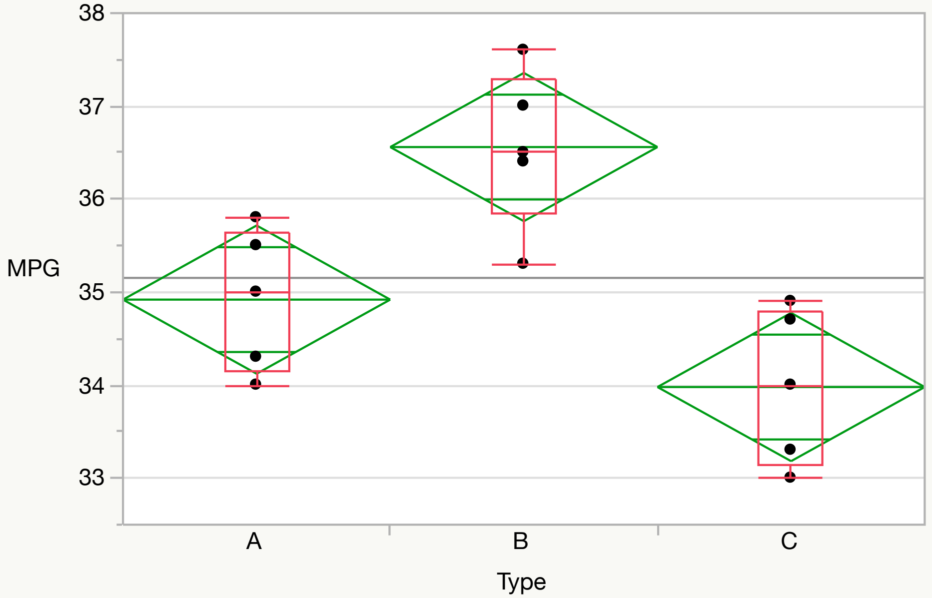
Level	Number	Mean	Std Error	Lower 95%	Upper 95%
A	5	34.920	0.366	34.123	35.717
B	5	36.560	0.366	35.763	37.357
C	5	33.980	0.366	33.183	34.777

Std Error uses a pooled estimate of error variance

F-Test: Gas Mileages

Creating plots in JMP: under the red triangle, choose **Display Options** and select **Boxplot**.

The summaries are found under **Means/ANOVA**, as well.



Means for Oneway Anova

Level	Number	Mean	Std Error	Lower 95%	Upper 95%
A	5	34.920	0.366	34.123	35.717
B	5	36.560	0.366	35.763	37.357
C	5	33.980	0.366	33.183	34.777

Std Error uses a pooled estimate of error variance

Summary of Fit

Rsquare	0.680
Adj Rsquare	0.627
Root Mean Square Error	0.818
Mean of Response	35.153
Observations (or Sum Wgts)	15.000

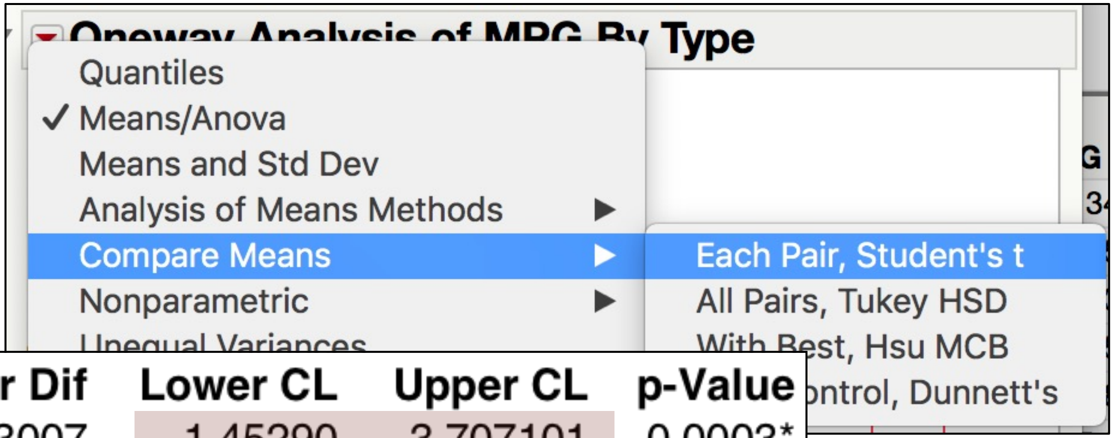
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Type	2	17.049	8.525	12.742	0.0011*
Error	12	8.028	0.669		
C. Total	14	25.077			

