Decision Tree & Empirical Distribution
(Bertsimas & Freund 2004, Chapter 1)

Howard Hao-Chun Chuang (莊皓鈞)

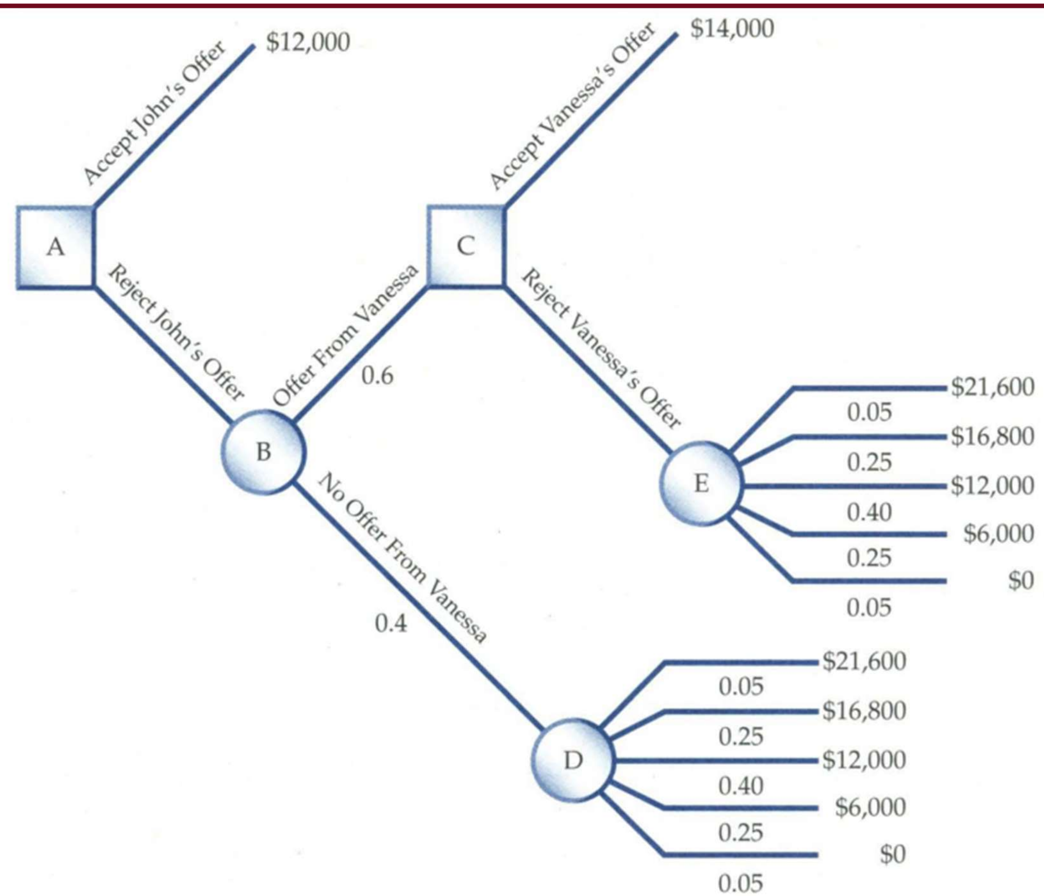
Professor
College of Commerce
National Chengchi University

September, 2024
Taipei, Taiwan

Key Characteristics

Decision tree: A roadmap that represents different choices for decision-making **(see page 8)**

