Topic : ANOVA and Related Designs

Dr. Lee Fawcett

Friday, September 17th, 2021

Introduction

Introduction

ANOVA is an important technique for interpreting the overall variability within an experiment or other study.

It allows an assessment of the significance of the effects of various 'factors' and permits an estimate to be made of underlying random variability.

The simplest version is **one–way ANOVA**, in which we compare several population means in a single comparative experiment. *This can be regarded as an extension of the two–sample t–test*.

Introduction

The term analysis of variance is often a source of confusion: In spite of its name, ANOVA is concerned with differences between **means** of groups, not differences between variances.

The name analysis of variance comes from the way the procedure uses variances to decide whether the means are different.

Illustrative examples

1. In an experiment to determine the relative effectiveness of four brands of sun cream[1], six volunteers each tried four creams (A, B, C and D) on various parts of their legs.

   They were then subjected to ultra–violet radiation and an estimate of the degree of burning was made (higher figures indicate greater burning).

   The results are given below, along with some summaries:

							Total	mean	s.d.
A	8.6	8.8	7.9	6.4	6.8	8.1	46.6	7.77	0.969
B	5.9	6.2	6.6	6.6	7.4	6.3	39.0	6.50	0.514
C	6.7	9.0	8.4	8.6	6.9	7.2	46.8	7.80	0.982
D	8.2	7.4	8.4	6.9	7.5	7.5	45.9	7.65	0.554

[1] Miksa *et al.* (2016); *International Journal of Cosmetic Science*

Illustrative examples

2. The following measurements were made of the width (in mm) of cuckoo's eggs laid in the nest of various types of birds.

Robin	16.0	16.3	15.9	17.0	17.1	16.0	16.6
	16.9	16.4	16.1	17.0	17.2	16.2	16.9
Hedge sparrow	17.0	16.1	16.9	16.9	15.8	16.7	17.3
	17.5	17.5	16.2	16.5	16.2	17.1	
Wren	15.0	16.0	16.0	15.7	16.2	15.9	15.9
	16.2	16.0	15.5	16.0	15.9	15.5	15.9
Reed Warbler	16.9	17.1	17.0	16.1	16.5	16.5	16.1
	17.9	16.5					
Yellowhammer	16.0	16.0	16.1	17.2	16.5	17.0	17.0
	16.8						

Illustrative examples

Some summaries are given below:

	n_i	T_i	$\bar{x}_i$	s_i
Robin	16	263.2	16.45	0.548
Hedge sparrow	14	234.7	16.76	0.534
Wren	15	237.7	15.85	0.311
Reed Warbler	10	167.1	16.71	0.538
Yellowhammer	9	149.6	16.62	0.482

Illustrative examples

Example **1** is an example of a **planned experiment**, while example **2** is an example of an **observational study**.

An observational study is an investigation in which processes are observed in their natural environment without direct intervention by the investigator. It is difficult to draw authoritative conclusions because extraneous random variation may obscure or even distort the effects of primary interest.

A planned experiment is an investigation in which the sources of extraneous random variation are, as far as possible, identified beforehand and accommodated within the design and analysis of the experiment.

Both **1** and **2** may be analysed using **one–way ANOVA**.

One-way ANOVA

Method outline

To answer questions such as those posed in the previous examples, it is tempting to look at all pairs of groups and perform a two–sample t–test each time (multiple t–tests).

However, though this is sometimes done in disciplines such as biology and psychology, this is not a valid procedure:

- If there are lots of groups to compare, there will be lots of t–tests, and so there is the possibility of a **high number of "false positives"** – such a procedure is "less powerful" at detecting any real differences
- Tests would not be mutually independent

Method outline

The idea behind ANOVA is to obtain two estimates of the underlying variance:

- One based on the variability *between* sample means (known as the **between treatments mean square**)
- The other based on the average variability *within* samples (known as the **residual mean square**)

The first is, essentially, the sample variance of the group means. The second is the average of the group variances, with more weight given to bigger samples.