Pairwise Tests: Gas Mileages

USC School of Business

JMP creates confidence intervals, too. Under the red triangle, use **Compare Means – Each Pair, Student's t.**



Level	- Level	Difference	Std Err Dif	Lower CL	Upper CL	p-Value
B	C	2.580000	0.5173007	1.45290	3.707101	0.0003*
B	A	1.640000	0.5173007	0.51290	2.767101	0.0081*
A	C	0.940000	0.5173007	-0.18710	2.067101	0.0942

To calculate confidence intervals for $\mu_B - \mu_A$ manually:

$$(\bar{x}_B - \bar{x}_A) \pm t_{\alpha/2}\sqrt{MSE\left(\frac{1}{n_B}+\frac{1}{n_A}\right)}$$

$t_{\alpha/2}$ is based on $n - k$ degrees of freedom.

$$95\%CI(\bar{x}_B - \bar{x}_A):\ (36.56 - 34.92) \pm 2.179\sqrt{0.669\left(\tfrac{1}{5}+\tfrac{1}{5}\right)} \rightarrow 1.64 \pm 1.127 = \boxed{[0.513, 2.767]}$$

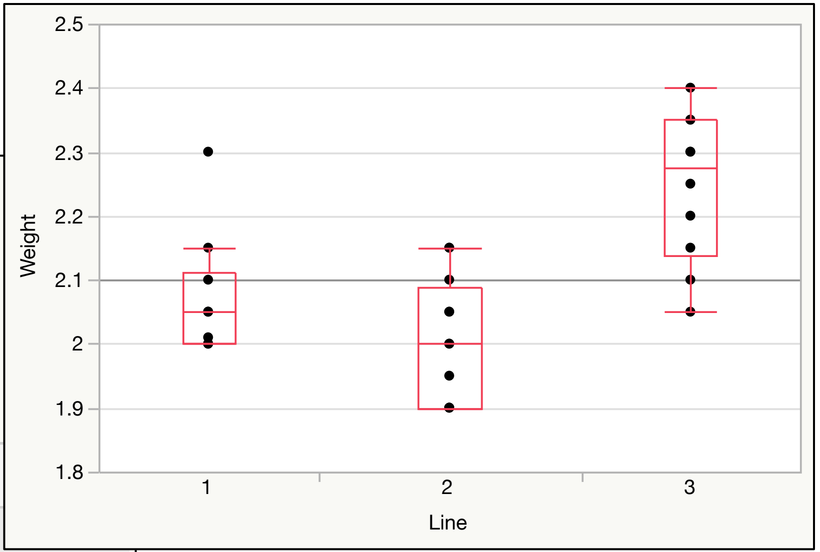
Gourmet Cookie Case

Weight	Line
2	1
2.05	1
2.01	1
2	1
2.1	1
2.05	1
2.05	1
2.3	1
2	1
2.15	1
1.9	2
1.9	2
2.1	2
2	2
1.95	2
2.05	2
1.9	2
2.05	2
2.15	2
1.9	2
2	2
2.15	2
2.2	3
2.1	3
2.15	3
2.3	3
2.35	3
2.05	3
2.25	3
2.3	3
2.4	3
2.35	3

The Gourmet Cookie Company produces its best-selling cookie, Chocolate Macadamia Delites, on all three of its cookie production lines and ovens. The quality control manager is concerned that the cookies on one or more of the lines weigh more than they are supposed to. The average weight of the cookies on all three lines should be 2.0 ounces.

Test the hypothesis at the 0.05 level of significance that there is no difference in cookie weights between the three cookie production lines.

Gourmet Cookie Case JMP



Summary of Fit

Rsquare	0.519
Adj Rsquare	0.486
Root Mean Square Error	0.102
Mean of Response	2.100
Observations (or Sum Wgts)	32.000

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Line	2	0.329	0.164	15.6741	<.0001*
Error	29	0.304	0.010		
C. Total	31	0.633			

There does appear to be a significant difference, but between which lines?

