

A/B testing

-or-

Randomized controlled trial

-or-

Completely Randomized  
Designs

Matthew Wheler

In many cases, we have one **factor** with multiple different **treatments** that we control and we want to find out which is better (or worse).

### **We Need To:**

1. Define the tests we are interested in making.
2. Determine the size of our sample using a power analysis.
3. Randomly assign the experimental unit to the treatment.
4. Run the experiment.
5. Analyze the data.

For now, we will assume 2 is done until we revisit this in the future.

## Issues we will deal with for our experimental design:

### 1. Multiple testing

- When we specify a specific  $\alpha$  and do multiple testing we are not actually testing at that  $\alpha$  for **ALL** tests.

### 2. What effect is significant?

- Suppose we are looking at increase in spending and there is a true effect of \$0.01, do we really care?
- How much variability can we expect?

### 3. If we design an experiment, how many “experimental units” do we need to 1 and 2? That is how much do we want to spend to see 2 with “reasonable probability?”

# Example 1

Let's analyze at the full dataset we looked at last class

0 $\mu\text{g/L}$	20 $\mu\text{g/L}$	40 $\mu\text{g/L}$
60	58	40
90	74	58
74	50	25
82	65	30
	68	42

# Our Plan:

## **Contrasts:**

1. Difference between 0  $\mu\text{g/L}$  and 20  $\mu\text{g/L}$
2. Difference between 0  $\mu\text{g/L}$  and 20  $\mu\text{g/L}$
3. Difference between 0  $\mu\text{g/L}$  , 20  $\mu\text{g/L}$  -vs- 40  $\mu\text{g/L}$

## **Overall Test Level:**

$$\alpha = 0.05$$

```
/* data for example 2
   in Lecture 1*/
data dubia_exp;
    input dose age;
    cards;
    0 60
    0 90
    0 74
    0 82
    20 58
    20 74
    20 50
    20 65
    20 68
    40 40
    40 58
    40 25
    40 30
    40 42
;
```

```

/* Using proc glm to answer questions
   for a 1-way anova*/
proc glm data = dubia_exp;
    class dose; /*treatment effects*/
    model age = dose; /*main model*/
    lsmeans dose/cl stderr; /*get model based estimates
                               of the effect
                               -cl specifies we are requesting
                               confidene limts*/

    estimate '0ug -vs- 20ug' dose 1 -1 0; *ANOVA BASED ESTIMATES;
    estimate '0ug -vs- 40ug' dose 1 0 -1;
    estimate '0ug, 20ug -vs- 40ug' dose 0.5 0.5 -1; *must add to 0;
    contrast '0ug -vs- 20ug' dose 1 -1 0; *ANOVA BASED TESTING;
    contrast '0ug -vs- 40ug' dose 1 0 -1;
    contrast '0ug, 20ug -vs- 40ug' dose 0.5 0.5 -1; *must add to 0;

run;
quit;

```

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	3297.857143	1648.928571	12.23	0.0016
Error	11	1483.000000	134.818182		
Corrected Total	13	4780.857143			

Estimates the model variance.

This says there is something going on,  
it doesn't say what. More on this later.



# Model Estimates

dose	age LSMEAN	95% Confidence Limits	
0	76.500000	63.722045	89.277955
20	63.000000	51.571050	74.428950
40	39.000000	27.571050	50.428950

These are essentially the mean estimates, but  
as our models get more complicated they will not be!  
Use lsmeans to get the model based estimate.