Method outline

It turns out that if there are no real differences between our groups, then the **variance ratio**

$$VR = \frac{\text{Between treatments mean square}}{\text{Residual mean square}}$$

should be close to one.

If the groups really are different, then this ratio should be bigger than one. The more differences between groups there are, the bigger the variance ratio will get.

Test structure

Suppose we have k populations and hence k samples.

In population i we have random variable $X_i \sim N\left(\mu_i, \sigma_i^2\right)$ and there are n_i observations: $x_{i1}, x_{i2}, \ldots, x_{in_i}$ (and $N = \sum n_i$).

The **assumptions** are as for the two–sample t–test, i.e.

- **Normality**: Histograms, Q-Q plots
- **Independence**: Good experimental design!
- **Homogeneity** (common variance, i.e. $\sigma_i^2 = \sigma^2$ for all i): Check that no standard deviation is more than twice any other

Test structure

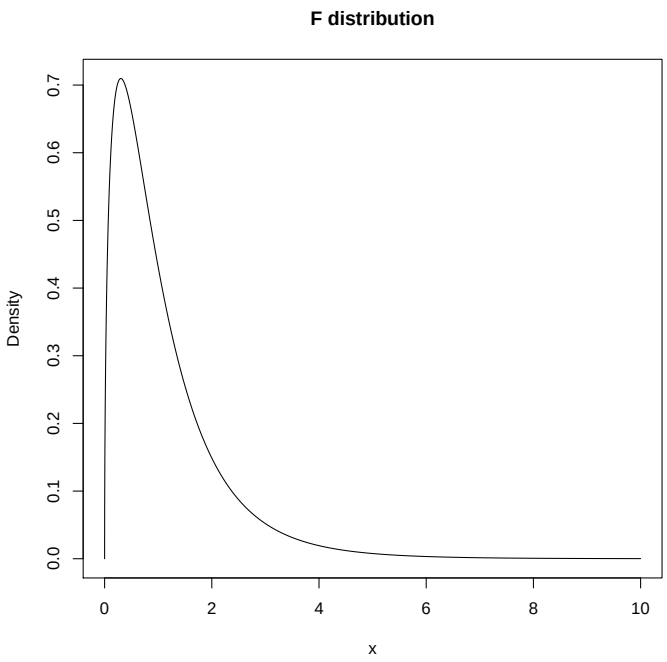
The **null hypothesis** is

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$$

and we assess this in the usual way – via a p-value associated with the test statistic

