SA421 – Simulation Modeling
Asst. Prof. Nelson Uhan

Spring 2013

Lesson 9. Kolmogorov-Smirnov Test for Uniformity, Testing for Independence

1 Overview

- Today: another test for uniformity: the **Kolmogorov-Smirnov (K-S) Test**
- Advantages over the Chi-squared test:
  - No intervals need to be specified
  - Designed for continuous data, like values sampled from a Uniform$[0,1]$ random variable
- Disadvantages: more involved

2 The Kolmogorov-Smirnov Test

- Let $Y_1, \dots, Y_n$ be n independent random variables in $[0,1]$
- Let $y_1, \dots, y_n$ be the observations of $Y_1, \dots, Y_n$
- Let F be the cumulative distribution function (cdf) of a Uniform$[0,1]$ random variable U, e.g.

- The null hypothesis H_0 for the K-S test:

- Let F_e be the **empirical cdf** of $Y_1, \dots, Y_n$:

- The **(adjusted) test statistic** D is
  - D is small $\Rightarrow$ evidence in favor of H_0
  - D is large $\Rightarrow$ evidence against H_0
- After observing $Y_1, \dots, Y_n$, we can compute the **observed (adjusted) test statistic** d using $y_1, \dots, y_n$
- The **p-value** is $\mathbb{P}(D \geq d)$
  - D does not have a closed form!
  - Critical values: $\mathbb{P}(D \geq d_\alpha) = \alpha$

α	0.150	0.100	0.050	0.025	0.010
d_α	1.138	1.224	1.358	1.480	1.628

  - For a given α, if $d > d_\alpha$, then reject H_0

3 Computing the test statistic D

- Let $y_{(j)}$ be the jth smallest of $y_1, \dots, y_n$, for $j = 1, \dots, n$
- Using the properties of the empirical cdf, one can show that

- Intuitively:
  - Remember: for Uniform$[0,1]$, the cdf is $F(x) = x$ for $0 \le x \le 1$
  - If $y_1, \dots, y_n$ are observations from the Uniform$[0,1]$ distribution, then we expect $y_{(j)}$ to be in the interval $\left[\frac{j-1}{n}, \frac{j}{n}\right]$
  - D measures how far $y_{(j)}$ falls from $\left[\frac{j-1}{n}, \frac{j}{n}\right]$
- In the Excel workbook for today's lesson, the "K-S" sheet contains 20 psuedo-random numbers
- Let's conduct the K-S test for uniformity on these numbers
- Copy and paste the numbers, use Data → Sort to get the numbers in ascending order (i.e., the $y_{(j)}$'s)
- Compute the differences, use the MAX function to get D

4 Testing for independence

- Many tests have been devised to determine whether a set of psuedo-random numbers are independent
- Here's a simple test that will serve our purposes for now
- The **sample correlation** between $(x_1, \dots, x_n)$ and $(y_1, \dots, y_n)$ is

$$\rho = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

  where $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ and $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$
  - Can be computed using the CORREL function in Excel
- Simple test:
  - Compute correlation between $(x_1, \dots, x_n)$ and $(x_2, \dots, x_{n+1})$ (consecutive observations)
  - Compute correlation between $(x_1, \dots, x_n)$ and $(x_3, \dots, x_{n+2})$ (every other observation)
  - If these correlations are "small" (rule of thumb: absolute value less than 0.3), then do not reject the hypothesis that the observations are independent
- In the "indep" sheet, there are 22 pseudo-random numbers
- Let's conduct this test for independence on these numbers