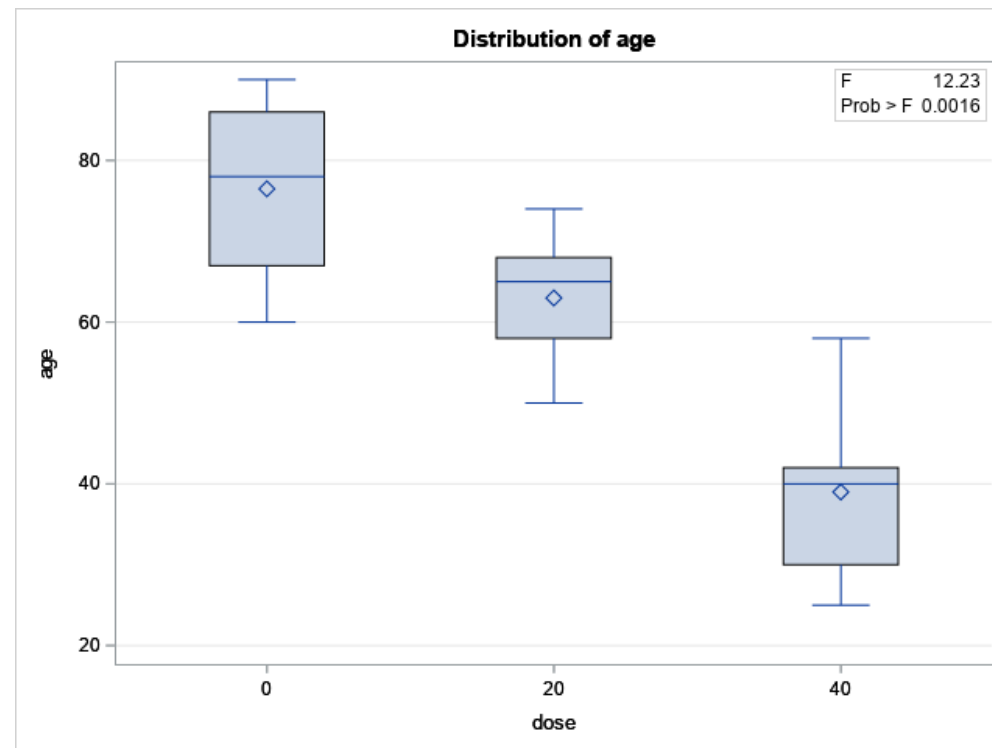
dose	age LSMEAN	Standard Error	Pr >  t
0	76.5000000	5.8055616	<.0001
20	63.0000000	5.1926522	<.0001
40	39.0000000	5.1926522	<.0001

# Model Based Estimated Differences

Parameter	Estimate	Standard Error	t Value	Pr >  t
0ug -vs- 20ug	13.5000000	7.78897823	1.73	0.1110
0ug -vs- 40ug	37.5000000	7.78897823	4.81	0.0005
0ug, 20ug -vs- 40ug	30.7500000	6.49081519	4.74	0.0006

# Contrasts

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
0ug -vs- 20ug	1	405.000000	405.000000	3.00	0.1110
0ug -vs- 40ug	1	3125.000000	3125.000000	23.18	0.0005
0ug, 20ug -vs- 40ug	1	3025.800000	3025.800000	22.44	0.0006



So we can conclude **individually** at  $\alpha = 0.05$ :

- ☐ There is not enough evidence to reject the null that 20µg of copper leads to reduced life spans vs. 0µg of copper.
- ☐ 40µg decreases the live span vs 0µg
- ☐ 40µg decreases the life span vs 0 and 20µg
- ☐ Organic food has problems too. Copper is the number 1 pesticide used for organic food.

**Note:** I said **individually**, that is when we test only one hypothesis the type I error rate is controlled to be our 5% level. When we have multiple tests it is actually higher than that.

# On Multiple Comparisons:

It is important to think about what setting  $\alpha = 0.05$  means.

It means that **BY CHANCE** if nothing is going on I should see a 'significant' result 5% of the time.

**So what happens if I do multiple tests with each test being independent?**

Number of Tests	Probability of Spurious Result
1	0.05
2	0.10
3	0.14
4	0.19
5	0.23
1	0.05
6	0.26
7	0.30
8	0.34
9	0.37
10	0.40
20	0.64
30	0.79
40	0.87
50	0.92

That is if **nothing is going** on and I do a “bunch of tests” after a while I am almost guaranteed to report a significant finding even when **nothing is happening!**

- Most ‘science’ (especially medical and psychology) does this – it is why I never ever trust a report about ‘health science’ news.
- It is why you have to **plan** your analyses before hand!
- It is also why you have to **adjust** for multiple comparisons!

# Adjusting for multiple comparisons

**First Way Bonferroni adjustment:**

**Benefits:** Can be used in any setting we deal with in this class

**Negatives:** If you are doing a lot of comparisons it kills power. Thus you need a ton of resources to conclude anything : use with caution, or if sampling is cheap, or if your boss doesn't care about money.