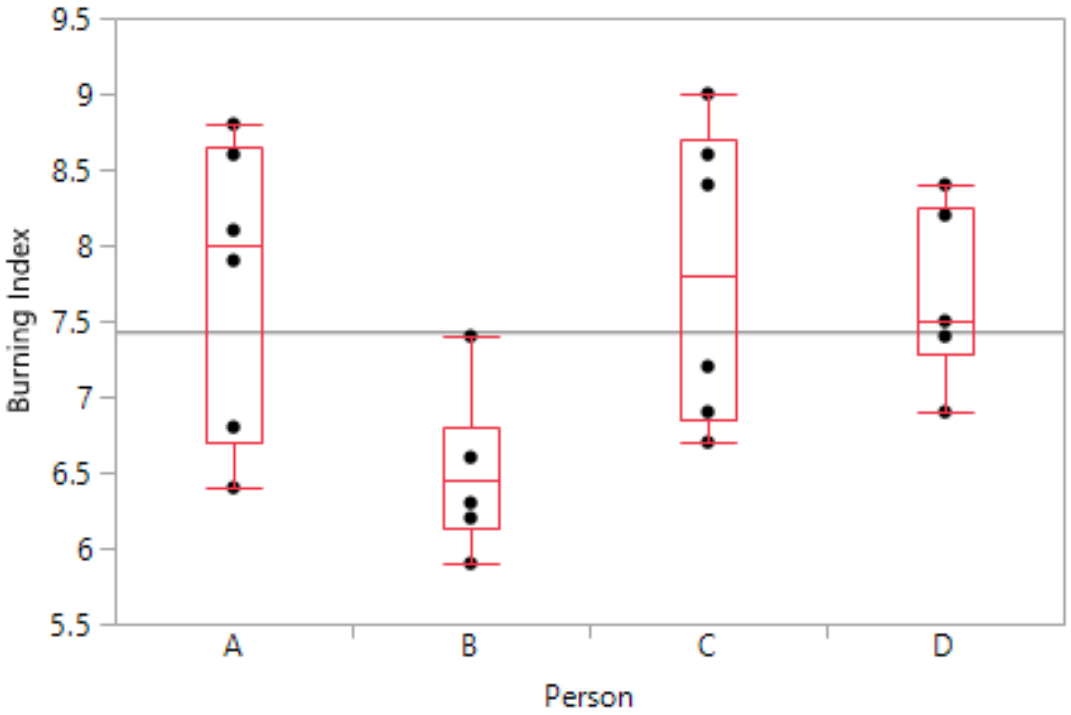
$$VR \sim F_{k-1,N-k},$$

where the F distribution looks like that shown in the plot below.



Sun cream example

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



Sun cream example

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

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Person	3	6.981250	2.32708	3.7630	0.0272*
Error	20	12.368333	0.61842		
C. Total	23	19.349583			

Sun cream example

We compare $VR = 3.763$ to the $F_{3,20}$ distribution:

- Reject H_0 at the 5% level of significance ($p = 2.72\%$)
- There is evidence to suggest the (average) effect of the sun creams is not the same
- But: Where do the differences lie?

One-way ANOVA: Multiple comparisons

If a one-way ANOVA suggests a rejection of H_0 – that all population means are the same – the obvious question to ask is:

"Are all the means different to each other? If not, where do the differences lie?"

- Many follow-up procedures have been devised; the most popular is **Tukey's Highest Significant Difference** method
- Tukey's method is based on the calculation of a 'yardstick', and a comparison of pairwise sample means to this yardstick
- As always, we can let our favourite software take care of all the calculations – we just need to interpret the output

Sun cream example

Means Comparisons

Comparisons for all pairs using Tukey-Kramer HSD

Confidence Quantile

q*	Alpha
2.79894	0.05

HSD Threshold Matrix

Abs(Dif)-HSD

	C	A	D	B
C	-1.2708	-1.2375	-1.1208	0.0292
A	-1.2375	-1.2708	-1.1541	-0.0041
D	-1.1208	-1.1541	-1.2708	-0.1208
B	0.0292	-0.0041	-0.1208	-1.2708

Positive values show pairs of means that are significantly different.

Connecting Letters Report

Level			Mean
C	A		7.8000000
A	A	B	7.7666667
D	A	B	7.6500000
B		B	6.5000000

Levels not connected by same letter are significantly different.

Sun cream example

Conclusions

- We rejected H_0 at the 5% level of significance
- There was evidence to suggest a difference (on average) between the four different sun creams
- In fact, the only significant difference was between creams B and C; no real differences were found between any other pairs
- Anything else we might have done?

Randomised Blocks Designs

Illustrative example

In this example, not only have we got different treatments, but each treatment is split up into **blocks**.

The data below are the results of a **randomised block design** and give the yields (in kilograms) of ten strains of carrots grown in four different plots.

		Plot				Treatment
		1	2	3	4	means
Treatment (strain of carrot)	A	30.0	28.0	28.3	31.7	29.5
	B	33.5	29.6	31.9	29.8	31.2
	C	32.8	29.0	31.2	25.8	29.7
	D	22.2	25.2	24.7	22.3	23.6
	E	30.2	33.1	30.1	27.8	30.3
	F	30.5	30.0	28.5	31.8	30.2
	G	32.6	31.7	30.1	30.8	31.3
	H	30.0	29.7	27.6	29.5	29.2
	I	36.7	35.7	35.6	32.4	35.1
	J	28.5	35.0	32.0	28.1	30.9

Illustrative example

The analysis is exactly the same as the one–way analysis, but this time we have two sets of population means to check: **treatment means** and **block means**.

