Section 12.2 - Comparing the Means

Statistics 104

Autumn 2004

Comparing the Means

The F test in an ANOVA analysis only answers the question whether all of the means are the same or not. It doesn't tell which groups or treatments are different.

Often there will be a set of comparisons that are of interest.

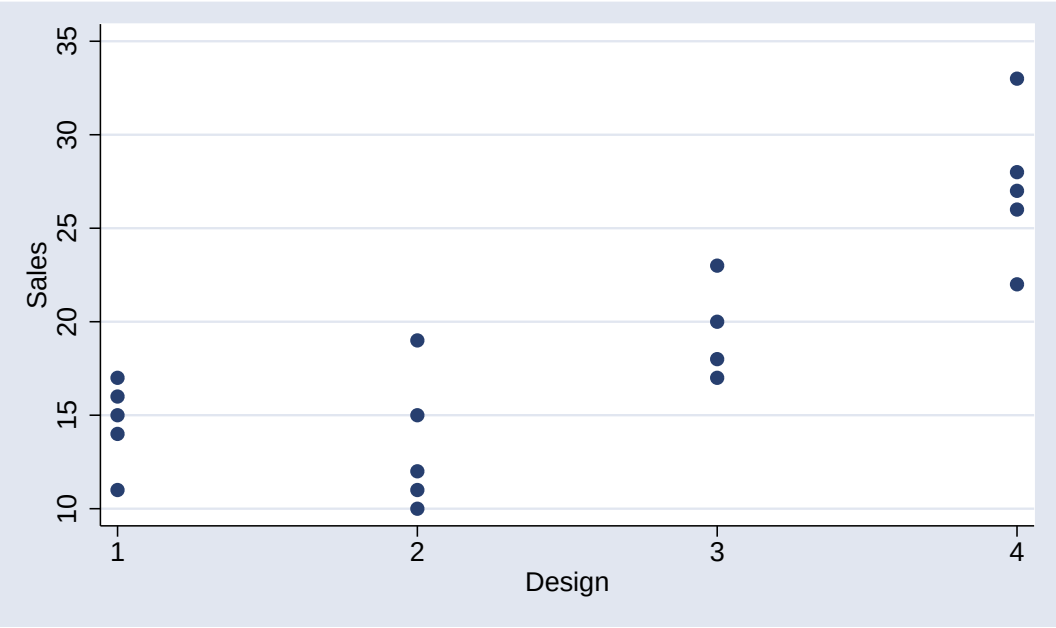
Want to focus on these comparisons (assuming they exist).

These comparisons should be developed before looking at the data. They describe the research questions of interest.

Example: Kenton Food Company

The Kenton Food Company was interested in the effect of 4 package designs for a new breakfast cereal. They picked 20 similar stores with respect to location and sales volumes as the experimental units. Each package was randomly allocated to 5 stores. The response of interest, y, was the number of cases of cereal sold in each of the stores.

Package Design	Characteristics
1	3 Colour with Cartoons
2	3 Colour with No Cartoons
3	5 Colour with Cartoons
4	5 Colour with No Cartoons



