CTL.SC0x - Supply Chain Analytics

Key Concepts Document

This document contains the Key Concepts for the SC0x course. The document is based on a past run of the course and does not include all the material that will be included in this course; please check back for newer versions of the document throughout the course.

These are meant to complement, not replace, the lesson videos and slides. They are intended to be references for you to use going forward and are based on the assumption that you have learned the concepts and completed the practice problems.

The draft was created by Dr. Alexis Bateman in the Spring of 2017.

This is a draft of the material, so please post any suggestions, corrections, or recommendations to the Discussion Forum under the topic thread "Key Concept Documents Improvements."

Thanks,
Chris Caplice, Eva Ponce and the SC0x Teaching Community
Spring 2017 v1

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Table of Contents

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Supply Chain Intro

Summary

Supply Chain Basics is an overview of the concepts of Supply Chain Management and logistics. It demonstrates that product supply chains as varied as bananas to women's shoes to cement have common supply chain elements. There are many definitions of supply chain management. But ultimately supply chains are the physical, financial, and information flow between trading partners that ultimately fulfill a customer request. The primary purpose of any supply chain is to satisfy a customer's need at the end of the supply chain. Essentially supply chains seek to maximize the total value generated as defined as: the amount the customer pays minus the cost of fulfilling the need along the entire supply chain. All supply chains include multiple firms.

Key Concepts

While Supply Chain Management is a new term (first coined in 1982 by Keith Oliver from Booz Allen Hamilton in an interview with the Financial Times), the concepts are ancient and date back to ancient Rome. The term "logistics" has its roots in the Roman military. Additional definitions:

- Logistics involves… "managing the **flow** of information, cash and ideas through the coordination of supply chain processes and through the strategic addition of place, period and pattern values" – MIT Center for Transportation and Logistics
- "Supply Chain Management deals with the management of materials, information and financial flows in a **network** consisting of suppliers, manufacturers, distributors, and customers" - Stanford Supply Chain Forum
- "Call it distribution or logistics or supply chain management. By whatever name it is the sinuous, gritty, and cumbersome process by which companies move materials, parts and products to customers" – Fortune 1994

Logistics vs. Supply Chain Management

According to the Council of Supply Chain Management Professionals…

- **Logistics management** is that part of supply chain management that plans, implements, and controls the efficient, effective forward and reverse flow and storage of goods, services and related information between the point of origin and the point of consumption in order to meet customers' requirements.
- **Supply chain management** encompasses the planning and management of all activities involved in sourcing and procurement, conversion, and all logistics management activities. Importantly, it also includes coordination and collaboration with channel partners, which can be suppliers, intermediaries, third party service providers, and customers. In essence, supply chain management integrates supply and demand management within and across companies.

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Supply Chain Perspectives

Supply chains can be viewed in many different perspectives including process cycles (Chopra & Meindl 2013) and the SCOR model (Supply Chain Council).

The Supply Chain Process has four Primary Cycles: Customer Order Cycle, Replenishment Cycle, Manufacturing Cycle, and Procurement Cycle, Not every supply chain contains all four cycles.

The Supply Chain Operations Reference (SCOR) Model is another useful perspective. It shows the four major operations in a supply chain: source, make, deliver, plan, and return. (See Figure below)

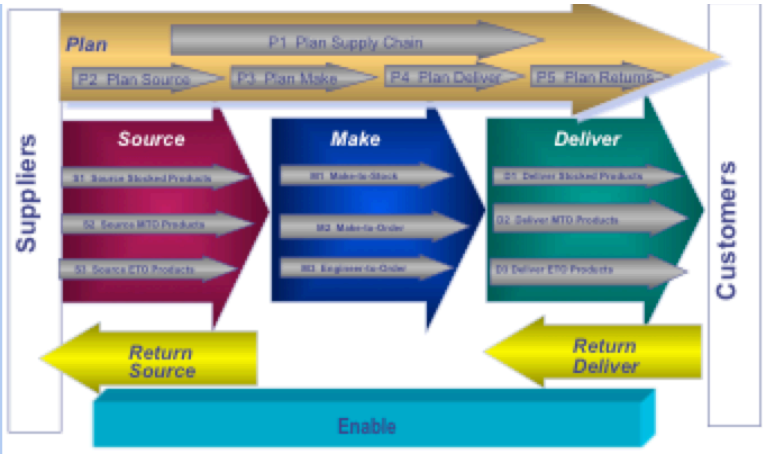


Figure: SCOR Model. Source: Supply Chain Council

Additional perspectives include:

- Geographic Maps - showing origins, destinations, and the physical routes.
- Flow Diagrams – showing the flow of materials, information, and finance between echelons.
- Macro-Process or Software – dividing the supply chains into three key areas of management: Supplier Relationship, Internal, and Customer Relationship.
- Traditional Functional Roles – where supply chains are divided into separate functional roles (Procurement, Inventory Control, Warehousing, Materials Handling, Order Processing, Transportation, Customer Service, Planning, etc.). This is how most companies are organized.
- Systems Perspective – where the actions from one function are shown to impact (and be impacted by) other functions. The idea is that you need to manage the entire system rather than the individual siloed functions. As one expands the scope of management, there are more opportunities for improvement, but the complexity increases dramatically.

Supply Chain as a System

It is useful to think of the supply chain as a complete system. This means one should:

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

- Look to **maximize value** across the supply chain rather than a specific function such as transportation.
- Note that while this increases the potential for improvement, **complexity** and **coordination** requirements increase as well.
- Recognize new challenges such as:
    - Metrics — how will this new system be measured?
    - Politics and power — who gains and loses influence, and what are the effects
    - Visibility — where data is stored and who has access
    - Uncertainty — compounds unknowns such as lead times, customer demand, and manufacturing yield
    - Global Operations — most firms source and sell across the globe

Supply chains must adapt by acting as both a bridge and a shock absorber to connect functions as well as neutralize disruptions.

Learning Objectives

- Gain multiple perspectives of supply chains to include process and system views.
- Identify physical, financial, and information flows inherent to supply chains.
- Recognize that all supply chains are different, but have common features.
- Understand importance of analytical models to support supply chain decision-making.

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Models, Algebra, & Functions

Summary

This review provides an overview of the building blocks to the analytical models used frequently in supply chain management for decision-making. Each model serves a role; it all depends on how the techniques match with need. First, a classification of the types of models offers perspectives on when the use a model and what type of output they generate. Second, a review of the main components of models, beginning with an overview of types of functions, the quadratic and how to find its root(s), logarithms, multivariate functions, and the properties of functions. These "basics" will be used continuously throughout the remainder of the courses.

Key Concepts

Models

Decision-making is at the core of supply chain management. Analytical models can aid in decision-making to questions such as "what transportation option should I use?" or "How much inventory should I have?" They can be classified into several categories based on degree of abstraction, speed, and cost.

Models can be further categorized into three categories on their approach:

- Descriptive – what has happened?
- Predictive – what could happen?
- Prescriptive – What should we do?

Functions

Functions are one the main parts of a model. They are "a relation between a set of inputs and a set of permissible outputs with the property that each input is related to exactly one output."(Wikipedia)

$$y=f(x)$$

Linear Functions

With Linear functions, "y changes linearly with x." A graph of a linear function is a straight line and has one independent variable and one dependent variable. (See figure below)

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

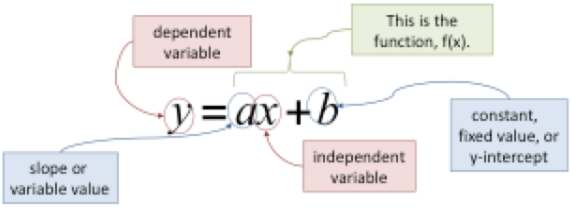


Figure: Components of a Linear Function

**Typically constants are denotes by letters from the start of the alphabet, variables are letters from the end of the alphabet.

Quadratic Functions

A quadratic function takes the form of $y=ax^2 + bx + c$, where a, b, and c are numbers and are not equal to zero. In a graph of a quadratic function, the graph is a curve that is called a parabola, forming the shape of a "U". (See Figures)

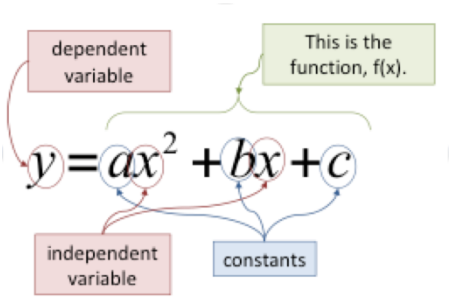


Figure: Components of a Quadratic Function

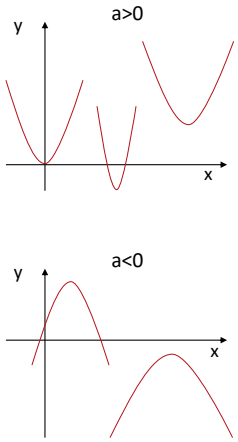


Figure: Graph of Quadratic Functions

Roots of Quadratic

A root, or solution, satisfies the quadratic. The equation can have 2, 1, or 0 roots. The roots must be a real number. There are two methods for finding roots:

Factoring: Find r_1 and r_2 such that $ax^2+bx+c = a(x-r_1)(x-r_2)$

Quadratic equation

$$r_1, r_2 = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Other Common Functional Forms

Power Function

A power function is a function where a ≠ zero, is a constant, and b is a real number. The shape of the curve is dictated by the value of b.

$$y = f(x) = ax^b$$

Euler's number, or e, is a constant number that is the base of the natural logarithm.

e = 2.7182818…

$$Y=e^x$$

Exponential Functions

Exponential functions have very fast growth. In exponential functions, the variable is the power.

$$y=ab^x$$

Multivariate Functions

Function with more than one independent x variable (x_1, x_2, x_3).

Logarithms: A logarithm is a quantity representing the power to which a fixed number must be raised to produce a given number. It is the inverse function of an exponential.

$$y=b^x \quad \leftrightarrow \quad \log_b(y)=x$$

Convexity and Continuity

Properties of functions:

- Convexity: Does the function "hold water"?
- Continuity: Function is continuous if you can draw it without lifting pen from paper!

Learning Objectives

- Recognize decision-making is core to supply chain management.
- Gain perspectives on when to use analytical models.
- Understand building blocks that serve as the foundation to analytical models.

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Optimization

Summary

This is an introduction and overview of optimization. It starts with an overview of unconstrained optimization and how to find extreme point solutions, keeping in mind first order and second order conditions. It also reviews rules in functions such as the power rule. Next the lesson review constrained optimization that shares similar objectives of unconstrained optimization but adds additional decision variables and constraints on resources. To solve constrained optimization problems, the lesson introduces mathematical programs that are widely used in supply chain for many practices such as designing networks, planning production, selecting transportation providers, allocating inventory, scheduling port and terminal operations, fulfilling orders, etc. The overview of linear programming includes how to formulate the problem, how to graphically represent them, and how to analyze the solution and conduct a sensitivity analysis.

In real supply chains, you cannot .5 bananas in an order or shipment. This means that we must add additional constraints for integer programming where either all of the decision variables must be integers, or in a mixed integer programming where some, but not all, variables are restricted to be an integer. We review the types of numbers you will encounter. Then we introduce integer programs and how they are different. We then review the steps to formulating an integer program and conclude with conditions for working with binary variables.

Key Concepts

Unconstrained Optimization

Unconstrained optimization considers the problem of minimizing or maximizing an objective function that depends on real variables with no restrictions on their values.

Extreme points

- Extreme points are when a function takes on an extreme value - a value that is much smaller or larger in comparison to nearby values of the function.
- They are typically a min or a max (either global or local), or inflection points.
- Extreme point occur where slope (or rate of change) of function = 0.
- Test for Global vs. Local
    - Global min/max – for whole range
    - Local min/max – only in certain area

Finding Extreme Point Solutions

Use differential calculus to find extreme point solutions, look for where slope is equal to zero

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

To find the extreme point, there is a three-step process:

1. Take the first derivative of your function
2. Set it equal to zero, and
3. Solve for X*, the value of x at extreme point.

This is called the First Order Condition.

Instantaneous slope (or first derivative) occurs when:

- dy/dx is the common form, where d means the rate of change.

