Lectures 28 – 30: One-Way Analysis of Variance

Motivation and an Example

Example: As city manager, one of your responsibilities is purchasing. The city is looking to buy lightbulbs for the city's streetlights. Aware that some brands' lightbulbs might outlive other brands' lightbulbs, you decide to conduct an experiment. Seven lightbulbs each are purchased from four brands (GE, Dot, West, and a generic) and placed in streetlights. The lifetime of each of the 28 lightbulbs, in thousands of hours, is then recorded in the file *Lightbulbs*.

In this example the lifetime of a lightbulb, in thousands of hours, is the quantitative dependent variable of interest. Unlike regression, however, the independent variable is the brand under which a lightbulb is marketed, which is a qualitative (or categorical) variable. Now, we learned how to incorporate categorical independent variables in regression by defining dummy variables, and we could do that here. However, when the only independent variable is categorical, One-Way Analysis of Variance, abbreviated One-Way ANOVA, is a more flexible alternative to regression. (At the end of these lectures in one-way ANOVA we will rerun the analysis as a regression so that we may compare the results.)

Some Terminology

We begin by defining the terms we'll use in this section:

- Response: the dependent variable. In our example, the response variable is bulb lifetime, in hours
- Factor: the independent variable. In ANOVA, independent variables are called factors rather than predictors. The factor variable in our example is the brand under which a lightbulb is marketed.
- Levels: the possible values of a factor. In our example, the levels are the brands: GE, Dot, West, and generic.
- Treatments: another name for levels in one-way ANOVA. In these notes I use term Treatments to refer to the levels of a factor, e.g., the four treatments in our example are GE, Dot, West, and generic. The term treatments derives from medicine, where the different treatments are the drugs or procedures being tested on patients, and agriculture, where the treatments were the different fertilizers or pesticides being tested on crops. (**Note:** When I introduce two-way ANOVA later, I will have to distinguish factor levels from treatments.)

The (Cell Means) Model

The model used in one-way ANOVA is similar in many respects to the model employed in regression. In fact, you may find it useful to make analogies between the model and formulas in one-way ANOVA and the corresponding model and formulas in regression. The most commonly used one-way analysis of variance model is

$\boldsymbol{Y} = \mu_i + \boldsymbol{\varepsilon}$**,** where

- $\boldsymbol{Y}$ is the quantitative response variable.
- μ_i is the true mean of the i^{th} treatment, where there are k treatments being compared. For example, μ_{GE} is the true mean lifetime of the lightbulbs sold by General Electric.
- $\boldsymbol{\varepsilon}$ is the random error in the response not attributable to the independent variable? As in regression, the error is assumed to be normally distributed with constant variance.

The **ANOVA** Table: **S**ums of **S**quares and **D**egrees of **F**reedom

At the heart of any analysis of variance is the ANOVA Table. The formulas for the sums of squares in ANOVA are simplified if the *k* samples are all of the same size n_s. Experimental designs incorporating equal sample sizes are said to be Balanced. In the interests of simplicity, therefore, the following discussion assumes that all *k* samples contain the same number of observations n_s. (Of course, observational studies may not have the luxury of equally sized treatment samples, but much of what follows holds for unbalanced designs as well with minor, and usually obvious, changes to formulas.)

Notation: Analysis of Variance is very notation "dense," but the notation is rational and should be easy to remember.

- The index *i* represents the i^{th} treatment or level, where *i* ranges from 1 to *k.*
- The index *j* represents the j^{th} observation within a treatment sample, where *j* ranges from 1 to n_s.
- *n* is the total number of observations from all samples. $n = kn_s$ for a balanced design.
- y_{ij} is the value of the j^{th} observation in the i^{th} treatment sample.
- $\bar{y}_i$ is the mean of the i^{th} treatment sample.
- $\bar{\bar{y}}$ (read "*y* double-bar") is the mean of all *n* observations, $\bar{\bar{y}} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_S} y_{ij}$, or the mean of the sample means (hence the "double-bar" in the name), $\bar{\bar{y}} = \frac{\bar{y}_1 + \bar{y}_2 + \cdots + \bar{y}_k}{k}$, *provided* all samples are the same size n_s, i.e., provided the experimental design is balanced.

