A formula for tail probabilities of Cox distributions

Ger Koole
Vrije Universiteit Amsterdam

Shortened version published in
Journal of Applied Probability **41**: 935–938, 2004

Abstract

We derive a simple recursive scheme for calculating tail probabilities of Cox distributions. This is particularly useful for the computation of certain performance measures in queueing systems. An example of a call center model is provided.

A Cox distribution with $n > 0$ phases can be defined as the time until absorption into state 0, starting from state n, of the Markov process depicted in Figure 1. The process remains in state $1 \le k \le n$ an exponentially distributed amount of time with parameter μ_k. Upon departure from state k the process moves to state 0 with probability α_k and moves to state $k-1$ with probability $\bar{\alpha}_k = 1 - \alpha_k$. To avoid trivial situations we assume that $\alpha_k < 1$ for all k.

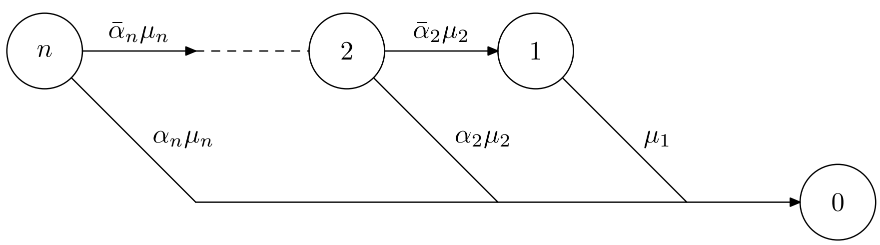

Figure 1: A Cox distribution with n phases.

Cox distrubitions are useful when approximating general non-negative distributions using exponential phases. It has been shown that the class of Cox distributions is dense in the class of all non-negative distributions (Schassberger [4]).

Define Y_k as the time until absorption in 0 starting from state k, and let $F_k = 1 - \bar{F}_k$ be the distribution function of Y_k.

Theorem 1 *Let all μ_k, $1 \le k \le n$, be different. Then $\bar{F}_k$ for $1 \le k \le n$ is given by*

$$\bar{F}_k(t) = \sum_{i=1}^{k} c_{ik} \mathrm{e}^{-\mu_i t}$$

for all $t \geq 0$*, with* c_{ik} *as follows:*

$$c_{ik} = \begin{cases} 1 & \text{if } i = k = 1; \\ \dfrac{\mu_k c_{ik-1} \bar{\alpha}_k}{\mu_k - \mu_i} & \text{if } k > 1, i < k; \\ 1 - \displaystyle\sum_{j=1}^{k-1} c_{jk} & \text{otherwise, i.e., if } i = k > 1. \end{cases}$$

Theorem 1 is a special case of Theorem 2, that deals with the case of general parameter values. Because of its simplicity and relevance for applications, we formulated it. Before continuing with Theorem 2, we introduce some additional notation.

Define:

- $m(j) = \#\{i | \mu_i = \mu_j, 1 \leq i < j\}$, i.e., the number of times μ_j occurs in $\mu_1, \ldots, \mu_{j-1}$;
- $h(j,k) = \min_i\{\mu_i = \mu_j, j < i \leq k\}$ if such an i exists, 0 otherwise, i.e., the lowest higher numbered phase with the same parameter in the Cox distribution with k phases;
- $n(j) = \max_i\{\mu_i = \mu_j, 1 \leq i < j\}$ if $m(j) > 0$, 0 otherwise, i.e., the highest lower numbered phase in the Cox distribution with the same parameter;
- $l(k) = \min_i\{\mu_i = \mu_k, 1 \leq i \leq k\}$, i.e., the lowest numbered phase with parameter μ_k.

For convenience we also take $c_{0k} = 0$ for all k.