We want to test for any differences in mean yield between strains of carrot, i.e.

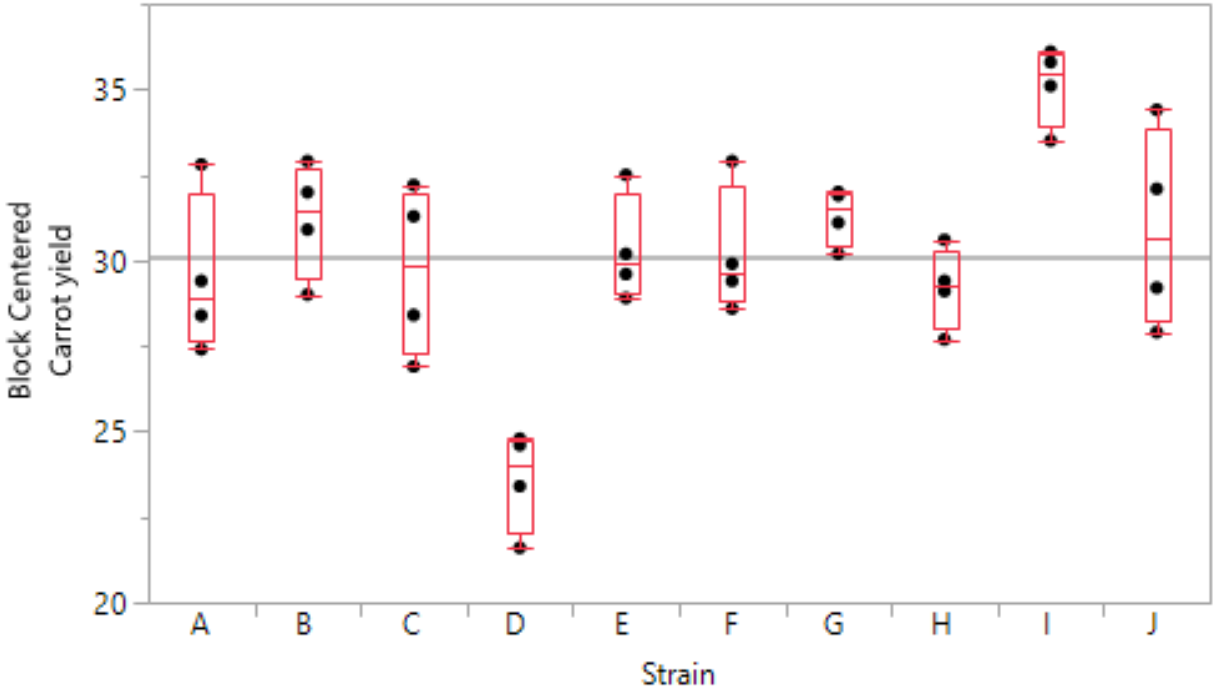
$$H_0 : \mu_A = \mu_B = \cdots = \mu_J.$$

However, we should also check to see if the blocks have had any effect on mean yield, and so we also test

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4.$$

Assumptions are checked in the usual way.

Illustrative example



Illustrative example

Block Means

Plot	Mean	Number
1	30.7000	10
2	30.7000	10
3	30.0100	10
4	29.0000	10

Means and Std Deviations

Level	Number	Mean	Std Dev	Std Err Mean	Lower 95%	Upper 95%
A	4	29.5000	2.34820	1.1741	25.763	33.237
B	4	31.2000	1.67770	0.8389	28.530	33.870
C	4	29.7000	2.47036	1.2352	25.769	33.631
D	4	23.6000	1.46698	0.7335	21.266	25.934
E	4	30.3000	1.56014	0.7801	27.817	32.783
F	4	30.2000	1.88079	0.9404	27.207	33.193
G	4	31.3000	0.84105	0.4205	29.962	32.638
H	4	29.2000	1.19584	0.5979	27.297	31.103
I	4	35.1250	1.15961	0.5798	33.280	36.970
J	4	30.9000	2.91868	1.4593	26.256	35.544

