SA421 – Simulation Modeling | Spring 2013
Asst. Prof. Nelson Uhan

Lesson 8. Chi-Square Test for Uniformity

1 Last time...

- A **pseudo-random number generator (PRNG)** is a (deterministic) algorithm that uses mathematical formulas or precalculated tables to produce sequences of numbers that appear to be independently sampled from a Uniform$[0,1]$ distribution
- A good PRNG must pass statistical tests for **uniformity** and **independence**
  - These numbers should not be statistically differentiable from a sequence of truly independently sampled values from the Uniform$[0,1]$ distribution
- Today: how can we test for uniformity?

2 Motivation: testing Excel's RAND function

- We don't know what PRNG Excel's RAND function uses
- How uniform are the pseudo-random numbers generated by Excel's RAND function?
- In the spreadsheet for today's lesson, we have 150 pseudo-random numbers generated by RAND
- Idea: if the interval $[0,1]$ is divided into m subintervals of equal length, and n values are sampled, then the expected number of values in each interval is n/m
  - ⇒ If we divide $[0,1]$ into 10 subintervals – $[0,0.1]$, $[0.1,0.2]$, etc. – we should expect to see about 15 numbers in each subinterval
- Set up "bins" for each interval
- Use the FREQUENCY function to figure out how many numbers fall into each bin
  - FREQUENCY(data_array, bins_array)
    - data_array = array/reference to a set of values for which you want to count frequencies
    - bins_array = array/reference to intervals into which you want to group the values in data_array
  - Highlight the "# observations" column, enter the formula, then press **CTRL-SHIFT-ENTER**
    - FREQUENCY is an array function
- Can plot a histogram by highlighting the "# observations" column, and then selecting Insert → Column → Clustered Column
- Are there roughly 15 numbers in each bin?
- Is this "close enough" to what we expect?
- Let's be more rigorous about this...

3 Chi-squared test for uniformity

- Let $Y_1, \dots, Y_n$ be n independent random variables in $[0, 1]$
- Let $y_1, \dots, y_n$ be observations of $Y_1, \dots, Y_n$
- Divide $[0, 1]$ into m subintervals of equal length: $\left[0, \frac{1}{m}\right], \left[\frac{1}{m}, \frac{2}{m}\right], \dots, \left[\frac{m-1}{m}, 1\right]$
- The hypothesis we are testing: if Y represents any of the Y_j's,

  $\boxed{}$

  - We call this the **null hypothesis** H_0 (for this test)
- Let O_i be the number of Y_j's in $\left[\frac{i-1}{m}, \frac{i}{m}\right]$ for $i = 1, \dots, m$
  - O_i is a random variable: uncertain quantity before Y_j's are observed
  - Let $e_i = \mathbb{E}[O_i]$ = expected number of observations in $\left[\frac{i-1}{m}, \frac{i}{m}\right] = \boxed{}$
- Let $o_1, \dots, o_m$ be the observations of $O_1, \dots, O_m$
- The **test quantity** T is $\boxed{}$
  - T is a random variable that has approximately a chi-squared distribution with $m - 1$ degrees of freedom when H_0 is true
  - Rule of thumb: $e_i \geq 5$ for $i = 1, \dots, m$
- The **observed test quantity** t is $\boxed{}$
- Small values of $t \Rightarrow$ evidence in favor of H_0
- Large values of $t \Rightarrow$ evidence against H_0
- How large does t have to be to reject H_0?
- The p-**value** is $\boxed{}$
  - Interpretation: probability that such a large value of T would have been observed if H_0 is true
  - Small p-values ($< \alpha$, where α is typically 0.05 or even 0.01) $\Rightarrow$ reject H_0
- Computing this in Excel:
  - `CHIDIST(t, m-1)` $= \mathbb{P}(\chi^2_{m-1} \geq t)$