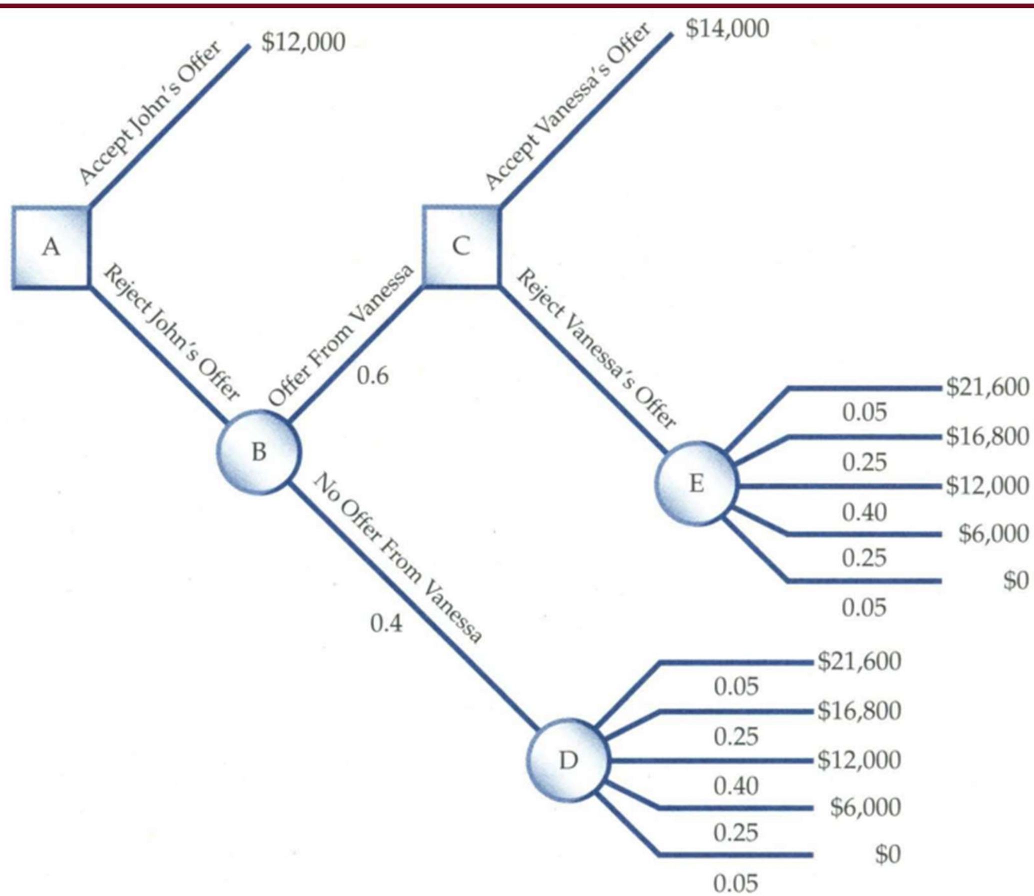
- **From left to right, each square node (A & C) is a decision point, where one has to make a choice.**
- **Decisions are mutually exclusive and collectively exhaustive?! What do you have to say about this?**
- **What are the circle nodes?**



Key Characteristics

Decision tree: A roadmap that represents different choices for decision-making (see page 8)

