Historic Minnesota Waste Management Statistical Analysis – ANOVA

# Introduction

This report explores some possible relationships found within the provided Wastedata2.xlsx file. Within this excel file contains the Amount of Recycling, Organic Recycling, Onsite Recycling, Waste to Energy, and Landfilled waste, each in tons, over the years 1991 to 2017 for 86 Counties in Minnesota. We will use R Studio for all analyses.

We have two research questions to answer. Our first research question: Is there a significant difference in percent change in recycling across all of Minnesota between the years 1991, 1997, 2002, 2007, 2012, and 2017?

This question was chosen to explore the possibility that over the years recycling efforts have waned over the years. To test this hypothesis, we’ll use the variables Year as a categorical IV, Recycling as a continuous DV, and create our variable of interest: Percent Change, which measures the percent change of recycling between the before and after time stamp. The Year variable is also adjusted to reflect the time range, rather than ending year. We’ve selected these specific years to approximate equally timed data. The exact formula used to calculate this percent change is:

The Null Hypothesis is:

With an Alternative Hypothesis of:

Our second research question is: Between the lowest, middle, and top tier counties in terms of Onsite Recycling, is there a significant difference in the mean Landfilled waste for the year 2017?

The variable of interest we will use here is Landfilled Waste. To determine which counties to include in the analysis, we’ll gather the data by Counties, find the average Onsite Recycling for each county across all years, sort the data by this average, and pick the 15 counties from the ends and center of the data, and then split them into each three tiers: Low, Middle, and High. This will be our independent variable. The Null Hypothesis will be:

The Alternative Hypothesis is:

For reader comprehension these research questions will be consecutively answered in full.

# Research Question 1

## Methods

For this question, a One-way repeated measure ANOVA is the sound choice. The data is recorded with over time thus requiring a repeated measure method. The counties can be thought of as individuals, which keeps the data paired together. The percent change is the single dependent variable. As there is only one IV, Year, this makes sense to use a One-way repeated measure ANOVA. If this test comes up significant, we can further explore this by performing a multiple comparison analysis to determine which pairs of years are significantly different.

The assumptions for the One-way repeated measure ANOVA are:

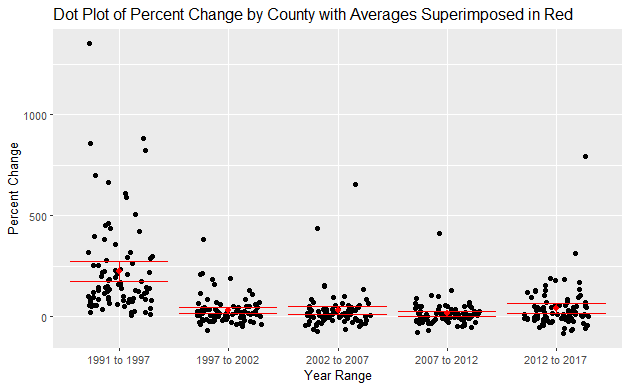
* Independence, which can be assumed by large data.
* Normality, which can be assumed by large data.
* Sphericity, which can be tested with Mauchly’s test of Sphericity.

## Conclusion

Considering we must test for sphericity, which is equal variance between Years, we use the function anova\_test() from the rstatix library to run both the one way repeated measure ANOVA model, as well as Mauchly’s Test of Sphericity. For Sphericity, the resulting test statistics, W, is 0.19 with a p-value of 0. This indicates we do not have equal variance between Years, and simply must use a degree of freedom adjustment for the ANOVA model. The F score initially is 40 with a P-value of 1.27e-27, but after the Greenhouse-Geisser degree of freedom adjustment changing the DF from 340 to 177.75, the P-value is 1.2e-15. As this is essentially 0 and less than an of 0.05, this means that we reject the null hypothesis. There is evidence to support the claim that at least two years pairs have different mean percentage change in Recycling.

Continuing to the multiple comparison analysis, we utilize the glht() function from the multcomp library to perform Tukey HSD test, and determine that the change from years 1991 to 1997 are statistically different from all other year ranges.

What this means intuitively, is that after the first 6 years of recording the recycling data, the amount of recycling has slowed down since. This can easily be viewed in this graph.