# Bonferroni Adjustment

**Idea:** We assume that each test is independent and we have a fixed Type I error rate  $\alpha$ , and that we are doing  $k$  tests. We want to control the probability that all tests combined have a type I error rate of **at least**  $\alpha$ . This can be done by making each individual test at  $\alpha^* = \alpha/k$ .

For **Example 1** this means that our individual tests are at an  $\alpha^* = 0.05/3$ , which is 0.017

# Revisiting this problem with Bonferroni adjustment.

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
0ug -vs- 20ug	1	405.000000	405.000000	3.00	0.1110
0ug -vs- 40ug	1	3125.000000	3125.000000	23.18	0.0005
0ug, 20ug -vs- 40ug	1	3025.800000	3025.800000	22.44	0.0006

These tests are significant at values less than 0.017, which implies that we can reject the null at our given Type I error rate and still be “safe.”

# Adjusting for multiple Comparisons

**Second Way Tukey-Kramer adjustment:**

**Benefits:** Very useful for finding differences between all possible comparisons between means.

**Negatives:** Can only be used in this situation, you can't compare multiple groups, and when you have data/analyses that are not normal you can't use it.

# Tukey-Kramer Adjustment

**Idea:** If there is nothing going on (i.e. under the null) we can assume that the differences between means are essentially normal random variables with mean zero and some variance. Then the distance between the maximum mean and the minimum mean, divided by the standard deviation, has some distribution. We call this the “**Studentized range distribution**,” which can be thought of as just another distribution like the T-distribution that can be used for testing purposes. This distribution, by definition, controls for multiple testing.

## Doing this in SAS

```
/* Using proc glm to answer questions
   for a 1-way anova*/
proc glm data = dubia_exp;
    class dose; /*treatment effects*/
    model age = dose; /*main model*/
    lsmeans dose/cl stderr adjust=tukey; /*get model based estimates
                                         of the effect
                                         -cl specifies we are requesting
                                         confidene limts
                                         -tukey adjusts for multiple
comparisons*/
run;
quit;
```

