**Determining Reasons for Discrepancies in Percentage of Valid Sales in Ratio Tests Across Indiana Counties**

 The percent of sales data used in a ratio study can be indicative of the quality of the study. In the state of Indiana, there is a lot of variance from county to county in the percent of sales used in ratio studies. In this report, we investigate possible reasons for why there may be such a discrepancy among counties, including county size and composition of parcels within the county.

**Number of Employees**

 Pearson’s Correlation Coefficient (PCC) was calculated for the correlation between number of full-time employees and the percentage of sales marked as valid for the ratio study and a 95% confidence interval was calculated. The PCC was calculated to be 0.2986 with a 95% confidence interval of (0.0988, 0.4753). This indicates that a positive correlation does exists between these two variables with statistical significance. However, this correlation is estimated to be 0.2986 which is a weak correlation.



 The scatterplot above shows the relationship between these two variables. It appears that the number of full-time employees contains several outliers, so we will perform the same analysis without those outliers. We determine outliers by creating upper and lower fences and excluding those values outside of these fences (the fences are determined using the generally accepted method of extending 1.5 times the interquartile range below and above the first and third quartiles respectively). When excluding these outliers, a PCC of 0.2726 is obtained with a 95% confidence interval of (0.0605, 0.4612). These results are similar to above and would be interpreted in the same way.



 Above shows the scatterplot with the outliers removed and a line of best fit. While this does indicate that a larger number of full-time employees means a larger percentage of sales will be included in the ratio study, it is important to note that a significant number of counties with four or fewer employees are able to include greater than 60% of their sales in their ratio study. This decreasing variation seems to indicate that a larger full-time staff may mitigate issues with underrepresentation of sales data in ratio studies, but also a smaller full-time staff does not seem to doom a county to low percentage of sales used.

Similar results occur when number of part-time employees is considered so those are omitted.

**Total Number of Parcels**



 Above is the scatter plot comparing the total number of parcels to the percent of sales used in the ratio test. The PCC was computed to be 0.3291 with a 95% confidence interval of (0.1321, 0.5011), which indicates a weak positive correlation with statistical significance at the 95% confidence level. However, it is clear that much of this correlation is driven by several larger parcels. If we consider only the counties is 50,000 or fewer parcels, the PCC is 0.1600 with a confidence interval of (-0.0728, 0.3762) which does not indicate a correlation at the 95% confidence level. It is important to note that the cut-off of 50,000 parcels includes 74 of the 92 Indiana counties.

**Population**

Next, we will compare the population of each county to the percent of sales used in the ratio test. The PCC was calculated to be 0.3361 with a 95% confidence interval of (0.1398, 0.5069). Once again this is statistically significant, but the correlation is minor. The interpretation is similar to that of parcel count.

**Real Property Gross Assessed Value**

Now, we will compare the gross assessed value of real property to the percent of sales used. This time the PCC was calculated as 0.3711 with a 95% confidence interval of (0.1788, 0.5361), which is interpreted in the exact same way as population. The results for population, gross assessed value, parcel count, and number of full-time employees are all very similar since these measures are highly correlated.

 We can consider all of these to be different measures of county size. The above results indicate that larger counties tend to include a larger percent of sales in their ratio studies. They also show that the variance in percent of sales used increases as counties get smaller.



**Compared to Previous Years**

 The above results may indicate that smaller counties have more variance in the percent of sales used in the ratio study from year to year. To investigate this possibility, we will look at the comparison between the percent used in the ratio study and the average of the percent used in the ratio study the prior 4 years. The data we have contains the 5-year average percent of sales used for each county. In order make this an accurate comparison, we need to exclude the current years percent of sales used, since this number is used in computing the average (simply multiply by 5, subtract this year’s average, and divide by 4).



 The PCC was calculated to be 0.6577 with a 95% confidence interval of (0.5226, 0.7607). This means that there is a moderate to strong correlation between the percent of sales used in the four previous years and the percent of sales used in the next year’s ratio study. This can be seen in the scatterplot shown above. This significant correlation indicates that many counties that use a small percentage of their sales do so in a consistent manner. The PCC was also calculated using only those counties with fewer than 60,000 parcels (the smallest 74 counties). The value was calculated as 0.6100 with a 95% confidence interval of (0.4471, 0.7337), which is similar to the above values calculated for all counties. This indicates that it is unlikely that there is simply more variation in the percentage of sales used in smaller counties as this would be indicated by low or no correlation.