δ (delta) = rate of change.

$$y' = f'(x) = \frac{dy}{dx}$$

The Product Rule: If function is constant, it doesn't have any effect.

$$y = f(x) = a \rightarrow y' = f'(x) = 0$$

Power Rule is commonly used for finding derivatives of complex functions.

$$y = f(x) = ax^n \rightarrow y' = f'(x) = anx^{n-1}$$

First and Second Order Conditions

In order to determine x* at the max/min of an unconstrained function

- First Order (necessary) condition – the slope must be 0

  f'(x*)=0

- Second order (sufficiency) condition - determines where extreme point is min or max by taking the second derivative, **f"(x).**
  - If f"(x) > 0 extreme point is a local min
  - If f"(x) < 0 extreme point is a local max
  - If f"(x) = 0 it is inconclusive
- Special cases
  - If f(x) is convex –> global min
  - If f(x) is concave –> global max

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Constrained Optimization

Similarities with unconstrained optimization

- Requires a prescriptive model
- Uses an objective function
- Solution is an extreme value

Differences

- Multiple decision variables
- Constraints on resources

Math Programming: Math programming is a powerful family of optimization methods that is widely used in supply chain analytics. It is readily available in software tools, but is only as good as the data input. It is the best way to identify the "best" solution under limited resources.

Some types of math programming in SCM:

- Linear Programming (LPs)
- Integer Programming
- Mixed Integer and Linear Programming (MILPs)
- Non-linear Programming (NLPs)

Linear Programs

1. Decision Variables
   - What are you trying to decide?
   - What are their upper or lower bounds?
2. Formulate objective function
   - What are we trying to minimize or maximize?
   - Must include the decision variables and the form of the function determines the approach (linear for LP)
3. Formulate each constraint
   - What is my feasible region? What are my limits?
   - Must include the decision variable and will almost always be linear functions

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

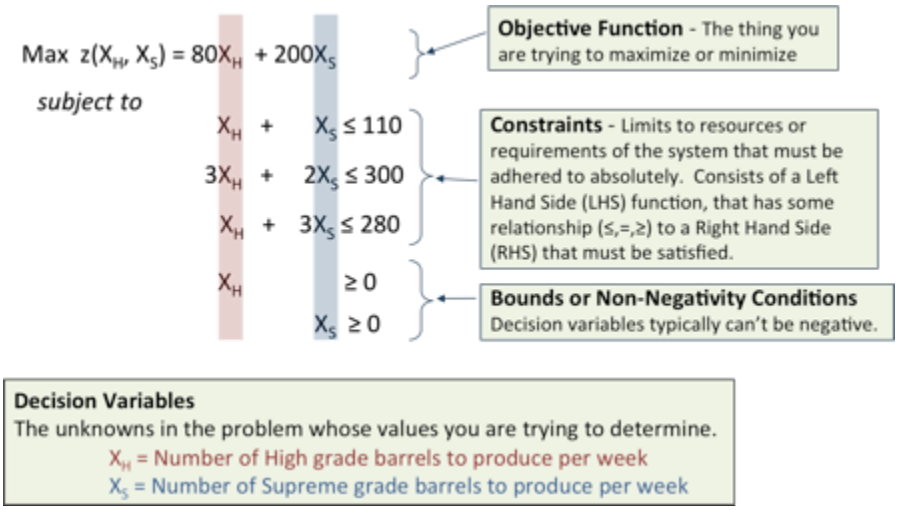


Figure: Linear Program Example

Solution

The solution of a linear program will always be in a "corner" of the Feasible Region:

- Linear constraints form a convex feasible region.
- The objective function determines in which corner is the solution.

The Feasible Region is defined by the constraints and the bounds on the decision variables (See Figure).

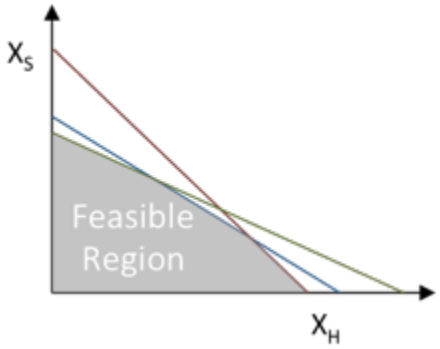


Figure: Graphical Representation of the Feasible Region

Analysis of the Results

Sometimes the original question is the least interesting one, it is often more interesting to dive a little deeper into the structure of the problem.

Additional Questions:

- Am I using all of my resources?
- Where do I have slack?

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

- Where I am constrained?
- How robust is my solution?

Sensitivity Analysis: what happens when data values are changed.

- Shadow Price or Dual Value of Constraint: What is the marginal gain in the profitability for an increase of one on the right hand side of the constraint?
- Slack Constraint – For a given solution, a constraint is not binding if total value of the left hand side is not equal to the right hand side value. Otherwise it is a binding constraint
- Binding Constraint – A constraint is binding if changing it also changes the optimal solution

Anomalies in Linear Programming

- Alternative or Multiple Optimal Solutions (see Figure)

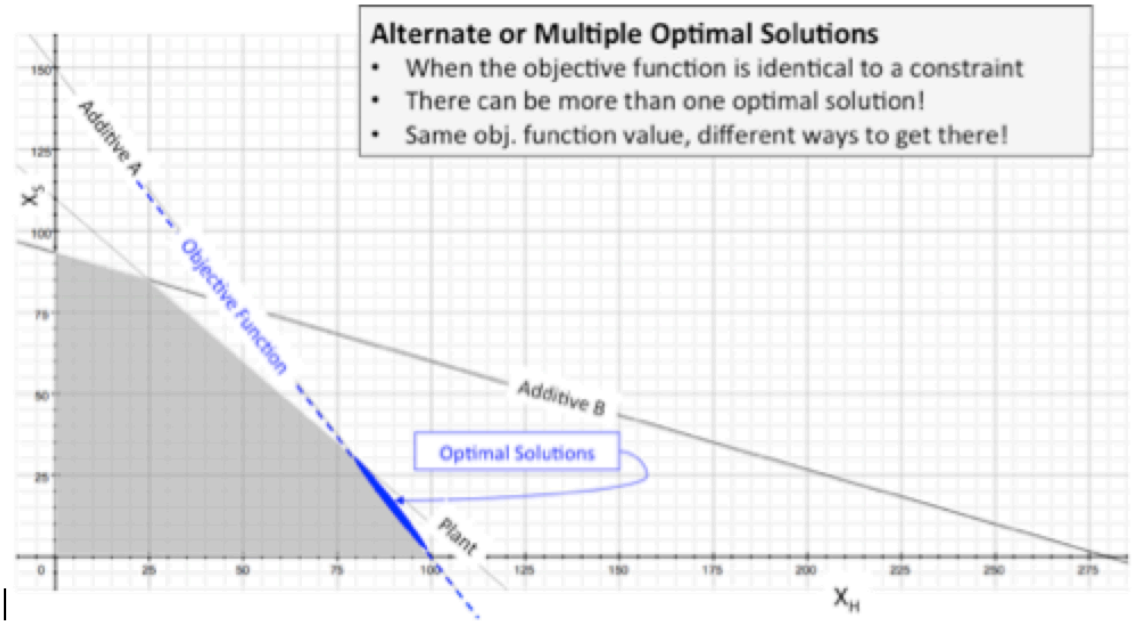


Figure: Alternative of Multiple Optimal Solutions

- Redundant Constraints - Does not effect the Feasible Region; it is redundant.
- Infeasibility - There are no points in the Feasible Region; constraints make the problem infeasible.

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Integer and Mixed Integer Programs

Although in some cases a linear program can provide an optimal solution, in many it cannot. For example in warehouse location selection, batch orders, or scheduling, fractional answers are not acceptable. In addition, the optimal solution cannot always be found by rounding the linear program solution. This is where integer programs are important. However, integer program solutions are never better than a linear program solution, they lower the objective function. In general, formulating integer programs is much harder than formulating linear program.

Numbers

- N = Natural, Whole or Counting numbers 1, 2, 3, 4
- Z-Integers = -3, -2, -1, 0, 1, 2, 3
- Q = Rational Number, continuous numbers = Any faction of integers ½ , -5/9
- R = Real Numbers = all Rational and Irrational Numbers, ex: e, pie, e
- Binary Integers = 0, 1

- To identify the solution in integer programs – the Feasible Region becomes a collection of points, it is no longer a convex hull (see Figure)
- In addition, cannot rely on "corner" solutions anymore – the solution space is much bigger

Mass enumeration - Unlike linear programs, integer programs can only take a finite number of integer values. This means that one approach is to enumerate all of these possibilities – calculating the objective function at each one and choosing the one with the optimal value. As the problem size increases, this approach is not feasible.

Formulating Integer Programs

To formulate an integer program, we follow the same approach for formulating linear programs – variables, constraints and objective. The only significant change is to formulating integer programs is in the definition of the variables. See example formulation in Figure below with integer specification.

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

$$\text{Max } z(X_{HL}, X_{SL}) = 8X_{HL} + 20X_{SL}$$

s.t.

Plant $X_{HL} + X_{SL} \leq 11$

Add. A $3X_{HL} + 2X_{SL} \leq 30$

Add. B $X_{HL} + 3X_{SL} \leq 28$

$X_{HL}, X_{SL} \geq 0$ Integers

Figure: Formulating an integer program

Binary Variables

Suppose you had the following formulation of a minimization problem subject to capacity at plants and meeting demand for individual products:

$$Min \quad z = \sum_i \sum_j c_{ij} x_{ij}$$

$$s.t.$$

$$\sum_i x_{ij} \leq C_j \quad \forall j$$

$$\sum_j x_{ij} \geq D_i \quad \forall i$$

$$x_{ij} \geq 0 \quad \forall ij$$

where:

x_{ij} = Number of units of product i made in plant j

c_{ij} = Cost per unit of product i made at plant j

C_j = Capacity in units at plant j

D_i = Demand for product i in units

We could add binary variables to this formulation to be able to model several different logical conditions. Binary variables are integer variables that can only take the values of 0 or 1. Generally, a positive decision (do something) is represented by 1 and the negative decision (do nothing) is represented by the value of 0.

Introducing a binary variable to this formulation, we would have:

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

$$Min \quad z = \sum_i \sum_j c_{ij} x_{ij} + \sum_j f_j y_j$$

$$s.t.$$

$$\sum_i x_{ij} \le C_j \quad \forall j$$

$$\sum_j x_{ij} \ge D_i \quad \forall i$$

$$\sum_i x_{ij} - M y_j \le 0 \quad \forall j$$

$$x_{ij} \ge 0 \quad \forall ij$$

$$y_j = \{0,1\}$$

where:
x_{ij} = Number of units of product i made in plant j
y_j = 1 if plant j is opened; = 0 otherwise
c_{ij} = Cost per unit of product i made at plant j
f_j = Fixed cost for producing at plant j
C_j = Capacity in units at plant j
D_i = Demand for product i in units
M = a big number (such as C_j in this case)

Note that the not only we have added the binary variable in the objective function, we have also added a new constraint (the third one). This is known as a linking constraint or a logical constraint. It is required to enforce an if-then condition in the model. Any positive value of x_{ij} will force the y_j variable to be equal to one. The "M" value is a big number – it should be as small as possible, but at least as big as the values of what sum of the x_{ij}'s can be. There are also more technical tricks that can be used to tighten this formulation.

We can also introduce **Either/Or Conditions**- where there is a choice between two constraints, only one of which has to hold; it ensures a minimum level, L_j, if y_j=1.

$$\sum_i x_{ij} - M y_j \le 0 \quad \forall j \qquad \sum_i x_{ij} - L_j y_j \ge 0 \quad \forall j$$

For example:

$$\sum_i x_{ij} \le C_j \quad \forall j$$

$$\sum_j x_{ij} \ge D_i \quad \forall i$$

$$\sum_i x_{ij} - M y_j \le 0 \quad \forall j$$

where:
x_{ij} = Number of units of product i made in plant j
y_j = 1 if plant j is opened; = 0 o.w.
M = a big number (such as C_j in this case)
C_j = Maximum capacity in units at plant j
L_j = Minimum level of production at plant j
D_i = Demand for product i in units

We need to add a constraint that ensures that if we DO use plant j, that the volume is between the minimum allowable level, L_j, and the maximum capacity, C_j. This is sometimes called an Either-Or condition.