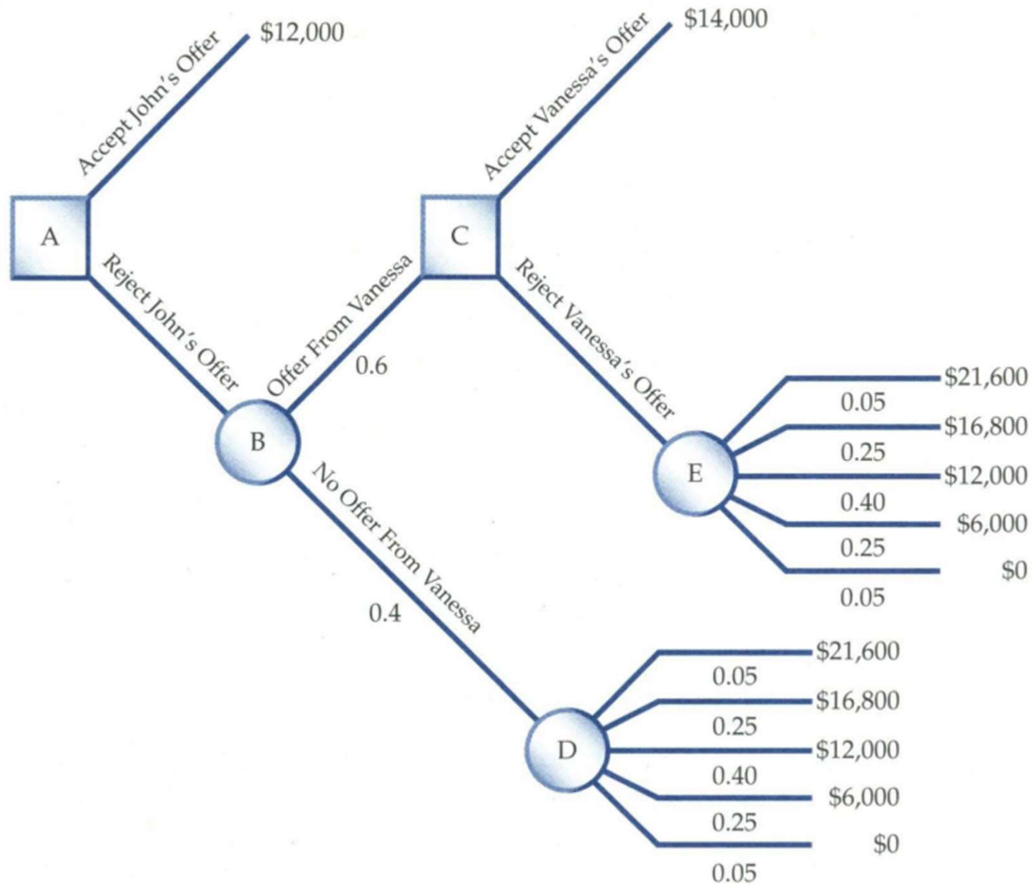
- **Each circle node (B, D, & E) is the event node that leads to uncertain outcomes.**
- **Sum of the probabilities of all possible outcomes should be 1.**
- **The final branches should have numerical values ($).**



Expected Monetary Value

EMV: An weighted-average of all possible numerical outcomes/payoffs($)

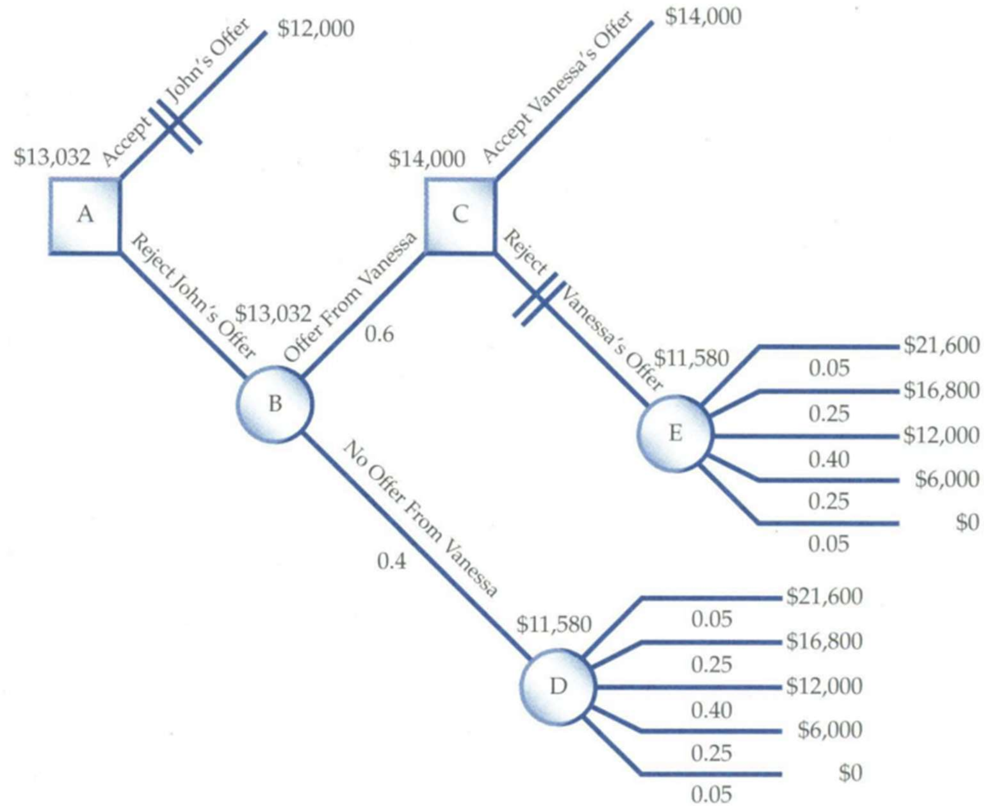
- **EMV of Node E = 21600*0.05 + 16800*0.25 + 12000*0.4 + 6000*0.25 + 0*0.05 = 11580.**
- **EMV of node D: 11580.**
- **What are the key parameters?**



Optimal Decision Strategy

Solution: Backward induction & choose the local optimal.