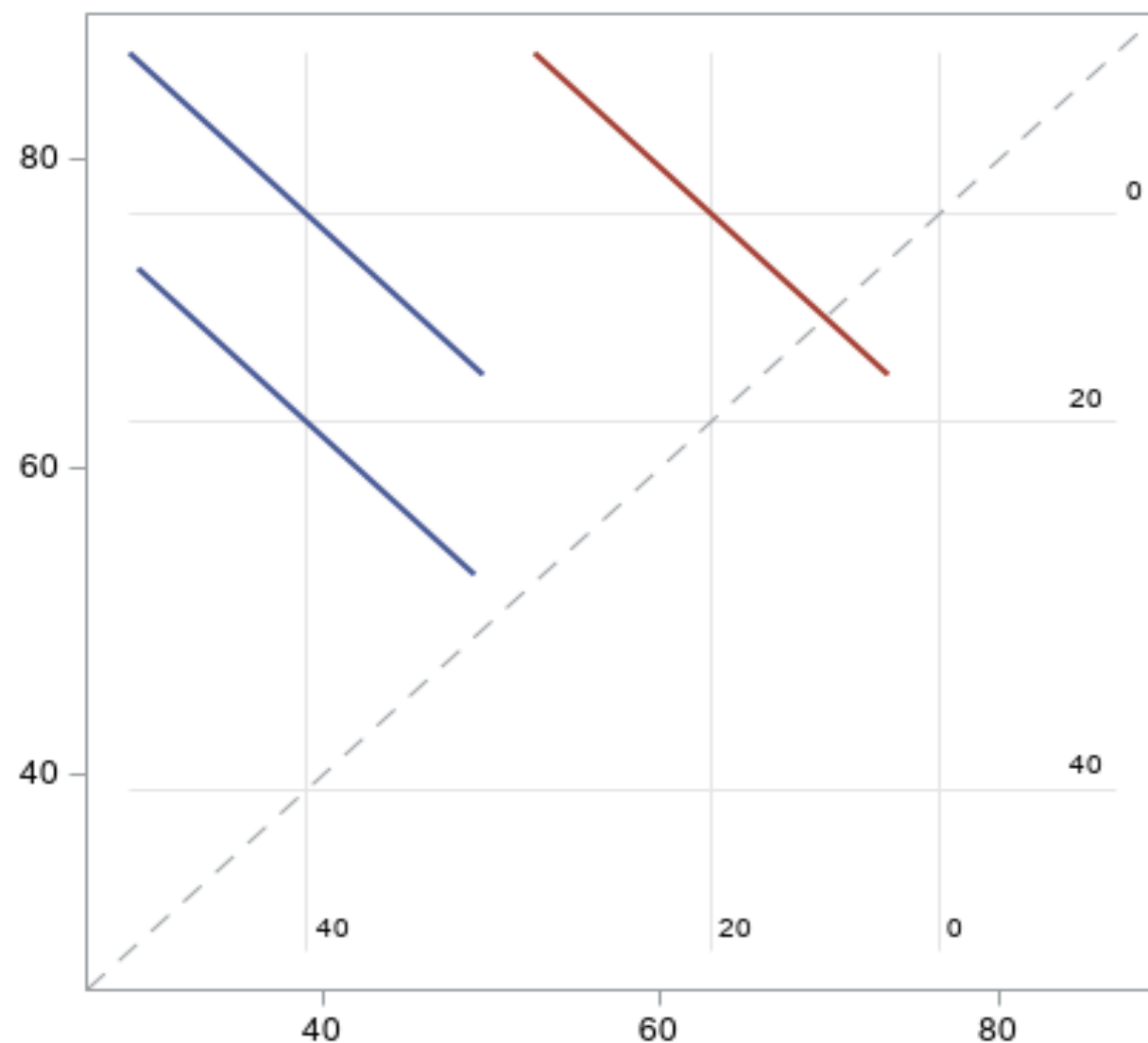
# New output

Least Squares Means for effect dose  
Pr > |t| for H0: LSMean(i)=LSMean(j)  
Dependent Variable: age

i/j	1	2	3
1		0.2368	0.0014
2	0.2368		0.0189
3	0.0014	0.0189	

Least Squares Means for Effect group				
i	j	Difference Between Means	Simultaneous 95% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	-9.750000	-25.937197	6.437197
1	3	-36.125000	-52.312197	-19.937803
2	3	-26.375000	-42.562197	-10.187803

age Comparisons for dose



Differences for alpha=0.05 (Tukey-Kramer Adjustment)

— Not significant — Significant

With the Tukey-Kramer adjustment, we get the same result as with the Bonferonni adjustment, but this isn't always the case.



# Adjusting for multiple comparisons

**Third Way: Control using an ANOVA (not really an adjustment):**

**Benefits:** The ANOVA controls the overall error for a factor, so you are saying something is going on with that factor:

**Negatives:** It isn't really controlling the error for individual tests between treatments in that factor. Use as a last resort.

# ANOVA

**Idea:** The ANOVA test is an overall test for a given factor. It says some treatment in the factor is different, but it doesn't say which one. Use this to narrow down the options, and report the most 'significant' tests individually. You are assured through your ANOVA that something is happening with the specified Type I error rate, you just don't know what.

# ANOVA Table



Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	3297.857143	1648.928571	12.23	0.0016
Error	11	1483.000000	134.818182		
Corrected Total	13	4780.857143			

**Something is going on.**

## Example 2



A marketing firm is considering two different incentives to increase a grocery store's frequent shopper's weekly purchases. The first is a mail marketing campaign and the second is a campaign that incentivizes purchases through a new smartphone app. We are interested in finding out if either of these marketing approaches will increase spending.

# Tests

Before we analyze the data, we must define what we are interested in:

We have 3 treatments:

1. Doing nothing.
2. Mail marketing.
3. App Marketing.

What are the comparisons we are interested in making?

**Test 1:** Doing nothing –vs- mail marketing

**Test 2:** Doing nothing –vs- new app

**Test 3:** Doing nothing or mail marketing -vs- new app

The first two tests are similar to what you might do with a **t-test**.

The third test compares the first two treatments to the newly **developed app**.

