Input Data Modeling

Histogram

- **Input data modeling** is the process of collecting and analyzing input data to increase the accuracy of a simulation. Incorrect input data lead to wrong results and can mislead stakeholders.
- The collection and analysis of input data require time and resource commitment, particularly in discrete event simulation.
- In a queuing system simulation, the typical input data are the distribution of time between arrivals and service time.
- Input data models provide a driving force for a simulation. The following are the possible steps in developing an input data model (Chaturvedi, 2010):
  - Collecting data from a system
  - Identifying the probability distribution to represent the input process
  - Choose appropriate/applicable parameters for the distribution
  - Evaluating the chosen distribution

- **Data collection** is a big part of solving real-life problems, particularly in input data modeling.
- The following practices are highly suggested to be observed during data collection:
  - Plan ahead and/or conduct a pre-observing session.
  - Perform general analysis of data as it is being collected.
  - Construct a scatter diagram to check for relationship(s) between variables.
  - Check for *autocorrelation* – a mathematical representation of the degree of similarity between a given time series and a lagged version of itself over successive intervals (Smith, 2020).

- A **histogram** is a graphical representation that organizes a group of data points into user-specified ranges. It represents the frequency distribution of variables in a data set.
- A histogram condenses data series into an easily interpreted visual by taking numerous data points and grouping them into logical ranges or bins.
- The appearance of a histogram can be customized in several ways by developers and analysts. It looks very similar to a bar graph.
- Histograms are commonly used in Statistics to demonstrate how many of a certain type of variable occurs within a specific range (Chen, 2021).
- The common shape of the distribution, both for natural and industrial settings, is the *normal distribution*, which looks like a bell-shaped curve, as shown below.

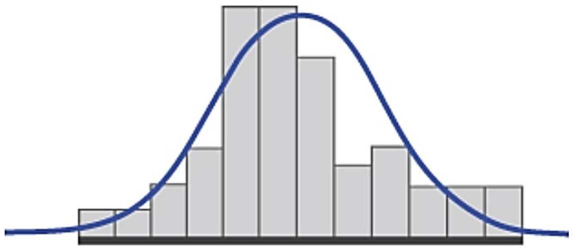

Figure 1. A histogram is depicting a normal frequency distribution.
Source: https://www.moresteam.com/toolbox/histogram.cfm

- The following are the other distribution shapes that you may encounter:

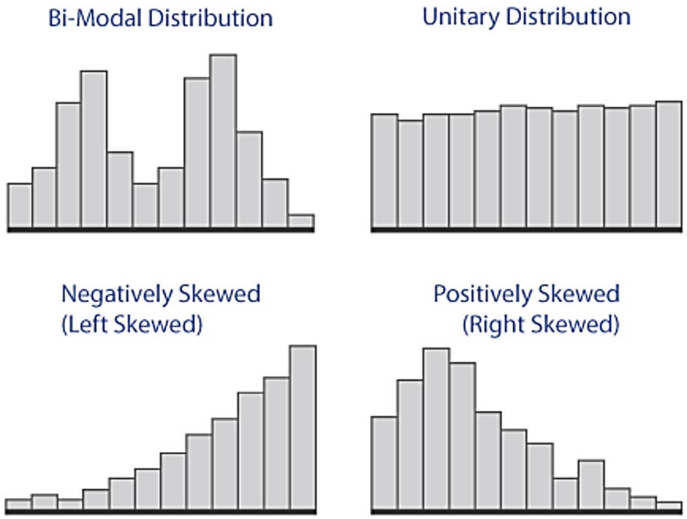

Figure 2. Other frequency distribution shapes.
Source: https://www.moresteam.com/toolbox/histogram.cfm

- In histograms, the number of class intervals, or the class boundaries, depends on the number of observations and the dispersion of the data, and the use of the square root of the sample size is suggested to be used in the computation process.

Example 1:
The Moving Average Convergence Divergence (MACD) Histogram. A MACD histogram is plotted on a chart, aiding traders to easily monitor the rising or falling of stock values in the market.