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

$$\sum_i x_{ij} \le My_j \quad \forall j$$
$$\sum_i x_{ij} \ge L_j y_j \quad \forall j$$

where:

x_{ij} = Number of units of product i made in plant j
y_j = 1 if plant j is opened; = 0 o.w.
M = a big number (such as C_j in this case)
L_j = Minimum level of production at plant j

Finally, we can create a **Select From Condition** – that allows us to select the best *N* choices. Note that this can be formulated as "choose at least N" or "choose no more than N" by changing the inequality sign on the second constraint.

$$\sum_i x_{ij} - My_j \le 0 \quad \forall j \qquad\qquad \sum_j y_j \le N$$

Difference between Linear Programs and Integer Programs/Mixed Integer Programs

- Integer programs are much harder to solve since the solution space expands.
  - For linear programs, a correct formulation is generally a good formulation.
  - For integer programs a correct formulation is necessary but not sufficient to guarantee solvability.
- Integer programs require solving multiple linear programs to establish bounds – relaxing the Integer constraints.
- While it seems the most straightforward approach, you often can't just "round" the linear programs solution – it might not be feasible.
- When using integer (not binary) variables, solve the linear program first to see if it is sufficient.

Learning Objectives

- Learn the role of optimization.
- Understand how to optimize in unconstrained conditions.
- Identify how to find Extreme Point Solutions.
- Understand how to formulate problems with decision variables and resource constraints and graphically present them.
- Review how to interpret results and conduct sensitivity analysis.
- Understand the different "types" of numbers and how they change the approach to problems.
- Review the approach of formulating integer and mixed integer program problems and solving them.

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Advanced Optimization

Summary

This review concludes the learning portion on optimization with an overview of some frequently used advanced optimization models. Network models are key for supply chain professionals. The review begins by first defining the terminology used frequently in these networks. It then introduces common network problems including the Shortest Path, Traveling Salesman Problem (TSP), and Flow problems. These are used frequent in supply chain management and understanding when they arise and how to solve them is essential. We then introduce non-linear optimization, highlighting its differences with linear programming, and an overview of how to solve non-linear problems. The review concludes with practical recommendations of for conducting optimization, emphasizing that supply chain professionals should: know their problem, their team and their tool.

Key Concepts

Network Models

Network Terminology

- Node or vertices – a point (facility, DC, plant, region)
- Arc or edge – link between two nodes (roads, flows, etc.) may be directional
- Network or graph – a collection of nodes and arcs

Common Network Problems

Shortest Path – Easy & fast to solve (LP or special algorithms)

Result of shortest part problem is used as the base of a lot of other analysis. It connects physical to operational network.

- Given: One, origin, one destination.
- Find: Shortest path from single origin to single destination,
- Challenges: Time or distance? Impact of congestion or weather? How frequently should we update the network?
- Integrality is guaranteed.
- Caveat: Other specialized algorithms leverage the network structure to solve much faster.

Traveling Salesman Problem (TSP) – Hard to solve (heuristics)

- Given: One origin, many destinations, sequential stops, one vehicle.
- Objective: Starting from an origin node, find the minimum distance required to visit each node once and only one and return to the origin.

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

- Importance: TSP is at the core of all vehicle routing problems; local routing and last mile deliveries are both common and important.
- Challenges: It is exceptionally hard to solve exactly, due to its size; possible solutions increase exponentially with number of nodes.
- Primary approach: special algorithms for exact solutions (smaller problems) – Heuristics (many available).
    - Two examples: Nearest Neighbor, Cheapest Insertion

Nearest Neighbor Heuristic

This algorithm starts with the salesman at a random city and visits the nearest city until all have been visited. It yields a short tour, but typically not the optimal one.

- Select any node to be the active node.
- Connect the active node to the closest unconnected node; make that the new active node.
- If there are more unconnected nodes go to step 2, otherwise connect to the origin and end.

Cheapest Insertion Heuristic

One approach to the TSP is to start with a subtour – tour of small subsets of nodes, and extend this tour by inserting the remaining nodes one after the other until all nodes have been inserted. There are several decisions to be made in how to construct the initial tour, how to choose next node to be inserted, where to insert chosen node.

- Form a sub tour from the convex hull.
- Add to the tour the unconnected node that increases the cost the least; continue until all nodes are connected.

Flow Problems (Transportation & Transshipment) – Widely used (MILPs)

- Given: Multiple supply and demand nodes with fixed costs and capacities on nodes and/or arcs.
- Objective: Find the minimum cost flow of product from supply nodes that satisfy demand at destination nodes.
- Importance: Transportation problems are everywhere; transshipment problems are at the heart of larger supply chain network design models. In transportation problems, shipments are between two nodes. For transshipment problems, shipments may go through intermediary nodes, possibly changing mode of transport. Transshipment problems can be converted into transportation problems.
- Challenges: data requirements can be extensive; difficult to draw the line on “realism” vs. “practicality”.
- Primary approaches: mixed integer linear programs; some simulation – usually after optimization.

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Non-Linear Optimization

A nonlinear program is similar to a linear program in that it is composed of an objective function, general constraints, and variable bounds. The difference is that a nonlinear program includes at least one nonlinear function, which could be the objective function, or some or all of the constraints.

- Many systems are nonlinear – important to know how to handle them.
- Harder to solve than linear programs – lose 'corner' solutions (See Figure).
- Shape of objective function and constraints dictate approach and difficulty.

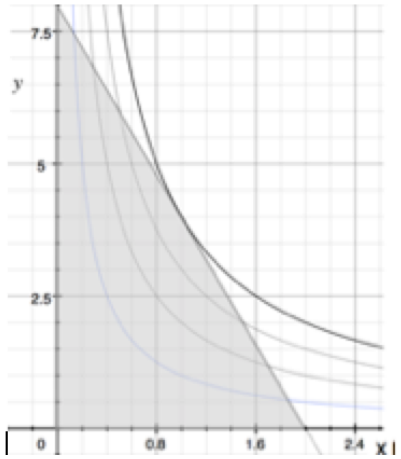


Figure: Example of NLP with linear constraint and non-linear objective function (z=xy).

Practical Tips for Optimization in Practice

- Know your problem:
  - Determining what to solve is rarely readily apparent or agreed upon by all stakeholders.
  - Establish and document the over-riding objective of a project early on.
- Level of detail & scope of model:
  - Models cannot fully represent reality, models will never represent all factors, determine problem boundaries and data aggregation levels.
- Input data:
  - Collecting data is hardest, least appreciated, and most time consuming task in an optimization project.
  - Data never complete clean, or totally correct.
  - Ever hour spent on data collection, cleaning and verification saves days later on in the project.
- Sensitivity and Robustness Analysis
  - These are all deterministic models – data assumed perfect & unchanging.
  - Optimization models will do anything for a dollar, yuan, peso, euro, etc.

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

    - Run multiple “what-if” scenarios changing uncertain input values and testing different conditions.
- Models vs. People (models don’t make decisions, people do!)
    - Optimization models are good at making trade-offs between complicated options and uncovering unexpected insights and solutions.
    - People are good at:
        - Considering intangible and non-quantifiable factors,
        - Identifying underlying patterns, and
        - Mining previous experience and insights.
        - Models should be used for Decision SUPPORT not for the decision.

Learning Objectives

- Introduction to advanced optimization methods.
- Understand the conditions and when to apply network models.
- Differentiate nonlinear optimization and when it should be used.
- Review recommendations for running optimization in practice – emphasizing importance of knowing the problem, team and tool.

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Algorithms and Approximations

Summary

In this lesson we will be reviewing Algorithms and approximations. The first half of the lesson will be a review of algorithms – which you technically have already be introduced to, but perhaps not in these terms. We will be reviewing the basics of algorithms, their components, and how they are used in our everyday problem solving! To demonstrate these we will be looking at a few common supply chain problems such as the Shortest Path problem, Traveling Salesman Problem, and Vehicle Routing Problem while applying the appropriate algorithm to solve them.

In this next part of the less we will be reviewing approximations. Approximations are good first steps in solving a problem because they require minimal data, allow for fast sensitivity analysis, and enable quick scoping of the solution space. Recognizing how to use approximation methods are important in supply chain management because commonly optimal solutions require large amounts of data and are time consuming to solve. So if that level of granularity is not needed, approximation methods can provide a basis to work from and to see whether further analysis is needed.

Algorithms

Algorithm - a process or set of rules to be followed in calculations or other problem-solving operations, especially by a computer.

Desired Properties of an Algorithm

- should be unambiguous
- require a defined set of inputs
- produce a defined set of outputs
- should terminate and produce a result, always stopping after a finite time.

Algorithm Example: find_max
Inputs:

- L = array of N integer variables
- v(i) = value of the i^{th} variable in the list

Algorithm:

1. set max = 0 and i = 1
2. select item i in the list
3. if v(i)>max, then set max=v(i)
4. if i<N, then set i = i+1 and go to step 2, otherwise go to step 5
5. end

Output:

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

maximum value in array L (max)

Shortest Path Problem

Objective: Find the shortest path in a network between two nodes
Importance: Its result is used as base for other analysis, and connects physical to operational network
Primary approaches:

- Standard Linear Programming (LP)
- Specialized Algorithms (Dijkstra's Algorithm)

Minimize: $\sum_i \sum_j c_{ij} x_{ij}$

Subject to:

$$\sum_i x_{ji} = 1 \qquad \forall\ j = s$$
$$\sum_i x_{ji} - \sum_i x_{ij} = 0 \qquad \forall\ j \neq s, j \neq t$$
$$\sum_i x_{ij} = 1 \qquad \forall\ j = t$$
$$x_{ij} \geq 0$$

where:

x_{ij} = Number of units flowing from node i to node j
c_{ij} = Cost per unit for flow from node i to node j
s = Source node – where flow starts
t = Terminal node – where flow ends

Dijkstra's Algorithm

Djikstra's algorithm (named after its discover, E.W. Dijkstra) solves the problem of finding the shortest path from a point in a graph (the source) to a destination.

L(j) = length of path from source node s to node j
P(j) = preceding node for j in the shortest path
S(j) = 1 if node j has been visited, = 0 otherwise
d(ij) = distance or cost from node i to node j

Inputs:

- Connected graph with nodes and arcs with positive costs, d(ij)
- Source (s) and Terminal (t) nodes

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Algorithm:

1. for all nodes in graph, set L()=∞, P()=Null, S()=0
2. set s to i, S(i)=1, and L(i)=0
3. For all nodes, j, directly connected (adjacent) to node i; if L(j) > L(i) + d(ij), then set L(j) = L(i) + d(ij) and P(j)=i
4. For all nodes where S()=0, select the node with lowest L() and set it to i, set S(i)=1
5. Is this node t, the terminal node? If so, go to end. If not, go to step 3
6. end – return L(t)

Output:
L(t) and P array

To find path from s to t, start at the end.

- Find P(t) – say it is j
- If j=source node, stop, otherwise, find P(j)
- keep tracing preceding nodes until you reach source node

Traveling Salesman Problem (TSP)

Starting from an origin node, find the minimum distance required to visit each node once and only once and return to the origin.

Nearest Neighbor Heuristic

1. Select any node to be the active node
2. Connect the active node to the closest unconnected node, make that the new active node.
3. If there are more unconnected nodes go to step 2, otherwise connect to the starting node and end.

2-Opt Heuristic

1. Identify pairs of arcs (i-j and k-l), where d(ij) + d(kl) > d(ik) + d(jl) – usually where they cross
2. Select the pair with the largest difference, and re-connect the arcs (i-k and j-l)
3. Continue until there are no more crossed arcs.

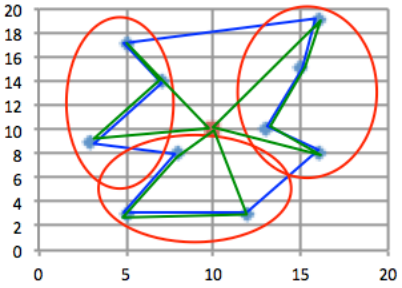
Vehicle Routing Problem

Find minimum cost tours from single origin to multiple destinations with varying demand using multiple capacitated vehicles.

