Dr. Sudhakar Raju

BIA 6309: STATISTICS & MACHINE LEARNING

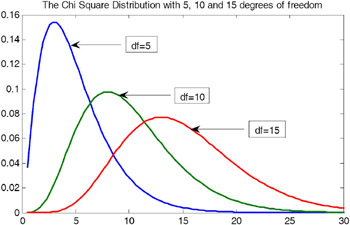
SPRING 2018

**ANSWERS TO ASSIGNMENT 7**

I. The Normal (Z) distribution is used for population distributions. It can also be used for sampling distributions provided the sample size is over 30. The normal distribution can be used for smaller samples IF the underlying distribution from which the samples are drawn are known to be normal. But usually we do not know this – thus, one is best off using the normal distribution for populations and large sample sizes.

The t distribution, like the normal distribution, is bell shaped but the distribution is both flatter and fatter than the normal distribution. The t distribution applies to samples. Since sample sizes differ, there is a specific t distribution for each sample size. That is why there is a specific t distribution that is dependent of the “degrees of freedom”. For large samples, the normal and t are indistinguishable. But when sample sizes are small, the t distribution is a better choice. Since the t distribution can be used both for and large and small samples, computer programs use the t distribution rather than the normal distribution.

The Chi-Square distribution is a skewed distribution, the degree of skew depending on the degrees of freedom. For small degrees of freedom (df = 5) the distribution is highly skewed to the right. As the degrees of freedom increase, the χ2 distribution becomes more and more normal. For df = 100, the χ2 distribution is indistinguishable from the normal distribution.



When do we use the χ2 distribution? The chi-square distribution is used to test if an observed (experimental) value is statistically different from a theoretical (expected) value. For instance, we can use the chi-square to test if the observed distribution is statistically different from the theoretical Benford distribution. The chi-square distribution is called a "goodness of fit" statistic since it measures how well the observed distribution fits the expected distribution. You may want to see the Khan Academy’s explanation of the chi-square distribution:

<https://www.khanacademy.org/math/statistics-probability/inference-categorical-data-chi-square-tests/chi-square-goodness-of-fit-tests/v/chi-square-distribution-introduction>

II. The frequency bar chart contrasting the actual data with the theoretical predictions from Benford’s Law is shown below.

b.) Visually, the data conforms to a distribution that is almost opposite to Benford’s prediction. The numbers 8 and 9 appear as first digits much more frequently as compared to smaller numbers like 3 and 4.

Generally speaking, chi-square values of less than 5% suggest that there is little likelihood that the data conform to the hypothesized (Benford) distribution while values of 10% or less suggest that there is at least a 90% probability that the data are not generated by a genuine process. The zero value here for the chi-square test indicates that there is a near 100% probability that the data in this sample was invented and not generated by a genuine process.

c.) There are two other indicators that imply fraud:

i.) Several dollar amounts are just below $100,000. There are 9 instance out of 23 where the dollar amounts are $90,000 or greater. The reason could be that amounts greater than $100,000 would receive greater financial scrutiny that numbers below $100,000.

ii.) There are no round numbers in the checks – almost all the values involve cents and subconsciously there seems to have been a preference for 8’s and 9’s.

III. a.)

b.)

|  |  |  |  |
| --- | --- | --- | --- |
| **DIGIT** | **NUMBER OF TIMES OCCURS IN DATA (ACTUAL COUNT)** | **BENFORD PREDICTION (%)** | **NUMBER OF TIMES OCCURS PER BENFORD (EXPECTED COUNT)** |
| 1 | 55 | 30.10% | 64 |
| 2 | 45 | 17.61% | 38 |
| 3 | 35 | 12.49% | 27 |
| 4 | 27 | 9.69% | 21 |
| 5 | 21 | 7.92% | 17 |
| 6 | 10 | 6.69% | 14 |
| 7 | 8 | 5.80% | 12 |
| 8 | 8 | 5.12% | 11 |
| 9 | 5 | 4.58% | 10 |
|  | **214** | **100%** | **214** |
|  |  |  |  |
| **CHI-SQUARE TEST** | **7.61%** |  |  |

c.) The very low chi-squared valued should alert an investigator to the fact that the probability of fraud here is high.

IV. Examples where Benford’s Law is unlikely to be satisfied:

* Numbers that are deliberately assigned such as check numbers
* Zip codes
* Social Security Numbers
* Prices that are set particular thresholds such as $.99, $999, etc.
* Values that have a maximum or minimum value