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

Lesson 26. Input Data Analysis

1 Overview

- In order to construct a model of a system, we need to make assumptions about the model inputs
  - 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?
- We can perform many of these tasks in Excel
- There is also specialized software: **Stat::Fit** automates steps 2 to 4
  - Stat::Fit is available in ProModel at launch, or via Tools » Stat::Fit

2 Rules of thumb

- Even though Stat::Fit can perform automated distribution identification, 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 Using Stat::Fit – an example

Example 1. The Midville Manufacturing Company has 2 planers for performing two different types of jobs, A and B. Each planer can process 1 job at a time. The time required to perform each job depends largely upon the number of passes that must be made: according to the planer operators, most jobs require 5 or fewer passes on a planer, and a pass takes about 10 minutes. Unfortunately, the planer department has had a difficult time keeping up with its workload. You have been asked to investigate the effect of obtaining 1 additional planer.

You have collected data for 1 week of the planer department's operation. The shop operates 8 hours per day. You have created a simulation model that accurately represents the flow of material through the shop, but you need the distributions for the interarrival times for each job type, and the service times for each job type on each planer.

- The Excel file for today's lesson contains the raw data that you collected in the "raw data" worksheet
- Let's first compute the **clock time** of each event in minutes: that is, let's normalize the times so that time 0 corresponds to 9:00 on the first day of the week

$$\text{clock time} = \big((\text{day} - 1) \times 8 + \text{hour} - 9\big) \times 60 + \text{min}$$

3.1 Interarrival time distribution fitting

- Let's try to fit a distribution to the interarrival times of type A jobs

Manipulating the data

- Make a copy of the raw data with clock times and get rid of all records except arrivals for type A jobs
  - This can be done easily by sorting
- Sort the remaining records by clock time in increasing order and compute the interarrival times
- Copy and paste the last 50 interarrival times into a new data table in Stat::Fit
  - Note: the student version of Stat::Fit only allows for 50 data points

Histogram and descriptive statistics

- Create a histogram of the data: Input » Input Graph
- Get descriptive statistics of the data: Statistics » Descriptive

Compare with known distributions and goodness-of-fit

- In general, we expect the random arrivals to have exponentially distributed interarrival times
- The histogram seems to indicate this as well

⇒ Let's restrict Stat::Fit's search to exponential distributions
  - Select Fit » Setup
  - In the "Distributions" tab, make sure only "Exponential" is selected
  - In the "Calculations" tab, select the Chi-squared and KS (Kolmogorov-Smirnov) tests
    - The KS goodness-of-fit test is similar to the KS test for uniformity

- Perform the goodness-of-fit tests: Fit » Goodness of Fit
- Null hypothesis H_0 of goodness-of-fit tests: data is from the specified distribution
- Suppose we want to test with 95% confidence: do the p-values indicate that H_0 should be rejected?

Testing for independence

- Visually test for correlation between consecutive data values using a scatter plot: Statistics > Independence > Scatter Plot

- Conduct **runs tests** for independence: Statistics > Independence > Runs Tests

  - The **above/below median test** measures the number of consecutive sequences of values above and below the median
  - The **turning points test** measures the number of times the values change direction (increasing to decreasing or decreasing to increasing)

- Do the tests indicate that the data are independent?

3.2 Service time distribution fitting

- Let's try to fit a distribution to the service times of type A jobs on planer 1

Manipulating the data

- Copy the raw data with clock times and get rid of all records except begin/end service for type A jobs on planer 1

- Sort the begin service records and the end service records separately in increasing order of clock time

- Match the begin and end service records by cutting and pasting

  - This works because each planer can only process 1 job at a time

- Compute the service times

- Copy and paste the last 50 service times into a new data table in Stat::Fit

Histogram, descriptive statistics, testing for independence

- Just like before, we can visually inspect the data using a histogram and a scatter plot

- Change the number of intervals used in the histogram: Input > Input Options

- We can also get a feel for the data via descriptive statistics

- Is the data independent? What distributions are likely?

Using Auto::Fit

- We don't have any prior knowledge about service time distributions, except that they are related to the number of passes the job requires

- Try all continuous distributions:

  - Fit > Auto::Fit, select the "continuous distributions" radio button
  - Click on the results for histograms
  - Fit > Goodness of Fit for more details on the tests performed

- Try all discrete distributions in a similar way
- Known distributions in Stat::Fit don't seem particularly convincing
- Another option: use an **empirical distribution**
    - Use the % frequency that the values occur in the data as the probability distribution