Risk Theory and the Single-server Queue

By Hilary L. Seal, Yale University

In certain circumstances the probability that (a) a risk business with positive sums at risk and an initial risk-reserve of w will not be technically ruined in an interval (0,t) is exactly the same as the probability that (b) a potential (virtual) customer arriving at an orderly queue at an epoch t after the server commenced business with an empty queue would not have to wait longer than an interval w for his service to commence. Under the same conditions the probability that (c) a risk business whose latest claim has just left it with a risk-reserve of w will not be technically ruined through the nth following claim is equivalent to the probability that (d) the nth customer arriving after the single server commenced business by serving a customer (number zero) would have not more than an interval w to wait for the commencement of his service.

There is a substantial literature on queueing theory with excellent texts that include detailed treatment of the first problem (*Beneš*, 1963; *Cohen*, 1969; *Prabhu*, 1965) or the second (*Cohen*, 1969; *Feller*, 1971; *Pollaczek*, 1957; *Prabhu*, 1965) or both. Many of the theoretical results of risk theory were discovered independently by queueing theoreticians (*Seal*, 1969, *passim*) and there are queueing formulas that could well be applied to risk theoretical models. It may be convenient for actuaries to have a short, simple article demonstrating the probability equivalences mentioned in the first paragraph so that they may save themselves from developing risk formulas that are already in the queueing literature.

Basic assumptions

There are several random variables that are basic to both risk and queueing models:

(i) A series of random variables representing successive interclaim (inter-arrival) intervals. In the general formulation the process of claim (arrival) epochs is supposed stationary – and stationarity may sometimes be achieved only after rescaling of time measurements. Successive intervals are not necessarily independent and multiple claims (arrivals) are not excluded.

Because of stationarity the intervals all have the same distribution function A(·) with mean *a*. The mathematics of such stationary point processes is found in *McFadden* (1962), *McFadden & Weissblum* (1963) and *Kuznecov & Stratonovič* (1956). Simple examples of stationary point processes are Poisson and renewal processes, the latter introduced in queueing theory by *Palm* (1943). Others are listed by *Seal* (1969, Ch. 4).

Expressions for probabilities (a) and (b) can be obtained in the general case (*Beneš*, 1963; *Seal*, 1969) but they are only equivalent (as we shall see) when the intervals between claims (arrivals) are independently and identically distributed (iid) or, more exactly, are exchangeable random variables. This restriction applies also to probabilities (c) and (d).

(ii) The random variable B denoting the individual claims sizes (service times) supposed to be independent of one another and of the epoch of occurrence of the claim (customer arrival). We write B(·) for the distribution function of B and *b* for its mean.

(iii) The random variable X(t) representing the aggregate claim amounts (service times) that are presented for payment (service) in the interval (0,t). This random variable is the sum of the N(t) random claim (customer arrival) variables B that have occurred in the interval. Note that N(t) is itself a random variable whose distribution can be determined in terms of the distribution function of the interval between an arbitrary epoch and the occurrence of the mth following claim (arrival), m = 1, 2, 3, ... (*McFadden*, 1962). Write $p_n(t) = \Pr\{N(t) = n\}$.

The basic results presented in this paper are obtained under the assumption that X(t) has independent and stationary increments.

The random variable X(t) is related to the A(·) and B(·) of (i) and (ii) above by means of the following formula common to risk and queueing theory:

$$\Pr\{X(t) \leq x\} \equiv F(x,t) = \sum_{n=0}^{\infty} p_n(t)\, B^{n*}(x) \qquad (1)$$

$B^{n*}(\cdot)$ being the distribution function of the aggregate of n claims (service times). Let us adopt the convention that the Laplace-Stieltjes transform of a nonnegative random variable A(·), say, is written as $\check{A}(s)$ so that

$$\check{A}(s) = \int_{0-}^{\infty} e^{-sx} dA(x)$$

the 0– in the lower limit of the integral denoting that any spike of probability at x = 0 is included. Then

$$\check{F}_t(s) = \sum_{n=0}^{\infty} p_n(t)\{\check{B}(s)\}^n \tag{2}$$

Nonruin by epoch t

We will now consider the first pair of probabilities (a) and (b) introduced in the first paragraph. The risk-reserve of a risk business at an epoch *t* after the business commenced with a capital of $w \geqslant 0$ is defined by

$$R(t) = w + (1+\eta)\, bt/a - X(t) \tag{3}$$

where $(1+\eta)\, bt/a$ is $1+\eta$ times the expected claim outgo in the interval (0,t) and $\eta \geqslant 0$ is called the risk loading. An essential feature of (3) is that the premiums are supposed to be payable continuously throughout the interval.

