EQUITABLE RETIREMENT INCOME TONTINES: MIXING COHORTS WITHOUT DISCRIMINATING

M.A. MILEVSKY AND T.S. SALISBURY

Abstract. There is growing interest in the design of annuities that insure against idiosyncratic longevity risk while pooling and sharing systematic risk; for example Piggott, Valdez and Detzel (2005) or Donnelly, Guillen and Nielsen (2014). In this paper we generalize the natural retirement income tontine introduced by Milevsky and Salisbury (2015), by combining heterogeneous cohorts into one pool. We engineer this scheme by allocating tontine shares at either a premium or a discount to par based on both the age of the investor and the amount they invest. For example, a 55 year-old allocating $10,000 to the tontine might be told to pay $200 per share and receive 50 shares, while a 75 year-old allocating $8,000 might pay $40 per share and receive 200 shares. They would all be mixed together into the same tontine pool and each tontine share would have equal income rights. On a historical note, this echoes a proposal by Charles Compton (1833) almost two centuries years ago, which hasn’t received much attention in the literature. The current paper addresses existence and uniqueness issues and discusses the conditions under which Compton’s scheme can be constructed equitably – which is distinct from fairly – even though it isn’t optimal for any cohort. As such, this also gives us the opportunity to differentiate between arrangements that are socially equitable, vs. actuarially fair vs. economically optimal.

Date: August 31, 2015.

Milevsky is an Associate Professor of Finance at the Schulich School of Business, York University, and Executive Director of the IFID Centre. Salisbury is a Professor in the Department of Mathematics and Statistics at York University. The authors acknowledge funding from the IFID Centre (Milevsky) and from NSERC (Salisbury) as well as helpful comments from seminar participants at University of New South Wales, Macquarie University, the CFA Society of Sydney and in particular Anthony Asher, Geoff Kingston, Sachi Purcal, John Piggott and Michael Sherris. The contact author (Milevsky) can be reached at: milevsky@yorku.ca.
1. Introduction

The tontine annuity – which was first promoted as a retirement income vehicle by Lorenzo di Tonti in the year 1653 – hasn’t benefited from the best publicity over the last three and a half centuries. Although at first tontines were used by British and French governments to finance their wars (against each other), conventional fixed interest bonds ended up superseding them as the preferred method of deficit financing. Private sector insurance companies in the 18th century offered tontine-like products, but they too were superseded by more familiar guaranteed life annuities and pensions. In fact, by the early 19th century regulators in the U.S. and the U.K. banned (a derivative product called) tontine insurance, although there is some debate over whether the ban actually applies to the tontines envisioned by Tonti. Legalities aside, in the words of the well-known financial writer Edward Chancellor, tontines are “one of the most discredited financial instruments in history.” We refer the interested reader to the book by Milevsky (2015) in which a slice of the tontine’s colorful history is addressed. In this article our focus (and contribution) is actuarial as opposed to political or historical.

In its purest financial form, a tontine annuity can be viewed as a perpetual (i.e. infinite maturity) bond that is purchased from an issuer by a group of investors who agree to share periodic coupons, only amongst survivors. As investors die and leave the tontine pool, the coupons or cash flows earned by those who avoid death increase (super) exponentially over time. And, in theory, the last remaining survivor receives all of the coupons until he or she finally dies and the issuer’s obligations to make payments are terminated. One can alternatively think of the tontine annuity as consisting of a portfolio of zero coupon bonds (ZCBs) with staggered maturities or face values, in which the final ZCB matures at the maximum possible lifespan of the investors in the pool, e.g. age 125. Cash flows from maturing ZCBs are distributed equally among survivors. From this perspective, the cash-flow pattern can be fined-tuned to any desired profile as long as its present value is equal to the amount invested by the group. The tontine pool retains longevity risk in the sense that if people live longer than expected and assumed at time zero, their payments are reduced relative to what they might have expected at time zero. Like other pooled schemes, such as those described in Piggott et. al. (2005) or Donnelly et. al. (2014), with a retirement income tontine there is no entity guaranteeing fixed payments for life, thus eliminating capital requirements. We use the term retirement income tontine to differentiate our scheme from a “winner take all” bet in which the payoff is deferred to the last survivor and to remind the reader of the pension-like structure.
1.1. **Problems with Tontines.** Over the past few centuries there has been quite a bit of popular and scholarly criticism leveled against tontine schemes. These writings have had no small part in accelerating their extinction. Overall, we believe that these concerns can be placed into three broad categories.

The first concern is that tontines themselves are amoral because one immediately benefits from someone else’s death. A more refined version of this concern is that they create an incentive for fraud, murder and other criminal activity. These critics contend that as the size of the tontine pool shrinks, the surviving members are incentivized to kill each other to gain a bigger share of the tontine pool. This colorful perception of the tontine permeates literary fiction and is the subject of many novels, but has no basis in reality. Most historical tontine schemes capped or limited payouts once a small fraction (say 5 to 10 members) remained in the pool. Despite all the novels and movies, there simply is no documented evidence that the last few survivors of a national tontine ever murdered each other. In fact, with hundreds of people in the tontine pool the economic benefit from nefariously reducing the pool size is minimal. The ethical concern that investors would benefit directly from others demise can be dismissed outright within the actuarial community, since that is the foundation of all pension and annuity pricing. As for the concern with fraud, this indeed was a problem in the 17th and 18th century, when documenting life and death was unreliable, but can also be dismissed in the 21st century with modern record-keeping systems.

The second concern with tontine annuities relates to economic optimality and cash flow patterns. In Lorenzo’s tontine scheme the cash paid to the group remained relatively constant (i.e. the numerator) but the number of survivors (i.e. the denominator) declined more-than exponentially fast over time, resulting in a rapidly increasing payout to surviving members. This explosive profile of income is at odds with the prevailing economic philosophy of consumption smoothing. It is assumed that older retired investors would want a stable or perhaps even declining real cash-flow over time, notwithstanding concerns about health-care expenses and inflation for the elderly. Either way, according to these critics, the tontine annuity isn’t an optimal contract.

This concern can in fact be remedied with proper product design. There is absolutely no reason why the cash-flows to the group should be structured to remain constant over time. As mentioned above, the ZCBs could be purchased so that payments to the pool decline (faster than) exponentially over time at the same rate as (expected) mortality. In fact, this is the essence of the proposal made in Milevsky and Salisbury (2015) under the term of *natural retirement income tontine*. To sum up, we believe the second objection around the
issue of economic optimality can also be overcome with 21st century financial markets. This argument was recently made by Professor William Sharpe in a presentation to the French Finance Association, quoted in Milevsky (2015).

The third concern is a more-subtle one and has to do with the pooling of cohorts and the age profile of the tontine. That is the core focus on this paper. Indeed, if one allows anyone, regardless of age, to participate equally in the same tontine annuity pool – which we do not – there would be an immediate transfer of wealth from the members who are expected to die early (i.e. the old) to those who are expected to live the longest (i.e. the young). Historical tontines – such as the one first issued by the English government in the late 17th century – therefore discriminated against the old in favor of the young. Thus, for example, in the earliest English tontine schemes the nominees on whose life the tontine was contingent ranged in age from a few months to over the age of 50. In equilibrium everyone should nominate the healthiest possible age (females age 10, approximately), but that leads to design problems when the nominee and annuitant are not the same person. If indeed one requires homogenous mortality pools to run a non-discriminating tontine scheme, then this limits the possible size of the pool and the efficacy of large number diversification. The 18th century tontine schemes in which investors were placed into small tailored classes – with different payouts based on age – suffered from reduced pooling and risk diversification. A modern day pension fund trying to implement a tontine payout structure with a predictable cash flow would face the same concern unless it had a very large pool of willing retirees – with identical ages and health profiles.

In some sense, we believe this is the most serious criticism against resurrecting retirement income tontines in the context of modern pension schemes. One would require hundreds of people of exactly the same age retiring on exactly the same day, to reduce the variability of payouts. In fact, this is why authors such as Piggot, Valdez and Detzel (2005) or Donnelly, Guillen and Nielsen (2014) have proposed more complex pooled annuitization schemes or overlays that allow for mixing of different cohorts and possibly over multiple generations. Note that we will not wade into the debate over which among the many pooling schemes is ideal. In reality there will be a tradeoff – a scheme that squeezes out the highest possible utility for the group may also be more complex to analyze and harder to explain.

In sum, we choose to analyze tontines, which offer a design that provides “good” utility, while remaining sufficiently simple that we can establish a range of qualitative results. We understand and acknowledge that other designs have their own appeal and role.
1.2. Making heterogeneous tontines equitable. As stated above, it is this third criticism leveled against (retirement income) tontines that we address in the current paper. Our remedy to this concern is to allow cohorts of different ages (and mortality) to mix in the same pool by allocating different participation rates or shares based on their age at the time of purchase. For example, a 55 year-old allocating $10,000 to the tontine might be told to pay $200 per share and receive 50 shares, while a 75 year-old allocating $8,000 might pay $40 per share and receive 200 shares. They would all be mixed together into the same tontine pool and each tontine share would have equal income rights.

On a historical note, this proposal to “fix” tontines was actually made almost 200 years ago by Mr. Charles Compton (1833), who was the Accountant General of the Royal Mail in the UK. In a document that hasn’t received much attention by the actuarial literature, he wrote:

“...His Majesty’s Government should create a Tontine Stock bearing interest at a certain rate per annum and to permit persons of a certain age to purchase such stock at par, those younger or older to purchase the same stock above or below par according to their several ages.”

Compton (1833) argued that this was preferable to forming the contributors into classes which ends-up creating very small groups with few benefits from pooling. He argued that:

“...Younger purchasers would give more money for £100 stock than the elder and would give up part of their income for the benefit of elder members, who in turn would bequeath their annuities to the younger as compensation.”

Compton (1833) goes on to list a table of values mapping ages into share prices, which is quite similar (in spirit) to our numbers, which we present later on in the paper.

Now, it might seem rather trivial (actuarially) to allocate shares in the tontine based on the age of the investor and the size of their investment. After all, with an immediate annuity, $1 of lifetime will cost $45 for a 45-year old and $75 for a 75 year-old. The relative prices of mortality-contingent claims are well understood in the actuarial literature. But what may not be obvious is that in fact there are situations (i.e. counterexamples) in which this cannot be done in a fair (or even equitable) manner, especially when the groups are small. In other words, there are cases in which no mapping or share price will allow groups to be mixed without discrimination, and our objective is to understand when this is (or is not) possible. This wasn’t addressed by Compton (1833) – who assumed that a proper price always existed in theory – and is something that emerges naturally from our analysis. The
need for large pools to diversify risk is intimately linked to the issues address in this paper, and is a question and that has recently been highlighted by Donnelly (2015) as well. We return to this later.

