**Completely Randomized Design (CRD)**

A CRD is a particular kind of experimental design where treatments are randomly applied to the experimental units.

The treatments are “completely randomized” when assigned to an experimental unit.

|  |
| --- |
| **Treatments** |
| 1 | 2 |  | t |
| \_\_\_\_\_\_\_\_\_ | \_\_\_\_\_\_\_\_\_ |  | \_\_\_\_\_\_\_\_\_ |

where \_\_\_\_ is an experimental unit.

There are n1 experimental units receiving treatment #1, n2 experimental units receiving treatment #2, …, and nt experimental units receiving treatment #t.

Example: Wheaties Cereal

The design types (treatments) of the boxes are randomly assigned to the stores (experimental units):

|  |
| --- |
| **Design** |
| 1 | 2 | 3 | 4 |
| Store 2Store 7 | Store 3Store 1Store 9 | Store 8Store 4Store 6 | Store 10Store 5 |

Stores 1 and 7 are randomly chosen to receive design type 1. Stores 3, 1, and 9 are randomly chosen to receive design type 2. Stores 8, 4, and 6 are randomly chosen to receive design type 3. Stores 10 and 5 are randomly chosen to receive design type 4.

Question: How can this type of assignment be done in R?

sample() function: Use sample(x = 1:10, size = 10, replace = FALSE). The first 2 numbers chosen are for treatment 1, the next three numbers are for treatment 2, ...

Example: Allergy Medications

Which drugs provide better relief on average to allergy sufferers: Seldane, Allegra, or Claritin?

Suppose there are 9 people for a study and each allergy medication receives an equal number of patients. The patients (experimental unit) are randomly given an allergy medication type (treatment):

|  |
| --- |
| **Allergy Medication** |
| Seldane | Allegra | Claritin |
| Patient 4Patient 8Patient 9 | Patient 2Patient 1Patient 5 | Patient 3Patient 7Patient 6 |

Goals of ANOVA: Determine if the population means for the factor levels (treatments) are different. This is done by a hypothesis test of

Ho: μ1 = μ2 = … = μt

Ha: At least two means are unequal

We can use the test statistic and p-value methods to perform the hypothesis test. We cannot use the confidence interval method because there would not be just one confidence interval to compare the means.

If we reject Ho, we will want to determine which means caused the rejection and which means are the smallest or the largest. We will discuss how to do this later.

Test Statistic Method:

1. Ho: μ1 = μ2 = … = μt
Ha: At least two means are unequal
2. Test statistic: Fobs = MST/MSE
3. Critical Value:  where N is the number of experimental units (sample size for us)
4. If Fobs >  reject Ho; otherwise don’t reject Ho
5. Conclusion

Reject Ho: There is at least one pair of \_\_\_\_ means that are different.

Don’t Reject Ho: There is not sufficient evidence to conclude that any of the \_\_\_\_ means are different.

where \_\_\_\_ means to insert the factor under investigation

P-value Method:

1. Ho: μ1 = μ2 = … = μt
Ha: At least two means are unequal
2. P-value is P(X > Fobs) where X is a random variable from a F distribution with ν1 = t – 1 and ν2 = N – t degrees of freedom and f is the observed value of
3. State α
4. Reject Ho if p-value < α; otherwise do not reject Ho
5. Conclusion

Assumptions:

* The populations for each treatment have approximate normal distributions.
* The population variances for each treatment are equal.

Without these assumptions holding, the type I error rate could be higher or lower than what is stated in the hypothesis test; however, small to moderate violations still allow the ANOVA methods to work well.

There are options available when one is worried about these assumptions:

* A nonparametric procedure could be used
* We can use a different distributional assumption. For example, a categorical data analysis course would show how to work with binary and count response data.

Why are we using these ANOVA methods now given the potential assumption problems?

This is an introductory statistics course! You need to start somewhere. The background you get from this course gives you the foundations for taking other courses. Plus, these ANOVA methods here often work well (and are WIDELY applied) even if there are problems with the assumptions.

Note: If t = 2, the hypothesis test is equivalent to testing the difference between two population means with independent samples (assuming equal variances).

