Proceedings of the 2020 Winter Simulation Conference
K.-H. Bae, B. Feng, S. Kim, S. Lazarova-Molnar, Z. Zheng, T. Roeder, and R. Thiesing, eds.

METHODS FOR ESTIMATING INCIDENCE RATES AND PREDICTING INCIDENT NUMBERS IN MILITARY POPULATIONS

Stephen Okazawa

Defence Research and Development Canada
Centre for Operational Research and Analysis
101 Colonel By Drive Ottawa, ON K1A 0K2, CANADA

ABSTRACT

Monitoring of the health of military populations and developing effective personnel management plans relies on the ability to measure and predict the incidence of important events such as attrition, training failures, promotions and transfers between groups. Incidence rates are widely relied on to report the prevalence of these events and for modelling to predict future events. However, calculating and using incidence rates in real-world scenarios is not straightforward, and challenges are frequently encountered. This paper provides a detailed mathematical development of equations that define incidence rates, Bayesian techniques for estimating rates based on the available evidence and quantifying how certain the estimate is, and a beta-binomial model for predicting the variation in future event numbers. These methods do not require significant additional effort or resources to apply in typical military workforce modelling applications, but produce meaningful improvements in the depth and accuracy of the analysis.

1 INTRODUCTION

Quantifying and predicting the incidence of events in populations of military personnel are important aspects of monitoring and maintaining the health of a military workforce. Examples of significant events that are indicators of population health include attrition (releases from the military population), training failures, transfers between groups, and promotions. Changes in the frequencies of these events may indicate a problem that requires action, and making good planning decisions often relies on being able to predict these events several years into the future.

The metric most often used to quantify the occurrence of such events is the incidence rate, i.e. the per capita number of events in the population in a given time period. Incidence rates are useful because they are intuitive, they can be compared between different populations, and they can be used to model how a population will evolve over time. In particular, attrition rates are a critical health metric that are monitored closely across the Canadian Armed Forces (CAF). Any significant upward change in the rate may require prompt action in the form of adjustments to plans, policies and compensation to stem losses.

In the context of military operational research, incidence rates are employed extensively in analytical models of military populations. For example, attrition rates and course failure rates have been used in a stochastic model of Australian Naval aviation training (Pike et al. 2018), multiple discrete event (DE) models of CAF occupation training (Henderson and Bryce 2019; Zegers and Isbrandt 2010; and Straver et al. 2009), and a system dynamics model of the Canadian Air Force pilot occupation which also includes transfer rates between aircraft fleets (Séguin 2015). Markov chain models have a long history of application in military workforce modelling and rely heavily on measured incidence rates (particularly retention rates and group transition rates) to establish state transition probabilities, for example, Merck and Hall (1971), Skulj et al. (2008), Straver et al. (2013), and Hoecherl et al. (2016). Boileau (2012) and Vincent and Okazawa (2019) describe models based on solving systems of equations to determine the steady state

condition of Canadian military and civilian occupations based on rates of attrition and occupation transfer. The use of incidence rates is also not limited to events affecting personnel, for example, Talafuse et al. (2019) employ flight cancellation rates in a stochastic model of a United States Air Force flying training unit.

Models such as these are used routinely to inform critical personnel planning decisions, and their validity is dependent on the ability to accurately measure incidence rates and predict future event numbers. Yet this apparently simple concept is complex under the surface and should be approached carefully.

This paper provides a detailed development of the mathematics for measuring and using incidence rates in military workforce analysis. First, it will review existing work on basic equations for calculating incidence rates. Second, it will apply Bayesian techniques to estimate the incidence rate given the available data and to quantify how certain the rate estimate is. Finally, it will demonstrate the beta-binomial distribution as a model for predicting the variability in event numbers over time.

2 WHAT IS DIFFICULT ABOUT MEASURING AND USING INCIDENCE RATES?

At first glance, the incidence rate of an event is simply the number of events observed in some time interval divided by the population size, and we can then predict the number of events that will occur in a future interval by multiplying the rate by the current population. So where do difficulties arise?

Firstly, in most cases, we must account for the fact that the population is not constant over the interval. The event itself usually affects the population (such as in attrition events) and there are likely to be other flows occurring into and out of the population at the same time, such as recruitment and transfers, which may be causing the population to grow or shrink. So the question is, what population to use in the denominator when calculating the incidence rate? How to correctly account for these other flows while measuring the incidence rate is not obvious.

Secondly, the event we are interested in may be relatively rare, the population itself may be small to begin with, or we may not have access to a long history of data. The result is that we often have fewer incidents of the event in our data than we would like in order to be certain of the measured rate. But exactly how does the amount of data affect how much certainty we should have? Further, if the rate is uncertain, what can we do about this? Intuitively, we might include data from other populations, but exactly how do we incorporate this information and how does this affect our conclusion about the incidence rate for the population we are interested in?