In this graph we see from 1991 to 1997, there’s much more spread in the data as well as a much higher average. We include a table of the means to zoom in on the x=0 axis.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Mean Percent Change Of Recycling Between Year Ranges** | | | | | |
| Year Range | 1991 to 1997 | 1997 to 2002 | 2002 to 2007 | 2007 to 2012 | 2012 to 2017 |
| Mean (%) | 224 | 28 | 28 | 13 | 39 |

The most interesting finding is how the later year ranges seem to have a mean much closer to 0, which might imply that there was a recycling boom early on, which has since died down to a steady growth over the years. The limitation of this analysis is that picking such wide year ranges smooths the data too much, and that percentage changes are over generous at start up when near 0 tons recycling. Time series analysis might be better for this case.

# Research Question 2

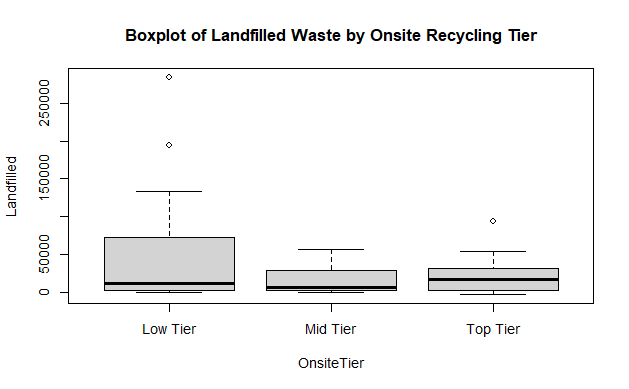
## Methods

For this research question, we use One-way ANOVA. This makes sense as we have a single dependent variable of interest, Tons Landfilled, and one independent variable with more than 2 levels; Low tier, Mid tier, and High tier On Site recycling. With this analysis, the assumptions that must be met are:

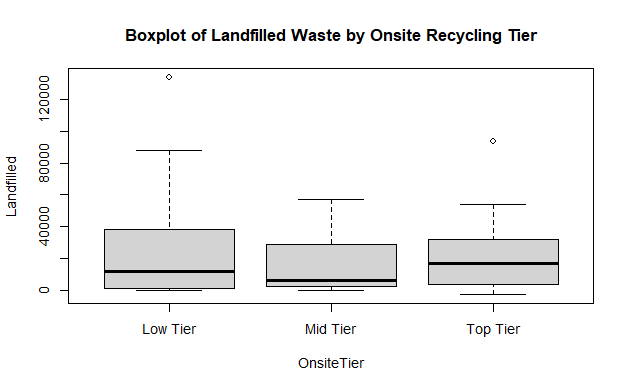
* Independence, assumed true as each county does not overlap.
* Normality, which can be assumed with the sample size of 45.
* Equal variance, which can be assessed in a box plot and Barlett’s Test.

## Conclusion

The result of the box plot is as follows:



From this, we can see very unequal variances, as depicted by the height of the boxes, between Low Tier and the others. We follow this up with Bartlett’s test of homogeneity of variance, which tests the alternative hypothesis: at least one tier has a variance that is unequal to the others. This results in a p-value of 3e-9, which is quite severe, so we adjust our approach.



In this graph we’ve removed the two outliers, Hennepin and Dakota County, from the Low Tier, and added the two next lowest counties. In the reassessed Barlett’s test, the p-value is now 0.01. The assumption is violated, but “In general, a one-way ANOVA is considered to be fairly robust against violations of the equal variances assumption as long as each group has the same sample size,” (Zach, 2019) which is the case for our data set. We proceed with the One Way ANOVA.

The F-statistic from the analysis is 0.73, which is a p-value of 0.49. This is larger than an of 0.05, which means we fail to reject the null hypothesis. There is insufficient evidence to support the claim that at least one tier of Onsite Recycling has a significantly different mean Landfilled waste. The primary limitation of this finding is the need to bar some of the most populous counties in the state. This sort of county-based analysis gives many more entries to rural counties, while Hennepin County is one of the most populous, while being approximately as large as other rural counties. The most interesting finding is how generally similar the variance is across the wide spectrum of on site recycling.

# References

Zach. (2019, December 22). *How to Check ANOVA Assumptions*. Retrieved from Statology: https://www.statology.org/anova-assumptions/

GitHub: https://github.com/TroyRodriguez/Projects-Portfolio/tree/main/Projects%20Portfolio