Illustrative example

Some comments so far

- Yields for strain D seem very low, and for strain I quite high, relative to yields for all other strains (which all seem similar)
- Yields for strains D and I seem very different, even after accounting for variability in the data
- There does not appear to be a difference in average yields across the different plots (blocks)
- Most boxplots indicate reasonable symmetry (and perhaps **Normality**) in the yield distributions, perhaps with the exception of strains A, D, G and I – although this is difficult to tell with such small samples
- There could be an issue with assuming equal variances across the 10 strains, with some standard deviations more than twice some others

Illustrative example

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Strain	9	288.68225	32.0758	8.3541	<.0001*
Plot	3	19.38075	6.4602	1.6826	0.1943
Error	27	103.66675	3.8395		
C. Total	39	411.72975			

Illustrative example

The ANOVA table tells us that:

- We **reject** the 'main' null hypothesis at the 0.01% level of significance...
- ... there is **overwhelming evidence** to reject the claim that the average yields are the same across all strains of carrot
- We **retain** the null hypothesis associated with the blocking by plot...
- ... effects of the strain of carrots aside, the plots themselves have had no impact upon the average yield of the carrots

So we must now investigate further, to see where the differences in yields lie!

Illustrative example

Comparisons for all pairs using Tukey-Kramer HSD

HSD Threshold Matrix

Abs(Dif)-HSD

	I	G	B	J	E	F	C	A	H	D
J	-0.5409	-4.3659	-4.4659	-4.7659	-4.1659	-4.0659	-3.5659	-3.3659	-3.0659	2.5341
E	0.0591	-3.7659	-3.8659	-4.1659	-4.7659	-4.6659	-4.1659	-3.9659	-3.6659	1.9341
F	0.1591	-3.6659	-3.7659	-4.0659	-4.6659	-4.7659	-4.2659	-4.0659	-3.7659	1.8341
C	0.6591	-3.1659	-3.2659	-3.5659	-4.1659	-4.2659	-4.7659	-4.5659	-4.2659	1.3341
A	0.8591	-2.9659	-3.0659	-3.3659	-3.9659	-4.0659	-4.5659	-4.7659	-4.4659	1.1341
H	1.1591	-2.6659	-2.7659	-3.0659	-3.6659	-3.7659	-4.2659	-4.4659	-4.7659	0.8341
D	6.7591	2.9341	2.8341	2.5341	1.9341	1.8341	1.3341	1.1341	0.8341	-4.7659

Positive values show pairs of means that are significantly different.

Connecting Letters Report

Level				Mean
I	A			35.125000
G	A	B		31.300000
B	A	B		31.200000
J	A	B		30.900000
E		B		30.300000
F		B		30.200000
C		B		29.700000
A		B		29.500000
H		B		29.200000
D			C	23.600000

Levels not connected by same letter are significantly different.

Illustrative example

Conclusions

- Average yields are significantly different between strain D and every other strain
- Three main groupings have emerged, **within which** average yields *are not* significantly different:
  - Strains (B, G, I, J)
  - Strains (A, B, C, E, F, G, H, J)
  - Strain (D)

Factorial Experiments

Illustrative example

Rats again!

The data in the file `rats.jmp`, which you will examine in our upcoming practical session, show the gains in weight of male rats under six feeding treatments in a **completely randomised experiment**.

The factors were:

(3 levels)	**Source of protein:**	1 = Beef	2 = Cereal	3 = Pork
(2 levels)	**Level of protein:**	1 = High	2 = Low	

The experimenter is particularly interested in comparing mean weight gains among sources of protein and how these comparisons differ between levels of protein.

Illustrative example

Note that there are **six possible source/level combinations**: Beef/High, Beef/Low, Cereal/High, Cereal/Low, Pork/High and Pork/Low.

