

Analysis of Variance



So far: Z-test and the t -test have been used to compare the mean of a sample to that of a population, or to compare the means of two samples

ANOVA

HINT!!

Module 10

Do you remember testing the difference between two variances?

How do you proceed to test the difference when you have more than two samples?

The technique we used on module 10, the F test, can also be used to compare three or more means, this is ANOVA: Analysis of variance.

ANOVA is used to compare the means from three or more samples.

It is more powerful than performing multiple t tests, which increase possibility of Type I error.

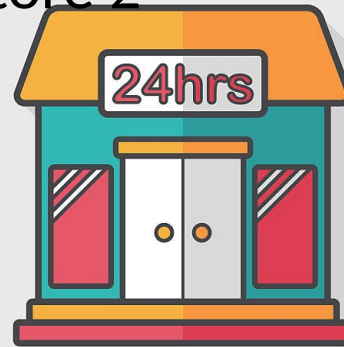
Example

We want to compare the average daily sales of three stores.

Store 1



Store 2



Store 3



Store 1



Store 2



Daily sales:

\$9,800

\$8,000

\$8,200

\$9,400

\$7,400

\$9,900

\$8,400

\$8,700

\$9,200

\$8,300

\$7,400

\$7,600

\$8,800

\$6,800

\$8,500

\$7,800

\$7,300

\$8,400

Mean daily sales:

\$8,778**\$7,878**

If you want to compare
the mean of two stores:

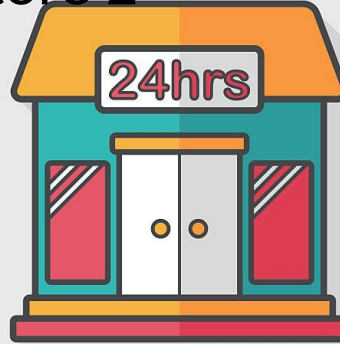
If $n < 30 = t$ test

If $n > 30 = z$ test

Store 1



Store 2



Store 3



Daily sales:

\$9,800

\$8,000

\$8,200

\$9,400

\$7,400

\$9,900

\$8,400

\$8,700

\$9,200

\$8,300

\$7,400

\$7,600

\$8,800

\$6,800

\$8,500

\$7,800

\$7,300

\$8,400

\$10,500

\$9,900

\$9,800

\$10,700

\$8,800

\$10,700

\$9,900

\$9,100

\$10,800

Mean daily sales:

\$8,778**\$7,878****\$10,022**

Not when you
have three or
more groups.

Using individual
test to compare:
1 versus 2
1 versus 3
2 versus 3
It is wrong!

Variance within groups

Variance between groups

Observe the variance on group 2.

In the first case, you might conclude that the differences are caused by variances *between* groups.

In the second case, by variances *within* the groups.

	GROUP 1	GROUP 2
	\$9,800	\$10,500
	\$8,000	\$9,900
	\$8,200	\$9,800
	\$9,400	\$10,700
	\$7,400	\$8,800
	\$9,900	\$10,100
	\$8,400	\$9,900
	\$8,700	\$9,100
	\$9,200	\$10,800
Mean	\$8,778	\$9,956
Variance	\$731,944	\$460,278
SD	\$856	\$678

	GROUP 1	GROUP 2
	\$9,800	\$12,500
	\$8,000	\$9,900
	\$8,200	\$9,800
	\$9,400	\$13,700
	\$7,400	\$8,800
	\$9,900	\$10,100
	\$8,400	\$9,900
	\$8,700	\$9,100
	\$9,200	\$10,800
Mean	\$8,778	\$10,511
Variance	\$731,944	\$2,568,611
SD	\$856	\$1,603

F test

The F test in this case is the ratio of variances between groups over variances within groups.
The larger this ratio is, the more likely that the groups have different means.
In this case, H_0 is rejected.

$$F = \frac{\text{Variances between groups}}{\text{Variances within groups}}$$

Therefore, you will know if differences exists, but you do not know where.

Is it between 1 and 2? 2 and 3? 1 and 3?

One-way versus two-way ANOVA

ANOVA is a simple test that uses the sample variance to compare the means across categorical variables with two or more categories.

One-way ANOVA test: It compares 3 or more means. It can be used to test the null hypothesis that all the means are equal because it has one categorical variable, store type.

