**STAT452: Lab 3**

**GoF Tests**

Terence Henriod

STAT 452

2/21/2013

**Introduction**

It is important to know how well data from a process matches the distribution or function we are using to make predictions about that process with. For this reason, we use Goodness of Fit (GoF) tests to evaluate whether or not it is feasible to apply a certain probability distribution to data, or to verify if it is likely that data came from a certain distribution. The Hypotheses we are testing are the following:

.

[1]

One of the most common methods for evaluating GoF is the Chi-Squared test. The Chi-Square statistic is computed by comparing observed cell counts with counts that would be expected under the null hypothesis, and summing the absolute differences, represented by the following equation:

[2]

where O is the Observed cell count (the actual data), and E is the Expected cell count (as predicted under the null hypothesis. The only parameter of the Chi-Square statistic is Degrees of Freedom (df), which represents the number of cells that are free to vary. Degrees of Freedom are found by multiplying the number of levels for each variable minus 1:

for one variable, for two, and so on… [3]

Other GoF tests of interest might include the Anderson-Darling, Kolmogorov-Smirnov, or Shapiro-Wilk tests, but are mostly used for testing Normality, although the Anderson-Darling test can be used to test many distributions.

The data to be used for the following exercises was provided by the instructor and features data generated from Uniform, Exponential, and Normal distributions. The objective is to apply GoF tests to these data, and interpret the results in terms of whether or not the data actually came from these distributions.

**Assignments**

1. Chi-square test for a uniform [0,1] sample in column C1. Perform the chi-square test with 3, 5, and 10 categories. Find the Pvalues, compare, discuss.

|  |  |  |
| --- | --- | --- |
| χ2 Test for the Uniform[0, 1] Sample in C1 | | |
| # of Categories (bins) Used | df | P-Value |
| 3 | 2 | 0.827 |
| 5 | 4 | 0.645 |
| 10 | 9 | 0.760 |

Table 1: A Chi-Square test on Uniform[0, 1] data using varied numbers of bins. As can be seen, the results fluctuate based on the number of bins (or degrees of freedom) used.

The results displayed in Table 1 highlight one of the drawbacks of using the Chi-Square GoF test: results of the test may vary, depending on the way the data is binned. For this task, although a high variation in P-Value is observed, it should also be noted that the conclusions that might be drawn from using these P-Values do not differ, we fail to reject the null hypothesis that the data came from the U[0, 1] distribution. However, this may be the difference between failure to reject and rejection with other data/distributions, possibly with a more significant outcome, which is why one should be aware of this phenomenon when applying the Chi-Square GoF test.

1. Anderson-Darling test for two exponential samples in columns C2 and C3. (Use the probability plot option, which shows the AD test results.) The two samples are from the exponential distribution with mean 5; C2 has length 100, while C3 has length 1000. For each sample, perform the AD test several times for different means of the theoretical distribution and find the limits of the mean corresponding to the P-value of above 5%. Compare results, discuss.

|  |  |  |  |
| --- | --- | --- | --- |
| Anderson-Darling Test for Two Exponential(5) Samples | | | |
| Sample | n | Mean (λ) | P-Value |
| C2 | 100 | 5 | > 0.250 |
| 6 | > 0.250 |
| 6.5 | 0.069 |
| 6.6 | 0.048 |
| 7 | 0.013 |
| C3 | 1000 | 5 | > 0.250 |
| 5.25 | > 0.250 |
| 5.4 | 0.053 |
| 5.5 | 0.014 |
| 6 | < 0.001 |

Table 2: Displays the results of various trials using the Anderson-Darling GoF test (implicit in Minitab’s Probability Plot) to see how well Exponential(2) samples fit Exponential distributions of varying means.

A theoretical distribution mean of slightly less than 6.6 would satisfy the condition outlined in this task if we use the data with only 100 observations, but when we use the 1000 observation sample, we are restricted to a theoretical mean of slightly more than 5.4.

As with most statistical processes, a higher number of observations decreased our margin of freedom, allowing for more precise results to report. Knowing that the data in both cases came from an Exponential(5) distribution, it is clear that with more observations we are limited in our freedom in determining what sort of distribution the data come from. Suppose that we cared more about the type of distribution the data come from than the exact parameters, then we may reject that the data even comes from the distribution of interest if we fail to perform enough trials with different parameter values, if we use too much data. A different GoF method besides the Chi-Square should be applied as well to prevent this.

