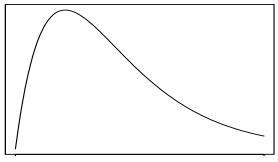
Modelling and Design of Service-Time; **Phase-Type** Distributions.

References.

- Issaev, Eva,: "Fitting phase-type distributions to data from a call center", M.Sc. thesis, Technion IE&M, 2003. (With ample references and a literature review.) http://iew3.technion.ac.il/serveng/References/Thesis.pdf
- Neuts, M.F., *Matrix Geometric Solutions in Stochastic Models*, John Hopkins University Press, 1981.
- Mandelbaum, A. and Reiman, M., "On pooling in queueing networks", Management Science, 44, 971-981, 1996. http://iew3.technion.ac.il/serveng/References/pooling.pdf
- Buzacott, J.A. and Shanthikumar, J., *Stochastic Models of Manufacturing Systems*, pages 63–64, pages 540–541; pages 64–67.

Buzacott and Shanthikumar, on pages 154–155, provide an IE-discussion and references on human task-time, stochastic variability and working rates. Their sources are likely to be manufacturing-based. These are their **key points:**

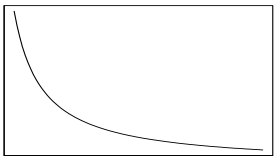
- Much of the variability is beyond control of the operator.
- There exists minimum time of task duration.
- Paced vs. *Unpaced* work. (In Services, it is typically unpaced: no upper bound is imposed on task duration.)
- Typically, for experienced operators service-time distribution is skewed with $P\{T \leq \text{ mean }\} \approx 0.65 \approx 1 - \frac{1}{e}$, which is consistent with an exponential distribution. (For many practical purposes, a distribution with CV ≈ 1 "is exponential".)

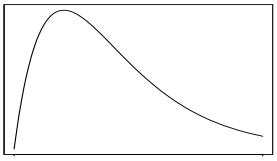


- For inexperienced operators: greater mean and less skewness. (My understanding of this: greater mean and more symmetry.)

- It is not clear from the literature whether there is serial correlation in performing successive tasks.
- Variation of working rate is rarely due to physical fatigue or exhaustion, but rather to inserted-idleness, while the distribution of task-time is stationary over the day.

Personal Experience. Two patterns of service-duration density are prevalent:





One pattern "is exponential" and the other "is lognormal".
It is interesting to understand what does the shape depend on: is it experience-related? job-related? Research is needed to answer such questions.

Examples: Service (Process) Design

- **Pooling Resources** without changing the service process. (At the City Hall of Haifa, moving the Treasury Department to a new location gave rise to phase-type service durations, and motivated M. and Reiman.)
- **IVR/VRU:** Design of search protocols (Comverse).
- **Phone Scripts:** Design of phone/chat services at call/contact centers. (Electric Company).

Contents

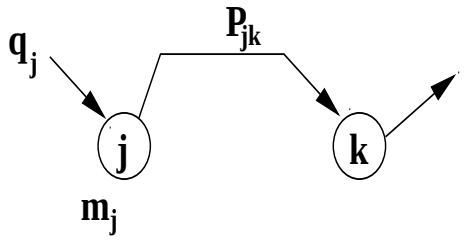
- Design of a service system (service time): Pooling.
- What is Service Time?
  - Single- vs. multiple-visits.
  - Time- and State-dependency.
  - Sample Size
  - Estimation and Prediction, Workload.
  - Production of Health: Hernia, Even-Doctors-Can-Manage.
  - After-service work: managing accessibility.
  - Averages do not tell the whole story: the need for the distribution.
  - Service Time is a Statistical Distribution: for example, Log-Normal, Exponential, Mixture.
  - Heterogeneity of Servers.
- Stochastic Ordering (of distributions).
- Exponential Service Times in Human Services:
  - How does one recognize an exponential distribution in a histogram.
  - mean = standard deviation ($CV = 1$) sometimes suffices, but *not* always. (For example, in the QED regime things are yet unclear).
  - Meta Theorem: Durations of human *homogeneous* services are either exponential or lognormal.
- Service Time is a Process: Phase-type models natural and useful.
- Beyond CV's: Some subtle effects of the service-time distribution in the QED regime.

Phase-Type Service Times (Durations).

Service-Time = a sequence/collection of tasks, of an *exponential* duration.
There are K types of tasks, indexed by $k = 1, \ldots, K$.

m_k = expected duration of task k;	$m = (m_k)$
q_k = % of services in which k is first;	$q = (q_k)$
P_{jk} = % of incidences in which task j is immediately followed by k.	$P = [P_{jk}]$

$1 - \sum_{\ell=1}^{K} P_{k\ell}$ = probability to end service at k.



Fact: service = *finite* number of tasks $\Leftrightarrow \exists [I - P]^{-1}$
Indeed, $[I - P]^{-1}_{jk}$ = expected number of "visits to k", given j was first.
$(q[I - P]^{-1})_k$ = expected number of "visits to k").