Thirdly, actual event numbers will vary (sometimes significantly) over time. When we calculate a single average rate and apply this to a population, we obtain a prediction of the mean number of events that will occur, but there is additional value in quantifying the probability of deviations from the mean. Stochastic DE models and Monte Carlo analysis are frequently employed to predict and analyze the variability of future events. However, we must be careful to understand how the variability in the model arises, and confirm that it is supported by evidence.

The following sections of this report will explore each of these three issues in detail and will present techniques that address the challenges and questions raised.

3 MEASURING AN INCIDENCE RATE IN THE PRESENCE OF OTHER FLOWS AFFECTING THE POPULATION

Let us first define exactly what the incidence rate means, consistent with intuition and how we expect to be able to use the rate. We will assume that the event results in a loss from the population which is the case for most events that we are interested in, including attrition, transfers, promotions, and training failures. We will also assume that the time interval of interest is a one year period. So all incidence rates will be per capita annual rates. These are common assumptions in many applications, but only slight changes to the math are required if these assumptions do not hold.

If we define $E(t)$ as the cumulative sum of events over time, then the notation $E|_0^1$ is the number of events that occurred during the interval from $t = 0$ to $t = 1$. The populations at the start and end of this

interval are P_0 and P_1 respectively. This notation clearly distinguishes quantities that are measured over an interval of time from quantities that are measured at a point in time.

We define the incidence rate γ as the coefficient relating the number of events over the interval to the population at the start of the interval.

$$E|_0^1 = P_0\gamma \tag{1}$$

This is the calculation that one expects to be able to perform using an incidence rate as it answers the question of how many events will occur in a particular population during the next interval. Hence, we ensure that all of the subsequent math is consistent with this definition.

Note, this is in the absence of other flows affecting the population, and in this simple case the incidence rate equals the observed events in the interval divided by the population at the start of the interval.

$$\gamma = \frac{E|_0^1}{P_0} \tag{2}$$

Following from equation (1), the population at the end of the interval is

$$P_1 = P_0(1-\gamma). \tag{3}$$

In general, if there are no other flows, the population over time is an exponential decay.

$$P(t) = P_0(1-\gamma)^t \tag{4}$$

Now we introduce other flows into and out of the population, which will be represented by the flow rate $f(t)$ in units of people per unit time. This must include all flows entering and exiting the population other than the events counted in $E|_0^1$. For simplicity, we will refer to these other flows as "intake" as this reflects the direction of flow corresponding to positive f. However, this flow can be negative if it consists of a net outflow from the population. We can determine the contribution of the intake to the population at the end of the interval by convolving $f(t)$ with $P(t)$ from equation (4).

$$P_{f1} = \int_0^1 f(t)(1-\gamma)^{1-t}dt \tag{5}$$

If $f(t)$ was known, we could potentially solve this integral. However, it is most often the case that these flows, such as recruitment and transfers, are only known in aggregate for each year. If we make the assumption that $f(t)$ is constant over the interval, meaning the distribution of the intake is approximately uniform, we can replace $f(t)$ in the integral with the quantity $F|_0^1$, i.e. the net number of individuals that entered or left the population during the interval other than those counted in $E|_0^1$. We can then solve the integral analytically.

$$P_{f1} = F|_0^1 \frac{-\gamma}{\ln(1-\gamma)} \tag{6}$$

When γ is small, to a very good approximation, this can be simplified to the following using the first two terms of its Taylor series expansion.

$$P_{f1} = F|_0^1(1 - ½\gamma) \tag{7}$$

In practice, "small γ" may be taken to mean $\gamma < 30\%$. In this range, the error between equations (6) and (7) is less than about 1%. For reference, at $\gamma = 50\%$ the approximation error is about 4%. For the vast majority of cases that we study in practice, incidence rates are well below 30%.

Then the final equation for the population P_1 at the end of the interval is the sum of those that remain from the initial population and those that remain from the intake.

$$P_1 = P_0(1-\gamma) + F|_0^1(1-½\gamma) \tag{8}$$

Correspondingly, the events that occur due to γ from both the initial population and the intake are

$$E|_0^1 = (P_0 + ½F|_0^1)\gamma. \tag{9}$$

We can understand the appearance of the ½ term in these equations by considering that the intake into the population is spread out over the interval. In the absence of more detailed information, we assume that the distribution of the intake is approximately uniform, and therefore members of this group are present in the population for roughly half the interval time on average. We therefore expect the number of events in this group to be approximately half of what it would be had these entries occurred at the start of the interval. This also works for negative intake where F represents a loss of individuals from the population. In this case, we should observe fewer events due to γ, but we assume that, on average, these individuals are still present in the population for approximately half the interval before they leave. Therefore, the reduction in events is half of what it would be had they departed the population at the start of the interval.

Finally, from equation (9), the incidence rate, including intake is

$$\gamma = \frac{E|_0^1}{P_0 + ½F|_0^1}. \tag{10}$$

Note the modified denominator for the incidence rate is $P_0 + ½F|_0^1$. This represents the effective population in which the events $E|_0^1$ originate. These equations form a practical and mathematically consistent basis for both measuring incidence rates (equation 10) and for population forecasting (equations 8 and 9), and these are the equations that underlie the rest of the work in this paper. Collectively, we refer to these equations as the "half-intake method" because of the halving of the intake term.