1. **Reject John's offer.**
2. **If Vanessa offers him a summer job, accept it.**
3. **If Vanessa does not extend an offer, go for the school's recruitment.**



- **The optimal strategy has an EMV of ?**
- **Read page 11 for solution procedures**

v = 0 ~ 8

v = 0

$$\text{vanessaioffer} = [\overbrace{1, 0, \ldots\ldots 1, 0, 1}^{1000}]$$

$$\text{VYes id} = [0, \ldots\ldots 999, 999]$$

No id = . . .

Sensitivity Analysis

How will the optimal solution change when parameters change?

- **The probability that Vanessa would offer Bill a summer job.**
- **The cost of Bill's time and effort in participating in the school's corporate summer recruiting.**
- **The distribution of summer salaries that Bill could expect to receive.**

	Spreadsheet Representation of Bill Sampras' Decision Problem		
Data			
Value of John's offer	$12,000		
Value of Vanessa's offer	$14,000		
Probability of offer from Vanessa's firm	0.60		
Cost of participating in Recruiting	$0		
	Distribution of Salaries from Recruiting		
	Weekly Salary	**Total Summer Pay**	**Percentage of Students**
		(based on 12 weeks)	**who Received this Salary**
	$1,800	$21,600	5%
	$1,400	$16,800	25%
	$1,000	$12,000	40%
	$500	$6,000	25%
	$0	$0	5%
	EMV of Nodes		
	Nodes	**EMV**	
	A	$13,032	
	B	$13,032	
	C	$14,000	
	D	$11,580	
	E	$11,580	

Departure from EMV

Decision Analysis Steps

1. List all the decisions, uncertain events, and outcomes.
2. Construct a decision tree & estimate probabilities of uncertain events.
3. Compute the EMV.
4. Solve the decision tree through backward induction.
5. Perform sensitivity analysis.

EMV as a decision criterion, what is missing?

- Most of the people are risk averse.
- 100% to get ? millions versus 50% to get 10 millions & 50% get nothing.
  - ? denotes CE (certainty equivalent)
  - Risk Premium = EMV - CE
- Great for repeated decisions, NOT so good for one-shot decision