ANOVA Table:

Summarizes the calculations for the hypothesis test

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Source** | **SS** | **DF** | **MS** | **F** |
| Treatments | SST | t – 1 | MST | Fobs |
| Error | SSE | N – t | MSE |  |
| Total | TSS | N – 1  |  |  |

where

* “DF” denotes degrees of freedom
* “SS” denotes sums of squares
* “MS” denotes mean squares
* “Source” denotes the source of variation.
* “Treatments” denotes the between treatment variation
* “Error” denotes the within treatment variation
*  measures the average variation between the treatments
*  measures the average variation within the treatments
*  measures the total variation in the experiment (not averaged)
* Fobs = MST/MSE

Make sure to know the relationships between quantities in the ANOVA table!

Example: Wheaties Cereal (wheaties.R, wheaties.csv)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Source** | **SS** | **DF** | **MS** | **F** |
| Treatments | SST | t – 1 | MST | Fobs |
| Error | SSE | N – t | MSE |  |
| Total | TSS | N – 1  |  |  |

=

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Source** | **SS** | **DF** | **MS** | **F** |
| Design | 258 | 4 – 1 = 3 | 258/3 = 86 | Fobs = 86/7.67 = 11.22 |
| Error | 46 | 10 – 4 = 6 | 46/6 = 7.67 |  |
| Total | 304 | 10 – 1 = 9 |  |  |

Hypothesis Test with α = 0.05:

Computations performed by the aov() function:

> mod.fit <- aov(formula = Response ~ factor(Design), data

 = wheaties)

> summary(object = mod.fit)

 Df Sum Sq Mean Sq F value Pr(>F)

factor(Design) 3 258 86.000 11.217 0.007135 \*\*

Residuals 6 46 7.667

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Notes:

* The syntax used for the formula argument of aov() is similar to what we have seen with other functions. One difference is that we need to use factor(Design) instead of just Design because Design is coded numerically. Simply, the factor() function tells R that we have a factor rather than treating its components as numerical values. If I would have used the levels of “A”, “B”, “C”, and “D” in the data set, factor() would not be necessary. When we discuss regression models later in the course, we will see what R does with a numerical value instead of a character value.
* The results from aov()are saved into an object named mod.fit (means “model fit”). I could have used a different name for the object if desired.
* The summary()function is used to summarize what is inside of mod.fit.
* To see more information of what is stored within mod.fit, use the names() function. We can access components of the object with mod.fit$<name>. We will discuss the items stored within the object later.

> names(x = mod.fit)

 [1] "coefficients" "residuals" "effects"

 [4] "rank" "fitted.values" "assign"

 [7] "qr" "df.residual" "contrasts"

[10] "xlevels" "call" "terms"

[13] "model"

Hypothesis test:

1. Ho: μ1 = μ2 = μ3 = μ4

Ha: At least two design type means are unequal

1. Test statistic: Fobs = 11.22
2. Critical Value:  = F0.05, 4-1, 10-4 = 4.76

> qf(p = 0.95, df1 = t - 1, df2 = N - t)

[1] 4.757063

1. Because 11.22 > 4.76, reject Ho.
2. There is at least one pair of design type means that are different.

Consequences:

Some design types are better than others. Thus, the company will make more money (on average) with one, two, or three of the designs than they would with the remaining design types.

Which design should General Mills use?

We will soon see!

ANOVA model:

By using ANOVA methods, we are actually estimating a statistical linear model of the form:

Yij = μ + αi + εij

for i = 1, …, t and j = 1, …, ni where

* μ is an overall mean
* αi = treatment effect
* εij = random error term that has a normal distribution with mean 0 and variance 

This model shows how we believe that the response Yij comes about. There is an overall mean for the treatments, where each treatment is a little different so we add a treatment effect. Because every response for a treatment will not be the same, we add an error term to it that accounts for perhaps other items we are not taking into account.

The μ and αi are parameters that we want to estimate. The εij is a random variable. Thus, for the (i, j) value in our sample, we could obtain Yij if we somehow knew μ, αi, and εij. Also, we can see that if α1 = … = αt, this implies there is no “treatment effect” so that μ1 = … = μt in our original notation. Because the model is written in this way, this is often referred to as a treatment effect model. Others will sometimes write the model as

Yij = μi + εij

which is referred to as a means effect model.