As you'll see in the practical session, ten rats were randomly allocated to each of these source/level combinations:

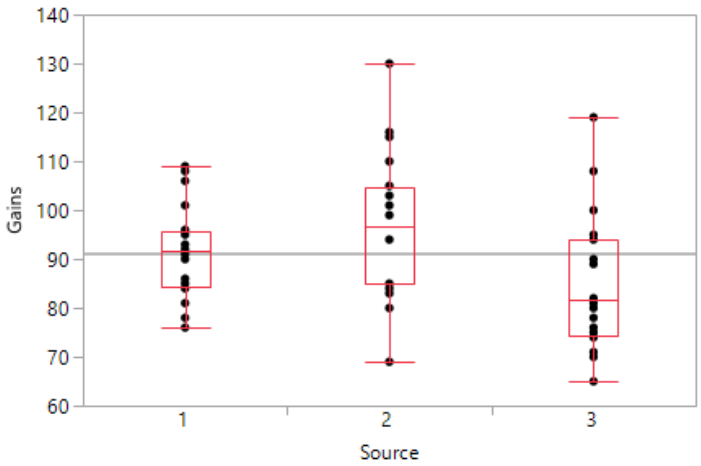
		Source		
		Beef	Cereal	Pork
Level	High	10	10	10
	Low	10	10	10

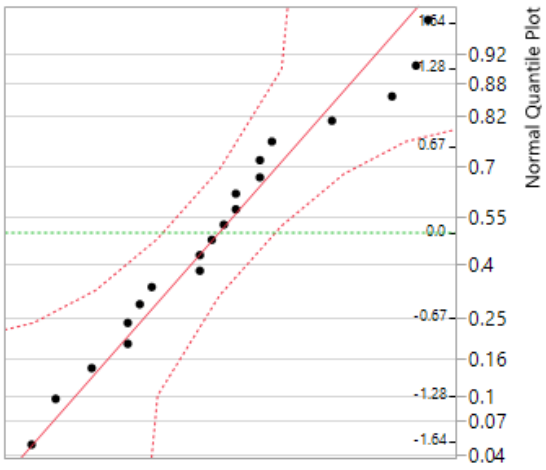
Illustrative example

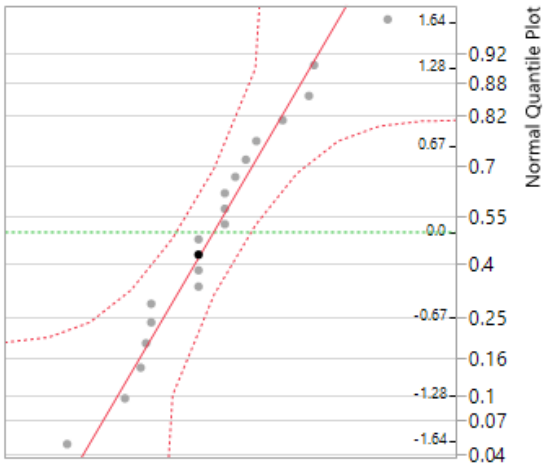
Such a design falls into the realms of **Two-way ANOVA**.

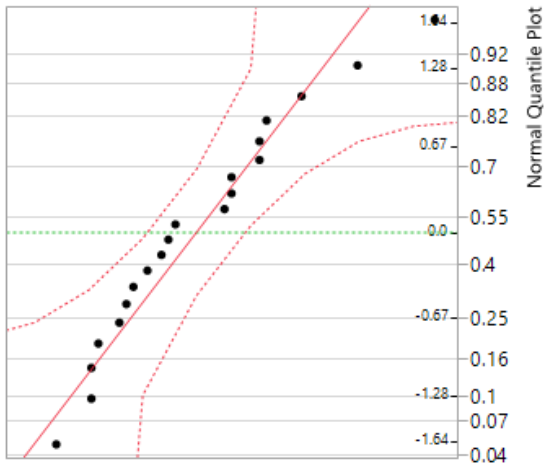
- We have **two factors** that might impact upon the main **response variable**
  - There might be differences in average weight gain depending on the **source** of protein
  - There might also be differences in average weight gain depending on the **level** of protein
- This leads to two null hypotheses:
  - Source: $H_0 : \mu_B = \mu_C = \mu_P$
  - Level: $H_0 : \mu_H = \mu_L$
- We might also want to check for an **interaction**: for example, any dependence of weight gain on source of protein might itself depend on the level of protein given
  - Interaction:
    H_0 : There is no interaction between source and level

