Lesson 8. Input Data Analysis – Discrete Distributions

1 Overview

- In order to construct a simulation model of a system, we need to make assumptions about the parameters
  - e.g. "customer interarrival times are exponentially distributed with a mean of 12.2 minutes"
- How do we go about making these assumptions in an intelligent way?
  1. **Collect raw data**: e.g. manually, automatically, expert opinion
  2. **Histogram and descriptive statistics**: e.g. visual tests, knowledge of probability distributions
  3. **Compare with known distributions**: e.g. discrete/continuous, shape of density/mass function, independence, homogeneous/non-homogeneous with respect to time
  4. **Goodness-of-fit tests**: statistically speaking, how good is our choice of distribution?

2 Which distribution should I choose? – Rules of thumb

- Certain types of processes have been found to be best modeled using certain distributions

Distribution	Where it may be useful	Examples
uniform	Values on an interval are "equally likely"	random number generator
exponential	Time between occurrences of a random (Poisson) process	arrivals
normal	Where averaging seems likely	measurements
gamma (Erlang)	Similar process repeated several times	service times
beta	Proportion of time something happens	% downtime, % impurities
triangular	Little information about distribution (only min, max, most likely)	no data available, expert opinion only

3 Discrete random variables and distributions: review

- A random variable is **discrete** if it can take on only a finite or countably infinite number of values
- Let X be a discrete random variable that takes on values $a_0, a_1, a_2, \ldots$ such that $a_0 < a_1 < a_2 < \ldots$
- The **cumulative distribution function (cdf)** F_X of X is

- The **probability mass function (pmf)** p_X of X is

- The pmf and cdf of a discrete random variable are related:

4 The chi-squared goodness-of-fit test

- Let $Y_0, \ldots, Y_{n-1}$ be n independent discrete random variables that take on values $a_0, a_1, a_2, \ldots, a_{m-1}$
- Let $y_0, \ldots, y_{n-1}$ be observations of $Y_0, \ldots, Y_{n-1}$
- Let X be the proposed discrete random variable with pmf p_X
- Question: Do the Y_j's share the same distribution same as X? More formally...
- Null hypothesis H_0: for any Y_j,
- Let O_i be the number of Y_j's equal to a_i, for $i = 0, \ldots, m-1$
  - O_i is a random variable: an uncertain quantity before the Y_j's are observed
  - Let $e_i = \mathbb{E}[O_i]$ = expected number of Y_j's equal to a_i under H_0 = , for $i = 0, \ldots, m-1$
- Let $o_0, \ldots, o_{m-1}$ be the observations of $O_0, \ldots, O_{m-1}$
- The **test statistic** is
  - T approximately follows a chi-squared distribution with $m-1$ degrees of freedom when H_0 is true
  - Rule of thumb: approximation holds when $e_i \geq 5$ for $i = 0, \ldots, m-1$
- The **observed test statistic** is
- 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
  - 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