This result, in the context of measuring attrition rates, was derived by the author in previous work (Okazawa 2007), but the derivation presented above is simpler, uses improved notation, and is generalized to apply to other types of events and other flows. Vincent et al. (2018) conducted a review of various methods for calculating attrition rates and proposed an improved method that can be used when high-resolution (e.g. daily) inflow, outflow and population counts are available. Their method, while producing results very similar to the half-intake method in practice, has the advantage that it makes no assumptions regarding the distribution of the intake and events over time. However, it is not well-suited to modelling because the equations cannot be rearranged to solve for a future population as a function of the present population, planned intake, and the incidence rate. The half-intake method has the advantage that it works well for typical time intervals (e.g. annual data), and it provides a self-consistent set of equations for both calculating rates and for population modelling.

Vincent et al. also draw attention to a body of analogous research from the field of financial analysis where the goal is to estimate the rate of return of an investment portfolio based on its past performance. They note that the financial analog of the half-intake method is known as the "simple Dietz method" after Dietz (1966).

4 QUANTIFYING THE UNCERTAINTY OF THE INCIDENCE RATE

While the equations developed in the preceding section allow us to calculate the incidence rate taking the data at face value, we have not addressed how strong the evidence is for the rate. Intuitively, we know that

if we observe three events in a population of 30, this is different from observing 300 events in a population of 3000. Both cases suggest a nominal incidence rate of 10%, but in the latter case we are much more certain of this conclusion than in the former. How can we quantify this?

We begin by proposing that the population has an underlying long-term average rate γ. We cannot directly measure γ, but as we observe events occurring in the population, we gather evidence that we can use to estimate it. If each individual in the population has probability γ of experiencing the event, then the probability of observing k events in a population n, given the rate γ, is a binomial distribution.

$$P(k, n|\gamma) = Bin(n, \gamma) \tag{11}$$

The mean and variance of the binomial distribution are given by the following

$$\mu_{bin} = n\gamma, \tag{12}$$

$$\sigma_{bin}{}^2 = n\gamma(1-\gamma). \tag{13}$$

However, the underlying incidence rate γ is unknown, and our objective is to estimate it based on the evidence we have. We can calculate a probability distribution for γ given the evidence k and n by applying Bayes' theorem.

$$P(\gamma|k, n) = \frac{P(k, n|\gamma)P(\gamma)}{\int_0^1 P(k, n|\gamma)P(\gamma)d\gamma} \tag{14}$$

$P(\gamma|k, n)$ is the posterior probability density function describing the likely values for the population's underlying incidence rate. The term $P(\gamma)$ is a distribution representing our prior knowledge of likely values for the incident rate before considering the evidence. The conjugate prior for the binomial distribution is the beta distribution. Hence, we will propose that $P(\gamma)$ can be modelled as a beta distribution with prior parameters α_0 and β_0.

$$P(\gamma) = Beta(\alpha_0, \beta_0) \tag{15}$$

If we consider the estimation of attrition rates in CAF occupations as an example, experience shows that almost any sub-population of the CAF has an attrition rate close to 7% with some variation around this typical value. Even if we had no information about a particular sub-population, we would still expect that an attrition rate around 7% is likely, and that attrition rates of 3% or 15% are very unlikely. Thus, we can empirically define our prior knowledge of the incidence rate for a given population as the frequency distribution of incidence rates observed among other associated populations in other years. We can reasonably expect the incidence rate for a randomly chosen population to follow this distribution.

The inclusion of the prior accomplishes what practitioners often do instinctively, which is to include data from other populations when we have insufficient data about the population we are interested in. But Bayes' theorem incorporates the prior information in a mathematically formal way.

We can apply the method of moments to determine the beta distribution parameters that fit the prior incidence rate frequency distribution. For example, Figure 1 shows the attrition rate frequency data for all CAF Non-Commissioned Member (NCM) occupations for each year over the last 15 years, calculated using equation (10). We filtered this data on those occupations that have at least 200 members to eliminate noise that arises in smaller populations. Also shown in the figure is the beta distribution fit to the frequency data. Subjectively, we can see that a beta prior with parameters $\alpha_0 = 9.45$ and $\beta_0 = 125$ fits the real data reasonably well. The mean of the distribution is 7.04%.

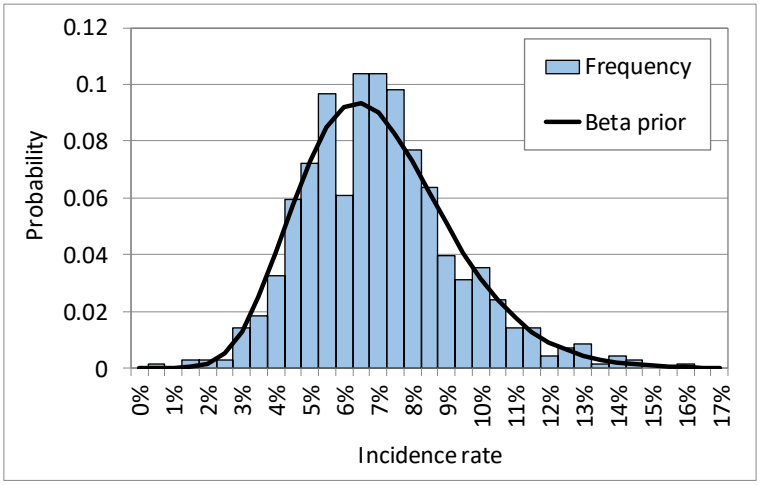

Figure 1. Distribution of attrition rates for occupations with 200 or more members over the last 15 years.

