

MEM6810 工程系统建模与仿真

案例 软件

第三讲：典型系统建模与仿真 I

沈海辉

中美物流研究院
上海交通大学

🏠 shenhaihui.github.io/teaching/mem6810p
✉ shenhaihui@sjtu.edu.cn

2023年春 (MEM非全日制)



上海交通大学
SHANGHAI JIAO TONG UNIVERSITY

董浩云智能制造与服务管理研究院
CY TUNG Institute of Intelligent Manufacturing and Service Management
(中美物流研究院)
(Sino-US Global Logistics Institute)



1 排队系统建模与仿真

- ▶ 引言
- ▶ 术语
- ▶ 符号
- ▶ 守恒方程
- ▶ 泊松到达过程
- ▶ 等待时间“悖论”
- ▶ $M/M/1$
- ▶ $M/M/s$
- ▶ $M/M/\infty$
- ▶ $M/M/1/K$
- ▶ $M/G/1$
- ▶ Excel 仿真实践

2 库存系统建模与仿真

- ▶ 引言
- ▶ 基本概念
- ▶ 恒定库存量模型
- ▶ 经济订货批量 (EOQ) 模型
- ▶ 单周期随机库存模型 (报童模型)



1 排队系统建模与仿真

- ▶ 引言
- ▶ 术语
- ▶ 符号
- ▶ 守恒方程
- ▶ 泊松到达过程
- ▶ 等待时间“悖论”
- ▶ $M/M/1$
- ▶ $M/M/s$
- ▶ $M/M/\infty$
- ▶ $M/M/1/K$
- ▶ $M/G/1$
- ▶ Excel 仿真实践

2 库存系统建模与仿真

- ▶ 引言
- ▶ 基本概念
- ▶ 恒定库存量模型
- ▶ 经济订货批量 (EOQ) 模型
- ▶ 单周期随机库存模型 (报童模型)



- 排队在现实世界中无处不在!
- 排队是现代生活中不可避免的一个现象。
 - 如, 在医院看病、在商店买东西、在银行取钱、通过线上客服中心进行咨询等.
 - 尽管人们都不喜欢排队, 但是大家都能认可排队机制的公平性.
- 实际上, 不仅仅只有人需要排队。
 - 例如, 邮件系统、打印机、生产线等内部也存在排队现象.
 - 生产系统在制造的过程中为原材料、半成品、成品维持队列, 即为库存.
 - 物流管理中的排队现象也很多, 如, 运输工具在仓储中心的卸货和装车过程, 电商按订单分拣出库的过程等.
- 物流管理中的排队现象也很多, 如
 - 运输工具在仓储中心的卸货和装车过程;
 - 电商按订单分拣出库的过程;
 - 顾客去快递自提点取货的过程, 等等.





图：医院中的队列





图: 商店中的队列 (from [The Sun](#))



图：银行中的队列





图：银行中的队列（有时人无需真的站到队列中）



图: 在线服务中的队列

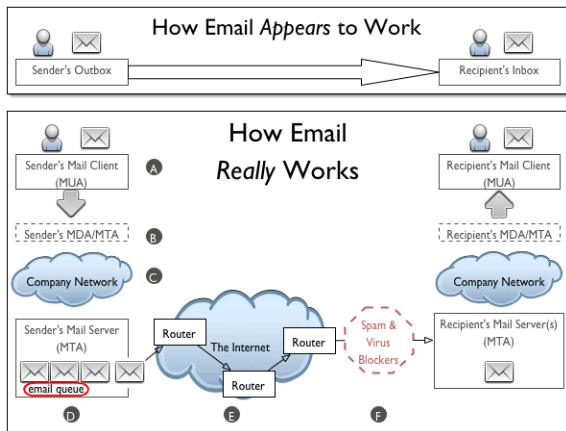


图: 邮件服务器中的队列 (from [OASIS](#))



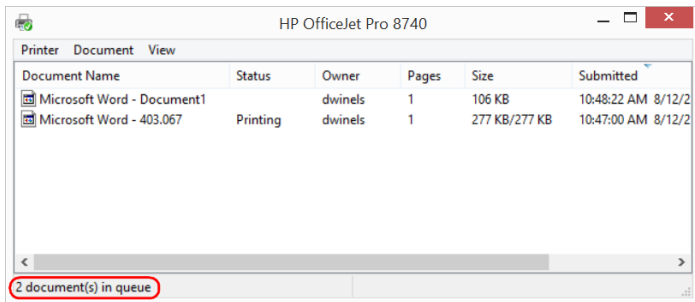


图: 打印机中的队列

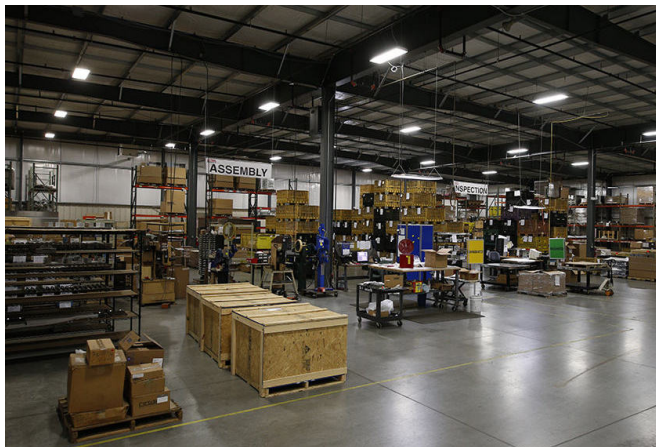


图: 生产线上的队列 (库存) (from [Estes](#))

- 一般来说, 一个排队系统包含一连串的“顾客”(可以是人、货物或信息), 他们
 - 以某种方式到达;
 - 根据特定的规则在队列中等待;
 - 接受服务;
 - 最终离开.
- 许多现实世界中的系统都可被视为排队系统, 例如,
 - 生活服务设施
 - 生产系统
 - 维修和保养设施
 - 通信与计算系统
 - 运输和物料管理系统
- 排队模型是对排队系统的数学化表示.

- 排队模型可以
 - 解析地求解, 当它比较简单时 (做了高度的简化处理);
 - 通过仿真进行分析, 当它比较复杂时 (更加贴近实际).
- 排队系统的仿真, 是典型的离散事件系统仿真.
- 无论采用哪种方式进行研究, 排队模型都是一个强大的工具, 可用以设计和评估排队系统的性能.
- 该目的可借由回答下列 (以及其它更多) 问题实现:
 - ① 平均来说, 有多少顾客在队列 (或者系统) 中?
 - ② 平均来说, 一个典型的顾客需要在队列 (或者系统) 中停留多长时间?
 - ③ 服务台的繁忙程度如何?