As will be articulated below, service-time duration is *Phase-type* (PH).
(Assuming independence among task-durations.)

Definition. Phase-type distribution = absorption time of a finite-space continuous-time Markov chain, with a single absorbing state.

Formally: $X = \{X_t, t \geq 0\}$ Markov on states $\{1, 2, \ldots, K, \Delta\}$, with infinitesimal generator

$$Q = \begin{array}{c} 1 \\ \vdots \\ K \\ \Delta \end{array} \begin{bmatrix} & & R & & r \\ & & & & \\ 0 & & \ldots 0 & & 0 \end{bmatrix}$$

- Δ absorbing (since $q_{\Delta\Delta} = 0$)
- $r = -R1$ (since $Q1 = 0$)
- $1, \ldots K$ transient $\Leftrightarrow \exists R^{-1}$ (fact)

and initial distribution (of X_0) is given by $(q_1, \ldots, q_k, 0) = (q, 0)$.

Recall:

$$P\{X_t = k\} = \sum_j q_j [\exp(tR)]_{jk} = q[\exp(tR)]_k$$

Define: T $= \inf\{t > 0 : X_t = \Delta\}$ has phase-type distribution, say $F_T(\cdot)$.
Claim: $F_T(t)$ $= 1 - qe^{tR}1,\ t \geq 0\,.$

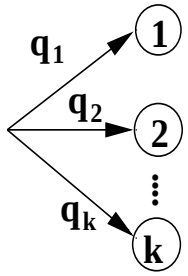
Proof. $P(T > t) = P\{X_t \neq \Delta\} = \sum_k q(e^{tR})_k = qe^{tR}1$.

Parameters:

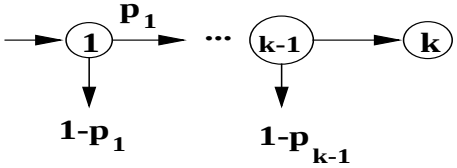
density	$f_T(t) = qe^{Rt}r$
Laplace transform	$\int_0^\infty e^{-xt}F_T(dt) = q[xI - R]^{-1}r$
nth moment	$\int_0^\infty t^n F_T(dt) = (-1)^n n!\, qR^{-n}1$
(mean $= -qR^{-1}1$)	

Special Cases:

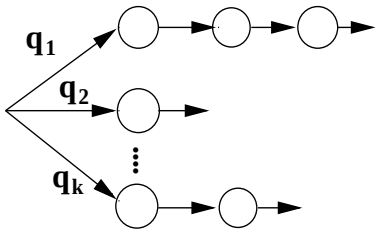
- Exponential (μ) : $R = [-\mu]$ and $q = 1$.
- Erlang: $\rightarrow \boxed{1} \rightarrow \boxed{2} \rightarrow \boxed{\text{K}}$ iid tasks / phases $\left(C^2(T) = \frac{1}{K}\right)$.
- Generalized Erlang: exponential phases in series (tandem) ($C^2 < 1$).
- Hyperexponential: K tasks in parallel (mixture) ($C^2 > 1$).



- Coxian: K phases; end at phase k with probability p_k.



- Minimum of exponential random variables is exponential.
- Max of exponential random variables is phase-type: e.g., $X_i \sim \exp(1)$ iid. This easily implies that $E(\max X_i) = \sum_i \frac{1}{i}$, Var $(\max\ X_i) = \sum_i \frac{1}{i^2}$ bounded!
- Erlang mixtures:



Importance of Phase-type distributions.

- Empirical + wishful thinking: homogeneous human tasks are exponential.
- Richness: the family of phase-type distributions is dense among all distributions on $[0, \infty)$. For every non-negative distribution G, there exists a sequence of phase-type distributions $F_n \ni F_n \Rightarrow G$.
  (In particular, we can guarantee convergence of any finite number of moments.)

  Dense subfamilies: Coxian, Erlang mixtures.

  For Erlang mixtures, this can be explained by the following two facts:

  1. The family of discrete distributions is dense.

  2. Constants can be approximated by Erlang distributions. Therefore, discrete distributions can be approximated by Erlang mixtures.
- Modelling, via the *method* of *phases*. For example, consider M/PH/1 queue (see HW).

  M/PH/1: state-space is (i, k) $(i =$ number in queue; $k =$ phase of service) or 0; e.g., $0 \xrightarrow{\lambda q_k} (1, k)$.

Representation directly in terms of (q, P, m).

Denote here $R = [I - P]^{-1}$ (as in Mandelbaum & Reiman).
Average work content $E(T) = qRm$ $(= \sum_j q_j R_{jk} m_k)$.

$$\text{Moments:} \qquad E(T^n) = n!\, q(RM)^n q\ , \qquad \text{where } M = \begin{bmatrix} m_1 & & 0 \\ & \ddots & \\ 0 & & m_K \end{bmatrix}$$

$$\frac{E(T^2)}{2(E(T))^2} = \frac{1 + C^2(T)}{2} = \frac{q(RM)^2 1}{(qRM1)^2}$$