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Heuristics

- Route first Cluster second
    - Any earlier TSP heuristic can be used

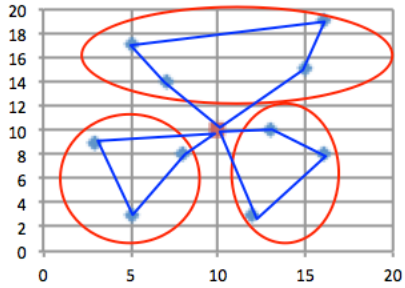


Optimal

- Mixed Integer Linear Program (MILP)
- Select optimal routes from potential set

Cluster first Route second

- Sweep Algorithm
- Savings (Clarke-Wright)



VRP Sweep Heuristic

Find minimum cost tours from DC to 10 destinations with demand as shown using up to 4 vehicles of capacity of 200 units.

Sweep Heuristic

1. Form a ray from the DC and select an angle and direction (CW vs CCW) to start
2. Select a new vehicle, j, that is empty, $w_j=0$, and has capacity, c_j.
3. Rotate the ray in selected direction until it hits a customer node, i, or reaches the starting point (go to step 5).
4. If the demand at i (D_i) plus current load already in the vehicle (w_j) is less than the vehicle capacity, add it to the vehicle, $w_j=D_i + w_j$ and go to step 3. Otherwise, close this vehicle, and go to step 2 to start a new tour.
5. Solve the TSP for each independent vehicle tour.

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Different starting points and directions can yield different solutions! Best to use a variety or a stack of heuristics.

Clark-Wright Savings Algorithm

The Clarke and Wright savings algorithm is one of the most known heuristic for VRP. It applies to problems for which the number of vehicles is not fixed (it is a decision variable)

- Start with a complete solution (out and back)
- Identify nodes to link to form a common tour by calculating the savings:

Example: joining node 1 & 2 into a single tour

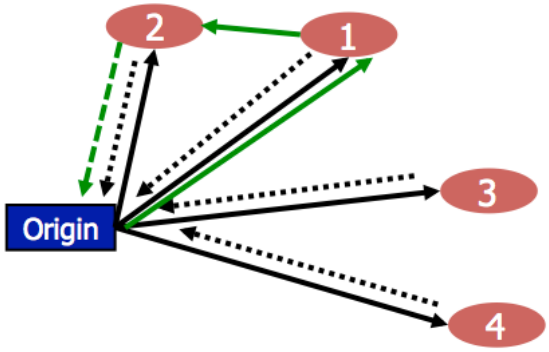
Current tours cost = $2c_{O1} + 2c_{O2}$

Joined tour costs = $c_{O1} + c_{12} + c_{2O}$

So, if $2c_{O1} + 2c_{O2} > c_{O1} + c_{12} + c_{2O}$ then join them

That is: $c_{O1} + c_{2O} - c_{12} > 0$

- This savings value can be calculated for every pair of nodes
- Run through the nodes pairing the ones with the highest savings first
- Need to make sure vehicle capacity is not violated
- Also, "interior tour" nodes cannot be added – must be on end



This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Savings Heuristic

1. Calculate savings $s_{i,j} = c_{O,i} + c_{O,j} - c_{i,j}$ for every pair (i, j) of demand nodes.
2. Rank and process the savings $s_{i,j}$ in descending order of magnitude.
3. For the savings $s_{i,j}$ under consideration, include arc (i, j) in a route only if:
    - No route or vehicle constraints will be violated by adding it in a route and
    - Nodes i and j are first or last nodes to/from the origin in their current route.
4. If the savings list has not been exhausted, return to Step 3, processing the next entry in the list; otherwise, stop.

Solving VRP with MILP

Potential routes are an input and can consider different costs, not just distance.

Mixed integer linear program used to select routes:

- Each column is a route
- Each row is a node/stop
- Total cost of each route is included

$$Min \sum_j C_j Y_j$$

$$s.t.$$

$$\sum_{j=1}^{J} a_{ij} Y_j \geq D_i \ \forall i$$

$$\sum_{j=1}^{J} Y_j \leq V$$

$$Y_j = \{0,1\} \ \forall j$$

Indices

Demand nodes i

Vehicle routes j

Input Data

C_j = Total cost of route j ($)

D_i = Demand at node i (units)

V = Maximum vehicles

a_{ij} = 1 if node i is in route j;

= 0 otherwise

Decision Variables

Y_j = 1 if route j is used,

=0 otherwise

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Approximation Methods

In this second half we will discuss examples of approximation and estimation. In particular we will review estimation of One-to-Many Distribution through linehaul distance, traveling salesman and vehicle routing problems.

Approximation: a value or quantity that is nearly but not exactly correct.

Estimation: a rough calculation of the value, number, quantity, or extent of something. synonyms: estimate, approximation, rough calculation, rough guess, evaluation, back-of-the envelope

Why use approximation methods?

- Faster than more exact or precise methods,
- Uses minimal amounts of data, and
- Can determine if more analysis is needed: Goldilocks Principle: Too big, Too little, Just right.

Always try to estimate a solution prior to analysis!

Quick Estimation

Simple Estimation Rules:

1. Break the problem into pieces that you can estimate or determine directly
2. Estimate or calculate each piece independently to within an order of magnitude
3. Combine the pieces back together paying attention to units

Example:

How many piano tuners are there in Chicago?"

- There are approximately 9,000,000 people living in Chicago.
- On average, there are two persons in each household in Chicago.
- Roughly one household in twenty has a piano that is tuned regularly.
- Pianos that are tuned regularly are tuned on average about once per year.
- It takes a piano tuner about two hours to tune a piano, including travel time.
- Each piano tuner works eight hours in a day, five days in a week, and 50 weeks in a year.

Tunings per Year = (9,000,000 ppl) ÷ (2 ppl/hh) × (1 piano/20 hh) × (1 tuning/piano/year) = 225,000
Tunings per Tuner per Year = (50 wks/yr) × (5 day/wk) × (8 hrs/day) ÷ (2 hrs to tune) = 1000
Number of Piano Tuners = (225,000 tunings per year) ÷ (1000 tunings per year per tuner) = 225
Actual Number = 290

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

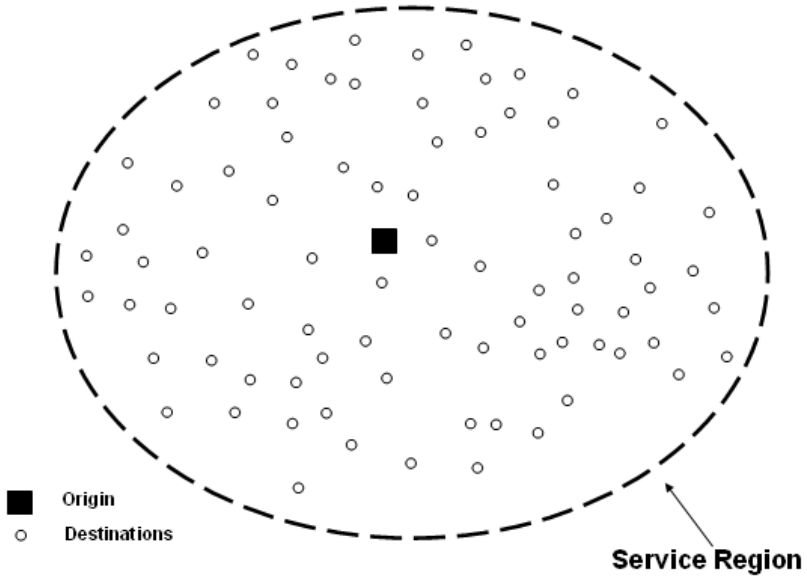
Estimation of One to Many Distribution

Single Distribution Center:

- Products originate from one origin
- Products are demanded at many destinations
- All destinations are within a specified Service Region
- Ignore inventory (same day delivery)

Assumptions:

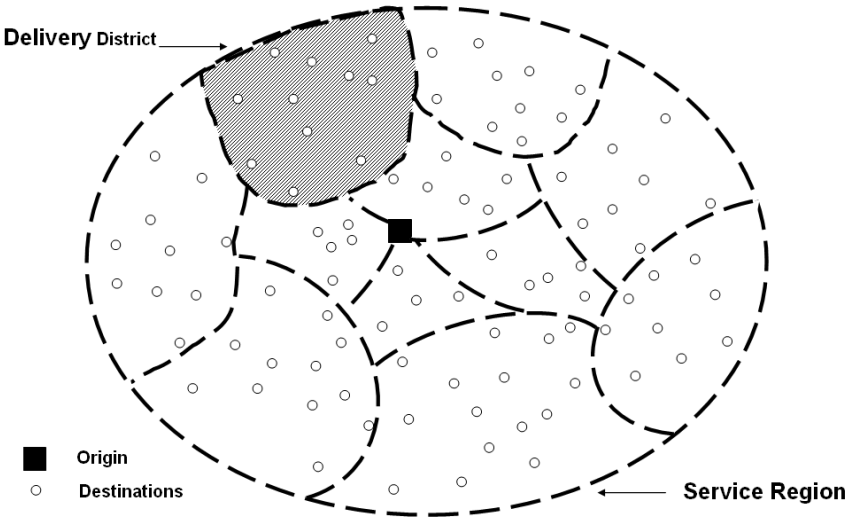
- Vehicles are homogenous
- Same capacity, Q_{MAX}
- Fleet size is constant



This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

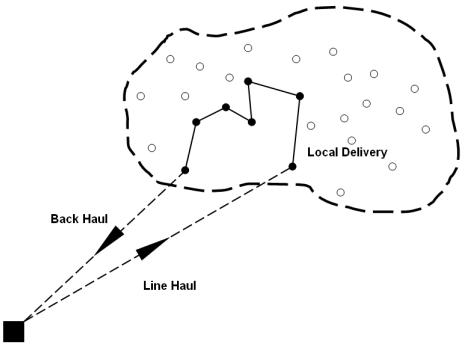
Finding the estimated total distance:

- Divide the Service Region into Delivery Districts
- Estimate the distance required to service each district



Route to serve a specific district:

- Line haul from origin to the 1st customer in the district
- Local delivery from 1st to last customer in the district
- Back haul (empty) from the last customer to the origin



$$d_{TOUR} \approx 2d_{LineHaul} + d_{Local}$$

$d_{LineHaul}$ = Distance from origin to center of gravity (centroid) of delivery district
d_{Local} = Local delivery between customers in one district

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

How do we estimate distances?

- Point to Point
- Routing or within a Tour

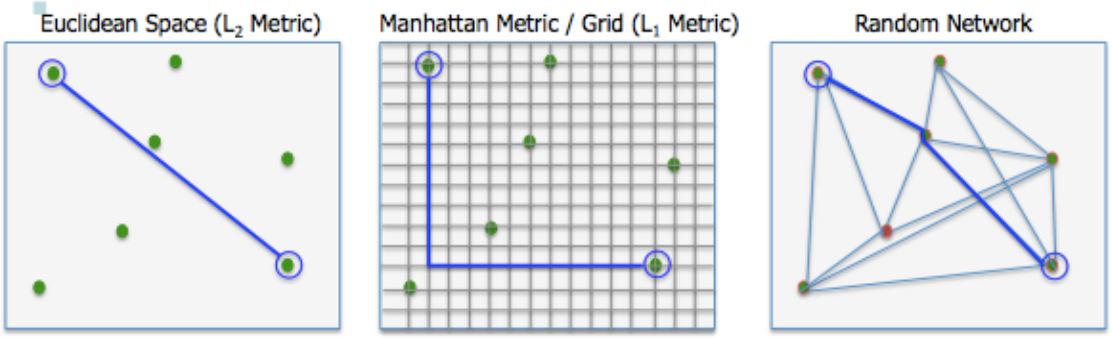
Estimating Point to Point Distances

Depends on the topography of the underlying region

Euclidean Space: $d_{A-B} = \sqrt{[(x_A-x_B)^2+(y_A-y_B)^2]}$

Grid: $d_{A-B} = |x_A-x_B|+|y_A-y_B|$

Random Network: different approach



For Random (real) Networks use: $D_{A-B} = k_{CF}\, d_{A-B}$

Find d_{A-B} - the "as crow flies" distance.