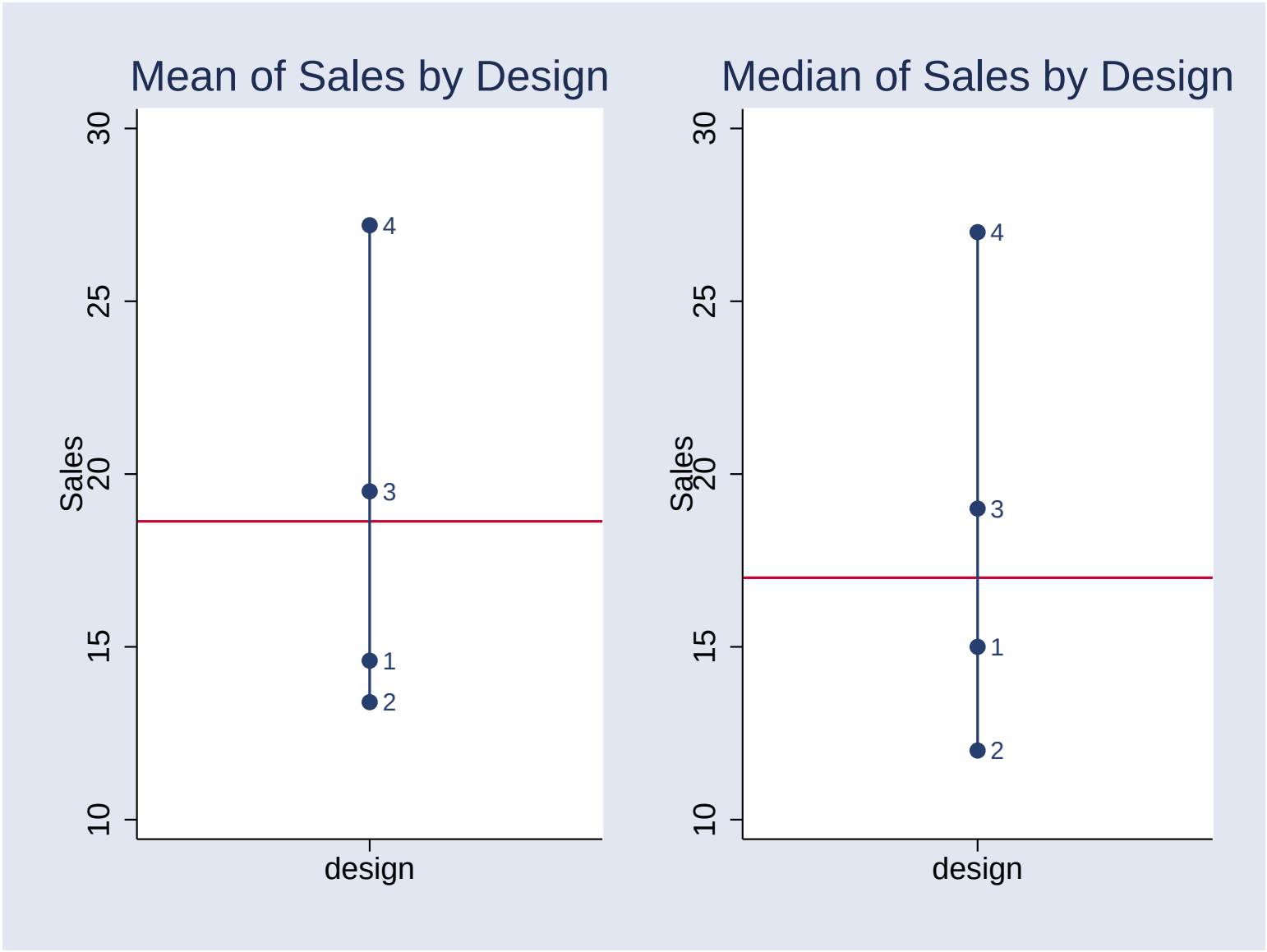
Note that one observation for package 3 is missing due to a fire during the study period.

```
. oneway sales design, tabulate

           |          Summary of Sales
    Design |        Mean   Std. Dev.       Freq.
-----------+------------------------------------
         1 |        14.6   2.3021729           5
         2 |        13.4   3.6469165           5
         3 |        19.5   2.6457513           4
         4 |        27.2   3.9623226           5
-----------+------------------------------------
     Total |   18.631579   6.4395525          19

                       Analysis of Variance
    Source              SS         df      MS             F     Prob > F
------------------------------------------------------------------------
Between groups      588.221053      3   196.073684     18.59     0.0000
 Within groups           158.2     15   10.5466667
------------------------------------------------------------------------
    Total           746.421053     18   41.4678363

Bartlett's test for equal variances:  chi2(3) =   1.3144  Prob>chi2 = 0.726
```


Mean of Sales by Design
Sales
10
15
20
25
30
4
3
1
2
design
Median of Sales by Design
Sales
10
15
20
25
30
4
3
1
2
design