Solving Bayes' theorem using the beta prior with parameters α_0 and β_0 and the binomial model for the evidence n and k, we find that the posterior is given by another beta distribution with parameters $\alpha = \alpha_0 + k$ and $\beta = \beta_0 + n - k$.

Returning to the notation developed in the previous section, n corresponds to the effective population $P_0 + ½F|_0^1$, and k corresponds to the observed events $E|_0^1$ in the interval. Thus, the posterior probability distribution for the population's incidence rate is given by

$$Beta(\alpha_0 + E|_0^1, \beta_0 + P_0 + ½F|_0^1 - E|_0^1). \tag{16}$$

The posterior distribution describes exactly how we should update our conclusion about the population's incidence rate based on both prior knowledge from other populations and the evidence we have for the population we are interested in. The updated best estimate for the population's incidence rate $\hat{\gamma}$ is the mean of the posterior beta distribution, given by

$$\hat{\gamma} = \mu_{beta} = \frac{\alpha}{\alpha+\beta} = \frac{\alpha_0 + E|_0^1}{\alpha_0 + \beta_0 + P_0 + ½F|_0^1}. \tag{17}$$

Note that this is similar to equation (10) derived in the preceding section, but with the inclusion of the prior parameters. Importantly, the shape of the posterior distribution also tells us how certain the estimate is and how our conclusions about the incidence rate should change depending on how much data we have. Figure 2 shows the posterior distributions for the attrition rates of three hypothetical CAF occupations. As in the example used above, the evidence in each case suggests a nominal rate of 10% but the amount of data ranges from observing three events in a population of 30, to 300 events in a population of 3000. We can see that in the first case (shown in red) the small amount of evidence results in a posterior that does not deviate far from the prior and remains uncertain (the distribution is spread out). This weighting toward the prior minimizes noise that arises when we attempt to calculate a rate from too little data. In the third case (shown in blue) where we have ample evidence, the posterior has diverge from the prior and is tightly clustered around the nominal rate of 10% indicating a high degree of certainty in this estimate.

We can also use the posterior beta distribution to define credible intervals within which the true rate is expected to lie with a prescribed probability. For the three cases shown in the figure (red, green and blue lines), the 90% credible intervals are [4.50%, 11.2%], [6.93%, 11.4%] and [9.01%, 10.8%] respectively. This provides a concrete measure of how uncertain the estimated rate is given the available data.

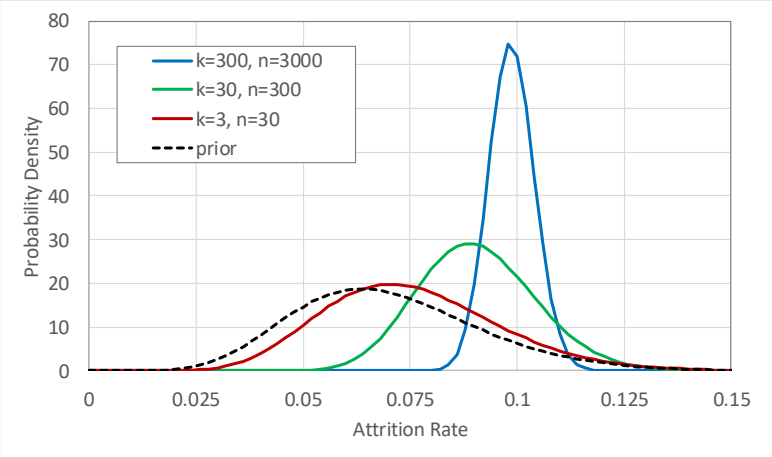

Figure 2. Posterior probability distributions for the attrition rates of three hypothetical CAF occupations.

A useful outcome of equation (15) is that it also tells us what to do when we do not have prior information. If no appropriate empirical data exists to inform the prior, we can use $\alpha_0 = \beta_0 = 1$ which corresponds to a uniform distribution. In other words, we assume all values of γ are equally likely prior to considering the evidence. In this case, equation (17) simplifies to

$$\hat{\gamma} = \frac{1+E|_0^1}{2+P_0+½F|_0^1}. \tag{18}$$

This is a suitable basic equation for estimating incidence rates in many applications. It accounts for the presence of other flows, and it incorporates a uniform prior which avoids common problems that arise in small populations without any additional effort. For example, if we are analyzing a small population and we observe no events in the interval, it avoids the naïve conclusion that the incidence rate is zero. Similarly, if all members of this small population happen to experience the event in some unlucky interval, it avoids the naïve conclusion that the incidence rate is 100%. Such problems are common when estimating incidence rates for CAF members in very specific circumstances. For example, if we want to measure the attrition rate among officers with 40 or more years of service, this will be a very small population, and the behavior of other populations is unlikely to be informative as a prior.