The probability of nonruin of the business during the interval (0,t) is the probability that $R(\tau)$, $\tau \leqslant t$, is always nonnegative or, equivalently, that the smallest value of $R(\tau)$, $\tau \leqslant t$, is nonnegative. Write

$$\begin{aligned}
U(w,t) &= \Pr\{\inf_{\tau<t} R(\tau) \geqslant 0\} \\
&= \Pr\{\inf_{\tau<t} [w + (1+\eta)\, \tau b/a - X(\tau)] \geqslant 0\} \\
&= \Pr\{w + \inf_{\tau<t} [(1+\eta)\, \tau b/a - X(\tau)] \geqslant 0\} \\
&= \Pr\{\sup_{\tau<t} [X(\tau) - (1+\eta)\, \tau b/a] \leqslant w\}
\end{aligned} \tag{4}$$

The last expression is the probability that the excess of claims over premiums does not exceed *w*, the initial risk-reserve, at any epoch in the interval (0,t). We notice particularly that both functions within the brackets refer to the first part of the interval, namely to $(0,\tau)$.

Turning now to the single-server queue in which the server is supposed idle at epoch 0 we introduce an indicator event $\bar{Q}\,(\cdot)$ (read as "no queue") such that

$$\bar{Q}(u) = \begin{cases} 0 \text{ if server busy with a customer at epoch } u \\ 1 \text{ if server idle at epoch } u \end{cases}$$

and a random variable W(t) representing the time a customer would have to wait for the commencement of his service if he arrived at the server (or the end of the queue) at an arbitrary epoch t.

The interval (0,t) commences with an idle period during which the server is waiting for his first customer. This customer arrives and, in effect, presents the server with a service load that is a realization of the random variable B. The server immediately starts to reduce this service load (the size of which is unknown to him) and will again become idle unless a second customer arrives before the server has finished with the first. The waiting time of a customer arriving for service at epoch t is thus equal to the aggregate service load imposed on the server by customers arriving during the interval (0,t) minus the time during which the server has been reducing this load by providing service. The deductive item is equal to the time elapsed t less the aggregate of the server's idle periods. We thus have

$$W(t) = X(t) - t + \int_0^t \bar{Q}(u)\, du \tag{5}$$

and will demonstrate (*Beneš*, 1963 Lemma 1.1) that

$$W(t) = \sup_{\tau < t} [X(t) - X(\tau) - t + \tau] \tag{6}$$

Using (5) for any epoch $\tau \leqslant t$ we have

$$W(\tau) = X(\tau) - \tau + \int_0^\tau \bar{Q}(u)\, du$$

and on subtracting this from (5)

$$W(t) = W(\tau) + X(t) - X(\tau) - t + \tau + \int_\tau^t \bar{Q}(u) du$$

$$\geqslant X(t) - X(\tau) - t + \tau \quad \text{since the other terms are nonnegative.}$$

On the other hand supposing that σ is the last epoch prior to t at which the server was idle.

$$W(t) = X(t) - X(\sigma) - t + \sigma \qquad \text{namely the overload since epoch } \sigma$$
$$\leqslant \sup_{\tau<t}[X(t) - X(\tau) - t + \tau] \qquad \text{by definition of a supremum}$$

Combining the inequalities for W(t) (6) results.

Now in general $\sup\limits_{\tau<t} [X(t) - X(\tau) - t + \tau]' \neq \sup\limits_{\tau<t} [X(\tau) - \tau]$.