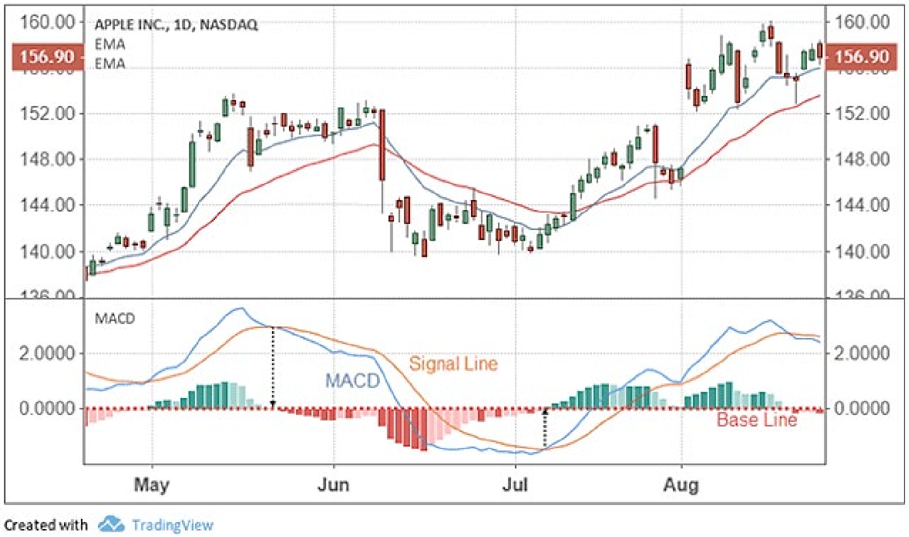

Figure 3. An example of a MACD chart with a histogram.
Source: https://www.investopedia.com/terms/m/macd.asp

Example 2:
The collected height data from a training class of 50 individuals.

69.9	69.0	69.6	68.5	65.0	65.9	67.2	67.5	68.0	68.6
68.9	70.0	69.5	70.4	71.1	71.0	72.5	73.1	68.8	71.3
68.2	68.5	70.0	66.8	69.0	69.3	69.1	69.4	68.5	65.5
66.0	66.5	67.5	68.3	68.2	69.1	70.2	69.5	70.5	70.8
71.0	72.5	73.0	69.0	71.3	68.2	68.5	70.0	67.0	69.2

There are only 50 values above, but it is difficult to draw any specific conclusion about the data without further analysis. By determining the appropriate class intervals or boundaries and tallying the frequency, a histogram can be constructed to provide more useful information.

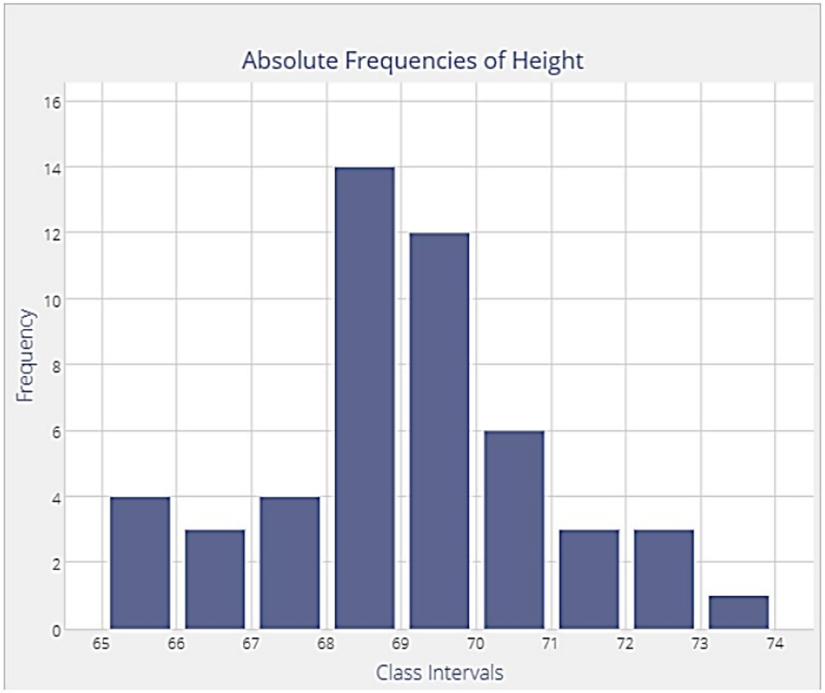

Figure 4: The histogram for the collected height data from a training class of 50 individuals.
Source: https://www.moresteam.com/toolbox/histogram.cfm

