**The logic behind ANOVA**

Example: Wheaties cereal (wheaties.R, wheaties.csv, wheaties\_demo.xlsx)

General Mills wants to test 4 different box designs (1, 2, 3, and 4) for the next design of Wheaties. Ten grocery stores, with approximately the same sales volumes, are selected as the experimental units. Each store is randomly assigned one of the package designs. Other relevant conditions besides package design, such as: price of cereal, amount of shelf space, location of shelf space, etc… are kept constant for all the stores in the experiment. The data consists of the number of cereal boxes sold in a one-month period.

Let Yij be the number of cereal boxes sold in a one-month period. The i subscript denotes the design (factor level) type, and the j subscript denotes the store (experimental unit) within the ith design type level.

|  |  |  |  |
| --- | --- | --- | --- |
| **Design** | | | |
| 1 | 2 | 3 | 4 |
| y11 = 12  y12 = 18 | y21 = 14  y22 = 12  y23 = 13 | y31 = 19  y32 = 17  y33 = 21 | y41 = 24  y42 = 30 |
| = 15 | = 13 | = 19 | = 27 |

Notice the • used in the subscript for the sample means. This is common notation used to indicate that the values corresponding to the second subscript are being averaged over.

We will use ni to denote the number of experimental units per design type (treatment level). For example, the number of experimental units (observations for us here) for design type #1 is n1 = 2.

* Response variable: The number of cereal boxes sold
* Factor: The cereal box design
* Factor levels: 1, 2, 3, and 4
* Treatments: 1, 2, 3, and 4
* Experimental unit: A grocery store

Hypotheses:

Ho: μ1 = μ2 = μ3 = μ4

Ha: At least one pair of means are different

where μi denotes the population mean sales of design i = 1, 2, 3, 4.

If don’t reject Ho:

There is not sufficient evidence to show there is a difference between population means. There is not enough evidence to show that one design is better on average than any of the others.

If reject Ho:

At least one pair of population means are different. Thus, there are some designs better (in terms of sales) than other designs on average.

Box and dot plots are very useful to better understand the data. For our case here, there are a small number of observations for each design level, so only a dot plot is of interest:

> wheaties <- read.csv(file = "wheaties.csv")

> wheaties

Design Store Response

1 1 1 12

2 1 2 18

3 2 1 14

4 2 2 12

5 2 3 13

6 3 1 19

7 3 2 17

8 3 3 21

9 4 1 24

10 4 2 30

> stripchart(x = wheaties$Response ~ wheaties$Design, lwd =

2, col = "red", method = "jitter", vertical = TRUE, pch

= 1, main = "Dot plot", ylab = "Number sold", xlab =

"Design type")



> aggregate(x = Response ~ Design, data = wheaties,

FUN = mean)

Added after video recording: R has changed the syntax for aggregate(). In the video, I show formula = Response ~ Design. Now, the proper syntax is x = Response ~ Design. I made the correction here and in the program.

Design Response

1 1 15

2 2 13

3 3 19

4 4 27

> aggregate(x = Response ~ Design, data = wheaties,

FUN = sd)

Design Response

1 1 4.242641

2 2 1.000000

3 3 2.000000

4 4 4.242641

Notice the separation between the data values for some of the designs on the dot plot. For example,

* All of design 4’s data values are larger than all of the other design’s data values. Obviously, design 4’s sample mean is larger than the other sample means.
* All of design 3’s data values are larger than design 2’s data values. Obviously, design 3’s sample mean is larger than design 2’s sample mean.
* Design 3’s data values are shifted farther up (more boxes sold) than design 1’s data values. Design 3’s sample mean is larger than design 1’s sample mean.

From these observations, is there sufficient evidence to indicate a difference between the **population mean** number of boxes sold for the different design types?

This question may be answered through using ANOVA! First, examine the within-sample variation:

* This is the variation of the sample values for design 1, design 2, design 3, and design 4 about their own means. In other words, it measures the variability within each design type.
* We call this value the mean square error (MSE):



= 46/6

= 7.67

Compare this expression to a sample variance calculated for a sample from one population.