Example: Wheaties Cereal (wheaties.R, wheaties.csv)

We will use information from the sample to estimate the parameters. We observed that

 = 15,  = 13,  = 19, and  = 27

Also, the treatment effects model gives us

Y1j = μ + α1 + ε1j

Y2j = μ + α2 + ε2j

Y3j = μ + α3 + ε3j

Y4j = μ + α4 + ε4j

by writing out the four treatments. The estimate of μ + αi is ! Thus,



where I use ^’s to help denote now that are dealing with parameter estimates. We have a system of equations that we can use to solve for . Unfortunately, there are 5 “unknowns” and only 4 equations leading to multiple solutions. To obtain one solution, I am going to set  (this is what R does) and solve for the remaining values:



We see then that  for i = 2, 3, 4 measure the deviation from the treatment level 1 mean.

There are other ways to obtain to parameter estimates. For example, the SAS software package would set . The most important part is that the same overall mean estimates  (where  is the mean estimate for treatment i based on the model) come about no matter how the solutions are found. Solving for the  values is used here only as a tool to help you understand the model better.

What role does εij? This describes the deviation of the individual observations from the treatment means (see previous plots). These deviations should have a normal distribution with the same variance across the treatment types (based on our model assumptions). How could you examine these model assumptions more closely?

Below is a more general way to perform the computations in R:

> mod.fit.lm <- lm(formula = Response ~ factor(Design),

 data = wheaties)

> names(mod.fit.lm)

 [1] "coefficients" "residuals" "effects" "rank"

 "fitted.values"

 [6] "assign" "qr" "df.residual"

 "contrasts" "xlevels"

[11] "call" "terms" "model"

> mod.fit.lm$coefficients

 (Intercept) factor(Design)2 factor(Design)3 factor(Design)4

 15 -2 4 12

> summary(object = mod.fit.lm)

Call:

lm(formula = Response ~ factor(Design), data = wheaties)

Residuals:

 Min 1Q Median 3Q Max

 -3.00 -1.75 0.00 1.75 3.00

Coefficients:

 Estimate Std. Error t value Pr(>|t|)

(Intercept) 15.000 1.958 7.661 0.000258 \*\*\*

factor(Design)2 -2.000 2.528 -0.791 0.458922

factor(Design)3 4.000 2.528 1.583 0.164620

factor(Design)4 12.000 2.769 4.334 0.004908 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.769 on 6 degrees of freedom

Multiple R-squared: 0.8487, Adjusted R-squared: 0.773

F-statistic: 11.22 on 3 and 6 DF, p-value: 0.007135

> anova(object = mod.fit.lm)

Analysis of Variance Table

Response: Response

 Df Sum Sq Mean Sq F value Pr(>F)

factor(Design) 3 258 86.000 11.217 0.007135 \*\*

Residuals 6 46 7.667

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Notes:

* The lm() function estimates a “linear model”, which is a more general form of an ANOVA model. The “linear” name comes from adding the components inside the model. We will use this function a lot for regression models later in the course. The key aspect here to make note of is that an ANOVA model is a special case of a regression model!
* The results from lm()are saved into an object named mod.fit.lm. I could have used a different name for the object if desired.
* The names() function allows us to see what is stored in the object. We can access components of the object with mod.fit$<name>. One example is given showing the parameter estimates ! There are number of extra items given to us by lm() that we do not need at this time.
* The summary()function is used to summarize what is inside of mod.fit.lm. The “Estimate” column matches what we had for  earlier. Also, the “F-statistic” and corresponding degrees of freedom and p-value match what we had before for a test of Ho: μ1 = μ2 = μ3 = μ4 vs. Ha: At least two means are unequal.
* The anova() function with mod.fit.lm provides the ANOVA table.

Why did I show you both the aov() and lm() function then?

Two reasons:

1. Both are used in practice.
2. A problem that some students have is relating an ANOVA model to a regression model. As mentioned before, an ANOVA model is just a special case of a regression model. When we discuss regression models, I plan to come back here to emphasize this point!