- Euclidean: for really short distances
  - $d_{A-B} = SQRT((x_A-x_B)^2+(y_A-y_B)^2)$
- Great Circle: for locations within the same hemisphere
  - $d_{A-B} = 3959(\arccos[\sin[LAT_A]\sin[LAT_B] +\cos[LAT_A]\cos[LAT_B]\cos[LONG_A-LONG_B]])$
- Where:
  - LAT_i = Latitude of point i in radians
  - $LONG_i$ = Longitude of point i in radians
  - Radians = (Angle in Degrees)$(\pi/180^o)$

Apply an appropriate circuity factor (k_{CF})

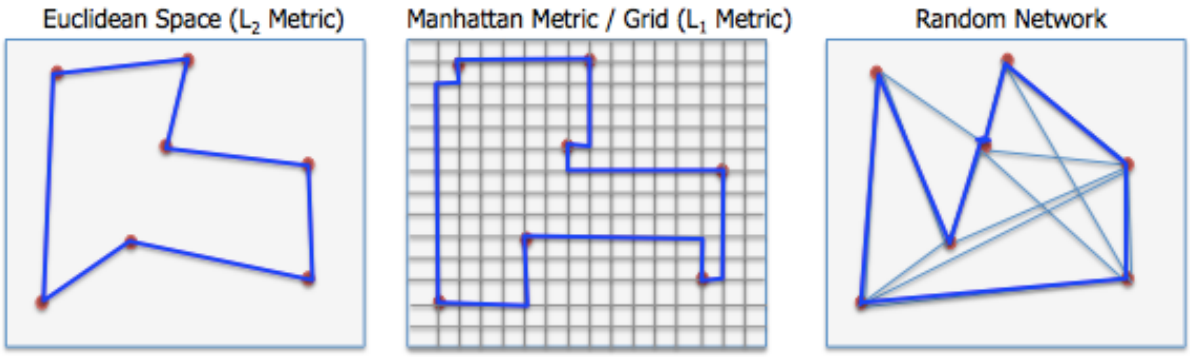
- How do you get this value?
- What do you think the ranges are?
- What are some cautions for this approach?

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Estimating Local Route Distances

Traveling Salesman Problem

- Starting from an origin, find the minimum distance required to visit each destination once and only once and return to origin.
- The expected TSP distance, d_{TSP}, is proportional to √(nA) where n= number of stops and A=area of district
- The estimation factor (k_{TSP}) is a function of the topology



One to Many System

What can we say about the expected TSP distance to cover **n** stops in district with an area of **A**?
A good approximation, assuming a "fairly compact and fairly convex" area, is:

A = Area of district
n = Number of stops in district
δ = Density (# stops/Area)
k_{TSP} = TSP network factor (unitless)
d_{TSP}= Traveling Salesman Distance
d_{stop}= Average distance per stop

$$d_{TSP} \approx k_{TSP}\sqrt{nA} = k_{TSP}\sqrt{n\left(\frac{n}{\delta}\right)} = k_{TSP}\left(\frac{n}{\sqrt{\delta}}\right)$$

$$d_{stop} \approx \frac{k_{TSP}\sqrt{nA}}{n} = k_{TSP}\sqrt{\frac{A}{n}}$$

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

What values of k_{TSP} should we use?

- Lots of research on this for L_1 and L_2 networks - depends on district shape, approach to routing, etc.
- Euclidean (L_2) Networks
  - k_{TSP} = 0.57 to 0.99 depending on clustering & size of N (MAPE~4%, MPE~-1%)
  - k_{TSP}=0.765 commonly used and is a good approximation!
- Grid (L_1) Networks
  - k_{TSP} = 0.97 to 1.15 depending on clustering and partitioning of district

Estimating Vehicle Tour Distances

Finding the total distance traveled on all tours, where:

- l = number of tours
- c = number of customer stops per tour and
- n=total number of stops = c*l

$$d_{TOUR} = 2d_{LineHaul} + \frac{ck_{TSP}}{\sqrt{\delta}}$$

$$d_{AllTours} = ld_{TOUR} = 2ld_{LineHaul} + \frac{nk_{TSP}}{\sqrt{\delta}}$$

Minimize number of tours by maximizing vehicle capacity

$$l = \left[\frac{D}{Q_{MAX}}\right]^+$$

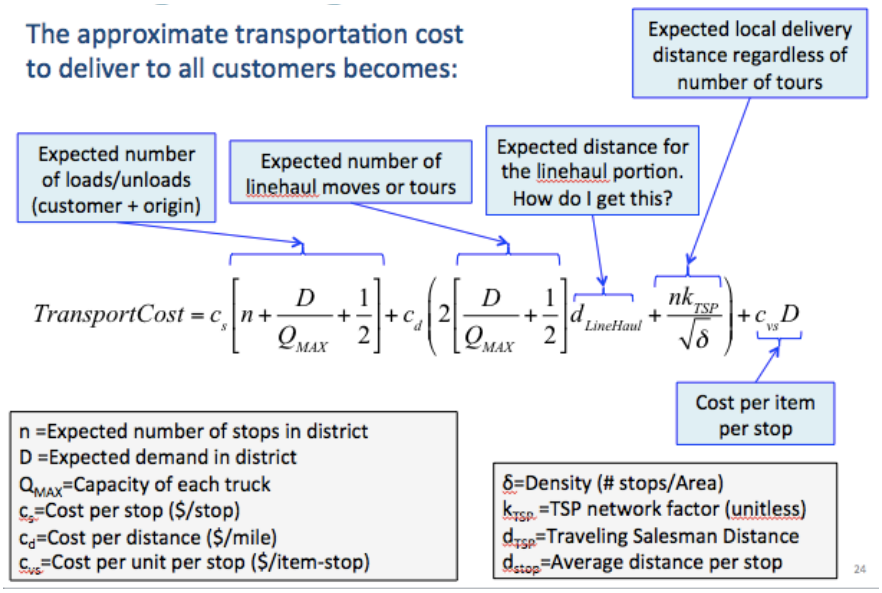
$$d_{AllTours} = 2\left[\frac{D}{Q_{MAX}}\right]^+ d_{LineHaul} + \frac{nk_{TSP}}{\sqrt{\delta}}$$

$[x]^+$ = lowest integer value > x. This is a step function

Estimate this with continuous function:

$[x]^+ \sim x + ½$

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.



Key Points

- Review the basis and components of algorithms
- Recognize desired properties of an algorithm
- Review different network algorithms
- Recognize how to solve the Shortest Path Problem
- Recognize which algorithms to use for the Traveling Salesman Problem
- Recognize how to solve a Vehicle Routing Problem (Cluster First – Route Second)
- Review how to use approximations
- Recognize steps to quick estimation

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Distributions and Probability

Summary

We review two very important topics in supply chain management: probability and distributions. Probability is an often-reoccurring theme in supply chain management due to the common conditions of uncertainty. On a given day, a store might sell 2 units of a product, on another, 50. To explore this, the probability review includes an introduction of probability theory, probability laws, and proper notation. Summary or descriptive statistics are shown for capturing central tendency and the dispersion of a distribution. We also introduce two theoretical discrete distributions: Uniform and Poisson.

We then introduce three common continuous distributions: Uniform, Normal, and Triangle. The review then goes through the difference between discrete vs. continuous distributions and how to recognize these differences. The remainder of the review is a exploration into each type of distribution, what they look like graphically and what are the probability density function and cumulative density function of each.

Key Concepts

Probability

Probability defines the extent to which something is probable, or the likelihood of an event happening. It is measured by the ratio of the case to the total number of cases possible.

Probability Theory

- Mathematical framework for analyzing random events or experiments.
- Experiments are events we cannot predict with certainty (e.g., weekly sales at a store, flipping a coin, drawing a card from a deck, etc.).
- Events are a specific outcome from an experiment (e.g., selling less than 10 items in a week, getting 3 heads in a row, drawing a red card, etc.)

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Notation

- P (A) – the probability that event A occurs
- P (A') = complement of P (A) – probability some other event that is not A occurs. This is also the probability that something other than A happens.

$\cup$ = Union of sets (OR)

$\cap$ = Intersection of sets (AND)

$\varnothing$ = Null or Empty set

Probability Laws

1. The probability of any event is between 0 and 1, that is $0 \leq P(A) \leq 1$
2. If A and B are mutually exclusive events, then P (A or B) = P (A **U** B) = P (A) + P (B)
3. If A and B are any two events, then

$$P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)} = \frac{P(A \cap B)}{P(B)}$$

Where P(A I B) is the **conditional probability** of A occurring given B has already occurred.

4. If A and B are independent events, then

$$P(A \mid B) = \mathrm{P(A)}$$
$$P(A \text{ and } B) = P(A \cap B) = P(A \mid B)P(B) = P(A) \times P(B)$$

Where events A and B are independent if knowing that B occurred does not influence the probability of A occurring.

Summary statistics

Descriptive or summary statistics play a significant role in the interpretation, presentation, and organization of data. It characterizes a set of data. There are many ways that we can characterize a data set, we focused on two: Central Tendency and Dispersion or Spread.

Central Tendency

This is, in rough terms, the "most likely" value of the distribution. It can be formally measured in a number of different ways to include:

- **Mode** – the specific value that appears most frequently
- **Median** – the value in the "middle" of a distribution that separates the lower from the higher half. This is also called the 50th percentile value.
- **Mean** (μ) – the sum of values multiplied by their probability (called the expected value). This is also the sum of values divided by the total number of observations (called the average).

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

$$E[X] = \bar{x} = \mu = \sum_{i=1}^{n} p_i x_i$$

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Dispersion or Spread

This captures the degree to which the observations "differ" from each other. The more common dispersion metrics are:

- **Range** – the maximum value minus the minimum value.
- **Inner Quartiles** – 75th percentile value minus the 25th percentile value - captures the "central half" of the entire distribution.
- **Variance** (σ^2) – the expected value of the squared deviation around the mean; also called the Second Moment around the mean

$$Var[X] = \sigma^2 = \sum_{i=1}^{n} p_i \left(x_i - \bar{x}\right)^2 = \sum_{i=1}^{n} p_i \left(x_i - \mu\right)^2$$

- **Standard Deviation** (σ) – the square root of the variance. This puts it in the same units as the expected value or mean.
- **Coefficient of Variation** (CV) – the ratio of the standard deviation over the mean = σ/μ. This is a common comparable metric of dispersion across different distributions. As a general rule:
  - $0 \le CV \le 0.75$, low variability
  - $0.75 \le CV \le 1.33$, moderate variability
  - $CV > 1.33$, high variability

Population versus Sample Variance

In practice, we usually do not know the true mean of a population. Instead, we need to estimate the mean from a sample of data pulled from the population. When calculating the variance, it is important to know whether we are using all of the data from the entire population or just using a sample of the population's data. In the first case we want to find the **population variance** while in the second case we want to find the **sample variance** .

The only differences between calculating the population versus the sample variances (and thus their corresponding standard deviations) is that for the population variance, σ^2, we divide the sum of the observations by *n* (the number of observations) while for the sample variance, s^2, we divide by *n-1*.

$$\sigma^2 = \frac{\sum_{i=1}^{n} \left(x_i - \mu\right)^2}{n} \qquad s^2 = \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}{n-1}$$

Note that the sample variance will be slightly larger than the population variance for small values of n. As n gets larger, this difference essentially disappears. The reason for the use n-1 is due to having to use a degree of freedom in calculating the average (x bar) from the same sample that we are estimating the variance. It leads to an unbiased estimate of the population variance. In practice, you should just use the sample variance and standard deviation unless you are dealing with specific probabilities, like flipping a coin.

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Spreadsheet Functions for Summary Statistics

All of these summary statistics can be calculated quite easily in any spreadsheet tool. The table below summarizes the functions for three widely used packages.

Function	Microsoft Excel	Google Sheets	LibreOffice->Calc
Minimum	=MIN(array)	=MINA(array)	=MIN(array)
Median	=MEDIAN(array)	=MEDIAN(array)	=MEDIAN(array)
Mode	=MODE(array)	=MODE(array)	=MODE(array)
Mean (μ)	=AVERAGE(array)	=AVERAGE(array)	=AVERAGE(array)
Maximum	=MAX(array)	=MAX(array)	=MAX(array)
Percentile	=PERCENTILE.INC(array, k)	=PERCENTILE(array, percentile)	=PERCENTILE.INC(array, alpha)
Population Variance (σ^2)	=VAR.P(array)	=VARP(array)	=VAR.P(array)
Sample Variance (σ^2)	=VAR.S(array)	=VAR(array)	=VAR.S(array)
Pop. Std Deviation (σ)	=STDEV.P(array)	=STDEVP(array)	=STDEV.P(array)
Sample Std Deviation (σ)	=STDEV(array)	=STDEV(array)	=STDEV.S(array)

Table: Spreadsheet Functions for Descriptive Statistics

Probability Distributions

Probability distributions can either be empirical (based on actual data) or theoretical (based on a mathematical form). Determining which is best depends on the objective of the analysis. Empirical distributions follow past history while theoretical distributions follow an underlying mathematical function. Theoretical distributions do tend to allow for more robust modeling since the empirical distributions can be thought of as a sampling of the population data. The theoretical distribution can be seen as better describing the assumed population distribution. Typically, we look for the theoretical distribution that best fits the data

We presented five distributions. Two are discrete (Uniform and Poisson) and three are continuous (Uniform, Normal, and Triangle). Each is summarized in turn.