Another convenient feature of Bayes' theorem is that it allows for repeatedly updating the rate estimate as more evidence becomes available. After each year, we collect a new set of data, and the previous year's posterior becomes the prior that we update with the new evidence. This yields the following general solution for the posterior, summing over each year i included in the evidence.

$$Beta\left(\alpha_0 + \textstyle\sum_i E|_i^{i+1}, \beta_0 + \sum_i P_i + ½F|_i^{i+1} - E|_i^{i+1}\right) \tag{19}$$

The corresponding general equation for estimating the incidence rate based multiple years of data is

$$\hat{\gamma} = \frac{\alpha_0+\sum_i E|_i^{i+1}}{\alpha_0+\beta_0+\sum_i P_i+½F|_i^{i+1}}. \tag{20}$$

This result is similar to what is termed the Weight Average Attrition Rate (WAAR) described by Vincent et al. (2018), but here we have derived it using Bayes' theorem and incorporated prior information. Vincent et al. also discuss other techniques for estimating a rate from multiple years of data, but they found that the choice does not make a significant difference in practice. However, the incorporation of the prior

information in the above equation does have practical importance when measuring rates for small populations or rare events.

Finally, to estimate the number of events in a future interval, we take the mean of the binomial distribution using the estimated incidence rate. From equation 12, this is

$$\widehat{E|_0^1} = (P_0 + ½F|_0^1)\hat{\gamma}. \tag{21}$$

This is identical to equation (9), but here we are deriving the expected value for a stochastic outcome.

This procedure for estimating incidence rates is mathematically consistent with the equations developed in the previous section, but is also robust in dealing with situations where we have varying amounts of evidence about the populations we are analyzing, and it quantifies the uncertainty of the rate estimate.

5 MODELLING THE VARIATION IN EVENT NUMBERS OVER TIME

So far, our objective has been to estimate a single underlying incidence rate for the events in a population and to predict the corresponding number of events that will occur in a future interval. This is sufficient if we are mainly interested in forecasting the mean number of events. However, there is additional value in predicting the variability in the number of events. For example, this would allow us to predict that the expected value of the attrition from an occupation next year is 600, but that there is also a 25% chance that the attrition will exceed 800. In this case, if the occupation managers develop their force generation plans based only on the expected value of attrition, there is a 25% chance that the population will incur a net loss of at least 200 members. This risk may be unacceptable, and it might be wiser to establish force generation plans so that the risk of such a deficit is reduced to an acceptable level. Some occupations do experience significant variability in attrition numbers from year to year, so this is a real concern in practice.

In the model developed in the preceding section, variability in the number of events arises from the assumption that the events are binomially distributed given the incidence rate. DE models make it easy to implement this variability by subjecting each entity in the model to a random decision with probability $\hat{\gamma}$ of being true. This approach seems reasonable on the surface, and we can implement a multi-replication DE simulation and use Monte Carlo analysis to quantify the variation in future event numbers.

However, we must ask whether this variability is meaningful. After all, it is a consequence of a modelling assumption, and we did not do anything to ensure that it matches the actual variation in event numbers observed historically.

For example, consider the historical attrition numbers in the CAF occupation shown in Figure 3 as a solid black line. We can employ the techniques developed in the preceding sections to estimate the attrition rate and forecast the mean attrition numbers (solid blue line), assuming intake continues at the average historical rate. We also implement a simple DE model where each member of the population has probability $\hat{\gamma}$ of releasing each year, and plot the resulting attrition numbers from one forecast replication as an example (solid red line). We also show a two-standard deviation window (dashed orange lines) derived from the variance of the binomial distribution that should contain approximately 95% of the predicted attrition numbers. Note that the variability in the predicted numbers is significantly less than the variability in historical numbers. This means that the binomial model is under-dispersed compared to the real data in this example, and the stochastic output of the DE simulation is in fact not saying anything meaningful about the true probability of deviations from the mean number of predicted events. Thus DE simulation practitioners should be cautious in their interpretation of stochastic model outputs when there is an assumed uniform probability of the event occurring in the population derived from a measured incidence rate.

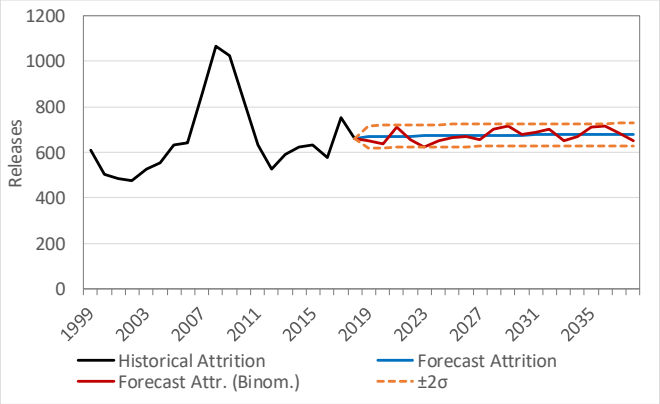

Figure 3. Historical and forecast releases for a CAF occupation (binomial model).