* If we wanted to predict the value of yij, we would use  if we did not have any other information than the design type (treatment level). Thus, we can think of  as a measure of how far off we are. This quantity is called the sum of squares for error (SSE). It is the amount of variability left over or unexplained by using this method of prediction.

There is also another way to measure variability called between-sample variation:

* Compare the  to the overall mean  by using their sum of squared deviations from each other. We will then weight these values  by ni to account for possible differences in sample sizes from the ith design level. This produces a measure of variability between the design types.
* We call this value the mean square for the treatments (MST):

=258/3

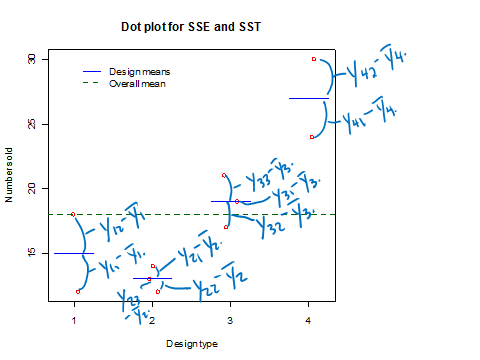
=86

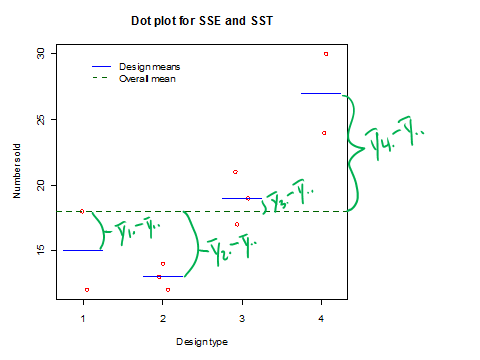
* If we wanted to predict the value of , we would use  if we did NOT have any additional information. Thus, we can think of  as a measure of how far off we are. This quantity is called the sum of squares for treatments (SST). Some book will refer to this as the “between-treatment sum of squares”.
* SST measures the variability explained by the differences between the sample means for each treatment and the overall sample mean.

Here are some plots demonstrating what MSE and MST are measuring:



Please see the program for the code to produce the plot.





Finally, the total sums of squares in the experiment is:



Note that

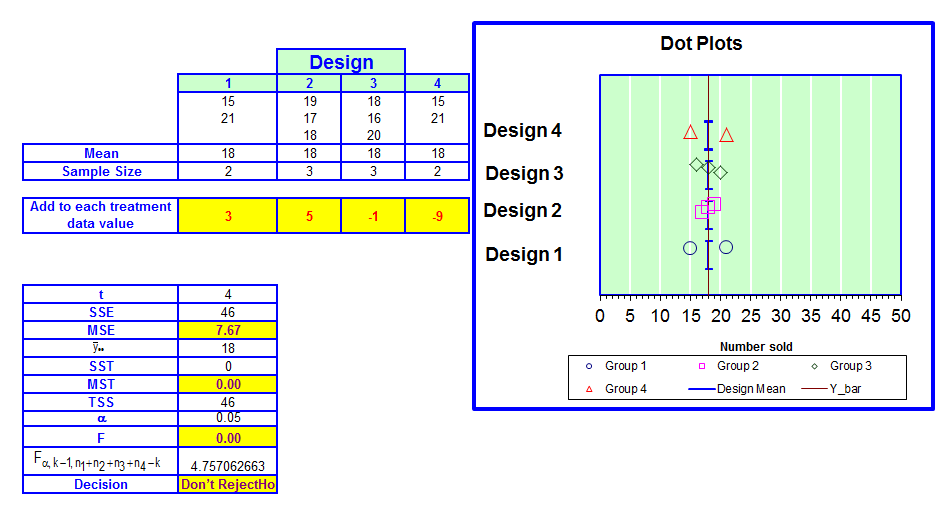


can be shown through using algebra. Also, note that TSS is how we would usually calculate the numerator of the sample variance!

