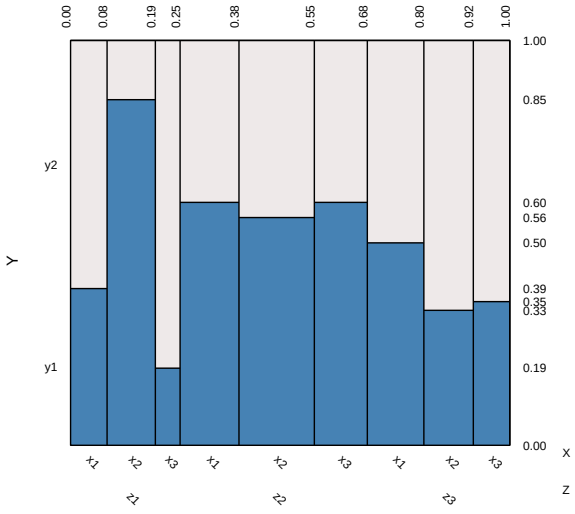
Using Modified Eikosograms to Describe Dependence

Steven Hobbs

December 20, 2013

Eikosograms are a great tool to analyze conditional probability. In this essay, I discuss how we can use eikosograms to analyze conditional independence. Given a k dimensional contingency table, we can test various hypotheses of conditional independence by using a chi-squared test. In a typical chi-squared test for independence we would first assume some form of independence or conditional independence as a null hypothesis. Then, in each cell of the contingency table we would have an expected frequency $E_{ij\ldots k}$ under the null hypothesis, and an observed frequency $O_{ij\ldots k}$. From this we can compute the Pearson residual $r_{ij\ldots k} = \frac{O_{ij\ldots k} - E_{ij\ldots k}}{\sqrt{E_{ij\ldots k}}}$ associated with each cell and the test statistic $D = \sum_{\text{all } ij\ldots k} r^2_{ij\ldots k}$, where D follows a chi-squared distribution under the null hypothesis. Then we compute a p-value and make inference. One of the shortcomings of the chi-squared test is that we lose information when we sum the squared Pearson residuals since we aren't looking at the individual residuals anymore. Perhaps the null hypothesis is only violated on a subset of the data. This partial independence can be captured using eikosograms.

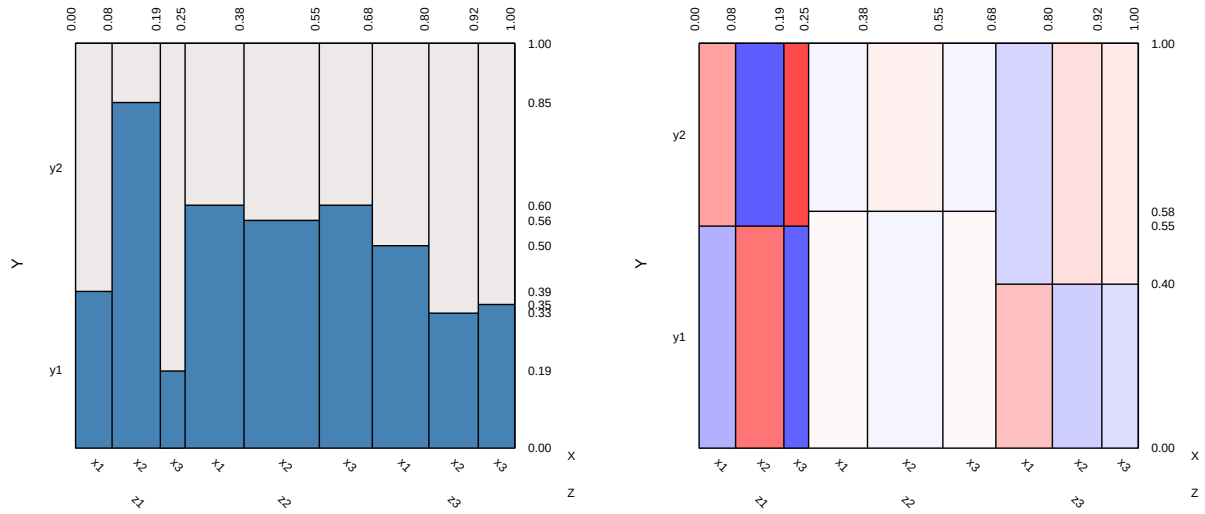
For example, suppose we have variables X, Y, and Z that take on values $\{x_1, x_2, x_3\}, \{y_1, y_2\}$, and $\{z_1, z_2, z_3\}$ respectively. And we want to test if X and Y are conditionally independent given Z. We can plot the following eikosogram to see the conditional independence.