But if the discrete increments in $X(\tau)$ are independently and identically distributed we can, as it were, reverse the time scale and write $X(t)-X(\tau) = X(t-\tau)$ in probability. Then

$$\Pr\{W(t) \leqslant w\} = \Pr\{\sup_{\tau<t}[X(t) - X(\tau) - t + \tau] = \Pr\{\sup_{\tau<t}[X(\tau) - \tau] \leqslant w\} \qquad (7)$$

Comparing (4) and (7) we see that they are equivalent provided

$$1 + \eta = \frac{a}{b} \qquad (8)$$

This is not a restriction since, in the risk model, b is in monetary units which may be chosen arbitrarily but consistently so that w and $R(\tau)$ are also expressed in those units. In both models a is in time units and is often chosen as unity. We note that the ratio b/a is sometimes written as ρ in queueing theory. Furthermore U(w,t) in risk theory, defined as zero when $w = 0-$, written as U(0,t) > 0 when $w = 0$, and increasing monotonically to unity as $w \to \infty$, can be regarded as the distribution function of a random variable W(t).

Nonruin through the nth claim

The initial conditions are different for the second pair of probabilities (c) and (d) specified in the first paragraph. The time origin must be chosen at the occurrence of a claim (customer arrival). In the risk theoretic case w is the risk reserve after paying this claim and in the queueing model the initial customer is assumed to be served before the count of n ($n = 1, 2, 3, \ldots$) further customers begins.

We now write A_j for the random variable representing the interval between the occurrence (arrival) of the (j–1)th and jth claim (customer), $j = 1, 2, 3, \ldots$, A_1 thus being the epoch of the first claim (customer arrival) after the new time origin. Let B_j be the amount (service time) of the jth claim (customer). The Aj's and Bj's have distribution functions A(•) and B(•), respectively. For convenience we choose a, the mean of A(·), to be the unit of time in both models. In the risk theoretic model $B_j - (1 + \eta)bA_j$ is the reduction in the risk-reserve between the (j–1)th and jth claims. Write

$$S_n = \sum_{j=1}^{n} \{B_j - (1+\eta)bA_j\} \tag{9}$$

for the aggregate of such depletions through the nth claim; then the risk reserve after the nth claim is defined to be

$$R_n = w - S_n \qquad n = 1, 2, 3, \ldots \tag{10}$$

and technical ruin will have occurred at or before the nth claim if any member of the series $R_1, R_2, \ldots R_n$ is negative.

Writing $W_n(w)$ for the probability of nonruin through the nth claim

$$W_n(w) = \Pr\{\min[R_1, R_2, \ldots R_n] \geqslant 0\}$$

$$= \Pr\{\max[0, S_1, S_2, \ldots S_n] \leqslant w\} \qquad \text{by (10)} \tag{11}$$

Reverting to the queueing model write W_j for the waiting time of customer j(j = 1, 2, ...). This customer arrives at epoch $A_1 + A_2 + \ldots + A_j$, his service commences at epoch $A_1 + A_2 + \ldots + A_j + W_j$ and it terminates at epoch $A_1 + A_2 + \ldots + A_j + W_j + B_j$. The next customer arrives at epoch $A_1 + A_2 + \ldots + A_j + A_{j+1}$ and has a waiting time

$$W_{j+1} = (A_1 + A_2 + \ldots A_j + W_j + B_j) - (A_1 + A_2 + \ldots + A_{j+1})$$

$$= W_j + B_j - A_{j+1}$$

provided this is positive; and in the opposite case $W_{j+1} = 0$ implying that the customer is served immediately.

It is convenient to write $B_j - A_{j+1} = Z_{j+1}$ (j = 0, 1, 2, ...), B_0 being the service time of the customer whose service commences at the time origin. Assuming that both A's and B's are iid the Z's are iid. We then have

$$\begin{aligned} W_n &= \max(W_{n-1} + Z_{n-1}, 0) \\ &= \max[\{\max(W_{n-2} + Z_{n-2}), 0\} + Z_{n-1}, 0] \\ &= \max[\max(W_{n-2} + Z_{n-2} + Z_{n-1}, Z_{n-1}), 0] \\ &= \max[W_{n-2} + Z_{n-2} + Z_{n-1}, Z_{n-1}, 0] \\ &= \max[W_{n-2} + S^*_{n-2}, S^*_{n-1}, 0] \end{aligned}$$

where $$S^*_k = \sum_{i=k}^{n-1} Z_i$$

namely the sum of the last n–k Z's through Z_{n-1} where each Z has the same distribution. Continuing the recurrence it is seen that

$$W_n = \max[S^*_0, S^*_1, \dots S^*_{n-1}, 0] \qquad n = 1, 2, 3, \dots$$

If we now write $$\underline{S}_{k-1} = \sum_{i=0}^{k-1} Z_i \qquad k = 1, 2, 3, \dots$$