Reasons for calling this ANOVA:

* Components of the total variance are examined.
* By examining MSE and MST, ANOVA methods can determine if μ1 = μ2 = μ3 = μ4 (no difference between the design types) or at least one pair of the means are not equal.
* Why does MSE and MST allow us to do this? Please see my wheaties\_demo.xlsx for an interactive demonstration.

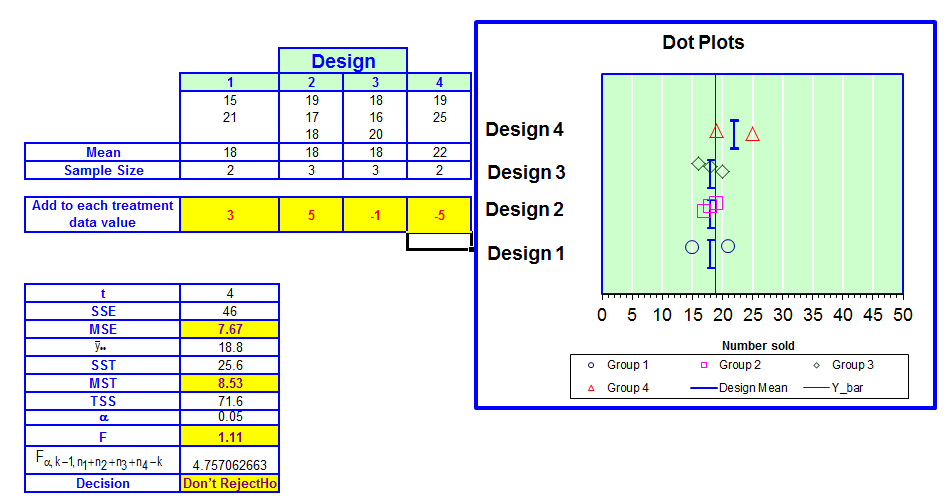
1. Suppose the design type means  are all equal. Obviously, in this case, you would not have any evidence from the sample to say that at least one pair of means was different (i.e., everything in the sample suggests μ1 = μ2 = μ3 = μ4)



obs

Notice that SST = MST = 0 because all of the design type means are equal to each other.

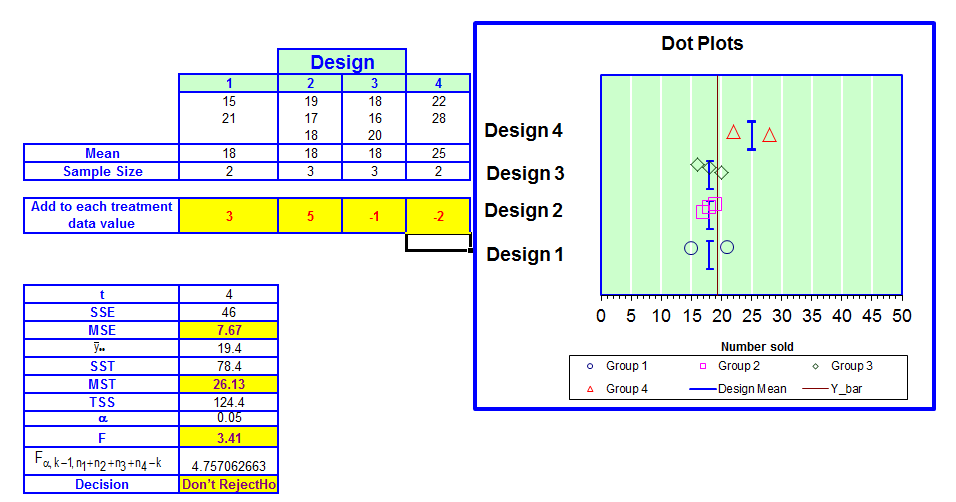
Now, let’s increase the mean for design type = 4 (subtract less from the original data):



obs

Notice that SSE and MSE remain the same because we still have the same variability within each design type’s responses. The only items that change are SST and MST. We can see that the design type sample means are now all a little different form the overall mean.