- **可解析求解的简单排队模型:**
 - 以可忽略的时间和费用, 得到系统性能的粗略估计.
 - 更重要的是, 理解排队系统的动态特性和不同性能度量之间的关系.
 - 为验证仿真模型是否被正确地编程实现提供了一种手段.
- **通过仿真进行分析的复杂排队模型:**
 - 使我们可以将实际系统任意精细的细节引入到模型中.
 - 估计任意感兴趣的性能度量, 并且具有高精度.

- 排队系统中最关键的两个要素是**顾客** (customer) 和**服务台** (server).
 - “顾客”一词可以指代任何到达系统并且需要服务的人或物.
 - “服务台”一词可以指代任何为顾客提供服务所需的资源.
- 相同的服务台及其前面的**队列** (queue) 构成**站点** (station), 它可能是排队系统的全部或者部分.
- **容量**是指一个站点所能容纳的顾客数量的最大值.
 - 队列中的数量 + 正在接受服务的数量.
 - 它可能是有限或无限的.

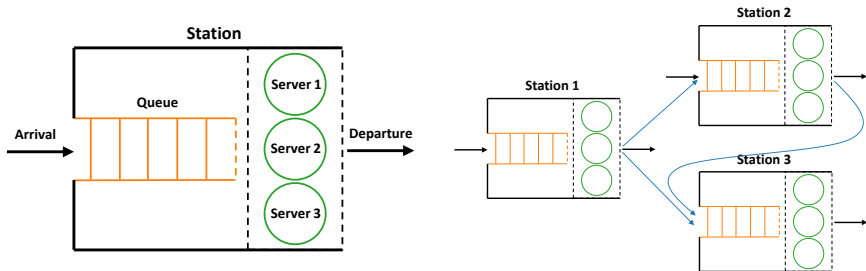


● 单站点排队系统.

- 顾客在接受完服务之后直接离开.
- 例如, 顾客来到咖啡店买完咖啡后便离开.

● 多站点排队系统 (排队网络).

- 顾客可能从一个站点去另一个站点 (接受不同的服务).
- 例如, 病人在医院中要去不同的科室排队并接受服务.



- **到达过程**模型描述了顾客是以何种方式到来。
 - 他们可能是在**预先指定的时刻**或**随机的时刻**到来。
 - 对于随机到达的情况, **到达时间间隔** (interarrival times) 通常可由概率分布来刻画。
 - 该概率分布可能依赖于当下的时间点。
 - 顾客可能一个一个到达, 或者以批次形式到达 (确定性的或随机性的批量大小)。
 - 顾客可能有多种类别。
- **当顾客达到站点之后:**
 - 如果站点容量已满:
 - 外部到达的客户只能马上离开 (lost);
 - 内部到达的客户可能在他上一个站点中等待。
 - 如果站点容量未满, 则进入站点:
 - 如果有空闲的服务台, 立刻接受服务;
 - 如果有所有服务台都处于繁忙, 则进入队列等待。



- 排队规则: 哪个顾客先服务.
 - 先进先出 (first-in-first-out, FIFO), 或称为先到先服务 (first-come-first-served, FCFS).
 - 后进先出 (last-in-first-out, LIFO), 或称为后到先服务 (last-come-first-served, LCFS).
 - 最短服务时间优先.
 - 根据优先权 (当有多个类别的顾客的时候).
- 队列行为: 在队列中的顾客的行为.
 - 畏缩 (balk): 当看到队列太长的时候选择直接离开.
 - 放弃 (renege): 在队列中等了一段时间, 感觉队伍移动得太慢而选择离开.
- **服务时长**是指一个服务台中的服务所持续的时长.
 - 确定的或随机的时长.
 - 可能依赖于顾客类别.
 - 可能依赖于当下的时间点或当前队列的长度.



- 现有的排队理论通常考虑满足下列假设的排队模型, 以便解析地求解出一些**稳态**性能指标:
 - ① 只有一类顾客.
 - ② 随机的到达 (即, 随机的到达时间间隔), 并且到达时间间隔独立同分布.
 - ③ 不是批次到达 (即, 批量大小为 1).
 - ④ 在一个站点中只有一条队列.
 - ⑤ 先到先服务.
 - ⑥ 没有畏缩 (balk) 和放弃 (renege) 现象.
 - ⑦ 随机的服务时长 (不与其他任何事物有关), 并且独立同分布.
- 即便如此, 分析求解排队模型也不是一件容易的事情.
- 而更加复杂的排队模型, 往往只能诉诸于仿真.

- 由 Kendall (1953) 提出的经典的符号体系: $X/Y/s/K$.
 - X 表示到达时间间隔的分布.
 - M : 无记忆, 即, 达时间间隔服从指数分布;
 - G : 一般的分布;
 - D : 确定性的分布.
 - Y 表示服务时长的分布.
 - 和达时间间隔的分布的记号一样.
 - s 表示并行的服务台的数量.
 - 是一个有限值.
 - 当服务台的数量无穷多时, s 被替换为 ∞ .
 - K 表示站点的容量.
 - 是一个有限值.
 - 当站点的容量为无限时, K 被替换为 ∞ , 或直接省略.
- 例子: $M/M/1$, $M/G/1$, $M/M/s/K$.



- 令 $L(t)$ 为在时刻 t 时站点内的顾客数量.

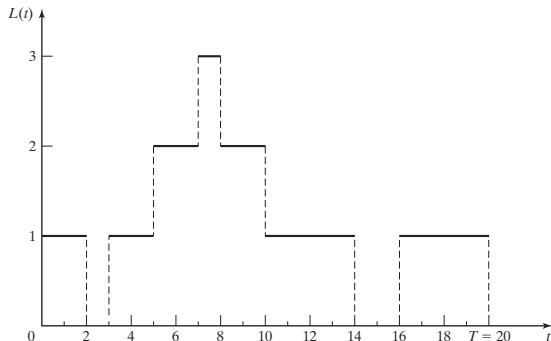


图: $L(t)$ 的图解 (from [Banks et al. \(2010\)](#))

- 令 $\hat{L}(T)$ 为截止到时刻 T 为止在站点内的 (时间加权的) 平均顾客数量:

$$\hat{L}(T) := \frac{1}{T} \int_0^T L(t) dt.$$

- $\hat{L}(T)$ 还可以用另一种方式来计算.
- 令 T_n 为在 $[0, T]$ 时间段中站点内恰好有 n 名顾客的总时长.