As far as terminology is concerned, in this paper we are careful to distinguish between a scheme that is fair and a scheme that is equitable, which is a somewhat weaker requirement. A retirement income tontine scheme in which there is a possibility of everyone in the pool dying before the maximum age and thus leaving left-overs can never be made fair, in the sense of Donnelly (2015), regardless of the cash flow payout structure itself. We will show rigorously that the expected present value of income will always be less than the amount contributed or invested into the tontine. However, a heterogeneous tontine scheme can often (though not always) be made equitable by ensuring that the present value of income is the same for all participants in the scheme, regardless of age. This scheme will not discriminate against any one cohort, although it won’t be fair in the strictest sense. All of this will be addressed in detail, including an analysis of scenarios in which equity is impossible to achieve.

To recap then, in this paper we investigate how to construct a tontine scheme including many ages – and determine the proper share prices to charge participants – so that it is equitable and doesn’t discriminate against any age or any group. The tontine we propose is a closed pool that does not allow anyone to enter (or obviously exit) after the initial set-up. We could in theory use a similar weighting scheme to add or introduce new members over time – and have explored the mathematical properties of an open scheme – but have not included them in this paper. Again, this is one further place we differ from the designs of Piggott et. al. (2005), Donnelly et. al. (2015), or (in the tontine context) Sabin (2010). That is, we advocate closing the group to newcomers, but allow multiple ages and contribution levels within the closed pool. We refrain from arguing that this is better or worse than any other design. That said, our model requires and assumes little (if any) actuarial discretion as time evolves. The rules are set at time zero and the cash-distribution algorithm is crystal clear. We believe this design has merit.

1.3. Outline. The remainder of this paper is organized as follows. In Section 2 we review what is meant by a natural retirement income tontine as introduced by Milevsky and Salisbury (2015). The latter discussed the relation with economic optimality, all within a single cohort pool. Indeed, there are many possible payout functions $d(t)$ that one can use to construct a tontine scheme – historically the $d(t)$ curve was constant, for example 8% per year – and the optimal function depends (at a minimum) on the representative investor’s
coefficient of risk aversion. The natural payout function is defined as the \( d(t) \) that optimizes logarithmic utility preferences. Section 3 moves from a review of single cohorts to the introduction of multiple cohorts, which is the contribution of this particular paper. It describes in precise terms what is meant by an equitable share price for all participants in a tontine scheme, given a particular payout function \( d(t) \). In this section we also address the notion of fairness and how it differs from being equitable. Section 4 returns to the matter of economic optimality. It is typically impossible to locate a payout function \( d(t) \) that is optimal for all cohorts, as was possible in the case of a single cohort, even if all participants have the same level of risk aversion. Indeed, the best that one can hope for when mixing cohorts is an equitable scheme and not a uniformly optimal one. Finally, in Section 5 we suggest that a good selection from among all the possible equitable schemes is one in which the payout function \( d(t) \) is set equal to the expected number of shares outstanding at any point in the future. Alas, we can’t yet prove uniqueness for this scheme and leave this as a conjecture. We do however discuss welfare gains and losses from the scheme and provide some numerical examples. Section 6 gives conclusions and offers some suggestions and avenues for further research. Proofs appear in the appendix (Section 7) and the tables and figures are at the end of the paper.

2. Annuities vs. Optimal Tontine Payout Functions

In this section we briefly review the scheme proposed by Milevsky and Salisbury (2015) and what is meant by the term optimal tontine scheme. We assume an objective survival function \( t_p_x \), for an individual aged \( x \) to survive \( t \) years. One purpose of the tontine structure is to insulate the issuer from the risk of a stochastic (or simply mis-specified) survival function, but in this paper we assume \( t_p_x \) is given, and applies to all individuals. We intend to address the stochastic case in a subsequent paper. We assume that the tontine pays out continuously, as opposed to quarterly or monthly. For simplicity, we assume a constant risk-free interest rate \( r \), though it would be simple to incorporate a term structure since what is really needed to construct the tontine is a portfolio of ZCBs.

2.1. Optimal annuities. The basic comparator for a tontine is an annuity involving annuitants each paying \$1 to the insurer initially, and receiving in return an income stream of \( c(t) \; dt \) for life. The constraint on these annuities is that they are fairly priced, in other words that with a sufficiently large client base, the initial payments invested at the risk-free rate will fund the called-for payments in perpetuity. This implies a constraint on the annuity
payout function $c(t)$, namely that
\begin{equation}
\int_0^\infty e^{-rt} tp_x c(t) \, dt = 1.
\end{equation}
Again, $c(t)$ is the payout rate per survivor. The payout rate per initial dollar invested is $tp_x c(t)$. Letting $U(c)$ denote the instantaneous utility of consumption, a rational annuitant (with lifetime $\zeta$) having no bequest motive will choose a life annuity payout function for which $c(t)$ maximizes the discounted lifetime utility:
\begin{equation*}
E\left[ \int_0^\zeta e^{-rt} U(c(t)) \, dt \right] = \int_0^\infty e^{-rt} tp_x U(c(t)) \, dt
\end{equation*}
where $r$ is (also) the subjective discount rate (SDR), all subject to the constraint (1). Provided $u$ is strictly concave, the following now follows from the Euler-Lagrange theorem.

**Theorem 1** (Theorem 1 of Milevsky and Salisbury (2015)). *Optimized life annuities have constant $c(t) \equiv c_0$, where*
\begin{equation*}
c_0 = \left[ \int_0^\infty e^{-rt} tp_x \, dt \right]^{-1}.
\end{equation*}

This result can be traced back to Yaari (1965) who showed that the optimal (retirement) consumption profile is constant (flat) and that 100% of wealth is annuitized when there is no bequest motive, subjective discount rates are equal to interest rates and complete annuity markets (actuarial notes) are available.

2.2. Optimal Tontine Payouts. In practice, insurance companies funding the life annuity $c(t)$ are exposed to both systematic longevity risk (due to randomness or uncertainty in $tp_x$), model risk (the risk that $tp_x$ is mis-specified), as well as re-investment or interest rate risk (which is the uncertainty in $r$ over long horizons). The latter is not our focus here, so we will continue to assume that $r$ is a given constant for most of what follows.

This brings us to the tontine structure introduced in Milevsky and Salisbury (2015), in which a predetermined dollar amount is shared among survivors at every $t$. Let $d(t)$ be the rate at which funds are paid out per initial dollar invested, a.k.a. the tontine payout function. There is no reason for tontine payout function to be a constant fixed percentage of the initial dollar invested (e.g. 4% or 7%), as it was historically. Getting back to the issue of optimality, we can pose the same question as considered above for annuities: what $d(t)$ is best for subscribers? The comparison is now between $d(t)$ and $tp_x c(t)$, where $c(t)$ is the optimal annuity payout found above.
Suppose there are initially \( n \) subscribers to the tontine scheme, each depositing a dollar with the tontine sponsor. Let \( N(t) \) be the random number of live subscribers at time \( t \). Consider one of these subscribers. Given that this individual is alive, \( N(t) - 1 \sim \text{Bin}(n - 1, p_x) \). In other words, the number of other (live) subscribers at any time \( t \) is binomially distributed with probability parameter \( t p_x \).

As in the Yaari (1965) model, which we extend to tontines, this individual’s discounted lifetime utility is

\[
E\left[ \int_0^\infty e^{-rt} u\left( \frac{nd(t)}{N(t)} \right) dt \right] = \int_0^\infty e^{-rt} t p_x E\left[ u\left( \frac{nd(t)}{N(t)} \right) \mid \zeta > t \right] dt
\]

\[
= \int_0^\infty e^{-rt} t p_x \sum_{k=0}^{n-1} \binom{n-1}{k} t^k p_x(1-t p_x)^{n-1-k} u\left( \frac{nd(t)}{k+1} \right) dt.
\]

The constraint on the tontine payout function \( d(t) \) is that the initial deposit of \( n \) should be sufficient to sustain withdrawals in perpetuity. Of course, at some point all subscribers will have died, so in fact the tontine sponsor will eventually be able to cease making payments, leaving a small remainder or windfall. This gets to the issue of fairness, which we revisit in the next section. But this time is not predetermined, so we treat that profit as an unavoidable feature of the tontine. Remember that we do not want to expose the sponsor to any longevity risk. It is the pool that bears this risk entirely.

Our budget or pricing constraint is therefore that

\[
\int_0^\infty e^{-rt} d(t) dt = 1.
\]

So, for example, if \( d(t) = d_0 \) is forced to be constant, i.e. a flat tontine as was typical historically, then the tontine payout function (rate) is simply \( d_0 = r \) (or somewhat more if a cap on permissible ages is imposed, replacing the upper bound of integration in (2) by a value less than infinity).

The optimal \( d(t) \) is in fact far from constant. Milevsky and Salisbury (2015) find this optimum in some generality. The following summarizes the conclusion in the case of CRRA utility \( U(c) = c^{1-\gamma} \) for \( \gamma > 0, \gamma \neq 1 \) (or \( U(c) = \log c \) for \( \gamma = 1 \)). Set

\[
\beta_{n,\gamma}(p) = p \sum_{k=0}^{n-1} \binom{n-1}{k} p^k (1-p)^{n-1-k} \left( \frac{n}{k+1} \right)^{1-\gamma}.
\]

**Theorem 2** (Corollary 2 of Milevsky and Salisbury (2015)). The optimal retirement income tontine structure has \( d(t) = d(0) \beta_{n,\gamma}(t p_x)^{1/\gamma} \), where \( d(0) \) is chosen to make (2) hold.
Additional analysis, discussion and numerical examples of the natural retirement income tontine in the context of only age group, is available from Milevsky and Salisbury (2015). In particular, for logarithmic utility ($\gamma = 1$) we have $\beta_n(p) = p$. The optimal payout in this case has $d(t) = c_0 \cdot \delta_x$, where $c_0$ is as in Theorem 1. Following the above paper, we call this design the natural tontine for the age-$x$ cohort. Figure 1 displays one such payout function $d(t)$ for a particular age, mortality assumption and interest rate.

One of the main conclusions from Milevsky and Salisbury (2015) is that annuities yield higher utility than tontines, but that the utility gains are actually quite marginal, once a reasonable number of individuals participate in the tontine. Another observation is that the utility differential between the optimal tontine and the natural tontine (which is only optimal when $\gamma = 1$) are also marginal. Therefore that paper advocates the natural tontine as a good alternative to a standard annuity, when the latter is priced fairly. If capital risk charges are imposed upon an annuity, in the form of a loading factor that protects against systematic mortality risk, the tontine can easily have higher utility to the consumer than the annuity. We refer the reader to the above paper for details. Note that all of the above is based on one assumed (baseline) age $x$. There is no room for older or younger people. Figure #1 (at the end of the paper) provides a graphical illustration of the payout function.