How can the model be revised such that the variation in event numbers over time is supported by evidence? We have so far made two important assumptions whose validity we will revisit. First, we investigate the assumption that the population is uniform, in other words, that all individuals in the population have the same probability of the event occurring. It might seem that in reality certain members of the population have a high probability of the event occurring (for example, a high probability of attrition because they are attracted to opportunities in the private sector) while most others have a low probability. However, this segmentation of the population into groups with different incidence rates must nevertheless average out to the overall rate of the whole population. From the binomial sum variance inequality (Nedelman and Wallenius 1986), the combined variances of these sub-groups will always be less than or equal to the variance of the uniform population case. So we are not any better off in terms of addressing the under-dispersion of the predicted event numbers. Further, if there is in fact a segment of the population with a high probability of the event occurring, we would have to model how individuals join this segment. We might logically propose that there is some probability that this happens, but this scenario then becomes difficult to distinguish from the original uniform population assumption. Thus, we are not addressing the problem at hand by considering alternatives to the uniform population assumption. In practice, this is a conservative assumption that allows the maximum variation in event numbers due to the randomness of which individuals experience the event in a given interval.

The second assumption we can revisit is the idea that the population's underlying incidence rate is constant in time. We can instead propose that the incidence rate varies randomly over time according to another distribution having a mean value equal to the estimated incidence rate $\hat{\gamma}$. This approach follows logically from the observation that event rates in military populations can be affected by unpredictable factors such as changes to compensation, personnel policies, and conditions in the private sector economy.

We demonstrated in the previous section that the beta distribution is an appropriate model for the variation in incidence rates, and we use it again here to describe the time variation of the incidence rate within a given population. Taken together, the random selection of which individuals experience the event in an interval is modelled by the binomial distribution, while the binomial rate parameter randomly varies from one interval to the next according to a beta distribution. The resulting compound distribution is the beta-binomial which describes the distribution of the number of events as a function of the effective population size n and the beta distribution describing the time variance of the incidence rate given by the parameters α and β. This motivation for the beta-binomial has been described by researchers in other fields. For example, Liggett and Delwiche (2005) employed the beta-binomial to model the number of individuals in a population that prefer one product over another. Their work similarly observed the insufficiency of the binomial distribution to model variation in the number of chosen products.

It will be convenient to define the beta component of the beta-binomial in terms of its mean (see equation (17)), that we set equal to the estimate for the incidence rate $\hat{\gamma}$, and its weight $\omega = \alpha + \beta$. The

weight parameter is a rough indicator of the precision of the beta distribution. The larger the value for the weight, the less the incidence rate varies over time.

The mean and variance of the beta-binomial, as functions of $\hat{\gamma}$ and ω are given by the following, where n is the effective population size $P_0 + ½F|_0^1$.

$$\mu_{bb} = n\hat{\gamma} \tag{22}$$

$$\sigma_{bb}{}^2 = \frac{n\hat{\gamma}(1-\hat{\gamma})(\omega+n)}{\omega+1} \tag{23}$$

The mean of the beta-binomial is unchanged from the mean of the binomial distribution so equation (21) still holds for predicting the average number of events. However, the variance of the beta-binomial depends on ω. For large ω (small variation of the incidence rate over time), the variance approaches that of the binomial distribution. For smaller values of ω, the variance becomes larger than that of the binomial. Thus, the weight parameter provides an additional degree of freedom to model the observed variation in event numbers.

We have specified the mean of the beta component of the beta-binomial as the estimate for the incidence rate, but we must also determine an appropriate value for the weight. Kruschke (2011) describes several approaches of varying complexity that can be taken to accomplish this in beta-binomial models. We will demonstrate an empirical approach that was used in an analysis of attrition in CAF NCM occupations. In the CAF, it is not uncommon for new occupations to be created and for old occupations to be split, merged or disbanded. Thus, many occupations have only a few years of data available in practice. While this is sufficient for estimating the mean value for γ, estimating a unique value for ω for each occupation was not feasible. Therefore, we attempted to determine a single value for ω that can be used for all NCM occupations. This can interpreted as a constant attribute that indicates, in general, how variable attrition rates are over time.

Our method for determining ω begins with a guessed value. A reasonable initial guess is the weight of the prior distribution, $\alpha_0 + \beta_0$. For each NCM occupation and every year over the last 15 years, we used equation (17) to estimate the incidence rate $\hat{\gamma}$. We then calculated the mean and variance of the occupation's attrition for the next year using equations (22) and (23). We then calculated the difference between the predicted mean attrition and the actual attrition for the next year measured in units of the predicted standard deviation σ_{bb}. This produces a result set of normalized model errors for each occupation and each predicted year. If the beta-binomial model fits the real data, the distribution of all normalized errors will be approximately a standard normal distribution. As a final step, we adjusted the weight parameter ω until we observed that the standard deviation of the normalized errors equals one. The left panel of Figure 4 shows the result of this process for the CAF NCM occupations. The weight parameter that achieved this result is 330. Subjectively, we can see that the histogram of the normalized errors closely matches the standard normal distribution shown by the black line. This demonstrates that the beta-binomial accurately models the variation in event numbers observed in the real data.

