FINAL LAW OF STATISTICAL RELATIONSHIPS OF CATEGORICAL VARIABLES

ZULFIKAR MOINUDDIN AHMED

1. On the Sufferings of Man

Ever since Stone Age, or more formally the Pleistocene Era, Man has suffered. Man could count, but did not know the relationship between any two counted variables statistically. The Advent of Space Age had no effect and Man remained confused on the Moon. The Landing of the Moon had to rely only on correlations between continuous variables, because for categorical variables Man simply did not have any valid statistical measure of relations between variables.

Only on April 28 2021, I, Zulfikar Moinuddin Ahmed, finally realised that one could indeed produce a valid measurement of relationship between categorical variables X and Y that is not a totally meaningless fudge.

You see, correlations between categorical or factor variables are meaningless. This did not prevent yours truly from computing correlations with abandon and bamboozling the gullible Human Race with results based on such meaningless numbers. Well, I have an excuse. I simply did not realise that there was a serious problem at all. I only realised there was a serious problem in the last 24 hours when I did not like Karl Pearson's chi-square independence test and realised that I needed to invent something that is sensible and correct. I did that and suddenly I realised that Finally, the great Darkness and Ignorance that had plagued Man since Pleistocene regarding categorical variable relations has passed. Enlightenment had arrived through necessity of making any sense of the World Values Survey relationship between variables.

2. Re-Iteration of the Replacement for Correlation

We are going to be using the p-value of the Zulf Chisquared Independence test to determine a measurement of independence and dependence of categorical variables $X \in \{x_1, \ldots, x_m\}$ and $Y \in \{y_1, \ldots, y_n\}$.

3. Crux of the Solution

With categorical or factor variables, the statistical data are frequencies. With notation from the last section, one draws a rectangular array of $m \times n$ cells and just counts the occurrences of the various combinations of X and Y as random variables. The data then are either row-by-row probabilities q_k^j for $1 \leq j \leq m$ and $1 \leq k \leq n$ with each row summing to one, or each column summing to one. These are the marginal distributions; or one considers the joint which is just normalizing the counts for total sum 1.

Date: April 28, 2021.

4. Our Approach to a Valid Correlation Replacement

Categorical variables do not have any sense of correlations at all. I, like many others, have been quite cavalier in the past and just recoded the categories to numerical variables in some ad hoc manner to gain information about the correlation. This can produce useful results but it is invalid and meaningless and unprincipled.

The key point of a valid measure of correlation is the path from the finite set probability simplex

$$\Sigma_n = \{(p_1, \ldots, p_n) \in \mathbf{R}^n : \sum_j p_j = 1, 0 \leq p_j \leq n\}$$

The transformation

$$g(p) = \log(-\log(p))$$

is what allows the correct mapping of $I = (0, 1)$ to $\mathbf{R}$. Then we consider Gaussian on $\mathbf{R}$ to bring in the chi-square distribution in a valid manner. I think that this issue of when and why the Chi-Square distributions have any relevance had simply not been resolved previously in a correct manner.

Recall that chi-square distributions are just the distribution of $S_m = X_1^2 + \cdots + X_m^2$ with $X_\ell \sim N(0, 1)$. You cannot bring in chi-square without explaining what is going to be normal in the problem, and my own impression has been that all previous statistical work has just not given any clear explanation of why chi-square distributions can be justified in categorical variable problems at all. Karl Pearson's chi-square independence test is not sensible or justified. With the g transformation we are able to justify normal distributions.

5. The p-value of Null Hypothesis of Independence is a Correlation Replacement

We consider the test that all rows of probabilities are equal. This is equivalent to pure probabilistic independence. The p-value being closer to zero then is then the replacement for correlation.

6. Syntactic Sugar

For a measure that behaves more like correlation, we can *define* discrete correlation as

$$C_d(X, Y) = 1 - p(X, Y)$$

where $p(X, Y)$ is the p-value of the independence test. This gives 1 when p-value is very small.

7. Code

```
g<-function(x){
  print(x)
  log(-log(x)+0.01)
}

zulf.sigma<-function(z){
  n<-length(z)
  D<-matrix(0, nrow=n, ncol=n)
```

```
  print(dim(D))
  s<-0
  for (j in 2:n){
    for (k in 1:(n-1)){
      gg = abs(z[j] - z[k])
      D[j,k] <- gg
      D[k,j] <- gg
      s<-s + gg
    }
  }
  out<- 2*s/(n^2-n)
  out
}

zulf.chisq<-function( data ){
  m<-dim(data)[1]
  n<-dim(data)[2]
  t<-0
  eps<-0.00001
  w0 <- (data[1,]+eps)/sum(data[1,]+eps)
  print('this')
  print(length(w0))

  v0<- log(-log(w0+eps))
  print(v0)
  sigma0 <- zulf.sigma( v0 )
  mu0 <- mean(v0)
  for (j in 2:m){
    w<-(data[j,]+eps)/sum(data[j,]+eps)
    print(w)
    v<-g(w)
    print('works')
    print(v)
    print('---')
    sigma <- zulf.sigma( v )
    mu <- mean(v)
    z <- v/sigma
    z0 <- v0/sigma0
    t<- t + sum( ( z - z0 )^2)
  }
  df<-(n-1)*(m-1)
  t0<-qchisq(0.95, df=df)
  pval <- 1 - pchisq( t, df=df)
  list(tstat=t,pval=pval,crit=t0)
}
```