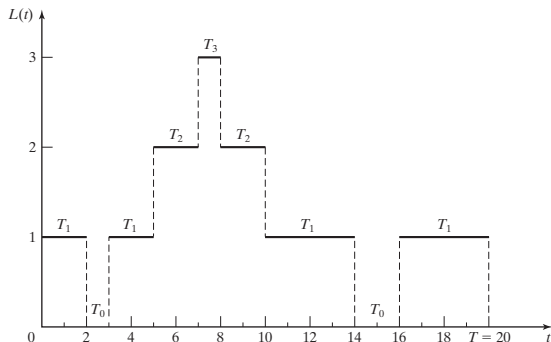


图: $L(t)$ 的图解 (from [Banks et al. \(2010\)](#))

- $\hat{L}(T) := \frac{1}{T} \int_0^T L(t) dt = \frac{1}{T} \sum_{n=0}^{\infty} n T_n = \sum_{n=0}^{\infty} n \left(\frac{T_n}{T} \right)$.

- 假设在 $[0, T]$ 时间段中, 一共有 $N(T)$ 名顾客进入站点, 并且用 $W_1, W_2, \dots, W_{N(T)}$ 表示截止到时刻 T 为止每位顾客在站点中的逗留时长.[†]
- 令 $\widehat{W}(T)$ 为截止到时刻 T 为止在站点内的平均逗留时长:

$$\widehat{W}(T) := \frac{1}{N(T)} \sum_{i=1}^{N(T)} W_i.$$

- 用类似的方式, 我们可以定义
 - $\widehat{L}_Q(T)$ - 截止到时刻 T 为止在队列中的平均顾客数量.
 - $\widehat{W}_Q(T)$ - 截止到时刻 T 为止在队列中的平均等待时长.

[†] 逗留时长包含等待时长和接受服务时长; 超出时刻 T 的部分不计入.

- 现在我们来考虑一些**长期的** (long-run) 性能度量.

- L - 在站点内的长期的平均顾客数量:

$$L := \lim_{T \rightarrow \infty} \widehat{L}(T).$$

- W - 在站点内的长期的平均逗留时长:

$$W := \lim_{T \rightarrow \infty} \widehat{W}(T).$$

- L_Q - 在队列中的长期的平均顾客数量:

$$L_Q := \lim_{T \rightarrow \infty} \widehat{L}_Q(T).$$

- W_Q - 在队列中的长期的平均等待时长:

$$W_Q := \lim_{T \rightarrow \infty} \widehat{W}_Q(T).$$

- L , W , L_Q 和 W_Q 也是对站点和队列的**期望性能度量**.

- 问题: 什么时候 L , W , L_Q 和 W_Q 存在 (且 $< \infty$), 即, 排队系统是稳定的?



- 对于一个任意的排队系统 $X/Y/s/K$:

- 令 λ 记为到达速率, 即,

$$\mathbb{E}[\text{到达时间间隔}] = \frac{1}{\lambda}.$$

- 令 μ 记为单个服务台的服务速率, 即,

$$\mathbb{E}[\text{服务时长}] = \frac{1}{\mu}.$$

定理 (稳定性条件)

对于一个 $X/Y/s/\infty$ 排队系统 (即, 无限容量), 若它的到达速率为 λ , 服务速率为 μ , 那么该系统是稳定的如果

$$\lambda < s\mu.$$

另一方面, 一个 $X/Y/s/K$ 排队系统 (即, 有限容量) 一定是稳定的.

- 守恒方程 (Little's Law) 是排队论中最一般最通用的定律之一。
 - 以 John D.C. Little 命名, 他最早在 1961 年首次证明了该定律的一个版本.
 - 当被巧妙地运用时, 守恒方程可以使一些推导变得十分简化.

定理 (守恒方程 – 经验版本)

定义 $\hat{\lambda} := N(T)/T$, 即为观测到的进入速率. 那么,

$$\hat{L}(T) = \hat{\lambda}\hat{W}(T), \quad \hat{L}_Q(T) = \hat{\lambda}\hat{W}_Q(T).$$

- 验证守恒方程*

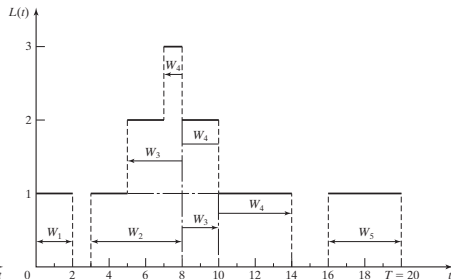
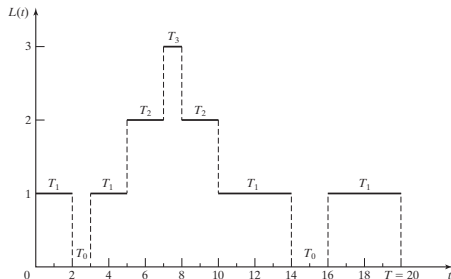


图: $L(t)$ 和 W_i 的图解 (from [Banks et al. \(2010\)](#))

$$\hat{\lambda} = N(T)/T = 5/20 = 0.25.$$

$$\widehat{W}(T) = \frac{1}{N(T)} \sum_{i=1}^{N(T)} W_i = \frac{1}{5}(2 + 5 + 5 + 7 + 4) = \frac{23}{5} = 4.6.$$

$$\widehat{L}(T) = \frac{1}{T} \sum_{n=0}^{\infty} nT_n = \frac{1}{20}(0 \times 3 + 1 \times 12 + 2 \times 4 + 3 \times 1) = \frac{23}{20} = 1.15.$$

所以, $\hat{\lambda}\widehat{W}(T) = 0.25 \times 4.6 = 1.15 = \widehat{L}(T)$. **(为什么它们总是相等?)**

- 验证守恒方程*

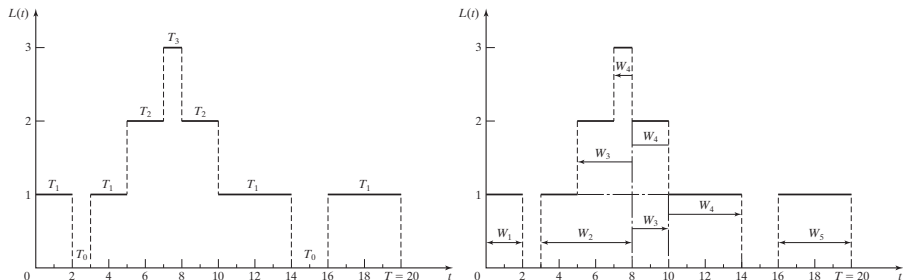


图: $L(t)$ 和 W_i 的图解 (from [Banks et al. \(2010\)](#))

- 为什么它们总是相等?

$$\widehat{L}(T) = \frac{1}{T} \sum_{n=0}^{\infty} nT_n = \frac{1}{T} \times \text{面积.}$$

$$\widehat{\lambda} \widehat{W}(T) = \frac{N(T)}{T} \frac{1}{N(T)} \sum_{i=1}^{N(T)} W_i = \frac{1}{T} \sum_{i=1}^{N(T)} W_i = \frac{1}{T} \times \text{面积.}$$

所以, $\widehat{L}(T) = \widehat{\lambda} \widehat{W}(T)$ 总是成立.