Two-way ANOVA test: Involves two variables. An example would be if we further broke down the mean monthly store sales data by region, across four regions, etc.

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

H_1 : At least one mean is different from the others.

Assumptions for the F Test for Comparing Three or More Means

1. The populations from which the samples were obtained must be normally or approximately normally distributed.
2. The samples must be independent of one another.
3. The variances of the populations must be equal.

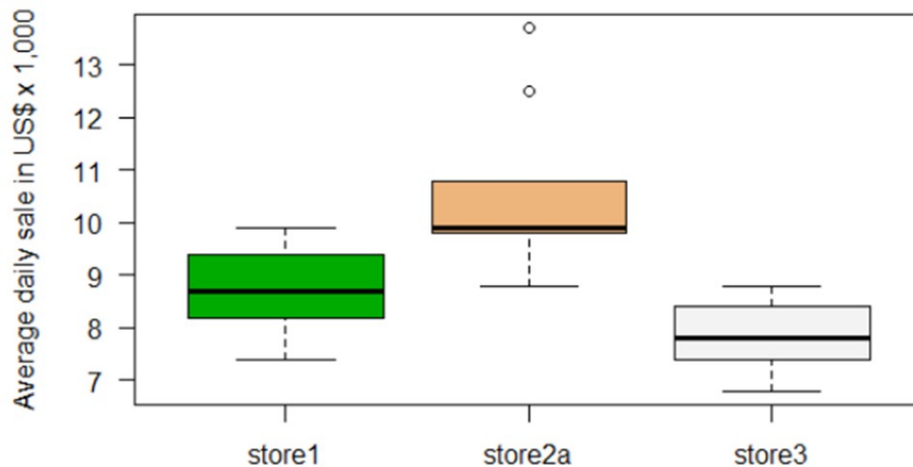
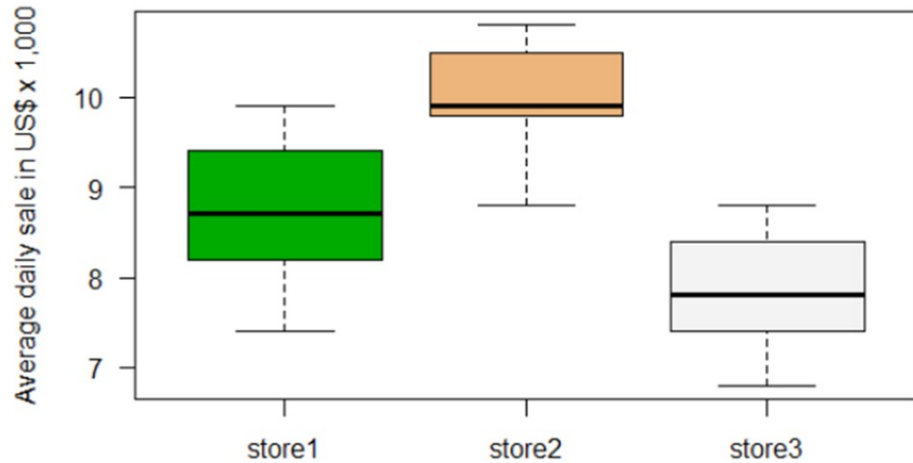
$$H_0: \mu_1 = \mu_2 = \cdots = \mu_k$$

H_1 : At least one mean is different from the others.

- It uses F test.
- Between-group variance: Variance of the means
- Within-group variance: variance of all data, not affected by means
- If no difference between the means, the between-group variance estimate will be approximately equal to the within-group variance estimate, and the F test value will be approximately equal to 1.
- When the means differ significantly, the between-group variance will be much larger than the within-group variance; the F test value will be significantly greater than 1
- H_0 is rejected if $F > CV$



