Sage Research Methods Datasets Part 1

Learn to Use One-Way ANOVA in SPSS With Data From the Canadian Fuel Consumption Report (2015)

For the most optimal reading experience we recommend using our website.

A free-to-view version of this content, which is easy to navigate and search, and includes interactive questions, is available by clicking on this link

Pub. Date: 2015

Product: Sage Research Methods Datasets Part 1

DOI: https://doi.org/10.4135/9781473947351

Methods: One-way analysis of variance, SPSS, Continuous variables

Keywords: fuel consumption, fuel, automobiles, natural resources, cities, Canada, gases

Disciplines: Economics, Social Policy and Public Policy, Science, Engineering

Access Date: March 3, 2025

Publisher: SAGE Publications, Ltd.

City: London

Online ISBN: 9781473947351

Student Guide

Introduction

This dataset example introduces readers to one-way Analysis of Variance (ANOVA). One-way ANOVA allows researchers to test the differences between means for a particular variable across two or more groups. This allows researchers to evaluate whether the mean of a given variable differs across two or more subsets of the data. One-way ANOVA is therefore an extension of the independent sample t-test for differences between two means.

This example describes one-way ANOVA, discusses the assumptions underlying it, and shows how to compute and interpret the results. We illustrate the method using a subset of data from the 2015 Fuel Consumption Report from Natural Resources Canada. Specifically, we test whether the average city fuel consumption of an automobile differs based on the type of fuel it uses.

What Is One-Way ANOVA?

One-way ANOVA is a method used to test whether the mean values of a continuous variable differ across two or more subsets of the data. Those subsets are generally defined by categories from a categorical variable. For example, you might compute the average weight for people living in three regions of a country to see if the average weight differs. In this way, one-way ANOVA allows researchers to explore whether a continuous variable (e.g. weight) and a categorical variable (e.g. region) are related to each other. If region proves to be a good way to distinguish between the average weight of people, then the mean value of weight will differ across regions and the variance in weight within regions will be small relative to the variance in weight across all regions.

When computing formal statistical tests, it is customary to define the null hypothesis (H_0) to be tested. In this case, the standard null hypothesis is that the mean of the continuous variable in question does not differ across the different groups defined by the categorical variable in question. Some difference is expected to appear in any one sample of data due to random chance in sampling. The F-test conducted here is designed

to help us determine if the differences are large enough to declare the test statistically significant.

“Large enough” is typically defined as selecting a critical value for the test statistic such that there is less than a 0.05 probability that the result observed in the sample of data occurred strictly due to random chance. When this probability, or *p*-value, is estimated to be less than 0.05, this would generally lead researchers to reject the null hypothesis (H_0) of no difference and conclude that there likely is a relationship between the two variables.

ANOVA is widely used in social science, both for observational and experimental data. In experiments, the grouping of observations is generally defined by the various treatment and control groups established by the experimental design.

Estimating One-Way ANOVA

The estimation of one-way ANOVA focuses on the variability in the variable of interest between the groups as determined by the categorical variable compared to variation in the variable of interest within those groups. In both cases, this variability is expressed as a sum of squares. A sum of squares simply computes some set of numbers, squares those numbers, and then adds them together.

We can start with the total sum of squares to express the total variability in a variable, which is calculated as follows:

(1)

$$SS_{total} = \sum_{i=1}^{N} (Y_i - \bar{Y})^2$$

Where:

- N = the sample size
- Y_i = the individual values of Y for all of the observations in the dataset
- $\bar{Y}$ = the overall, or grand, mean of Y for the entire sample of data.

The degrees of freedom for SS_{total} will equal $N - 1$.

Next we compute the sum of squares between the groups as follows:

(2)

$$SS_{total} = \sum_{i=1}^{N} (Y_i - \bar{Y})^2$$

Where:

- j = the number of groups defined by the categorical variable
- n_i = the sample size for the i^{th} group
- $\bar{Y}_i$ = the mean of the variable in question for the i^{th} group
- $\bar{Y}$ = the overall, or grand, mean of Y for the entire sample of data.

$SS_{between}$ will be relatively large when the individual group means differ greatly from each other, and thus from the grand mean. The degrees of freedom for $SS_{between}$ will equal $j - 1$.

Finally, we can compute the sum of squares within the groups as follows:

(3)

$$SS_{within} = SS_{total} - SS_{between}$$

The degrees of freedom for $SS_{between}$ will equal $N - j$.

Once we have computed $SS_{between}$ and SS_{within}, we need to compute the mean sum of squares for each of them. We do this because the size of the simple sums of squares we have computed thus far will depend in part on how many observations we have. Computing means will compensate for that.

In both cases, the mean sum of squares is computed by just dividing the sum of squares in question by its degrees of freedom.