Discrete Uniform Distribution ~U(a,b)

Finite number (N) of values observed with a minimum value of **a** and a maximum value of **b**. The probability of each possible value is 1/N where N= b – a +1

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Probability Mass Function (pmf):

$$P\left[X = x\right] = f(x \mid a,b) = \begin{cases} \frac{1}{n} & \text{for } a \le x \le b \\ 0 & \text{otherwise} \end{cases}$$

Summary Metrics

- Mean = (a + b) / 2
- Median = (a + b) / 2
- Mode N/A (all values are equally likely)
- Variance = $((b - a + 1)^2 - 1)/12$

Poisson Distribution ~P(λ)

Discrete frequency distribution that gives the probability of a number of independent events occurring in a fixed time where the parameter **λ** = mean = variance. Widely used to model arrivals, slow moving inventory, etc. Note that the distribution only contains non-negative integers and can capture non-symmetric distributions. As the number of observations increase, the distribution becomes "bell like" and approximates the Normal Distribution.

Probability Mass Function (pmf):

$$P\left[X = x\right] = f(x \mid \lambda) = \begin{cases} \frac{e^{-\lambda}\lambda^{x}}{x!} & \text{for } x = 0,1,2,\ldots \\ 0 & \text{otherwise} \end{cases}$$

where

- e = Euler's number ~ 2.71828 . . .
- λ = mean value (parameter)
- x! = factorial of x, e.g., 3! = 3×2×1=6 and 0!=1

Summary Metrics

- Mean = λ
- Median ≈ $\lfloor(\lambda + 1/3 - 0.02/\lambda)\rfloor$
- Mode = $\lfloor\lambda\rfloor$
- Variance = λ

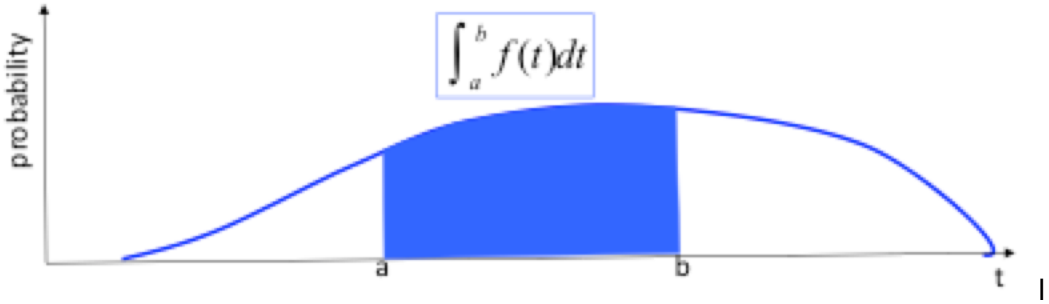
Spreadsheet	Function	Prob 1	Prob 2
Microsoft Excel	=POISSON.DIST(x, mean, cumulative)	=POISSON.DIST(0, 2.2, 0)	=POISSON.DIST(2, 2.2, 1)
Google Sheets	=POISSON(x, mean, cumulative)	=POISSON(0, 2.2, 0)	=POISSON(2, 2.2, 1)
LibreOffice->Calc	=POISSON(Number; Mean; C)	=POISSON(0; 2.2; 0)	=POISSON(2; 2.2; 1)

Table: Spreadsheet Functions for Poisson distribution

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Probability Density Function (pdf)

The pdf is function of a continuous variable. The probability that X lies between values a and b is equal to area under the curve between a and b. Total area under the curve equals 1, but the P(X= t) = 0 for any specific value of t.



Cumulative Density Function (cdf)

- F(t) = P(X ≤ t) or the probability that X does not exceed t
- ≤ F (t) ≤ 1.0
- F (b) ≥ F (1) if b > a – it is increasing

Simple rules

- P (X ≤ t) = F(t)
- P (X > t) = 1 – F(t)
- P(c ≤ X ≤ d) = f (d) – F(c)
- P (X = t) = 0

Continuous Uniform Distribution ~U(a,b)

Sometimes also called a rectangular distribution

- "X if uniformly distributed over the range a to b, or X~ U (a,b)".

pdf: $$f(t \mid a,b) = \begin{cases} \frac{1}{b-a} & if\ a \le t \le b \\ 0 & otherwise \end{cases}$$

cdf:

$$F(t \mid a,b) = \begin{cases} 0 & if\ t < a \\ \frac{t-a}{b-a} & if\ a \le t \le b \\ 1 & if\ t > b \end{cases}$$

Summary Metrics

- Mean = (a+b)/2
- Median = (a+b)/2
- Mode N/A all values equally likely
- Variance = $(b-a)^2/12$

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Normal Distribution ~N(μ,σ)

Widely used bell-shaped, symmetric continuous distribution with **mean μ** and **standard deviation σ**. Most commonly used distribution in practice.

Summary Metrics

- Mean = μ
- Median = μ
- Mode =μ
- Variance = σ^2

pdf: $$f(x \mid \mu, \sigma) = \frac{1}{(2\pi)^{1/2}\sigma} e^{\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]}$$

Common dispersion values ~N (μ, σ)

- P (X w/in 1σ around μ) = 0.6826
- P (X w/in 2σ around μ) = 0.9544
- P (X w/in 3σ around μ) = 0.9974
- +/- 1.65 σ around μ = 0.900
- +/- 1.96 σ around μ = 0.950
- +/- 2.81 σ around μ = 0.995

Unit or Standard Normal Distribution Z~N(0,1)

- The transformation from any ~N(μ,σ) to the unit normal distribution = Z=(x- μ)/σ
- Z score (standard score) gives the number of standard deviations away from the mean
- Allows for use of standard tables and is used extensively in inventory theory for setting safety stock

Function	Microsoft Excel	Google Sheets	LibreOffice->Calc
cdf of Normal Distribution	=NORM.DIST(X, μ, σ, 1)	=NORMDIST (X, μ, σ, 1)	=NORM.DIST (X, μ, σ, 1)
pdf of Normal Distribution	=NORM.DIST(X, μ, σ, 0)	=NORMDIST (X, μ, σ, 0)	=NORM.DIST (X, , μ, σ, 0)
Inverse of Normal cdf	=NORM.INV(Probability, μ, σ)	=NORMINV (Probability, μ, σ)	=NORM.INV (Probability, μ, σ)
Standard Normal cdf	=NORM.S.DIST(z,1)	=NORMSDIST (z)	=NORM.S.DIST (z,1)
Inverse Standard Normal cdf	=NORM.S.INV(Probability)	=NORMSINV (Probability)	=NORM.S.INV (Probability)

Table: Spreadsheet Functions for Normal Distribution

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Triangle Distribution ~T(a,b,c)

This is a continuous distribution with a minimum value of **a**, maximum value of **b**, and a mode of **c**. It is a good distribution to use when dealing with an anecdotal or unknown distribution. It can also handle non-symmetric distributions with long tails.

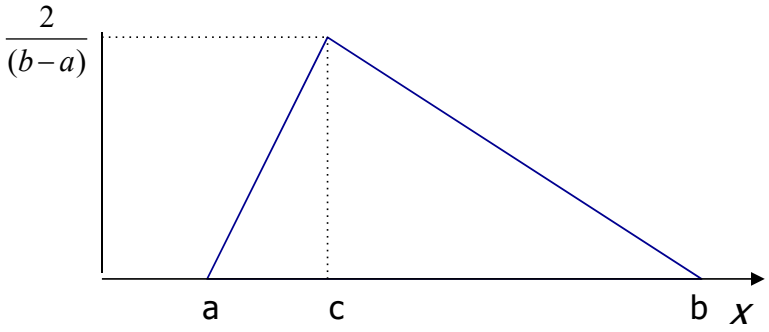


Figure 1 Triangle Distribution

pdf:

$$f(x)=\begin{cases}0 & x<a\\ \dfrac{2(x-a)}{(b-a)(c-a)} & a\le x\le c\\ \dfrac{2(b-x)}{(b-a)(b-c)} & c\le x\le b\\ 0 & x>b\end{cases}$$

cdf:

$$\begin{cases}0 & \text{for } x\le a,\\ \dfrac{(x-a)^2}{(b-a)(c-a)} & \text{for } a<x\le c,\\ 1-\dfrac{(b-x)^2}{(b-a)(b-c)} & \text{for } c<x<b,\\ 1 & \text{for } b\le x.\end{cases}$$

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Summary Metrics

$$E[x] = \frac{a+b+c}{3}$$

$$Var[x] = \left(\frac{1}{18}\right)\left(a^2+b^2+c^2-ab-ac-bc\right)$$

$$P[x>d] = \left(\frac{(b-d)^2}{(b-a)(b-c)}\right) \quad for\ c \le d \le b$$

$$d = b - \sqrt{P[x>d](b-a)(b-c)} \quad for\ c \le d \le b$$

Differences between Continuous and Discrete Distributions

Just like variables, distributions can be classified into continuous (pdf) and discrete (pmf) probability distributions. While discrete distributions have a probability for each outcome, the probability for a specific point in a continuous distribution makes no sense and is zero. Instead for continuous distributions we look for the probability of a random variable falling within a specific interval. Continuous distributions use a function or formula to describe the data and thus instead of summing (as we did for discrete distributions) to find the probability, we integrate over the region.

Discrete Distributions

$$\mu = E(X) = \sum_{i=1}^{n} p_i x_i$$

$$\sigma^2 = \sum_{i=1}^{n} p_i (x_i - \mu)^2$$

Continuous Distributions

$$\mu = \int_a^b t \cdot f(t)\,dt$$

$$\sigma^2 = \int_a^b (t-\mu)^2 \cdot f(t)\,dt$$

Learning Objectives

- Understand probabilities, importance and application in daily operations and extreme circumstances.
- Understand and apply descriptive statistics.
- Understand difference between continuous vs. discrete random variable distributions.
- Review major distributions: Uniform (discrete and continuous), Poisson, Normal and Triangle.
- Understand the difference between discrete vs. continuous distributions.
- Recognize and apply probability mass functions (pmf), probability density functions (pdf), and cumulative density functions (cdf).

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Regression

Summary

In this review we expand our tool set of predictive models to include ordinary least squares regression. This equips us with the tools to build, run and interpret a regression model. We are first introduced with how to work with multiple variables and their interaction. This includes correlation and covariance, which measures how two variables change together. As we review how to work with multiple variables, it is important to keep in mind that the data sets supply chain managers will deal with are largely samples, not a population. This means that the subset of data must be representative of the population. The later part of the lesson introduces hypothesis testing, which allows us to answer inferences about the data.

We then tackle linear regression. Regression is a very important practice for supply chain professionals because it allows us to take multiple random variables and find relationships between them. In some ways, regression becomes more of an art than a science. There are four main steps to regression: choosing with independent variables to include, collecting data, running the regression, and analyzing the output (the most important step).

Key Concepts

Multiple Random Variables

Most situations in practice involve the use and interaction of multiple random variables or some combination of random variables. We need to be able to measure the relationship between these RVs as well as understand how they interact.

Covariance and Correlation

Covariance and correlation measure a certain kind of dependence between variables. If random variables are positively correlated, higher than average values of X are likely to occur with higher than average values of Y. For negatively correlated random variables, higher than average values are likely to occur with lower than average values of Y. It is important to remember as the old, but necessary saying goes: correlation does not equal causality. This means that you are finding a mathematical relationship – not a causal one.