$\underline{S}_{k-1}$ has the same distribution as S^*_{n-k} since each is the sum of k random variables Z. Hence W_n has the same distribution as $\max[\underline{S}_{n-1}, \underline{S}_{n-2}, \dots \underline{S}_1, 0]$ and

$$\Pr\{W_n \leqslant w\} = \Pr\{\max[0, \underline{S}_1, \dots \underline{S}_{n-1}] \leqslant w\} \tag{12}$$

If we write $(1+\eta)b = 1$ [cp.(8)] S_n of relation (9) is the sum of n random variables each of which has the same distribution as B–A = Z. With this choice of monetary units in the risk theoretic case S_n has the same distribution as $\underline{S}_{n-1}$ and (11) and (12) are identical.

References

Beneš, V. E. (1963): "General Stochastic Processes in the Theory of Queues." Addison-Wesley, Reading, Mass., viii + 88 pp.

Cohen, J. W. (1969): "The Single Server Queue." North-Holland, Amsterdam, xiv + 657 pp.

Cramér, H. (1955): "Collective Risk Theory." Esselte, Stockholm, 91 pp.

Feller, W. (1971): "An Introduction to Probability Theory and Its Applications", Vol. II, Second Ed. Wiley, New York, xxiv + 669 pp.

Kuznecov, P. I. & R. L. Stratonovič (1956): "On the mathematical theory of correlated random events." Selected Trans. Math. Statist. Prob. *7*, 1968, 1–16.

McFadden, J. A. (1962): "On the lengths of intervals in a stationary point process." J. Roy. Statist. Soc., B, *24*, 364–382.

McFadden, J. A. & W. Weissblum (1963): "Higher-order properties of a stationary point process." J. Roy. Statist. Soc., B, *25*, 413–431.

Palm, C. (1943): "Intensitätsschwankungen im Fernsprechverkehr." Esselte, Stockholm, 189 pp.

Pollaczek, F. (1957): "Problèmes Stochastiques Posés par le Phénomène de Formation d'une Queue d'Attente à un Guichet et par des Phénomènes Apparentés." Gauthier-Villars, Paris, 123 pp.

Prabhu, N. U. (1965): "Queues and Inventories: A Study of Their Basic Stochastic Processes." Wiley, New York, xi + 275 pp.

Seal, H. L. (1969): "Stochastic Theory of a Risk Business." Wiley, New York, xiii + 210 pp.

Zusammenfassung

Die Wahrscheinlichkeitsverteilung der Wartezeit eines Kunden in einer Warteschlange und diejenige des Maximums des Verlustes im Risikoprozess sind eng miteinander verbunden. In dieser Arbeit wird dieser Zusammenhang unter besonderer Berücksichtigung versicherungsmathematischer Terminologie und Interessen dargestellt.

Summary

The probability distribution of the waiting time of a customer in a single-server queue and that of the maximum loss in the risk process are closely related. The connection is discussed with particular emphasis on actuarial language and interest.

Résumé

Les fonctions de distribution de la période d'attente d'un client qui se joint à une queue et du maximum de la perte dans un processus stochastique du risque sont étroitement liées. Dans cet article on déduit ce rapport en tenant compte spécialement de la terminologie et de l'intérêt dans les assurances.

Riassunto

Le funzioni di distribuzioni del tempo di attesa di un cliente che si aggiunge a una fila d'attesa e del massimo della perdita in un processo aleatorio del rischio sono strettamente unite.
In questo articolo si dimostra questo rapporto con particolare riguardo alla terminologia e all'interesse assicurativo.