Sums of Squares

In regression, the sums of squares were located in the ANOVA table, so it should come as no surprise that similar sums of squares play a central role in the analysis of variance.

Note: In the discussion below, be advised that what I'm calling *SSR* and *MSR* are often called *SST* and *MST*, respectively, in the literature. I've chosen my notation to emphasize the similarities between regression and the analysis of variance. *MSE* remains the same for both regression and the analysis of variance.

Treatment Sum of Squares, $SSR = n_s\sum_{i=1}^{k}\left(\bar{y}_i - \bar{\bar{y}}\right)^2$ is the "Between Treatments" variation, i.e., the variation in the treatment means about the mean of means $\bar{\bar{y}}$. If the sample means differ substantially, then *SSR* will be large and we'll have evidence for a statistically significant Treatment Effect on the mean response.

Error Sum of Squares, $SSE = \sum_{i=1}^{k}\sum_{j=1}^{n_S}\left(y_{ij} - \bar{y}_i\right)^2$ is the "Within Treatment" variation, i.e., the variation of the response variable about the treatment means.

Total Sum of Squares, $SST = \sum_{i=1}^{k}\sum_{j=1}^{n_S}\left(y_{ij} - \bar{\bar{y}}\right)^2$ is the total variation in the values of the response variable over all **k** samples. (**Note:** *SST* is the same as in regression)

Degrees of Freedom

Degrees of freedom for treatments, $df_{SSR} = k - 1$. Rather than memorizing this formula, imagine the number of dummy variables that would have to be created to conduct the analysis in regression. Since you always leave one possibility out in regression, you would need to create $k - 1$ dummy variables to represent the k treatments. Since the resulting regression model would have $k - 1$ independent variables, *SSR* would have $k - 1$ degrees of freedom.

Degrees of freedom for error, $df_{SSE} = n - k$.

Total degrees of freedom, $df_{SST} = n - 1$. This is the same result obtained in regression.

Note: $df_{SSR} + df_{SSE} = df_{SST}$, just as in regression.

Mean Squares

As in regression, the mean squares estimate variances.

Treatment Mean Square, $MSR = \frac{SSR}{k - 1}$ is equivalent to *MSR* in regression

Error Mean Square, $MSE = \frac{SSE}{n - k}$ is the same as *MSE* in regression. As in regression, *MSE* is an unbiased estimator of the error variance σ^2.

Discussion: The analysis of variance, like linear regression, is an example of a linear statistical model. I will not emphasize the use of matrices in ANOVA to the extent that I did in simple and multiple linear regression, but you will have an opportunity to explore the topic in an upcoming quiz or homework.

The **ANOVA** Table Summarized

The ANOVA Table below summarizes some of the information to this point.

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Between Treatments	*SSR*	$k - 1$	$MSR = SSR/(k-1)$	$F = MSR/MSE$	
Within Treatments	*SSE*	$n - k$	$MSE = SSE/(n-k)$		
Total (Corr.)	*SST*	$n - 1$			

ANOVA Table for One-Way Analysis of Variance with *k* Treatments (Levels)

Hypotheses

As in regression, a hypothesis test is conducted to determine if the treatments being compared have differential effects on the mean response. The hypotheses below are reminiscent of the hypotheses in multiple regression.

- $\mathbf{H}_0$ **:** $\mu_1 = \cdots = \mu_k$, i.e., all treatment means are equal. This is equivalent to saying that the k treatments have no differential effect upon the mean value of the response.
- $\mathbf{H}_A$ **:** At least two of the means differ. This says that different treatments produce different mean responses.

Example: It is important to understand why probability is important. It is not unusual for one manufacturer to source a product marketed under many brand-names. For example, there are only a handful of companies manufacturing denim jeans, but there are dozens of brand-name jeans available to the consumer. Similarly, not all lightbulbs are manufactured by the companies marketing them. It is not inconceivable, therefore, that all four brands of lightbulbs being tested by the city come off of the same assembly line. Yet, when tested, samples drawn from the four brands would have different mean lifetimes due to sample-to-sample variation. As city manager, you might be more than a little embarrassed to discover that the brand that you've touted as superior to all others is actually different in name only!