Finally, the F-statistic we used to determine whether the groups are statistically significantly different from each other is calculated as the mean sum of squares between the groups divided by the mean sum of squares within the groups. This ratio will be relatively large if there is more variance on average in the variable in question between the groups than there is within the groups. The numerator degrees of freedom for this F-test equals $j - 1$ and the denominator degrees of freedom equals $N - j$. The resulting F-statistic is compared to what would be predicted to occur strictly due to random chance, and if the score is larger than expected, the null hypothesis of no difference between the groups is rejected.

Assumptions Behind the Method

Nearly every statistical test relies on some underlying assumptions, and they are all affected by the mix of data you happen to have. Critical considerations for one-way ANOVA include:

- The observations are independent of each other.
- The residuals of the continuous variable in question are (approximately) normally distributed.
- The variances are equal across groups.
- Values of the continuous variable within each group are independent and identically distributed.

These are important assumptions to consider, though we note that ANOVA is known to be somewhat robust to violations of the normality assumption.

Illustrative Example: Fuel Consumption by Fuel Type

This example illustrates one-way ANOVA using two variables from the 2015 Fuel Consumption Report from Natural Resources Canada. Specifically, it examines whether automobile city fuel consumption differs depending on the type of fuel an automobile uses. This is useful because it allows researchers to explore whether different types of fuel result in different rates of city fuel consumption.

Thus this example addresses the following research question:

> Does the average city fuel consumption rate for automobiles in the United States differ across different types of fuel?

This can be stated in the form of a null hypothesis:

> $H0$ = There is no difference in the average city fuel consumption rates of automobiles across different types of fuel.

The Data

This example uses a subset of data from 2015 drawn from the Fuel Consumption Report from Natural Re-

sources Canada. This extract includes data from 1082 automobiles. The two variables we examine are:

- The city fuel consumption rate for each automobile (fuelusecity, measured in liters per 100km).
- The type of fuel each automobile uses (fuel).

The variable fuelusecity ranges between 4.5 and 30.60 in this sample dataset, with a mean of 12.53. The variable fuel divides automobile fuel into four categories:

- Regular Gas
- Premium Gas
- Diesel
- Ethanol-E85

The city fuel consumption rate is a continuous variable and the type of fuel is a categorical variable, making these variables appropriate for a one-way ANOVA.

Analyzing the Data

Before conducting a one-way ANOVA, we should first look at each variable in isolation. We start by presenting a histogram of the city fuel consumption rate in Figure 1. The values are clustered around the mean of 12.10. There are a few extreme values, with the largest being 30.6 liters per 100 km. Researchers may want to explore whether cases with these extreme values have undue influence on the regression.

Figure 1: Histogram showing the distribution of the city fuel consumption rate for automobiles (l/100km), 2015 Fuel Consumption Report, Natural Resources Canada

