**SiO2 and TiO2 Analysis from the Hualalai Volcano, HI**

Report by Jonathon Sevy || Geologist

September 21, 2021

New silicone and titanium data from a Hualalai volcanic flow were compared to data on record from previous flows. Additionally, two deliverable MATLAB scripts were created for statistically comparing any two datasets of the same variable. This program is specifically designed to perform the following: Calculate the most common statistics of the two datasets; plot data in appropriate histograms; Calculate statistics about the shape of the histogram; calculate statistics about the relation between the two sets of data. This report will function as both an analysis on the statistics of the Hualalai flow, and also a description of the used statistics meaning so it can be easily applied to diferent datasets in the future.

Characteristics of the Data Sets

Statistical numbers are a way of numerically describing a collection of data points. These characteristics can give insight into the population the sample data came from, as well as enable the comparison between two or more sets of sample data. Some of the characteristics that can be described are the central tendency, the dispersion, and the shape of the data. The characteristic for our data can be seen in figure 1. The characteristics for central tendency are described in the statistics of the mean, median and mode. The Characteristics of dispersion are described in the standard deviation, and variance. The shape of the data is described by the skewness, and kurtosis.

Hypothesis Testing, and Comparison of the Data Sets

Hypothesis testing is looking at the probability that the population had a certain characteristic, or that two data sets might have come from an identical population. In our case we are looking to if we can state two things: 1) If we can say with 95% confidence that the samples are from a normally distributed population; 2) If we can say with 95% confidence that the samples are from the same, or identical populations. As a note on p-values, which are reported in Figure 2. Hypothesis testing is relying on p-values, and so are important to have a brief understanding of what they mean. P-value is the probability that the measured result of data would come from a population with a given characteristic, or when comparing two datasets that those sets of points could have come from the same (or identical) populations. Because of this, the smaller the p-value is, the less likely that the datapoint or dataset came from the hypothesized population. For example, the p-value for the Kolmogorov-Smirnov test for SiO2 was fairly high at 4.89E-1, leading us to fail to reject the hypothesis that samples come from an identical population. Whereas the same test for TiO2 had an extremely low p-value of 2.19E-23, leading to the rejection of the hypothesis that samples come from an identical population.

To look at the first one, or the probability of the samples coming from a normally distributed population, we use a Lilliefors test. The Lilliefors test for the SiO2 came back as a 1 for both, so we reject that the SiO2 population is normally distributed with 95% confidence. The Lilliefors test for TiO2 came back 0 for the reference data, and 1 for the new data. This means that we rejects that the reference TiO2 data came from a population that was normally distributed, at the 95% confidence level, but we fail to reject the same for the new data.

The second one, or the probability of the sets of samples coming from a the same, or identical population. For this test we use two tests, the two sided t-test, and the two sample Kolmogorov-Smirnov test. The t-test is included because it has historically been used, robust and generally well understood. The t-test has some starting assumptions, making the math simpler. Some of the main assumptions of the t-test is that both samples come from a population that is normally distributed, and that the variance is approximately between the samples. Although, for most purposes the t-test is robust enough to handle if the t-test is not normally distributed so long as it is approximately normally distributed. The Kolmogorov-Smirnov test is more complex and does not make any starting assumptions about the samples. The Kolmogorov-Smirnov test is more adaptable and can handle non-normally distributed samples. because we rejected that the samples were from normally distributed populations in at least one of the datasets for both titanium and silicone we will rely more on the Kolmogorov-Smirnov test.

Conclusions

At the 95% confidence level we can reject that the new and reference SiO2 datasets are from a population that is normally distributed. Similarly, we reject the refence data for TiO2, but fail to reject the new data for TiO2. At a 95% confidence level, we fail to reject the SiO2 levels of the new sample came from the same population as the reference SiO2 data. However, also at the 95% confidence level we reject that the new data TiO2 is from an identical population as the reference data.

The two MATLAB scripts that were used to produce data can be used for comparing two datasets of univariate data. The scripts will produce common statistical characteristics, and statistics for comparing the data to the population. Although these statistics are somewhat simple, there is some care that should be taken to prevent from misinterpreting the meaning of the data.

Figures:

Text, table

Description automatically generatedText, table

Description automatically generated with medium confidence

Figure 1. Image of command line print out of data set characteristics. Left: SiO2 data. Right TiO2 data. This is the output from the first script, run twice for both the SiO2 and TiO2 data. These can be used to understand, discuss the datasets more easily.

Text

Description automatically generated with medium confidence

Text

Description automatically generated with medium confidence

Figure 2. Above: SiO2, Below TiO2. Images of Hypothesis testing for the datasets. These are images of command line printout of the second script, run twice for both TiO2 and SiO2 datasets. Listing both the P-value and test result (0 or 1). Printout below test statistic is meant to help in interpretation.