3. Mixing cohorts: Equitable share prices when $d(t)$ is given

Now suppose a retirement income tontine pays out $d(t)$ per initial dollar invested, which may (or may not) be optimal for people of a given age. In other words, an inhomogeneous group of individuals subscribe to purchase shares in the tontine. Each subscriber will be entitled to a share of the total funds disbursed in proportion to the number of shares owned, with the sole caveat that the subscriber must be alive at the time of disbursement. Once the list of subscribers is known, together with the dollar value they will invest, they are each quoted a price per share, depending on their age and the size of their investment (and more generally, the ages and investments of all subscribers). Of course, the price then determines the number of shares they will receive in return for their announced investment. The issue we wish to address is how to assign prices in an equitable manner, which we will take to mean that the expected present value of funds received by the various subscribers (per initial dollar invested) are all equal. The mathematical question becomes whether there exists a collection of share prices that realize this, and whether such prices are unique.
Let \( n \) be the number of subscribers. For computational purposes it will sometimes be convenient to group them into \( K \) homogeneous cohorts (i.e., with the same age and contribution level), though this is not actually a restriction since we could choose to take \( K = n \) and deem each cohort to consist of a single individual. We will use notation that permits grouping, but in many proofs will without loss of generality take cohorts to consist of single individuals. For \( i = 1, \ldots, K \) let \( x_i \) be the initial age of individuals in the \( i \)th cohort, and let \( w_i \) be the number of dollars each of them invests. Let \( n_i \) be the size of the \( i \)th cohort, so \( n = \sum n_i \) and the total initial investment is \( w = \sum n_i w_i \). Therefore the total time-\( t \) payouts occur at rate \( wd(t) \). For notational convenience, we choose to base prices on participation rates \( \pi_i \), in other words, \( 1/\pi_i \) is the \( i \)th subscriber’s price per share. Let \( u_i = \pi_i w_i \) be the number of shares purchased by each individual in the \( i \)th cohort, and let \( u = \sum n_i u_i \) be the total number of shares purchased. Let \( N_i(t) \) be the number in the \( i \)th cohort who survive to time \( t \). The \( d(t) \) function satisfies the budget constraint \( \int_0^\infty e^{-rt}d(t) \, dt = 1 \), where \( r \) is the interest rate. An individual in the \( i \)th cohort who survives to time \( t \) will receive payments at a rate of

\[
  u_i \times \frac{wd(t)}{\sum_j u_j N_j(t)} = wd(t) \frac{\pi_i w_i}{\sum_j \pi_j w_j N_j(t)}.
\]

Summing this over all subscribers of course gives back a total payout rate of \( wd(t) \), as long as at least one subscriber survives. In other words, \( \sum_i wd(t)N_i(t) \frac{\pi_i w_i}{\sum_j \pi_j w_j N_j(t)} = wd(t) \), as long as \( \sum_j \pi_j w_j N_j(t) > 0 \).

Let \( \zeta \) be the lifetime of an individual in the \( i \)th cohort. The present value of their payments is

\[
  E \left[ \int_0^\zeta e^{-rt} u_i \frac{wd(t)}{\sum_j u_j N_j(t)} \, dt \right] = \int_0^\infty e^{-rt} i p_x \pi_i w_i E \left[ \frac{wd(t)}{\sum_j \pi_j w_j N_j(t)} \mid \zeta > 0 \right] \, dt
\]

\[
  = \int_0^\infty e^{-rt} i p_x wd(t) E_i \left[ \frac{\pi_i w_i}{\sum_j \pi_j w_j N_j(t)} \right] \, dt.
\]

We use the notation \( E_i \) to remind us that this is a conditional expectation, in which \( N_i - 1 \sim \text{Bin}(n_i - 1, i p_x) \), while the other \( N_j \sim \text{Bin}(n_j, i p_x) \). Call the above expression \( w_i F_i(\pi_1, \ldots, \pi_K) \), so if \( \pi = (\pi_1, \ldots, \pi_K) \) then

\[
  F_i(\pi) = \int_0^\infty e^{-rt} i p_x wd(t) E_i \left[ \frac{\pi_i}{\sum_j \pi_j w_j N_j(t)} \right] \, dt
\]

represents the present value of the returns per dollar invested by the \( i \)th cohort.

Subscribers in the \( i \)th cohort invest \( w_i \), so ideally, “fairness” would mean that the present value of each person’s payments equals their initial fee, in other words, that \( F_i(\pi) = 1 \) for
each $i$. This is not possible, for the simple reason that there is always a positive probability of money being left on the table once everyone dies. Let $A_{i,k}(t)$ be the event that the $k$th individual in the $i$th cohort survives till time $t$. Then $\sum n_i w_i = w$ but

$$\sum n_i w_i F_i(\pi) = \int_0^\infty e^{-rt} d(t) \sum_i n_i \cdot i px_i E_i \left[ \frac{\pi_i w_i}{\sum_j \pi_j w_j N_j (t)} \right] dt$$

$$= \int_0^\infty e^{-rt} d(t) E \left[ \sum_i \sum_{k=1}^{n_i} \frac{\pi_i w_i}{\sum_j \pi_j w_j N_j (t)} 1_{A_{i,k}(t)} \right] dt$$

$$= \int_0^\infty e^{-rt} d(t) E \left[ \sum_i \frac{\pi_i w_i N_i (t)}{\sum_j \pi_j w_j N_j (t)} 1_{\{N_i (t) > 0\}} \right] dt$$

$$= \int_0^\infty e^{-rt} d(t) P \left( \sum_j N_j (t) > 0 \right) dt < \int_0^\infty e^{-rt} d(t) dt = w.$$  

In other words, we have proved that

**Lemma 3.** Regardless of $\pi$, at least one cohort must have its present value $F_i(\pi) < 1$.  

The closest we can come to being truly fair is to have all the $F_i(\pi)$ equal. In other words, each subscriber loses the same tiny percentage of their investment, in present value terms. We say that $\pi$ is *equitable* if

$$F_i(\pi) = F_j(\pi) \forall i, j.$$  

Equivalently,

$$F_i(\pi) = 1 - \epsilon$$ for each $i$, where $\epsilon = \int_0^\infty e^{-rt} d(t) P \left( \sum_j N_j (t) = 0 \right) dt.$$

If we want to make the tontine fair in the absolute sense, we’d need to return any monies remaining after the last death to the estates of the subscribers (as a whole, or simply to the estate of the last survivor). In practice, the remaining funds will be miniscule, once $n$ is large, so we prefer to work with the notion as defined above. Nevertheless, this is precisely why Donnelly et. al. (2013, 2014) operate schemes and overlays that include a death benefit, which then ensures no left-overs (using our terminology) and make the scheme fair. We eliminate the death benefit, but keep things *equitable.*

Since $\sum \frac{n_i w_i}{w} F_i(\pi) = 1 - \epsilon$ by the above argument, it is clear that either the tontine is equitable, or there are some indices for which $F_i(\pi) > 1 - \epsilon$ and some for which $F_i(\pi) < 1 - \epsilon$. We call

$$\theta(\pi) = \max_{i \neq j} |F_i(\pi) - F_j(\pi)|$$

the *inequity* of the tontine. We say that $\pi$ is *more equitable* than $\pi'$ if $\theta(\pi) < \theta(\pi').
There is an obstruction to equity, as the following example shows. Suppose \( n = K = 2 \).
The first subscriber will receive all the available income during the period they outlive the second subscriber. Therefore if \( \frac{w}{w_1} \) is sufficiently large,
\[
F_1(\pi) > \frac{w}{w_1} \int_0^\infty e^{-rt}d(t)p_{x_1}tq_{x_2} dt > \int_0^\infty e^{-rt}d(t)[1 - tq_{x_1}tq_{x_2}] dt = 1 - \epsilon.
\]
For example, it using reasonable ages and mortality rates it is impossible to make equitable a tontine in which one subscriber invests one dollar, and another invests a million. The most equitable such a tontine could be is in the limiting case \( \pi_1 = 0 \), so that the first subscriber only starts receiving payments once the second subscriber has died. We will address such contingent tontines in the appendix (Section 7).

The main theorem of this paper is as follows.

**Theorem 4.** Fix \( d(t) \) as well as the \( n_i, x_i \) and \( w_i, i = 1, \ldots, K \).

(a) If there exists an equitable choice of \( \pi = (\pi_1, \ldots, \pi_K) \) such that \( 0 < \pi_i < \infty \) for each \( i \), then this choice is unique up to an arbitrary multiplicative constant.

(b) A necessary and sufficient condition for such a \( \pi \) to exist is the following:

\[
\int_0^\infty e^{-rt}d(t)\prod_{i \notin A} tq_{x_i}^{n_i}(1 - \prod_{i \in A} tq_{x_i}) dt < \alpha_A(1 - \epsilon)
\]

for every \( A \subset \{1, \ldots, K\} \) with \( 0 < |A| < K \), where \( \alpha_A = \frac{1}{w} \sum_{k \in A} n_kw_k \).

We will prove this result in the appendix, where we also expand on the meaning of condition (3). In heuristic terms: if there is a cohort who find the tontine favourable even if they have to wait for income until all subscribers from other cohorts have died, then equity is impossible.

Note that our formulation tacitly assumed that all members of a cohort share the same participation rate. If equitable rates exist, then this must in fact be the case. To see this, subdivide cohort \( i \) into \( n_i \) cohorts, each with a single member, and apply the uniqueness conclusion of the above theorem.

In Section 5 we will examine some plausible scenarios, with utility included. Here we treat some fairly extreme examples, to illustrate how equitable rates may vary, as well as giving some circumstances in which they will fail to exist. We exhibit values in the two-cohort case \( (K = 2) \), using Gompertz hazard rates, ie \( \lambda_x = \frac{1}{b}e^{\frac{x-m}{b}} \) at age \( x \). Parameters are \( m = 88.72 \), \( b = 10 \), and \( r = 4\% \). In Table 1 (table at the end of the paper) we look at age disparities, and in Table 2 we look at disparities in investment levels.
These tables show the spread in $\pi$ (ratio of the largest to the smallest) narrowing as the population size increases. This is not a general rule however. If in Table 1 we had taken $x_1 = 90$ and $x_2 = 65$ the spread would narrow at first but then widen. With $x_1 = 65$, $x_2 = 85$, and $n_1 = 1$ there would be equitable rates for small values of $n_2$ but not for large ones.