Each one is called a **contrast**.

```

/*data marketing
   Example 2 data*/
data marketing;
    input group spent @@;
    cards ;
    1 100 1 110 1 92
    1 122 1 118 1 98
    1 130 1 110 2 115
    2 121 2 110 2 130
    2 142 2 108 2 112
    2 120 3 125 3 140
    3 153 3 142 3 130
    3 162 3 157 3 160
;

/*test the marketing data
   using an ANOVA based glm model*/
proc glm data = marketing;
    class group; * types of groups;
    model spent = group; *model;
    lsmeans group/cl stderr adjust=tukey; *multiple mean comparisons;
    contrast 'Nothing -vs- Mail' group 1 -1 0; *ANOVA BASED TESTING;
    contrast 'Nothing -vs- App ' group 1 0 -1; *For contrasts;
    contrast 'Nothing, Mail -vs- App' group 0.5 0.5 -1;

run;

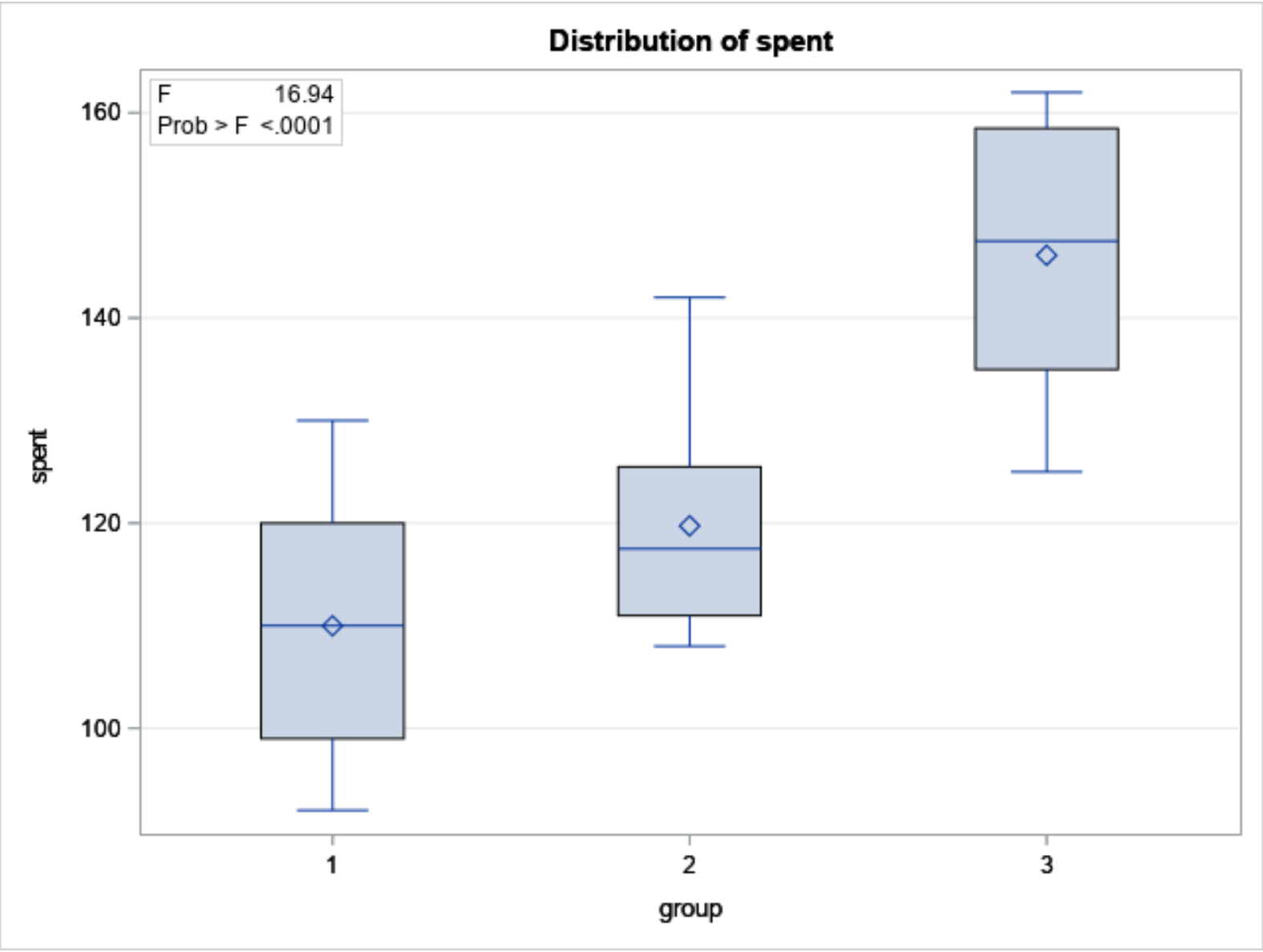
```

