1 Chi-squared test

See `https://en.wikipedia.org/wiki/Chi-squared_test`

Without other qualification, 'chi-squared test' often is used as short for Pearson's chi-squared test. The chi-squared test is used to determine whether there is a significant difference between the expected frequencies and the observed frequencies in one or more categories.

Example chi-squared test for categorical data:

Suppose there is a city of 1,000,000 residents with four neighborhoods: A, B, C, and D. A random sample of 650 residents of the city is taken and their occupation is recorded as "white collar", "blue collar", or "no collar". The null hypothesis is that each person's neighborhood of residence is independent of the person's occupational classification. The data are tabulated as follows:

```
import numpy as np
import pandas as pd
import scipy.stats as stats

cols = ['A', 'B', 'C', 'D']
data = pd.DataFrame(columns=cols)

data.loc['White Collar'] = [90, 60, 104, 95]
data.loc['Blue Collar'] = [30, 50, 51, 20]
data.loc['No collar'] = [30, 40, 45, 35]

print (data)

OUT:

               A   B    C   D
White Collar  90  60  104  95
Blue Collar   30  50   51  20
No collar     30  40   45  35
```

We can use the chi-squared test whether area effects the numbers of each collar type. That is, are the values in each row affected by the column? We can reported an expected distribution if column does not effect the values in each row (other than each column having a different total).

```
V, p, dof, expected = stats.chi2_contingency(data)
# add correction=False for uncorrected Chi-square

print ('P value for effect of area on proportion of each collar:')
print (p)
print ('\nExpected numbers if area did not effect proportion of each collar:')
print (expected)

OUT:

P value for effect of area on proportion of each collar:
0.0004098425861096696

Expected numbers if area did not effect proportion of each collar:
[[ 80.53846154  80.53846154 107.38461538  80.53846154]
 [ 34.84615385  34.84615385  46.46153846  34.84615385]
 [ 34.61538462  34.61538462  46.15384615  34.61538462]]
```

Note. In Chi-square at least 80% of the of the cells should have a value of at least 5, and all cells where values are expected should be at least 1. If this is not the case then use Fisher exact test.