Note that equity being infeasible is not purely a phenomenon of small populations. A poorly designed tontine can also produce this effect. For example, suppose we have two cohorts of size $n_1 = n_2 = 100$ with ages $x_1 = 65$ and $x_2$ to be specified. If each member of the second cohort contributes $w_2 = 1$, then for a large enough value of $w_1$ the tontine must be inequitable. With a well designed tontine it typically takes a large value of $w_1$ to destroy equity. But if we take a flatter tontine than is desirable – say a tontine whose $d(t)$ would be natural for a population of age 50 – then quite modest values of $w_2$ will produce inequity, especially once there is a disparity in cohort ages. For example, if $x_2 = 80$ (resp. 75/70/65) then even $w_2 = 7$ (resp. 14/37/209) will accomplish this, according to the criterion of Theorem 4. What this all means practically speaking is that Compton’s (1833) scheme to charge different share prices for tontine stock might not work for all ages and investment amounts. The equitable price is most definitely not linear in the amount invested, which is in contrast to a tontine scheme with homogenous ages. And, while one certainly can’t fault Compton (1833) for not realizing this fact, we believe its an interesting aspect of his rather-clever proposal.

4. Utility, asymptotics, and optimality

4.1. Utility and loading factors. Back to the general case involving any weighting function $\pi$, whether it is equitable or not, we now consider the utility of the cash flow received by an individual from the $i$th cohort, namely

$$\int_0^\infty e^{-rt} \mathbb{E}_i \left[ U \left( \sum_{i,j} w_j \pi_j \mathbb{E}_j \right) \right] dt.$$

We are interested in the effect of inhomogeneity in the subscriber population. In particular, we would like to understand whether adding individuals to a tontine raises or lowers utility (and by how much), when the added individuals are no longer homogeneous (in homogeneous populations, adding individuals always increases utility). In particular, for a cohort of size $n_i$ in a heterogeneous tontine with payout $d(t)$, the natural comparison involves contrasting their utility with that from an optimized tontine $\hat{d}(t)$ in which only those $n_i$ homogeneous individuals participate. Thus we define a loading factor $\delta_i$, which (when applied to the
homogeneous tontine) makes the two utilities equal. In other words,
\[
\int_0^\infty e^{-rt}tp_{x_i}E_i\left[u\left(\frac{wd(t)}{\sum w_j\pi_j(t)N_j(t)\pi_jw_j}\right)\right]dt = \int_0^\infty e^{-rt}tp_{x_i}E_i\left[u\left(\frac{n_i\hat{d}(t)}{N_i(t)}(1-\delta_i)w_i\right)\right]dt.
\]
If \(\delta_i > 0\) this means that the cohort loses utility from the addition of heterogeneous individuals to the pool. If \(\delta_i < 0\) then the cohort gains utility from the addition of these individuals. For a different comparison, between tontines and annuities, see Milevsky and Salisbury (2015), although \(\delta\) is used to denote a different loading. See also the related work by Hanewald et. al. (2015) which examines how product loadings might affect the choice between different mortality-contingent claims.

In Section 5, we will give numerical calculations of loadings for various choices of \(d(t)\), and we will see that in a well-designed tontine, adding participants increases utility (ie loadings are negative). We will work with \(\gamma = 1\) so \(U(c) = \log c\), and the above formula simplifies considerably. By results of Milevsky and Salisbury (2015), the optimal \(d(t)\) is the tontine that is natural for the age-\(x_i\) cohort, in other words, \(\hat{d}(t) = c_0 \cdot tp_{x_i}\).

4.2. Asymptotics and the proportional tontine. We fix \(K\), the \(x_i\), and the \(w_i\) and consider the limit of the \(\pi_i\) when the total number of subscribers \(n = \sum n_i \to \infty\). Let \(\alpha_i > 0\) and \(\sum_{i=1}^K \alpha_i = 1\). Assume that the \(n_i \to \infty\) in such a way that \(\frac{n_iw_i}{w} \to \alpha_i\), so \(\alpha_i\) represents the fraction of the initial investment attributable to the \(i\)th cohort. Then
\[
F_i(\pi) = \int_0^\infty e^{-rt}tp_{x_i}wd(t)E_i\left[\frac{\pi_i}{\sum \pi_jw_jN_j(t)}\right]dt \to \int_0^\infty e^{-rt}d(t)\frac{\pi_i \cdot tp_{x_i}}{\sum_{j=1}^K \pi_j \alpha_j \cdot tp_{x_j}} dt.
\]
A particular case of this bears particular attention. Let \(a_x = \int_0^\infty e^{-rt}tp_x dt\) be the standard annuity price of $1 for life for age \(x\) individuals. Consider \(d(t) = \sum_{j} \alpha_j tp_{x_j}\) (which clearly satisfies the condition that \(\int_0^\infty e^{-rt}d(t) = 1\)). In this case, \(F_i(\pi) \to a_{x_i} \pi_i\), so the equitable participation rates asymptotically become \(\pi_i = \frac{1}{a_{x_i}}\). We call a tontine with this \(d(t)\) and these \(\pi_i\) a proportional tontine, and emphasize that it is equitable only in the limit as \(n \to \infty\). In the case of a homogeneous population (ie. \(K = 1\)), the proportional tontine agrees with what we have earlier called the natural tontine for this cohort.

One motivation for this particular design is that the payout rate to a surviving individual from the \(i\)th group, at time \(t\), is asymptotically
\[
\frac{d(t)}{\sum \pi_j \alpha_j \cdot tp_{x_j}} \pi_i = \frac{1}{a_{x_i}}
\]
per unit. In other words, the rate of payment to a surviving individual remains constant in time, and is simply the standard annuity factor of \(\frac{1}{a_{x_i}}\) per dollar of initial premium. In this
sense, a proportional tontine reproduces in the limit, the payment structure and cost of a
standard fixed annuity for each subscriber. We will shortly see a further motivation, when
we show that it is asymptotically optimal.

How do our utility loadings behave when \( n \to \infty \) as above? The above equation becomes that
\[
\int_0^\infty e^{-rt}tp_x u \left( \sum \pi_j \alpha_j \cdot \overline{t}p_{x_j} \pi_i w_i \right) dt = \int_0^\infty e^{-rt}tp_x u \left( \frac{d(t)}{\overline{t}p_{x_i}} (1 - \delta_i) w_i \right) dt.
\]
Take \( d(t) \) to be the proportional tontine, so in the limit, so \( \pi_j = \frac{1}{a_{x_j}} \) is equitable in the limit.
As above, take \( u \) to be logarithmic, and \( \hat{d}(t) \) to be natural for the age-\( x_i \) cohort. We obtain that
\[
\int_0^\infty e^{-rt}tp_x u \left( \frac{w_i}{a_{x_i}} \right) dt = \int_0^\infty e^{-rt}tp_x u \left( \frac{w_i}{a_{x_i}} (1 - \delta_i) \right) dt,
\]
from which the following immediately follows:

**Lemma 5.** Asymptotically, the proportional tontine has utility loadings \( \delta_i = 0 \).

### 4.3. Can a tontine be optimal for multiple cohorts?

A natural question is whether it is possible to design a tontine to be optimal for multiple age cohorts. This turns out not to be possible, except in the limit as \( n \to \infty \). To formulate the question, we include equity as an additional set of constraints in the optimization problem. In particular, we wish to choose \( d(t) \) and the \( \pi_j \) to maximize the utility of the \( i \)th cohort
\[
\int_0^\infty e^{-rt}tp_x E_i \left[ U \left( \frac{wd(t)}{\sum_j \pi_j w_j N_j(t)} \pi_i w_i \right) \right] dt
\]
over \( d(t) \geq 0 \), subject to the budget constraint \( \int_0^\infty e^{-rt}d(t) dt = 1 \) and the equity constraints
\[
\int_0^\infty e^{-rt}tp_x w d(t) E_i \left[ \frac{\pi_i}{\sum_j K \pi_j w_j N_j(t)} \right] dt = \int_0^\infty e^{-rt}tp_x \frac{\pi_i}{\sum_j K \pi_j w_j N_j(t)} dt
\]
for \( \ell \neq i \).

In the limit as \( n \to \infty \) we wish to maximize
\[
\int_0^\infty e^{-rt}tp_x U \left( \frac{d(t)}{\sum \alpha_j \pi_j \cdot \overline{t}p_{x_j} \pi_i w_i} \right) dt
\]
over \( d(t) \geq 0 \), subject to the budget constraint \( \int_0^\infty e^{-rt}d(t) dt = 1 \) and the equity constraints
\[
\int_0^\infty e^{-rt}tp_x d(t) \left( \frac{\pi_i}{\sum_j K \alpha_j \pi_j \cdot \overline{t}p_{x_j}} \right) dt = \int_0^\infty e^{-rt}tp_x d(t) \left( \frac{\pi_i}{\sum_j K \alpha_j \pi_j \cdot \overline{t}p_{x_j}} \right) dt
\]
for \( \ell \neq i \). This version of the problem simplifies if reformulated in terms of \( \Gamma(t) = \sum_{j=1}^{\infty} a_j \pi_j t p_x_j \). Now we seek to maximize \( \int_0^\infty e^{-rt} t p_x U(\pi_i w_i \Gamma(t)) \, dt \) over \( \Gamma(t) \geq 0 \), subject to the budget constraint \( \int_0^\infty e^{-rt} \Gamma(t) \sum_j \alpha_j \pi_j t p_x_j \, dt = 1 \) and the equity constraints
\[
\int_0^\infty e^{-rt} \Gamma(t) \pi_i t p_x_i \, dt = \int_0^\infty e^{-rt} \Gamma(t) \pi_\ell t p_x_\ell \, dt \quad \text{for } \ell \neq i.
\]

The equity constraints become merely that
\( \pi_\ell = \pi_i \frac{\int_0^\infty e^{-rt} \Gamma(t) t p_x_i \, dt}{\int_0^\infty e^{-rt} \Gamma(t) t p_x_\ell \, dt} \).

and substituting back, the budget constraint becomes that
\( \pi_i \frac{\int_0^\infty e^{-rt} \Gamma(t) t p_x_i \, dt}{\int_0^\infty e^{-rt} \Gamma(t) \pi_\ell t p_x_\ell \, dt} = 1 \). This puts us back in the context of optimizing the simple annuity of Theorem 1, which implies that the optimal \( \Gamma(t) \) is constant. If we normalize so \( \pi_i = \frac{1}{a_x} \) then the \( \pi_\ell = \frac{1}{a_x} \), and we get \( \Gamma(t) = 1 \).

In particular, optimizing the utility of the \( i \)th cohort, in the presence of equity constraints, asymptotically gives precisely the proportional tontine described in the last section, ie \( d(t) = \sum_j \frac{\alpha_j}{a_x} t p_x_j \). Therefore this optimal tontine (in this case, really a type of annuity) has the same design, regardless of which \( i \) one chooses to optimize for. We have shown that

**Proposition 6.** Assume a strictly concave utility function. In the limit as \( n \to \infty \), the proportional tontine optimizes the utility of each cohort simultaneously.

The original optimization problem (i.e. in the setting of finite \( n \)) can also be solved, though not so cleanly. We do not present this here, except to note that when we optimize even the logarithmic utility of the \( i \)th cohort, the results turn out to no longer be consistent when we vary \( i \). In other words, it is typically impossible to make everyone happy simultaneously.