For comparison, the right panel of Figure 4 shows the histogram of normalized errors produced using the binomial model for the variation in event numbers. In this case the normalized errors are over-dispersed compared to the standard normal distribution, having a standard deviation of 1.80. This confirms that the binomial model underestimates the true variation in attrition numbers observed in CAF NCM occupations.

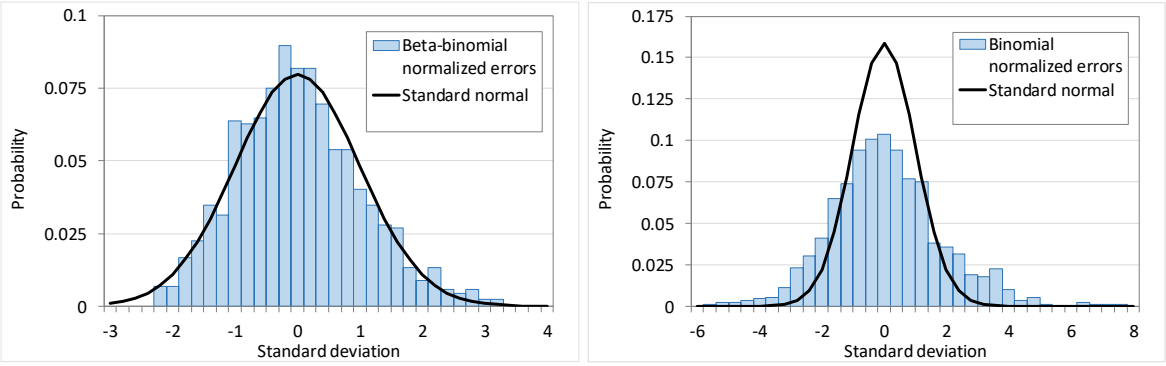

Figure 4. Distribution of normalized errors for the predicted attrition among NCM occupations for the beta-binomial model (left) and binomial model (right).

For the purposes of DE modelling, the beta-binomial model can be implemented by randomly sampling the incidence rate in a given interval from a beta distribution having mean $\hat{\gamma}$ and weight ω. Then each entity can be subjected to a random decision with a probability equal to the sampled incidence rate. This process is repeated in each interval.

Returning the to the example at the start of this section, if we apply the beta-binomial model to the prediction of the occupation's attrition numbers, shown in Figure 5, we observe that the variation in predicted numbers now resembles the variation observed historically. These results demonstrate that the output of stochastic simulations based on the beta-binomial model are informative of the real probability (and thus risk) of deviations from the mean number of predicted events.

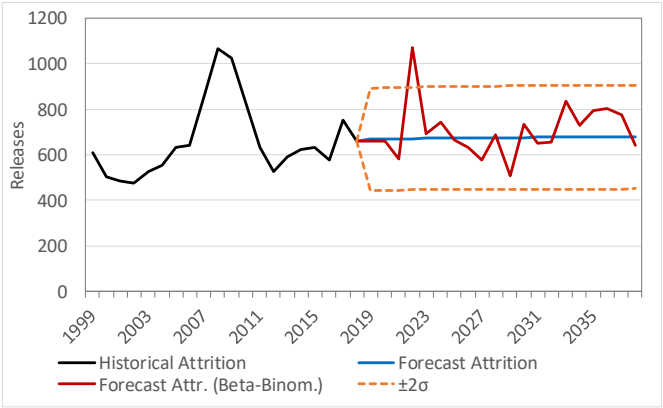

Figure 5. Historical and forecast releases for a CAF occupation (beta-binomial model).

6 CONCLUSION

Incidence rates are a simple and widely used concept in military workforce analysis, yet they are complex under the surface. Practitioners must pay careful attention to the exact meaning of the incidence rate, how it is used, how certain the rate estimate is, and how the rate varies over time. This paper provided a detailed development of techniques that address these challenges. These techniques were devised and applied primarily to analyze attrition in the CAF, but they are applicable to other populations and other events. An important advantage of the methods described herein is their robustness in cases where data is limited. This commonly occurs when the population is small, events are rare or little historical data is available. A second important result is the application of the beta-binomial distribution as a model for the variation in events over time. The correct use of this model enables the output of stochastic simulations to produce meaningful

conclusions about the probability of deviations from the average number of events. We showed that the often-assumed binomial model for the variation in event numbers cannot be relied on to do this.

Significantly, the techniques presented in this paper are straightforward to apply. The equations for estimating incidence rates and predicting future events remain simple, but they provide a deeper and more accurate picture of what is happening in the population. Without much additional analytical effort, this will enable better decision-making and improved planning in the context of military personnel management.