1. Chi-square test for a Normal(10, 4) sample in column C4. Perform the chi-square test for the normal sample, using the cdf transformation that produces a uniform sample. Compare the chisquare P-value with that of KS, AD, and SW tests. Discuss.

|  |  |  |
| --- | --- | --- |
| GoF Testing for the Normal(10, 4) sample in C4 | | |
| Test Used | Test Statistic | P-Value |
| χ2 with df = 9 | 5.8 | 0.760 |
| Kolmogorov-Smirnov | 0.067 | > 0.150 |
| Anderson-Darling | 0.333 | 0.506 |
| Shapiro-Wilk | 0.993 | > 0.100 |

Table 3: Displays the results of various GoF tests to see if the provided data from the Normal(10, 4) distribution could be considered to reasonably belong to said distribution.



Figure 1: A probability plot of the Normal(10, 4) data, along with the Kolmogorov-Smirnov GoF test results. 25, 50, and 75 percentile lines are included for reference.



Figure : A probability plot of the Normal(10, 4) data, along with the Anderson-Darling GoF test results. 25, 50, and 75 percentile lines are included for reference.



Figure 3: A probability plot of the Normal(10, 4) data, along with the Ryan-Joiner (Minitab’s substitute for the Shapiro-Wilk) GoF test results. 25, 50, and 75 percentile lines are included for reference.

As can be seen from the probability plots, intuitively the data is normal, and the computed statistics for the data are very close to that of the parameters of the hypothesized Normal(10, 4) distribution. All of the tests are in agreement that the null hypothesis should not be rejected, with all of them having P-Values that indicate that the null hypothesis cannot be rejected with any reasonable level of confidence. It is unfortunate that Minitab does not display the higher P-Values that resulted with the Kolmogorov-Smirnov and Shapiro-Wilk Tests, but the Chi-Square and Anderson-Darling tests’ P-Values are in good agreement of one another.

1. Chi-square test for three samples in columns C7-C9. One of these samples in from the U[0,1] distribution, second significantly deviates from the U[0,1], and the third is overfit to the U[0,1]. Use the chi-square test to decide which sample is which.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| χ2 Test for the Uniform[0,1] Samples in C7-C9 | | | | |
| Sample | df | χ2 Statistic | P-Value | Assertion (which sample it is) |
| C7 | 9 | 25.8 | 0.002 | Sample not from U[0, 1] |
| C8 | 9 | 6.6 | 0.679 | Typical U[0, 1] Sample |
| C9 | 9 | 1.8 | 0.994 | Over-fit Sample |

Table 4: The results of using the Chi-Square test to provide P-Values, used to determine which set of data belonged to which classification.

The Chi-Square test allows us to see which samples likely came from the U[0, 1] distribution, and which does not. The C8 and C9 samples have high P-Values, meaning that they very likely fit the distribution, while the C7 sample has a very low P-Value, allowing us to reject the null hypothesis that the C7 data belongs to the U[0, 1] distribution with a high level of confidence.

However, in most cases, the P-Value corresponding to the C9 data would warrant rejection of the null hypothesis as well, because P-Value is so high, indicating that the data is over-fit. When over-fitting has occurred, we suspect the data is not truly random because it matches our expectations too perfectly, and therefore is not valid data to perform analysis on. In this case, though, we were expecting an over-fit sample, so the high P-Value is our indicator that C9 is our over-fit sample.

There is little to say concerning the C8 sample, it has a moderate P-Value, far from warranting rejection, but far from raising over-fit suspicion, so it likely is a typical U[0, 1] sample.

**Conclusion**

In summary, these exercises were practice in applying and interpreting Goodness of Fit (GoF) tests. In general, a GoF test produces a P-Value which indicates the likelihood of observing a similar or greater statistic value under the null hypothesis; in the case of GoF tests, this hypothesis is that the data come from a certain probability distribution.

The Chi-Square is one of, if not the, most commonly applied GoF tests, and is easy to use, but results may vary depending on the data is binned.

Anderson-Darling, Kolmogorov-Smirnov, and Shapiro-Wilk are also GoF tests that can be applied to data, usually to supposedly Normally distributed data.

For the first task, the variability of the Chi-Square test was observed by binning data in different ways.

In the second, through trial and error, it was observed how much freedom, so to speak, one has in using the Anderson-Darling test when using data containing different numbers of observations.

The third task involved applying all four of the discussed GoF tests to a supposedly normal sample; all four tests agreed in that the null hypothesis could not be rejected.

Finally, using the Chi-Square test, P-Values were produced for three U[0, 1] samples, and used to determine which sample was which, with one being significantly deviant from the distribution, one fitting normally, and one being over-fit, in that respective order.