Applying R



Compare the means among three stores.

```
store1 = c(9800, 8000, 8200, 9400, 7400, 9900, 8400, 8700, 9200)
store2 = c(10500, 9900, 9800, 10700, 8800, 10100, 9900, 9100, 10800)
store3 = c(8300, 7400, 7600, 8800, 6800, 8500, 7800, 7300, 8400)
store2a = c(12500, 9900, 9800, 13700, 8800, 10100, 9900, 9100, 10800)
```

```
storegroupA = stack(data.frame(cbind(store1, store2, store3)))
storegroupB = stack(data.frame(cbind(store1, store2a, store3)))
```

```
summary(aov(values~ind, data = storegroupA))
```

```
##              Df    Sum Sq Mean Sq F value    Pr(>F)
## ind              2 19542963  9771481    17.99 1.68e-05 ***
## Residuals      24 13033333   543056
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(aov(values~ind, data = storegroupB))
```

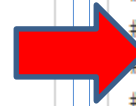
```
##              Df    Sum Sq Mean Sq F value    Pr(>F)
## ind              2 32246667 16123333    12.94 0.000154 ***
## Residuals      24 29900000   1245833
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

	Variance
Store 1	\$731,944
Store 2	\$460,278
Store 3	\$436,944
Store 2a	\$2,568,611

To run ANOVA in R

```
store1 = c(9800, 8000, 8200, 9400, 7400, 9900, 8400, 8700, 9200)
store2 = c(10500, 9900, 9800, 10700, 8800, 10100, 9900, 9100, 10800)
store3 = c(8300, 7400, 7600, 8800, 6800, 8500, 7800, 7300, 8400)
data.frame(cbind(store1, store2, store3))
```

##	store1	store2	store3
## 1	9800	10500	8300
## 2	8000	9900	7400
## 3	8200	9800	7600
## 4	9400	10700	8800
## 5	7400	8800	6800
## 6	9900	10100	8500
## 7	8400	9900	7800
## 8	8700	9100	7300
## 9	9200	10800	8400



```
store1 = c(9800, 8000, 8200, 9400, 7400, 9900, 8400, 8700, 9200)
store2 = c(10500, 9900, 9800, 10700, 8800, 10100, 9900, 9100, 10800)
store3 = c(8300, 7400, 7600, 8800, 6800, 8500, 7800, 7300, 8400)
stack(data.frame(cbind(store1, store2, store3)))
```

##	values	ind
## 1	9800	store1
## 2	8000	store1
## 3	8200	store1
## 4	9400	store1
## 5	7400	store1
## 6	9900	store1
## 7	8400	store1
## 8	8700	store1
## 9	9200	store1
## 10	10500	store2
## 11	9900	store2
## 12	9800	store2
## 13	10700	store2
## 14	8800	store2
## 15	10100	store2
## 16	9900	store2
## 17	9100	store2
## 18	10800	store2
## 19	8300	store3
## 20	7400	store3
## 21	7600	store3
## 22	8800	store3
## 23	6800	store3
## 24	8500	store3
## 25	7800	store3
## 26	7300	store3
## 27	8400	store3

Data must be converted to a table using code `data.frame()` and then stacked using code `stack()`.

Example

$H_0: \mu_1 = \mu_2 = \mu_3$ (claim)

H_1 : At least one mean is different from the others.

A researcher wishes to try three different techniques to lower the blood pressure of individuals diagnosed with high blood pressure.

The subjects are randomly assigned to three groups; the first group takes medication, the second group exercises, and the third group follows a special diet.

After four weeks, the reduction in each person's blood pressure is recorded.

At $\alpha = 0.05$, test the claim that there is no difference among the means.

	Medication	Exercise	Diet
	10	6	5
	12	8	9
	9	3	12
	15	0	8
	13	2	4
Mean:	11.8	3.8	7.6
Variance:	5.7	10.2	10.3

Example

	Medication	Exercise	Diet
	10	6	5
	12	8	9
	9	3	12
	15	0	8
	13	2	4
Mean:	11.8	3.8	7.6
Variance:	5.7	10.2	10.3

$$\text{d.f.N.} = k - 1$$

$$\text{d.f.D.} = N - k$$

k	3
N	15
d.f.N.	2
d.f.D.	12

Critical value.