- 相同的论证可以得到 $\widehat{L}_Q(T) = \widehat{\lambda} \widehat{W}_Q(T)$.

定理 (守恒方程 – 极限/期望版本)

对于稳定的排队系统, 令 λ^* 表示到达速率或进入速率, 那么,

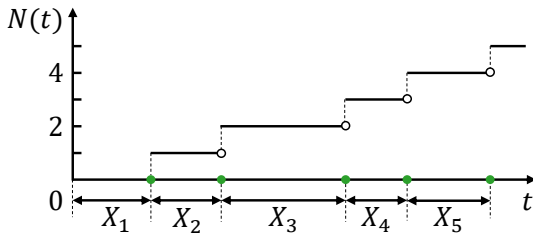
$$L = \lambda^* W, \quad L_Q = \lambda^* W_Q.$$

注意: 如果 λ^* 是到达速率, 那么时间均值 (W, W_Q) 是基于所有 (进入或未进入站点的) 顾客而言; 如果 λ^* 是进入速率, 那么时间均值只是基于所有进入站点的顾客而言.

● 附注:

- 对于未进入站点的顾客 (由于有限容量), 他在系统或者队列中花费的时间为 0.
- 一旦我们知道了 L, W, L_Q 和 W_Q 中的任意一个, 便可借助守恒方程来计算其余的量.

- 一个随机过程 $\{N(t), t \geq 0\}$ 被称为是一个计数过程, 如果 $N(t)$ 表示截止到时刻 t 为止随机到达 (或称发生的“事件”) 的总数量.



- 以 $\{X_n, n \geq 1\}$ 记到达时间间隔 (interarrival times):
 - X_1 指第一个到达的时刻 (即, 他与 0 时刻的间隔);
 - 对于 $n \geq 2$, X_n 指第 $(n-1)$ 个和第 n 个到达之间的时间间隔.

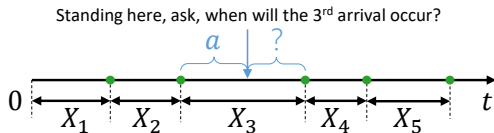
- **定义 1:** 一个计数过程 $\{N(t), t \geq 0\}$ 被称为**泊松过程**, 速率参数为 $\lambda > 0$, 如果它满足下列条件:
 - ① $N(0) = 0$;
 - ② 过程具有**独立且平稳**的增量;
 - ③ 给定 t , 随机变量 $N(t)$ 服从参数为 λt 的泊松分布, 即,
 $N(t) \sim \text{Poisson}(\lambda t)$:

$$\mathbb{P}(N(t) = n) = e^{-\lambda t} \frac{(\lambda t)^n}{n!}, \quad n = 0, 1, 2, \dots$$

- **独立的增量:** 在不相交的时间区间中到达的数量是独立的.
- **平稳的增量:** 在任一时间区间中到达的数量的分布只依赖于时间区间的长度, 即, 对 $s < t$, $N(t) - N(s)$ 的分布只依赖于 $t - s$.

- **定义 2:** 一个计数过程 $\{N(t), t \geq 0\}$ 被称为**泊松过程**, 速率参数为 $\lambda > 0$, 如果它满足下列条件:
 - ① $N(0) = 0$;
 - ② 过程具有独立且平稳的增量;
 - ③ $\mathbb{P}(N(t) = 1) = \lambda t + o(t)$;
 - ④ $\mathbb{P}(N(t) \geq 2) = o(t)$.
- **定义 3:** 一个计数过程 $\{N(t), t \geq 0\}$ 被称为**泊松过程**, 速率参数为 $\lambda > 0$, 如果它满足下列条件:
 - ① $N(0) = 0$;
 - ② $\{X_n, n \geq 1\}$ 是均值为 $1/\lambda$ 的独立同分布的指数随机变量, 即, $X_n \sim \text{exponential}(\lambda), n = 1, 2, \dots$
- 可以证明, **定义 1, 定义 2 和 定义 3 是等价的!**
- 在 Excel 中验证泊松分布和指数分布之间的关系.

- 问: 下一个什么时候到来?



$$\begin{aligned}
 \mathbb{P}(X_3 - a > x | X_3 > a) &= \frac{\mathbb{P}(X_3 - a > x, X_3 > a)}{\mathbb{P}(X_3 > a)} \\
 &= \frac{\mathbb{P}(X_3 > a + x, X_3 > a)}{\mathbb{P}(X_3 > a)} \\
 &= \frac{\mathbb{P}(X_3 > a + x)}{\mathbb{P}(X_3 > a)} \\
 &= \frac{e^{-\lambda(a+x)}}{e^{-\lambda a}} = e^{-\lambda x}. \quad (\text{与 } a \text{ 无关!})
 \end{aligned}$$

- 泊松过程具有**无记忆性**, 这是因为到达时间间隔服从指数分布, 而指数分布具有无记忆性.

- 真实数据: 以色列某电话客服中心的来电记录.
- 我们为上午 10:30 至 10:35 时段内的来电时间间隔拟合分布 (1991.11.01 – 1991.12.31, 共 43 个工作日)

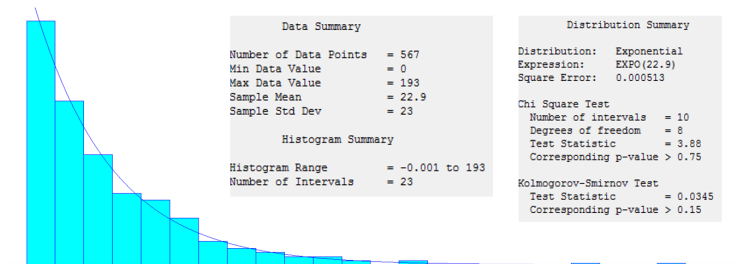


图: 数据的经验 pdf vs 指数分布 pdf (by Input Analyzer in Arena)



- 假设某车站发车间隔平均为 60 分钟, 乘客到达的时间在时间轴上呈均匀分布. 问, 乘客的平均等车时长是多少?
 - 如发车间隔为确定的 60 分钟, 易知平均等车时长为 30 分钟.
 - 如发车间隔是随机变量 X , $\mathbb{E}[X] = 60$ 分钟, 则平均等车时长 > 30 分钟.
 - 如发车间隔服从指数分布, 即, 汽车到来服从泊松过程, 则平均等车时长为 60 分钟!