Possible comparisons of interest

- Cartoon vs No Cartoon for 3 Colour designs (μ_1 vs μ_2)
- 3 Colour vs 5 Colour for Cartoon designs (μ_1 vs μ_3)
- 3 Colour average vs 5 Colour average $\left(\frac{\mu_1+\mu_2}{2} \text{ vs } \frac{\mu_3+\mu_4}{2}\right)$
- Cartoon average vs Non-cartoon average $\left(\frac{\mu_1+\mu_3}{2} \text{ vs } \frac{\mu_2+\mu_4}{2}\right)$

These comparisons can be examined with contrasts.

Contrasts

A contrast is a linear combination of the population means of the form

$$\psi = \sum a_i \mu_i$$

such that $\sum a_i = 0$

The previous comparisons are described by the following contrasts:

- Cartoon vs No Cartoon for 3 Colour designs (μ_1 vs μ_2)

$$\mu_1 - \mu_2 \qquad a = (1, -1, 0, 0)$$

- 3 Colour vs 5 Colour for Cartoon designs (μ_1 vs μ_3)

$$\mu_1 - \mu_3 \qquad a = (1, 0, -1, 0)$$

- 3 Colour average vs 5 Colour average $\left(\frac{\mu_1+\mu_2}{2} \text{ vs } \frac{\mu_3+\mu_4}{2}\right)$

$$\frac{\mu_1+\mu_2}{2} - \frac{\mu_3+\mu_4}{2} \qquad a = \left(\frac{1}{2}, \frac{1}{2}, \frac{-1}{2}, \frac{-1}{2}\right)$$

- Cartoon average vs Non-cartoon average $\left(\frac{\mu_1+\mu_3}{2} \text{ vs } \frac{\mu_2+\mu_4}{2}\right)$

$$\frac{\mu_1+\mu_3}{2} - \frac{\mu_2+\mu_4}{2} \qquad a = \left(\frac{1}{2}, \frac{-1}{2}, \frac{1}{2}, \frac{-1}{2}\right)$$

- Another possible contrast is

$$\mu_1 - \frac{\mu_2+\mu_3+\mu_4}{3} \qquad a = \left(1, \frac{-1}{3}, \frac{-1}{3}, \frac{-1}{3}\right)$$

Note that contrasts for a particular comparison aren't unique. For example for the comparison

- Cartoon average vs Non-cartoon average $\left(\frac{\mu_1+\mu_3}{2} \text{ vs } \frac{\mu_2+\mu_4}{2}\right)$ the following contrasts are all equally good

$$\frac{\mu_1+\mu_3}{2} - \frac{\mu_2+\mu_4}{2} \qquad a_1 = \left(\frac{1}{2}, \frac{-1}{2}, \frac{1}{2}, \frac{-1}{2}\right)$$

$$(\mu_1 + \mu_3) - (\mu_2 + \mu_4) \qquad a_2 = (1, -1, 1, -1)$$

$$(\mu_2 + \mu_4) - (\mu_1 + \mu_3) \qquad a_3 = (-1, 1, -1, 1)$$

You just need to be careful when interpreting them.

The sample contrast (e.g. the estimate) is

$$c = \sum a_i \bar{y}_i$$

The standard error of this estimate is

$$SE_c = s_p \sqrt{\sum \frac{a_i^2}{n_i}}$$

For example, for the 3 Colour average vs 5 Colour average $\left(\frac{\mu_1+\mu_2}{2} - \frac{\mu_3+\mu_4}{2}\right)$

$$\begin{aligned} c &= \frac{14.6 + 13.4}{2} + \frac{19.5 + 27.2}{2} \\ &= 14.0 - 23.35 = -9.35 \end{aligned}$$

$$\begin{aligned} SE_c &= 3.248\sqrt{\frac{0.5^2}{5}+\frac{0.5^2}{5}+\frac{(-0.5)^2}{4}+\frac{(-0.5)^2}{5}} \\ &= 3.248\sqrt{0.2125} = 1.497 \end{aligned}$$