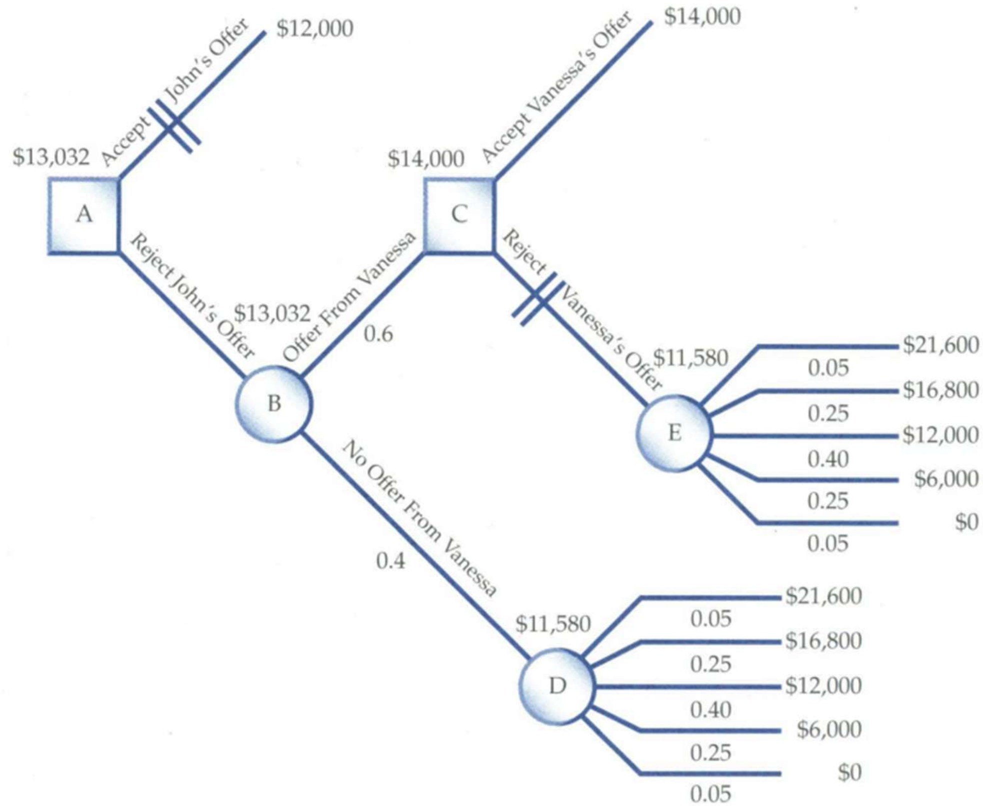
The Need for Monte-Carlo Simulation

Based on EMV, we should reject John's offer. But…

- **This is a sure thing, and he must respond to John now. Vanessa will NOT inform him until next week.**

- **Must learn the consequences of saying NO to John. Should he**
  1. **wait for Vanessa?**
  2. **bet on the school's offer**

- **What are possible consequences?** **What can be learnt from simulation?**



Empirical Distribution

Distribution of Salaries from Recruiting		
Weekly Salary	Total Summer Pay (based on 12 weeks)	Percentage of Students who Received this Salary
$1,800	$21,600	5%
$1,400	$16,800	25%
$1,000	$12,000	40%
$500	$6,000	25%
$0	$0	5%

How to form such a distribution?

- **experience/belief from (often small) information**
- **objective (maybe big) data**

Example: MLB Baseball

- **hot zone of a hitter**
- **strike zone of a pitcher**

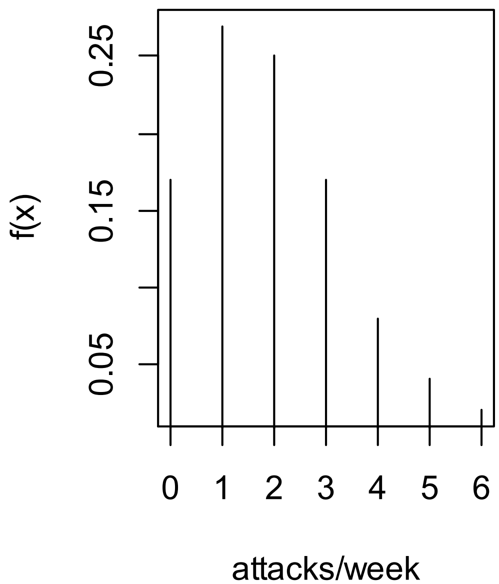
Distance of two empirical distributions

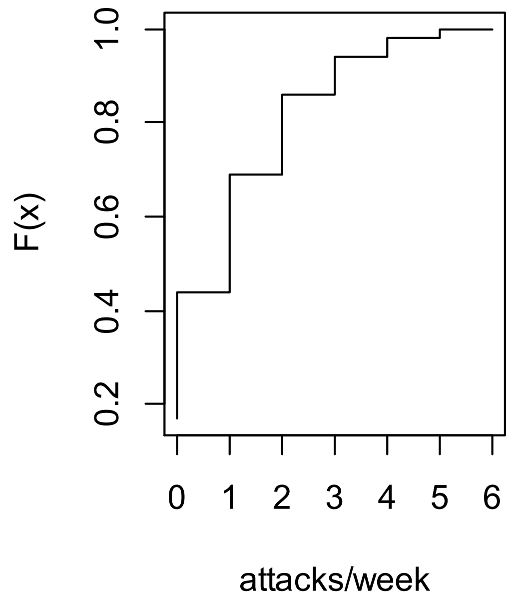
- **Wassertein distance (earth mover)**
- **JS divergence**
- **KL divergence**

Empirical Distribution

What is the limit of such a distribution?