Correlation Coefficient: is used to standardize the covariance in order to better interpret. It is a measure between -1 and +1 that indicates the degree and direction of the relationship between two random variables or sets of data.

$$\text{CORR}(X,Y) = \frac{COV(X,Y)}{\sigma_X \sigma_Y}$$

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Covariance

$$\text{Cov}(X,Y) = \sum_{i=1}^{n} P(X = x_i, Y = y_i)[(x_i - \mu_X)(y_i - \mu_Y)] = \frac{\sum_{i=1}^{n}(x_i - \mu_X)(y_i - \mu_Y)}{n}$$

Spreadsheet Functions

Function	Microsoft Excel	Google Sheets	LibreOffice->Calc
Covariance	=COVAR(array,array)	=COVAR(array,array)	=COVAR(array;array)
Correlation	=CORREL(array,array)	=CORREL(array, array)	=CORREL(array;array)

Linear Function of Random Variables

A linear relationship exists between X and Y when a one-unit change in X causes Y to change by a fixed amount, regardless of how large or small X is. Formally, this is: Y = aX + b.
The summary statistics of a linear function of a Random Variable are:

Expected value: $E[Y] = \mu_Y = a\mu_X + b$
Variance: $VAR[Y] = \sigma^2_Y = a^2\sigma^2_X$
Standard Deviation: $\sigma_Y = |a|\sigma_X$

Sums of Random Variables

IF X and Y are independent random variables where W = aX + bY, then the summary statistics are:

Expected value: $E[W] = a\mu_X + b\mu_Y$
Variance: $VAR[W] = a^2\sigma^2_X + b^2\sigma^2_Y + 2abCOV(X,Y)$
$= a^2\sigma^2_X + b^2\sigma^2_Y + 2ab\sigma_X\sigma_Y CORR(X,Y)$
Standard Deviation: $\sigma_W = \sqrt{VAR[W]}$

These relations hold for any distribution of X and Y. However, if X and Y are ~N, then W is ~N as well!

Central Limit Theorem

Central limit theorem states that the sample distribution of the mean of any independent random variable will be normal or nearly normal, if the sample size is large enough. Large enough is based on a few factors – one is accuracy (more sample points will be required) as well as the shape of the underlying population. Many statisticians suggest that 30, sometimes 40, is considered large enough. This is important because is doesn't matter what distributions the random variable follows.

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Can be interpreted as follows:

- $X_i,..X_n$ are iid with mean=μ and standard deviation = σ
  - The sum of the n random variables is $S_n=\Sigma X_i$
  - The mean of the n random variables is $\bar{X} = S_n/n$
- Then, if **n** is "large"(say > 30)
  - S_n is Normally distributed with mean = $\mathbf{n\mu}$ and standard deviation $\mathbf{\sigma\sqrt{n}}$
  - $\bar{X}$ is Normally distributed with mean = $\mathbf{\mu}$ and standard deviation $\mathbf{\sigma/\sqrt{n}}$

Inference Testing

Sampling

We need to know something about the sample to make inferences about the population. The inference is a conclusion reached on the basis of evidence and reasoning. To make inferences we need to ask testable questions such as if the data fits a specific distribution or are two variables correlated? To understand these questions and more – we need to understand sampling of a population. If sampling is done correctly, the sample mean should be an estimator of the population mean as well as corresponding parameters.

- Population: is the entire set of units of observation
- Sample: subset of the population.
- Parameter: describes the distribution of random variable.
- Random Sample: is a sample selected from the population so that each item is equally likely.

Things to keep in mind

- X is a rv ~?(μ,σ, . . .) for the entire population
- $X_1, X_2, \ldots X_n$ are iid
- $\bar{X}$ is an estimate of the population parameter, the mean or μ
- Remember that $\bar{X}$ is also a rv by itself!
- $x_1, x_2, \ldots x_n$, are the realizations or observations of rv X
- $\bar{x}$ is the sample statistic – the mean
- We want to find how $\bar{x}$ relates to $\bar{X}$ relates to μ

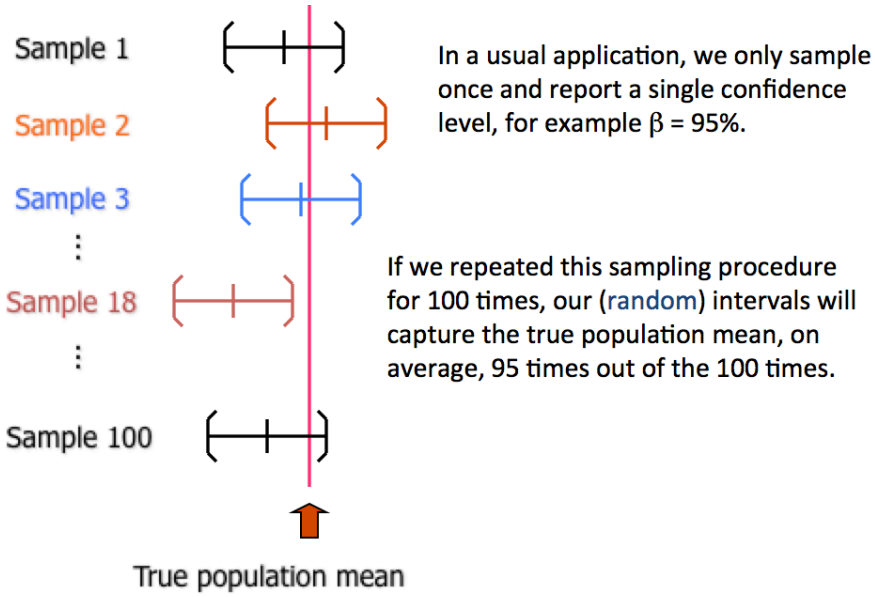
Why do we care?

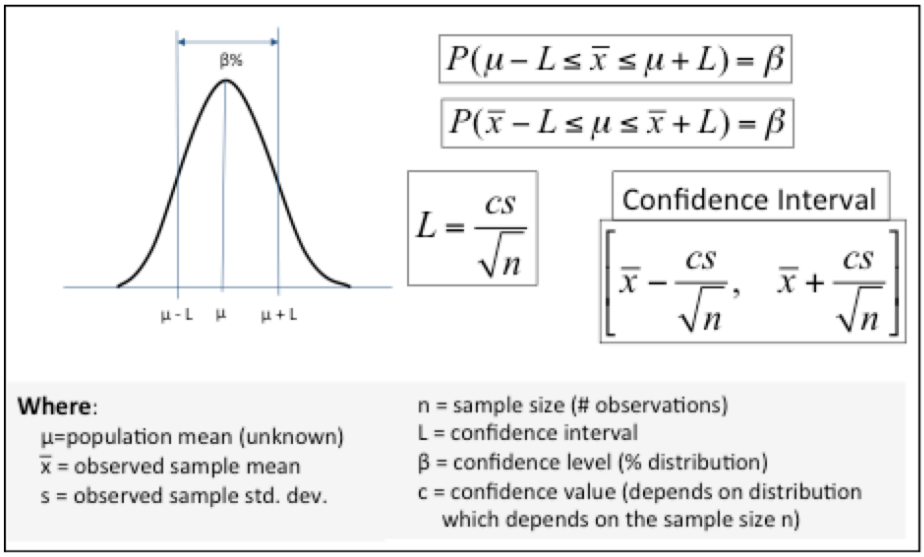
- We can show that $E[\bar{X}]=\mu$ and that $S = \sigma/\sqrt{n}$
- Note: standard deviation decreases as sample size gets bigger!
- Also, the Central Limit Theorem says that sample mean $\bar{X}$ is ~N(μ, $\sigma/\sqrt{n}$)

Confidence Intervals

Confidence intervals are used to describe the uncertainty associated with a sample estimate of a population parameter.

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.





Calculating Confidence Intervals

When the n>30

We can assume: $\bar{X} \sim N(\mu, \sigma/\sqrt{n})$

The level c of a confidence interval gives the probability that the interval produced includes the true value of the parameter.

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Where **z** is the corresponding z-score corresponding to the area around the mean:

z=1.65 for β =.90 ,
z=1.96 for β =.95,
z=2.81 for β =.995

$$\left[\bar{x} - \frac{zs}{\sqrt{n}}, \quad \bar{x} + \frac{zs}{\sqrt{n}}\right]$$

For spreadsheets use: z = NORM.S.INV((1+β)/2)

When n≤30
Then we need to use the t-distribution, which is bell-shaped and symmetric around 0.

- Mean = 0, but Std Dev = √(k/k-2)
- Where k is the degrees of freedom and, generally, k=n-1
- The value of c is a function of β and k

Where **c** is the corresponding t-statistic corresponding to the area around the mean.

$$\left[\bar{x} - \frac{cs}{\sqrt{n}}, \bar{x} + \frac{cs}{\sqrt{n}}\right]$$

For spreadsheets, use: c =T.INV.2T(1-β, k)

There are some important insights for confidence intervals around the mean. There are tradeoffs between interval (I), sample size (n) and confidence (b):

- When n is fixed, using a higher confidence level **b** leads to a wider interval, **L**.
- When confidence level is fixed (**b**), increasing sample size n, leads to smaller interval, L.
- When both n and confidence level are fixed, we can obtain a tighter interval, L, by reducing the variability (i.e. small **s** and **s**).

When interpreting confidence intervals, a few things to keep in mind:

- Repeatedly taking samples and finding confidence intervals leads to different intervals each time,
- But **b**% of the resulting intervals would contain the true mean.
- To construct a **b**% confidence interval that is within (+/-) L of μ , the required sample size is: $n=z^2*s^2/L^2$

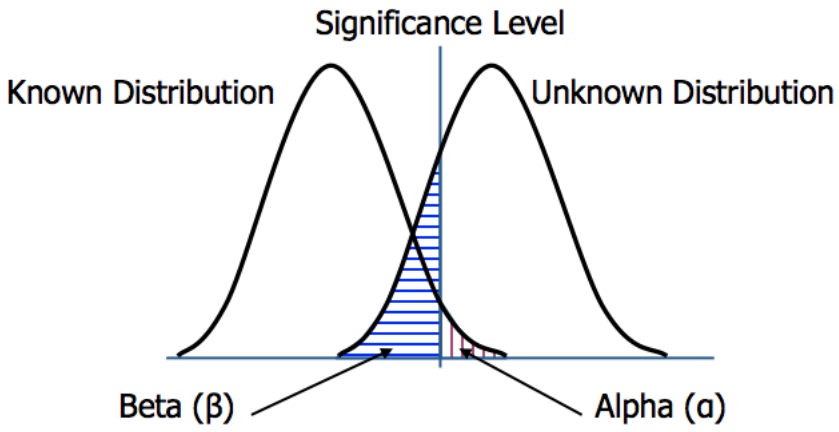
This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Hypothesis Testing

Hypothesis testing is a method for making a choice between two mutually exclusive and collectively exhaustive alternatives. In this practice, we make two hypotheses and only one can be true. Null Hypothesis (H_0) and the Alternative Hypothesis (H_1). We test, at a specified significance level, to see if we can Reject the Null hypothesis, or Accept the Null Hypothesis (or more correctly, "do not reject").

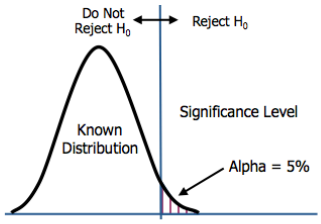
Two types of Mistakes in hypothesis testing:

- Type I: Reject the Null hypothesis when in fact it is True (Alpha)
- Type II: Accept the Null hypothesis when in fact it is False (Beta)
- We focus on Type I errors when setting significance level (.05, .01)

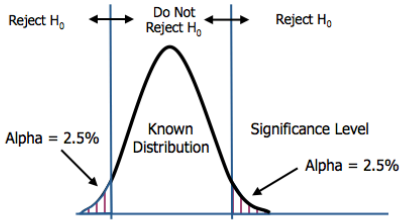


Three possible hypotheses or outcomes to a test

- Unknown distribution is the same as the known distribution (Always H_0)
- Unknown distribution is 'higher' than the known distribution
- Unknown distribution is 'lower' than the known distribution



One Tailed Predictions
(test for direction)
H_1 is (2) Unknown distribution is 'higher' than the known
H_1 is (3) Unknown distribution is 'lower' than the known



Two Tailed Predictions
(test for any difference)
H1 is either (2) or (3)
Alpha value (significance) is divided on each side of the distribution

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Example of Hypothesis Testing

I am testing whether a new information system has decreased my order cycle time. We know that historically, the average cycle time is 72.5 hours +/- 4.2 hours. We sampled 60 orders after the implementation and found the average to be 71.4 hours. We select a level of significance to be 5%.