Means for Oneway Anova

Level	Number	Mean	Std Error	Lower 95%	Upper 95%
1	10	2.071	0.032	2.005	2.137
2	12	2.004	0.030	1.944	2.065
3	10	2.245	0.032	2.179	2.311

Std Error uses a pooled estimate of error variance

Gourmet Cookie Case

USC School of Business

Means for Oneway Anova

Level	Number	Mean	Std Error	Lower 95%	Upper 95%
1	10	2.071	0.032	2.005	2.137
2	12	2.004	0.030	1.944	2.065
3	10	2.245	0.032	2.179	2.311

Std Error uses a pooled estimate of error variance

Notice that some of these intervals don't overlap…

Under the red triangle, choose **Compare Means** and then **Each Pair, Student's t.**

Level	- Level	Difference	Std Err Dif	Lower CL	Upper CL	p-Value
3	2	0.2408	0.0439	0.1511	0.3305	<.0001*
3	1	0.1740	0.0458	0.0803	0.2677	0.0007*
1	2	0.0668	0.0439	-0.0229	0.1565	0.1384

The differences appear to be between lines 2 & 3 and also lines 1 & 3. This is verified with *t*-tests between each pair of treatments.

Apple Juice Marketing

A new apple juice product was entering the marketplace. It had three distinct advantages relative to existing apple juices.

1. **Quality:** It was not a concentrate and was therefore considered to be of *higher quality* than many similar products.
2. **Price:** As one of the first juices packaged in cartons, it was *cheaper* than competing products.
3. **Convenience:** Partly because of the packaging, it was more *convenient*.

The Marketing Director wants to know:

a) Which **advantage** should be emphasized in advertisements?
b) Which **medium**, local television or newspapers, is better for sales?

Apple Juice Marketing

Six cities with similar demographics are chosen, and a different combination of *Media* and *Marketing Strategy* is tried in each.

Sales of apple juice for ten weeks following the start of the ad campaigns are recorded for each city.

	Convenience	Quality	Price
Local Television	City 1	City 3	City 5
Newspaper	City 2	City 4	City 6

Differences: Between Factors

Are there statistically significant differences between the factors?

Factor i

- H_0: $\mu_{1j} = \mu_{2j} = \ldots = \mu_{kj}$
- H_a: at least two of $\mu_{1j}, \mu_{2j}, \ldots, \mu_{kj}$ differ

Factor j

- H_0: $\mu_{i1} = \mu_{i2} = \ldots = \mu_{im}$
- H_a: at least two of $\mu_{i1}, \mu_{i2}, \ldots, \mu_{im}$ differ

Interaction ij

- H_0: $\mu_{11} = \mu_{12} = \ldots = \mu_{im}$
- H_a: at least two of $\mu_{11}, \mu_{12}, \ldots, \mu_{im}$ differ