The *F–Ratio* Test Statistic

The statistic used to test the null hypothesis $\mu_1 = \cdots = \mu_k$ is $F = \frac{MSR}{MSE}$. Notice that the test statistic for the model is an *F-Ratio* as in linear regression. The decision rule, as in regression, is to reject the null hypothesis that treatments don't affect the mean response when $F \gg 1$.

The Logic Behind the *F-Ratio*

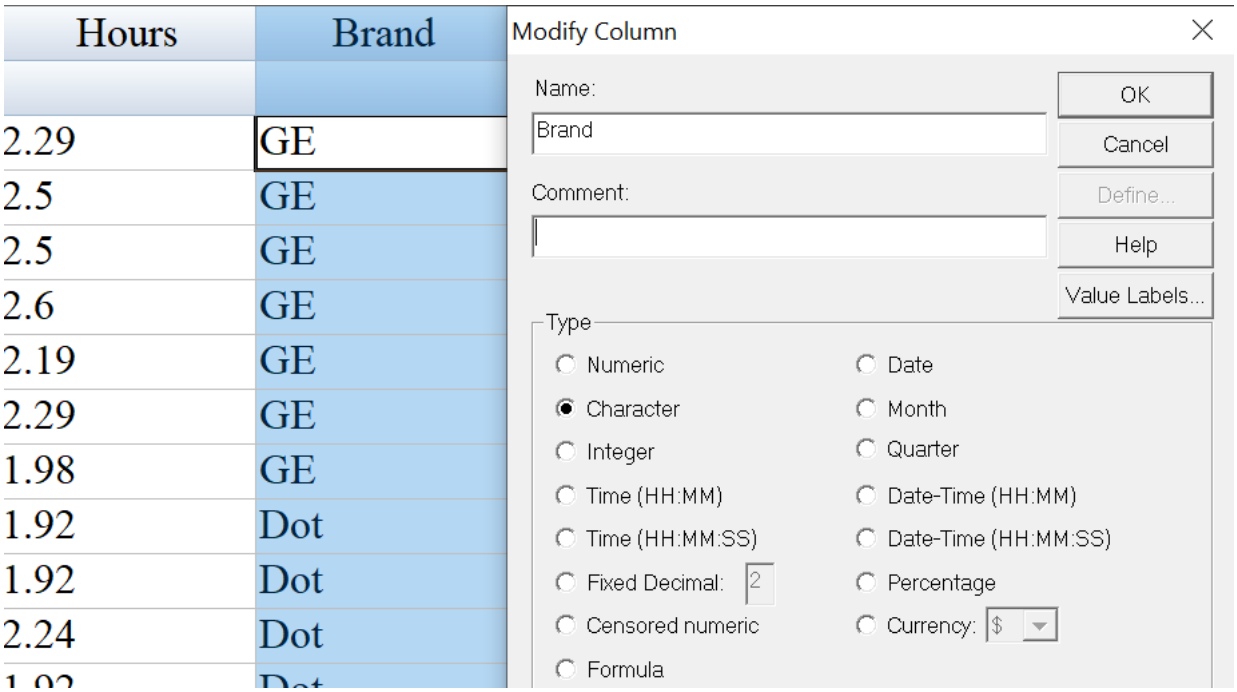
The analysis of variance uses the ratio of two sample ***variances***, *MSR* and *MSE*, to test whether treatment ***means*** differ: hence the name "analysis of variance." Recall that one of the assumptions of the model is that the variance σ^2 is the same for all treatments. The Mean Square Error (*MSE*) provides an unbiased estimate of σ^2 in ANOVA just as it does in regression (see regression notes).

If the null hypothesis $\mathbf{H}_0$ is correct, the ***true*** treatment means (the μ_i in the model) are all equal and it can be shown that *MSR* also provides an unbiased estimate of σ^2. Thus, if all treatment means are equal, we would expect $F = \frac{MSR}{MSE}$ to be nearly equal to 1 since *MSR* and *MSE* should yield similar estimates of the error variance σ^2. If the alternative hypothesis $\mathbf{H}_A$ is correct, however, then *MSR* is a biased estimator of σ^2 with a strictly positive bias (i.e., *MSR* tends to overestimate σ^2) and the *F-Ratio* will tend to be greater than 1. Thus, large values of the *F-Ratio* lead to the conclusion that at least two treatment means differ.