Family of Distributions

There are various types of distribution that can be used in input data modeling, and some of them are the following (Chaturvedi, 2010):

- ○ Exponential distribution
- ○ Normal distribution
- ○ Poisson distribution
- ○ Binomial distribution
- ○ Triangular distribution
- ○ Lognormal distribution
- ○ Gamma distribution
- ○ Beta distribution
- ○ Weibull distribution
- ○ Uniform distribution

- The frequently encountered distributions are the following:
  - ○ Easy: Exponential Distribution – It is often concerned with the amount of time until some specific event occurs. Commonly, in this type of distribution, the large values are fewer compared to the small values (Lumen Learning, n.d.).

- Easy: Normal Distribution – It is also known as the *Gaussian Distribution*. It is a probability distribution that is symmetric about the mean, showing that the data near the mean occur more frequently than the data far from the means. Hence the bell-shaped curve. In addition, the normal distribution is also considered the most common type of distribution assumed in technical stock market analysis and in other types of statistical analysis (Chen, 2021).
- Easy: Poisson Distribution – This distribution is used to show how many times an event is likely to occur within a specified period of time. It is technically a count distribution. This type of distribution is often used to understand independent events that occur at a constant rate within a given interval of time (Hayes, 2021).
- Complex: Beta, Gamma, and Weibull Distribution

- The family of distribution is selected based on the following:
  - Context of the input variable
  - Shape of the histogram

- In selecting the appropriate distribution function for modeling and simulation, always take into consideration the physical characteristics of a system or process, such as bounded or unbounded, contains discrete or continuous values, etc.
- In hypothesizing the distribution, use the physical basis of the distribution as a guide, such as the examples below (Banks, Carson, Nelson & Nicol, n.d.):
  - *For exponential distribution*: the time between independent events
  - *For normal distribution*: the distribution of a process that is the sum of component processes
  - *For Poisson distribution*: the number of independent events that occur in a fixed time or space
  - *For binomial distribution*: the number of success in *n* trials
  - *For triangular distribution*: a process wherein the minimum, most likely, and maximum values are known.

Parameter Estimation

- **Parameter estimation** is the process of approximating the model parameters based on the model's response to certain test inputs.
- Parameter estimation is technically the next step after selecting the appropriate family of distribution for the input data modeling.
- The following parameter estimation formulas can be used, depending on the data characteristics (Banks, Carson, Nelson & Nicol, n.d.):
  - If the observations in a sample of size *n* are $X_1, X_2, \ldots, X_n$ (either discrete or continuous), the sample mean and variance can be computed as:

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} \quad S^2 = \frac{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2}{n-1}$$

  - If the data are discrete and have been grouped by frequency distribution (where f_j is the observed frequency of X_j), the sample mean and variance can be computed as:

$$\overline{X} = \frac{\sum_{j=1}^{n} f_j X_j}{n} \qquad S^2 = \frac{\sum_{j=1}^{n} f_j X_j^2 - n\overline{X}^2}{n-1}$$

  - If the raw data are unavailable (data are grouped into class intervals), the approximate sample mean and variance can be computed as:

$$\overline{X} = \frac{\sum_{j=1}^{c} f_j X_j}{n} \qquad S^2 = \frac{\sum_{j=1}^{n} f_j m_j^2 - n\overline{X}^2}{n-1}$$

Where:
f_j – is the observed frequency of the *j*th class interval;
m_j – is the midpoint of the *j*th interval; and
c – is the number of class intervals.