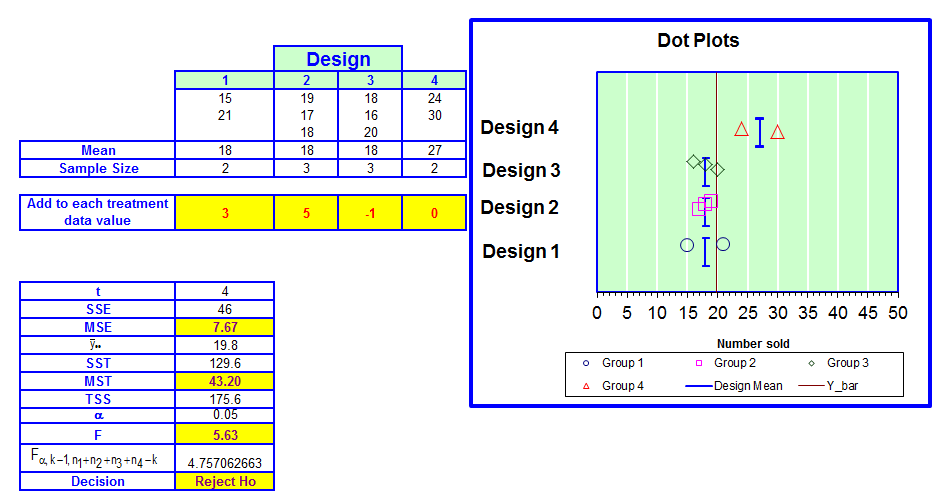
Now, let’s increase the sample mean more for design type = 4 (subtract less from the original data):



obs

Again, SSE and MSE remain the same because we still have the same variability within each design type’s responses. SST and MST increase again. We can see that the design type sample means are now even more different form the overall mean.

Finally, let’s increase the sample mean for design type = 4 even more and obtain the original observed data:



obs

Again, SSE and MSE remain the same because we still have the same variability within each design type’s responses. SST and MST are even greater now. We can see that the design type sample means are now even more different form the overall mean.

On each screen capture, we see a value Fobs calculated. The calculated value is

Fobs = MST/MSE

Thus, we are comparing the two sources of variation by their ratio. If μ1 = μ2 = μ3 = μ4 was true, we would expect MST and MSE to be similar, which was demonstrated in the above plots. However, when we started seeing more and more evidence of μ4 being different from the other means, MST continued to increase causing the ratio to be larger than 1. Therefore, Fobs is a measure of how much evidence there is against μ1 = μ2 = μ3 = μ4.

MST and MSE here are the observed values of the test statistic. Thus, “F” should be a lower case “f”. Because a capitol “F” is most often still used in these scenarios (due to the F distribution – see upcoming discussion), I am going to keep the capitol “F” and use “Fobs”.

When μ1 = μ2 = μ3 = μ4, one can show that the expected values of the random variable versions of MSE and MST are both , where I am using a subscript ε on σ2 to differentiate it from how we used σ2 in the past (it still means a population variance).

When μ1 = μ2 = μ3 = μ4 is NOT true, the expected value of the random variable version of MSE is still . However, the expected value of the random variable version of MST is now GREATER THAN . The amount that it is greater than grows as the differences among the means grows.

Thus, it is reasonable then to use Fobs = MST/MSE as a measure against μ1 = μ2 = μ3 = μ4. In fact, one can show that that Fobs = MST/MSE has an F distribution if μ1 = μ2 = μ3 = μ4 (and if some additional assumptions are satisfied – to be discussed later). Thus, we can make statements such as



where X has an F distribution with ν1 = t – 1 = 4 – 1 = 3 and ν2 = N – t = 10 – 4 = 6 degrees of freedom and  is the overall sample size. If we observe a Fobs value that is “large” relative to the F distribution, this leads us to believe that μ1 = μ2 = μ3 = μ4 is NOT true. We then use  as the critical value for a hypothesis test of

Ho: μ1 = μ2 = … = μt (No difference in pop. means)

Ha: At least one pair of means are not equal