Hypothesis Testing with Spark RDD of Labeled Point:

```
//an RDD of labeled points
            JavaRDD<LabeledPoint> labelPt = jsc.parallelize(
                  Arrays.asList(
                        new LabeledPoint(1.0,  Vectors.dense(10.0,1.0)),
                        new LabeledPoint(1.0,  Vectors.dense(1.0,2.0)),
                        new LabeledPoint(-1.0,  Vectors.dense(-1.0, 0.0))
                              )

                        );

            // The contingency table is constructed from the raw (feature, label)
pairs and used to conduct
            // the independence test. Returns an array containing the
ChiSquaredTestResult for every feature
            // against the label.
            ChiSqTestResult[] featureTestResults =
Statistics.chiSqTest(labelPt.rdd());
            int i = 1;
            for (ChiSqTestResult result : featureTestResults) {
              System.out.println("Column " + i + ":");
              System.out.println(result + "\n");  // summary of the test
              i++;
            }
```

Output:

```
Column 1:
Chi squared test summary:
method: pearson
degrees of freedom = 2
statistic = 3.0
pValue = 0.22313016014843035
No presumption against null hypothesis: the occurrence of the outcomes is statistically
independent..

Column 2:
Chi squared test summary:
method: pearson
degrees of freedom = 2
statistic = 3.0
pValue = 0.22313016014843035
No presumption against null hypothesis: the occurrence of the outcomes is statistically
independent..
```

Explanation:

Here we will see how to create the contingency table manually if we have the data like the above format i.e Labeled Point and then it is feed to calculate the chisquare test for every column in the Labeled point.

Step 1:

The given input is as follows:

Label	Features	
L	F1	F2
	C0	C1
1.0	10.0	1.0
1.0	1.0	2.0
-1.0	-1.0	0.0

Step 2:
Calculate the number of distinct labels and give index to it. Here we have only two distinct labels. i.e 1.0 , -1.0

(1.0, 0), (-1.0, 1).

This deceides the number of column we will have in our contingency matrix.Here two distinct features so our contingency matrix will have two columns.

Here 0 and 1 are the index we gave.

Step3:
Calculate the number of distinct features form each column and give the index.

For column 0:
(10.0, 0), (1.0, 1), (-1.0, 2)

Here three distinct features so our contingency matrix for column 0 feature will have three rows.

For coulmn 1:
(1.0, 0), (2.0, 1), (0.0, 2)
Here three distinct features so our contingency matrix for column 1 feature will have three rows.

Step 4:
Calculate the feature, label pair for each column. i.e (column,feature,label)
Here we have two columns column 0 and column 1.

For column 0:
(0, 10.0, 1.0), (0, 1.0, 1.0), (0, -1.0, -1.0)

For column 1:
(1, 1.0, 1.0), (1, 2.0, 1.0), (1, 0.0, -1.0)

Step 5:
Calculate the pair count from step 3.

For column 0:
(0, 10.0, 1.0) -> 1, (0, 1.0, 1.0) -> 1, (0, -1.0, -1.0) -> 1

For column 1:
(1, 1.0, 1.0)-> 1, (1, 2.0, 1.0)-> 1, (1, 0.0, -1.0)-> 1

Step6:
Creating contingency matrix for column 0:
Three distinct features and two distinct labels so our contingency matrix for column 0 feature will have three rows and two columns.

	C0	**C1**
R0	**1**	**0**
R1	**1**	**0**

R2 0 1

The above matrix is our contingency matrix for column 0 features. Here is how it is calculated:

Calculation of R0,C0 value:
Row is for feature and column is for label

-> We have to take the detatils from step 2 and step 3 and put it in step 5 to find out the values.
-> In step 2 for index 0 (since we want to find value for C0) value is 1.0 and in step 3 for index 0 (since we want to find value for R0)value is 10.0.So we have to take this 10.0 and 1.0 value from these steps and compare it in step 5 column 0 to calculate the count.

i.e (0,10.0,1.0) -> 1

So the value of R0,C0 is 1.

similarly we have to calculate for R0,C1 ; R1,C0 ; R1,C1 ; R2,C0 ; R2,C1.

Step7:

Once it is done we could use the above contingency matrix to find out the ChiSquare Test result.