We can see that conditional on $Z = z_1$, X and Y are far from independent since the three leftmost blue rectangles have very different heights. Similarly conditional on $Z = z_2, X$ and Y are close to independent, and conditional on $Z = z_3$, X and Y are somewhere in between. So we have a data set that appears to not satisfy $X \perp\!\!\!\perp Y|Z$, but none the less has some conditional independence for a subset of the data.

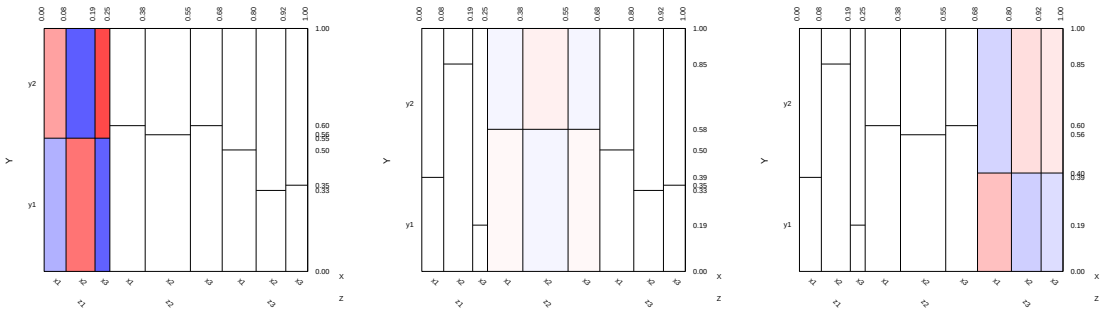
In this essay, Im going to discuss another way to communicate the conditional independencies in a data set. Suppose that instead of presenting the eikosogram of the data, we present the eikosogram of the expected data

given the null hypothesis and color the rectangles based on how different the actual data is from what is expected.The plot below compares these two ways of presenting the conditional independence of the same data.



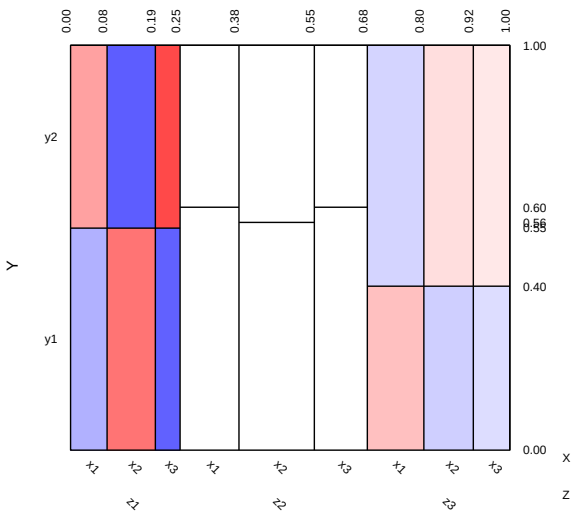
In the plot on the right, the dark colours in the left third of the plot indicates strong evidence that the assumption that $X \perp\!\!\!\perp Y|Z$ has been violated when $Z = z_1$. The faint shading in the middle suggests there is little evidence against the null hypothesis when $Z = z_2$, and the light shading on the right indicates there is some evidence against the null hypothesis when $Z = z_3$. The colour saturation of each rectangle in determined by the Pearson residual associated with that rectangle. For a given cell in the data, blue indicates that the observed frequency is less than expected under the null hypothesis (negative residual), and red indicates that the observed frequency is greater than expected (positive residual). Viewing this type of plot, we can very quickly see where the data seems to violate the null hypothesis by comparing the colour saturations in each rectangle.

These plots can be created for various null hypotheses of conditional independence. In the plots below, the left plot corresponds to testing $X \perp\!\!\!\perp Y|Z = z_1$, the middle plot is testing $X \perp\!\!\!\perp Y|Z = z_2$, and the rightmost plot is testing $X \perp\!\!\!\perp Y|Z = z_3$.