This is one reason we feel it is reasonable to fix a tontine structure \( d(t) \) (as we have done above), and allow people participate at equitable rates if they so wish. Naturally, this means one of the questions we will need to answer is how significant their utility loss is, when doing so.

5. Our Suggested \( d(t) \): The Natural and Equitable Tontine

In the context of a homogeneous population of age \( x \), all investing equal amounts, the design proposed in Milevsky and Salisbury (2015) had \( d(t) = \frac{1}{a_x} t p_x \), ie. the natural tontine for age \( x \). In this context, the design is optimal in the case of logarithmic utility, and near-optimal otherwise. In this section, we wish to propose a suitable generalization in the heterogeneous setting.

For heterogeneous tontines, we have seen that overall optimality is not feasible (except asymptotically). In that context, we propose adopting the following design, which performs
well in numerical experiments we have conducted, reduces to the above design in the case of a homogeneous population, and agrees with the proportional tontine in the limit as \( n \to \infty \) (so is optimal asymptotically).

Fix the \( x_i, w_i, \) and \( n_i \). We say that a tontine is natural if \( d(t) \) is at all times proportional to the mean number of surviving tontine shares. In other words, \( d(t) = c \sum u_j n_j \cdot t p_{x_j} = c \sum \pi_j w_j n_j \cdot t p_{x_j} \). Integrating, we see that

\[
d(t) = \sum_i \left[ \frac{\pi_i n_i w_i}{\sum_j a_{x_j} \pi_j w_j n_j} \right] t p_{x_i}.
\]

Note that once the \( \pi_i \) are given, the natural tontine is fully determined by the budget constraint. But to construct a tontine that is both natural and equitable, we must compute the \( \pi_i \) and \( d(t) \) simultaneously. In practice this is more complicated than (as above) simply fixing a \( d(t) \) and computing equitable \( \pi \)'s, but not unduly so (at least when the number \( K \) of types is small).

The following (at the end of the paper) two tables (Table 3 for \( K = 2 \) cohorts, Table 4 for \( K = 3 \)) display such natural and equitable tontines, and compare them to “natural” tontines that would have been chosen if the population had been homogeneous (but with equitable participation rates). We also compare with the corresponding proportional tontines, though those are not equitable. Though the theoretical basis of proportional tontines is not as appealing as that of natural ones, they are simpler to compute, and they do appear to perform reasonably in practice. We view them as an acceptable alternative if computational resources are not available to work out equitable \( \pi \)'s and natural \( d(t) \)'s. Note that since these tables normalize \( \pi \) to make \( \pi_i = 1 \) for some \( i \), this means that the proportional tontine has \( \pi_j = a_{x_i} / a_{x_j} \).

First consider Table 3. Rows labelled “A” and “D” use tontine designs that would be natural for homogeneous populations, of age 65 and 75 respectively. Equitable \( \pi \)'s are then computed. In row A, both \( \delta_1 \) and \( \delta_2 \) start negative \((n_1 = 1 = n_2)\), meaning that the benefit of the extra participant outweighs the impact of heterogeneity. As the common value of \( n_1 = n_2 \) rises, the \( \delta_i \) become positive, as defects in the design become more relevant. In particular, loadings remain strictly positive asymptotically – while the product becomes essentially an annuity, it is not an optimal one. Note that the loadings are not actually monotone. Adding participants is more beneficial for older (age 75) participants than for younger (age 65) ones. Surprisingly, in the presence of age 65 participants, even the age 75 ones get more benefit from an age-65 design than an age-75 one.
Rows labelled “B” correspond to a truly natural and equitable design. Rows labelled “C” are proportional designs. In most cases, either performs better for both cohorts than the homogeneous designs do. The problem with the A design is now clear – to be equitable it requires a higher participation rate $\pi_2$, which dilutes the benefit of adding individuals to the tontine, and produces utility loss. In contrast, rows B and C typically show a negative loading (ie a utility gain), though with different choices of parameters (not shown) this can in fact sometimes not be the case.

Comparing B and C, it is generally the case that the older cohort prefers a natural design, whereas the younger cohort prefers a proportional design. The proportional design comes closer to equalizing the utility gains between the cohorts. The two designs perform similarly asymptotically. The factor contributing most to their difference is the equitability of $\pi$ rather than the choice of $d(t)$ – using the proportional $d(t)$ but equitable $\pi$’s would turn out to give very similar utilities to the fully natural design.

Table 4 treats the three-cohort case, comparing the natural and proportional designs with a design that would be natural for the age-65 cohort alone. Now all three designs have very similar effects on utility. Otherwise the table is consistent with empirical observations made above: adding people to the tontine is generally favourable (despite heterogeneity); and the utility improvement is greater for the older participants. Note that the good performance of these designs may in part be a consequence of a balance between ages 60 and 70 – asymmetric designs (not shown) are less consistent.

Natural and equitable tontines appear to exist for a broad range of parameter values (though not universally – the obstruction raised in Section 3 still remains valid). But we have not yet succeeded in finding necessary and sufficient conditions for existence, or in establishing uniqueness, as in Theorem 4. Therefore resolving the following remains a topic for further research.

**Conjecture 7.** Fix the $x_i$, $n_i$, and $w_i$, $i = 1, \ldots, K$. Under broad conditions there will exist a choice of $\pi = (\pi_1, \ldots, \pi_K)$ such that the corresponding natural tontine is equitable. Up to an arbitrary multiplicative constant, there is at most one such $\pi$.

6. Conclusion

There is a growing interest among practitioners and academics in the optimal design of retirement de-accumulation products that insure against idiosyncratic longevity risk while sharing aggregate exposure within a group. In this paper we investigated the design of a retirement income tontine scheme that allows individuals with different mortality rates to
participate in the same pool. And while this scheme might not be actuarially fair, in the sense of Donnelly (2015), this scheme is equitable in that the scheme does not discriminate against any particular sub-group and all participants receive the exact same expected present value of benefits. And, although pooled annuity funds (PAF) and group self annuitization (GSA) schemes introduced by Piggot, Valdez and Detzel (2005) or Donnelly, Guillen and Nielsen (2014) can provide more utility than a tontine, the retirement income tontine has the advantage of transparency, simplicity and requiring little if any actuarial discretion. It pays a reasonably steady and predictable cash flow to a declining group of survivors. It might be simpler to analyze qualitatively, but leads to interesting mathematical properties and insights.

The structure we introduced in this paper – which is an extension of Milevsky and Salisbury (2015) – allows anyone of any age to participate in the scheme by adjusting the price of a tontine share to be a function of (i.) the number of investors, (ii.) their ages, and (iii.) the capital they have invested. In Lorenzo Tonti's original scheme – as well as the structure proposed in Milevsky and Salisbury (2015), all investors were assumed to be of the same age and paid the same price. When smaller groups were segmented into age bands they lost the benefit of large numbers. In this paper we proved that its possible to mix cohorts provided the diversity of the pool satisfies certain dispersion conditions, and we propose a specific design that appears to work well in practice. Our approach echoes a proposal made by Charles Compton (1833), which is why one might label this the Compton Tontine, when optimal weights actually exists. Future research will examine the impact of stochastic (or systematic) mortality on the design of such a scheme, which is yet another factor that Charles Compton can’t be faulted for ignoring.
REFERENCES

[1] Cannon, E. and I. Tonks (2008), *Annuity Markets*, Oxford University Press, UK.

[2] Compton, C. (1833), *A Treatise on Tontine: In Which The Evils of the Old System Are Exhibited and an Equitable Plan Suggested for Rendering the Valuable Principle of Tontine More Beneficially Applicable to Life Annuities*, Printed by J. Powell, Hand Court, Ipper Thames Street, London.

[3] Donnelly, C., M. Guillen and J.P. Nielsen (2013), Exchanging Mortality for a Cost, *Insurance: Mathematics and Economics*, Vol. 52(1), pg. 65-76.

[4] Donnelly, C., M. Guillén and J.P. Nielsen (2014), Bringing Cost Transparency to the Life Annuity Market, *Insurance: Mathematics and Economics*, Vol. 56 (1), pg. 14-27.

[5] Donnelly, C. (2015) Actuarial Fairness and Solidarity in Pooled Annuity Funds, *ASTIN Bulletin*, Vol. 45, pg. 49-74

[6] Doob, J.L. (1984), *Classical Potential Theory and its Probabilistic Counterpart*, Springer Verlag, New York.

[7] Hanewald, K, J. Piggott and M. Sherris (2013), Individual post-retirement longevity risk management under systematic mortality risk, *Insurance: Mathematics and Economics*, Vol. 52(1), pg. 87-97.

[8] Jennings, R. M. and A. P. Trout (1982), *The Tontine: From the Reign of Louis XIV to the French Revolutionary Era*, S.S. Huebner Foundation for Insurance Education, University of Pennsylvania, 91 pages.

[9] Milevsky, M.A. (2015), *King William’s Tontine: Why the Retirement Annuity of the Future Should Resemble Its Past*, Cambridge University Press, New York City.

[10] Milevsky, M.A. and T.S. Salisbury (2015), Optimal Retirement Income Tontines, *Insurance: Mathematics and Economics*, Vol. 64, pg. 91-105

[11] Piggott, J., E. A. Valdez, and B. Detzel (2005), The Simple Analytics of a Pooled Annuity Fund, *The Journal of Risk and Insurance*, Vol. 72(3), pg. 497-520.

[12] Qiao, C and M. Sherris (2013), Managing Systematic Mortality Risk with Group Self-Pooling and Annuitization Schemes, *Journal of Risk and Insurance*, Vol. 80(4), pg. 949-974.

[13] Sabin, M.J. (2010), Fair Tontine Annuity, *SSRN abstract #1579932*

[14] Stamos, M. Z. (2008), Optimal Consumption and Portfolio Choice for Pooled Annuity Funds, *Insurance: Mathematics and Economics*, Vol. 43 (1), pg. 56-68.

[15] Tonti, L. (1654), *Edict of the King for the Creation of the Society of the Royal Tontine*, Translated by V. Gasseau-Dryer, *History of Actuarial Science*, Volume V, Edited by S. Haberman and T.A. Sibbett, published by William Pickering, London, 1995

[16] Valdez, E. A., J. Piggott, and L. Wang (2006), Demand and Adverse Selection in a Pooled Annuity Fund, *Insurance: Mathematics and Economics*, Vol. 39(2), pg. 251-266.