# ANOVA

Something is going on:

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	5588.583333	2794.291667	16.94	<.0001
Error	21	3464.375000	164.970238		
Corrected Total	23	9052.958333			





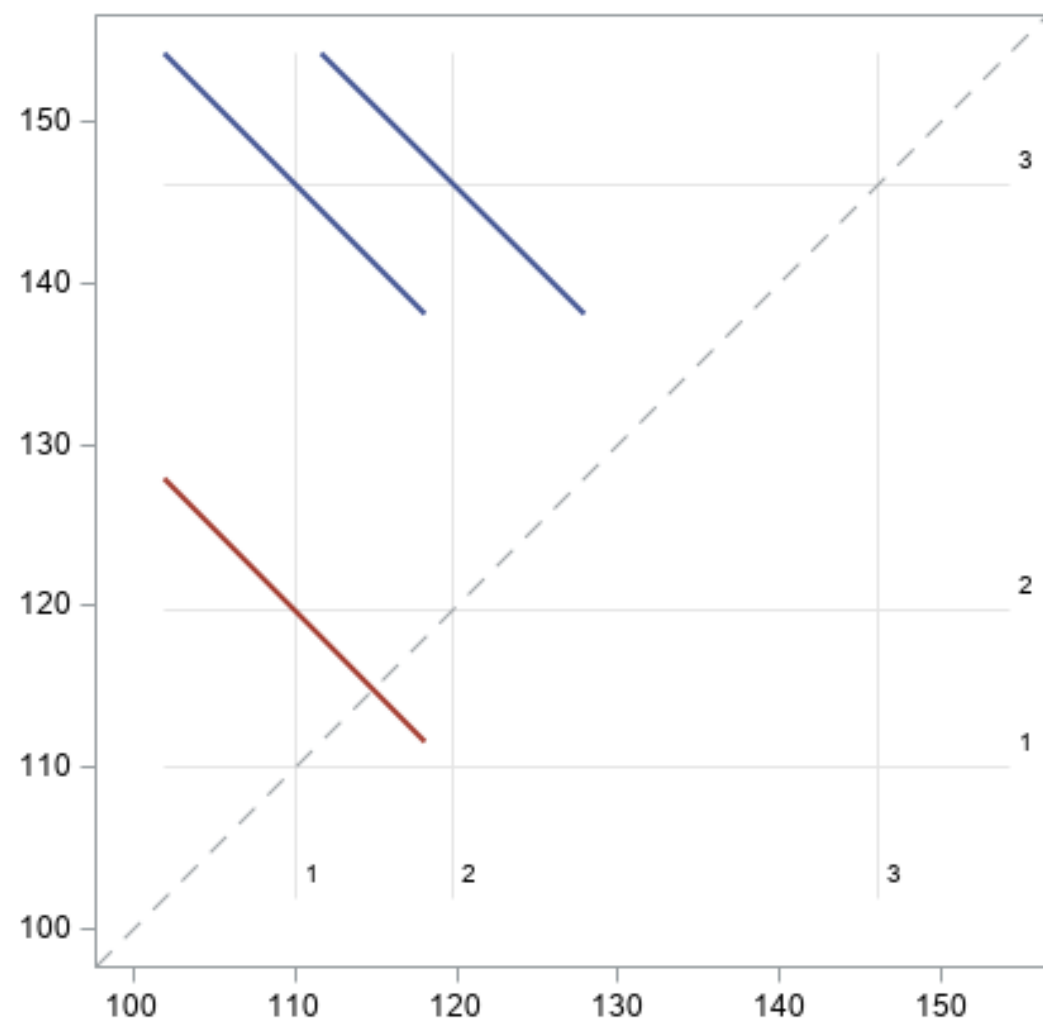
# Tukey's Difference

group	spent LSMEAN	Standard Error	Pr >  t	LSMEAN Number
1	110.000000	4.541066	<.0001	1
2	119.750000	4.541066	<.0001	2
3	146.125000	4.541066	<.0001	3

