Goodness of Fit Tests

Statistical tests for comparing a random sample with a theoretical probability distribution

Considerations

- Goodness of fit tests only provide guidance as to suitability of using a particular probability distribution (as opposed to falling back on an empirical table)
  - In real application it is unlikely there is a single correct theoretical distribution
- Tests typically check for frequency of sample data in cells (histogram), or deviation of data from that of a theoretical probability distribution
- If there is a lot of data, it is likely all candidate distributions will be rejected
  - Failure to reject favors the choice
  - Rejecting is only one piece of evidence against the choice and so does not rule it out

UNF

Common Tests

Chi-Square and Kolmogorov-Smirnov

- Chi-Square (Pearson)
  - Uses a histogram where columns are of equal width or of equal probability
  - A weakness is that a large data set (at least 50) is required
  - Columns should have at least 3, 4, or 5 elements, depending on the source
    - Adjacent columns may be grouped to meet this requirement so long as appropriate adjustments to formulas are made
- Kolmogorov-Smirnov
  - Since the underlying test is for testing uniformity of data in (0,1), each theoretical distribution requires its own transformation to convert the sample data for comparison to the uniform distribution
  - The test statistic is the largest deviation of the (transformed) cumulative data from the cumulative distribution of the uniform distribution function

UNF

Newer Tests

- Anderson-Darling (1952)
  - Similar in approach to Kolmogorov-Smirnov, but uses a more comprehensive measure of difference, so is more sensitive to the tails of the distribution than Kolmogorov-Smirnov
  - Particularly useful for determining if a small sample set conforms to a normal distribution
- G-test (1994)
  - Similar in application to chi-square, but is more sensitive when there is some wide dispersion of sample data from the theoretical distribution

UNF

Chi-Square

- The test objective is to formalize the approach of matching a theoretical distribution across a histographic representation of the data
- The data range is divided into k intervals and the frequency count O_i for each interval is tabulated
- The expected frequency E_i based on the theoretical pdf f with cumulative distribution F is calculated for each interval
  - $E_i = Np_i$ where N is the number of data points
    - $p_i = F(a_i) - F(a_{i-1})$ where a_i and a_{i-1} are the endpoints of the interval
    - or $p_i=f(i)$ if the interval is a single point
- The test statistic is $$\chi_0^2 = \sum_1^k \frac{(O_i - E_i)^2}{E_i}$$
- The test statistic then approximates the chi-square distribution with k-s-1 degrees of freedom, where s is the number of parameters of the theoretical probability distribution estimated from sample data

UNF

Test Procedure

- The null hypothesis is that the random variable corresponding to the sample conforms to the theoretical distribution (using the parameter estimates)
- The null hypothesis is rejected if the test statistic

$$\chi_0^2 > \chi_{\alpha,\text{k-s-1}}^2$$

  - Chi-square values for significance level α and degrees of freedom k-s-1 are in the table on page 584 or can be calculated using Excel

UNF

Worked Example

Normal Distribution

- Consider the 50 numbers:

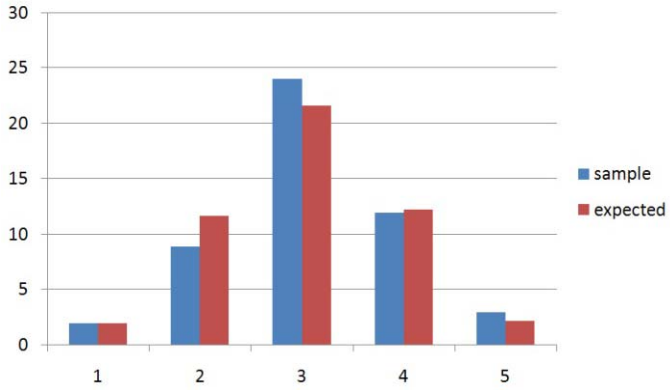
  81, 79, 75, 65, 76, 77, 85, 74, 73, 86, 65, 75, 66, 74, 83, 79, 74, 73, 81, 80, 85, 94, 78, 76, 89, 68, 80, 80, 78, 76, 71, 67, 63, 80, 70, 75, 76, 61, 86, 58, 74, 62, 94, 71, 58, 61, 70, 71, 74, 93

- 5 intervals 50-60, 60-70, 70-80, 80-90, 90-100 have frequencies

  $O_1=2$, $O_2=9$, $O_3=24$, $O_4=12$, $O_5=3$

- Sample mean = 75.2; sample standard deviation = 8.764492
- For the normal distribution $E_i = F(a_i) - F(a_{i-1})$ produces