Theorem 2 *For arbitrary* $\mu_k > 0$*,* $\bar{F}_k$ *for* $1 \leq k \leq n$ *is given by*

$$\bar{F}_k(t) = \sum_{i=1}^{k} c_{ik} t^{m(i)} \mathrm{e}^{-\mu_i t} \tag{1}$$

for all $t \geq 0$*, with* c_{ik} *as follows:*

$$c_{ik} = \begin{cases} 1 & \text{if } i = k = 1; \\ \dfrac{\mu_k c_{ik-1} \bar{\alpha}_k - c_{h(i,k)k}(m(i)+1)}{\mu_k - \mu_i} & \text{if } \mu_i \neq \mu_k; \\ \dfrac{\mu_k c_{n(i)k-1} \bar{\alpha}_k}{m(i)} & \text{if } \mu_i = \mu_k, m(i) > 0, k > 1; \\ 1 - \displaystyle\sum_{1 \leq j < k : m(j) = 0, j \neq i} c_{jk} & \text{otherwise, i.e., if } \mu_i = \mu_k, m(i) = 0, k > 1. \end{cases}$$

Proof We extend the proof of Riordan [3], p. 110–111, who treats a special case of Theorem 1 (see the paragraph on special cases below). From properties of the exponential distribution we find for $h > 0$ small:

$$F_k(t+h) = \mu_k h(\alpha_k + \bar{\alpha}_k F_{k-1}(t)) + (1 - \mu_k h) F_k(t) + o(h),$$

where $o(h)$ has the usual meaning of $\lim_{h\to 0} o(h)h^{-1} = 0$. Rewriting and taking the limit as $h \to 0$ gives

$$\bar{F}_k'(t) = \mu_k(\bar{\alpha}_k \bar{F}_{k-1}(t) - \bar{F}_k(t)),$$

for $k > 0$. Plugging in equation (1) leads to:

$$\sum_{i=1}^{k} c_{ik}\Big[m(i)t^{m(i)-1}\mathrm{e}^{-\mu_i t} - \mu_i t^{m(i)}\mathrm{e}^{-\mu_i t}\Big] = \sum_{i=1}^{k-1} \mu_k \bar{\alpha}_k c_{ik-1} t^{m(i)}\mathrm{e}^{-\mu_i t} - \sum_{i=1}^{k} \mu_k c_{ik} t^{m(i)}\mathrm{e}^{-\mu_i t}.$$

Equating coefficients of $t^m\mathrm{e}^{-\mu t}$ for equal m and μ leads to the expressions of c_{ik} for $\mu_i \neq \mu_k$ and for $\mu_i = \mu_k$, $m(i) > 0$. The expression for $\mu_i = \mu_k$, $m(i) = 0$ follows from $\bar{F}_k(0) = 1$. □

Numerical considerations Calculating the coefficients directly using (1) can lead to numerical problems. E.g., if $t > 1$ and many phases have the same parameter, then, for i big, c_{ik} will approach 0 and $t^{m(i)}$ will get very big, leading to numerical instabilities. In this case it is better to scale the parameters such that t can be omitted, i.e., μ_i should be replaced by $t\mu_i$. This has been done in the algorithm below. Likewise one should be careful when $\mu_i \approx \mu_j$ for certain i, j; taking them equal might give a very good approximation while avoiding numerical difficulties. Finally, when $\bar{F}_k(t)$ gets close to 1, then numerical problems can occur when computing $\bar{F}_{k+n}(t)$. If $\alpha_k = 0$ for all k, then $\bar{F}_k(t)$ is increasing in k. For this reason $\bar{F}_{k+n}(t)$ should be set to 1 if $\bar{F}_k(t) \approx 1$ in this case.