Checking Normality









Checking equality of variances

Source	Level	N Rows	Std Dev(Gains)	Mean(Gains)
1	1	10	9.4539821122	93.4
1	2	10	9.3386889396	89.9
2	1	10	10.870960297	89.2
2	2	10	13.199747472	104.7
3	1	10	17.352873601	85.3
3	2	10	10.400854666	84.8

Interpreting computer output

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	5	2685.483	537.097	3.6691
Error	54	7904.700	146.383	Prob > F
C. Total	59	10590.183		0.0063*

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Level	1	1	220.4167	1.5057	0.2251
Source	2	2	1421.7333	4.8562	0.0115*
Level*Source	2	2	1043.3333	3.5637	0.0352*

Conclusions

Source of protein

- $p = 1.15\%$, so we **reject** H_0 at the 5% level of significance
- There is evidence to suggest a difference in mean weight gains between the three sources of protein (Beef/Cereal/Pork)

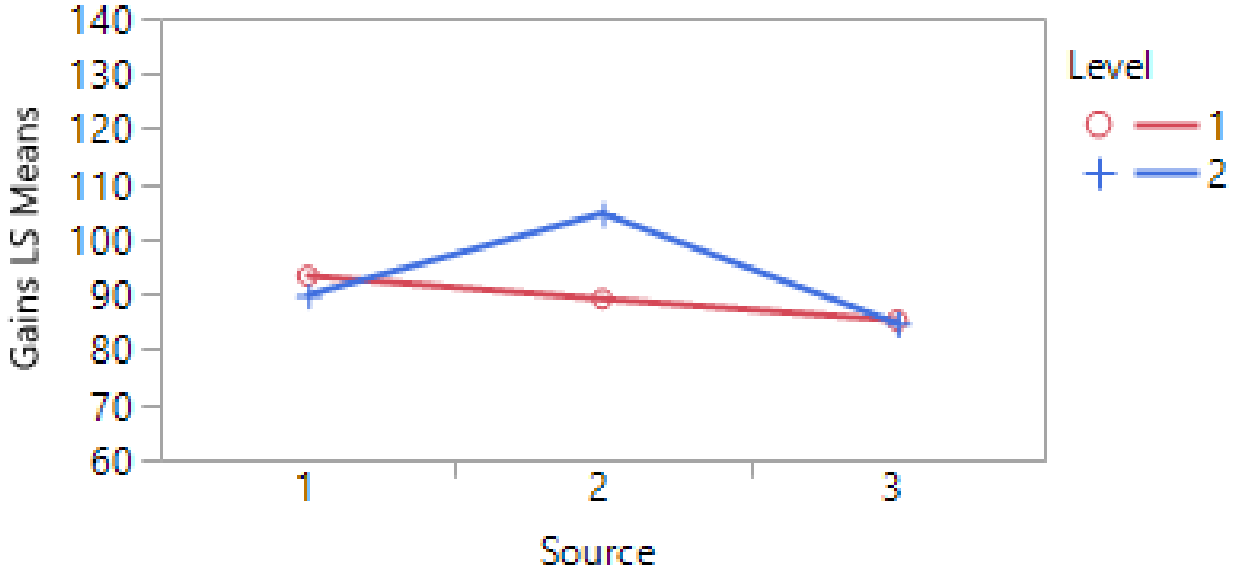
Level of protein

- $p = 22.51\%$, so we **retain** H_0
- There is no evidence to suggest a difference in mean weight gain between the two levels of protein (High/Low)

Interaction

- $p = 3.52\%$, so we **reject** H_0
- There is evidence in the data of an interaction between source and level

Interaction



- Weight gains are similar, irrespective of the level of protein given (High/Low), when the source of protein is meat-based (Beef/Pork)
- Weight gains are considerably higher when the level of protein is High only when the source of protein is non-meat-based (Cereal)