- **Example: To obtain a model for campus Internet security in NCCU, the number of cyber-attacks occurring each week was observed over a period of 1 year.**
  It was found that,
  0 attacks occurred in each of 9 weeks
  1 attack occurred in each of 14 weeks
  2 attacks occurred in each of 13 weeks
  3 attacks occurred in each of 9 weeks
  4 attacks s occurred in each of 4 weeks
  5 attacks occurred in each of 2 weeks
  6 attacks occurred in each of 1 weeks





Empirical Distribution: Transportation Problem

Weltman & Tokar (2019) paper

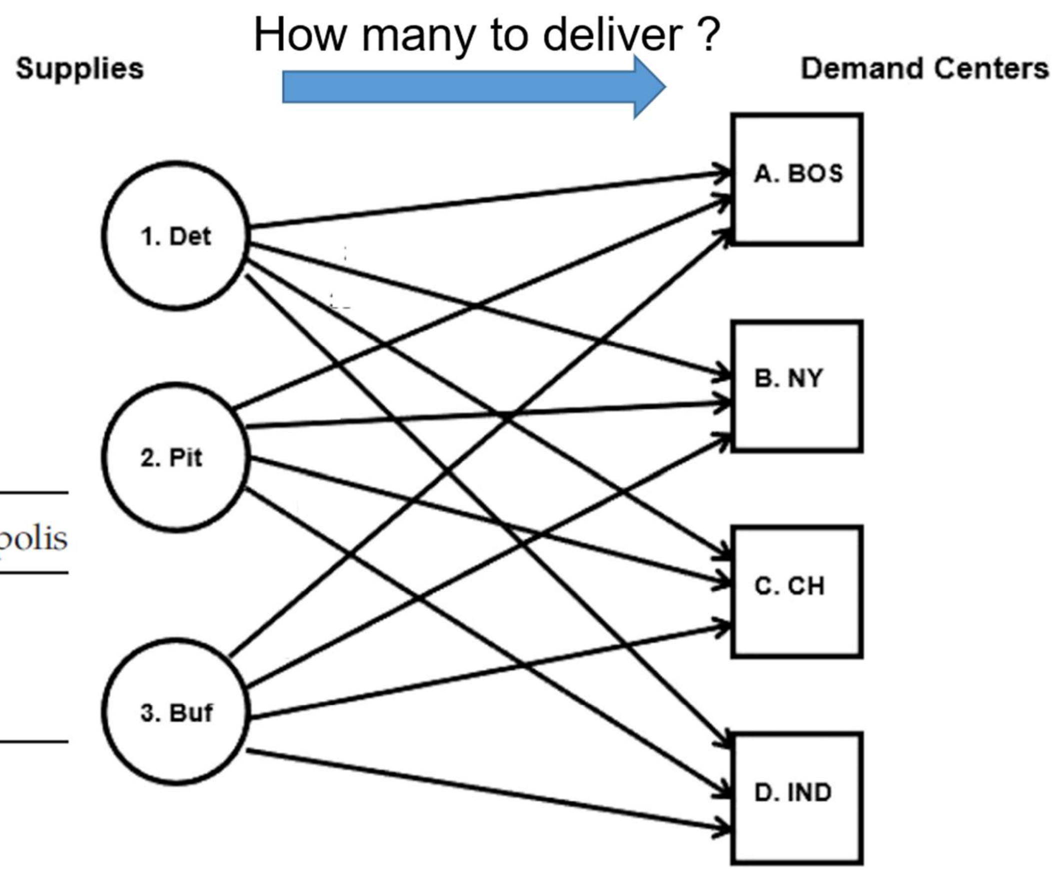
Three types of cost:

- **transportation cost**
- **stockout cost: $60 per ton**
- **inventory holding cost: $20 per ton**

Table 1. Transportation Costs ($/ton)