For the others, the sample contrasts and SE's are

Contrast	Estimate	SE
$\mu_1 - \mu_2$	1.20	2.054
$\mu_1 - \mu_3$	-4.90	2.179
$\frac{\mu_1+\mu_3}{2} - \frac{\mu_2+\mu_4}{2}$	-3.25	1.497
$\mu_1 - \frac{\mu_2+\mu_3+\mu_4}{3}$	-5.43	1.694

A confidence interval for a contrast ψ is given by

$$c \pm t^* SE_c$$

where t^* is based on the error $df = N - I$.

For example, a 95% confidence interval for the expected difference between 3 and 5 Colour sales for Cartoon designs $(\mu_1 - \mu_3)$ is

$$\begin{aligned} &-4.90 \pm 2.131 \times 2.179 \\ &\quad = -4.90 \pm 4.64 = (-9.54, -0.26) \end{aligned}$$

This interval suggests that the sales for the 5 Colour Cartoon version are better than the 3 Colour Cartoon version on average (0 is not in the interval).

Note that this confidence procedure matches with the pooled two-sample t procedure for $\mu_1 - \mu_2$.

Hypothesis Tests on Contrasts

Also of interest may be tests based on the null hypothesis, $H_0 : \psi = 0$.

A test statistic for examining this is

$$t = \frac{c}{SE_c}$$

The alternatives for this test can either be one- or two-sided.

p-values and critical values are based on the $t(N - I)$ distribution.

For example, to examine the difference between the average cartoon effect vs the average non-cartoon effect

$$t = \frac{-3.25}{1.497} = 2.171; \qquad p\text{−value} = 0.0464$$

It appears that sales are better with the non-cartoon version of the packaging.

Note that some stat packages (including Stata) will perform tests on contrasts using F tests instead of t tests. These tests are equivalent as

$$t^2 = F$$

If the alternative hypothesis is two-sided, the p-values will be the same. However, if the alternative hypothesis is one-sided, the p-value reported for the F test needs to be divided by 2 (assuming that the estimated contrast is in the direction of the alternative).

```
. matrix cartoon = (0 , 1, -1, 1, -1)
. test,test(cartoon)

 ( 1)  design[1] - design[2] + design[3] - design[4] = 0

       F(  1,    15) =    4.71
            Prob > F =    0.0464
```

Prespecified Contrasts vs Data Snooping

- The confidence intervals and test described earlier are only valid for prespecified constrasts.
- They are not valid when you take a look at the data and decide from the data which contrasts to look at.
- Contrasts that are picked by looking at the data tend to lead to p-values that are too small and confidence intervals that suggests effects that are larger than they really are.
- Its OK to look, but you need to be careful in doing your inference in these cases.

Multiple Comparisons

One approach for making the necessary adjustments when data snooping.

These procedures are usually only used when the F test for the ANOVA is declared significant (and some are only valid if this holds).

The focus will be on pairwise comparisons, but these approaches can be used for different types of contrasts as well.

Pairwise comparisons: $\mu_i - \mu_j$

If there are k different groups in the study, there are $\binom{k}{2} = \frac{k(k-1)}{2}$ different possible pairwise comparisons.

When performing these comparisons, we want to focus on setting an error rate α (or confidence level C) that holds for all of them jointly.

For example, with the fabric flammability data set there are 10 different comparisons ($k = 5$ labs). For example we want a procedure so that the confidence level is 95% that the true difference in means is contained in all 10 intervals.

The t procedures discussed earlier have the property for each interval we are 95% confident that the true difference is in each interval separately.

If we use these intervals, our confidence level could be as low as 50% that all of the intervals contain the truth.

Want to base inference on

$$t_{ij} = \frac{\bar{y}_i - \bar{y}_j}{s_p\sqrt{\frac{1}{n_i} + \frac{1}{n_j}}}$$

(Like the pooled two-sample t test).