Example - Lightbulb Lifetimes

A city is looking to buy lightbulbs for the city's streetlights. Seven lightbulbs from each of four brands (GE, Dot, West, and a generic) are purchased and placed in streetlights. The lifetime of each of the 28 lightbulbs is then recorded in the file *Lightbulbs*. Below we perform a one-way analysis of variance of the lifetimes in Statgraphics.

Notice when you open the file *Lightbulbs* in Statgraphics, the column for the factor brand doesn't have to be numeric. If you right-click on the column and select *Modify Column* you'll see the data-type is set to *Character*.



To perform a one-way analysis of variance in Statgraphics, follow *Compare* > *Analysis of Variance* > *One-Way ANOVA* and enter the response variable *Hours* as the Dependent Variable and the independent variable *Brand* as the Factor.

There a number of tables and graphs to consider, but start with the ANOVA Table reproduced below.

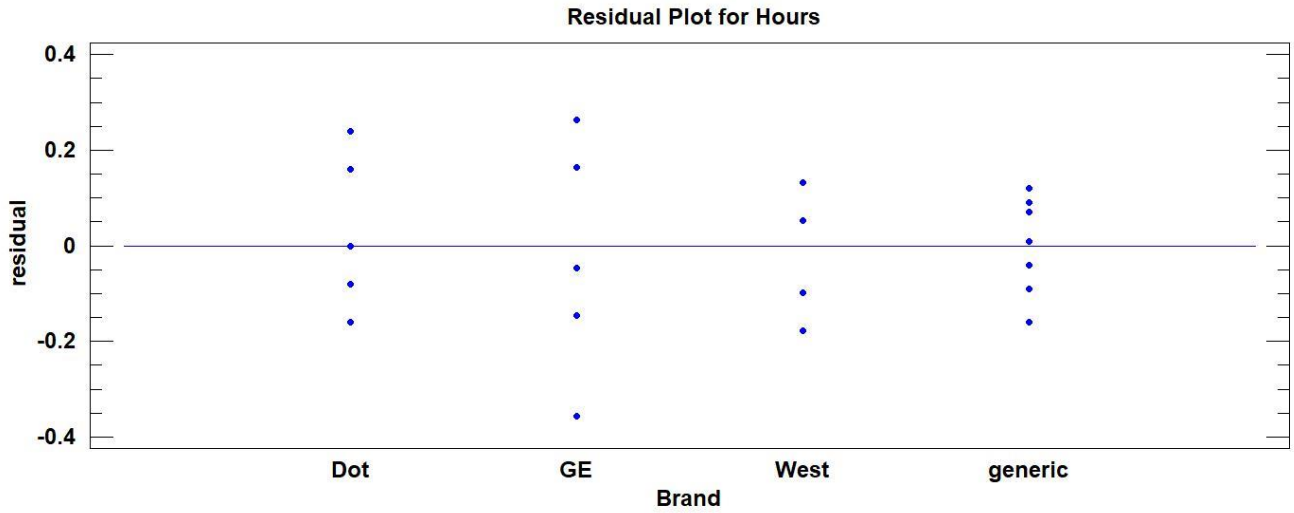
Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Between groups	1.08917	3	0.363057	15.62	0.0000
Within groups	0.557714	24	0.0232381		
Total (Corr.)	1.64689	27			

ANOVA Table for Hours by Brand

For this example, the treatment sample size is $n_s = 7$ for all brands, and there are four brands, so the total number of lightbulbs tested is $n = k \times n_s = 4 \times 7 = 28$. Following the previous discussion, the degrees of freedom are $df_{SSR} = 4-1$, $df_{SSE} = 28-4$, and $df_{SST} = 28-1$.

The small *P*-value supports the conclusion that some brands have longer mean lifetimes than others, but a lot of work remains to be done. We'll need to run some diagnostics to determine if the assumptions of the model appear reasonable, especially the assumption of equal variance across brands. Finally, we'll need to decide what we can say about the relative lifetimes of the brands. WARNING: There are often limitations on what conclusions we can justify with a reasonable amount of confidence. Don't claim evidence for differences that may be explainable by chance variation alone!