[17] Yaari, M. (1965) Uncertain lifetime, life insurance and the theory of the consumer, *Review of Economic Studies*, Vol. 32(2), pg. 137-150.
Figure 1. The payout function $d(t)$ can be visualized as a time-dependent yield which declines in proportion to the assumed survival probability for the group at time zero.
Equitable rate $\pi_2$ with a single outlier with respect to age

| $n_1$ | 1 | 2 | 3 | 4 | 5 | 7 | 10 |
|-------|---|---|---|---|---|---|----|
| 65    |   |   |   |   |   |   |    |
| 90    | NA| 23.180 | 7.219 | 5.286 | 4.536 | 3.892 | 3.512 |
| 50    |   |   |   |   |   |   |    |
| 95    | NA| NA | NA | NA | 35.455 | 13.492 | 9.199 |

$\pi_2$ when there are $n_1$ individuals in the cohort of age $x_1$, and a single outlier ($n_2 = 1$) with age $x_2$. Tontine is natural for the age $x_1$ cohort, and each individual invests $\$1$ ($w_1 = w_2 = 1$). Normalized so $\pi_1 = 1$. NA indicates that equity is infeasible.

Table 1. Shows the equitable participation rate $\pi_2$ when there are $n_1$ individuals in the cohort of age $x_1$, and a single outlier ($n_2 = 1$) with age $x_2$. Tontine is natural for the age $x_1$ cohort, and each individual invests $\$1$ ($w_1 = w_2 = 1$). Normalized so $\pi_1 = 1$. NA indicates that equity is infeasible.
Equitable rate $\pi_2$ with a single outlier with respect to $\$ invested

| $w_2$ | 1   | 2   | 5   | 20  | 100  | 1000 | $\infty$ |
|-------|-----|-----|-----|-----|------|------|---------|
| $n_1$ | 2   | 1.283 | 1.169 | 1.077 | 1.022 | 1.005 | 1.001 | 1       |
|       | 5   | 3.086 | 1.799 | 1.287 | 1.075 | 1.017 | 1.002 | 1       |
|       | 10  | NA   | 9.129 | 1.781 | 1.158 | 1.036 | 1.004 | 1       |
|       | 20  | NA   | NA   | 6.552 | 1.342 | 1.069 | 1.009 | 1       |

Minimal $n_1$ at which equity is feasible

| $w_2$ | 20  | 50  | 100 | 250 | 500  | 1000 |         |
|-------|-----|-----|-----|-----|------|------|---------|
| $n_1$ | 5   | 12  | 23  | 57  | 114  | 229  |         |

Assumes Gompertz Mortality ($m = 88.72, b = 10$) and $r = 4\%$.

Table 2. Shows the equitable participation rate $\pi_2$ when there are $n_1$ individuals who invest 1 dollar ($w_1 = 1$) and a single outlier who invests $w_2$ dollars. Shows the minimal $n_1$ at which equity can be achieved, for various $w_2$. All subscribers are the same age ($x_1 = x_2 = 65$), and the tontine is natural for that age. Normalized so $\pi_1 = 1$. NA indicates that equity is infeasible.
**Table 3.** Shows the participation rates $\pi_2$ (= inverse of the share price) and corresponding utility loadings $\delta_1, \delta_2$ when there are two cohorts of subscribers in the pool: ages $x_1 = 65$ and $x_2 = 75$. Utility is logarithmic. Everyone invests 1 dollar ($w_1 = w_2 = 1$). Tontine designs are:

A = Natural tontine based on age 65 cohort alone, equitable rates;
B = Natural tontine based on the range of ages, equitable rates;
C = Proportional tontine;
D = Natural tontine based on age 75 cohort alone, equitable rates;

Assumes $r = 4\%$, Gompertz Mortality ($m = 88.72, b = 10$);
$\delta_i$ are given in b.p.; Rates are normalized so $\pi_2 = 1$
## Participation rates and utility loadings in pools with three cohorts: $2n_1 = n_2 = 2n_3$

|       | age 60 | age 65 | age 70 | age 60 | age 65 | age 70 |
|-------|--------|--------|--------|--------|--------|--------|
|       | $\pi_1$ | $\pi_2$ | $\pi_3$ | $\delta_1$ | $\delta_2$ | $\delta_3$ |
| $n_1 = 5$ | $n_2 = 10$ | $n_3 = 5$ | $n_1 = 5$ | $n_2 = 10$ | $n_3 = 5$ |
| A     | 0.886  | 1      | 1.161  | -186.9 | -136.1 | -594.3 |
| B     | 0.884  | 1      | 1.161  | -216.0 | -136.6 | -586.8 |
| C     | 0.889  | 1      | 1.153  | -275.0 | -138.7 | -586.8 |
| $n_1 = 10$ | $n_2 = 20$ | $n_3 = 10$ | $n_1 = 10$ | $n_2 = 20$ | $n_3 = 10$ |
| A     | 0.889  | 1      | 1.157  | -79.4  | -68.9  | -301.0 |
| B     | 0.887  | 1      | 1.157  | -102.9 | -70.4  | -297.2 |
| C     | 0.889  | 1      | 1.153  | -133.3 | -71.3  | -264.5 |
| $n_1 = 20$ | $n_2 = 40$ | $n_3 = 20$ | $n_1 = 20$ | $n_2 = 40$ | $n_3 = 20$ |
| A     | 0.890  | 1      | 1.155  | -29.8  | -20.8  | -153.3 |
| B     | 0.888  | 1      | 1.155  | -49.7  | -23.0  | -151.8 |
| C     | 0.889  | 1      | 1.153  | -65.4  | -23.4  | -135.1 |

Assumes $r = 4\%$ and Gompertz Mortality $(m = 88.72, b = 10)$; $\delta_i$ are given in b.p.; Rates are normalized so $\pi_2 = 1$

|       |       |       |       |       |       |       |
|-------|-------|-------|-------|-------|-------|-------|

Table 4. Shows the participation rates $\pi_1, \pi_2, \pi_3$ (which are the inverse of the share prices) and corresponding utility loadings $\delta_1, \delta_2, \delta_3$ when there are three cohorts of subscribers in the pool: ages $x_1 = 60$, $x_2 = 65$, and $x_3 = 70$. Utility is logarithmic. Everyone invests 1 dollar ($w_1 = w_2 = w_3 = 1$). Tontine designs are:

A = Natural tontine based on age 65 cohort alone, equitable rates;

B = Natural tontine based on the range of ages, equitable rates;

C = Proportional tontine.
We will assume, for now, that each cohort consists of a single individual, i.e. \( n = K \), each \( n_i = 1 \), and the \( N_i(t) \) are Bernoulli. There will be no loss of generality in doing so, as far as the proof of uniqueness goes. To see this, simply split the cohorts up. For existence however, an additional argument will then be needed to establish the sufficiency of (3). In fact, a short proof of necessity could be extracted from the proof of Lemma 11 below (see formula (7)). The more complex structure we develop below is needed principally for sufficiency.

Let \( P = \{(\pi_1, \ldots, \pi_n) : 0 < \pi_i \leq 1 \text{ for each } i, \text{ and } \max_i \pi_i = 1 \} \). Scaling \( \pi \) does not affect the \( F_i \), so every value of \( F = (F_1, \ldots, F_n) \) can be realized with some \( \pi \in P \). We start with the uniqueness question.

Proof of (a) of Theorem 4. Suppose \( \pi \neq \tilde{\pi} \) are both equitable, and not multiples of each other. Interpolate between them using \( \pi(s) = s\pi + (1 - s)\tilde{\pi} \). Then

\[
\frac{d}{ds} F_i(\pi(s)) = \int_0^\infty e^{-rt}p_x wd(t) E_i \left[ \frac{d}{ds} \sum_j \pi_j(s) w_j N_j(t) \right] dt
\]

(4)

Moreover

\[
\pi'_i(s) \pi_j(s) - \pi_i(s) \pi'_j(s) = (\pi_i - \tilde{\pi}_i) [s(\pi_j - \tilde{\pi}_j) + \tilde{\pi}_j] - [s(\pi_i - \tilde{\pi}_i) + \tilde{\pi}_i] (\pi_j - \tilde{\pi}_j)
\]

\[
= (\pi_i - \tilde{\pi}_i) \tilde{\pi}_j - (\pi_j - \tilde{\pi}_j) \tilde{\pi}_i = \pi_i \tilde{\pi}_j - \pi_j \tilde{\pi}_i = \pi_i \pi_j \left( \frac{\tilde{\pi}_j}{\pi_j} - \frac{\tilde{\pi}_i}{\pi_i} \right)
\]

for each \( s \in [0, 1] \). In particular, if we choose \( i \) to minimize \( \frac{\tilde{\pi}_j}{\pi_j} \), it follows that this expression is \( \geq 0 \) for every \( j \), and \( > 0 \) for some \( j \) (since \( \pi \) is not a multiple of \( \tilde{\pi} \)). Therefore for this choice of \( i \) we have that \( \frac{d}{ds} F_i(\pi(s)) > 0 \). Thus \( F_i(\tilde{\pi}) > F_i(\pi) \), which is a contradiction. It follows that uniqueness holds.

To prove existence, and in particular to understand (3), we need to dig deeper into the structure of the optimal \( \pi \). In particular, to consider limiting cases when some of the \( \pi_i \to 0 \). To do that, we borrow an idea from the theory of Martin boundaries (see for example Doob (1984)) and embed \( P \) in a compact set \( P_0 \).

Set \( \eta = \{1, \ldots, n\} \). For non-empty \( A \subseteq \eta \) and \( \pi \in P \), let \( \pi_A = (\frac{\pi_i}{\max_{j \in A} \pi_j})_{i \in A} \). Set \( g(\pi) = (\pi_A)_{\emptyset \neq A \subseteq \eta} \), so \( g : P \to [0, 1]^m \), where \( m = \sum_{\emptyset \neq A \subseteq \eta} |A| = \sum_{k=1}^n k^n = \sum_{j=0}^{n-1} n^{n-1} = n2^{n-1} \).

Let \( P_0 \) be the closure of \( g(P) \) in \([0, 1]^m \), so \( g \) is a continuous embedding of \( P \) in \( P_0 \). We will think of \( P \) as a subset of \( P_0 \), so to some extent we will use notation that identifies \( \pi \in P \).
with \( g(\pi) \in \mathcal{P}_0 \). In particular, we will freely use \( \pi \) to denote either an element of \( \mathcal{P} \) or an element of \( \mathcal{P}_0 \), and in both cases will use the same notation \( \pi_A \) for its components, as defined above.