However we need to compare this to a difference distribution than a t so to get desired error rates for all comparisons.

The distribution needed for comparisons depends on the set of comparison being done. However, in all cases, two groups will be declared significantly different if

$$|t_{ij}| \geq t^{**}$$

or equivalently if

$$|\bar{y}_i - \bar{y}_j| \geq t^{**} s_p \sqrt{\frac{1}{n_i} + \frac{1}{n_j}} = MSD$$

(MSD = Minimum Significant Difference) where t^{**} depends on the error rate and the multiple comparison approach taken.

One approach for choosing t^{**} is the Bonferroni procedure.

It is based on the relationship

$$P[A_1 \text{ or } A_2 \text{ or } \ldots \text{ or } A_l] \leq P[A_1] + \ldots + P[A_l]$$

Suppose that the are l comparisons of interest. Then t^{**} comes from a $t(N-I)$ distribution with an upper tail probability of $\frac{\alpha}{2l}$. With this choice of t^{**}, the chance that any of the groups that have the same mean in reality are declared significantly different is at most α .

For 10 comparisons and an overall error rate of $\alpha = 0.05$, the critical value needs to be based on an upper tail area of 0.0025.

For the fabric testing example, the error df = 50 , giving t^{**} = 2.937 and

$$MSD = 2.937 \times 0.4058\sqrt{\frac{1}{11}+\frac{1}{11}} = 0.508$$

Note that when using multiple comparison procedures to examine all pairwise comparisons, you want to use the two-sided alternative for each comparison. Also MSD depends on the sample size so it won't be a constant unless all the sample sizes are the same, as they are for this example.

Stata will perform this Bonferroni procedure as part of the `oneway` command. However it reports things based on p-values. In this case, declare two groups statistically significantly different if

$$p\text{−value} \leq \alpha$$

The lower number in each cell is the p-value and the upper value is the estimated difference in means.

```
. oneway charred lab, bonferroni tabulate

                    Comparison of charred by lab
                            (Bonferroni)
Row Mean-|
Col Mean |          1          2          3          4
---------+--------------------------------------------
       2 |    .263636
         |      1.000
         |
       3 |   -.036364         -.3
         |      1.000      0.891
         |
       4 |   -.336364         -.6         -.3
         |      0.575      0.011      0.891
         |
       5 |    .309091    .045455    .345455    .645454
         |      0.801      1.000      0.513      0.005
```

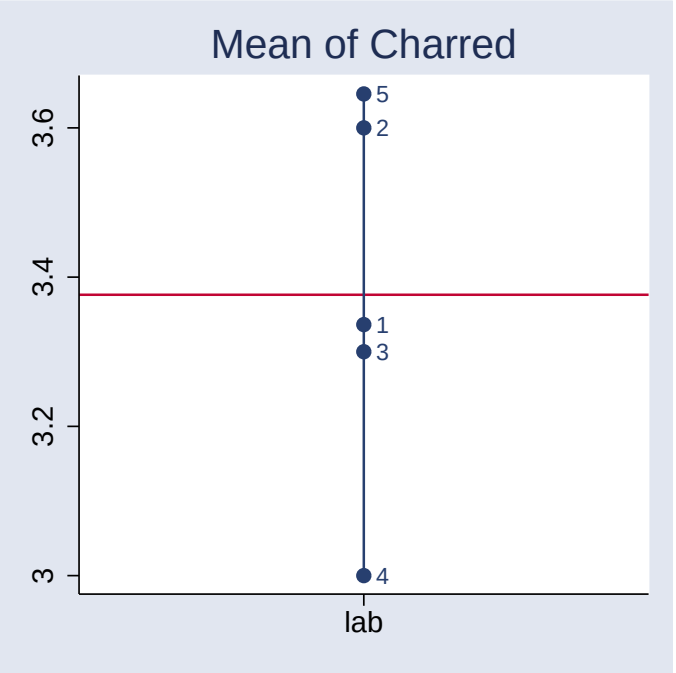
So for the fabric testing data set, with $\alpha = 0.05$, labs 2 & 4, and labs 5 & 4 are declared significantly different.