1. Select the test statistic of interest
   mean cycle time in hours – use Normal distribution (z-statistic)
2. Determine whether this is a one or two tailed test
   One tailed test
3. Pick your significance level and critical value
   alpha = 5 percent, therefore z= NORM.S.INV(.05) = -1.6448
4. Formulate your Null & Alternative hypotheses
   H_0: New cycle time is not shorter than the old cycle time
   H_1: New cycle time is shorter than the old cycle time
5. Calculate the test statistic
   $z = (\bar{X} - \mu_{\bar{X}})/\sigma_{\bar{X}} = (\bar{X} - \mu)/(\sigma/\sqrt{n}) = (71.4 - 72.5)/(4.2/\sqrt{60}) = -2.0287$
6. Compare the test statistic to the critical value
   $z = -2.0287 < -1.6448$ the test statistic < critical value, therefore, we reject the null hypothesis

Rather than just reporting that H_0 was rejected at a 5% significance level, we might want to let people know how strongly we rejected it. The p-value is the smallest level of alpha (level of significance) such that we would reject the Null hypothesis with our current set of data. Always report the p-value p-value = NORM.S.DIST(-2.0287) = .0212

Chi square test

Chi Square test can be used to measure the goodness of fit and determine whether the data is distributed normally. To use a chi square test, you typically will create a bucket of categories, c, count the expected and observed (actual) values in each category, and calculate the chi-square statistics and find the p-value. If the p=value is less than the level of significant, you will then reject the null hypothesis.

$$\chi^2 = \sum \left(\frac{(Observed - Expected)^2}{Expected} \right) \qquad df = c - 1$$

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Spreadsheet Functions:

Function	Returns p-value for Chi-Square Test
Microsoft Excel	=CHISQ.TEST(observed_values, expected_values)
Google Sheets	=CHITEST(observed_values, expected_values)
LibreOffice->Calc	=CHISQ.TEST(observed_values; expected_values)

Ordinary Least Squares Linear Regression

Regression is a statistical method that allows users to summarize and study relationships between a dependent (Y) variable and one or more independent (X) variables. The dependent variable Y is a function of the independent variables X. It is important to keep in mind that variables have different scales (nominal/ordinal/ratio). For linear regression, the dependent variable is always a ratio. The independent variables can be combinations of the different number types.

Linear Regression Model

The data (x_i, y_i) are the observed pairs from which we try to estimate the B coefficients to find the 'best fit'. The error term, ε, is the 'unaccounted' or 'unexplained' portion.

Linear Model:

$$y_i = \beta_0 + \beta_1 x_i$$
$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad for\ i = 1,2,...n$$

Residuals

Because a linear regression model is not always appropriate for the data, you should assess the appropriateness of the model by defining residuals. The difference between the observed value of the dependent variable and predicted value is called the residual.

$$\hat{y}_i = b_0 + b_1 x_i \quad for\ i = 1,2,...n$$
$$e_i = y_i - \hat{y}_i = y_i - b_0 + b_1 x_i \quad for\ i = 1,2,...n$$

Ordinary Least Squares (OLS) Regression

Ordinary least squares is a method for estimating the unknown parameters in a linear regression model. It finds the optimal value of the coefficients (b_0 and b_1) that minimize the sum of the squares of the errors:

$$\sum_{i=1}^{n}\left(e_i^2\right) = \sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 = \sum_{i=1}^{n}\left(y_i - b_0 - b_1 x_i\right)^2$$

$$= \bar{y} - b_1\bar{x} \qquad b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Multiple Variables

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

These relationships translate also to multiple variables.

$$Y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} + \varepsilon_i \quad for\ i = 1,2,\ldots n$$
$$E(Y \mid x_1, x_2, \ldots, x_k) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$
$$StdDev(Y \mid x_1, x_2, \ldots, x_k) = \sigma$$

$$\sum_{i=1}^{n}\left(e_i^2\right) = \sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 = \sum_{i=1}^{n}\left(y_i - b_0 - b_1 x_{1i} - \ldots - b_k x_{ki}\right)^2$$

Validating a Model

All statistical software packages will provide statistics for evaluation (names and format will vary by package). But the model output typically includes: model statistics (regression statistics or summary of fit), analysis of variance (ANOVA), and parameter statistics (coefficient statistics).

Overall Fit

Overall fit = how much variation in the dependent variable (y), can we explain?

Total variation of CPL – find the dispersion around the mean.

Total Sum of Squares

$$TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Make estimate for of Y for each x

.

Error or Residual Sum of Squares

$$e_i = y_i - \hat{y}_i$$

$$RSS = \sum e_i^2 = \sum \left(y_i - \hat{y}_i\right)^2$$

Model explains % of total variation of the dependent variables.

Coefficient of Determination or Goodness of Fit (R^2)

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\left(y_i - \hat{y}_i\right)^2}{\left(y_i - \bar{y}\right)^2}$$

Adjusted R^2 corrects for additional variables

$$adjR^2 = 1 - \left(\frac{RSS}{TSS}\right)\left(\frac{n-1}{n-k-1}\right)$$

Individual Coefficients

Each Independent variable (and b_0) will have:

- An estimate of coefficient (b_1),
- A standard error (s_{bi})

$$b_1 \pm t_{\alpha/2} s_{b_1} \quad \nu = n - 2$$

  - s_e = Standard error of the model

$$s_e = \sqrt{\frac{\sum \left(y_i - \hat{y}_i\right)^2}{N-2}}$$

  - s_x = Standard deviation of the independent variables = number of observations

$$s_{b_1} = \frac{s_\varepsilon}{\sqrt{(n-1)s_x^2}}$$

- The t-statistic
  - k = number of independent variables
  - b_i = estimate or coefficient of independent variable

$$t = \frac{b_1 - \beta_1}{s_{b_1}}$$

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Corresponding p-value – Testing the Slope

- We want to see if there is a linear relationship, i.e. we want to see if the slope (b_1) is something other than zero. So: H_0: $b_1 = 0$ and H_1 $b_1 \neq 0$

- Confidence intervals – estimate an interval for the slope parameter.

$$b_1 \pm t_{\alpha/2} s_{b_1} \quad \nu = n - 2$$

Multi-Collinearity, Autocorrelation and Heterscedasticity

Multi-Collinearity is when two or more variables in a multiple regression model are highly correlated. The model might have a high R^2 but the explanatory variables might fail the t-test. It can also result in strange results for correlated variables.

Autocorrelation is a characteristics of data in which the correlation between the values of the same variables is based on related objects. It is typically a time series issue.

Heterscedasticity is when the variability of a variable is unequal across the range of values of a second variable that predicts it. Some tell tale signs include: observations are supposed to have the same variance. Examine scatter plots and look for "fan-shaped" distributions.

Learning Objectives

- Understand how to work with multiple variables.
- Be aware of data limitations with size and representation of population.
- Identify how to test a hypothesis.
- Review and apply the steps in the practice of regression.
- Be able to analyze regression and recognize issues.

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Simulation

Summary

This review provides an overview of simulation. After a review of deterministic, prescriptive modeling (optimization, math programming) and predictive models (regression), simulation offers an approach for descriptive modeling. Simulations let you experiment with different decisions and see their outcomes. Supply chain managers typically use simulations to assess the likelihood of outcomes that may follow from different actions. This review outlines the steps in simulation from defining the variables, creating the model, running the model, and refining the model. It offers insight into the benefit of using simulation for open ended questions but warns of its expensive and time consuming nature.

Over the duration of the course we have reviewed several types of models including optimization, regression, and simulation. Optimization (LP, IP, MILP, NLP) is a prescriptive form of model that finds the "best solution" and/or provides a recommendation. Regression is a predictive form of model that measures the impact of independent variables on dependent variables. We now cover simulation, which captures the outcomes of different policies with an uncertain or stochastic environment.

Simulation

Simulation can be used in a variety of contexts; it is most useful in capturing complex system interactions, modeling system uncertainty, or generating data to analyze, describe and visualize interactions, outcomes, and sensitivities. There are several classes of simulation models including: system dynamics; Monte Carlo Simulation; discrete time simulation; and agent based simulations.

There are five main steps in developing a simulation study. Formulate and plan the study; collect data and define a model; construct model and validate; make experimental runs; and analyze output. The following will review how each of those steps can be conducted.

Steps in a Simulation Study

Formulate & plan the study

Once it has been determined that simulation is the appropriate tool for the problem under investigation, the next step is to formulate the plan and study. This involves a few main steps:

- Define the goals of the study and determine what needs to be solved
- Develop a model where daily demand varies, a "production policy" will be applied
- Based on demand and policy – calculate profitability
- Assess profitability and performance metrics of different policies

Collect data & define a model

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Once the plan has been formulated, the data needs to be collected and a model defined.
If you are faced with a lack of sample data – you will need to determine the "range" of the variable(s) by talking to stakeholders or experts to identify possible values. Some data can be derived from known distributions such as Poisson or Uniform/Triangular Distributions when little to no information is available.

If sample data is available, conduct steps as we have previously reviewed such as histograms, calculating summary statistics. Then conduct a Chi-Square test to fit the sample to "traditional" distributions. Or use a "custom" empirical distribution such as discrete empirical (use % of observation as probabilities), or continuous – use histogram to compute probabilities of each range and then "uniform" within the range.

Chi-Square Test

$$\chi^2 = \sum \left(\frac{(Observed - Expected)^2}{Expected}\right) \quad df = c - 1$$

In Excel:

=**CHITEST(Observed_Range,Actual_Range)** – returns p-value
=1-**CHIDIST(Chi-square, Degrees of freedom)** – returns the p-value

Next steps are to:

- Determine relationship between various variables
- Determine performance metrics
- Collect data & estimate probability

Construct model & validate

Make necessary inputs random, add a data table to automate runs of model, add summary statistics based on results from data table.

Generating random variables with the underlying principle of generating a random (or pseudo-random) number and transform it to fit the desired distribution:

- Manual Techniques: rolling die, turning a roulette wheel, random number tables
- Excel
  - RAND () = continuous variable between 0 and 1
    - Generate random number **u**
    - For each random **u**, calculate a value **y** whose cumulative distribution function is equal to **u;** assign value **y** as the generated number:

      F(y) = P(y)=u
- Uniform Distribution ~U(a,b)
  - ~U(a, b) = a + (b – a) * RAND()
- Normal Distribution ~N(μ, σ)
  - ~N(μ, σ) =NORMINV(RAND(), μ, σ)

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Validation of Model – is the process of determining the degree to which a model and its associated data are an accurate representation of the real world for the intended use of the model. Different ways of validating model including comparing to historical data or getting expert input. One primary method is parameter variability and sensitivity analysis:

- Generate statistical parameters with confidence intervals
- Hypothesis Testing (see Week 6)

Make experimental runs

You will need to make multiple runs for each policy; use hypothesis testing to evaluate the results. If spreadsheets contain a random input, we can use our data table to repeatedly analyze the model. An additional column for runs can be made.

Analyze output

Analyzing output deals with drawing inferences about the system model based on the simulation output. Need to ensure that the model output has maintained a proper balance between establishing model validation and statistical significance. Depending on the system examined, simulation will potentially be able to provide valuable insight with the output. Ability to draw inferences from results to make system improvements will be discussed further in future courses.

Learning Objectives

- Review the steps to developing a simulation model
- Understand when to use a simulation, and when to not
- Recognize different kinds of simulations and when to apply them

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

References

APICS Supply Chain Council - http://www.apics.org/sites/apics-supply-chain-council

Chopra, Sunil, and Peter Meindl. "Chapter 1." *Supply Chain Management: Strategy, Planning, and Operation*. 5th edition, Pearson Prentice Hall, 2013.

Council of Supply Chain Management Professionals (CSCMP) https://cscmp.org/

Hillier and Lieberman (2012) Introduction to Operations Research, McGraw Hill.

Law & Kelton (2000). Simulation Modeling & Analysis, McGraw Hill.

Taha, H.A. (2010). Operations Research. An introduction. 9th edition. Pearson Prentice Hall.

Winston (2003) Operations Research: Applications and Algorithms, Cengage Learning. There are many different books by Wayne Winston - they are all pretty good.

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.