REFERENCES

Boileau, M.L.A. 2012. "Workforce Modelling Tools Used by the Canadian Forces". In *Proceedings of the International Workshop on Applied Modelling and Simulation*, edited by A.G. Bruzzone, W. Buck, E. Cayirci, F. Longo, 18-23. Rende, Italy: DIME University of Genoa.

Dietz, P. 1966. "Pension Funds: Measuring Investment Performance". Graduate School of Business, Columbia University.

Henderson, J.A., R.M. Bryce. 2019. "Verification Methodology for Discrete Event Simulation Models of Personnel in the Canadian Armed Forces". In *Proceedings of the 2019 Winter Simulation Conference*, edited by N. Mustafee, K.-H.G. Bae, S. Lazarova-Molnar, M. Rabe, C. Szabo, P. Haas, Y-J. Son, 2479-2490. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc..

Kruschke, J.K. 2011. *Doing Bayesian Data Analysis*. Burlington, MA: Academic Press.

Ligget, R.E., J.F. Delwiche. 2005. "The Beta-Binomial Model: Variability in Overdispersion Across Methods and Over Time". *Journal of Sensory Studies*, 20(2005): 48-61.

Merck, J.W., K. Hall. 1971. "A Markovian Flow Model: The Analysis of Movement in Large Scale (Military) Personnel Systems". Report No. R-514-PR, Rand Corporation, Santa Monica, CA.

Nedelman, J., T. Wallenius. 1986. "Bernoulli Trials, Poisson Trials, Surprising Variances, and Jensen's Inequality". *The American Statistician* 40(4): 286-289.

Okazawa, S. 2007. "Measuring Attrition Rates and Forecasting Attrition Volume". Director General Military Personnel Research and Analysis Technical Memorandum CORA TM 2007-02, Defence Research and Development Canada, Ottawa, Canada.

Pike, C., A. Novak, B. Moran, D. Kirszenblat, B. Hill. 2018. "A Stochastic Programming Approach to Optimal Recruitment in Australian Naval Aviation Training". In *Proceedings of the 2018 Winter Simulation Conference*, edited by M. Rabe, A.A. Juan, N. Mustafee, A. Skoogh, S. Jain, and B. Johansson, 3753-3764. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc..

Séguin, R. 2015. "PARSim, a Simulation Model of the Royal Canadian Air Force (RCAF) Pilot Occupation". In *Proceedings of the 2015 International Conference on Operations Research and Enterprise Systems*, Jan 10-12, Lisbon, Portugal, 51-62.

Škulj, D., V. Vehovar, D. Štamfelj. 2008. "The Modelling of Manpower by Markov Chains – A Case Study of the Slovenian Armed Forces". *Informatica* 32(2008): 289–291.

Straver, M., S.A. Latchman, Major N. Tabbenor. 2013. "Estimating the Cost of a Proposed Change to Canadian Armed Forces Promotion Policy". In *Proceedings of the IASTED International Conference Modelling and Simulation (MS2013)*, July 17-19, Banff, Canada, 213-121.

Straver, M., S. Okazawa, A. Wind, P. Moorhead. 2009. "Training Pipeline Modelling Using the Production Management Tool". Director General Military Personnel Research and Analysis Technical Memorandum DGMPRA TM 2009-019, Defence Research and Development Canada, Ottawa, Canada.

Talafuse, T., C. Lance, E. Gilts. 2019. "A Simulation Approach to Address MQ-9 Flying Unit Manning Shortfalls". In *Proceedings of the 2019 Winter Simulation Conference*, edited by N. Mustafee, K.-H.G. Bae, S. Lazarova-Molnar, M. Rabe, C. Szabo, P. Haas, Y-J. Son, 2479-2490. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc..

Vincent, E., S. Okazawa. 2019. "Determining Equilibrium Staffing Flows in the Canadian Department of National Defence Public Servant Workforce". In *Proceedings of the 2019 International Conference on Operations Research and Enterprise Systems*, Feb 19-21, Prague, Czech Republic, 205-212.

Vincent, E., D. Calitoiu, R. Ueno. 2018. "Personnel Attrition Rate Reporting". Director General Military Personnel Research and Analysis Scientific Report DRDC-RDDC-2018-R238, Defence Research and Development Canada, Ottawa, Canada.

Zegers, A., S. Isbrandt. 2010. "The arena career modelling environment: a new workforce modelling tool for the Canadian forces". In *Proceedings of the 2010 Summer Computer Simulation Conference*, July 2010, Ottawa, Canada, 94-101.

AUTHOR BIOGRAPHIES

STEPHEN OKAZAWA is a defence scientist with the Centre for Operational Research and Analysis in Defence Research and Development Canada in Ottawa, Canada. He holds a masters degree in Electro-Mechanical Engineering from the University of British Columbia, Canada. His research focuses on modelling and simulation of military personnel systems, especially using discrete event simulation methods. His email address is stephen.okazawa@forces.gc.ca.