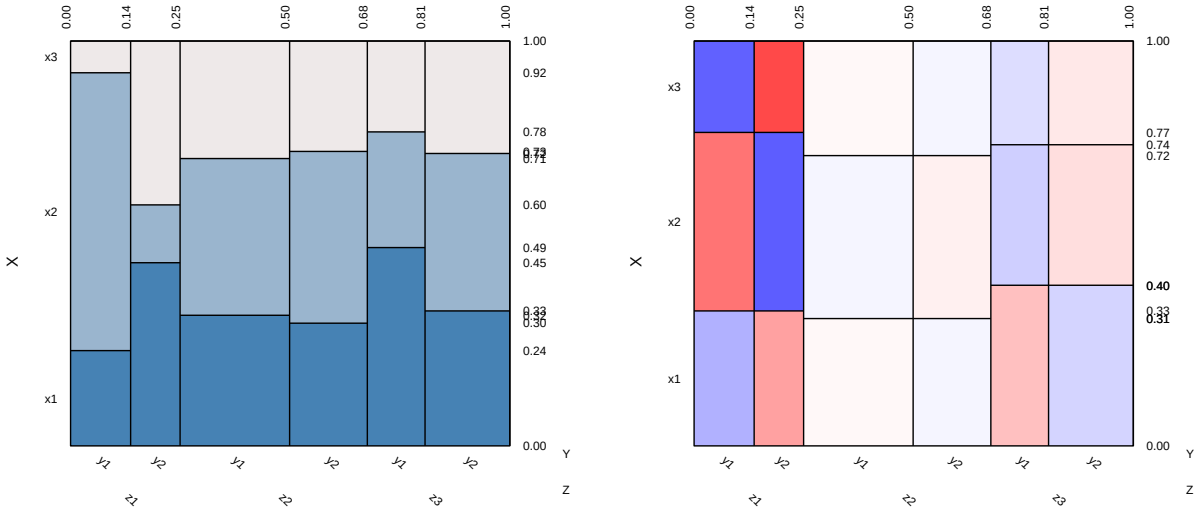
Observe that two thirds of each plot is white. This is because the null hypotheses say nothing about conditional independence of two of the three values of Z. Also notice how the location of the horizontal line and colours in the plots conveys the information of what null hypothesis is being tested.

We can create these eikosograms to test $X \perp\!\!\!\perp Y|Z \in S$ for any subset S of the possible values of Z. for example, the plot below presents the expected data given $X \perp\!\!\!\perp Y|Z \in \{z_1, z_3\}$.



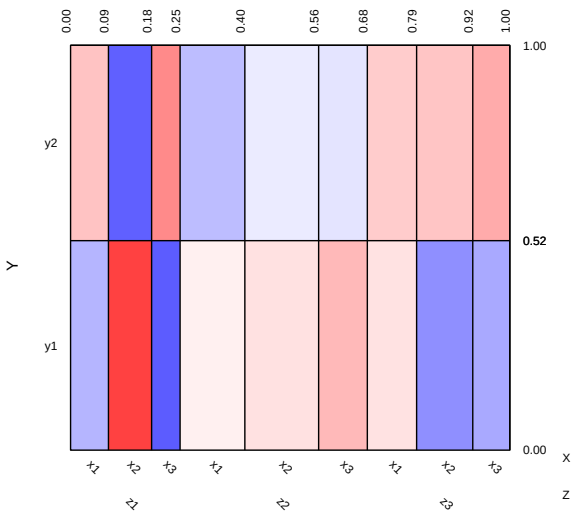
The fact that there are horizontal lines in the $Z = z_1$ and $Z = z_3$ rectangles, but not the $Z = z_2$ rectangle tells us that the purpose of this plot is to determine if $X \perp\!\!\!\perp Y|Z \in \{z_1, z_3\}$.

Switching the axes, so that X is on the vertical axis, we can again test $X \perp\!\!\!\perp Y|Z$.



As we would expect we get the same results from the eikosogram. When $Z = z_1$ there is a lot of heavy saturated colours, so there is strong evidence against test $X \perp\!\!\!\perp Y|Z = z_1$. Similarly we also see some evidence against the hypotheses that $X \perp\!\!\!\perp Y|Z = z_3$ and almost no evidence against $X \perp\!\!\!\perp Y|Z = z_2$. Comparing this to the eikosogram of the original data on the left also gives a consistent interpretation of the conditional independence relationship of the data.

We can also use these type of plots to test complete independence. That is, to test $X \perp\!\!\!\perp Y \perp\!\!\!\perp Z$. We can see from the plot below that there is strong evidence against this hyposthesis.



Earlier I mentioned that the colour of each rectangle is determined by its associated Pearson residual. But how exactly should the colour be chosen? We want a function that takes a real number and outputs a colour. We want large residuals to correspond to saturated colours, and small residuals to be close to white. The way that I decided to do this was to generate a discrete spectrum of a large number of colours from blue to white to red. Next, I created a bunch of intervals based on the quantiles of a standard normal distribution and assigned a colour to each interval. I used the quantiles of a standard normal distribution because Pearson residuals are normally distributed.