Goodness-of-fit Test

- **Goodness-of-fit test** is a statistical hypothesis test to see how well sample data fit a distribution. This test will show if the sample data represents the expected data from the population.
- Goodness-of-fit test is often used to make interpretations about the observed values. It determines how related actual values are to the predicted values in a model. In addition, the goodness-of-fit tests can also help in predicting future trends and patterns.
- There are multiple methods that can be performed in determining the goodness-of-fit. The two (2) commonly used methods are the following (Kenton, 2021).:
  - **Chi-Square Test** – It is an inferential statistics method that tests the validity of a claim made about a population based on a random sample, but it does not indicate the type or intensity of the relationship.

It is valid for large sample sizes when parameters are estimated by maximum likelihood. In performing this method, the *n* number of observations must be arranged into sets of *k* class intervals. This method is mathematically described as:

$$X^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \qquad ; E_i = n * p_i$$

Where:
k – is the class intervals;
O_i – is the frequency of observed values;
E_i – is the frequency of expected values;
n – is the number of observations; and
p_i – is the theoretical probability of the *i*th interval, with the suggested minimum value of five (5).

- The hypothesis of a chi-square test is as follows:
  H_0: The random variable *X* conforms to the distributional assumption with the parameter(s) given by the estimate(s).
  H_1: The random variable *X* does not conform.

- **Kolmogorov-Smirnov Test** – It is also known as the K-S Test. This method was named after Russian mathematicians *Andrey Kolmogorov* and *Nikolai Smirnov*. It is a statistical method that determines whether a sample is from a specific distribution within the population. This goodness-to-fit test is non-parametric, meaning it does not rely on any distribution to be valid and is applicable to continuous distributions with small sample size. The Kolmogorov-Smirnov test is considered a more powerful test and particularly useful when the sample sizes are small, and there are no parameters that have been estimated from the data. This method is mathematically described as (Tutorialspoint, n.d.):

$$D = Maximum \mid F_0(X) - F_r(X) \qquad ; F_0(X) = \frac{k}{n}$$

Where:
D – is the test statistic;
k – is the number of observations less than or equal to X;
n – is the total number of observations;
F_0 *(X)* – is the observed cumulative frequency distribution of a random sample of *n* distributions; and
F_r *(X)* – is the theoretical frequency distribution.

- **Acceptance Criteria***:* If the calculated value is less than the critical value, accept the null hypothesis.
- **Rejection Criteria**: If the calculated value is greater than the critical value, reject the null hypothesis.

References:
Banks, Carson, Nelson & Nicol. (n.d.). *Input modeling: Discrete-event system simulation*. Retrieved on April 7, 2021 from https://cs.wmich.edu/~alfuqaha/Spring09/cs6910/lectures/Chapter9.pdf
Chaturvedi, D. (2010). *Modeling and simulation of systems using MATLAB and Simulink*. CRC Press – Taylor & Francis Group, LLC
Chen, J. (2021, February 2). *Histogram*. Retrieved on April 6, 2021 from https://www.investopedia.com/terms/h/histogram.asp
Chen, J. (2021, March 31). *Normal distribution*. Retrieved on April 10, 2021 from https://www.investopedia.com/terms/n/normaldistribution.asp
Hayes, A. (2021, February 20). *Poisson distribution*. Retrieved on April 10, 2021 from https://www.investopedia.com/terms/p/poisson-distribution.asp
Kenton, W. (2021, March 24). *Goodness-of-fit*. Retrieved on April 11, 2021 from https://www.investopedia.com/terms/g/goodness-of-fit.asp#:~:text=The%20goodness%2Dof%2Dfit%20test,if%20it%20is%20somehow%20skewed.
Lumen Learning. (n.d.). *The exponential distribution*. Retrieved on April 10, 2021 from https://courses.lumenlearning.com/introstats1/chapter/the-exponential-distribution/
MoreSteam. (n.d.). *Histogram*. Retrieved on April 7, 2021 from https://www.moresteam.com/toolbox/histogram.cfm
Smith, T. (2020, March 10). *Autocorrelation*. Retrieved on April 7, 2021 from https://www.investopedia.com/terms/a/autocorrelation.asp
Tutorialspoint. (n.d). *Statistics – Kolmogorov Smirnov Test*. Retrieved on April 11, 2021 from https://www.tutorialspoint.com/statistics/kolmogorov_smirnov_test.htm