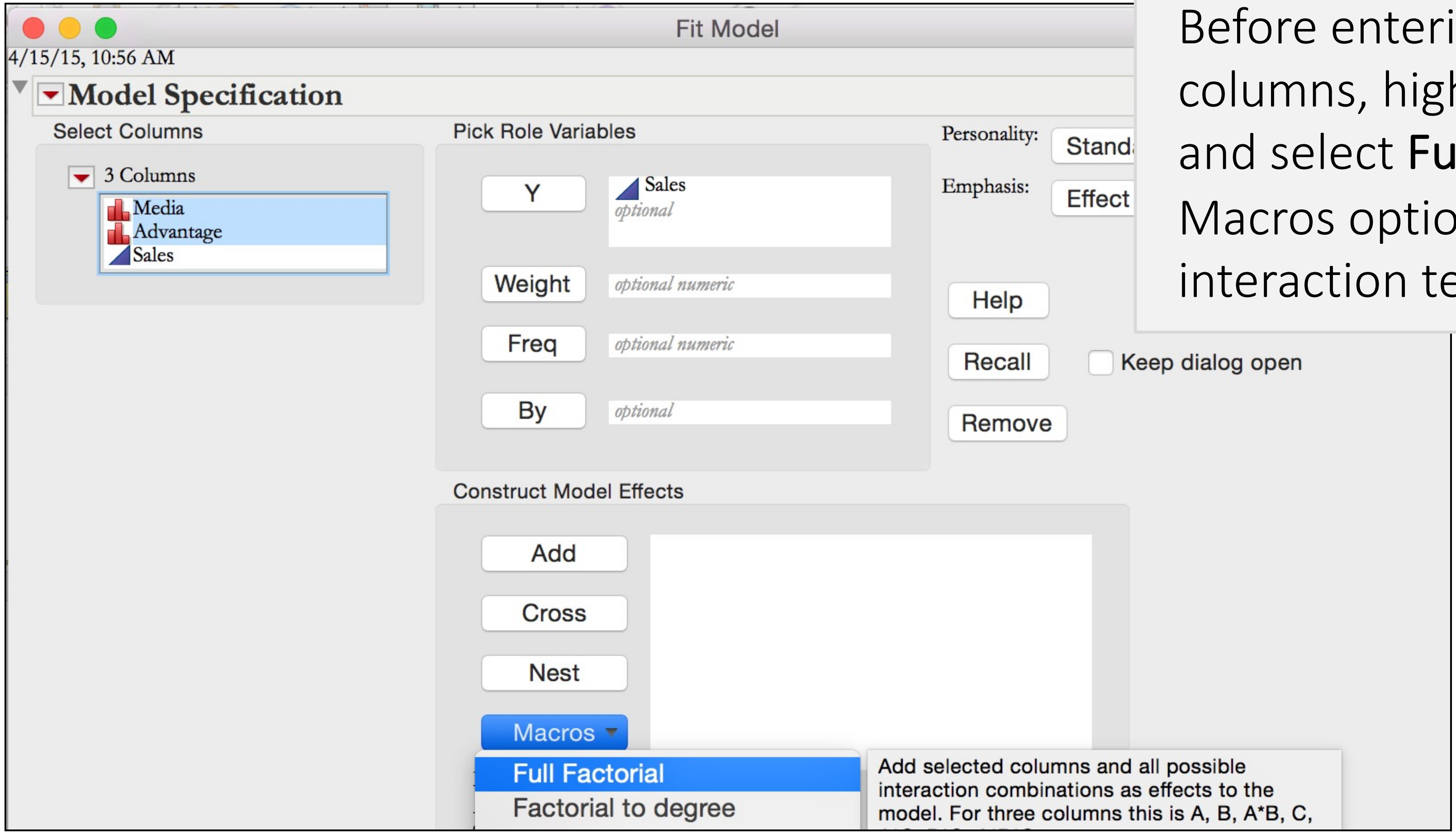
F-Test: Apple Juice Excel

Data needs to be in a *contingency table* format in Excel.

One column indicates the value of the first independent variable, then a separate column is needed for each *level* of the second variable (or factor).

	Advantage		
Media	Convenience	Quality	Price
Television	492	464	678
Television	712	559	628
Television	559	759	591
Television	447	558	633
Television	480	528	684
Television	624	670	761
Television	547	534	691
Television	444	657	549
Television	583	557	580
Television	672	474	645
Newspaper	690	577	803
Newspaper	650	616	583
Newspaper	705	708	525
Newspaper	653	486	497
Newspaper	577	480	813
Newspaper	837	652	564
Newspaper	629	585	708
Newspaper	799	538	545
Newspaper	498	581	616
Newspaper	842	797	586

F-Test: Apple Juice JMP



Before entering the *treatment* columns, highlight both of them and select **Full Factorial** from the Macros option. This will create the interaction term for you.

F-Test: Apple Juice JMP

Summary of Fit

RSquare	0.1854
RSquare Adj	0.1099
Root Mean Square Error	96.4171
Mean of Response	615.0000
Observations (or Sum Wgts)	60.0000

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	5	114220.0	22844.0	2.4573
Error	54	501998.0	9296.3	Prob > F
C. Total	59	616218.0		0.0445

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	615.0000	12.4474	49.41	<.0001
Media[Newspaper]	23.0000	12.4474	1.85	0.0701
Advantage[Convenience]	7.0000	17.6033	0.40	0.6925
Advantage[Price]	19.0000	17.6033	1.08	0.2852
Media[Newspaper]*Advantage[Convenience]	43.0000	17.6033	2.44	0.0179
Media[Newspaper]*Advantage[Price]	-33.0000	17.6033	-1.87	0.0663

Output gives general results, but not distinctions between groups.

It seems there is a statistically significant relationship here, though.

Effect Tests: Apple Juice JMP

Choosing **Effect Tests** from the red triangle and Regression Reports will give more details.

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Media	1	1	31740.00	3.4143	0.0701
Advantage	2	2	21720.00	1.1682	0.3187
Media*Advantage	2	2	60760.00	3.2680	0.0457

There seems to be a significant result with the *Interaction* term, and the *Media* factor is close to significance.