Least Squares Means for effect group  
Pr > |t| for H0: LSMean(i)=LSMean(j)  
Dependent Variable: spent

i/j	1	2	3
1		0.3029	<.0001
2	0.3029		0.0014
3	<.0001	0.0014	

spent Comparisons for group



Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
Nothing -vs- Mail	1	380.250000	380.250000	2.30	0.1439
Nothing -vs- App	1	5220.062500	5220.062500	31.64	<.0001
Nothing, Mail -vs- App	1	5208.333333	5208.333333	31.57	<.0001

**Bonferroni test is ok**

# The contrast statement

The contrast statement is designed to compare means across groups:

```
contrast 'Nothing -vs- Mail' group 1 -1 0; *ANOVA BASED TESTING;  
contrast 'Nothing -vs- App ' group 1 0 -1; *For contrasts;  
contrast 'Nothing, Mail -vs- App' group 0.5 0.5 -1;
```

First we must recognize that there are 3 group means we are interested in:

- 1) Money spent doing nothing  $\mu_1$
- 2) Money spent with direct mail marketing  $\mu_2$
- 3) Money spent with new mobile app  $\mu_3$

Let's break down the first question:

What is the difference between a customer's spending habits when I do not directly market to them ( $\mu_1$ ) in comparison to mail marketing ( $\mu_2$ )?

As an equation, this is:  $1\mu_1 + (-1)\mu_2$

The contrast statement says:

```
contrast 'Nothing -vs- Mail' group (1) (-1) 0;
```


$$1\mu_1 + (-1)\mu_2 + (0)\mu_3$$

I am literally subtracting the mean from group 1 to the mean of group 2

For a contrast to be correct, my numbers must sum to **zero**.

Now let's think about the following statement:

Is there a difference between a customer's spending habits when I do not directly market to them ( $\mu_1$ ) or I directly mail marketing ( $\mu_2$ ) in comparison to them using the new app ( $\mu_3$ )?

One might be tempted to do this:  $(1)\mu_1 + (1)\mu_2 + (-1)\mu_1$

However, it is wrong. We are not accounting for the fact I am comparing one group to two groups. Instead:

$$(1)\mu_1 + (1)\mu_2 + (-2)\mu_1$$

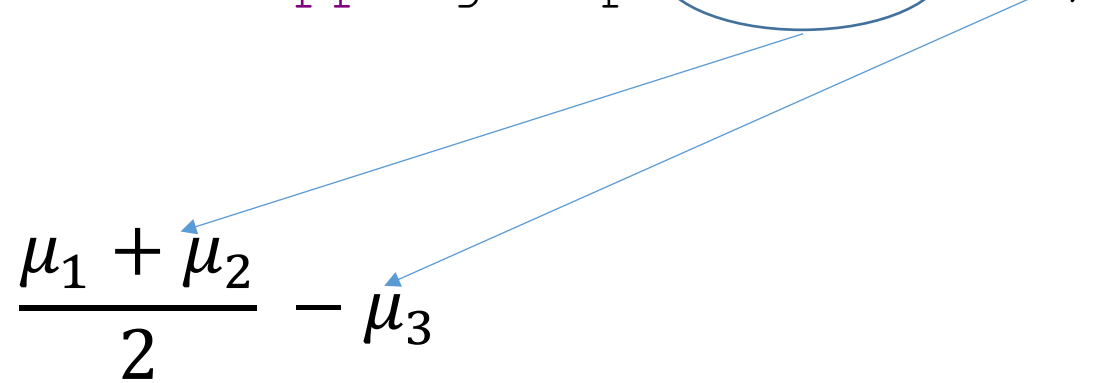
$$(0.5)\mu_1 + (0.5)\mu_2 + (-1)\mu_1$$

Both are correct, and will give the same answer.



I prefer the second statement, because it can be interpreted as the difference in spending of the 'Average' of the first two marketing campaigns when compared to the mobile app marketing campaign.

contrast 'Nothing, Mail -vs- App' group **0.5 0.5 -1**;

$$\frac{\mu_1 + \mu_2}{2} - \mu_3$$


## **Things we are missing still with A/B testing:**

1. What happens when we have 'extra variables' that we can not randomize.
2. How is SAS coding things is called a design matrix. Unfortunately different PROCs do different things with the design matrix. We have to disentangle things so we can go forward and no what we are doing.