From\to	A. Boston	B. New York	C. Chicago	D. Indianapolis
1. Detroit	8	2	35	11
2. Pittsburgh	12	3	31	5
3. Buffalo	10	1	34	9

Objective: Minimize total cost



Empirical Distribution Case: Transportation Problem

Decision variables:

From\to	A. Boston	B. New York	C. Chicago	D. Indianapolis
1. Detroit	x_{1A}	x_{1B}	x_{1C}	x_{1D}
2. Pittsburgh	x_{2A}	x_{2B}	x_{2C}	x_{2D}
3. Buffalo	x_{3A}	x_{3B}	x_{3C}	x_{3D}

$$\underline{8x_{1A} + 12x_{2A} + 10x_{3A} + \cdots + 11x_{1D} + 5x_{2D} + 9x_{3D}}$$

→ 運輸成本

$$\Big(+60 * \max(demand_{Bos} - (x_{1A} + x_{2A} + x_{3A}), 0) + \cdots + 60 * \max(demand_{Ind} - (x_{1D} + x_{2D} + x_{3D}), 0)$$

$$+20 * \max\big((x_{1A} + x_{2A} + x_{3A}) - demand_{Bos}, 0\big) + \cdots + 20 * \max\big((x_{1D} + x_{2D} + x_{3D}) - demand_{Ind}, 0\big)\Big)$$