- M/M/1 排队系统
 - 到达时间间隔是独立同分布的随机变量, 服从 $\text{exponential}(\lambda)$ 分布, 即, 顾客达到过程是泊松过程, 速率参数为 λ .
 - 服务时长是独立同分布的随机变量, 服从 $\text{exponential}(\mu)$ 分布.
 - 只有一个服务台, 且顾客以先到先服务的方式接受服务.
 - 容量是无限的, 即, 假设队列长度 (等候的区域) 可以是无限的.
 - 该排队系统是稳定的, 当且仅当 $\lambda < \mu$.
 - 由于无限容量, 到达速率 = 进入速率.
- 对于 M/M/1 排队系统, 我们可以相对容易地求出 L , W , L_Q 和 W_Q .

- 令 $\rho := \lambda/\mu$.
- $L = \frac{\rho}{1-\rho}$.
- $W = L/\lambda = \frac{1}{\mu-\lambda}$.
- $L_Q = \frac{\rho^2}{1-\rho}$.
- $W_Q = L_Q/\lambda = \frac{\rho}{\mu-\lambda}$.
- 或者, $W_Q = W - \mathbb{E}[\text{服务时长}] = \frac{1}{\mu-\lambda} - \frac{1}{\mu} = \frac{\rho}{\mu-\lambda}$.
- 由于无限容量, 到达速率 = 进入速率, 因此时间均值 (W , W_Q) 是基于所有顾客而言.
- 服务台利用率 = ρ , $\mathbb{P}[\text{服务台空闲}] = 1 - \rho$.
- 当 $\rho \rightarrow 1$, L , W , L_Q 和 W_Q 都趋向于 ∞ .



- $M/M/s$ 排队系统

- 到达时间间隔是独立同分布的随机变量, 服从 $\text{exponential}(\lambda)$ 分布, 即, 顾客达到过程是泊松过程, 速率参数为 λ .
- 服务时长是独立同分布的随机变量, 服从 $\text{exponential}(\mu)$ 分布.
- 有 s 个服务台.
- 顾客形成一条队列, 以先到先服务的方式, 在最先空出来的服务台接受服务 (若有多个同时空闲则等概率随机选择).
- 容量是无限的, 即, 假设队列长度 (等候的区域) 可以是无限的.
- 该排队系统是稳定的, 当且仅当 $\lambda < s\mu$.
- 由于无限容量, 到达速率 = 进入速率.

- $M/M/s$ 排队系统是 $M/M/1$ 排队系统的一般化版本; 如令 $s = 1$, 则退化成为 $M/M/1$.

- 令 $\rho := \lambda/(s\mu)$, $P_s = \left[\sum_{i=0}^s \frac{1}{i!} \left(\frac{\lambda}{\mu}\right)^i + \frac{s^s \rho^{s+1}}{s! (1-\rho)} \right]^{-1} \frac{s^s \rho^s}{s!}$.
- $L_Q = \frac{P_s \rho}{(1-\rho)^2}$.
- $W_Q = L_Q / \lambda = \frac{P_s}{s\mu(1-\rho)^2}$.
- $W = W_Q + \mathbb{E}[\text{服务时长}] = \frac{P_s}{s\mu(1-\rho)^2} + \frac{1}{\mu}$.
- $L = \lambda W = \lambda(W_Q + \frac{1}{\mu}) = L_Q + \frac{\lambda}{\mu} = \frac{P_s \rho}{(1-\rho)^2} + \frac{\lambda}{\mu}$.
- 由于无限容量, 到达速率 = 进入速率, 因此时间均值 (W , W_Q) 是基于所有顾客而言.
- 服务台利用率 = ρ ,
- $\mathbb{P}[\text{服务台全空闲}] = \left[\sum_{i=0}^s \frac{1}{i!} \left(\frac{\lambda}{\mu}\right)^i + \frac{s^s \rho^{s+1}}{s! (1-\rho)} \right]^{-1}$.
- 当 $\rho \rightarrow 1$, L , W , L_Q 和 W_Q 都趋向于 ∞ .



- 通过令 $s \rightarrow \infty$, 可以得到 $M/M/s$ 排队系统的极限情况, $M/M/\infty$ 排队系统.
- $M/M/\infty$ 排队系统总是稳定的! (服务台利用率总是 0.)
- 所有的性能度量都可以通过令 $M/M/s$ 的结果中的 $s \rightarrow \infty$ 得到.[†]
- $L = \frac{\lambda}{\mu}$.
- $W = L/\lambda = \frac{1}{\mu}$.
- $L_Q = 0, W_Q = 0$.

[†] 需要用到泰勒级数: $e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$.



• $M/M/1/K$ 排队系统

- 到达时间间隔是独立同分布的随机变量, 服从 $\text{exponential}(\lambda)$ 分布, 即, 顾客达到过程是泊松过程, 速率参数为 λ .
- 服务时长是独立同分布的随机变量, 服从 $\text{exponential}(\mu)$ 分布.
- 只有一个服务台, 且顾客以先到先服务的方式接受服务.
- 容量为 K , $K \geq 1$, 即, 队列长度 + 服务台中的顾客数量 $\leq K$.
- 如果一个从外部到达的顾客发现该站点是满的 (里面已经包含了 K 名顾客), 他会立即离开 (lost).
- 进入速率 (记为 λ_e) 小于到达速率 (记为 λ).
- 该排队系统永远是稳定的 (由于有限容量).

- 令 $\rho := \lambda/\mu$.
- $$L = \begin{cases} \frac{\rho}{1-\rho} \frac{1-(K+1)\rho^K + K\rho^{K+1}}{1-\rho^{K+1}}, & \text{如果 } \rho \neq 1, \\ \frac{K}{2}, & \text{如果 } \rho = 1. \end{cases}$$
- $$\mathbb{P}[\text{站点是满的}] = P_K = \begin{cases} \frac{(1-\rho)\rho^K}{1-\rho^{K+1}}, & \text{如果 } \rho \neq 1, \\ \frac{1}{K+1}, & \text{如果 } \rho = 1. \end{cases}$$
- 进入速率 $\lambda_e = \lambda(1 - P_K)$.
- 服务台利用率 = $\lambda_e/\mu = \rho(1 - P_K)$.
- 当 $\rho \rightarrow \infty$ 时, $L \rightarrow K$, $1 - P_K \rightarrow 0$, $\rho(1 - P_K) \rightarrow 1$.