Algorithm to compute $F_K(t)$
for $j, k = 1$ **to** K **do** *calculate* $m(j)$**,** $h(j,k)$**,** $n(j)$
$c_{11} = 1$
for $k = 2$ **to** K **do**
  $c_{0k} = 0$
  for $i = k$ **downto 1 do**
    if $\mu_i \neq \mu_k$ **then** $c_{ik} = \dfrac{t\mu_k c_{ik-1}\bar{\alpha}_k - c_{h(i,k)k}(m(i)+1)}{t(\mu_k - \mu_i)}$
    elseif $m(i) > 0$ **then** $c_{ik} = \dfrac{t\mu_k c_{n(i)k-1}\bar{\alpha}_k}{m(i)}$
    endif
  endfor
  $c_{l(k)k} = 1 - \displaystyle\sum_{1\le j<k: m(j)=0, j\neq l(k)} c_{jk}$
endfor
$F_K(t) = 1 - \displaystyle\sum_{i=1}^{K} c_{iK}\mathrm{e}^{-\mu_i t}$

Special cases If $\mu_i = \mu$ for all i then we obviously get a Gamma distribution; in this case $c_{1k} = 1$ for all k and $c_{ik} = \mu c_{i-1k-1}/(i-1)$, leading to $c_{ik} = \mu^{i-1}/(i-1)!$, in case $1 < i \leq k$. These are indeed the coefficients of the Gamma distribution.

The case $\mu_i = C + i$ is discussed on p. 110–111 of Riordan [3] (see Garnett et al. [2] for a discussion and other references). For convenience we start numbering at $k = 0$, this also corresponds to the application to which Riordan applies his results (for which k represents the number of waiting customers in a queue with abandonments). Riordan finds

$$c_{ik} = \frac{C(C+1)\cdots(C+k)}{k!}(-1)^i \binom{k}{i} \frac{1}{C+i}.$$

It is easily seen that indeed

$$c_{ik} = \frac{(C+k)c_{ik-1}}{k-i},$$

as was to be expected from our derivation.

In fact, Riordan derives a closed-form expression for $\sum_{k=0}^{\infty} p_k \bar{F}_k(t)$, with p_k the stationary probability of finding k waiting customers in the system under study. Note that it is computationally more efficient to compute the coefficients recursively, instead of utilizing a closed-form solution.

Example As an example, consider a standard Erlang delay queueing system with s servers, Poisson arrivals (with rate λ) and exponential service time distributions (with rate μ) with the additional feature that customers abandon while waiting. This is a good model for call centers (for an overview of call center models, see Gans et al. [1]). If the abandonment rate is constant for each customer then we get the model of Riordan [3]. However, it is not unlikely that the abandonment rate is a function of the position in the queue, for example because of announcements made to the callers. Let us assume that the abandonment rate is of the form $\alpha + k\gamma$, with k the position in the queue. Then the total abandonment rate when there are k customers waiting is $k\alpha + \frac{1}{2}k(k+1)\gamma$. We compute the service level which is defined in this case as the probability that a "test customer" (who has infinite patience) spends longer than 20 seconds (standard in call centers) in queue. This is done by calculating the stationary probabilities p_k and the tail probabilities $\bar{F}_k(t)$. The results for various situations can be found in Table 1. The conclusion to be drawn is that abandonments improve the service level (SL), with only a low decrease in server productivity (or, equivalently, a small abandonment probability). This phenomenon is even stronger in the case of abandonment rates that increase in the server position. This calls for stimulating callers to abandon when they enter a long queue.

Acknowledgment I would like to thank René Swarttouw for some useful discussions, and Auke Pot for finding a few typos and a mistake in the algorithm.

References

[1] N. Gans, G.M. Koole, and A. Mandelbaum. Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management*, 5:79–141, 2003.

α	γ	SL	abandonment probability
0	0	0.51	0
0.1	0	0.66	0.03
0.1	0.1	0.80	0.06
0.25	0	0.75	0.05
0.25	0.1	0.83	0.07
0.5	0	0.82	0.07
0.5	0.5	0.92	0.09

Table 1: The call center example with $\lambda = 10$ per minute, $\mu = 1$ per minute, and $s = 11$.

[2] O. Garnett, A. Mandelbaum, and M. Reiman. Designing a call center with impatient customers. *Manufacturing & Service Operations Management*, 4:208–227, 2002.

[3] J. Riordan. *Stochastic Service Systems*. Wiley, 1961.

[4] R. Schassberger. *Warteschlangen*. Springer-Verlag, 1973.