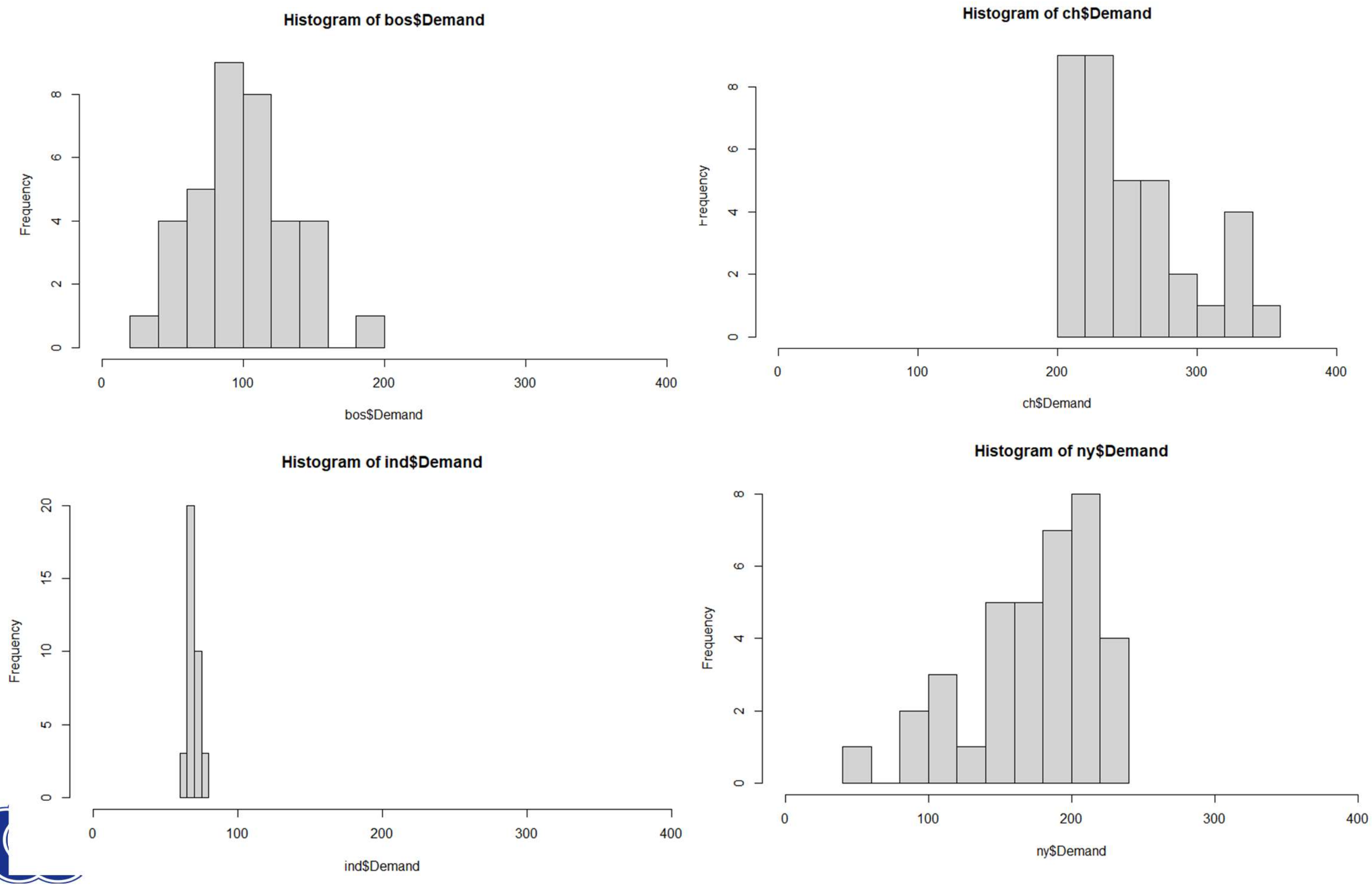
用過去數據平均

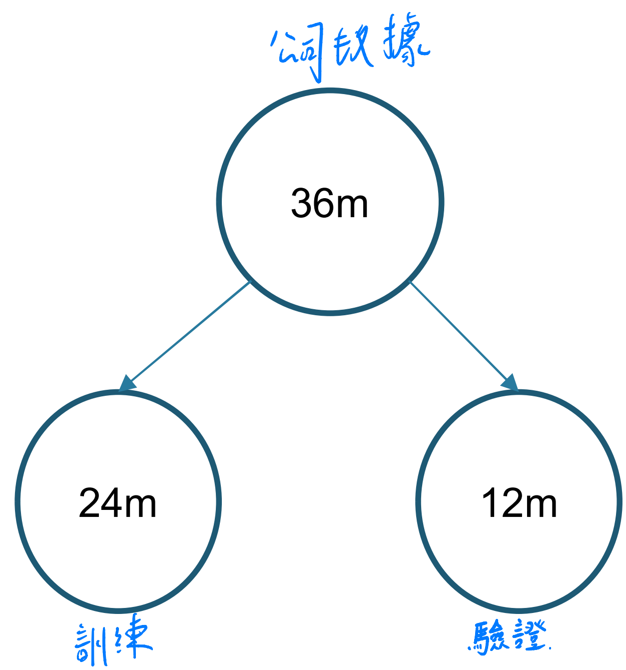
持有過多時的成本

$$x_{1A} + x_{1B} + x_{1C} + x_{1D} \leq 300$$
$$x_{2A} + x_{2B} + x_{2C} + x_{2D} \leq 180$$
$$x_{3A} + x_{3B} + x_{3C} + x_{3D} \leq 250$$

倉庫最多出的貨

Empirical Distribution Case: Transportation Problem





Empirical Distribution Case: Transportation Problem

- **Solution Strategy I: Fixed mean**
  assume future *D* is Mean of past 24-month demand
  find decisions that minimize Cost(E[*D*])

	BOS	NY	CH	IND
Det	96	47	25	0
Pitt	0	0	110	70
Buff	0	128	122	0

- **Solution Strategy II: Monthly forecast**
  use models/algorithms to predict *D* every month (e.g., Prophet) facebook?
  find decisions that minimize Cost($D^{forecast}$)

- **Solution Strategy III: Empirical simulation**
  use empirical distributions to simulate S scenarios
  find decisions that minimize E[Cost(*D*)]
  用过去經驗做隨機抽樣

	BOS	NY	CH	IND
Det	112	0	62	0
Pitt	0	0	109	71
Buff	0	197	53	0

- **Backtesting: Implement the strategies to months 25-36 & assess performance** plan

Model x_1 ⋮ x_2

1) ⇒ $\bar{D}$ (Sample Avg)

3) D_1 Cost$_1$, D_2 Cost$_2$, D_3 Cost$_3$, ⋮ ⋮, D_s Cost$_s$ ⇒ E(Cost)