```
k_BP = 3
n_BP = 15
dfN = k_BP-1
dfD = n_BP-k_BP
alpha = 0.05

# F Critical value (answer = 3.89)
cv_BP = qf(alpha, dfN, dfD, lower.tail = FALSE)

# Confirm CV by reversing to alpha (answer = 0.05)
alpha2 = pf(cv_BP, dfN, dfD, lower.tail = FALSE)
```

Critical value = 3.89

Confirm CV by reversing to alpha value = 0.05

Calculate the Grand Mean

$$\bar{X}_{GM} = \frac{\sum X}{N}$$

$$= \text{SUM}(C4:E8)/15 = 7.73$$

$$= 7.73$$

Calculate the between-group variance

$$s_B^2 = \frac{\sum n_i (\bar{X}_i - \bar{X}_{GM})^2}{k - 1}$$

$$= \frac{5*(11.8-7.73)^2 + 5*(3.8-7.73)^2 + 5*(7.6-7.73)^2}{3-1}$$

$$= 80.07$$

Calculate the within-group variance

$$s_W^2 = \frac{\sum (n_i - 1) s_i^2}{\sum (n_i - 1)}$$

$$= \frac{(5-1)*5.7 + (5-1)*10.2 + (5-1)*10.3}{(5-1) + (5-1) + (5-1)}$$

$$= 8.73$$

Calculate the F value

$$F = \frac{\text{Variances between groups}}{\text{Variances within groups}}$$

$$F = \frac{s_B^2}{s_W^2} = \frac{80.07}{8.73} = 9.17$$


```

Medication = c(10, 12, 9, 15, 13)
Exercise = c(6, 8, 3, 0, 2)
diet = c(5, 9, 12, 8, 4)
k_BP = 3
n_BP = 15
n_med = 5
n_exer = 5
n_diet = 5

# Calculate the grand Mean

gm_BP = sum(Medication, Exercise, diet)/n_BP

# Calculate the between-group variance

mean_med = mean(Medication)
mean_exc = mean(Exercise)
mean_diet = mean(diet)
n_med = 5
n_exer = 5
n_diet = 5

between_groups_top = (n_med*(mean_med-gm_BP)^2+n_exer*(mean_exc-gm_BP)^2+n_diet*(mean_diet-gm_BP)^2)
between_groups_bottom = (k_BP-1)
between_groups = between_groups_top/between_groups_bottom

# Calculate the within-group variance

var_med = var(Medication)
var_exc = var(Exercise)
var_diet = var(diet)

within_groups_top = (n_med-1)*var_med + (n_exer-1)*var_exc + (n_diet-1)*var_diet
within_groups_bottom = (n_med-1) + (n_exer-1) + (n_diet-1)
within_groups = within_groups_top/within_groups_bottom

Ftest_BP = between_groups/within_groups

```

```

# Calculate the grand Mean

gm_BP = sum(Medication, Exercise, diet)/n_BP

# Calculate the between-group variance

mean_med = mean(Medication)
mean_exc = mean(Exercise)
mean_diet = mean(diet)
n_med = 5
n_exer = 5
n_diet = 5

between_groups_top = (n_med*(mean_med-gm_BP)^2+n_exer*(mean_exc-gm_BP)^2+n_diet*(mean_diet-gm_BP)^2)
between_groups_bottom = (k_BP-1)
between_groups = between_groups_top/between_groups_bottom

```

$$\bar{X}_{\text{GM}} = \frac{\sum X}{N} \quad s_B^2 = \frac{\sum n_i (\bar{X}_i - \bar{X}_{\text{GM}})^2}{k - 1}$$

```
# Calculate the within-group variance

var_med = var(Medication)
var_exc = var(Exercise)
var_diet = var(diet)

within_groups_top = (n_med-1)*var_med + (n_exer-1)*var_exc + (n_diet-1)*var_diet
within_groups_bottom = (n_med-1) + (n_exer-1) + (n_diet-1)
within_groups = within_groups_top/within_groups_bottom

Ftest_BP = between_groups/within_groups
```

$$s_W^2 = \frac{\sum (n_i - 1) s_i^2}{\sum (n_i - 1)}$$

Notice:

In the previous slides, I showed you how I performed all my calculations in one single r chunk, and I make sure that all equations contain a name. This way I prevent them from displaying on my report unless I purposely call them.

In this case I call them in the main text of the report by using inline r codes, next slide.

Inline r codes