Let \( \pi \in \mathcal{P}_0 \). Then \( \pi_\eta \in [0,1]^n \) may have some (but not all) of its components = 0. Writing \( \pi_\eta = (\pi_{\eta,1}, \ldots, \pi_{\eta,n}) \), we let \( A_0 = \{ i \mid \pi_{\eta,i} \neq 0 \} \). If \( A_0 \neq \eta \), then \( \pi_{\eta \setminus A_0} \) may in turn have some (but not all) of its components = 0. Let \( A_1 = \{ i \in \eta \setminus A_0 \mid \pi_{\eta \setminus A_0,i} \neq 0 \} \). Continuing this way with \( \pi_{\eta \setminus A_0 \cup A_1} \) etc., we partition \( \eta \) into a finite number of non-empty subsets \( A_0, A_1, \ldots, A_J \) such that the components of each \( \pi_{A_j} \) are all non-zero. Of course, if \( \pi \) actually \( \in \mathcal{P} \) then \( A_0 = \eta \). In fact, \( \pi \) may be recovered from these \( A_j \) and \( \pi_{A_j} \), as the following Lemma shows.

**Lemma 8.** Let \( A \subset \eta \) be nonempty, and let \( \pi \in \mathcal{P}_0 \). Define the \( A_j \) as above, and let \( i = \min \{ j \mid A \cap A_j \neq \emptyset \} \). Then for \( k \in A \),

\[
\pi_{A,k} = \begin{cases} 
\frac{\pi_{A,k}}{\max_{j \in A \cap A_i} \pi_{A_j,j}}, & k \in A \cap A_i \\
0, & k \in A \setminus A_i.
\end{cases}
\]

**Proof.** To see this, choose \( \pi^{(m)} \in \mathcal{P} \) such that \( g(\pi^{(m)}) \to \pi \). Set \( B = \cup_{j \geq i} A_j \), so \( A \subset B \). For \( k \in B \), we have

\[
\pi_{B,k} = \lim_{m \to \infty} \pi^{(m)}_{B,k} = \lim_{m \to \infty} \frac{\pi^{(m)}_k}{\max_{j \in B} \pi^{(m)}_j}.
\]

By definition of \( \pi_{B,k} \), this is non-zero precisely for \( k \in A_i \), so if \( j \in A_i \) and \( k \in B \setminus A_i \) then \( \frac{\pi^{(m)}_k}{\pi^{(m)}_j} \to 0 \). Therefore \( \max_{j \in A} \pi^{(m)}_j = \max_{j \in A \cap A_i} \pi^{(m)}_j \) for sufficiently large \( m \), and if \( k \in A \)

\[
\pi_{A,k} = \lim_{m \to \infty} \pi^{(m)}_{A,k} = \lim_{m \to \infty} \frac{\pi^{(m)}_k}{\max_{j \in A} \pi^{(m)}_j} = \lim_{m \to \infty} \frac{\pi^{(m)}_k}{\max_{j \in A \cap A_i} \pi^{(m)}_j}.
\]

By the above, this = 0 if \( k \in A \setminus A_i \). If \( k \in A \cap A_i \) we may divide numerator and denominator by \( \max_{j \in A_i} \pi^{(m)}_j \) to see that it

\[
\pi_{A,k} = \lim_{m \to \infty} \frac{\pi^{(m)}_{A,k}}{\max_{j \in A \cap A_i} \pi^{(m)}_{A_i,j}} = \frac{\pi_{A_i,k}}{\max_{j \in A \cap A_i} \pi_{A_i,j}},
\]

as required. \( \square \)

What is going on in this argument is that for \( \pi \in \mathcal{P}_0 \) and \( A_0, A_1, \ldots, A_J \) as above, having a sequence \( \pi(n) \in \mathcal{P} \) converge to \( \pi \) in the topology of \( \mathcal{P}_0 \) means that the \( \pi^{(m)}_j \) converge to non-zero values for \( j \in A_0 \), they converge to 0 at a common rate for \( j \in A_1 \) (with \( \pi_{A_i} \) giving a suitably renormalized limit), they converge to 0 at a faster rate for \( j \in A_2 \), etc.
For a given payout function \( d(t) \) we defined a tontine above, corresponding to any \( \pi \in \mathcal{P} \). We can generalize this to any \( \pi \in \mathcal{P}_0 \). It pays only to individuals in \( A_0 \), as long as any of the them survive, using participation rates \( \pi_i \). As soon as the last of these individuals dies, it starts paying out to individuals in \( A_1 \), using participation rates \( \pi_{A_1,i} \). Once they all die, it starts paying out to individuals in \( A_2 \), using rates \( \pi_{A_2,i} \), etc. Since payments are contingent on the extinction of an earlier group, we call this generalization a contingent tontine.

If there is a contingent tontine that is favourable to the last group to start collecting, it is quite plausible that no \( \pi \) can achieve equity. The content of Theorem 4 is that these two statements are in fact equivalent. Moreover, we shall see in the course of the proof that equation (3) precisely captures the failure of the first statement.

We may now generalize the definition of the present value functions \( F_i(\pi) \). Let \( \pi \in \mathcal{P}_0 \), and suppose that \( i \in A_k \). Let \( T_0 = 0 \), and for \( 1 \leq j \leq J + 1 \) let \( T_k \) be the time the last survivor from \( A_0 \cup \cdots \cup A_{k-1} \) dies. Let \( \zeta_i \) be the lifetime of individual \( i \). Define

\[
F_i(\pi) = \int_0^\infty e^{-rt}d(t)E\left[ \frac{\pi_{A_k,i}}{\sum_{j \in A_k} \pi_{A_k,j} w_j N_j(t)} 1\{T_k < t < \zeta_i\} \right] dt.
\]

The point of passing to the more complicated index set \( \mathcal{P}_0 \) is the following:

**Lemma 9.** Each \( F_i : \mathcal{P}_0 \to \mathbb{R} \) is continuous.

**Proof.** Let \( \pi \in \mathcal{P}_0 \). Suppose that \( \pi^{(m)} \to \pi \). Assume to start with that each \( \pi^{(m)} \in \mathcal{P} \). Define \( A_0, \ldots, A_J \) as above (using \( \pi \)), and likewise \( T_0, \ldots, T_{J+1} \), and let \( i \in A_\ell \). Set \( B_k = A_k \cup \cdots \cup A_J \). Then

\[
F_i(\pi^{(m)}) = \int_0^\infty e^{-rt}d(t)E\left[ \frac{\pi^{(m)}_i}{\sum_{j \in \eta} \pi^{(m)}_j w_j N_j(t)} 1\{T_k < t < \zeta_i \} \right] dt
\]

\[
= \sum_k \int_0^\infty e^{-rt}d(t)E\left[ \frac{\pi^{(m)}_i}{\sum_{j \in B_k} \pi^{(m)}_j w_j N_j(t)} 1\{T_k < t < \zeta_i \} \right] dt
\]

\[
= \sum_{k \leq \ell} \int_0^\infty e^{-rt}d(t)E\left[ \frac{\pi^{(m)}_{i,k}}{\sum_{j \in B_k} \pi^{(m)}_{j,k} w_j N_j(t)} 1\{T_k < t < \zeta_i \} \right] dt
\]

where \( \pi^{(m)}_{q,k} = \frac{\pi^{(m)}_q}{\max_{j \in A_k} \pi^{(m)}_j} \). Now send \( m \to \infty \). If \( j \in A_k \) then \( \pi^{(m)}_{j,k} \to \pi_{A_k,j} \), while if \( j \in B_{k+1} \) then \( \pi^{(m)}_{j,k} \to 0 \). In particular, dominated convergence implies that the terms with \( k < \ell \) vanish in the limit, while the \( k = \ell \) term converges to to \( F_i(\pi) \).
The general case now follows. If \( \pi^{(m)} \to \pi \) but we no longer assume \( \pi^{(m)} \in \mathcal{P} \), simply choose \( \tilde{\pi}^{(m)} \in \mathcal{P} \) such that \( |F_i(\pi^{(m)}) - F_i(\tilde{\pi}^{(m)})| \to 0 \) and \( \|\pi^{(m)} - \tilde{\pi}^{(m)}\| \to 0 \). Then \( \tilde{\pi}^{(m)} \to \pi \) so \( \lim F_i(\pi^{(m)}) = \lim F_i(\tilde{\pi}^{(m)}) = F_i(\pi) \). \( \square \)

Define a very equitable participation rate to be any \( \pi \in \mathcal{P}_0 \) that minimizes \( \theta_0(\pi) = \max_{i,k} |F_i(\pi) - F_k(\pi)| \) over \( \pi \in \mathcal{P}_0 \). By Lemma 9 and compactness of \( \mathcal{P}_0 \), such a \( \pi \) exists. Of course, \( \pi \) is equitable if and only if \( \theta_0(\pi) = 0 \) and \( \pi \in \mathcal{P} \).

**Lemma 10.** Let \( \pi \) be very equitable. Let \( a_0 = \min_{i \in I} F_i(\pi) \), and \( a_J = \max_{i \in I} F_i(\pi) \). Then \( F_i(\pi) = a_0 \) for every \( i \in A_0 \), and \( F_i(\pi) = a_J \) for every \( i \in A_J \).

**Proof.** Let \( \pi \) be very equitable. As in the proof of uniqueness, we will perturb \( \pi \) to improve equity. Fix \( j \). Let \( \tilde{A} \) consist of the \( i \in A_j \) which minimize \( F_i(\pi) \) over \( A_j \), and set

\[
\pi_{A_j,i}(s) = \begin{cases} 
\pi_{A_j,i}(1 + s), & i \in \tilde{A} \\
\pi_{A_j,i}, & i \in A_j \setminus \tilde{A}.
\end{cases}
\]

We do not perturb \( \pi_{A_k,q} \) for any \( k \neq j \), so there is no impact on \( F_q \) for \( q \notin A_j \). Note that this \( \pi(s) \) may not lie in \( \mathcal{P}_0 \), as \( \max_{i \in A_j} \pi_{A_j,i}(s) \) may now be \( \neq 1 \). This will not turn out to matter, and could in any case be remedied by rescaling at a suitable point in the argument. A calculation as in (4) shows that for \( i \in A_j \),

\[
\frac{d}{ds} F_i(\pi(s)) = \int_0^\infty e^{-rt} t p_x w(t) E \left[ \sum_{k \in A_j} (\pi_{A_j,i}'(s) \pi_{A_j,k}(s) - \pi_{A_j,i}(s) \pi_{A_j,k}'(s)) w_k N_k(t) \right] \frac{1}{(\sum_{k \in A_j} \pi_{A_j,k}(s) w_k N_k(t))^2} dt.
\]

Assume that \( \tilde{A} \neq A_j \), i.e., that \( F_i(\pi) \) is not constant on \( A_j \). I claim that

\[
\frac{d}{ds} F_i(\pi(s)) = \begin{cases} 
> 0, & i \in \tilde{A} \\
< 0, & i \in A_j \setminus \tilde{A}.
\end{cases}
\]

In other words, this perturbation brings the lowest \( F_i \)'s up, and the other \( F_i \)'s down.