  $E_1=1.97$, $E_2=11.75$, $E_3=21.58$, $E_4=12.32$, $E_5=2.17$



- $\chi_0^2 = 1.246612$
- k-s-1 = 5-2-1 = 2 degrees of freedom
- Since $\chi^2_{0.10,2} = 4.61$ we cannot reject the Null hypothesis with 10% certainty; moreover we have to get to 55% for a chi square value exceeding the test statistic
- Conclusion: the test does not call into question that the data is normally distributed

UNF

Worked Example
Gamma Distribution

- With the same data the 5 intervals 50-60, 60-70, 70-80, 80-90, 90-100 have frequencies O_1=2, O_2=9, O_3=24, O_4=12, O_5=3 as before
- Sample mean μ = 75.2; sample standard deviation σ = 8.764492
  sample b = σ^2/μ = 1.021494; sample α = μ/β = 73.61768
- For the exponential distribution $E_i = F(a_i) - F(a_{i-1})$ produces
  E_1=1.66, E_2=12.55, E_3=21.63, E_4=11.54, E_5=2.36
- $\chi^2_{0.10,2}$ = 4.61
- k-s-1 = 5-2-1 = 2 degrees of freedom
- Since χ^2 = 1.652259 we cannot reject the Null hypothesis with 10% certainty; moreover we have to get to 48% for a chi square value exceeding the test statistic
- Conclusion: the test does not call into question that the data is gamma distributed although it appears the normal distribution might be the better choice

UNF

Comments on Chi-Square

- Chi-square incorporates estimation of parameters for the hypothesized distribution by decreasing degrees of freedom
- Data counts need to be fairly large
- Data frequency is tabulated in somewhat arbitrarily chosen intervals
- Changing the number of intervals could conceivably cause rejection of the Null hypothesis in one case but not the other; i.e., the test results could be manipulated
- The test statistic only approximates chi-square, potentially weakening the test

UNF

Kolmogorov-Smirnov

- Kolmogorov-Smirnov in works with smaller sample sizes and estimation of parameters from the sample data makes the test more conservative
- Unfortunately, the test requires special tables (or calculations) except when being used to test for the exponential or uniform distribution
  - The data has to be normalized for the unit interval (0,1) to use the table
- The test objective is to look at the largest (absolute) deviation D of the cumulative distribution function $S_N(x)$ of the (normalized) sample from the cumulative distribution function F(x) of the uniform distribution on (0,1) over the range of the data
  - Note that F(x) is just the straight line (given by y=x) through the data points of $S_N(x)$
- The test distribution has been determined and its values for different α values are given for different values of N (degrees of freedom) in the table on page 586

UNF

Test Procedure

- The null hypothesis is that the random variable corresponding to the sample conforms to the theoretical distribution
- The null hypothesis is rejected if the test statistic D determined from the data exceeds the corresponding Kolmogorov-Smirnov value D_α for significance level α
  - Kolmogorov-Smirnov values for significance level α and degrees of freedom N are in the table on page 586

UNF

Worked Example
Exponential Distribution

- Scenario: interarrival times are collected over a 30 minute interval as follows

  1.25, 0.01, 1.41, 1.86, 0.78, 0.27, 1.86, 1.89, 1.67, 0.05, 0.63, 0.59, 0.36, 4.30, 0.15, 1.00, 0.32, 1.02, 2.04, 0.98, 0.21, 0.62, 2.32, 1.00, 0.18, 0.41, 0.11, 0.63, 0.87, 0.70
- As mentioned earlier (Poisson process), if the times are exponentially distributed, the arrival times are uniformly distributed
- By normalizing arrival times (dividing by 30 minutes) to the unit interval (0,1) Kolmogorov-Smirnov can be applied

UNF

Worked Example (2)
Exponential Distribution

- The normalized arrival times are

  0.04, 0.04, 0.09, 0.15, 0.18, 0.19, 0.25, 0.31, 0.37, 0.37, 0.39, 0.41, 0.42, 0.56, 0.57, 0.60, 0.61, 0.65, 0.72, 0.75, 0.75, 0.78, 0.85, 0.89, 0.89, 0.91, 0.91, 0.93, 0.96, 0.98
- These are compared to $F(x) = x$ values $i/30$ for $i=1 \ldots 30$ so $D = \max|S_N(i) - i/30| = .098$
- $D < D_{0.1} = 0.22$
- Conclusion: the test does not call into question that the data is exponentially distributed
- Remark: even with 30 points, the data is too dispersed for a histographic plot to pick out the exponential distribution

UNF