```
152  
153 Medication: Mean = ``r round(mean_med, 2)`` , Variance = ``r round(var_med, 2)``  
154 Exercise: Mean = ``r round(mean_exc, 2)`` , Variance = ``r round(var_exc, 2)``  
155 Diet: Mean = ``r round(mean_diet, 2)`` , Variance = ``r round(var_diet, 2)``  
156  
157 Grand mean = ``r round(gm_BP, 2)``  
158 Between-group variance = ``r round(between_groups, 2)``  
159 Within-group variance = ``r round(within_groups, 2)``  
160  
161 <I>F</I> test = ``r round(Ftest_BP, 2)``  
162 Critical value at  $\alpha = 0.05$  = ``r round(cv_BP, 2)``  
163  
164 Is F higher than critical value? = ``r Ftest_BP>cv_BP`` |
```

HTML outcome

Medication: Mean = 11.8 , Variance = 5.7
Exercise: Mean = 3.8 , Variance = 10.2
Diet: Mean = 7.6 , Variance = 10.3

Grand mean = 7.73
Between-group variance = 80.07
Within-group variance = 8.73

F test = 9.17
Critical value at $\alpha = 0.05$ = 3.89

Is F higher than critical value? = TRUE

Using R aov() code

```
Medication = c(10, 12, 9, 15, 13)
Exercise = c(6, 8, 3, 0, 2)
diet = c(5, 9, 12, 8, 4)

BP_data = data.frame(cbind(Medication, Exercise, diet))

BP_stacking = utils::stack(BP_data)

# stack produces a data frame with two columns:
# values: the result of concatenating the selected vectors in x
# ind: a factor indicating from which vector in x the observation originated

BP_anova = aov(values~ind, data = BP_stacking)

BP_summary = summary(BP_anova)

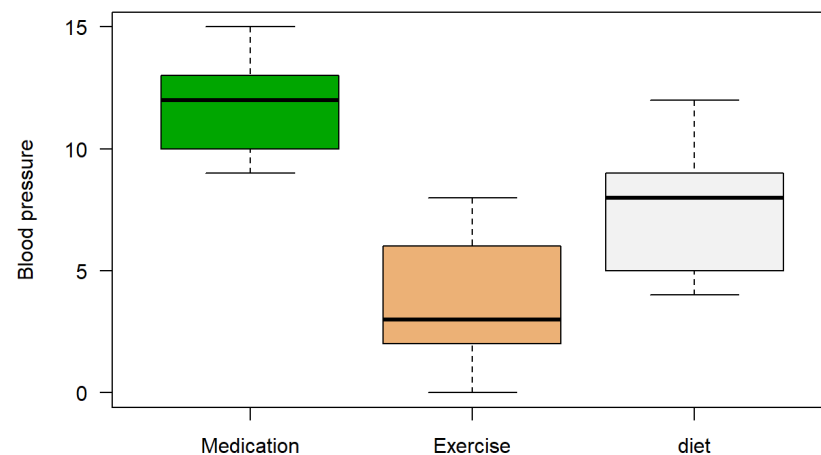
boxplot(BP_data,
        col = terrain.colors(3),
        las=1,
        ylab = "Blood pressure")
```

BP_anova

```
## Call:
## aov(formula = values ~ ind, data = BP_stacking)
##
## Terms:
##               ind Residuals
## Sum of Squares 160.1333   104.8000
## Deg. of Freedom      2       12
##
## Residual standard error: 2.955221
## Estimated effects may be unbalanced
```

BP_summary

```
##           Df Sum Sq Mean Sq F value Pr(>F)
## ind         2  160.1    80.07   9.168 0.00383 **
## Residuals   12  104.8     8.73
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



```
BP_summary
```

```
##           Df Sum Sq Mean Sq F value    Pr(>F)    **
## ind         2   160.1    80.07    9.168 0.00383
## Residuals   12   104.8     8.73
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

F test versus Critical values

```
# F Critical value
cv_05 = qf(0.05, 2, 12, lower.tail = FALSE)
cv_01 = qf(0.01, 2, 12, lower.tail = FALSE)
cv_001 = qf(0.001, 2, 12, lower.tail = FALSE)
```

$\alpha = 0.05$, CV = 3.89

$\alpha = 0.01$, CV = 6.93

$\alpha = 0.001$, CV = 12.97

$\alpha = 0.05$: $F > CV = \text{TRUE}$

$\alpha = 0.01$: $F > CV = \text{TRUE}$

$\alpha = 0.001$: $F > CV = \text{FALSE}$

P value versus α