Our next stop should probably be the *Residual Plots* to see if the assumption of constant variance is reasonable. In the plot below, there does appear to be more variation in the GE bulbs, but some of this is due to one bulb that failed mush early than other GE bulbs tested (an outlier), so the issue is probably minor and we'll proceed.



Staying with the assumption of equal variances across brands, the *Variance Check* window reports the result of Levene's test. As the output below suggests, the null hypothesis of equal variances remains plausible due to the relatively large *P*-value of 0.2089.

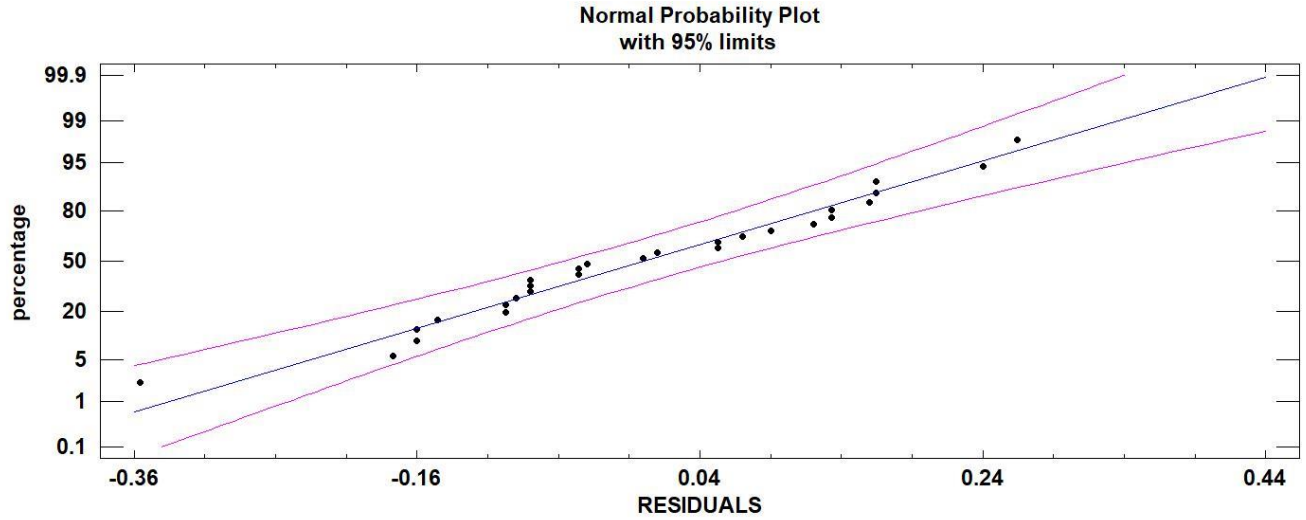
Variance Check

	Test	*P-Value*
Levene's	1.62908	0.2089

The StatAdvisor

The statistic displayed in this table tests the null hypothesis that the standard deviations of Hours within each of the 4 levels of Brand is the same. Of particular interest is the *P*-value. Since the *P*-value is greater than or equal to 0.05, there is not a statistically significant difference amongst the standard deviations at the 95.0% confidence level.

