Natalie Klang

STAT301 Project

04/24/2023

Analysis of Recycling Data

*Introduction 1*

For this project, I have chosen to look at the mean difference between select residential recycling categories across all years and counties. Using the dataset “Wastedata1”: Is there a significant mean difference of tons of material recycled for the top 4 residential recycling categories? I have chosen to reduce the recycling categories from seven to four in order to reduce the noise in the Tukey analysis discussed later.

To answer this question, we require two variables. The independent variable is categorical: the recycling Category. As defined by the question, the recycling category may take on four values which will be determined by charting the mean tons of materials recycled by category.

|  |
| --- |
|  |
| **Category**  <chr> | **Average Tons Recycled Residentially**  <dbl> |  |  |  |
| Glass | 706.83493 |  |  |  |
| Paper | 680.12151 |  |  |  |
| Organic | 528.85227 |  |  |  |
| Metal | 433.04136 |  |  |  |
| Other | 221.25914 |  |  |  |
| Hazardous | 100.20804 |  |  |  |
| Plastic | 91.68357 |  |  |  |

We can see that the top four categories of residentially recycled materials are Glass, Paper, Organic and Metal. These are the values that we will limit the independent variable to. The second variable, the dependent variable, is the tons of material recycled residentially. In the dataset, this variable is called “Res Tons”. This variable is numerical.

The null hypothesis is that the mean tons of materials recycled across the top four categories is the same. The alternative hypothesis is that at least two means are different.

at least two means are different

*Method 1*

Because we are comparing more than two means for a single independent variable, the appropriate analysis is One Way ANOVA. We will assume the following conditions are met:

* **Normality**: each group of samples are taken from a normally distributed population
* **Equal variance**: variances of data in different groups are not significantly different
* **Independence**: each sample is drawn independently of other samples

If the p-value of the ANOVA test is less than 0.05, we can reject the null hypothesis and perform a Tukey analysis to understand which categories have means that are significantly different from one another.

Before performing our analysis, we will filter the dataset by Category to only include Paper, Glass, Organic and Metal.

*Conclusion 1*

After performing a One Way ANOVA test, we find that the F statistic is 95.85 and the p-value is very small (statistically 0). Since the p-value is less than the declared significance level (0.05), we reject the null hypothesis and claim that there a significant mean difference of tons of material recycled for the top 4 residential recycling categories.

Furthermore, the Tukey analysis provided the following p-values (“p adj”) between category comparisons.

diff lwr upr p adj

Metal-Glass -273.79357 -429.25596 -118.33118 0.0000358

Organic-Glass -177.98266 -391.81601 35.85068 0.1410145

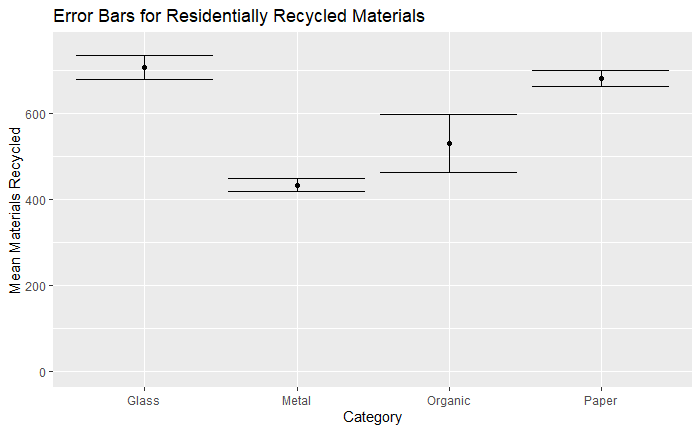
Paper-Glass -26.71342 -175.21420 121.78736 0.9672482

Organic-Metal 95.81091 -90.42910 282.05091 0.5490046

Paper-Metal 247.08015 142.13673 352.02357 0.0000000

Paper-Organic 151.26924 -29.20033 331.73881 0.1364949

From here, we can see that the p-values for metal and glass and metal and paper are very small. Therefore, we can conclude that the means for metal and glass and metal and paper are significantly different. This is further demonstrated with the below error bar plot.



I think it is interesting that glass is the most highly recycled material. From my own recycling habits, I recycle plastic and metal far more than glass. I also think it is interesting that organics recycling ranks higher than plastic given that not all cities have an organics recycling program. I wonder if the dataset reflects only the materials that are successfully recycled rather than the materials that residents put into their recycling bins. Hopefully all the plastic we use does not end up in the ocean!

*Introduction 2*

I am also interested in how landfilling habits may have changed over the years for all counties. For my second question, using “Wastedata2”, I want to know: Is there a significant mean difference in the tons of materials landfilled for the years 1991, 2000, 2009 and 2017?

This question requires three variables. The independent variable is the year the data was collected, which we will consider to be categorical. The dependent variable is the tons of material landfilled, which is numerical. The third variable is the individual, which in this case is the county.

The null hypothesis is that the mean landfilled materials by county across all years is the same. The alternative hypothesis is that at least two means are different.

*Method 2*

Because we are comparing the dependent measures of a single sample under more than two times, this question requires a One Way Repeated Measure ANOVA test. Again, we will assume the following conditions are met:

* **Normality**: each group of samples are taken from a normally distributed population
* **Equal variance**: variances of data in different groups are not significantly different
* **Independence**: each sample is drawn independently of other samples

Before performing our analysis, we will filter the dataset to include only the years 1991, 2000, 2009 and 2017.

*Conclusion 2*

The ANOVA test provides an F statistic of 20.42 and p-value of >0.0001. Since the p-value is less than , we can reject the null hypothesis and say there is evidence to support that claim that there a significant mean difference in the tons of materials landfilled for the years 1991, 2000, 2009 and 2017.

From the Tukey analysis, we see that the p-values (Pr(>|z|)) for the comparisons of 1991 to 2000, 1991 to 2009 and 1991 to 2017 are significantly different.

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

2000 - 1991 == 0 10376.6 2875.2 3.609 0.00178 \*\*

2009 - 1991 == 0 9770.0 2875.2 3.398 0.00378 \*\*

2017 - 1991 == 0 9488.4 2875.2 3.300 0.00529 \*\*

2009 - 2000 == 0 -606.7 2875.2 -0.211 0.99668

2017 - 2000 == 0 -888.2 2875.2 -0.309 0.98978

2017 - 2009 == 0 -281.6 2875.2 -0.098 0.99966

Going a step further, I am curious to know if the average amount of materials landfilled have increased or decreased since 1991. Charting the means by year:

| **Year**  <fctr> | **Average Landfilled in Tons**  <dbl> |  |  |
| --- | --- | --- | --- |
| 1991 | 11818.64 |  |  |
| 2000 | 22195.28 |  |  |
| 2009 | 21588.62 |  |  |
| 2017 | 21307.03 |  |  |

We see that the amount Landfilled each year has grown since 1991. This is interesting because one would assume that participation in recycling programs has grown over the years and therefore the amount in landfills would go down. However, we do see a decrease in landfilled amount from 2000 to 2009 and from 2009 to 2017. Therefore, we should consider that some data from 1991 may be under reported.