- 对于进入站点的顾客
 - $W = L/\lambda_e = \frac{L}{\lambda(1-P_K)}$.
 - $W_Q = W - \frac{1}{\mu} = \frac{L}{\lambda(1-P_K)} - \frac{1}{\mu}$.
- 对于全体顾客 (那些未进入站点的顾客, 他们的逗留时间和等待时间都为 0)
 - $W' = (1 - P_K)W + P_K \times 0 = \frac{L}{\lambda}$.
 - $W'_Q = (1 - P_K)W_Q + P_K \times 0 = \frac{L}{\lambda} - \frac{1-P_K}{\mu}$.
- $L_Q = \lambda_e W_Q = \lambda W'_Q = L - \rho(1 - P_K)$.
- 当 $\rho \rightarrow \infty$ 时, $L_Q \rightarrow K - 1$.
 - 如果 μ 固定而 $\lambda \rightarrow \infty$:
 $\lambda(1 - P_K) \rightarrow \mu$, $W \rightarrow \frac{K}{\mu}$, $W_Q \rightarrow \frac{K-1}{\mu}$, $W' \rightarrow 0$, $W'_Q \rightarrow 0$.
 - 如果 λ 固定而 $\mu \rightarrow 0$:
 $\frac{1}{\mu}(1 - P_K) \rightarrow \frac{1}{\lambda}$, $W \rightarrow \infty$, $W_Q \rightarrow \infty$, $W' \rightarrow \frac{K}{\lambda}$, $W'_Q \rightarrow \frac{K-1}{\lambda}$.

- M/G/1 排队系统

- 服务时长是独立同分布的随机变量, 服从一个任意的分布 (均值 $\frac{1}{\mu}$, 方差 σ^2).
- 其他都和 M/M/1 排队系统相同.

- 令 $m^2 := \left(\frac{1}{\mu}\right)^2 + \sigma^2$, 以及 $\rho := \lambda/\mu < 1$.

- 服务台利用率 = ρ , $\mathbb{P}[\text{服务台空闲}] = 1 - \rho$.
- $W_Q = \frac{\lambda m^2}{2(1-\rho)}$.
- $L_Q = \lambda W_Q = \frac{\lambda^2 m^2}{2(1-\rho)}$.
- $W = W_Q + \frac{1}{\mu} = \frac{\lambda m^2}{2(1-\rho)} + \frac{1}{\mu}$.
- $L = \lambda W = L_Q + \lambda/\mu = \frac{\lambda^2 m^2}{2(1-\rho)} + \rho$.

- 对于 M/G/ ∞ 排队系统, 那些性能度量与 M/M/ ∞ 排队系统中的性能度量是相同的.



- 在 Excel 中实现 $M/M/1$ 排队系统的仿真, $\lambda = 0.6$, $\mu = 1$.
 - 根据排队论, 已知 $L = 1.5$, $L_Q = 0.9$, $W = 2.5$, $W_Q = 1.5$, 服务台利用率 $= \rho = 0.6$.
- 任意 $G/G/1$ 排队系统的仿真可用相同的方法实现 (此时已经无法解析求解).
- 在 Excel 中实现 $M/M/2$ 排队系统的仿真, $\lambda = 0.6$, $\mu = 0.5$.
 - 根据排队论, 已知 $L = 1.875$, $L_Q = 0.675$, $W = 3.125$, $W_Q = 1.125$, 服务台利用率 $= \rho = 0.6$.
- 任意 $G/G/2$ 排队系统的仿真可用相同的方法实现 (此时已经无法解析求解).



● 单泊位港口仿真实例

- 某港口现有 1 个泊位, 可供船舶停靠, 进行装船、卸货作业.
- 船舶入港后, 如泊位是空闲的, 可立即使用; 否则需要遵循先到先服务原则在港口区排队等候.
- 邮轮到达时间间隔的分布, 邮轮类型的分布及其所需作业时长如下列表格所示:

到达时间间隔/天	概率	邮轮类型	所需时长/天	概率
1	0.05	巨型	4	0.40
2	0.15	中型	3	0.35
3	0.35	小型	2	0.25
4	0.25			
5	0.20			

- 计算该港口长时间连续运行下的性能度量: L , L_Q , W , W_Q 和泊位利用率.
- 提示: 可以建模为 $G/G/1$ 进行仿真.



- 双泊位港口仿真实例

- 由于发现该港口现有的服务水平太低, 现考虑对港口进行改造, 新增 1 个泊位.
- 由于新增泊位的技术水平较高, 所需作业时长较原泊位短, 如下表所示:

邮轮类型	原泊位所需时长/天	新泊位所需时长/天
巨型	4	3
中型	3	2
小型	2	1

- 计算改造后各性能度量的变化情况.
- 提示: 可以建模为 $G/G/2$ (两个服务台速率不同) 进行仿真.

1 排队系统建模与仿真

- ▶ 引言
- ▶ 术语
- ▶ 符号
- ▶ 守恒方程
- ▶ 泊松到达过程
- ▶ 等待时间“悖论”
- ▶ $M/M/1$
- ▶ $M/M/s$
- ▶ $M/M/\infty$
- ▶ $M/M/1/K$
- ▶ $M/G/1$
- ▶ Excel 仿真实践

2 库存系统建模与仿真

- ▶ 引言
- ▶ 基本概念
- ▶ 恒定库存量模型
- ▶ 经济订货批量 (EOQ) 模型
- ▶ 单周期随机库存模型 (报童模型)



- 库存 (inventory) 是一个组织中存储的为满足日后所需的物品或资源。
 - 制造库存主要分为原材料库存、零部件库存、在制品库存、成品库存。
 - 在服务行业, 库存通常指将来售出的有形商品和提供服务所需的供给。
- 持有库存的主要目的是
 - 利用订货的规模经济效应;
 - 应对供应方或需求方的不确定性。
- 库存系统是一套政策和控制机制, 它监控库存水平, 决定需要维持何种水平, 何时需要补货, 以及订单量多大。
- 库存模型是对库存系统的数学化表示, 可以帮助企业提高库存决策科学性和准确性。

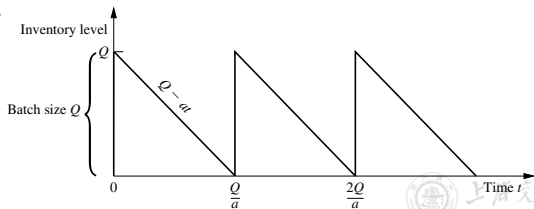
- 需求
 - 离散 vs 连续
 - 确定 vs 随机
- 补充 (向其他厂家购买或者自己生产)
 - 订货提前期 (lead time): 确定 vs 随机
 - 订货批量 (batch size)
 - 到货方式: 一次性到达 vs 连续到达
- 费用
 - 订货/生产费用
 - ① 订购/装配费用 (固定, 与数量无关)
 - ② 商品成本 (与数量成正比, 有时也有数量折扣等因素要考虑)
 - 缺货费用 (shortage cost)
 - 失去销售机会造成的损失, 停工待料造成的损失, 信誉的损失, 相关的赔偿等
 - 若不允许缺货, 则将缺货费用作无穷大处理
 - 库存费用 (holding cost, or carrying cost)