Note that for this example we do not want to, for example, to declare labs 1 & 4 or labs 1 & 5 different.

This might seem a bit illogical.

If labs 4 & 5 are really different, then either labs 1 & 4 or labs 1 & 5 (or both) are different.

However with this data, we don't have enough information about where the additional differences are.



Similarly, it is possible to calculate confidence intervals such that the confidence level that all intervals contain the truth is $C = 100(1-\alpha)\%$.

The intervals for $\mu_i - \mu_j$ are of the form

$$(\bar{y}_i - \bar{y}_j) \pm t^{**} s_p \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

where t^{**} is the same as for the multiple comparisons test.

You can also write this interval as

$$(\bar{y}_i - \bar{y}_j) \pm MSD$$

So for $\mu_2 - \mu_1$, the interval is

$$0.2636 \pm 0.5081 = (-0.2445, 0.7717)$$

For the fabric testing data set, the set of Bonferroni intervals are (as calculated in Minitab)

```
Bonferroni 95.0% Simultaneous Confidence Intervals
Response Variable Charred
All Pairwise Comparisons among Levels of Lab
```

```
Lab = Lab1  subtracted from:

Lab     Lower   Center   Upper  ------+---------+---------+---------+
Lab2  -0.2445   0.2636  0.7718              (------*------)
Lab3  -0.5445  -0.0364  0.4718         (------*-------)
Lab4  -0.8445  -0.3364  0.1718     (------*------)
Lab5  -0.1991   0.3091  0.8173              (------*-------)
                                  ------+---------+---------+---------+
                                     -0.70      0.00      0.70      1.40


Lab = Lab2  subtracted from:

Lab    Lower   Center     Upper  ------+---------+---------+---------+
Lab3  -0.808  -0.3000   0.20816     (-------*------)
Lab4  -1.108  -0.6000  -0.09184  (------*-------)
Lab5  -0.463   0.0455   0.55362          (-------*------)
                                 ------+---------+---------+---------+
                                    -0.70      0.00      0.70      1.40
```

```
Lab = Lab3  subtracted from:

Lab     Lower   Center   Upper  ------+---------+---------+---------+
Lab4  -0.8082  -0.3000  0.2082      (-------*------)
Lab5  -0.1627   0.3455  0.8536                (------*------)
                                 ------+---------+---------+---------+
                                    -0.70      0.00      0.70      1.40
```

```
Lab = Lab4  subtracted from:

Lab    Lower  Center  Upper  ------+---------+---------+---------+
Lab5  0.1373  0.6455  1.154                  (------*------)
                             ------+---------+---------+---------+
                                -0.70      0.00      0.70      1.40
```

In this example only 2 intervals don't contain 0, $\mu_4 - \mu_2$ and $\mu_5 - \mu_4$. This matches with the results of the tests.

Different Multiple Comparisons Procedures

- Tukey: Optimal when investigating all pairwise comparisons only. Gives narrower confidence intervals than Bonferroni while still maintaining desired confidence level (Not available in Stata without addon routine `prcomp`)

- Scheffé: Useful if you want to look at contrasts in addition to pairwise ones. Gives wider intervals than Tukey (reasonable since it can handle more intervals) but is often better than Bonferroni (narrower intervals and more powerful tests). Available in Stata's `oneway` command.

- Bonferroni: Can be expanded to look at more than just pairwise comparisons. The critical value depends on the total number of comparison and they can be contrasts of any type.

- Dunnett: Used for comparing individual treatments to a control. e.g. μ_1 with μ_2, μ_1 with μ_3, ... , μ_1 with μ_k where group 1 is the control group. (Not available in Stata)

  An example where Dunnett is appropriate is an experiment that was performed investigating melting rates for 3 brands of margarines. Butter was also added into the study as a control. The comparisons of interest are

  - Brand 1 vs Butter
  - Brand 2 vs Butter
  - Brand 3 vs Butter