```
# p value calculation
Ftest_BP = between_groups/within_groups
dfN = k_BP-1
dfD = n_BP-k_BP
pvalue_BP = pf(Ftest_BP, dfN, dfD, lower.tail = FALSE)
```

P value = 0.004

P value < $\alpha = 0.05 = \text{TRUE}$

P value < $\alpha = 0.01 = \text{TRUE}$

P value < $\alpha = 0.001 = \text{FALSE}$

Inline r codes

```
182
183  $\alpha$  = 0.05, CV = ``r round(cv_05, 2)``
184  $\alpha$  = 0.01, CV = ``r round(cv_01, 2)``
185  $\alpha$  = 0.001, CV = ``r round(cv_001, 2)``
186
187  $\alpha$  = 0.05: F > CV = ``r Ftest_BP>cv_05``
188  $\alpha$  = 0.01: F > CV = ``r Ftest_BP>cv_01``
189  $\alpha$  = 0.001: F > CV = ``r Ftest_BP>cv_001``
```

Inline r codes

```
227 <B>P value versus  $\alpha$ </B>
228
229 ``{r}
230 # p value calculation
231 Ftest_BP = between_groups/within_groups
232 dfN = k_BP-1
233 dfD = n_BP-k_BP
234 pvalue_BP = pf(Ftest_BP, dfN, dfD, lower.tail = FALSE)
235
236
237 P value = ``r round(pvalue_BP, 3)``
238
239 P value <  $\alpha$  = 0.05 = ``r pvalue_BP<0.05``
240 P value <  $\alpha$  = 0.01 = ``r pvalue_BP<0.01``
241 P value <  $\alpha$  = 0.001 = ``r pvalue_BP<0.001``
242
```


A significant test value means that there is a high probability that this difference in means is not due to chance (H_0 is rejected), but it does not indicate where the difference lies.

Tukey and Scheffé Tests

Tukey Test (q)

Used after the analysis of variance has been completed
Make pairwise comparisons between means when the groups have the same sample size.

If $q > CV$, differences between the two means is significant.

$$q = \frac{\bar{X}_i - \bar{X}_j}{\sqrt{s_W^2/n}}$$

Scheffé Test (F_s)

Used after the analysis of variance has been completed
Compares the means two at a time, using all possible combinations.

\bar{X}_1 versus \bar{X}_2 \bar{X}_1 versus \bar{X}_3 \bar{X}_2 versus \bar{X}_3

$$F_s = \frac{(\bar{X}_i - \bar{X}_j)^2}{s_W^2[(1/n_i) + (1/n_j)]}$$

Tukey test (q)

```
TukeyHSD(BP_anova)
```

```
##    Tukey multiple comparisons of means
##      95% family-wise confidence level
##
## Fit: aov(formula = values ~ ind, data = BP_stacking)
##
## $ind
##              diff          lwr          upr          p adj
## Exercise-Medication -8.0 -12.98636 -3.0136398 0.0028351
## diet-Medication      -4.2  -9.18636  0.7863602 0.1031329
## diet-Exercise         3.8  -1.18636  8.7863602 0.1465881
```

Only one has shows
a statistically
difference between
the means at $p < 0.05$

diff : is the difference in means between the two groups

lwr : is the lower estimate of the 95% confidence interval of the difference in means

upr : is the upper estimate of the same 95% confidence interval

p adj : is the significance of the test after correcting for family-wise error rate

The adjustment of the p-value is necessary in controlling for Type 1 error inflation.

Scheffé Test (F_s)

```
# Package DescTools  
library(DescTools)  
ScheffeTest(BP_anova)
```

```
##  
## Posthoc multiple comparisons of means: Scheffe Test  
## 95% family-wise confidence level  
##  
## $ind  
##  
##      diff      lwr.ci      upr.ci    pval  
## Exercise-Medication -8.0 -13.210111 -2.789889 0.0038 **  
## diet-Medication     -4.2  -9.410111  1.010111 0.1216  
## diet-Exercise       3.8  -1.410111  9.010111 0.1693  
##  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Again, only one has shows a statistically difference between the means at $p < 0.05$