- 若一个仓库只为一个客户供货, 该客户的需求是随机的, 服从 $[1, 10]$ 上的离散均匀分布. 若要保证至少在 80% 的情况中订单需求可以满足, 最低的库存水平是多少? **答案: 8**
- 如果新增一个客户, 该客户的需求也服从 $[1, 10]$ 上的离散均匀分布, 且两个客户的需求是独立的. 若仍要保证至少在 80% 的情况中订单需求可以满足, 最低的库存水平是多少?
答案: 16? 15 就已经足够!
- 理论计算 vs 仿真分析.
- 若要保证至少在 20% 的情况中订单需求可以满足呢?
- 如果该商品的数量为连续的, 需求服从 $\text{uniform}(1, 10)$ 呢?

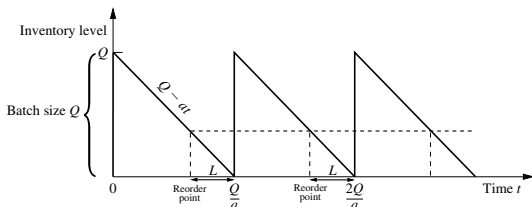
- 经济订货批量 (EOQ) 模型是库存理论中最基本的模型, 最早由 Ford W. Harris 于 1913 年提出.
- 基础 EOQ 模型 (不允许缺货, 补货一次到达)
 - 需求是连续且均匀的, 单位时间需求量为 a ;
 - 每次订货批量 Q 是恒定的, 订货提前期 $L = 0$ 并且是一次性到达 (即, 瞬时完成);
 - 固定订购费用为 C_0 ; 单位物资成本为 c ; 单位时间单位商品的库存费用为 h ; 不允许缺货 (缺货费用无穷大).

- 最优订货批量为 $Q^* = \sqrt{\frac{2aC_0}{h}}$; 订货周期也随之确定, 为

$$t^* = \frac{Q^*}{a} = \sqrt{\frac{2C_0}{ah}}$$



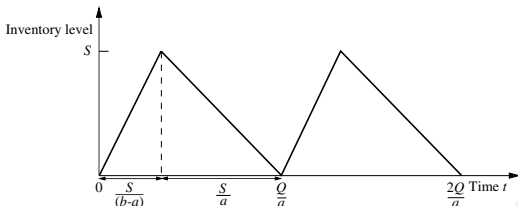
- 基础 EOQ 模型 (不允许缺货, 补货一次到达)
 - 订货提前期 L 固定但 $L > 0$, 并且是一次性到达;
 - 其余都不变.
- 最优订货批量不变, 只需将订货点 (reorder point) 往前平移时间 L 即可.



- 只要订货提前期 L 是确定的, 它便不会对分析造成影响 (只需先按 $L = 0$ 计算, 最终将订货点前移即可).

- EOQ 模型 (不允许缺货, 补货**连续到达**)
 - 需求是连续且均匀的, 单位时间需求量为 a ;
 - 每次订货批量 Q 是恒定的, 订货提前期 L 是确定的, 但**货物是连续到达的, 单位时间到达量为 b , $b > a$** .
 - 固定订购费用为 C_0 ; 单位物资成本为 c ; 单位时间单位商品的库存费用为 h ; 不允许缺货 (缺货费用无穷大).

- 最优订货批量为 $Q^* = \sqrt{\frac{2aC_0}{h}} \sqrt{\frac{b}{b-a}}$; 订货周期也随之确定, 为 $t^* = \frac{Q^*}{a} = \sqrt{\frac{2C_0}{ah}} \sqrt{\frac{b}{b-a}}$. 最优库存峰值为 $S^* = \sqrt{\frac{2aC_0}{h}} \sqrt{\frac{b-a}{b}}$.

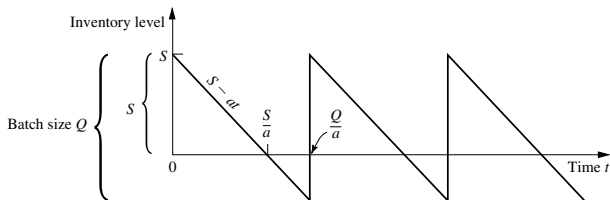


- 有些情况下, 允许一定数量的缺货可以减少库存费用, 降低总成本.
- 缺货时, 顾客的行为有两种: 流失 (去别家买) 或等待.
 - 如果顾客会因为缺货而流失, 在前两个模型其他假设不变的情况下, 可知, 即使无需因为缺货而赔偿顾客, 商家也不会允许缺货发生 (为了最大化利润), 因此最优解与前面保持一致.
 - 如果顾客愿意等待, 而商家需要为此付出一定的费用, 那么最优解会出现变化.

- EOQ 模型 (**允许缺货**, 补货一次到达)
 - 需求是连续且均匀的, 单位时间需求量为 a ;
 - 每次订货批量 Q 是恒定的, 订货提前期 L 是确定的, 并且是一次性到达;
 - 固定订购费用为 C_0 ; 单位物资成本为 c ; 单位时间单位商品的库存费用为 h ;
 - **允许缺货**: 顾客一直等待直到有货 (backorder), 但是每单位时间单位商品的缺货费用为 p ; 当有货时, 他们瞬时被满足。



- EOQ 模型 (**允许缺货**, 补货一次到达)



- 最优订货批量为 $Q^* = \sqrt{\frac{2aC_0}{h}} \sqrt{\frac{p+h}{p}}$; 订货周期也随之确定, 为 $t^* = \frac{Q^*}{a} = \sqrt{\frac{2C_0}{ah}} \sqrt{\frac{p+h}{p}}$.
- 最优库存峰值为 $S^* = \sqrt{\frac{2aC_0}{h}} \sqrt{\frac{p}{p+h}}$, 最大缺货量为 $Q^* - S^* = \sqrt{\frac{2aC_0}{p}} \sqrt{\frac{h}{p+h}}$.