**Percent of Agricultural Parcels**

 The PCC between percentage of agricultural parcels and total percent of sales used in the ratio study is calculated as -0.4664 with a 95% confidence interval of (-0.6134, -0.2881). This is a weak to moderate negative correlation that is statistically significant. This means that there may be some evidence that the larger percent of agricultural parcels in a county, the smaller the percent of sales used in the ratio study.

 To explore and quantify this, a simple linear regression model is fit. The model is as follows:

$$\hat{y}=67.26 – 69.39 x\_{ag}$$

However, this model has an R2 value of 0.2175. This is interpreted to mean that percent of agricultural parcels explains 21.75% of the variation in the percent of sales data used in ratio studies. If we still assume that percent of agricultural parcels is a good indicator of the percent of sales used in ratio studies, it is of interest to see the 95% confidence interval about this regression line. Shown below is the scatterplot with the fitted regression line in red and the upper and lower ends of the 95% confidence interval in black. Many of the counties with the lowest percent of sales used are not in the 95% confidence interval for the regression line.

 Overall, percent of agricultural parcels is not a good predictor for percent of sales used in ratio studies. It is also of note that there are many counties with more than 25% of their parcels as agricultural that utilize a significant portion of their sales data. This seems to indicate that there is still another underlying factor leading to smaller percentage of data used in ratio studies.



**Contract vs No Contract**

 Next, we investigate if there is a statistically significant difference between the mean percent of sales used in ratio studies for counties that did use a contract and counties that did not. The mean percentage of sales used in ratio studies for counties that did not use a contractor is 56.5267% compared to 49.8724% for counties that did use a contractor. In order to see if this difference is significant, we will perform a statistical test for the difference of means. The procedure by which this is done will be a nonparametric randomization test, which mitigates issues with nonnormality of the data as well as other assumptions that may be violated with a traditional t-test. The p-value obtained from the test was 0.0695. At the traditional 95% confidence level we fail to reject that these means are different. However, it is important to note that this is still a rather small p-value, and if a confidence level of 90% was chosen (which is another common confidence level), we would reject the null hypothesis and conclude that these means are different.



The boxplot above helps use visualize the difference in percent of sales used in ratio studies between counties that did and did not use a contractor. One point of interest here is the minimum value for counties that did not use a contractor is quite higher than for those who did use a contractor, suggesting that using a contractor increases the odds of using a smaller percent of sales in the ratio study.

 Next, we will see if controlling for the use of a contract allows us to find a better model for percent of sales used. To do this, we will subset the data into counties that use a contractor and those that do not then fit a linear model on these subsets using the percent of agricultural parcels. For the counties that use a contractor, the linear model has an R2 value of 16.82%, which means that the percent of agricultural parcels explains 16.82% of the variation in the percent of sales used in ratio studies among counties that use a contract. This is compared to the model fit among those counties that did not use a contract having an R2 value of 48.74%, which means 48.74% of the variation in percent of sales used can be explained by the percent of agricultural parcels. There are only 15 counties in Indiana that do not use a contractor. This means that 15 out of 92 counties seem to be driving most of the correlation between percent of sales used and percent of agricultural parcels.

**Summary:** It appears that there is a correlation between measures of county size and the percent of sales used in ratio studies, however that correlation is mostly driven by the largest 10 or so counties having high percentage of sales used. Among the smaller counties, the percent of sales used varies wildly. This could be simply due to higher variance of valid sales in these counties, but one would expect that to vary year to year within each county. Thus, the strong correlation between the most recent year’s percent of sales used and the average percent of sales used from the previous 4 years suggests this is not the case. So, it appears that there is some underlying reason that some counties have fewer percent of sales used in their ratio study. To see if percent of agricultural counties accounts for this variation, we consider a linear model to fit the percent of sales used in ratio studies. However, this model accounts for only approximately 21.75% of the variation in percent of sales used in ratio studies, indicating there is still some underlying factor leading some counties to use a smaller percent of sales in ratio studies. We also investigated the difference between having a contract vs not having a contract. Although not statistically significant, there is a greater spread among the counties that used a contractor. When controlling for if a county uses a contractor, the linear model for counties that use a contractor is much worse than for counties that do not, further supporting that there is some other underlying factor besides agricultural parcel percentage that is influencing the vast difference in percent of sales used in ratio studies, especially for counties that use a contractor.