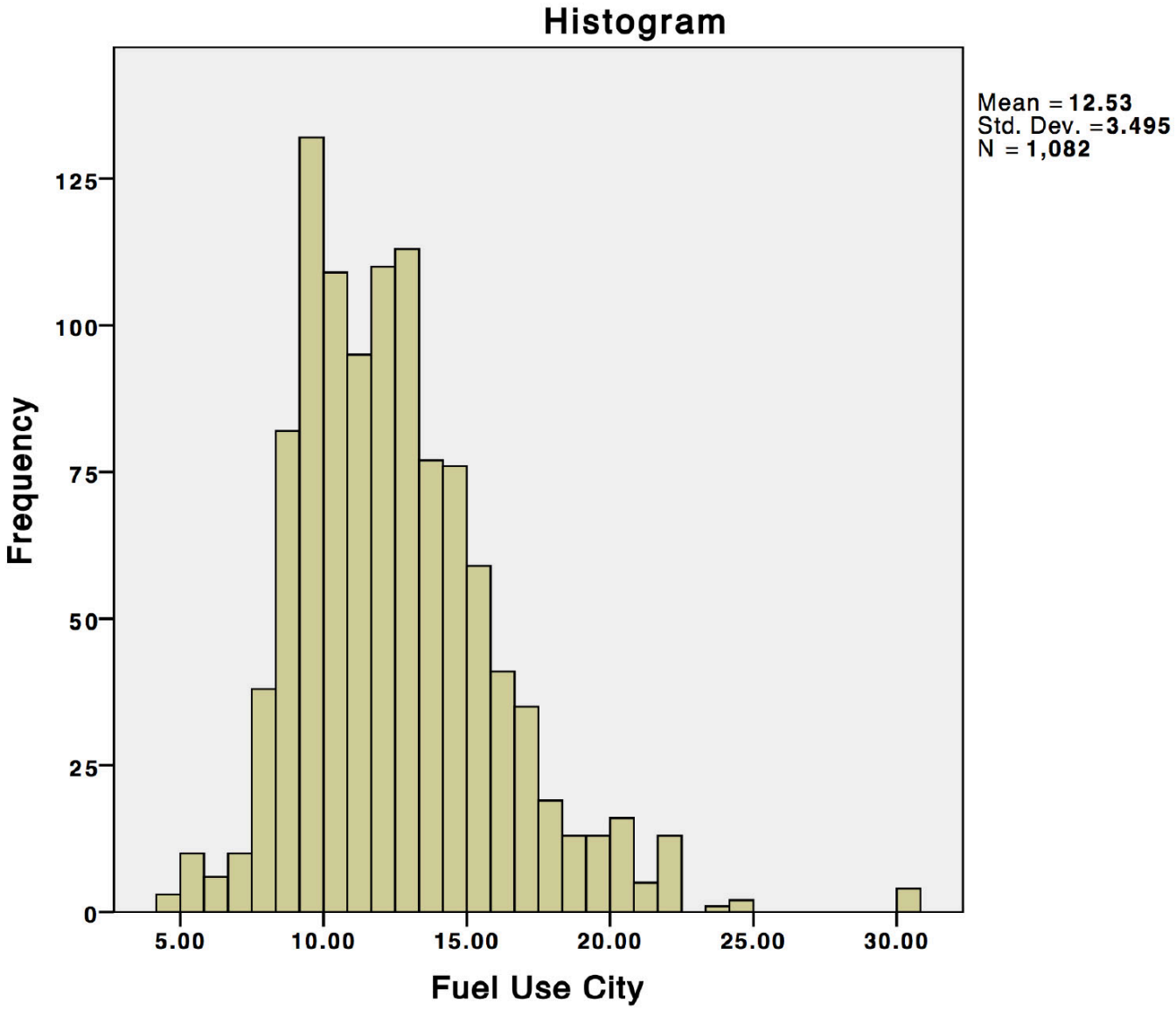

Table 1 presents a frequency distribution of the four types of automobile fuel in the sample. We have more than 400 each of Regular Gas and Premium Gas, respectively, while the remaining two types constitute close to 100 observations in the dataset.

Table 1: Frequency distribution of type of automobile fuel, 2015 Fuel Consumption Report, Natural Resources Canada.

	Frequency	Percent	Cumulative Percent
Regular Gas	518	47.87	47.87
Premium Gas	466	43.07	90.94
Diesel	35	3.23	94.18
Ethanol-E85	63	5.82	100.00
Total	1082	100.0	100.0

Figure 1 and Table 1 show the distribution of each of these variables by themselves. Next we explore whether the two variables are related.

Table 2 presents the results of the one-way ANOVA.

Table 2: Summary of results from one-way ANOVA analysis of differences in city fuel consumption rates by type of fuel used, 2015 Fuel Consumption Report, Natural Resources Canada.

	Sum of Squares	df	Mean Square	F	Sig. of F
Between Groups	3949.58	3	1316.53	153.32	0.000
Within Groups	9256.44	1078	336.5		
Total	13206.03	1081	12.22		

Table 2 reports the Between, Within, and Total sums of squares, along with their associated degrees of freedom. The Mean square for both the Between Groups and Within Groups are the respective sums of squares divided by the reported degrees of freedom. Again, the F-test shown in the table is the Mean Square between

groups divided by the Mean Square within groups. The results shown in Table 2 would lead us to reject the null hypothesis of no difference between the types of fuel in terms of their average city fuel consumption rates. Rather, we conclude that there are statistically significant differences between them. Notice that these results do not tell us how these types of automobile fuels differ or which differences are most important. Additional analysis, such as comparing the mean city fuel consumption rates for each type of fuel, should be carried out to explore those questions.

Presenting the Results

The results of a one-way ANOVA analysis can be presented as follows:

"We used a subset of data from the 2015 Fuel Consumption Report from Natural Resources Canada to test the null hypothesis:

> H_0 = There is no difference in the average city fuel consumption rates of automobiles across different types of fuel.

The data included 1082 automobiles, divided into four types of fuel: Regular Gas, Premium Gas, Diesel, and Ethanol-E85. Table 2 reports a statistically significant test of the null hypothesis, leading us to reject it. We therefore conclude that there are statistically significant differences between the city fuel consumption rates of automobiles with different types of fuel. Further analysis is necessary to determine exactly which differences are most important."

Review

The one-way ANOVA model provides a way to test whether there are differences in the mean of a continuous variable across two or more groups defined by values of a categorical variable. It tests the null hypothesis that there is no difference in the continuous variable across the groups.

You should know:

- What types of variable are suited for a one-way ANOVA.

- The basic assumptions underlying this method.
- How to compute and interpret a one-way ANOVA model.
- How to report the results of a one-way ANOVA model.

Your Turn

You can download this sample dataset along with a guide showing how to produce a one-way ANOVA using statistical software. The sample dataset also includes the variable fuelusehwy, which measures highway fuel consumption rate. See if you can reproduce the results presented here, and try producing your own one-way ANOVA using this other fuel consumption rate measure in place of the city fuel consumption rate used for this example.

https://doi.org/10.4135/9781473947351