- EOQ 模型 (**允许缺货, 补货连续到达**)
 - 需求是连续且均匀的, 单位时间需求量为 a ;
 - 每次订货批量 Q 是恒定的, 订货提前期 L 是确定的, 但**货物是连续到达的, 单位时间到达量为 $b, b > a$.**
 - 固定订购费用为 C_0 ; 单位物资成本为 c ; 单位时间单位商品的库存费用为 h ;
 - **允许缺货: 顾客一直等待直到有货 (backorder), 但是每单位时间单位商品的缺货费用为 p ; 当有货时, 他们瞬时被满足.**

- **数量折扣**: 供应商有时为了鼓励大批量订货, 会实行价格优惠, 订货批量越大, 单价越便宜.
- EOQ 模型 (不允许缺货, 补货一次到达, 有**数量折扣**)
 - 需求是连续且均匀的, 单位时间需求量为 a ;
 - 每次订货批量 Q 是恒定的, 订货提前期 L 是确定的, 并且是一次性到达;
 - 固定订购费用为 C_0 ; 单位时间单位商品的库存费用为 h ; 不允许缺货 (缺货费用无穷大).
 - **数量折扣**: 单位物资成本为 $c(Q)$, 与订货批量 Q 有关.

$$c(Q) = \begin{cases} c_1, & 0 \leq Q < Q_1, \\ c_2, & Q_1 \leq Q < Q_2, \\ \dots & \dots \\ c_n, & Q_{n-1} \leq Q < Q_n, \end{cases}$$

其中, $0 < Q_1 < Q_2 < \dots < Q_n$, $c_1 > c_2 > \dots > c_n$.

- 最优订货批量 Q^* 可通过下列方式计算.

- ① 令 $\tilde{Q} = \sqrt{\frac{2aC_0}{h}}$ (此为基础模型中的最优订货批量);
- ② 计算该订货批量下的平均总费用: 若 $Q_{i-1} \leq \tilde{Q} < Q_i$, 则平均总费用为

$$\tilde{C} = \frac{h\tilde{Q}}{2} + \frac{aC_0}{\tilde{Q}} + ac_i.$$

- ③ 计算其他订货批量下的平均总费用:

$$C^{(j)} = \frac{hQ_j}{2} + \frac{aC_0}{Q_j} + ac_{j+1}, \quad j = i, i+1, \dots, n-1.$$

- ④ $\{\tilde{C}, C^{(i)}, C^{(i+1)}, C^{(n-1)}\}$ 中最小者, 所对应的订货批量, 即为最优订货批量 Q^* .



- 例子: 某车间每月需要消耗零件 30000 个 (a), 固定订购费用为 500 元 (C_0), 每月每件库存费用为 0.2 元 (h). 订货单价 $c(Q)$ 与订货批量 Q 的关系如下:

$$c(Q) = \begin{cases} 1\text{元/个}, & 0 \leq Q < 10000, \\ 0.98\text{元/个}, & 10000 \leq Q < 30000, \\ 0.94\text{元/个}, & 30000 \leq Q < 50000, \\ 0.90\text{元/个}, & 50000 \leq Q < \infty. \end{cases}$$

求最优订货批量 Q^* .

- 单周期库存模型: 在某一时期内订货只有一次, 到此时期结束时要么所有的产品全卖光, 要么就折本销售剩余产品。
 - 典型例子: 报纸、时装、易腐烂产品等。
- 报童模型 Arrow et al. (1951)
 - 需求是一个随机变量 X , 其累积分布函数 (CDF) 为 $F(x)$ (注: X 可以是离散随机变量, 也可以是连续随机变量);
 - 在销售期开始之前确定订货批量 Q , 并于销售期开始之前送达;
 - 固定订购费用 C_0 可假设为 0 (若 $C_0 > 0$, 不会改变最优解, 除非 C_0 过大而使最优解变为“不订货”);
 - 商品的库存费用为一个**恒定的值** (与时间和数量无关), 可假设为 0 (若大于 0, 不会改变最优解, 除非过大而使最优解变为“不订货”);
 - 单位商品成本为 c ; 单位商品的售出价为 p , $p > c$, 商品售完即止 (后续的顾客全流失); 若销售期结束时仍有商品剩余, 则以单价 q , $q < c$, 一次性处理 (q 称为残值)。



• 简单分析

- 随机需求为 X , 订货量为 Q , 因此剩余量为 $\max\{Q - X, 0\}$, 销售量为 $Q - \max\{Q - X, 0\}$.
- 令 $a := p - c > 0$ 表示每售出单位商品的利润, 令 $b := c - q > 0$ 表示每剩余单位商品的亏损.
- 销售期结束后的总利润为

$$\begin{aligned}\pi(Q) &= a[Q - \max\{Q - X, 0\}] - b \max\{Q - X, 0\} \\ &= aQ - (a + b) \max\{Q - X, 0\} \\ &= (p - c)Q - (p - q) \max\{Q - X, 0\}.\end{aligned}$$

- 总利润的期望值为

$$\mathbb{E}[\pi(Q)] = (p - c)Q - (p - q) \mathbb{E}[\max\{Q - X, 0\}].$$

- 需要寻找 Q 使 $\mathbb{E}[\pi(Q)]$ 最大.

• 最优订货批量 Q^* 为

$$Q^* = F^{-1}\left(\frac{p - c}{p - q}\right) = \min \left\{ x \geq 0 : F(x) \geq \frac{p - c}{p - q} \right\}.$$

- 如果在做订货决策之时, 已有初始库存 I .
- 若 $I \geq Q^*$, 则肯定选择“不订货”.
- 若 $I < Q^*$, 令 q^* 满足 $q^* \leq Q^*$ 且 $\mathbb{E}[\pi(q^*)] = \mathbb{E}[\pi(Q^*)] - C_0$:
 - 如果 $I \geq q^*$, 选择“不补货”;
 - 如果 $I < q^*$, 选择补货至 Q^* .
- 可知当 $C_0 = 0$ 时, $q^* = Q^*$, 此时, 若 $I < Q^*$, 则必选择补货至 Q^* .
- 这便是经典的 (s, S) 订货策略: 当库存水平低于 s , 补货至 S ; 当库存水平高于或等于 s , 不补货.
- 在报童模型中, $S = Q^*$, $s = q^*$, 并且可以证明 (s, S) 订货策略是最优的策略.

- 实例分析

- 某报刊亭全年出售一种报纸, 每份售价 1.0 元, 每份进价 0.4 元, 当天剩余报纸残值为每份 0 元.
- 根据以往经验, 每天报纸的需求量的分布如下表所示:

需求/份	300	400	500	600	700	800
概率	0.05	0.10	0.25	0.30	0.20	0.10

- 每天的最优订货批量为多少?
 - 通过 Excel 进行理论计算 vs 仿真分析.
- 如果保亭与印刷厂达成协议, 当天剩余报纸可以每份 0.2 元卖回给印刷厂, 该协议的价值有多大?