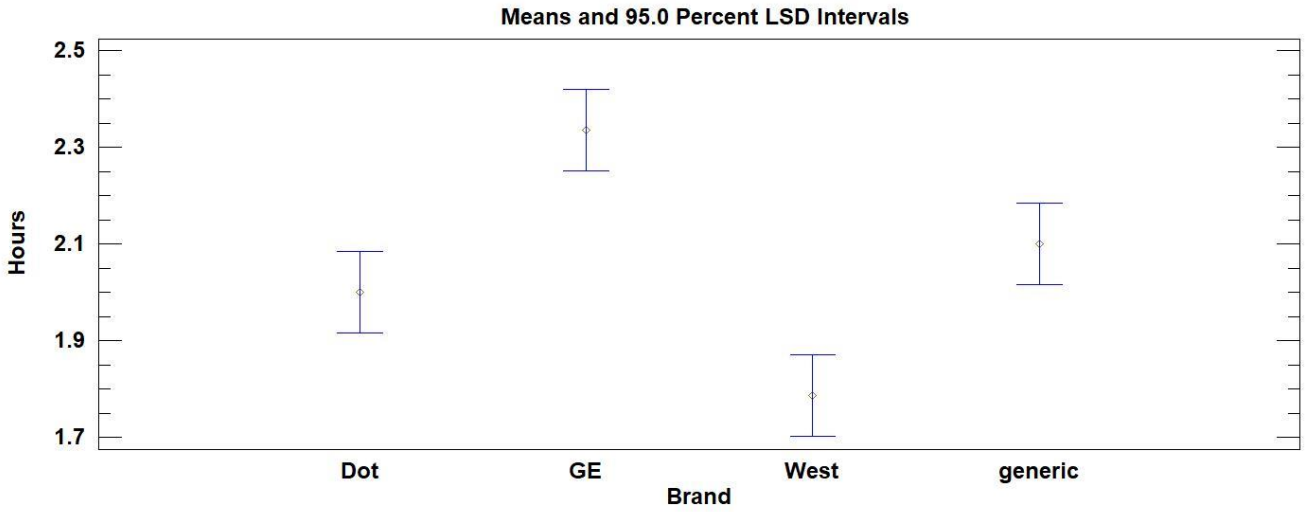
To check the assumption that errors are normally distributed, a normal probability plot was created. The plot below supports the reasonableness of the assumption, which is given further credence by the values for standardized skewness and kurtosis reported by Statgraphics (see output below the probability plot).



Stnd. skewness	-0.382337
Stnd. kurtosis	-0.0760129

The StatAdvisor: The standardized skewness and standardized kurtosis are within the range expected for data from a normal distribution.

Next, we'll look at the *Means Plot* which displays confidence intervals for the means of all four brands. The intervals all have the same width because Statgraphics assumes a common variance and uses all 28 lifetimes to estimate it. Statgraphics defaults to 95% LSD (Least Significant Difference) intervals. There are a number of options available by right-clicking and selecting *Pane Options*. Some of these are pretty self-explanatory, but there are also several other types of intervals to choose from besides LSD. You may want to play around with these to see what affect they have on the widths of intervals. We'll discuss the different intervals available to us in a later lecture, but for now we'll just accept the LSD intervals Statgraphics defaults to (they tend to be the narrowest).



Intervals that don't overlap provide evidence, at the stated level of confidence, for a difference in the means of the treatments represented by the intervals. For example, the samples provide evidence that GE lightbulbs last longer, on average, than those marketed under the West label. It may be difficult to tell from the graph whether two intervals overlap, so next we'll consult the *Multiple Range Tests* window which produces the tables below. The first table provides the sample mean for each treatment sample (which is a point estimate of the true mean lightbulb lifetime for the brand), and also groups the brands into homogeneous groups. Brands within a homogeneous group can't be "separated" at the stated level of confidence. (Remember, it's quite possible that brands flagged as belonging to a homogeneous group were all sourced from the same manufacturer and came off of the same assembly line, so don't try to force conclusions that can't be supported statistically!)

The good news for the city manager is that the brand with the greatest sample mean lifetime, GE, appears as its own group.

Method: 95.0 percent LSD

Brand	*Count*	*Mean*	*Homogeneous Groups*
West	7	1.78714	X
Dot	7	2.0	X
generic	7	2.1	X
GE	7	2.33571	X

The second table looks at differences between pairs of brands. This is how Statgraphics determines the homogeneous groups in the first table.

Contrast	*Sig.*	*Difference*	*+/- Limits*
Dot - GE	*	-0.335714	0.168173
Dot - West	*	0.212857	0.168173
Dot - generic		-0.1	0.168173
GE - West	*	0.548571	0.168173
GE - generic	*	0.235714	0.168173
West - generic	*	-0.312857	0.168173

* denotes a statistically significant difference.

To make a point, I've used Pane Options to create Tukey HSD intervals. These intervals are wider than corresponding LSD intervals, and as a result the Dot branded bulbs are assigned to two different groups, one with the brand West and another with the generic lightbulbs. Based on these intervals, we would not be able to justify conclusions differentiating Dot from West or Dot from generic. GE remains the clear winner, however.

Method: 95.0 percent Tukey HSD

Brand	*Count*	*Mean*	*Homogeneous Groups*
West	7	1.78714	X
Dot	7	2.0	XX
generic	7	2.1	X
GE	7	2.33571	X