Suppose \( i \in \tilde{A} \). If \( k \in \tilde{A} \) then \( \pi_{A_j,i}'(s) \pi_{A_j,k}(s) - \pi_{A_j,i}(s) \pi_{A_j,k}'(s) = \pi_{A_j,i} \pi_{A_j,k} - \pi_{A_j,i} \pi_{A_j,k} = 0 \). If \( k \in A_j \setminus \tilde{A} \) then \( \pi_{A_j,i}'(s) \pi_{A_j,k}(s) - \pi_{A_j,i}(s) \pi_{A_j,k}'(s) = \pi_{A_j,i} \pi_{A_j,k} > 0 \). Therefore the sum is \( > 0 \) so \( \frac{d}{ds} F_i(\pi(s)) > 0 \) too.

Now suppose \( i \in A_j \setminus \tilde{A} \). If \( k \in \tilde{A} \) then \( \pi_{A_j,i}'(s) \pi_{A_j,k}(s) - \pi_{A_j,i}(s) \pi_{A_j,k}'(s) = -\pi_{A_j,i} \pi_{A_j,k} < 0 \). If \( k \notin \tilde{A} \) then \( \pi_{A_j,i}'(s) \pi_{A_j,k}(s) - \pi_{A_j,i}(s) \pi_{A_j,k}'(s) = 0 \). Therefore the sum is \( < 0 \) so \( \frac{d}{ds} F_i(\pi(s)) < 0 \) too.
The result of this calculation is that for small $s$, our perturbation reduces the variation of $F_i(\pi)$ on $A_j$, unless $i \rightarrow F_i(\pi)$ is already constant on $A_j$.

Turning to the statement of the lemma, there must be a $j$ such that $F_i(\pi) = a_0$ for every $i \in A_j$, otherwise we could perturb so as to raise every $F_i(\pi)$ which equals $a_0$, while lowering or not changing the other $F_k(\pi)$. (If necessary, rescale to keep $\pi \in \mathcal{P}_0$.) This would reduce $\theta_0(\pi)$ which is impossible. By the same perturbation, we may also assume that for any $j$, either $F_i(\pi) = a_0$ for every $i \in A_j$, or $F_i(\pi) > a_0$ for every $i \in A_j$. Let $J_0$ be the set of $j$ of the former type. Our goal is to show that $0 \not\in J_0$.

Suppose that $j \geq 1$ belongs to $J_0$, but $j - 1$ does not. Consider the following perturbation. Combine $A_{j-1}$ and $A_j$, by setting

$$
\pi_{A_{j-1} \cup A_j}(i)(s) = \begin{cases} 
\pi_{A_{j-1}}(i), & i \in A_{j-1} \\
\pi_{A_j}(i), & i \in A_j
\end{cases}
$$

for $s > 0$. This will not impact $F_i(\pi(s))$ for $i$ other than $j - 1$ or $j$. For $i \in A_j$ this has no effect on the expectation in (5) representing payments after time $T_j$, but with positive probability it adds a non-zero contribution from the integral over $[T_{j-1}, T_j)$. Therefore $F_i(\pi(s))$ increases for each $i \in A_j$. A derivative calculation similar to that given above shows that the $F_i(\pi(s))$ decrease for $i \in A_{j-1}$.

This perturbation may or may not decrease $\theta_0(\pi)$. But if $0 \not\in J_0$ then we may apply it in turn to the first $j$ in $J_0$, then the second $j$ in $J_0$, etc., until eventually $\theta_0(\pi)$ will decrease. This would be a contradiction, so it follows that $0 \not\in J_0$.

We may apply a similar argument to $a_j$ to prove the remaining conclusions.

We are now ready to prove the existence portion of Theorem 4, under our additional restriction that $n = K$. For $A \subset \eta$, let $\alpha_A = \frac{1}{w} \sum_{i \in A} w_i$ be the percentage of the total initial investment contributed by members of $A$.

**Lemma 11.** Fix $d(t)$ as well as the $x_i$ and $w_i$, and assume that each cohort consists of a single individual. For there to exist a choice of $\pi \in \mathcal{P}$ with $F_i(\pi) = 1 - \epsilon$ for each $i$, it is necessary and sufficient that for every $A \subset \eta$ with $\emptyset \neq A \neq \eta$ we have

$$
\int_0^\infty e^{-\pi t} d(t) \left( \prod_{i \notin A} iq_{x_i} \right) \left( 1 - \prod_{i \in A} iq_{x_i} \right) dt < \alpha_A (1 - \epsilon).
$$

We may think of (6) failing for one of two reasons – the presence of particularly elderly individuals, or of individuals who contribute a disproportionately large fraction of the initial
investment. In either case, we let \( A \) consist of the remaining people (younger, or investing less). Condition (6) will fail if either \( \prod_{i \notin A} t q_{x_i} \) is large (ie. the \( A^c \) individuals all die early into the tontine), or if \( \alpha_A \) is small (ie. the \( A^c \) individuals over-invest). A well-designed \( d(t) \) will attempt to mitigate these possibilities, though we have seen that this is not always possible. In particular, if we fix a choice of \( d(t) \) but allow the \( w_i \) to vary, there will always be a choice for the \( w_i \) that makes one of the \( \alpha_A \) small enough to force (6) to fail.

Proof. Assume (6), and let \( \pi \in \mathcal{P}_0 \) be very equitable. Let the \( a_J \) be as in Lemma 10. By Lemma 10 we have \( a_J \geq 1 - \varepsilon \), since a suitably weighted average of the \( F_i(\pi) \) equals \( 1 - \varepsilon \). Suppose that there is no equitable \( \pi \), or that the equitable \( \pi \) belongs to \( \mathcal{P}_0 \setminus \mathcal{P} \). Then \( J > 0 \), so \( 0 < |A_J| < n \). Moreover

\[
(1 - \varepsilon)\alpha_{A_J} \leq \sum_{i \in A_J} w_i F_i(\pi) = \int_0^\infty e^{-rt}d(t) \sum_{i \in A_J} E\left[ \frac{\pi_{A_J,i} w_i}{\sum_{k \in A_J} \pi_{A_J,k} w_k N_k(t)} 1\{T_{A_J} \leq t \} \right] dt \\
= \int_0^\infty e^{-rt}d(t)E\left[ \sum_{i \in A_J} \sum_{k \in A_J} \frac{\pi_{A_J,i} w_i N_i(t)}{\pi_{A_J,k} w_k N_k(t)} 1\{N_i(t) \neq 0, T_{A_J} \leq t \} \right] dt \\
= \int_0^\infty e^{-rt}d(t)P\left( \sum_{i} N_i(t) > 0, T_{A_J} \leq t \right) dt \\
= \int_0^\infty e^{-rt}d(t)\left( \prod_{i \notin A_J} t q_{x_i} \right)(1 - \prod_{i \in A_J} t q_{x_i}) dt,
\]

which violates (6).

Conversely, suppose \( \pi \in \mathcal{P} \) is equitable, and let \( A \subset \eta \) be non-empty and \( \neq \eta \). Then as above,

\[
(1 - \varepsilon)\alpha_A = \sum_{i \in A} w_i F_i(\pi) = \int_0^\infty e^{-rt}d(t)E\left[ \sum_{i \in A} \sum_{k \in \eta} \frac{\pi_i w_i N_i(t)}{\pi_k w_k N_k(t)} 1\{N_i(t) \neq 0 \} \right] dt \\
> \int_0^\infty e^{-rt}d(t)P\left( \sum_{i \in A} N_i(t) > 0, \sum_{i \in \eta \setminus A} N_i(t) = 0 \right) dt,
\]

which shows (6). \( \square \)

The problem with condition (6) of Lemma 11 is that it involves checking \( 2^n - 2 \) conditions. Condition (3) brings this down to a manageable number, provided we have a modest number of cohorts. We therefore now abandon the assumption that \( n = K \), in which case recall that \( \alpha_A \) once more denotes \( \frac{1}{w} \sum_{i \in A} n_i w_i \).
Proof of (b) of Theorem 4. Assume condition (3), which we may restate as
\[
\int_0^\infty e^{-rt} d(t) \prod_{i \in A} i q_{x_i}^{n_i} dt < \alpha_A(1 - \epsilon) + \epsilon.
\]
Condition (6) involves a general collection of subscribers, which we will take to consist of
\[0 \leq k_i \leq n_i\] individuals from the ith group, \(i = 1, \ldots, K\). Stated in this way it becomes that
\[
\int_0^\infty e^{-rt} d(t) \prod_{i=1}^K i q_{x_i}^{n_i-k_i} dt < (1 - \epsilon) \sum_{i=1}^K \frac{k_i w_i}{w} + \epsilon
\]
for every choice of \(0 \leq k_i \leq n_i\) (other than \((0, \ldots, 0)\) and \((n_1, \ldots, n_K)\)). Observe that the
left hand side of (9) is a concave function of each \(k_i\) (when the others are fixed), while the
right hand side varies in an affine way. Observe also that (8) is precisely (9) for the extreme
points of \(I = [0, n_1] \times \cdots \times [0, n_K]\) other than \((0, \ldots, 0)\) and \((n_1, \ldots, n_K)\). It is easily verified
that the same inequality holds at the latter two points as well, but with = rather than <.
This is enough to conclude that the strict inequality (9) holds at all points of \(I\) other than
\((0, \ldots, 0)\) and \((n_1, \ldots, n_2)\), so Lemma 11 applies to give an equitable \(\pi \in \mathcal{P}\).

Necessity also follows from Lemma 11, since (3) is a special case of (6). \(\square\)

We close with a comment on computations. Motivated by the proof of Theorem 4, we
calculate equitable participation rates by successively raising those \(\pi_i\) for which
\(F_i(\pi) < 1 - \epsilon\). We choose a relaxation rate \(\eta < 1\), and cycle through the \(i\), raising \(\pi_i\) by \(\eta\) as often as possible
while preserving the above criterion. Then repeat the process this time raising by \(\eta^2\), then
repeat again raising by \(\eta^3\), etc. For \(K = 3\) this appears to converge to 5 digits accuracy,
using fewer than 100 evaluations of the vector \(F(\pi)\).

Each equitable \((20, 40, 20)\) entry of Table 4 took approximately 4 hours running under
R, and involved 100 evaluations of the vector \(F(\pi)\). Each evaluation required computing
two integrals, using Simpson’s rule with 400 time steps. Each time step in turn called for a
binomial sum of \(n_1 n_2 n_3\) terms. With \(K = 2\) the computations are faster (25 integrations,
with each time step calling for summing \(n_1 n_2\) terms); for example the equitable \((500, 500)\)
entries in Table 3 each ran in just over 3 hours, despite accounting for an order of magnitude
more people. Note that our code could easily be optimized to run faster, eg. by using a
more efficient integration method, or by starting with coarser time steps and refining them
as the iteration proceeds. For utility computations, we do seem to need the level of accuracy
provided by 400 time steps, but the \(\pi's\) are less sensitive and could have been found quicker.